References with abstracts for QWIM project: Network analysis and clustering 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

1 Motivation for the project **2**
1.1 Beware of correlations which are not consistent with intuition . . . 2
1.2 Portfolio optimization . . . 2
1.3 Portfolio diversification . . . 2

2 Relevant references **3**
2.1 Main references . . . 3
2.2 Comprehensive list of references . . . 4
2.2.1 Clustering within context of QWIM . . . 4
2.2.2 Network analysis within context of QWIM . . . 6
2.2.3 Network analysis and clustering . . . 7
2.2.4 Testing and comparison procedures for investment portfolios . . . 9
2.2.5 Software implementations and frameworks . . . 10

References **12**

1 Motivation for the project

1.1 Beware of correlations which are not consistent with intuition

In many cases, advanced quantitative approaches (including machine learning) may not necessarily deliver better results in $\mathbb{QWIM}$, especially when decision making relies on data-based spurious correlations rather than on real causality.

Vigen (*Spurious Correlations*, 2019): many examples of such data-based spurious correlations

- Divorce rate in Maine has 99% correlation with per capita consumption of margarin
- Per capita consumption of mozzarella cheese has 96% correlation with civil engineering doctorates awarded in US
- Per capita consumption of chicken has 90% correlation with total US crude oil imports

Example in $\mathbb{QWIM}$ Laurinaityte et al. ("Elephants and the Cross-Section of Expected Returns," 2019): population growth of captive Asian elephants explains cross-section of expected returns of usually sorted portfolios with $R^2 = 0.91$ and $tStat = 2.93$ for market price of risk.

Question: Does it mean that number of captive elephants is the new outstanding factor in empirical asset pricing?

Answer: Likely it is an artifact due to data mining rather than a proper factor for factor-based investing.

1.2 Portfolio optimization

While there are many challenges in a portfolio optimization process relying on the corelation (or covariance) matrix, some of the most important issues (including potential lack of robustness and diversification) are due to the fact that correlation matrix lacks the notion of hierarchy.

It was shown that many complex systems can be arranged in a natural hierarchy comprising nested substructures, and financial markets are no exception. While a correlation matrix makes no differentiation between assets, some assets seem closer substitutes of one another, while others seem complementary to one another. This can be better handled through network analysis and clustering.

Networks enable practical usage of high / low centrality concepts

- significant interconnectedness risk (tail events propagate more quickly) due to assets with high centrality scores
- "peripheral assets" carry relatively less interconnectedness risk

Network-based and clustering-based portfolio optimization is likely to deliver more robust and diversified portfolios, and achieve better risk-adjusted performances compared to portfolios obtained using commonly used portfolio optimization techniques. Since no single clustering algorithm can be said to perform best on all datasets, different strategies must be tested and compared.

1.3 Portfolio diversification

Diversification is one of the most important concepts in the financial world. It is often said that diversification is the only free lunch in finance. From a qualitative point of view, the concept of diversification is quite clear: a portfolio is well-diversified if shocks in the individual components do not heavily impact on the overall portfolio. Relatively simple to understand then but profoundly difficult to define. Indeed, there is no broadly accepted precise and quantitative definition of diversification.

One of the most vexing problems in investment management is that diversification seems to disappear when investors need it the most. A key challenge in the construction of diversified multi-asset portfolio strategies is that even a seemingly well-balanced allocation to many asset classes can eventually translate into a portfolio with a very concentrated set of underlying risk exposures.

Network analysis applied to structure of investment portfolios is very beneficial to analyze diversification properties. We can also consider a portfolio selection approach that combines diversification and optimization.

2 Relevant references

2.1 Main references

List of references:

Ackerman et al. ("Weighted clustering: Towards solving the user's dilemma," 2021)

Alfarra et al. ("Rethinking Clustering for Robustness," 2021)

Akansu et al. ("Quant investing in cluster portfolios," 2021)

Avellaneda and Serur ("Hierarchical PCA and Modeling Asset Correlations," 2020)

Baitinger ("Forecasting asset returns with network-based metrics: A statistical and economic analysis," 2021)

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

Bnouachir and Mkhadri ("Efficient cluster-based portfolio optimization," 2021)

Chua et al. ("The Myth of Diversification," 2009)

Clemente et al. ("Smart network based portfolios," 2019)

Clemente et al. ("Asset allocation: new evidence through network approaches," 2021)

Coraggio and Coretto ("Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score," 2021)

Dees et al. ("Portfolio Cuts: A Graph-Theoretic Framework to Diversification," 2020)

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

Duarte and De Castro ("A Framework to Perform Asset Allocation Based on Partitional Clustering," 2020)

Dugué et al. ("Evaluating clustering quality using features salience: a promising approach," 2021)

Eidenvall ("Hierarchical Clustering To Improve Portfolio Tail Risk Characteristics," 2021)

Flint et al. ("Defining and measuring portfolio diversification," 2021)

Fu and Perry ("Estimating the Number of Clusters Using Cross-Validation," 2020)

Fusai et al. ("Equally Diversified or Equally Weighted?" 2020)

Giudici et al. ("Network models to improve robot advisory portfolios," 2022)

Guan and Loew ("A Distance-based Separability Measure for Internal Cluster Validation," 2021)

Guo et al. ("A Time-Varying Network for Cryptocurrencies," 2021)

Heckens and Guhr ("New collectivity measures for financial covariances and correlations," 2022)

Herteliu et al. ("Network analysis of pension funds investments," 2021)

Horvath et al. ("Clustering Market Regimes Using the Wasserstein Distance," 2021)

Huang et al. ("Financial risk propagation between Chinese and American stock markets based on multilayer networks," 2022)

Jaeger et al. ("Understanding machine learning for diversified portfolio construction by explainable AI," 2020)

Jaeger et al. ("Interpretable Machine Learning for Diversified Portfolio Construction," 2021)

Jaeger et al. ("Adaptive Seriational Risk Parity and other Extensions for Heuristic Portfolio Construction using Machine Learning and Graph Theory," 2021)

Katsouris ("Optimal Portfolio Choice and Stock Centrality for Tail Risk Events," 2021)

Kawamoto and Kabashima ("Cross-validation estimate of the number of clusters in a network," 2017)

Kaya ("Eccentricity in Asset Management," 2015)

Kaya ("Managing ambiguity in asset allocation," 2017)

Kinlaw et al. ("The Myth of Diversification Reconsidered," 2021)

Konstantinov et al. ("A network and machine learning approach to factor, asset, and blended allocation," 2020)

Koumou ("Diversification and portfolio theory: a review," 2020)

Kurtti ("How many stocks make a diversified portfolio in a continuous-time world?" 2020)

Laur ("Portfolio Optimization - Can Optimizing Portfolio Outperform Naive Diversification?" 2020)

Leon et al. ("Clustering algorithms for Risk-Adjusted Portfolio Construction," 2017)

Lim and Ong ("Portfolio Diversification Using Shape-Based Clustering," 2021)

Lohre et al. ("Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations," 2020)

Lopez de Prado ("Building Diversified Portfolios that Outperform Out of Sample," 2016)

Lopez de Prado ("Estimation of Theory-Implied Correlation Matrices," 2019)

Lopez de Prado ("Clustering," 2020)

Lopez de Prado (*Machine learning for asset managers*, 2020)

Lu et al. ("A Cluster Representative Selection Method for Stock Portfolio Based on Efficient Frontier," 2021)
Marti et al. ("A review of two decades of correlations, hierarchies, networks and clustering in financial markets," 2021)
Massahi et al. ("Development of an efficient cluster-based portfolio optimization model under realistic market conditions," 2020)
Millington and Niranjan ("Stability and similarity in financial networks – How do they change in times of turbulence?" 2021)
Molyboga ("A Modified Hierarchical Risk Parity Framework for Portfolio Management," 2020)
Olmo ("Optimal portfolio allocation and asset centrality revisited," 2021)
Page and Panariello ("When Diversification Fails," 2018)
Papenbrock et al. ("Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios," 2021)
Papenbrock et al. ("Can Adaptive Seriational Risk Parity Tame Crypto Portfolios?" 2021)
Parmentier ("Measures of Portfolio Diversification," 2018)
Peralta and Zareei ("A network approach to portfolio selection," 2016)
Raffinot ("Hierarchical Clustering-Based Asset Allocation," 2017)
Raffinot ("The Hierarchical Equal Risk Contribution Portfolio," 2018)
Roncalli ("Advanced Course in Asset Management," 2021)
Sakurai et al. ("Correlation diversified passive portfolio strategy based on permutation of assets," 2021)
Sass and Thos ("Risk reduction and portfolio optimization using clustering methods," 2022)
Scherer ("Adding alternative assets: return enhancement, diversification or hedging?" 2021)
Schwendner et al. ("Adaptive Seriational Risk Parity and Other Extensions for Heuristic Portfolio Construction Using Machine Learning and Graph Theory," 2021)
Serur and Avellaneda ("Hierarchical PCA and Modeling Asset Correlations," 2021)
Shirota and Murakami ("Long-term Time Series Data Clustering of Stock Prices for Portfolio Selection," 2021)
Snow ("Machine Learning in Asset Management - Part 2: Portfolio Construction - Weight Optimization," 2020)
Swedroe ("The Importance of Diversification in Achieving Long-Term Goals," 2020)
Tang et al. ("Asset Selection via Correlation Blockmodel Clustering," 2021)
Thiagarajan et al. ("Financial Globalization and Its Implications for Diversification of Portfolio Risk," 2021)
Thrun and Stier ("Fundamental clustering algorithms suite," 2021)
Vankwikelberge et al. ("ExClus: Explainable Clustering on Low-dimensional Data Representations," 2021)
Vyrost et al. ("Network-based asset allocation strategies," 2019)
Wang and Aste ("Dynamic Portfolio Optimization with Inverse Covariance Clustering," 2022)
Yang et al. ("Portfolio optimization based on empirical mode decomposition," 2019)
Zaimovic et al. ("How Many Stocks Are Sufficient for Equity Portfolio Diversification? A Review of the Literature," 2021)
Zhan et al. ("Graphical Models for Financial Time Series and Portfolio Selection," 2021)
Zhao et al. ("Stock market as temporal network," 2018)
Zhao et al. ("Robust portfolio rebalancing with cardinality and diversification constraints," 2021)

2.2 Comprehensive list of references

2.2.1 Clustering within context of QWIM

List of references:

Akansu et al. ("Quant investing in cluster portfolios," 2021)
Alokley and Albarrak ("Clustering of Extremes in Financial Returns: A Study of Developed and Emerging Markets," 2020)
Avellaneda and Serur ("Hierarchical PCA and Modeling Asset Correlations," 2020)
Begusic and Kostanjcar ("Cluster-Based Shrinkage of Correlation Matrices for Portfolio Optimization," 2019)
Bennett et al. ("Lead-lag detection and network clustering for multivariate time series with an application to the US equity market," 2022)
Bnouachir and Mkhadri ("Efficient cluster-based portfolio optimization," 2021)
Cai et al. ("Clustering Approaches for Financial Data Analysis: a Survey," 2016)
Cajas ("Robust Portfolio Selection with Near Optimal Centering," 2019)

Custodio João et al. ("Clustering Dynamics and Persistence for Financial Multivariate Panel Data," 2021)
Duarte and De Castro ("A Framework to Perform Asset Allocation Based on Partitional Clustering," 2020)
Dugué et al. ("Evaluating clustering quality using features salience: a promising approach," 2021)
Eidenvall ("Hierarchical Clustering To Improve Portfolio Tail Risk Characteristics," 2021)
Emerson ("Machine learning for financial applications: self-organising maps, hierarchical clustering and dynamic time-warping for portfolio constructive," 2019)
Garvey and Madhavan ("Reconstructing Emerging and Developed Markets Using Hierarchical Clustering," 2019)
Guan and Loew ("A Distance-based Separability Measure for Internal Cluster Validation," 2021)
Gubu et al. ("Robust mean-variance portfolio selection with time series clustering," 2021)
Gupta and Chatterjee ("Financial Time Series Clustering," 2018)
Han and Ge ("Effect of dimensionality reduction on stock selection with cluster analysis in different market situations," 2020)
Horvath et al. ("Clustering Market Regimes Using the Wasserstein Distance," 2021)
Heckens and Guhr ("New collectivity measures for financial covariances and correlations," 2022)
Jaeger et al. ("Interpretable Machine Learning for Diversified Portfolio Construction," 2021)
Jaeger et al. ("Adaptive Seriational Risk Parity and other Extensions for Heuristic Portfolio Construction using Machine Learning and Graph Theory," 2021)
Jain and Jain ("Can Machine Learning-Based Portfolios Outperform Traditional Risk-Based Portfolios? The Need to Account for Covariance Misspecification," 2019)
Kolrep et al. ("Economic Versus Statistical Clustering in Multi-Asset Multi-Factor Strategies," 2020)
Lee and Seregina ("Optimal Portfolio Using Factor Graphical Lasso," 2022)
Leon et al. ("Clustering algorithms for Risk-Adjusted Portfolio Construction," 2017)
Lim and Ong ("Portfolio Diversification Using Shape-Based Clustering," 2021)
Lohre et al. ("Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations," 2020)
Lopez de Prado ("Clustering," 2020)
Lopez de Prado ("Building Diversified Portfolios that Outperform Out of Sample," 2016)
Lu et al. ("A Cluster Representative Selection Method for Stock Portfolio Based on Efficient Frontier," 2021)
Mattera et al. ("Distribution-Based Entropy Weighting Clustering of Skewed and Heavy Tailed Time Series," 2021)
Marti et al. ("A review of two decades of correlations, hierarchies, networks and clustering in financial markets," 2021)
Millington and Niranjan ("Partial correlation financial networks," 2020)
Molyboga ("A Modified Hierarchical Risk Parity Framework for Portfolio Management," 2020)
Murialdo et al. ("Inferring multi-period optimal portfolios via detrending moving average cluster entropy," 2021)
Nanakorn and Palmgren ("Hierarchical Clustering in Risk-Based Portfolio Construction," 2021)
Naraoka et al. ("Detecting and explaining changes in various assets' relationships in financial markets," 2020)
Papenbrock et al. ("Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios," 2021)
Papenbrock et al. ("Can Adaptive Seriational Risk Parity Tame Crypto Portfolios?" 2021)
Park ("Clustering Approaches for Global Minimum Variance Portfolio," 2020)
Pharasi et al. ("Dynamics of the market states in the space of correlation matrices with applications to financial markets," 2021)
Poletaev and Spiridonova ("Hierarchical Clustering as a Dimension Reduction Technique for Markowitz Portfolio Optimization," 2020)
Puerto et al. ("Clustering and portfolio selection problems: A unified framework," 2020)
Putra et al. ("Portfolio Selection of KOMPAS-100 Stocks Index Using B-Spline Based Clustering," 2021)
Raffinot ("Hierarchical Clustering-Based Asset Allocation," 2017)
Raffinot ("The Hierarchical Equal Risk Contribution Portfolio," 2018)
Sass and Thos ("Risk reduction and portfolio optimization using clustering methods," 2022)
Schwendner et al. ("Adaptive Seriational Risk Parity and Other Extensions for Heuristic Portfolio Construction Using Machine Learning and Graph Theory," 2021)
Serur and Avellaneda ("Hierarchical PCA and Modeling Asset Correlations," 2021)

Sjostrand and Behnejad ("Exploration of Hierarchical Clustering in Long-only Risk-based Portfolio Optimization," 2020)
Snow ("Machine Learning in Asset Management - Part 2: Portfolio Construction - Weight Optimization," 2020)
Tang et al. ("Asset Selection via Correlation Blockmodel Clustering," 2021)
Tola et al. ("Cluster analysis for portfolio optimization," 2008)
Turner ("Graph Auto-Encoders for Financial Clustering," 2021)
Vojtko and Cisár ("An Analysis of Volatility Clustering of Equity Factor Strategies," 2021)

2.2.2 Network analysis within context of QWIM

List of references:
Ahelegbey and Giudici ("Market Risk, Connectedness and Turbulence: A Comparison of 21st Century Financial Crises," 2020)
Baitinger ("Forecasting asset returns with network-based metrics: A statistical and economic analysis," 2021)
Baitinger and Maier ("The (Mis)Behavior of Hedge Fund Strategies: A Network-Based Analysis," 2019)
Baitinger and Papenbrock ("Interconnectedness Risk and Active Portfolio Management," 2017)
Baitinger and Flegel ("New Concepts in Financial Forecasting: Network-Based Information, Topological Data Analysis and their Combination," 2021)
Bardoscia et al. ("The Physics of Financial Networks," 2021)
Barunik and Ellington ("Dynamic Networks in Large Financial and Economic Systems," 2021)
Castilho et al. ("Forecasting Financial Market Structure from Network Features using Machine Learning," 2021)
Chen et al. ("Constructing a multilayer network for stock market," 2020)
Clemente et al. ("Smart network based portfolios," 2019)
Clemente et al. ("Asset allocation: new evidence through network approaches," 2021)
de Carvalho and Gupta ("A network approach to unravel asset price comovement using minimal dependence structure," 2018)
de Miranda Cardoso et al. ("Algorithms for Learning Graphs in Financial Markets," 2020)
Dey et al. ("Community detection in complex networks: From statistical foundations to data science applications," 2021)
Di Cerbo and Taylor ("Graph theoretical representations of equity indices and their centrality measures," 2021)
Duan ("Predicting with Structured Data: Graphs, Ranks, and Time Series," 2021)
Escanciano and Hualde ("Measuring Asset Market Linkages: Nonlinear Dependence and Tail Risk," 2021)
Giudici et al. ("Network Models to Enhance Automated Cryptocurrency Portfolio Management," 2020)
Giudici et al. ("Network models to improve robot advisory portfolios," 2022)
Jackson and Pernoud ("Systemic Risk in Financial Networks: A Survey," 2020)
Jiang et al. ("Tail-event driven network of cryptocurrencies and conventional assets," 2022)
Kalyagin et al. ("Reliability of MST identification in correlation-based market networks," 2021)
Katsouris ("Optimal Portfolio Choice and Stock Centrality for Tail Risk Events," 2021)
Kaya ("Eccentricity in Asset Management," 2015)
Kaya ("Managing ambiguity in asset allocation," 2017)
Loistl and Konstantinov ("Interactions and Interconnectedness Shape Financial Market Research," 2020)
Konstantinov et al. ("A network and machine learning approach to factor, asset, and blended allocation," 2020)
Konstantinov and Rusev ("The Bond-Equity-Fund Relation Using the Fama-French-Carhart Factors: A Practical Network Approach," 2020)
Konstantinov and Simonian ("A Network Approach to Analyzing Hedge Fund Connectivity," 2020)
Kukreti et al. ("A Perspective on Correlation-Based Financial Networks and Entropy Measures," 2020)
Kumar et al. ("Ripples on financial networks," 2021)
Magner et al. ("The Volatility Forecasting Power of Financial Network Analysis," 2021)
Marti et al. ("A review of two decades of correlations, hierarchies, networks and clustering in financial markets," 2021)
Millington and Niranjan ("Stability and similarity in financial networks – How do they change in times of turbulence?" 2021)
Olmo ("Optimal portfolio allocation and asset centrality revisited," 2021)
Pang et al. ("An analysis of network filtering methods to sovereign bond yields during COVID-19," 2021)

Peralta and Zareei ("A network approach to portfolio selection," 2016)
Samal et al. ("Network-centric indicators for fragility in global financial indices," 2021)
Son and Lee ("Graph-based multi-factor asset pricing model," 2022)
Stavroglou ("Finding Hidden Structures in Financial Networks," 2020)
Vyrost et al. ("Network-based asset allocation strategies," 2019)
Yang et al. ("Portfolio optimization with idiosyncratic and systemic risks for financial networks," 2021)
Yang et al. ("Portfolio optimization based on empirical mode decomposition," 2019)
Zhan et al. ("Graphical Models for Financial Time Series and Portfolio Selection," 2021)
Zhao et al. ("Stock market as temporal network," 2018)
Zhao et al. ("Community detection and portfolio optimization," 2021)

2.2.3 Network analysis and clustering

List of references:
Abboud et al. ("Subquadratic High-Dimensional Hierarchical Clustering," 2019)
Ackerman et al. ("Weighted clustering: Towards solving the user's dilemma," 2021)
Adolfsson et al. ("To cluster, or not to cluster: An analysis of clusterability methods," 2019)
Alfarra et al. ("Rethinking Clustering for Robustness," 2021)
Bandara et al. ("Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," 2020)
Bouveyron et al. (*Model-Based Clustering and Classification for Data Science: With Applications in R*, 2019)
Brécheteau et al. ("Robust Bregman clustering," 2021)
Campello et al. ("Density-based clustering," 2020)
Casa et al. ("Better than the best? Answers via model ensemble in density-based clustering," 2021)
Chavent et al. ("Combining clustering of variables and feature selection using random forests," 2021)
Chehreghani ("Shift of Pairwise Similarities for Data Clustering," 2021)
Chung et al. ("Statistical Connectomics," 2022)
Coraggio and Coretto ("Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score," 2021)
Dalmia and Sia ("Clustering with UMAP: Why and How Connectivity Matters," 2021)
De Luca and Zuccolotto ("Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach," 2021)
Den Teuling et al. ("Clustering of longitudinal data: A tutorial on a variety of approaches," 2021)
Djouzi and Beghdad-Bey ("A Review of Clustering Algorithms for Big Data," 2019)
Doreian et al. (*Advances in Network Clustering and Blockmodeling*, 2020)
Duan and Dunson ("Bayesian Distance Clustering," 2021)
Elliott et al. ("Core–periphery structure in directed networks," 2020)
Esmaeili et al. ("Probabilistic Fair Clustering," 2021)
Exarchakis et al. ("A sampling-based approach for efficient clustering in large datasets," 2022)
Ezhilmaran and Indira ("A survey on clustering techniques in pattern recognition," 2020)
Faletto and Bien ("Cluster Stability Selection," 2022)
Feng and Simon ("Ensembled sparse-input hierarchical networks for high-dimensional datasets," 2020)
Franti and Sieranoja ("How much can k-means be improved by using better initialization and repeats?" 2019)
Fu and Perry ("Estimating the Number of Clusters Using Cross-Validation," 2020)
Gagolewski ("genieclust: Fast and robust hierarchical clustering," 2021)
Gao and Tsay ("Divide-and-Conquer: A Distributed Hierarchical Factor Approach to Modeling Large-Scale Time Series Data," 2021)
Gherbaoui et al. ("Generation of Gaussian sets for clustering methods assessment," 2021)
Ghosal et al. ("A Short Review on Different Clustering Techniques and Their Applications," 2019)
Goel and Majumdar ("Transformed K-means Clustering," 2021)
Grun ("Model-based Clustering," 2018)
Guan and Loew ("A Distance-based Separability Measure for Internal Cluster Validation," 2021)
Guijo-Rubio et al. ("Time-Series Clustering Based on the Characterization of Segment Typologies," 2020)
Hua ("Clusterability, Model Selection and Evaluation," 2019)

Huang et al. ("Learning From Networks: Algorithms, Theory, and Applications," 2019)
Irani et al. ("Clustering Techniques and the Similarity Measures used in Clustering: A Survey," 2016)
Javed et al. ("A Benchmark Study on Time Series Clustering," 2020)
Jose-Garcia and Gomez-Flores ("A survey of cluster validity indices for automatic data clustering using differential evolution," 2021)
Kawamoto and Kabashima ("Comparative analysis on the selection of number of clusters in community detection," 2017)
Kawamoto and Kabashima ("Cross-validation estimate of the number of clusters in a network," 2017)
Keranovic et al. ("Estimating the Number of Latent Factors in High-Dimensional Financial Time Series," 2020)
Kumari and Sharma ("A review for the efficient clustering based on distance and the calculation of centroid," 2020)
Landi et al. ("reval: a Python package to determine best clustering solutions with stability-based relative clustering validation," 2020)
Lemenkova ("R Libraries (dendextend and magrittr) and Clustering Package scipy.cluster of Python For Modelling Diagrams of Dendrogram Trees," 2020)
Leopold and Rose ("UNIC: A fast nonparametric clustering," 2020)
Li et al. ("Network cross-validation by edge sampling," 2020)
Li and Liu ("Multivariate time series clustering based on complex network," 2021)
Lipor and Balzano ("Clustering quality metrics for subspace clustering," 2020)
Ma (*Network Analysis Made Simple*, 2020)
Ma and Seth (*Network Analysis Made Simple*, 2020)
Maharaj et al. (*Time Series Clustering and Classification*, 2019)
Mahfuz et al. ("Review of single clustering methods," 2019)
Mehta et al. ("Analytical review of clustering techniques and proximity measures," 2020)
Millington and Niranjan ("Construction of Minimum Spanning Trees from Financial Returns using Rank Correlation," 2020)
Peng et al. ("Multi-dimensional clustering through fusion of high-order similarities," 2022)
Pimentel and de Carvalho ("A Meta-learning approach for recommending the number of clusters for clustering algorithms," 2020)
Policastro et al. ("ROBustness In Network (robin): an R package for Comparison and Validation of communities," 2021)
Rahgoshay and Salavatipour ("Hierarchical Clustering: New Bounds and Objective," 2021)
Rehman and Belhaouari ("Divide well to merge better: A novel clustering algorithm," 2022)
Romashchenko ("Clustering with Respect to the Information Distance," 2021)
Sarda-Espinosa ("Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package," 2019)
Sato-Ilic ("Cluster-scaled principal component analysis," 2021)
Seabrook et al. ("Evaluating structural edge importance in temporal networks," 2021)
Silva et al. ("Novel Features for Time Series Analysis: A Complex Networks Approach," 2021)
Silva et al. ("Time series analysis via network science: Concepts and algorithms," 2021)
Sobczyk et al. ("VARCLUST: clustering variables using dimensionality reduction," 2020)
Stankovic et al. ("Data Analytics on Graphs Part I: Graphs, Graph Spectra, and Spectral Clustering," 2020)
Stankovic et al. ("Data Analytics on Graphs Part II: Signals on Graphs," 2020)
Stankovic et al. ("Data Analytics on Graphs Part III: Machine Learning on Graphs, from Graph Topology to Applications," 2020)
Tong et al. ("A density-peak-based clustering algorithm of automatically determining the number of clusters," 2021)
Thrun ("The Exploitation of Distance Distributions for Clustering," 2021)
Thrun and Stier ("Fundamental clustering algorithms suite," 2021)
Underwood et al. ("Motif-based spectral clustering of weighted directed networks," 2020)
Vankwikelberge et al. ("ExClus: Explainable Clustering on Low-dimensional Data Representations," 2021)
Vázquez et al. ("An ensemble solution for multivariate time series clustering," 2021)
Wadhwa and Scott ("Exploring complex networks with the ICON R package," 2020)
Wang and Tsay ("Clustering Multiple Time Series with Structural Breaks," 2019)

Wang et al. ("On the Efficiency of K-Means Clustering: Evaluation, Optimization, and Algorithm Selection," 2020)

Wu et al. ("Analyses and applications of optimization methods for complex network reconstruction," 2020)

Wu et al. ("Effective hierarchical clustering based on structural similarities in nearest neighbor graphs," 2021)

Yelibi and Gebbie ("Agglomerative Likelihood Clustering," 2021)

Zambelli ("Ensemble Method for Cluster Number Determination and Algorithm Selection in Unsupervised Learning," 2021)

Zhang ("Weighted Clustering Ensemble: A Review," 2021)

Zheng et al. ("Multi-view subspace clustering networks with local and global graph information," 2021)

Zhong et al. ("Ensemble clustering based on evidence extracted from the co-association matrix," 2019)

Zhong et al. ("Ensemble clustering based on evidence extracted from the co-association matrix," 2019)

Zhong and Pun ("Subspace clustering by simultaneously feature selection and similarity learning," 2020)

Zhou et al. ("Unsupervised feature selection for balanced clustering," 2020)

2.2.4 Testing and comparison procedures for investment portfolios

References:

Adcock et al. ("Portfolio Performance Measurement: Monotonicity with Respect to the Sharpe Ratio and Multivariate Tests of Correlation," 2017)

Arnott et al. ("A backtesting protocol in the era of machine learning," 2019)

Bailey et al. ("Stock Portfolio Design and Backtest Overfitting," 2017)

Bessler and Wolff ("Portfolio Optimization with Industry Return Prediction Models," 2017)

Bessler et al. ("Multi-asset portfolio optimization and out-of-sample performance: an evaluation of Black Litterman, mean-variance, and naive diversification approaches," 2017)

Bjerring et al. ("Feature selection for portfolio optimization," 2017)

Bruni et al. ("On exact and approximate stochastic dominance strategies for portfolio selection," 2017)

Bruni et al. ("Real-world datasets for portfolio selection and solutions of some stochastic dominance portfolio models," 2016)

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

Cesarone et al. ("On the stability of portfolio selection models," 2019)

Cesarone et al. ("Why Small Portfolios Are Preferable and How to Choose Them," 2018)

Chaudhuri and Lo ("Dynamic Alpha: A Spectral Decomposition of Investment Performance Across Time Horizons," 2019)

Diris et al. ("Long-Term Strategic Asset Allocation: An Out-of-Sample Evaluation," 2015)

Fabozzi and Lopez de Prado ("Being Honest in Backtest Reporting: A Template for Disclosing Multiple Tests," 2018)

Greiner and Stoyanov ("Portfolio scoring by expected risk premium," 2019)

Guidolin et al. ("Portfolio performance of linear SDF models: an out-of-sample assessment," 2018)

Guo ("A Statistical Response to Challenges in Vast Portfolio Selection," 2019)

Guo et al. ("When Does The 1/N Rule Work?" 2019)

Haley ("K-fold cross validation performance comparisons of six naive portfolio selection rules: how naive can you be and still have successful out-of-sample portfolio performance?" 2017)

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

Hens et al. ("Escaping the backtesting illusion," 2020)

Hsu et al. (*Do Cross-Sectional Stock Return Predictors Pass the Test without Data-Snooping Bias?* 2017)

Hsu et al. ("Asset allocation strategies, data snooping, and the 1 / N rule," 2018)

Huang and Yu ("A new procedure for resampled portfolio with shrinkaged covariance matrix," 2020)

Hwang et al. ("Naive versus optimal diversification: Tail risk and performance," 2018)

Ielpo et al. (*Engineering Investment Process: Making Value Creation Repeatable*, 2017)

Jaeger et al. ("Understanding machine learning for diversified portfolio construction by explainable AI," 2020)

Kazak and Pohlmeier ("Testing out-of-sample portfolio performance," 2019)

Kazak and Pohlmeier (*Portfolio Pretesting with Machine Learning*, 2020)

Kuntz ("Portfolio Strategies with Classical and Alternative Benchmarks," 2018)

Lohre et al. ("Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations," 2020)

Lopez de Prado ("A Data Science Solution to the Multiple-Testing Crisis in Financial Research," 2019)

Lopez de Prado and Lewis ("Detection of false investment strategies using unsupervised learning methods," 2019)

Malavasi et al. ("Second order of stochastic dominance efficiency vs mean variance efficiency," 2021)

Mooney et al. ("Dynamic Regime Strategy for Stress Testing and Optimizing Institutional Investor Portfolios," 2020)

Platanakis et al. ("Horses for Courses: Mean-Variance for Asset Allocation and 1/N for Stock Selection," 2021)

Radovanov and Marcikic ("Testing The Performance Of The Investment Portfolio Using Block Bootstrap Method," 2014)

Rebonato ("A financially justifiable and practically implementable approach to coherent stress testing," 2019)

Schumann ("Backtesting," 2019)

Seymour et al. ("Dynamic portfolio management strategies: A framework for historical analysis," 2018)

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

Taljaard and Maré ("Why has the equal weight portfolio underperformed and what can we do about it?" 2021)

Tayali ("A novel backtesting methodology for clustering in mean–variance portfolio optimization," 2020)

Traccucci et al. ("A Triptych Approach for Reverse Stress Testing of Complex Portfolios," 2019)

Valentine et al. ("Beyond p values: utilizing multiple methods to evaluate evidence," 2019)

Vincent et al. ("Analyzing the Performance of Multifactor Investment Strategies under a Multiple Testing Framework," 2018)

Vovk and Wang ("True and false discoveries with e-values," 2020)

Vovk and Wang ("E-values: Calibration, combination, and applications," 2021)

Wiecki et al. ("All That Glitters Is Not Gold: Comparing Backtest and Out-of-Sample Performance on a Large Cohort of Trading Algorithms," 2016)

Yu ("Comparing Classical Portfolio Optimization and Robust Portfolio Optimization on Black Swan Events," 2021)

Yuan and Zhou ("Why Naive 1/N Diversification Is Not So Naive, and How to Beat It?" 2022)

Zhang et al. ("DoubleEnsemble: A New Ensemble Method Based on Sample Reweighting and Feature Selection for Financial Data Analysis," 2020)

Zhang et al. ("Information Coefficient as a Performance Measure of Stock Selection Models," 2020)

Zhang et al. ("Deep Learning for Portfolio Optimization," 2020)

2.2.5 Software implementations and frameworks

List of references:

Bonald et al. ("Scikit-network: Graph Analysis in Python," 2020)

Charrad et al. ("NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," 2014)

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

Ferraro et al. ("fclust: An R Package for Fuzzy Clustering," 2019)

Fischer et al. ("REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit," 2021)

Fop and Murphy ("Variable Selection Methods for Model-based Clustering," 2017)

Gagolewski ("genieclust: Fast and robust hierarchical clustering," 2021)

Haddad and Bouguessa ("TopoDetect: Framework for Topological Features Detection in Graph Embeddings," 2021)

Landi et al. ("reval: a Python package to determine best clustering solutions with stability-based relative clustering validation," 2020)

Lemenkova ("R Libraries (dendextend and magrittr) and Clustering Package scipy.cluster of Python For Modelling Diagrams of Dendrogram Trees," 2020)

Louiset et al. ("UCSL : A Machine Learning Expectation-Maximization framework for Unsupervised Clustering driven by Supervised Learning," 2021)

McCabe et al. ("netrd: A library for network reconstruction and graph distances," 2020)

Miranda et al. ("HiClass: a Python library for local hierarchical classification compatible with scikit-learn," 2022)
Montero and Vilar ("TSclust: An R Package for Time Series Clustering," 2015)
Mori et al. ("Distance Measures for Time Series in R: The TSdist Package," 2016)
Rusch et al. ("Cluster Optimized Proximity Scaling," 2021)
Ruta et al. ("SAX Navigator: Time Series Exploration through Hierarchical Clustering," 2020)
Sarda-Espinosa ("Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package," 2019)
Sekula et al. ("optCluster: An R Package for Determining the Optimal Clustering Algorithm," 2017)
Sobczyk et al. ("VARCLUST: clustering variables using dimensionality reduction," 2020)
Tellaroli et al. ("Cross-Clustering: A Partial Clustering Algorithm with Automatic Estimation of the Number of Clusters," 2016)
Valk and Cybis ("U-Statistical Inference for Hierarchical Clustering," 2021)
Wang et al. ("Thresher: determining the number of clusters while removing outliers.," 2018)
Weylandt et al. ("Dynamic Visualization and Fast Computation for Convex Clustering via Algorithmic Regularization," 2019)
Yu et al. ("Bootstrapping estimates of stability for clusters, observations and model selection," 2018)

References

Abboud, A., Cohen-Addad, V., and Houdrouge, H. (2019). "Subquadratic High-Dimensional Hierarchical Clustering." In: *Advances in Neural Information Processing Systems 32 (NIPS 2019)*.
We consider the widely-used average-linkage, single-linkage, and Ward's methods for computing hierarchical clusterings of high-dimensional Euclidean inputs. It is easy to show that there is no efficient implementation of these algorithms in high dimensional Euclidean space since it implicitly requires to solve the closest pair problem, a notoriously difficult problem. However, how fast can these algorithms be implemented if we allow approximation? More precisely: these algorithms successively merge the clusters that are at closest average (for average-linkage), minimum distance (for single-linkage), or inducing the least sum-of-square error (for Ward's). We ask whether one could obtain a significant running-time improvement if the algorithm can merge -approximate closest clusters (namely, clusters that are at distance (average, minimum, or sum-of-square error) at most times the distance of the closest clusters). We show that one can indeed take advantage of the relaxation and compute the approximate hierarchical clustering tree using approximate nearest neighbor queries. This leads to an algorithm running in time for dimensional Euclidean space. We then provide experiments showing that these algorithms perform as well as the non-approximate version for classic classification tasks while achieving a significant speedup.

Ackerman, M., Ben-David, S., Brânzei, S., and Loker, D. (2021). "Weighted clustering: Towards solving the user's dilemma." In: *Pattern Recognition* 120, p. 108152.
This paper makes a major step towards addressing a long-standing challenge in cluster analysis, known as the user's dilemma, which is the problem of selecting an appropriate clustering algorithm for a specific task. A formal approach for addressing this challenge relies on the identification of succinct, user-friendly properties that capture formal differences amongst clustering techniques. While helpful for gaining insight into the nature of clustering paradigms, there is a theory-practice gap that has so far limited the utility of this approach: Formal properties typically highlight advantages of classical linkage-based algorithms, while practical experience shows that center-based methods are preferable for many applications. We present simple new properties that delineate core differences between common clustering paradigms and overcome this theory-practice gap. The properties we present give a formal understanding of the advantages of center-based approaches for some applications and insight into when different clustering paradigms should be used. These properties address how sensitive algorithms are to changes in element frequencies, which we capture in a generalized setting where every element is associated with a real-valued weight. To complement extensive formal analysis, we discuss how these properties can be applied in practice.

Adcock, C., Areal, N., Armada, M. R., Cortez, M. C., Oliveira, B., and Silva, F. (2017). "Portfolio Performance Measurement: Monotonicity with Respect to the Sharpe Ratio and Multivariate Tests of Correlation." In: *SSRN e-Print*.
This paper reports an investigation into methods of portfolio performance measurement. The work is motivated first by equivocal empirical evidence reported by several authors about the correlation of performance measures with the Sharpe ratio. Secondly it is motivated by recent work which specifies that performance measures will be monotone functions of the Sharpe ratio if portfolio returns follow the same location-scale distribution. The paper demonstrates that the class of location-scale distributions is broader than previously reported. It presents conditions under which monotonicity with respect to the Sharpe ratio will fail. The paper shows that for large sample sizes the correlation between pairs of performance measures that are functions of the Sharpe ratio is unity. The correct null hypothesis for tests of correlation is therefore rho=1. Two multivariate tests of this null hypothesis are presented. The new tests are used to carry out of a comprehensive study of performance measurement for a set over ninety UK investment trusts.

Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). "To cluster, or not to cluster: An analysis of clusterability methods." In: *Pattern Recognition* 88, pp. 13–26.
Abstract Clustering is an essential data mining tool that aims to discover inherent cluster structure in data. For most applications, applying clustering is only appropriate when cluster structure is present. As such, the study of clusterability, which evaluates whether data possesses such structure, is an integral part of cluster analysis. However, methods for evaluating clusterability vary radically, making it challenging to select a suitable measure. In this paper, we perform an extensive comparison of measures of clusterability and provide guidelines that clustering users can reference to select suitable measures for their applications.

Ahelegbey, D. F. and Giudici, P. (2020). "Market Risk, Connectedness and Turbulence: A Comparison of 21st Century Financial Crises." In: *SSRN e-Print*.

We construct a network-based turbulence score that proves useful for analyzing the relationship between financial interconnectedness, and global market risk, and for identifying systemically important markets, with the highest contribution to financial turbulence. We apply our measure to study the integration among the major stock markets over the first two decades of the 21st century, particularly during the tech, sub-prime, and ongoing COVID-19 crises. The result shows that the interconnectedness of the markets amplifies initial global market risks (on average almost four times), to cause financial turbulence. We also found evidence that the United States is central to global market turbulence, followed by Brazil, France, Hong Kong, and Germany.

Akansu, A., Avellaneda, M., and Xiong, A. (2021). "Quant investing in cluster portfolios." In: *The Journal of Investment Strategies* 9(4), pp. 61–78.
This paper discusses portfolio construction for investing in N given assets, eg, constituents of the Dow Jones Industrial Average (DJIA) or large cap stocks, based on partitioning the investment universe into clusters. The clusters are determined from the trailing correlation matrix via an information theoretic algorithm that uses thresholding of high-correlation pairs. We calculate the principal eigenvector of each cluster from its correlation matrix and the corresponding eigenportfolio. The cluster portfolios are combined into a single N-asset portfolio based on a weighting scheme for the clusters. Various tests conducted on components of the DJIA and a 30-stock basket of large cap stocks indicate that the new portfolios are superior to the DJIA and other mean-variance portfolios in terms of their risk-adjusted returns from 2009 to 2019. We also tested the cluster portfolios for a larger basket of 373 Standard & Poor's 500 components from 2001 to 2019. The test results provide convincing evidence that a cluster-based portfolio can outperform passive investing.

Alfarra, M., Perez, J. C., Bibi, A., Thabet, A., Arbelaez, P., and Ghanem, B. (2021). "Rethinking Clustering for Robustness." In: *arXiv e-Print*.
This paper studies how encouraging semantically-aligned features during deep neural network training can increase network robustness. Recent works observed that Adversarial Training leads to robust models, whose learnt features appear to correlate with human perception. Inspired by this connection from robustness to semantics, we study the complementary connection: from semantics to robustness. To do so, we provide a robustness certificate for distance-based classification models (clustering-based classifiers). Moreover, we show that this certificate is tight, and we leverage it to propose ClusTR (Clustering Training for Robustness), a clustering-based and adversary-free training framework to learn robust models. Interestingly, *ClusTR* outperforms adversarially-trained networks by up to 4% under strong PGD attacks.

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.

Arnott, R. D., Harvey, C. R., and Markowitz, H. (2019). "A backtesting protocol in the era of machine learning." In: *The Journal of Financial Data Science* 1(1), pp. 64–74.
Machine learning offers a set of powerful tools that holds considerable promise for investment management. As with most quantitative applications in finance, the danger of misapplying these techniques can lead to disappointment. One crucial limitation involves data availability. Many of machine learning early successes originated in the physical and biological sciences, in which truly vast amounts of data are available. Machine learning applications often require far more data than are available in finance, which is of particular concern in longer-horizon investing. Hence, choosing the right applications before applying the tools is important. In addition, capital markets reflect the actions of people, which may be influenced by others actions and by the findings of past research. In many ways, the challenges that affect machine learning are merely a continuation of the long-standing issues researchers have always faced in quantitative finance. While investors need to be cautious, more cautious than in past applications of quantitative methods new tools offer many potential applications in finance. In this article, the authors develop a research protocol that pertains both to the application of machine learning techniques and to quantitative finance in general.

Avellaneda, M. and Serur, J. A. (2020). "Hierarchical PCA and Modeling Asset Correlations." In: *arXiv e-Print*.
Modeling cross-sectional correlations between thousands of stocks, across countries and industries, can be challenging. In this paper, we demonstrate the advantages of using Hierarchical Principal Component Analysis (HPCA) over the classic PCA. We also introduce a statistical clustering algorithm for identifying of homogeneous clusters of stocks, or "synthetic sectors". We apply these methods to study cross-sectional correlations in the US, Europe, China, and Emerging Markets.

Bailey, D. H., Borwein, J. M., and Lopez de Prado, M. (2017). "Stock Portfolio Design and Backtest Overfitting." In: *Journal of Investment Management* 15(1), pp. 75–87.
In mathematical finance, backtest overfitting connotes the usage of historical market data to develop an investment strategy, where too many variations of the strategy are tried, relative to the amount of data available. Backtest overfitting is now thought to be a primary reason why investment models and strategies that look good on paper often disappoint in practice. Models and strategies suffering from overfitting typically target the specific idiosyncrasies of a limited dataset, rather than any general behavior, and, as a result, often perform erratically when presented with new data. In this study, we address overfitting in the context of designing a mutual fund or investment portfolio as a weighted collection of stocks. Very often a newly minted equity-based fund of this type has been designed by an exhaustive computer-based search of some sort to obtain an optimal weighting that exhibits excellent performance based, say, on the past 10 or 20 years' historical market data, and the fund often highlights this backtest performance.

Baitinger, E. (2021). "Forecasting asset returns with network-based metrics: A statistical and economic analysis." In: *Journal of Forecasting*.
One of the main challenges facing researchers and industry professionals for decades is the successful prediction of asset returns. This paper enriches this endeavor by applying topological metrics of correlation networks to the challenge of financial forecasting. These network-based metrics are retrieved with the help of graph theory and quantify the interconnectedness of financial assets. In this paper, we show that this network-based information statistically significantly predicts future asset returns. Because industry professionals are more interested in the economic value-added of competing forecasting approaches, we also devote our attention to an economic analysis. Considering economic performance metrics, network-based predictors generate a clear value-added, which also applies to the multi-asset allocation case.

Baitinger, E. and Flegel, S. (2021). "New Concepts in Financial Forecasting: Network-Based Information, Topological Data Analysis and their Combination." In: *SSRN e-Print*.
This paper introduces novel financial predictors that are derived from the interaction profile of financial markets. These predictors utilize network-based and topological information. Since these predictors are derived from the inner dynamics (microstructure) of financial markets, they can be best described as microstructural predictors. After equipping the reader with the methodological background of the novel predictors, we perform an extensive in-sample and out-of-sample performance analyses. The in-sample studies demonstrate that microstructural predictors and their combinations are informative with regard to future asset returns. In the out-of-sample studies, we combine microstructual predictors with state of the art machine learning and statistical factor extraction methods. The resulting active forecasting models dominate the benchmark mean model in terms of profitability, but not it terms of statistical precision. Since an investor is usually more concerned with profitability of active investment strategies, the out-of-sample results confirm the value-added of the novel predictors.

Baitinger, E. and Maier, T. (2019). "The (Mis)Behavior of Hedge Fund Strategies: A Network-Based Analysis." In: *The Journal of Alternative Investments* 22 (1), pp. 57–74.
The authors discuss a network-based methodology that models hedge fund strategies across the superordinate-subordinate dimension to gain new insights into their interrelation. This methodology uncovers considerable misbehavior of various hedge fund strategies from the network perspective. Simply speaking, a misbehaving hedge fund strategy has undesired network proximity (similarity) with strategies from other classifications and/or undesired network-based risk properties. The authors conduct extensive static and dynamic (bootstrapping) analyses demonstrating misbehaviors for the full-sample data set. In addition, they demonstrate that numerous network-based behavioral properties of hedge fund strategies can explain future hedge fund returns. This aspect is of significant relevance, as it shows that network-based information has the potential to act as a value-adding warning indicator for funds of hedge funds. Summing up, the authors think that this article provides novel and valuable tools for hedge fund investors, managers, and analysts.

Baitinger, E. and Papenbrock, J. (2017). "Interconnectedness Risk and Active Portfolio Management." In: *Journal of Investment Strategies* 6(2), pp. 63–90.

Interconnectedness is an alternative risk concept that so far has earned little attention in the asset management academia and industry. In this paper, we show that this neglect is not justified, as interconnectedness risk (i) has only moderate or no connection to conventional portfolio optimization inputs and (ii) active investment strategies based on interconnectedness information outperform their conventional peers. Utilizing a multi asset dataset, we measure interconnectedness risk by the embeddedness intensity, i.e. centrality, of assets in a correlation network, a concept from graph theory. Using the most common centrality measures, we first conduct empirical similarity studies analyzing how different centrality scores relate to each other and to conventional portfolio optimization inputs. Next, we outline how centrality can be incorporated in a risk-based as well as in a risk-return-based framework. Out-of-sample performance studies of centrality-optimized portfolios prove their competitiveness.

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

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

Bardoscia, M., Barucca, P., Battiston, S., Caccioli, F., Cimini, G., Garlaschelli, D., Saracco, F., Squartini, T., and Caldarelli, G. (2021). "The Physics of Financial Networks." In: *arXiv e-Print*.

The field of Financial Networks is a paramount example of the novel applications of Statistical Physics that have made possible by the present data revolution. As the total value of the global financial market has vastly outgrown the value of the real economy, financial institutions on this planet have created a web of interactions whose size and topology calls for a quantitative analysis by means of Complex Networks. Financial Networks are not only a playground for the use of basic tools of statistical physics as ensemble representation and entropy maximization; rather, their particular dynamics and evolution triggered theoretical advancements as the definition of DebtRank to measure the impact and diffusion of shocks in the whole systems. In this review we present the state of the art in this field, starting from the different definitions of financial networks (based either on loans, on assets ownership, on contracts involving several parties – such as credit default swaps, to multiplex representation when firms are introduced in the game and a link with real economy is drawn) and then discussing the various dynamics of financial contagion as well as applications in financial network inference and validation. We believe that this analysis is particularly timely since financial stability as well as recent innovations in climate finance, once properly analysed and understood in terms of complex network theory, can play a pivotal role in the transformation of our society towards a more sustainable world.

Barunik, J. and Ellington, M. (2021). "Dynamic Networks in Large Financial and Economic Systems." In: *arXiv e-Print*.

We propose new measures to characterize dynamic network connections in large financial and economic systems. In doing so, our measures allow one to describe and understand causal network structures that evolve throughout time and over horizons using variance decomposition matrices from time-varying parameter VAR (TVP VAR) models. These methods allow researchers and practitioners to examine network connections over any horizon of interest whilst also being applicable to a wide range of economic and financial data. Our empirical application redefines the meaning of big in big data, in the context of TVP VAR models, and track dynamic connections among illiquidity ratios of all S&P500 constituents. We then study the information content of these measures for the market return and real economy.

Begusic, S. and Kostanjcar, Z. (2019). "Cluster-Based Shrinkage of Correlation Matrices for Portfolio Optimization." In: *11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, pp. 301–305.

The estimation of correlation and covariance matrices from asset return time series is a critical step in financial portfolio optimization. Although sample estimates are reliable when the length of time series is very large compared to the number of assets, in high-dimensional settings estimation issues arise. To reduce estimation errors and mitigate their propagation to out-of-sample performance of portfolios based on noisy estimates, shrinkage methods are applied. In this paper we consider several shrinkage methods for correlation matrix estimation and define a cluster-based shrinkage procedure which introduces information about the structures of communities identified in asset dependence graphs. To test the considered shrinkage methods we apply them in a portfolio optimization scenario using the global minimum variance portfolio, and perform backtests on a large sample of NYSE daily stock return data. We find that shrinkage methods generally improve out-of-sample portfolio performance, and the proposed cluster-based method yields improved results and portfolios which outperform other considered methods.

Bennett, S., Cucuringu, M., and Reinert, G. (2022). "Lead-lag detection and network clustering for multivariate time series with an application to the US equity market." In: *arXiv e-Print*.

In multivariate time series systems, it has been observed that certain groups of variables partially lead the evolution of the system, while other variables follow this evolution with a time delay; the result is a lead-lag structure amongst the time series variables. In this paper, we propose a method for the detection of lead-lag clusters of time series in multivariate systems. We demonstrate that the web of pairwise lead-lag relationships between time series can be helpfully construed as a directed network, for which there exist suitable algorithms for the detection of pairs of lead-lag clusters with high pairwise imbalance. Within our framework, we consider a number of choices for the pairwise lead-lag metric and directed network clustering components. Our framework is validated on both a synthetic generative model for multivariate lead-lag time series systems and daily real-world US equity prices data. We showcase that our method is able to detect statistically significant lead-lag clusters in the US equity market. We study the nature of these clusters in the context of the empirical finance literature on lead-lag relations and demonstrate how these can be used for the construction of predictive financial signals.

Bessler, W., Opfer, H., and Wolff, D. (2017). "Multi-asset portfolio optimization and out-of-sample performance: an evaluation of Black Litterman, mean-variance, and naive diversification approaches." In: *The European Journal of Finance* 23(1), pp. 1–30.

The Black Litterman model aims to enhance asset allocation decisions by overcoming the problems of mean-variance portfolio optimization. We propose a sample-based version of the Black Litterman model and implement it on a multi-asset portfolio consisting of global stocks, bonds, and commodity indices, covering the period from January 1993 to December 2011. We test its out-of-sample performance relative to other asset allocation models and find that Black Litterman optimized portfolios significantly outperform na ive-diversified portfolios (1/N rule and strategic weights), and consistently perform better than mean-variance, Bayes Stein, and minimum-variance strategies in terms of out-of-sample Sharpe ratios, even after controlling for different levels of risk aversion, investment constraints, and transaction costs. The BL model generates portfolios with lower risk, less extreme asset allocations, and higher diversification across asset classes. Sensitivity analyses indicate that these advantages are due to more stable mixed return estimates that incorporate the reliability of return predictions, smaller estimation errors, and lower turnover.

Bessler, W. and Wolff, D. (2017). "Portfolio Optimization with Industry Return Prediction Models." In: *SSRN e-Print*.

We postulate that utilizing return prediction models with fundamental, macroeconomic, and technical indicators instead of using historical averages should result in superior asset allocation decisions. We investigate the predictive power of individual variables for forecasting industry returns in-sample and out-of-sample and then analyze multivariate predictive regression models including OLS, a regularization technique, principal components, a target-relevant latent factor approach, and forecast combinations. The gains from using industry return predictions are evaluated in an out-of-sample Black-Litterman portfolio optimization framework. We provide empirical evidence that portfolio optimization utilizing industry return prediction models significantly outperform portfolios using historical averages and those being passively managed.

Bjerring, T., Ross, O., and Weissensteiner, A. (2017). "Feature selection for portfolio optimization." In: *Annals of Operations Research* 256, pp. 21–40.

Most portfolio selection rules based on the sample mean and covariance matrix perform poorly out-of-sample. Moreover, there is a growing body of evidence that such optimization rules are not able to beat simple rules of thumb, such as 1/N. Parameter uncertainty has been identified as one major reason for these findings. A strand of literature addresses this problem by improving the parameter estimation and/or by relying on more

robust portfolio selection methods. Independent of the chosen portfolio selection rule, we propose using feature selection first in order to reduce the asset menu. While most of the diversification benefits are preserved, the parameter estimation problem is alleviated. We conduct out-of-sample back-tests to show that in most cases different well-established portfolio selection rules applied on the reduced asset universe are able to improve alpha relative to different prominent factor models.

Bnouachir, N. and Mkhadri, A. (2021). "Efficient cluster-based portfolio optimization." In: *Communications in Statistics - Simulation and Computation* 50, pp. 3241–3255.
The sample mean and covariance matrix of historical data provide a disappointing out-of-sample performance in mean-variance portfolio rules. This poor performance is certainly due to the high estimation error incurred in the optimization model. Our purpose in this article is to find a method that enhances the out-of-sample performance of the portfolio weights. Using hierarchical clustering, we propose an alternative cluster-based portfolio to obtain a sequence of cluster assets. On the basis of Gram-Schmidt orthogonalization, the estimation risk of the data set becomes the sum of the estimations of the clusters in the sequence. The performance of our method and its competitors is compared empirically and via some simulations in high dimension.

Bonald, T., Lara, N. de, Lutz, Q., and Charpentier, B. (2020). "Scikit-network: Graph Analysis in Python." In: *Journal of Machine Learning Research* 21(185), pp. 1–6.
Scikit-network is a Python package inspired by scikit-learn for the analysis of large graphs. Graphs are represented by their adjacency matrix in the sparse CSR format of SciPy. The package provides state-of-the-art algorithms for ranking, clustering, classifying, embedding and visualizing the nodes of a graph. High performance is achieved through a mix of fast matrix-vector products (using SciPy), compiled code (using Cython) and parallel processing. The package is distributed under the BSD license, with dependencies limited to NumPy and SciPy. It is compatible with Python 3.6 and newer. Source code, documentation and installation instructions are available online.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press. 446 pp.
Model-Based Clustering and Classification for Data Science written by Bouveyron et al. provides a comprehensive overview of the model-based approach for clustering and classification. The model-based approach sounds old-fashioned in the era of big data, but has attractive advantages such as "interpretability and simplicity with good if not optimal performance" according to the authors of the book. One potential drawback of the model-based approach is model misspecification. To alleviate it, the authors employ mixture models After a historical overview of clustering and classification given in Chapter 1, Chapter 2 introduces a family of Gaussian mixture models (GMM), that is, (1) with Gaussian fg, for clustering. The GMM serves as a building block for the rest of chapters. Under the GMM, the likelihood principle provides natural means for statistical inference. The computation of the maximum likelihood estimates becomes handy thanks to the well-known expectation and maximization (EM) algorithm, which greatly facilitates the use of GMM in various applications. Chapter 3 is devoted to address three frequently encountered difficulties when using the GMM for clustering, which are outliers, degenerated solution, and violation of the normality. The authors also described their respective solutions in great detail. Chapters 4 and 5 expand the use of the mixture model (1) to supervised and semi-supervised clustering and classification. Most textbooks start from the classification problem due to its popularity. However, in the context of a generative approach that seeks the DGP, classification can be viewed as a special case of clustering problem with known membership denoted by z. This is the reason why this book talks about classification after clustering. Chapter 6 describes how to modify the mixture model when y is discrete. For categorical y, the multinomial distribution is a natural choice that yields the multinomial mixture model, also known as the latent class model. If y are count data, the Poisson mixture model is canonical. Chapter 7 covers variable selection in the mixture model. The variable selection problem is solved by employing information criteria such as AIC and BIC, as well as sparsity-pursuing penalties. The mixture model (1) suffers from the curse of dimensionality when data are high-dimensional. In Chapter 8, the authors review a class of mixture models that efficiently reduce the parameter space. The last three chapters focus on the extension of the mixture models with more complex data. In Chapter 10, model-based clustering methods for the network data are introduced. Network data are now getting popular, but are quite challenging to handle due to its unique structure. In this chapter, the authors mainly focused on the illustration of various real data examples without analytical details. Chapter 11 describes the case when additional covariates are available. The last chapter introduces model-based clustering techniques for various types of data with complex structures: function (functional clustering method and its variants), text (latent Dirichlet allocation), and image. In addition, the co-clustering that seeks clusters of both

the observations and the variables simultaneously are also briefly mentioned. In conclusion, this book covers a wide range of statistical problems that are tailored for statistical analysis based on mixture models which help alleviate model misspecification issues. For those who expect a broad overview of various statistical techniques in data science including many recent black-box type approaches, this book may not be the best choice. However, for data scientists who need to improve the outcome quality and also to understand the DGP of the data in order to capture the complete picture of both the problem of interest and its solution, this book provides an excellent guide and is definitely recommendable.

Brécheteau, C., Fischer, A., and Levrard, C. (2021). "Robust Bregman clustering." In: *Annals of Statistics* 49(3).
Clustering with Bregman divergences encompasses a wide family of clustering procedures that are well suited to mixtures of distributions from exponential families (J. Mach. Learn. Res. 6 (2005) 1705-1749). However, these techniques are highly sensitive to noise. To address the issue of clustering data with possibly adversarial noise, we introduce a robustified version of Bregman clustering based on a trimming approach. We investigate its theoretical properties, showing for instance that our estimator converges at a sub-Gaussian rate 1/sqrt(n), where n denotes the sample size, under mild tail assumptions. We also show that it is robust to a certain amount of noise, stated in terms of breakdown point. We also derive a Lloyd-type algorithm with a trimming parameter, along with a heuristic to select this parameter and the number of clusters from sample. Some numerical experiments assess the performance of our method on simulated and real datasets.

Bruni, R., Cesarone, F., Scozzari, A., and Tardella, F. (2016). "Real-world datasets for portfolio selection and solutions of some stochastic dominance portfolio models." In: *Data in Brief* 8, pp. 858–862.
A large number of portfolio selection models have appeared in the literature since the pioneering work of Markowitz. However, even when computational and empirical results are described, they are often hard to replicate and compare due to the unavailability of the datasets used in the experiments. We provide here several datasets for portfolio selection generated using real-world price values from several major stock markets. The datasets contain weekly return values, adjusted for dividends and for stock splits, which are cleaned from errors as much as possible. The datasets are available in different formats, and can be used as benchmarks for testing the performances of portfolio selection models and for comparing the efficiency of the algorithms used to solve them. We also provide, for these datasets, the portfolios obtained by several selection strategies based on Stochastic Dominance models (see "On Exact and Approximate Stochastic Dominance Strategies for Portfolio Selection" (Bruni et al. [2])). We believe that testing portfolio models on publicly available datasets greatly simplifies the comparison of the different portfolio selection strategies.

Bruni, R., Cesarone, F., Scozzari, A., and Tardella, F. (2017). "On exact and approximate stochastic dominance strategies for portfolio selection." In: *European Journal of Operational Research* 259(1), pp. 322–329.
New type of approximate stochastic dominance designed for portfolio selection. Equivalent to minimizing the expected shortfall of the portfolio below the benchmark. An easily solvable LP model for the practical implementation of our approach. Extensive empirical comparison of stochastic dominance models for portfolio selection. One recent and promising strategy for Enhanced Indexation is the selection of portfolios that stochastically dominate the benchmark. We propose here a new type of approximate stochastic dominance rule which implies other existing approximate stochastic dominance rules. We then use it to find the portfolio that approximately stochastically dominates a given benchmark with the best possible approximation. Our model is initially formulated as a Linear Program with exponentially many constraints, and then reformulated in a more compact manner so that it can be very efficiently solved in practice. This reformulation also reveals an interesting financial interpretation. We compare our approach with several exact and approximate stochastic dominance models for portfolio selection. An extensive empirical analysis on real and publicly available datasets shows very good out-of-sample performances of our model.

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

apply our method to a very large set of factors proposed in the literature, and analyse 2.25 quadrillion possible models, gaining novel insights on the empirical drivers of asset returns.

Cai, F., Le-Khac, N.-A., and Kechadi, T. (2016). "Clustering Approaches for Financial Data Analysis: a Survey." In: *arXiv e-Print*.
Nowadays, financial data analysis is becoming increasingly important in the business market. As companies collect more and more data from daily operations, they expect to extract useful knowledge from existing collected data to help make reasonable decisions for new customer requests, e.g. user credit category, confidence of expected return, etc. Banking and financial institutes have applied different data mining techniques to enhance their business performance. Among these techniques, clustering has been considered as a significant method to capture the natural structure of data. However, there are not many studies on clustering approaches for financial data analysis. In this paper, we evaluate different clustering algorithms for analysing different financial datasets varied from time series to transactions. We also discuss the advantages and disadvantages of each method to enhance the understanding of inner structure of financial datasets as well as the capability of each clustering method in this context.

Cajas, D. (2019). "Robust Portfolio Selection with Near Optimal Centering." In: *SSRN e-Print*.
Quantitative asset allocation models have not been widely adopted by practitioners because they suffer from two problems: the lack of robustness and diversification of portfolios obtained through these models. To solve these problems, I developed a new portfolio selection method that can be applied to any convex risk measure. The procedure begins selecting an optimal portfolio in the efficient frontier, then I define a near optimal region and finally I define the analytic center as the new optimal portfolio. I com- pare 30 portfolio optimization models for 4 asset samples, and the results suggest that the new method overcomes traditional methods in robustness and diversification.

Campello, R. J. G. B., Kroger, P., Sander, J., and Zimek, A. (2020). "Density-based clustering." In: *WIREs Data Mining and Knowledge Discovery* 10(2).
Clustering refers to the task of identifying groups or clusters in a data set. In density-based clustering, a cluster is a set of data objects spread in the data space over a contiguous region of high density of objects. Density-based clusters are separated from each other by contiguous regions of low density of objects. Data objects located in low-density regions are typically considered noise or outliers. In this review article we discuss the statistical notion of density-based clusters, classic algorithms for deriving a flat partitioning of density-based clusters, methods for hierarchical density-based clustering, and methods for semi-supervised clustering. We conclude with some open challenges related to density-based clustering.

Casa, A., Scrucca, L., and Menardi, G. (2021). "Better than the best? Answers via model ensemble in density-based clustering." In: *Advances in Data Analysis and Classification* 15(3), pp. 599–623.
With the recent growth in data availability and complexity, and the associated outburst of elaborate modelling approaches, model selection tools have become a lifeline, providing objective criteria to deal with this increasingly challenging landscape. In fact, basing predictions and inference on a single model may be limiting if not harmful; ensemble approaches, which combine different models, have been proposed to overcome the selection step, and proven fruitful especially in the supervised learning framework. Conversely, these approaches have been scantily explored in the unsupervised setting. In this work we focus on the model-based clustering formulation, where a plethora of mixture models, with different number of components and parametrizations, is typically estimated. We propose an ensemble clustering approach that circumvents the single best model paradigm, while improving stability and robustness of the partitions. A new density estimator, being a convex linear combination of the density estimates in the ensemble, is introduced and exploited for group assignment. As opposed to the standard case, where clusters are typically associated to the components of the selected mixture model, we define partitions by borrowing the modal, or nonparametric, formulation of the clustering problem, where groups are linked with high-density regions. Staying in the density-based realm we thus show how blending together parametric and nonparametric approaches may be beneficial from a clustering perspective.

Castilho, D., Souza, T. T. P., Kang, S. M., Gama, J., and Carvalho, A. C. P. L. F. de (2021). "Forecasting Financial Market Structure from Network Features using Machine Learning." In: *arXiv e-Print*.
We propose a model that forecasts market correlation structure from link- and node-based financial network features using machine learning. For such, market structure is modeled as a dynamic asset network by quantifying time-dependent co-movement of asset price returns across company constituents of major global market indices. We provide empirical evidence using three different network filtering methods to estimate market structure, namely Dynamic Asset Graph (DAG), Dynamic Minimal Spanning Tree (DMST) and Dynamic Thresh-

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

Cesarone, F., Moretti, J., and Tardella, F. (2018). "Why Small Portfolios Are Preferable and How to Choose Them." In: *SSRN e-Print*.

One of the fundamental principles in portfolio selection models is minimization of risk through diversification of the investment. However, this principle does not necessarily translate into a request for investing in all the assets of the investment universe. Indeed, following a line of research started by Evans and Archer almost 50 years ago, we provide here further evidence that small portfolios are sufficient to achieve almost optimal in-sample risk reduction with respect to variance and to some other popular risk measures, and very good out-of-sample performances. While leading to similar results, our approach is significantly different from the classical one pioneered by Evans and Archer. Indeed, we describe models for choosing the portfolio of a prescribed size with the smallest possible risk, as opposed to the random portfolio choice investigated in most of the previous works. We find that the smallest risk portfolios generally require no more than 15 assets. Furthermore, it is almost always possible to find portfolios that are just 1% more risky than the smallest risk portfolios and contain no more than 10 assets. The preference for small optimal portfolios is also justified by recent theoretical results on the estimation errors for the parameters required by portfolio selection models. Our empirical analysis is based on some new and on some publicly available benchmark data sets often used in the literature.

Cesarone, F., Mottura, C., Ricci, J. M., and Tardella, F. (2019). "On the stability of portfolio selection models." In: *SSRN e-Print*.

One of the main issues in portfolio selection models consists in assessing the effect of the estimation errors of the parameters required by the models on the quality of the selected portfolios. Several studies have been devoted to this topic for the minimum variance and for several other minimum risk models. However, no sensitivity analysis seems to have been reported for the recent popular Risk Parity diversification approach, nor for other portfolio selection models requiring maximum gain-risk ratios.Based on artificial and real-world data, we provide here empirical evidence showing that the Risk Parity model is always the most stable one in all the cases analyzed. Furthermore, the minimum risk models are typically more stable than the maximum gain-risk models, with the minimum variance model often being the preferable one.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." In: *Journal of Statistical Software*.

Clustering is the partitioning of a set of objects into groups (clusters) so that objects within a group are more similar to each others than objects in dierent groups. Most of the clustering algorithms depend on some assumptions in order to dene the subgroups present in a data set. As a consequence, the resulting clustering scheme requires some sort of evaluation as regards its validity. The evaluation procedure has to tackle dicult problems such as the quality of clusters, the degree with which a clustering scheme ts a specic data set and the optimal number of clusters in a partitioning. In the literature, a wide variety of indices have been proposed to nd the optimal number of clusters in a partitioning of a data set during the clustering process. However, for most of indices proposed in the literature, programs are unavailable to test these indices and compare them. The R package NbClust has been developed for that purpose. It provides 30 indices which determine the number of clusters in a data set and it oers also the best clus- tering scheme from dierent results to the user. In addition, it provides a function to perform k-means and hierarchical clustering with dierent distance measures and aggregation methods. Any combination of validation indices and clustering methods can be requested in a single function call. This enables the user to simultaneously evaluate sev- eral clustering schemes while varying the number of clusters, to help determining the most appropriate number of clusters for the data set of interest.

Chaudhuri, S. E. and Lo, A. W. (2019). "Dynamic Alpha: A Spectral Decomposition of Investment Performance Across Time Horizons." In: *Management Science* 65(9), pp. 4440–4450.

The value added by an active investor is traditionally measured using alpha, tracking error, and the information ratio. However, these measures do not characterize the dynamic component of investor activity, nor do they consider the time horizons over which weights are changed. In this paper, we propose a technique to measure the value of active investment that captures both the static and dynamic contributions of an investment process. This

dynamic alpha is based on the decomposition of a portfolio's expected return into its frequency components using spectral analysis. The result is a static component that measures the portion of a portfolio's expected return resulting from passive investments and security selection and a dynamic component that captures the manager's timing ability across a range of time horizons. Our framework can be universally applied to any portfolio and is a useful method for comparing the forecast power of different investment processes. Several analytical and empirical examples are provided to illustrate the practical relevance of this decomposition.

Chavent, M., Genuer, R., and Saracco, J. (2021). "Combining clustering of variables and feature selection using random forests." In: *Communications in Statistics - Simulation and Computation* 50(2), pp. 426–445.
Standard approaches to tackle high-dimensional supervised classification often include variable selection and dimension reduction. The proposed methodology combines clustering of variables and feature selection. Hierarchical clustering of variables allows to built groups of correlated variables and summarizes each group by a synthetic variable. Originality is that groups of variables are unknown a priori. Moreover clustering approach deals with both numerical and categorical variables. Among all the possible partitions, the most relevant synthetic variables are selected with a procedure using random forests. Numerical performances are illustrated on simulated and real datasets. Selection of groups of variables provides easier interpretation of results.

Chehreghani, M. H. (2021). "Shift of Pairwise Similarities for Data Clustering." In: *arXiv e-Print*.
Several clustering methods (e.g., Normalized Cut and Ratio Cut) divide the Min Cut cost function by a cluster-dependent factor (e.g., the size or the degree of the clusters), in order to yield a more balanced partitioning. We, instead, investigate adding such regularizations to the original cost function. We first consider the case where the regularization term is the sum of the squared size of the clusters, and then generalize it to adaptive regularization of the pairwise similarities. This leads to shifting (adaptively) the pairwise similarities which might make some of them negative. We then study the connection of this method to Correlation Clustering and then propose an efficient local search optimization algorithm with fast theoretical convergence rate to solve the new clustering problem. In the following, we investigate the shift of pairwise similarities on some common clustering methods, and finally, we demonstrate the superior performance of the method by extensive experiments on different datasets.

Chen, W., Jiang, M., and Jiang, C. (2020). "Constructing a multilayer network for stock market." In: *Soft computing* 24, pp. 6345–6361.
In this paper, we discuss the stock network construction problem under simultaneous consideration of linear and nonlinear relations between stocks. A novel method based on the conditional probability is proposed to describe the nonlinear relation between stocks. Furthermore, by considering both the linear and nonlinear relations between stocks, a multilayer network is constructed to characterize stock market, in which Pearson correlation network, Granger causality network, and our proposed nonlinear relation network are combined. Finally, several experiments are conducted to illustrate the effectiveness of the proposed approaches. The results show that the proposed multilayer network not only covers more nodes than the Pearson correlation network, but also better balances the relation between prediction accuracy and the number of predictable nodes.

Chua, D. B., Kritzman, M., and Page, S. (2009). "The Myth of Diversification." In: *The Journal of Portfolio Management* 36(1), pp. 26–35.
Perhaps the most universally accepted precept of prudent investing is to diversify, yet this precept grossly oversimplifies the challenge of portfolio construction. Correlations, as typically measured over the full sample of returns, often belie an asset's diversification properties in market environments when diversification is most needed. Moreover, upside diversification is undesirable. The authors first describe the mathematics of conditional correlations assuming returns are normally distributed. Then they present empirical results across a wide variety of assets, which reveal that, unlike the theoretical conditional correlations, empirical correlations are significantly asymmetric. Finally, the authors show that a portfolio construction technique called full-scale optimization produces portfolios in which the component assets exhibit relatively lower correlations on the downside and higher correlations on the upside than mean-variance optimization portfolios.

Chung, J., Bridgeford, E., Arroyo, J., Pedigo, B. D., Saad-Eldin, A., Gopalakrishnan, V., Xiang, L., Priebe, C. E., and Vogelstein, J. T. (2022). "Statistical Connectomics." In: *Annual Review of Statistics and Its Application* 8(1), pp. 463–492.
The data science of networks is a rapidly developing field with myriad applications. In neuroscience, the brain is commonly modeled as a connectome, a network of nodes connected by edges. While there have been thousands of papers on connectomics, the statistics of networks remains limited and poorly understood. Here, we provide an overview from the perspective of statistical network science of the kinds of models, assumptions, problems,

and applications that are theoretically and empirically justified for analysis of connectome data. We hope this review spurs further development and application of statistically grounded methods in connectomics.

Clemente, G. P., Grassi, R., and Hitaj, A. (2019). "Smart network based portfolios." In: *arXiv e-Print*.
In this article we deal with the problem of portfolio allocation by enhancing network theory tools. We use the dependence structure of the correlations network in constructing some well-known risk-based models in which the estimation of correlation matrix is a building block in the portfolio optimization. We formulate and solve all these portfolio allocation problems using both the standard approach and the network-based approach. Moreover, in constructing the network-based portfolios we propose the use of two different estimators for the covariance matrix: the sample estimator and the shrinkage toward constant correlation one. All the strategies under analysis are implemented on two high-dimensional portfolios having different characteristics, covering the period from January 2001 to December 2017. We find that the network-based portfolio consistently better performs and has lower risk compared to the corresponding standard portfolio in an out-of-sample perspective.

Clemente, G. P., Grassi, R., and Hitaj, A. (2021). "Asset allocation: new evidence through network approaches." In: *Annals of Operations Research* 299, pp. 61–80.
The main contribution of the paper is to unveil the role of the network structure in the financial markets to improve the portfolio selection process, where nodes indicate securities and edges capture the dependence structure of the system. Three different methods are proposed in order to extract the dependence structure between assets in a network context. Starting from this modified structure, we formulate and then we solve the asset allocation problem. We find that the optimal portfolios obtained through a network-based approach are composed mainly of peripheral assets, which are poorly connected with the others. These portfolios, in the majority of cases, are characterized by an higher trade-off between performance and risk with respect to the traditional global minimum variance portfolio. Additionally, this methodology benefits of a graphical visualization of the selected portfolio directly over the graphic layout of the network, which helps in improving our understanding of the optimal strategy.

Coraggio, L. and Coretto, P. (2021). "Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score." In: *arXiv e-Print*.
Cluster analysis requires many decisions: the clustering method and the implied reference model, the number of clusters and, often, several hyper-parameters and algorithms' tunings. In practice, one produces several partitions, and a final one is chosen based on validation or selection criteria. There exist an abundance of validation methods that, implicitly or explicitly, assume a certain clustering notion. Moreover, they are often restricted to operate on partitions obtained from a specific method. In this paper, we focus on groups that can be well separated by quadratic or linear boundaries. The reference cluster concept is defined through the quadratic discriminant score function and parameters describing clusters' size, center and scatter. We develop two cluster-quality criteria called quadratic scores. We show that these criteria are consistent with groups generated from a general class of elliptically-symmetric distributions. The quest for this type of groups is common in applications. The connection with likelihood theory for mixture models and model-based clustering is investigated. Based on bootstrap resampling of the quadratic scores, we propose a selection rule that allows choosing among many clustering solutions. The proposed method has the distinctive advantage that it can compare partitions that cannot be compared with other state-of-the-art methods. Extensive numerical experiments and the analysis of real data show that, even if some competing methods turn out to be superior in some setups, the proposed methodology achieves a better overall performance.

Custodio João, I., Lucas, A., and Schaumburg, J. (2021). "Clustering Dynamics and Persistence for Financial Multivariate Panel Data." In: *SSRN e-Print*.
We introduce a new method for dynamic clustering of panel data with dynamics for cluster location and shape, cluster composition, and for the number of clusters. Whereas current techniques typically result in (economically) too many switches, our method results in economically more meaningful dynamic clustering patterns. It does so by extending standard cross-sectional clustering techniques using shrinkage towards previous cluster means. In this way, the different cross-sections in the panel are tied together, substantially reducing short-lived switches of units between clusters (flickering) and the birth and death of incidental, economically less meaningful clusters. In a Monte Carlo simulation, we study how to set the penalty parameter in a data-driven way. A systemic risk surveillance example for business model classification in the global insurance industry illustrates how the new method works empirically.

Dalmia, A. and Sia, S. (2021). "Clustering with UMAP: Why and How Connectivity Matters." In: *arXiv e-Print*.

Topology based dimensionality reduction methods such as t-SNE and UMAP have seen increasing success and popularity in high-dimensional data. These methods have strong mathematical foundations and are based on the intuition that the topology in low dimensions should be close to that of high dimensions. Given that the initial topological structure is a precursor to the success of the algorithm, this naturally raises the question: What makes a "good" topological structure for dimensionality reduction? Insight into this will enable us to design better algorithms which take into account both local and global structure. In this paper which focuses on UMAP, we study the effects of node connectivity (k-Nearest Neighbors vs *mutual* k-Nearest Neighbors) and relative neighborhood (Adjacent via Path Neighbors) on dimensionality reduction. We explore these concepts through extensive ablation studies on 4 standard image and text datasets; MNIST, FMNIST, 20NG, AG, reducing to 2 and 64 dimensions. Our findings indicate that a more refined notion of connectivity (*mutual* k-Nearest Neighbors with minimum spanning tree) together with a flexible method of constructing the local neighborhood (Path Neighbors), can achieve a much better representation than default UMAP, as measured by downstream clustering performance.

de Carvalho, P. J. C. and Gupta, A. (2018). "A network approach to unravel asset price comovement using minimal dependence structure." In: *Journal of Banking & Finance* 91, pp. 119–132.
We develop a network representation-based methodology to aid an exploratory analysis of temporally evolving comovement in asset prices. This parsimonious order-n representation of the most significant comovement in asset prices, filtered by common factors, allows tackling a large number of assets and unraveling their complex comovement structure. Flexibility in choosing explanatory factors to suit the specific objectives of a study makes this methodology useful for portfolio analysis, risk parity approaches, and risk management decisions. We illustrate the features of the methodology for a set of major industry equity indices and to blue chip stocks, where we analyze the dynamic relevance of Fama-French factors. Investigating the network for more than 20 years, including the dot-com bust, global financial crisis, and European debt crisis, helps draw many insights. For instance, unexpected industries are seen to connect idiosyncratically through the dot-com bust. We demonstrate that a network factor model based portfolio allocation performs better than a regular factor model based allocation.

De Luca, G. and Zuccolotto, P. (2021). "Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach." In: *International Journal of Approximate Reasoning* 139, pp. 88–103.
Time series clustering with a dissimilarity matrix based on tail dependence coefficients estimated by copula functions has been proposed in 2011 by De Luca and Zuccolotto, who used a two-step procedure allowing to resort to the k-means algorithm. The possibility to carry out hierarchical clustering directly on the dissimilarity matrix is still an open issue and the main concerns are relative to the meaning of the most common linkage methods in the context of tail dependence. In this paper, in a multivariate copula approach, we propose a linkage method based on the tail dependence coefficients between the clusters that are agglomerated at each iteration of the hierarchical clustering algorithms.

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.

Dees, B. S., Stankovic, L., Constantinides, A. G., and Mandic, D. P. (2020). "Portfolio Cuts: A Graph-Theoretic Framework to Diversification." In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
Investment returns naturally reside on irregular domains, however, standard multivariate portfolio optimization methods are agnostic to data structure. To this end, we investigate ways for domain knowledge to be conveniently incorporated into the analysis, by means of graphs. Next, to relax the assumption of the completeness of graph topology and to equip the graph model with practically relevant physical intuition, we introduce the portfolio cut paradigm. Such a graph-theoretic portfolio partitioning technique is shown to allow the investor to devise robust and tractable asset allocation schemes, by virtue of a rigorous graph framework for considering smaller, computationally feasible, and economically meaningful clusters of assets, based on graph cuts. In turn, this makes it possible to fully utilize the asset returns covariance matrix for constructing the portfolio, even without the requirement for its inversion. The advantages of the proposed framework over traditional methods are demonstrated through numerical simulations based on real-world price data.

Den Teuling, N., Pauws, S., and Heuvel, E. van den (2021). "Clustering of longitudinal data: A tutorial on a variety of approaches." In: *arXiv e-Print*.
During the past two decades, methods for identifying groups with different trends in longitudinal data have become of increasing interest across many areas of research. To support researchers, we summarize the guidance from the literature regarding longitudinal clustering. Moreover, we present a selection of methods for longitudinal clustering, including group-based trajectory modeling (GBTM), growth mixture modeling (GMM), and longitudinal k-means (KML). The methods are introduced at a basic level, and strengths, limitations, and model extensions are listed. Following the recent developments in data collection, attention is given to the applicability of these methods to intensive longitudinal data (ILD). We demonstrate the application of the methods on a synthetic dataset using packages available in R.

Dey, A. K., Tian, Y., and Gel, Y. R. (2021). "Community detection in complex networks: From statistical foundations to data science applications." In: *WIREs Computational Statistics*.
Identifying and tracking community structures in complex networks are one of the cornerstones of network studies, spanning multiple disciplines, from statistics to machine learning to social sciences, and involving even a broader range of application areas, from biology to politics to blockchain. This survey paper aims to provide an overview of some most popular approaches in statistical network community detection as well as the newly emerging research directions such as community extraction with higher-order features and community discovery in multilayer and multiscale networks. Our goal is to offer a unified view at methodological interconnections and the wide spectrum of interdisciplinary data science applications of network community analysis.

Di Cerbo, L. F. and Taylor, S. (2021). "Graph theoretical representations of equity indices and their centrality measures." In: *Quantitative Finance* 21(4), pp. 523–537.
The time dependent notion of equity market centrality can uncover the influence of the pairwise and risk evolution of securities with respect to system stability.

Diris, B., Palm, F., and Schotman, P. (2015). "Long-Term Strategic Asset Allocation: An Out-of-Sample Evaluation." In: *Management Science* 61(9), pp. 2185–2202.
We evaluate the out-of-sample performance of a long-term investor who follows an optimized dynamic trading strategy. Although the dynamic strategy is able to benefit from predictability out-of-sample, a short-term investor using a single-period market timing strategy would have realized an almost identical performance. The value of intertemporal hedge demands in strategic asset allocation appears negligible. The result is caused by the estimation error in predicting the predictors. A myopic investor only needs to predict one-period-ahead expected returns, but hedge demands also require accurate predictions of the predictor variables. To reduce the problem of errors in optimized portfolio weights, we consider Bayesian procedures. Myopic and dynamic portfolios are similarly affected by such modifications, and differences in performance become even smaller.

Djouzi, K. and Beghdad-Bey, K. (2019). "A Review of Clustering Algorithms for Big Data." In: *International Conference on Networking and Advanced Systems (ICNAS)*. IEEE.
Big data is usually defined by five (05) characteristics called 5Vs +1C (Volume, Velocity, Variety, Veracity, Value and Complexity). It means to data that are too large, dynamic and complex with certain degree of accuracy. For that, data become difficult to analyze using traditional data analysis techniques because of their high complexity and computational cost. Clustering analysis technique is the most used method for cope with huge amount of data. The main goal of clustering is to classify data into clusters in manner that data grouped are more similar. In this paper, we provide an overview of various clustering techniques used for data analysis.

Doreian, P., Batagelj, V., and Ferligoj, A. (2020). *Advances in Network Clustering and Blockmodeling*. Wiley. 432 pp.
This book offers an integrated treatment of network clustering and blockmodeling, covering all of the newest approaches and methods that have been developed over the last decade. Presented in a comprehensive manner, it offers the foundations for understanding network structures and processes, and features a wide variety of new techniques addressing issues that occur during the partitioning of networks across multiple disciplines such as community detection, blockmodeling of valued networks, role assignment, and stochastic blockmodeling. Written by a team of international experts in the field, Advances in Network Clustering and Blockmodeling offers a plethora of diverse perspectives covering topics such as: bibliometric analyses of the network clustering literature; clustering approaches to networks; label propagation for clustering; and treating missing network data before partitioning. It also examines the partitioning of signed networks, multimode networks, and linked networks. A chapter on structured networks and coarsegrained descriptions is presented, along with another on scientific coauthorship networks. The book finishes with a section covering conclusions and directions for future work. In addition, the editors provide numerous tables, figures, case studies, examples, datasets, and more.

Duan, J. (2021). "Predicting with Structured Data: Graphs, Ranks, and Time Series." PhD thesis. Kyoto University.
Predictive models have received wide attractions in modern engineering, financial, and social problems. We develop models and algorithms to analyse the patterns in big data, and try to leverage these patterns to make better predictions on what may happen in the next. In this process, structured data and patterns play an important role, because they are the intersections between the raw observations that we collect from the real world and the algorithms that we build and operate on computing resources. In the pursuit of favourable predictive performance, the models are mainly challenged by the variety and veracity of structured data and representations, because we neither have trustworthy raw observations that unveil the ground truth nor have enough amount of qualitative techniques to infer the values of interest given the heterogeneous observations.
In this dissertation, we attempt to address the variety and veracity problem in predicting with structured data by developing learning algorithms that can discover hidden structured patterns and use the patterns to facilitate the learning algorithm and the prediction. There are three essential challenges involved: (1) develop a method for the graph-structured data in materials informatics so that the fine-grained similarity of the graph elements can be accounted and contributes to the prediction, (2) construct learning algorithms and proper data representations so that the risk of conflicting observations can be mitigated, (3) develop time series forecasting algorithm that is robust to noisy temporal observation and is complementary to legacy evaluation metrics.
To tackle these challenges, we propose innovative structured learning techniques that can incorporate structured patterns as complementary input, output, and model components that can be learned directly from raw observations. For the first challenge, we develop fine grained kernels to describe the similarity of graph elements, e.g., labeled vertices and labeled edges. The proposed technique is tailored for domain experts to contribute and transmit their knowledge to the kernel construction. The proposed method achieves favourable predictive performance compared to the methods that use handmade features. In the small data scenario in materials informatics, the proposed method exhibits a significant advantage when the availability of annotations is limited. For the second challenge, we propose two techniques ranging from shallow model to deep neural networks to handle the intransitive relationships in raw observations. The models generalise the related works via incorporating cross-sectional numerical interactions between model parameters. For the shallow model, we define the numerical interactions via additional matrices and show constructively how the model generalise to the legacy related works. For the deep neural networks, we devise a structured neural network that simplifies the legacy models while maintaining the notion of pairwise comparison and matchup. A thorough investigation on real-world datasets show an universal presence of conflicting observations and highlights the importance of developing algorithms that can handle intransitive relationships. For the final challenge, we develop an alternative perspective, i.e., the ranking perspective, that is robust to noisy temporal observations in comparison with the legacy methods that optimize tracking errors. The proposed forecasting algorithm leverages the learning-to-rank technique and can simultaneously analyze the time series from both ranking and tracking error perspective by adopting a local learning algorithm that enables augmented inference. We are able to obtain improved ranking performance for the temporal observations, and moreover, this improvement is complementary to the tracking errors that are optimized by the conventional methods.

Duan, L. L. and Dunson, D. B. (2021). "Bayesian Distance Clustering." In: *Journal of Machine Learning Research* 22(224), pp. 1–27.
Model-based clustering is widely used in a variety of application areas. However, fundamental concerns remain about robustness. In particular, results can be sensitive to the choice of kernel representing the within-cluster data

density. Leveraging on properties of pairwise differences between data points, we propose a class of Bayesian distance clustering methods, which rely on modeling the likelihood of the pairwise distances in place of the original data. Although some information in the data is discarded, we gain substantial robustness to modeling assumptions. The proposed approach represents an appealing middle ground between distance- and model-based clustering, drawing advantages from each of these canonical approaches. We illustrate dramatic gains in the ability to infer clusters that are not well represented by the usual choices of kernel. A simulation study is included to assess performance relative to competitors, and we apply the approach to clustering of brain genome expression data.

Duarte, F. G. and De Castro, L. N. (2020). "A Framework to Perform Asset Allocation Based on Partitional Clustering." In: *IEEE Access* 8, pp. 110775–110788.

Over the past years, many approaches to perform asset allocation have been proposed in the literature. Most of them tackle this problem as an optimization task, where the goal is to maximize return, whilst minimizing the risk. However, such approaches require the inversion of a positive-definite covariance matrix, usually resulting in the concentration of allocation, instability and low performance. Some methods have been recently introduced to solve this problem by facing it as a clustering problem. This paper introduces a framework for asset allocation based on partitional clustering algorithms. The idea is to segment the assets into clusters of correlated assets, allocate resources for each cluster and then within each cluster. The framework allows the use of different partitional clustering algorithms, intragroup and intergroup allocation methods. Also, various assessment criteria are considered, and a specialized initialization method is proposed for the clustering algorithm. The framework is evaluated with the Brazilian Stock Exchange (B3) data from the period 12/2005 to 04/2020. Different initialization methods are used for the clustering algorithm together with two intergroup and two intragroup techniques, resulting in five experimental scenarios. The results are compared with the Ibovespa index, the mean-variance model of Markowitz, and the risk-parity model recently proposed by Lopez de Prado.

Dugué, N., Lamirel, J.-C., and Chen, Y. (2021). "Evaluating clustering quality using features salience: a promising approach." In: *Neural Computing and Applications* 33(19), pp. 12939–12956.

This paper focuses on using feature salience to evaluate the quality of a partition when dealing with hard clustering. It is based on the hypothesis that a good partition is an easy to label partition, i.e. a partition for which each cluster is made of salient features. This approach is mostly compared to usual approaches relying on distances between data, but also to more recent approaches based on entropy or stability. We show that our feature-based approach outperforms the compared indexes for optimal model selection: they are more efficient from low- to high-dimensional range as well as they are more robust to noise. To show the efficiency of our indexes on a real-life application, we consider the task of diachronic analysis on a textual dataset. We demonstrate that our approach allows to get some interesting and relevant results in that context, while other approaches mostly lead to unusable results.

Eidenvall, A. (2021). "Hierarchical Clustering To Improve Portfolio Tail Risk Characteristics." MA thesis. Lund University.

Many agree that estimating portfolio risks has better estimation possibilities, than estimations on returns. Therefore investors attempts to construct better, more efficient riskmanaged portfolios by diversifying portfolios through factors rather than traditional asset classes. This entails very often in estimations of correlation matrices so complex it cannot be fully analyzed. Hierarchical clustering reduces the complexity, by only focusing on the correlations that matters.

Hierarchical clustering uses graph theory, linked to unsupervised machine learning techniques. Hierarchical clustering is obtained by the suggested data and is a formation of a recursive clustering. Several hierarchical clustering methods are presented and evaluated against traditional riskbased portfolios with focus on left hand tail risk. A regime shift, based on momentum is applied to minimize drawdowns. The portfolios are tested on simulated data derived from Bootstrapping simulations and on historical data in a Walk forward optimization process. The results indicate that hierarchical clustering based portfolios are truly diversified and achieve statistically better riskadjusted performances than commonly used portfolio optimization techniques.

Elliott, A., Chiu, A., Bazzi, M., Reinert, G., and Cucuringu, M. (2020). "Core–periphery structure in directed networks." In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 476(2241), p. 20190783.

Empirical networks often exhibit different meso-scale structures, such as community and core-periphery structures. Core-periphery structure typically consists of a well-connected core and a periphery that is well connected to the core but sparsely connected internally. Most core-periphery studies focus on undirected networks. We

propose a generalization of core-periphery structure to directed networks. Our approach yields a family of core-periphery block model formulations in which, contrary to many existing approaches, core and periphery sets are edge-direction dependent. We focus on a particular structure consisting of two core sets and two periphery sets, which we motivate empirically. We propose two measures to assess the statistical significance and quality of our novel structure in empirical data, where one often has no ground truth. To detect core-periphery structure in directed networks, we propose three methods adapted from two approaches in the literature, each with a different trade-off between computational complexity and accuracy. We assess the methods on benchmark networks where our methods match or outperform standard methods from the literature, with a likelihood approach achieving the highest accuracy. Applying our methods to three empirical networks-faculty hiring, a world trade dataset and political blogs-illustrates that our proposed structure provides novel insights in empirical networks.

Emerson, S. (2019). "Machine learning for financial applications: self-organising maps, hierarchical clustering and dynamic time-warping for portfolio constructive." PhD thesis. University College Cork.
This study investigates how modern machine learning (ML) techniques can be used to advance the field of quantitative investing. A broad literature review evaluated the common applications for ML in finance, and what ML algorithms are being used. The results show ML is commonly applied to the areas of Return Forecasting, Portfolio Construction, Ethics, Fraud Detection Decision Making Language Processing and Sentiment Analysis. Neural Network technology and support vector machine are identified as popular ML algorithms. A second review was carried out, focusing in the area of ML for quantitative finance in recent years finds three primary areas; Return forecasting, Portfolio construction and Risk management. A practical ML experiment carried out as a proof of concept of ML for financial applications. This experiment was informed by the results of the broad and more focused literature searches. Two forms of ML techniques are used to analyse market return data and equity flow data (provided by State Street Global Markets) and create a portfolio from insights derived from the ML technology. The ML technologies employed are those of Self-Organising Maps and Hierarchical Clustering. The portfolios created were tested in terms of risk, profitability and stability. Stable regimes and profitable portfolios are created. Results show that portfolios obtained by analysing equity flow data consistently outperform those created by analysing return data.

Escanciano, J. C. and Hualde, J. (2021). "Measuring Asset Market Linkages: Nonlinear Dependence and Tail Risk." In: *Journal of Business & Economic Statistics* 39(2), pp. 453–465.
Traditional measures of dependence in time series are based on correlations or periodograms. These are adequate in many circumstances but, in others, especially when trying to assess market linkages and tail risk during abnormal times (e.g., financial contagion), they might be inappropriate. In particular, popular tail dependence measures based on exceedance correlations and marginal expected shortfall (MES) have large variances and also contain limited information on tail risk. Motivated by these limitations, we introduce the (tail-restricted) integrated regression function, and we show how it characterizes conditional dependence and persistence. We propose simple estimates for these measures and establish their asymptotic properties. We employ the proposed methods to analyze the dependence structure of some of the major international stock market indices before, during, and after the 2007-2009 financial crisis. Monte Carlo simulations and the application show that our new measures are more reliable and accurate than competing methods based on MES or exceedance correlations for testing tail dependence. Supplementary materials for this article are available online.

Esmaeili, S. A., Brubach, B., Tsepenekas, L., and Dickerson, J. P. (2021). "Probabilistic Fair Clustering." In: *arXiv e-Print*.
In clustering problems, a central decision-maker is given a complete metric graph over vertices and must provide a clustering of vertices that minimizes some objective function. In fair clustering problems, vertices are endowed with a color (e.g., membership in a group), and the features of a valid clustering might also include the representation of colors in that clustering. Prior work in fair clustering assumes complete knowledge of group membership. In this paper, we generalize prior work by assuming imperfect knowledge of group membership through probabilistic assignments. We present clustering algorithms in this more general setting with approximation ratio guarantees. We also address the problem of "metric membership", where different groups have a notion of order and distance. Experiments are conducted using our proposed algorithms as well as baselines to validate our approach and also surface nuanced concerns when group membership is not known deterministically.

Exarchakis, G., Oubari, O., and Lenz, G. (2022). "A sampling-based approach for efficient clustering in large datasets." In: *arXiv e-Print*.
We propose a simple and efficient clustering method for high-dimensional data with a large number of clusters. Our algorithm achieves high-performance by evaluating distances of datapoints with a subset of the cluster cen-

tres. Our contribution is substantially more efficient than k-means as it does not require an all to all comparison of data points and clusters. We show that the optimal solutions of our approximation are the same as in the exact solution. However, our approach is considerably more efficient at extracting these clusters compared to the state-of-the-art. We compare our approximation with the exact k-means and alternative approximation approaches on a series of standardised clustering tasks. For the evaluation, we consider the algorithmic complexity, including number of operations to convergence, and the stability of the results.

Ezhilmaran, D. and Indira, D. V. (2020). "A survey on clustering techniques in pattern recognition." In: *AIP Conference Proceedings*. AIP Publishing.

In the recent era, technological approaches to a problem can be made to solve it easily. There are many technical ways to solve according to the types of problem. Classification makes that identify the problem of character. Clustering is one of the technique to classify and grouping the objects. Clustering has hundreds of different approaches to group the objects. Clustering has more than 50 years of history behind it. In this article, we take a study on cluster's definitions, types, algorithms and try to analyze the results. As well as give a literature survey based on data clustering, pattern recognition, and image classifications.

Fabozzi, F. J. and Lopez de Prado, M. (2018). "Being Honest in Backtest Reporting: A Template for Disclosing Multiple Tests." In: *The Journal of Portfolio Management* 45(1), pp. 141–147.

Selection bias under multiple testing is a serious problem. From a practitioner perspective, failure to disclose the impact of multiple tests of a proposed investment strategy to clients and senior management can lead to the adoption of a false discovery. Clients will lose money, senior management will misallocate resources, and the firm may be exposed to reputational, legal, and regulatory risks. From the perspective of academic journals that publish evidence supporting an investment strategy, the failure to address selection bias under multiple testing threatens to invalidate large portions of the literature in empirical finance. In this article, the authors propose a template that practitioners should use to fairly disclose multiple tests involved in an alleged discovery when pitching strategies to clients and senior management. The same template could be used by contributors to academic journals so that referees, and ultimately readers, can assess the strategy. By disclosing this information, those who are charged with making the final decision about a discovery can evaluate the probability that the purported discovery is false.

Faletto, G. and Bien, J. (2022). "Cluster Stability Selection." In: *arXiv e-Print*.

Stability selection (Meinshausen and Buhlmann, 2010) makes any feature selection method more stable by returning only those features that are consistently selected across many subsamples. We prove (in what is, to our knowledge, the first result of its kind) that for data containing highly correlated proxies for an important latent variable, the lasso typically selects one proxy, yet stability selection with the lasso can fail to select any proxy, leading to worse predictive performance than the lasso alone. We introduce cluster stability selection, which exploits the practitioner's knowledge that highly correlated clusters exist in the data, resulting in better feature rankings than stability selection in this setting. We consider several feature-combination approaches, including taking a weighted average of the features in each important cluster where weights are determined by the frequency with which cluster members are selected, which we show leads to better predictive models than previous proposals. We present generalizations of theoretical guarantees from Meinshausen and Buhlmann (2010) and Shah and Samworth (2012) to show that cluster stability selection retains the same guarantees. In summary, cluster stability selection enjoys the best of both worlds, yielding a sparse selected set that is both stable and has good predictive performance.

Feng, J. and Simon, N. (2020). "Ensembled sparse-input hierarchical networks for high-dimensional datasets." In: *arXiv e-Print*.

Neural networks have seen limited use in prediction for high-dimensional data with small sample sizes, because they tend to overfit and require tuning many more hyperparameters than existing off-the-shelf machine learning methods. With small modifications to the network architecture and training procedure, we show that dense neural networks can be a practical data analysis tool in these settings. The proposed method, Ensemble by Averaging Sparse-Input Hierarchical networks (EASIER-net), appropriately prunes the network structure by tuning only two L1-penalty parameters, one that controls the input sparsity and another that controls the number of hidden layers and nodes. The method selects variables from the true support if the irrelevant covariates are only weakly correlated with the response; otherwise, it exhibits a grouping effect, where strongly correlated covariates are selected at similar rates. On a collection of real-world datasets with different sizes, EASIER-net selected network architectures in a data-adaptive manner and achieved higher prediction accuracy than off-the-shelf methods on average.

Ferraro, M. B., Giordani, P., and Serafini, A. (2019). "fclust: An R Package for Fuzzy Clustering." In: *The R Journal*.
Fuzzy clustering methods discover fuzzy partitions where observations can be softly assigned to more than one cluster. The package fclust is a toolbox for fuzzy clustering in the R programming language. It not only implements the widely used fuzzy k-means (FkM) algorithm, but also many FkM variants. Fuzzy cluster similarity measures, cluster validity indices and cluster visualization tools are also offered. In the current version, all the functions are rewritten in the C++ language allowing their application in large-size problems. Moreover, new fuzzy relational clustering algorithms for partitioning qualitative/mixed data are provided together with an improved version of the so-called Gustafson-Kessel algorithm to avoid singularity in the cluster covariance matrices. Finally, it is now possible to automatically select the number of clusters by means of the available fuzzy cluster validity indices.

Fischer, D., Berro, A., Nordhausen, K., and Ruiz-Gazen, A. (2021). "REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit." In: *Communications in Statistics - Simulation and Computation*.
The R-package REPPlab is designed to explore multivariate data sets using one-dimensional unsupervised projection pursuit. It is useful as a preprocessing step to find clusters or as an outlier detection tool for multivariate data. Except from the packages tourr and rggobi, there is no implementation of exploratory projection pursuit tools available in R. REPPlab is an R interface for the Java program EPP-lab that implements four projection indices and three biologically inspired optimization algorithms. It also proposes new tools for plotting and combining the results and specific tools for outlier detection. The functionality of the package is illustrated through some simulations and using some real data.

Flint, E., Seymour, A., and Chikurunhe, F. (2021). "Defining and measuring portfolio diversification." In: *South African Actuarial Journal* 20(1), pp. 17–48.
It is often said that diversification is the only 'free lunch' available to investors; meaning that a properly diversified portfolio reduces total risk without necessarily sacrificing expected return. However, achieving true diversification is easier said than done, especially when we do not fully know what we mean when we are talking about diversification. While the qualitative purpose of diversification is well known, a satisfactory quantitative definition of portfolio diversification remains elusive. In this research, we summarise a wide range of diversification measures, focusing our efforts on those most commonly used in practice. We categorise each measure based on which portfolio aspect it focuses on: cardinality, weights, returns, risk or higher moments. We then apply these measures to a range of South African equity indices, thus giving a diagnostic review of historical local equity diversification and, perhaps more importantly, providing a description of the investable opportunity set available tofund managers in this space. Finally, we introduce the idea of diversification profiles. These regime dependent profiles give a much richer description of portfolio diversification than their single-value counterparts and also allow one to manage diversification proactively based on one's view of future market conditions.

Fop, M. and Murphy, T. B. (2017). "Variable Selection Methods for Model-based Clustering." In: *arXiv e-Print*.
Model-based clustering is a popular approach for clustering multivariate data which has seen applications in numerous fields. Nowadays, high-dimensional data are more and more common and the model-based clustering approach has adapted to deal with the increasing dimensionality. In particular, the development of variable selection techniques has received a lot of attention and research effort in recent years. Even for small size problems, variable selection has been advocated to facilitate the interpretation of the clustering results. This review provides a summary of the methods developed for variable selection in model-based clustering. Existing R packages implementing the different methods are indicated and illustrated in application to two data analysis examples.

Franti, P. and Sieranoja, S. (2019). "How much can k-means be improved by using better initialization and repeats?" In: *Pattern Recognition* 93, pp. 95–112.
In this paper, we study what are the most important factors that deteriorate the performance of the k-means algorithm, and how much this deterioration can be overcome either by using a better initialization technique, or by repeating (restarting) the algorithm. Our main finding is that when the clusters overlap, k-means can be significantly improved using these two tricks. Simple furthest point heuristic (Maxmin) reduces the number of erroneous clusters from 15% to 6%, on average, with our clustering benchmark. Repeating the algorithm 100 times reduces it further down to 1%. This accuracy is more than enough for most pattern recognition applications. However, when the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization. Therefore, if high clustering accuracy is needed, a better algorithm should be used instead.

Fu, W. and Perry, P. O. (2020). "Estimating the Number of Clusters Using Cross-Validation." In: *Journal of Computational and Graphical Statistics* 29(1), pp. 162–173.
Many clustering methods, including k-means, require the user to specify the number of clusters as an input parameter. A variety of methods have been devised to choose the number of clusters automatically, but they often rely on strong modeling assumptions. This article proposes a data-driven approach to estimate the number of clusters based on a novel form of cross-validation. The proposed method differs from ordinary cross-validation, because clustering is fundamentally an unsupervised learning problem. Simulation and real data analysis results show that the proposed method outperforms existing methods, especially in high-dimensional settings with heterogeneous or heavy-tailed noise. In a yeast cell cycle dataset, the proposed method finds a parsimonious clustering with interpretable gene groupings. Supplementary materials for this article are available online.

Fusai, G., Mignacca, D., Nardon, A., and Human, B. (2020). "Equally Diversified or Equally Weighted?" In: *Risk (Cutting Edge)*.
Gianluca Fusai, Domenico Mignacca, Andrea Nardon and Ben Human show how to decompose portfolio volatility into undiversified volatility and a diversification component. The authors' decomposition has a clear statistical interpretation because it relates the diversification component to partial covariances. On this basis, they advocate the construction of an equally diversified portfolio. An empirical analysis illustrates the superior out-of-sample performance of the equally diversified portfolio with respect to an equally weighted portfolio.

Gagolewski, M. (2021). "genieclust: Fast and robust hierarchical clustering." In: *SoftwareX* 15, p. 100722.
genieclust is an open source Python and R package that implements the hierarchical clustering algorithm called Genie. This method frequently outperforms other state-of-the-art approaches in terms of clustering quality and speed, supports various distances over dense, sparse, and string data domains, and can be robustified even further with the built-in noise point detector. As domain-independent software, it can be used for solving problems arising in all data-driven research and development activities, including environmental, health, biological, physical, decision, and social sciences as well as technology and engineering. The Python version provides a scikit-learn-compliant API, whereas the R variant is compatible with the classic hclust(). Numerous tutorials, use cases, non-trivial examples, documentation, installation instructions, benchmark results and timings can be found at `https://genieclust.gagolewski.com`.

Gao, Z. and Tsay, R. S. (2021). "Divide-and-Conquer: A Distributed Hierarchical Factor Approach to Modeling Large-Scale Time Series Data." In: *arXiv e-Print*.
This paper proposes a hierarchical approximate-factor approach to analyzing high-dimensional, large-scale heterogeneous time series data using distributed computing. The new method employs a multiple-fold dimension reduction procedure using Principal Component Analysis (PCA) and shows great promises for modeling large-scale data that cannot be stored nor analyzed by a single machine. Each computer at the basic level performs a PCA to extract common factors among the time series assigned to it and transfers those factors to one and only one node of the second level. Each 2nd-level computer collects the common factors from its subordinates and performs another PCA to select the 2nd-level common factors. This process is repeated until the central server is reached, which collects common factors from its direct subordinates and performs a final PCA to select the global common factors. The noise terms of the 2nd-level approximate factor model are the unique common factors of the 1st-level clusters. We focus on the case of 2 levels in our theoretical derivations, but the idea can easily be generalized to any finite number of hierarchies. We discuss some clustering methods when the group memberships are unknown and introduce a new diffusion index approach to forecasting. We further extend the analysis to unit-root nonstationary time series. Asymptotic properties of the proposed method are derived for the diverging dimension of the data in each computing unit and the sample size T. We use both simulated data and real examples to assess the performance of the proposed method in finite samples, and compare our method with the commonly used ones in the literature concerning the forecastability of extracted factors.

Garvey, G. and Madhavan, A. (2019). "Reconstructing Emerging and Developed Markets Using Hierarchical Clustering." In: *The Journal of Financial Data Science* 1 (4), pp. 84–102.
The distinction between emerging and developed markets is of first-order importance for investors. In this article, the authors use hierarchical clustering to objectively identify the countries or regions that cluster from an investment viewpoint. They go beyond classifications based on economic fundamentals and group countries based on returns in equity and bond markets. The authors find an important geographical footprint that differs significantly from the groupings that are used by most practitioners. This analysis has practical implications for both active and index investors.

Gherbaoui, R., Ouali, M., and Benamrane, N. (2021). "Generation of Gaussian sets for clustering methods assessment." In: *Data & Knowledge Engineering* 131-132, p. 101876.
Clustering methods are generally used to study the homogeneity in a set of observations. The results obtained from the clustering process differ from one method to another, to the extent that the same method or validity index gives different outcomes depending on the initial parameters. Analytical evaluation appears to be insufficient for studying the behavior of clustering methods due to its ad hoc nature. Even if the real data set is used in evaluating clustering methods, artificial data is fundamental for assessing the performance since it allows creating different scenarios of test with known structures. The main drawback of existing methods of artificial data is that they do not take into consideration the problem of sensitivity to the size of clusters. In this paper, we propose an automatic method: the high-dimensional artificial Gaussian mixture generator. By formally quantifying the overlap, the generator preserves the notion of the overlap rate between the mixture components. The advantages of this generator are its use of the notion of overlap rate, the unlimited number of mixture components, high-dimensionality of the observations, and the non-utilization of visual inspection as a criterion to quantify the overlap. In addition, we evaluate the k-means, fuzzy c-means (FCM), FCM-based splitting algorithm (FBSA), and expectation maximization (EM) in different dimensions. The results obtained confirm previous work and reveal new findings that are not pointed out when using 1D and 2D artificial data The source code of the implementation is publicly available at the following URL: `http://193.194.88.10/bitstream/123456789/392/1/cgeg.zip`.

Ghosal, A., Nandy, A., Das, A. K., Goswami, S., and Panday, M. (2019). "A Short Review on Different Clustering Techniques and Their Applications." In: *Advances in Intelligent Systems and Computing*. Springer Singapore, pp. 69–83.
In modern world, we have to deal with huge volumes of data which include image, video, text and web documents, DNA, microarray gene data, etc. Organizing such data into rational groups is a critical first step to draw inferences. Data clustering analysis has emerged as an effective technique to accurately accomplish the task of categorizing data into sensible groups. Clustering has a rich association with researches in various scientific domains. One of the most popular clustering algorithms, k-means algorithm was proposed as early as 1957. Since then, many clustering algorithms have been developed and used, to group data in various commercial and non-commercial sectors alike. In this paper, we have given concise description of the existing types of clustering approaches followed by a survey of the fields where clustering analytics has been effectively employed in pattern recognition and knowledge discovery.

Giudici, P., Pagnottoni, P., and Polinesi, G. (2020). "Network Models to Enhance Automated Cryptocurrency Portfolio Management." In: *Frontiers in Artificial Intelligence* 3.
The usage of cryptocurrencies, together with that of financial automated consultancy, is widely spreading in the last few years. However, automated consultancy services are not yet exploiting the potentiality of this nascent market, which represents a class of innovative financial products that can be proposed by robo-advisors. For this reason, we propose a novel approach to build efficient portfolio allocation strategies involving volatile financial instruments, such as cryptocurrencies. In other words, we develop an extension of the traditional Markowitz model which combines Random Matrix Theory and network measures, in order to achieve portfolio weights enhancing portfolios' risk-return profiles. The results show that overall our model overperforms several competing alternatives, maintaining a relatively low level of risk.

Giudici, P., Polinesi, G., and Spelta, A. (2022). "Network models to improve robot advisory portfolios." In: *Annals of Operations Research* 313, pp. 965–989.
Robot advisory services are rapidly expanding, responding to a growing interest people have in directly managing their savings. Robot-advisors may reduce costs and improve the quality of asset allocation services, making user's involvement more transparent. Against this background, there exists the possibility that robot advisors underestimate market risks, especially during crisis times, when high order interconnections arise. This may lead to a mismatch between investors' expected and actual risk. The aim of this paper is to overcome this issue, taking into account not only investors' risk preference but also their attitude towards interconnectdness. To achieve this aim, we combine random matrix theory with correlation networks and extend the Markowitz' optimisation problem to a third dimension. To demonstrate the practical advantage of our proposed approach we employ daily returns of a large set of Exchange Traded Funds, which are representative of the financial products employed by robot-advisors.

Goel, A. and Majumdar, A. (2021). "Transformed K-means Clustering." In: *arXiv e-Print*.
In this work we propose a clustering framework based on the paradigm of transform learning. In simple terms the representation from transform learning is used for K-means clustering; however, the problem is not solved in

such a naïve piecemeal fashion. The K-means clustering loss is embedded into the transform learning framework and the joint problem is solved using the alternating direction method of multipliers. Results on document clustering show that our proposed approach improves over the state-of-the-art.

Greiner, S. P. and Stoyanov, S. V. (2019). "Portfolio scoring by expected risk premium." In: *The Journal of Portfolio Management* 45(4), pp. 83–90.
In this article, the authors discuss a general method for ranking portfolios that places few limitations on the portfolio constituents other than using publicly traded assets. The ranking scores reflect the expected reward investors would require for accepting the risks of the portfolio in the context of an asset pricing framework. The scores are computed through a factor model that acknowledges the factor return correlations. The authors illustrate the approach with a large universe of exchange-traded funds assuming a linear model with Fama-French-Carhart factors wherein factor premiums (i.e., expected returns) are proportional to factor volatilities. The empirical analysis implies that the most significant factors from the Fama-French-Carhart factor set driving the premiums are the market and the momentum factors.

Grun, B. (2018). "Model-based Clustering." In: *arXiv e-Print*.
Mixture models extend the toolbox of clustering methods available to the data analyst. They allow for an explicit definition of the cluster shapes and structure within a probabilistic framework and exploit estimation and inference techniques available for statistical models in general. In this chapter an introduction to cluster analysis is provided, model-based clustering is related to standard heuristic clustering methods and an overview on different ways to specify the cluster model is given. Post-processing methods to determine a suitable clustering, infer cluster distribution characteristics and validate the cluster solution are discussed. The versatility of the model-based clustering approach is illustrated by giving an overview on the different areas of applications.

Guan, S. and Loew, M. (2021). "A Distance-based Separability Measure for Internal Cluster Validation." In: *arXiv e-Print*.
To evaluate clustering results is a significant part of cluster analysis. Since there are no true class labels for clustering in typical unsupervised learning, many internal cluster validity indices (CVIs), which use predicted labels and data, have been created. Without true labels, to design an effective CVI is as difficult as to create a clustering method. And it is crucial to have more CVIs because there are no universal CVIs that can be used to measure all datasets and no specific methods of selecting a proper CVI for clusters without true labels. Therefore, to apply a variety of CVIs to evaluate clustering results is necessary. In this paper, we propose a novel internal CVI – the Distance-based Separability Index (DSI), based on a data separability measure. We compared the DSI with eight internal CVIs including studies from early Dunn (1974) to most recent CVDD (2019) and an external CVI as ground truth, by using clustering results of five clustering algorithms on 12 real and 97 synthetic datasets. Results show DSI is an effective, unique, and competitive CVI to other compared CVIs. We also summarized the general process to evaluate CVIs and created the rank-difference metric for comparison of CVIs' results.

Gubu, L., Rosadi, D., and Abdurakhman (2021). "Robust mean-variance portfolio selection with time series clustering." In: *AIP Conference Proceedings*. AIP Publishing.
This study presents a robust portfolio selection with time series clustering. The stocks are firstly grouped into several clusters using Partitioning Around Medoids (PAM) time series clustering base on autocorrelation function (ACF) dissimilarity. After clustering process, stocks are chosen to represent each cluster to build a portfolio. The stock chosen from each cluster is the stock that has the best Sharpe ratio. The optimum portfolio is determined using the robust Fast Minimum Covariance Determinant (FMCD) and S estimation. Using this procedure, we can efficiently obtain the best portfolio when there are large number of stocks involved in portfolio formulation process. This procedure is also robust against the probability of outlier presence in the data. To measure the performance of portfolios that are formed we use the Sharpe ratio. For empirical study, we used the daily closing price of stocks listed on the Indonesia Stock Exchange, which included in the LQ-45 indexed for the period of August 2017-July 2018. Results of this study showed that the performance of portfolio generated by the use of PAM time series clustering combined with robust FMCD estimation was better than performance of portfolio generated by other methods that we tested.

Guidolin, M., Hansen, E., and Lozano-Banda, M. (2018). "Portfolio performance of linear SDF models: an out-of-sample assessment." In: *Quantitative Finance* 18(8), pp. 1425–1436.
We evaluate linear stochastic discount factor models using an ex-post portfolio metric: the realized out-of-sample Sharpe ratio of mean-variance portfolios backed by alternative linear factor models. Using a sample of monthly US portfolio returns spanning the period 1968-2016, we find evidence that multifactor linear models have

better empirical properties than the CAPM, not only when the cross-section of expected returns is evaluated in-sample, but also when they are used to inform one-month ahead portfolio selection. When we compare portfolios associated to multifactor models with mean-variance decisions implied by the single-factor CAPM, we document statistically significant differences in Sharpe ratios of up to 10 percent. Linear multifactor models that provide the best in-sample fit also yield the highest realized Sharpe ratios.

Guijo-Rubio, D., Duran-Rosal, A. M., Gutierrez, P. A., Troncoso, A., and Hervas-Martinez, C. (2020). "Time-Series Clustering Based on the Characterization of Segment Typologies." In: *IEEE Transactions on Cybernetics*.
Time-series clustering is the process of grouping time series with respect to their similarity or characteristics. Previous approaches usually combine a specific distance measure for time series and a standard clustering method. However, these approaches do not take the similarity of the different subsequences of each time series into account, which can be used to better compare the time-series objects of the dataset. In this article, we propose a novel technique of time-series clustering consisting of two clustering stages. In a first step, a least-squares polynomial segmentation procedure is applied to each time series, which is based on a growing window technique that returns different-length segments. Then, all of the segments are projected into the same dimensional space, based on the coefficients of the model that approximates the segment and a set of statistical features. After mapping, a first hierarchical clustering phase is applied to all mapped segments, returning groups of segments for each time series. These clusters are used to represent all time series in the same dimensional space, after defining another specific mapping process. In a second and final clustering stage, all the time-series objects are grouped. We consider internal clustering quality to automatically adjust the main parameter of the algorithm, which is an error threshold for the segmentation. The results obtained on 84 datasets from the UCR Time Series Classification Archive have been compared against three state-of-the-art methods, showing that the performance of this methodology is very promising, especially on larger datasets.

Guo, D. (2019). "A Statistical Response to Challenges in Vast Portfolio Selection." PhD thesis. University of Waterloo.
The thesis is written in response to emerging issues brought about by an increasing number of assets allocated in a portfolio and seeks answers to puzzling empirical findings in the portfolio management area. Over the years, researchers and practitioners working in the portfolio optimization area have been concerned with estimation errors in the first two moments of asset returns. The thesis comprises several related chapters on our statistical inquiry into this subject. Chapter 1 of the thesis contains an introduction to what will be reported in the remaining chapters. A few well-known covariance matrix estimation methods in the literature involve adjustment of sample eigenvalues. Chapter 2 of the thesis examines the effects of sample eigenvalue adjustment on the out-of-sample performance of a portfolio constructed from the sample covariance matrix.

Guo, D., Boyle, P. P., Weng, C., and Wirjanto, T. S. (2019). "When Does The 1/N Rule Work?" In: *SSRN e-Print*.
We propose a "1/N favorability index" to measure how favorable a market is to holding a 1/N portfolio. This index reflects the extent of difficulty for an optimized portfolio to outperform the 1/N portfolio in a specific market. A single-factor model predicts that bull markets are accompanied by a high 1/N favorability index and vice versa. We validate the model implication that the 1/N portfolio is more difficult to beat in bull markets using stock return datasets from a number of countries as well as the classic datasets used by DeMiguel et al. (2009). Our results imply that the reported good performance of the 1/N portfolio in the US equity market can be partially attributed to the long-run bullish trend in the market which gives rise to the high favorability of the market to the 1/N portfolio.

Guo, L., Hardle, W. K., and Tao, Y. (2021). "A Time-Varying Network for Cryptocurrencies." In: *arXiv e-Print*.
Cryptocurrencies return cross-predictability and technological similarity yield information on risk propagation and market segmentation. To investigate these effects, we build a time-varying network for cryptocurrencies, based on the evolution of return cross-predictability and technological similarities. We develop a dynamic covariate-assisted spectral clustering method to consistently estimate the latent community structure of cryptocurrencies network that accounts for both sets of information. We demonstrate that investors can achieve better risk diversification by investing in cryptocurrencies from different communities. A cross-sectional portfolio that implements an inter-crypto momentum trading strategy earns a 1.08% daily return. By dissecting the portfolio returns on behavioral factors, we confirm that our results are not driven by behavioral mechanisms.

Gupta, K. and Chatterjee, N. (2018). "Financial Time Series Clustering." In: *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2*. Ed. by S. C. Satapathy and A. Joshi. Vol. 84. Smart Innovation, Systems and Technologies. Springer International Publishing, pp. 146–156.

Financial time series clustering finds application in forecasting, noise reduction and enhanced index tracking. The central theme in all the available clustering algorithms is the dissimilarity measure employed by the algorithm. The dissimilarity measures, applicable in financial domain, as used or suggested in past researches, are correlation based dissimilarity measure, temporal correlation based dissimilarity measure and dynamic time wrapping (DTW) based dissimilarity measure. One shortcoming of these dissimilarity measures is that they do not take into account the lead or lag existing between the returns of different stocks which changes with time. Mostly, such stocks with high value of correlation at some lead or lag belong to the same cluster (or sector). The present paper, proposes two new dissimilarity measures which show superior clustering results as compared to past measures when compared over 3 data sets comprising of 526 companies.

Haddad, M. and Bouguessa, M. (2021). "TopoDetect: Framework for Topological Features Detection in Graph Embeddings." In: *arXiv e-Print*.
TopoDetect is a Python package that allows the user to investigate if important topological features, such as the Degree of the nodes, their Triangle Count, or their Local Clustering Score, are preserved in the embeddings of graph representation models. Additionally, the framework enables the visualization of the embeddings according to the distribution of the topological features among the nodes. Moreover, TopoDetect enables us to study the effect of the preservation of these features by evaluating the performance of the embeddings on downstream learning tasks such as clustering and classification.

Haley, M. R. (2017). "K-fold cross validation performance comparisons of six naive portfolio selection rules: how naive can you be and still have successful out-of-sample portfolio performance?" In: *Annals of Finance* 13, pp. 341–353.
Recent research reports that optimal portfolio selection models often perform worse than equal-weight naive diversification in out-of-sample testing. This paper extends this line of inquiry by comparing the out-of-sample performance of the equal-weight naive strategy to the out-of-sample performance of five alternative naive strategies, each of which derives from a simple heuristic that does not require any optimization. Out-of-sample portfolio performance is assessed by mean, standard deviation, skewness, and Sharpe ratio; k-fold cross validation is used as the out-of-sample testing mechanism. The results indicate that the proposed naive heuristic rules exhibit strong out-of-sample performance, in most cases superior to the equal-weight naive strategy. These findings are consequential for at least two reasons: first, if these simple heuristic-based rules outperform the equal-weight naive strategy, then by transitivity they can outperform the mean-variance- and shortfall-optimal portfolio rules that have been shown in the literature to be inferior to the equal-weight naive rule, which further emphasizes the out-of-sample fragility of "optimal"methods; and second, among naive diversification strategies, some appear more robust in out-of-sample testing than others, hence the proposed methods may be useful when forming mixed portfolio selection models wherein a naive strategy is combined with an optimal strategy to improve performance.

Han, J. and Ge, Z. (2020). "Effect of dimensionality reduction on stock selection with cluster analysis in different market situations." In: *Expert systems with applications* 147 (113226).
Dimensionality reduction is inevitable in stock selection with cluster analysis. Considering relations among dimensionality reduction, noise trading, and market situations, we empirically investigate the effect of dimensionality-reduction methods-principal component analysis, stacked autoencoder, and stacked restricted Boltzmann machine-on stock selection with cluster analysis in different market situations. Based on the index fluctuation, the market is divided into sideways and trend situations. For the CSI 100 and Nikkei 225 constituent stocks, experimental results show that: (1) in sideways situations, dimensionality reduction hardly improves the performance of stock selection with cluster analysis; (2) the advantage of dimensionality reduction is mainly reflected in trend situations, but whether it is in an up or down trend depends on the market analyzed. More importantly, according to the above findings and assuming that the dimensionality-reduction effect will continue, we propose a rotation strategy with and without dimensionality reduction. The results of experiments show that the proposed rotation strategy outperforms the stock market indices as well as the stock-selection strategies based on dimensionality reduction and cluster analysis. These findings offer practical insights into how dimensionality reduction can be efficiently used for stock selection.

Harvey, C. R., Liu, Y., and Saretto, A. (2020). "An Evaluation of Alternative Multiple Testing Methods for Finance Applications." In: *The Review of Asset Pricing Studies* 10(2), pp. 199–248.
In almost every area of empirical finance, researchers confront multiple tests. One high-profile example is the identification of outperforming investment managers, many of whom beat their benchmarks purely by luck. Multiple testing methods are designed to control for luck. Factor selection is another glaring case in which

multiple tests are performed, but numerous other applications do not receive as much attention. One important example is a simple regression model testing five variables. In this case, because five variables are tried, a t-statistic of 2.0 is not enough to establish significance. Our paper provides a guide to various multiple testing methods and details a number of applications. We provide simulation evidence on the relative performance of different methods across a variety of testing environments. The goal of our paper is to provide a menu that researchers can choose from to improve inference in financial economics.

Heckens, A. J. and Guhr, T. (2022). "New collectivity measures for financial covariances and correlations." In: *Physica A: Statistical Mechanics and its Applications* 604, p. 127704.

Complex systems are usually non-stationary and their dynamics is often dominated by collective effects. Collectivity, defined as coherent motion of the whole system or of some of its parts, manifests itself in the time-dependent structures of covariance and correlation matrices. The largest eigenvalue corresponds to the collective motion of the system as a whole, while the other large, isolated, eigenvalues indicate collectivity in parts of the system. In the case of finance, these are industrial sectors. By removing the collective motion of the system as a whole, the latter effects are much better revealed. We measure a remaining collectivity to which we refer as average sector collectivity. We identify collective signals around the Lehman Brothers crash and after the dot-com bubble burst. For the Lehman Brothers crash, we find a potential precursor. We analyze 213 US stocks over a period of more than 30 years from 1990 to 2021. We plot the average sector collectivity versus the collectivity corresponding to the largest eigenvalue to study the whole market trajectory in a two dimensional space spanned by both collectivities. Therefore, we capture the average sector collectivity in a much more precise way. Additionally, we observe that larger values in the average sector collectivity are often accompanied by trend shifts in the mean covariances and mean correlations. As of 2015/2016 the collectivity in the US stock markets changed fundamentally.

Hens, T., Schenk-Hoppe, K. R., and Woesthoff, M.-H. (2020). "Escaping the backtesting illusion." In: *The Journal of Portfolio Management* 46(4), pp. 81–93.

Two tests can help asset managers to develop more robust investment strategies: an impact test and a survival test. Both tests complement the backtest, in which one checks how a proposed investment strategy would have performed in the past. The impact test considers the performance of the strategy when assets under management grow (crowdedness), and it checks the impact that growth in assets under management in competing strategies has on the proposed strategy (cross impact). The survival test considers the effect of the long-term evolution of assets under management in competition for market capital. Using Shiller S&P 500 index and bond market data, we show that time-series momentum (relative strength) performs best in the backtest and the impact test but that an expected relative cash-flow rule (relative dividend yield) has the best long-term survival properties.

Herteliu, C., Levantesi, S., and Rotundo, G. (2021). "Network analysis of pension funds investments." In: *Physica A: Statistical Mechanics and its Applications* 579, p. 126139.

In this paper, we analyze the Italian pension funds and their declared benchmarks, which are market indexes. Within this perspective, the amounts invested in accord to the declared benchmarks can be analyzed like as a portfolio of benchmarks. We aim at understanding whether the pension funds investments are in line with the optimal portfolios which can be built through the declared benchmarks. To achieve the results, we set up a portfolio optimization problem building two networks of pension funds: one based on the (Pearson) correlation, and the other measuring the tail correlation. For each network, we use the local clustering coefficients to describe the level of connectivity, and we insert it in the risk function. This approach allows us to consider the network measures directly in the portfolio optimization model. We compare the results with the classical Markowitz setting, and we find a new efficient frontier overperforming the Markowitz one. A comparison among the performances of pension funds and their declared portfolio of benchmarks is also reported.

Horvath, B., Issa, Z., and Muguruza, A. (2021). "Clustering Market Regimes Using the Wasserstein Distance." In: *SSRN e-Print*.

he problem of rapid and automated detection of distinct market regimes is a topic of great interest to financial mathematicians and practitioners alike. In this paper, we outline an unsupervised learning algorithm for clustering financial time-series into a suitable number of temporal segments (market regimes). As a special case of the above, we develop a robust algorithm that automates the process of classifying market regimes. The method is robust in the sense that it does not depend on modelling assumptions of the underlying time series as our experiments with real datasets show. This method – dubbed the Wasserstein k-means algorithm – frames such a problem as one on the space of probability measures with finite p^{th} moment, in terms of the p-Wasserstein distance between (empirical) distributions. We compare our WK-means approach with a more traditional clus-

tering algorithms by studying the so-called maximum mean discrepancy scores between, and within clusters. In both cases it is shown that the WK-means algorithm vastly outperforms all considered competitor approaches. We demonstrate the performance of all approaches both in a controlled environment on synthetic data, and on real data.

Hsu, Y.-C., Lin, H.-W., and Vincent, K. (2017). *Do Cross-Sectional Stock Return Predictors Pass the Test without Data-Snooping Bias?* Tech. rep. Institute of Economics Academia Sinica.
This study examines the possible data-snooping bias as a competing explanation for the anomalies in the cross-section of stock returns. We posit that the exhaustive standalone searches for profitable strategies could lead to recommending spuriously predictive variables. In order to explore the severity of this problem, we use a multiple testing method to evaluate the profitability of portfolios constructed by these predictors. Our empirical analyses suggest that over half of the findings based on individual testing method are no longer statistically significant after we adjust for data-snooping bias. Excluding the micro-cap stocks before portfolios construction and applying the notion of economic significance in this study further weaken the evidence for predictability.

Hsu, P.-H., Han, Q., Wu, W., and Cao, Z. (2018). "Asset allocation strategies, data snooping, and the 1 / N rule." In: *Journal of Banking & Finance* 97, pp. 257–269.
Using a series of advanced tests from White's (2000) Check to correct for data-snooping bias, we assess the out-of-sample performance of various portfolio strategies relative to the naive 1/N rule. When we analyze 16 basic portfolio strategies, 126 learning strategies, and nearly 2,000 extended strategies, we find that some strategies outperform the 1/N rule in conventional tests that do not account for data-snooping bias. However, after we use the new tests that control for such bias, we find that none or very few of these strategies outperform the 1/N rule. Thus, our finding underscores the necessity to control for data-snooping bias when making asset allocation decisions.

Hua, K. (2019). "Clusterability, Model Selection and Evaluation." PhD thesis. University of Massachusetts Boston.
Clustering is a central topic in unsupervised learning and has a wide variety of applications. However, the increasing needs of clustering massive datasets and the high cost of running clustering algorithms poses difficult problems for users, while to select the best clustering model with a suitable number of clusters is also a primary focus. In this thesis, we mainly focus on determining whether a data set is clusterable, and what is the natural number of clusters in a dataset.
First, we approach data clusterability from an ultrametric-based perspective. A novel approach to determine the ultrametricity of a dataset is proposed via a special type of matrix product and via this measure, we can evaluate the clusterability of it. Then, we show that our method of matrix product on the distance matrix will finally generate a sub-dominant ultrametric distance space of the original dataset. In addition, if a dataset has a unimodal or poorly constructed structure, its ultrametricity will be lower than other datasets with the same cardinality. We also show that by promoting the clusterability of a dataset, a poor clustering algorithm will perform better on the same dataset.
Secondly, we present a technique grounded in information theory for determining the natural number of clusters existent in a data set. Our approach involves a bi-criterial optimization that makes use of the entropy and the cohesion of a partition. Additionally, the experimental results are validated by using two quite distinct clustering methods: the k-means algorithm and Ward hierarchical clustering and their contour curves. We also show that by modifying the parameter, our approach can handle dataset with heavily imbalanced clustering structure, which is further complicated in practice.

Huang, Q.-A., Zhao, J.-C., and Wu, X.-Q. (2022). "Financial risk propagation between Chinese and American stock markets based on multilayer networks." In: *Physica A: Statistical Mechanics and its Applications* 586, p. 126445.
Stock networks, which are constructed from stock price time series, are useful tools for analyzing complex behaviors in stock markets. Following former researches, the epidemic model has been usually used to detect dynamic characteristics in a stock price complex systems. Recently, multilayer networks have been demonstrated well when working on heterogeneous nodes rather than integrated networks. In this paper, we proposed a two-layer SIR propagation model with an infective medium to analyze the spread of financial shocks. In consideration of strict financial regulation in the A shares, the model assumed that capital cannot flow directly between layers but through the Hong Kong stock market. By applying the model to constituent stocks included in three prominent indices, Standard & Poor 500, Shanghai and Shenzhen 300, and Hang Seng(medium), we established a two-layer Granger networks. Betweenness showed that the Hong Kong stock market had a promoting transition function of financial shocks between the US stock markets and the mainland China stock markets. In addition, with a big basic reproduction number, stock markets system appeared to be vulnerable during extreme financial

shock such as the outbreak of COVID-19 epidemic and the meltdown of stock markets. Furthermore, sensitivity analysis and the spreading simulation indicated that the US stock markets were much more robust to financial shocks than the mainland China stock markets.

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 Michuad 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.

Huang, X., Cui, P., Dong, Y., Li, J., Liu, H., Pei, J., Song, L., Tang, J., Wang, F., Yang, H., and Zhu, W. (2019). "Learning From Networks: Algorithms, Theory, and Applications." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.
Arguably, every entity in this universe is networked in one wayr another. With the prevalence of network data collected, such as social media and biological networks, learning from networks has become an essential task in many applications. It is well recognized that network data is intricate and large-scale, and analytic tasks on network data become more and more sophisticated. In this tutorial, we systematically review the area of learning from networks, including algorithms, theoretical analysis, and illustrative applications. Starting with a quick recollection of the exciting history of the area, we formulate the core technical problems. Then, we introduce the fundamental approaches, that is, the feature selection based approaches and the network embedding based approaches. Next, we extend our discussion to attributed networks, which are popular in practice. Last, we cover the latest hot topic, graph neural based approaches. For each group of approaches, we also survey the associated theoretical analysis and real-world application examples. Our tutorial also inspires a series of open problems and challenges that may lead to future breakthroughs. The authors are productive and seasoned researchers active in this area who represent a nice combination of academia and industry.

Hwang, I., Xu, S., and In, F. (2018). "Naive versus optimal diversification: Tail risk and performance." In: *European Journal of Operational Research* 265(1), pp. 372–388.
It is well documented in portfolio optimization that naive diversification outperforms optimal mean-variance diversification because the latter is subject to severe estimation error. Our study provides an alternative explanation for the outperformance of naive diversification by examining the tail risk of naive diversification relative to optimal mean-variance diversification. We utilize a rolling-sample approach and compare the out-of-sample performance and tail risk of various optimal strategies to that of the naive diversification strategy. Using portfolios consisting of individual stocks, we show that for portfolios containing relatively small number of stocks, naive diversification outperforms optimal mean-variance diversification and is less exposed to tail risk. However, for relatively large number of stocks in the portfolio, naive diversification maintains its superior performance but increases tail risk and results in more concave portfolio returns. These results imply that the outperformance of naive diversification acts as compensation for the increase in tail risk and concavity.

Ielpo, F., Merhy, C., and Simon, G. (2017). *Engineering Investment Process: Making Value Creation Repeatable*. Elsevier. 430 pp.
The book explores the quantitative steps of a financial investment process. The authors study how these steps are articulated in order to make any value creation, whatever the asset class, consistent and robust. The discussion includes factors, portfolio allocation, statistical and economic backtesting, but also the influence of negative rates, dynamical trading, state-space models, stylized facts, liquidity issues, or data biases. Besides the quantitative concepts detailed here, the reader will find useful references to other works to develop an in-depth understanding of an investment process.

Irani, J., Pise, N., and Phatak, M. (2016). "Clustering Techniques and the Similarity Measures used in Clustering: A Survey." In: *International Journal of Computer Applications* 134(7), pp. 9–14.
Clustering is an unsupervised learning technique which aims at grouping a set of objects into clusters so that objects in the same clusters should be similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in other clusters. Cluster analysis aims to group a collection of patterns into clusters based on similarity. A typical clustering technique uses a similarity function for comparing various data items.

This paper covers the survey of various clustering techniques, the current similarity measures based on distance based clustering, explains the limitations associated with the existing clustering techniques and propose that the combination of the advantages of the existing systems can help overcome the limitations of the existing systems.

Jackson, M. O. and Pernoud, A. (2020). "Systemic Risk in Financial Networks: A Survey." In: *SSRN e-Print*.
We provide an overview of the relationship between financial networks and systemic risk. We present a taxonomy of different types of systemic risk, differentiating between direct externalities between financial organizations (e.g., defaults, correlated portfolios and firesales), and perceptions and feedback effects (e.g., bank runs, credit freezes). We also discuss optimal regulation and bailouts, measurements of systemic risk and financial centrality, choices by banks' regarding their portfolios and partnerships, and the changing nature of financial networks.

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.

Jaeger, M., Krugel, S., Marinelli, D., Papenbrock, J., and Schwendner, P. (2021a). "Interpretable Machine Learning for Diversified Portfolio Construction." In: *The Journal of Financial Data Science* 3(3), pp. 31–51.
In this paper, the authors construct a pipeline to benchmark Hierarchical Risk Parity (HRP) relative to Equal Risk Contribution (ERC) as examples of diversification strategies allocating to liquid multi-asset futures markets with dynamic leverage ("volatility target"). The authors use interpretable machine learning concepts ("explainable AI") to compare the robustness of the strategies and to back out implicit rules for decision making. The empirical dataset consists of 17 equity index, government bond and commodity futures markets across 20 years. The two strategies are backtested for the empirical dataset and for about 100 000 bootstrapped datasets. XGBoost is used to regress the Calmar ratio spread between the two strategies against features of the bootstrapped datasets. Compared to ERC, HRP shows higher Calmar ratios and better matches the volatility target. Using Shapley values, the Calmar ratio spread can be attributed especially to univariate drawdown measures of the asset classes.

Jaeger, M., Krugel, S., Papenbrock, J., and Schwendner, P. (2021b). "Adaptive Seriational Risk Parity and other Extensions for Heuristic Portfolio Construction using Machine Learning and Graph Theory." In: *SSRN e-Print*.
In this article, the authors present a conceptual framework named Adaptive Seriational Risk Parity (ASRP) to extend Hierarchical Risk Parity (HRP) as an asset allocation heuristic. The first step of HRP (quasi-diagonalization) determining the hierarchy of assets is required for the actual allocation in the second step of HRP (recursive bisectioning). In the original HRP scheme, this hierarchy is found using the single-linkage hierarchical clustering of the correlation matrix, which is a static tree-based method. The authors of this paper compare the performance of the standard HRP with other static and also adaptive tree-based methods, but also seriation-based methods that do not rely on trees. Seriation is a broader concept allowing to reorder the rows or columns of a matrix to best express similarities between the elements. Each discussed variation leads to a different time series reflecting portfolio performance using a 20-year backtest of a multi-asset futures universe. An unsupervised representation learning based on this time series data creates a taxonomy that groups the strategies in high correspondence to the structure of the various types of ASRP. The performance analysis of the variations shows that most of the static tree-based alternatives of HRP outperform the single linkage

clustering used in HRP on a risk-adjusted basis. Adaptive tree methods show mixed results and most generic seriation-based approaches underperform.

Jain, P. and Jain, S. (2019). "Can Machine Learning-Based Portfolios Outperform Traditional Risk-Based Portfolios? The Need to Account for Covariance Misspecification." In: *Risks* 7(3), pp. 74+.
The Hierarchical risk parity (HRP) approach of portfolio allocation, introduced by Lopez de Prado (2016), applies graph theory and machine learning to build a diversified portfolio. Like the traditional risk-based allocation methods, HRP is also a function of the estimate of the covariance matrix, however, it does not require its invertibility. In this paper, we first study the impact of covariance misspecification on the performance of the different allocation methods. Next, we study under an appropriate covariance forecast model whether the machine learning based HRP outperforms the traditional risk-based portfolios. For our analysis, we use the test for superior predictive ability on out-of-sample portfolio performance, to determine whether the observed excess performance is significant or if it occurred by chance. We find that when the covariance estimates are crude, inverse volatility weighted portfolios are more robust, followed by the machine learning-based portfolios. Minimum variance and maximum diversification are most sensitive to covariance misspecification. HRP follows the middle ground; it is less sensitive to covariance misspecification when compared with minimum variance or maximum diversification portfolio, while it is not as robust as the inverse volatility weighed portfolio. We also study the impact of the different rebalancing horizon and how the portfolios compare against a market-capitalization weighted portfolio.

Javed, A., Lee, B. S., and Rizzo, D. M. (2020). "A Benchmark Study on Time Series Clustering." In: *arXiv e-Print*.
This paper presents the first time series clustering benchmark utilizing all time series datasets currently available in the University of California Riverside (UCR) archive – the state of the art repository of time series data. Specifically, the benchmark examines eight popular clustering methods representing three categories of clustering algorithms (partitional, hierarchical and density-based) and three types of distance measures (Euclidean, dynamic time warping, and shape-based). We lay out six restrictions with special attention to making the benchmark as unbiased as possible. A phased evaluation approach was then designed for summarizing dataset-level assessment metrics and discussing the results. The benchmark study presented can be a useful reference for the research community on its own; and the dataset-level assessment metrics reported may be used for designing evaluation frameworks to answer different research questions.

Jiang, W., Xu, Q., and Zhang, R. (2022). "Tail-event driven network of cryptocurrencies and conventional assets." In: *Finance Research Letters* 46 (Part B) (102424).
We investigate the tail risk spillover effects between cryptocurrencies and conventional assets from a systemic risk perspective, by constructing a large tail-event driven network. The results provide strong evidence for the existence of tail-risk spillovers, which challenges most literature stating the detachment of Bitcoin from traditional assets. Moreover, this paper finds two significant network factors in explaining the return of cryptocurrencies. Specifically, the risk contagion occurs under extreme market conditions, while the network diversification happens only when the market is under distress. Further sub-market analysis finds that cryptocurrencies are impacted more than stocks by the massive selloff during bear markets.

Jose-Garcia, A. and Gomez-Flores, W. (2021). "A survey of cluster validity indices for automatic data clustering using differential evolution." In: *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM.
In cluster analysis, the automatic clustering problem refers to the determination of both the appropriate number of clusters and the corresponding natural partitioning. This can be addressed as an optimization problem in which a cluster validity index (CVI) is used as a fitness function to evaluate the quality of potential solutions. Different CVIs have been proposed in the literature, aiming to identify adequate cluster solutions in terms of intracluster cohesion and intercluster separation. However, it is important to identify the scenarios in which these CVIs perform well and their limitations. This paper evaluates the effectiveness of 22 different CVIs used as fitness functions in an evolutionary clustering algorithm named ACDE based on differential evolution. Several synthetic datasets are considered: linearly separable data having both well-separated and overlapped clusters, and non-linearly separable data having arbitrarily-shaped clusters. Besides, real-life datasets are also considered. The experimental results indicate that the Silhouette index consistently reached an acceptable performance in linearly separable data. Furthermore, the indices Calinski-Harabasz, Davies-Bouldin, and generalized Dunn obtained an adequate clustering performance in synthetic and real-life datasets. Notably, all the evaluated CVIs performed poorly in clustering the non-linearly separable data because of the assumptions about data distributions.

Kalyagin, V. A., Koldanov, A. P., and Koldanov, P. A. (2021). "Reliability of MST identification in correlation-based market networks." In: *arXiv e-Print*.

Maximum spanning tree (MST) is a popular tool in market network analysis. Large number of publications are devoted to the MST calculation and it's interpretation for particular stock markets. However, much less attention is payed in the literature to the analysis of uncertainty of obtained results. In the present paper we suggest a general framework to measure uncertainty of MST identification. We study uncertainty in the framework of the concept of random variable network (RVN). We consider different correlation based networks in the large class of elliptical distributions. We show that true MST is the same in three networks: Pearson correlation network, Fechner correlation network, and Kendall correlation network. We argue that among different measures of uncertainty the FDR (False Discovery Rate) is the most appropriated for MST identification. We investigate FDR of Kruskal algorithm for MST identification and show that reliability of MST identification is different in these three networks. In particular, for Pearson correlation network the FDR essentially depends on distribution of stock returns. We prove that for market network with Fechner correlation the FDR is non sensitive to the assumption on stock's return distribution. Some interesting phenomena are discovered for Kendall correlation network. Our experiments show that FDR of Kruskal algorithm for MST identification in Kendall correlation network weakly depend on distribution and at the same time the value of FDR is almost the best in comparison with MST identification in other networks. These facts are important in practical applications.

Katsouris, C. (2021). "Optimal Portfolio Choice and Stock Centrality for Tail Risk Events." In: *arXiv e-Print*.
We propose a novel risk matrix to characterize the optimal portfolio choice of an investor with tail concerns. The diagonal of the matrix contains the Value-at-Risk of each asset in the portfolio and the off-diagonal the pairwise Delta-CoVaR measures reflecting tail connections between assets. First, we derive the conditions under which the associated quadratic risk function has a closed-form solution. Second, we examine the relationship between portfolio risk and eigenvector centrality. Third, we show that portfolio risk is not necessarily increasing with respect to stock centrality. Forth, we demonstrate under certain conditions that asset centrality increases the optimal weight allocation of the asset to the portfolio. Overall, our empirical study indicates that a network topology which exhibits low connectivity is outperformed by high connectivity based on a Sharpe ratio test.

Kawamoto, T. and Kabashima, Y. (2017a). "Comparative analysis on the selection of number of clusters in community detection." In: *arXiv e-Print*.
We conduct a comparative analysis of various model selection criteria for community detection. While the exhaustive comparison requires the test of all possible combinations of the frameworks, algorithms, and model selection criteria, we mainly focus on the statistical inference using the stochastic block model and its implementation by the EM algorithm with belief propagation. For the model selection criteria, we consider the ones that have been commonly used in the literature and the ones that are directly related to the algorithm that we consider. As we evaluate the tendency of overfit and underfit for each criterion compared to the others, we also analyze how the model-parameter learning in the EM algorithm affects the performance of model assessment. In addition, we propose that the alluvial diagram is a suitable tool to visualize the result of inference and can be useful to determine the number of clusters.

Kawamoto, T. and Kabashima, Y. (2017b). "Cross-validation estimate of the number of clusters in a network." In: *Scientific Reports* 7(1).
Network science investigates methodologies that summarise relational data to obtain better interpretability. Identifying modular structures is a fundamental task, and assessment of the coarse-grain level is its crucial step. Here, we propose principled, scalable, and widely applicable assessment criteria to determine the number of clusters in modular networks based on the leave-one-out cross-validation estimate of the edge prediction error.

Kaya, H. (2015). "Eccentricity in Asset Management." In: *Journal of Network Theory in Finance* 1(1), pp. 1–32.
We describe how networks based on information theory can help measure and visualize systemic risk, enhance diversification, and help price assets. To do this, we first define a distance measure based on the mutual information between asset pairs and use this measure in the construction of minimum spanning trees. The dynamics of the shape and the descriptive statistics of these trees are analyzed in various investment domains. The method provides evidence of regime changes in dependency structures prior to market sell-offs, and as such, it is a potential candidate for monitoring systemic risk. We also provide empirical evidence that the assets that are located towards the center of the network tend to have higher returns. Finally, an investment strategy that utilizes network centrality information is shown to add value historically.

Kaya, H. (2017). "Managing ambiguity in asset allocation." In: *Journal of Asset Management* 18(3), pp. 163–187.
This paper is about the issue of input parameter uncertainty in portfolio optimization in a discrete setting with finite states (such as the case in a world with different macroeconomic regimes). In such a setting, being unable to assign reliable point estimates to the probabilities (or frequencies) of the states creates the ambiguity. We first

describe how this ambiguity can be modeled probabilistically. Then, we show how this added uncertainty can be dealt with in optimal asset allocation problems. In simple-yet-realistic example applications we demonstrate that without sacrificing much of the upside, ambiguity managed portfolios may enhance the uniformity of returns across different states when compared to portfolios constructed by traditional methods. We stress that a key conclusion to be taken from these methods builds the case for insurance-like and potentially negative-yielding investments such as bonds and commodities so as to hedge the unforeseeable macrouncertainties for a smoother portfolio performance. Finally, we offer a variety of problem domains in which ambiguity management can be nested including macroeconomic scenario-based asset allocation, investing with regime-switching models, momentum investing, and risk-based investing.

Kazak, E. and Pohlmeier, W. (2019). "Testing out-of-sample portfolio performance." In: *International Journal of Forecasting* 35(2), pp. 540–554.
This paper studies the quality of portfolio performance tests based on out-of-sample returns. By disentangling the components of the out-of-sample performance, we show that the observed differences are driven largely by the differences in estimation risk. Our Monte Carlo study reveals that the puzzling empirical findings of inferior performances of theoretically superior strategies result mainly from the low power of these tests. Thus, our results provide an explanation as to why the null hypothesis of equal performance of the simple equally-weighted portfolio compared to many theoretically-superior alternative strategies cannot be rejected in many out-of-sample horse races. Our findings turn out to be robust with respect to different designs and the implementation strategies of the tests. For the applied researcher, we provide some guidance as to how to cope with the problem of low power. In particular, we make use of a novel pretest-based portfolio strategy to show how the information regarding performance tests can be used optimally.

Kazak, E. and Pohlmeier, W. (2020). *Portfolio Pretesting with Machine Learning*. Tech. rep. University of Lancaster.
This paper exploits the idea of pretesting to choose between competing portfolio strategies. We propose an estimator for a portfolio weight vector, which optimally trades off between Type I and Type II errors when choosing the best investment strategy. Furthermore we accommodate the idea of bagging in the portfolio testing problems, which helps to avoid sharp thresholding and reduces the amount of portfolio turnover. Our approach borrows from both shrinkage and forecast combination literature. The portfolio weights of our strategy are weighted averages of the portfolio weights from a set of stand-alone strategies. More specifically, the weights are generated from a pseudo out-of-sample portfolio pretesting, such that they reflect the probability that a given strategy will be overall best performing. Contrary to previous approaches, the shrinkage intensity is continuously updated to incorporate the most recent information in the rolling window forecasting set-up. We show that the bagged pretest estimator performs exceptionally well, especially when combined with adaptive smoothing. The resulting strategy allows for a flexible and smooth switch between the underlying strategies and is shown to outperform the corresponding stand-alone strategies.

Keranovic, V., Begusic, S., and Kostanjcar, Z. (2020). "Estimating the Number of Latent Factors in High-Dimensional Financial Time Series." In: *International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE.
Various methods for modelling financial risk rely on factor models which assume that a smaller number of latent factors are responsible for a significant portion of the observed price dynamics. A critical step for accurate estimation of these factors is obtaining the true number of factors, which is additionally problematic in high-dimensional settings and in presence of heavy tailed data - both of which are common circumstances in financial time series. In this paper we propose a method for estimating the number of latent factors that tackles these issues. To find the number of factors, the method relies on properties of optimal portfolios estimated from the covariance matrices, given by the estimated factor structures. We also introduce a simulation environment for evaluating the selection of the number of factors on high-dimensional data with heavy tailed distributions, and test the performance of the proposed method against some well known estimators such as the Marcenko-Pastur law and parallel analysis. The results suggest that our method works very well and delivers more accurate and remarkably stable results.

Kinlaw, W. B., Kritzman, M., Page, S., and Turkington, D. (2021). "The Myth of Diversification Reconsidered." In: *The Journal Of Portfolio Management* 47(8).
That investors should diversify their portfolios is a core principle of modern finance. Yet there are some periods where diversification is undesirable. When the portfolio's main growth engine performs well, investors prefer the opposite of diversification. An ideal complement to the growth engine would provide diversification when it performs poorly and unification when it performs well. Numerous studies have presented evidence of asymmetric

correlations between assets. Unfortunately, this asymmetry is often of the undesirable variety: it is characterized by downside unification and upside diversification. In other words, diversification often disappears when it is most needed. In this article we highlight a fundamental flaw in the way that some prior studies have measured correlation asymmetry. Because they estimate downside correlations from subsamples where both assets perform poorly, they ignore instances of "successful" diversification; that is, periods where one asset's gains offset the other's losses. We propose instead that investors measure what matters: the degree to which a given asset diversifies the main growth engine when it underperforms. This approach yields starkly different conclusions, particularly for asset pairs with low full sample correlation. In this paper we review correlation mathematics, highlight the flaw in prior studies, motivate the correct approach, and present an empirical analysis of correlation asymmetry across major asset classes.

Kolrep, M., Lohre, H., Radatz, E., and Rother, C. (2020). "Economic Versus Statistical Clustering in Multi-Asset Multi-Factor Strategies." In: *Risk & Reward*, pp. 26–32.
Maximizing for diversification in the multi-asset multi-factor universe, the literature advances diversified risk parity strategies across economic clusters. For handling overly complex correlation matrices, hierarchical clustering techniques have recently been put forward to guide risk parity allocations. Indeed, such statistical clusters might be considered natural portfolio building blocks given that they automatically pick up the dependence structure and thus form meaningful ingredients to aid portfolio diversification. We explain the intuition and nature of hierarchical clustering techniques in the context of multi-asset multi-factor investing vis-a-vis the use of economic factors in diversified risk-based allocation paradigms such as 1 over N, minimum-variance and diversified risk parity.

Konstantinov, G., Chorus, A., and Rebmann, J. (2020). "A network and machine learning approach to factor, asset, and blended allocation." In: *The Journal of Portfolio Management* 46 (6), pp. 54–71.
The main idea of this article is to approach and compare factor and asset allocation portfolios using both traditional and alternative allocation techniques: inverse variance optimization, minimum-variance optimization, and centrality-based techniques from network science. Analysis of the interconnectedness between assets and factors shows that their relationship is strong. The authors compare the allocation techniques, considering centrality and hierarchal-based networks. They demonstrate the advantages of graph theory to explain the advantages to portfolio management and the dynamic nature of assets and factors with their importance score. They find that asset allocation can be efficiently derived using directed networks, dynamically driven by both US Treasuries and currency returns with significant centrality scores. Alternatively, the inverse variance weight estimation and correlation-based networks generate factor allocation with favorable risk-return parameters. Furthermore, factor allocation is driven mostly by the importance scores of the Fama-French-Carhart factors: SMB, HML, CMA, RMW, and MOM. The authors confirm previous results and argue that both factors and assets are interconnected with different value and momentum factors. Therefore, a blended strategy comprising factors and assets can be defensible for investors. As argued in previous research, factors are much more overcrowded than assets. Therefore, the centrality scores help to identify the crowded exposure and build diversified allocation. The authors run LASSO regressions and show how the network-based allocation can be implemented using machine learning.

Konstantinov, G. and Rusev, M. (2020). "The Bond-Equity-Fund Relation Using the Fama-French-Carhart Factors: A Practical Network Approach." In: *The Journal of Financial Data Science* 2 (1), pp. 24–44.
The main goal of this article is to show the relation between global equity and bond funds from a network perspective. The authors demonstrate the advantages of graph theory to explain the collective fund dynamics. The results show that equity and bond funds have a significant exposure to the Fama-French-Carhart factors. The authors argue that the network is dynamically driven by equity funds with their centrality scores and risk factor exposure and can transmit and amplify system-wide stress or inefficiencies in the factor bets. Using graph theory, the authors demonstrate that the return-based relationships between bond and equity funds are asymmetrical and the network is sufficiently clustered. Specifically, equity funds connect the different clusters. The HML factor is significant both on a single-fund level and as a web determinant. Therefore, investors should pay close attention to it when managing funds and deriving asset allocations. Finally, the authors provide a machine learning approach to how fund managers, plan sponsors, and analysts can derive equity-bond allocations, based on centrality scores, factor exposure, and hierarchical clustering of asymmetrically connected assets.

Konstantinov, G. S. and Simonian, J. (2020). "A Network Approach to Analyzing Hedge Fund Connectivity." In: *The Journal of Financial Data Science* 22(3) (3), pp. 55–72.

In this article, the authors investigate the hedge fund market as a network of interacting individual funds. The authors identify and analyze the most important hedge fund styles that could both affect the market and transmit systemwide shocks to other funds, individual asset classes, and beyond. The authors find that the most connected hedge fund database categories are global macro and equity long-short funds. A central result of the article is a classification of funds using clustering, in which seemingly different funds are shown to cluster based on their shared factor exposures. This finding demonstrates that investors should consider fund connectivity and their attendant importance scores rather than database classifications when measuring hedge fund risk across the business cycle. The authors also provide a forecasting framework that can be used to predict hedge fund network behavior and the impact of individual factors on the network.

Koumou, G. B. (2020). "Diversification and portfolio theory: a review." In: *Financial Markets and Portfolio Management* 34, pp. 267–312.
Diversification is one of the major components of investment decision-making under risk or uncertainty. However, paradoxically, as the 2007-2009 financial crisis revealed, the concept remains misunderstood. Our goal in writing this paper is to correct this issue by reviewing the concept in portfolio theory. The core of our review focuses on the following diversification principles: law of large numbers, correlation, capital asset pricing model and risk contribution or risk parity diversification principles. These four diversification principles are the DNA of the existing portfolio selection rules and asset pricing theories and are instrumental to the understanding of diversification in portfolio theory. We review their definition. We also review their optimality, with respect to expected utility theory, and their usefulness. Finally, we explore their measurement.

Kukreti, V., Pharasi, H. K., Gupta, P., and Kumar, S. (2020). "A Perspective on Correlation-Based Financial Networks and Entropy Measures." In: *Frontiers in Physics* 8.
In this mini-review, we critically examine the recent work done on correlation-based networks in financial systems. The structure of empirical correlation matrices constructed from the financial market data changes as the individual stock prices fluctuate with time, showing interesting evolutionary patterns, especially during critical events such as market crashes, bubbles, etc. We show that the study of correlation-based networks and their evolution with time is useful for extracting important information of the underlying market dynamics. Also, we present our perspective on the use of recently-developed entropy measures, such as structural entropy and eigen-entropy, for continuous monitoring of correlation-based networks.

Kumar, S., Bansal, A., and Chakrabarti, A. S. (2021). "Ripples on financial networks." In: *The European Journal of Finance*, Early View.
In the financial markets, asset returns exhibit collective dynamics masking individual impacts on the rest of the market. Hence, it is still an open problem to identify how shocks originating from one particular asset create spillover effects across other assets. The problem is more acute when there is a large number of simultaneously traded assets, making the identification of which asset affects which other assets even more difficult. In this paper, we construct a network of the conditional volatility series estimated from asset returns and estimate a many-dimensional VAR model with unique identification criteria based on the network topology. Because of the interlinkages across stocks, volatility shock to a particular asset propagates through the network creating a ripple effect. Our method allows us to find the exact path the ripple effect follows on the whole network of assets.

Kumari, I. and Sharma, V. (2020). "A review for the efficient clustering based on distance and the calculation of centroid." In: *International Journal of Advanced Technology and Engineering Exploration (IJATEE)* 7(63), pp. 48–52.
Clustering is helpful in different areas of interdisciplinary engineering. It helps in finding the alike element in a single label. The clustering efficiency depends on the centroid calculation and the nearest distance estimation. This paper's main aim is to review and analysis the method in finding the better clustering mechanism to extract the higher efficiency. In this regard different methods from the previous approaches have been discussed and their advantages have been highlighted. Based on the identified gaps, future suggestions have been listed for the efficient clustering mechanism.

Kuntz, L.-C. (2018). "Portfolio Strategies with Classical and Alternative Benchmarks." PhD thesis. Georg August University of Gottingen.
This dissertation addresses different key elements in portfolio management. It intends to improve and analyze influences on portfolio strategies and their performance. Likewise, it aims at the systematization and extension of benchmark specifications as well as their effect on portfolio strategies. Each chapter focuses on a different aspect of developing and implementing portfolio strategies. The dissertation seeks to contribute to the advancement of

portfolio strategies by making the performance generating process and influences on it more comprehensible and transparent. In doing so, it attempts to strengthen the awareness of the impact of the exact design of portfolio strategies and benchmarks on the resulting portfolio and its performance. The key findings of this dissertation can be summarized as follows: The benchmark specification, especially in terms of the investible universe and the inherent risk conception, has substantial influence on the explicit design and performance of portfolio strategies. In general, the specification of the benchmark and design of portfolio strategies should be carefully considered and the implementation should be well thought out. Alternative risk conceptions, such as regret risk, can be applied to portfolio selection and lead to clearly different portfolio compositions. Moreover, timing strategies can be improved by choosing a careful investment approach on the basis of distributional regressions. All empirical work 3 of this thesis has in common that it pursues different ideas to set up portfolio strategies while explicitly addressing the benchmark specification used for the implementation and evaluation of said strategies.

Kurtti, M. (2020). "How many stocks make a diversified portfolio in a continuous-time world?" MA thesis. University of Oulu.

This thesis aims to answer how many stocks make a diversified portfolio in a continuous-time world. The study investigates what are the factors determining diversification effects in a real, continuous-time, world as opposed to thoroughly studied theoretical single period world. Continuous-time world investors care about geometric, instead of arithmetic, rate of return.

We show how methodology based on information theory can be utilized in investing context. Geometric risk premium is explained by the Shannon limit and its derivative, fractional Kelly criterion. Investing world counterpart for the Shannon limit, compounding process capacity, is derived. Geometric risk premium is decomposed to single stock risk premium and diversification premium. Method for estimating diversification premium is provided. Concept of realizable risk premium is derived and used in risk averse investor diversification metrics. Diversification effect is measured as a (realizable) risk premium ratio and as a (realizable) gross compound excess wealth ratio. Both ratios are between a randomly selected portfolio of selected size and fully diversified benchmark.

We show, both analytically and empirically, that diversification in a continuous-time world is a negative price lunch as opposed to free lunch in a single period world. Investor is paid a diversification premium, implying higher geometric risk premium, for consuming a lunch. The magnitude of diversification premium difference to benchmark, the opportunity cost of foregone diversification, is shown to be equal to one half of portfolio's idiosyncratic variance scaled by squared investment fraction. To maintain a constant wealth ratio, required level of diversification for a long-term risk neutral investor is approximately directly proportional to investment time horizon length.

The factors determining required level of diversification in a continuous-time world are number of stocks in the benchmark, Sharpe ratio and variance of the benchmark, idiosyncratic variance of an average stock, investment fraction and time. At investment fraction 1.0, risk averse investor requires more than 100, 200 or 1000 stocks to achieve 90%, 95% or 99% of the maximum diversification benefit, respectively. For short-term risk neutral investor, the corresponding numbers are about 20, 40 or 200 stocks and yet significantly more for long-term risk neutral investor. The numbers increase and decrease as investment fraction increase and decrease, respectively. We find that small firms require substantially more diversification compared to large firms and that there are substantial and consistent differences in diversification premiums between investing styles.

Landi, I., Mandelli, V., and Lombardo, M. V. (2020). "reval: a Python package to determine best clustering solutions with stability-based relative clustering validation." In: *arXiv e-Print*.

Determining the best partition for a dataset can be a challenging task because of 1) the lack of a priori information within an unsupervised learning framework; and 2) the absence of a unique clustering validation approach to evaluate clustering solutions. Here we present reval: a Python package that leverages stability-based relative clustering validation methods to determine best clustering solutions as the ones that best generalize to unseen data. Statistical software, both in R and Python, usually rely on internal validation metrics, such as silhouette, to select the number of clusters that best fits the data. Meanwhile, open-source software solutions that easily implement relative clustering techniques are lacking. Internal validation methods exploit characteristics of the data itself to produce a result, whereas relative approaches attempt to leverage the unknown underlying distribution of data points looking for generalizable and replicable results. The implementation of relative validation methods can further the theory of clustering by enriching the already available methods that can be used to investigate clustering results in different situations and for different data distributions. This work aims at contributing to this effort by developing a stability-based method that selects the best clustering solution as the one

that replicates, via supervised learning, on unseen subsets of data. The package works with multiple clustering and classification algorithms, hence allowing both the automatization of the labeling process and the assessment of the stability of different clustering mechanisms.

Laur, B. (2020). "Portfolio Optimization - Can Optimizing Portfolio Outperform Naive Diversification?" In: *SSRN e-Print*.

In this study we examined the performances of mean-variance and tangency portfolio investment strategies in order to determine if optimal diversification has benefits over 1/N strategy.

Laurinaityte, N., Meinerding, C., Schlag, C., and Thimme, J. (2019). "Elephants and the Cross-Section of Expected Returns." In: *SSRN e-Print*.

The population growth of captive Asian elephants explains the cross-section of expected returns of size-value sorted portfolios with a cross-sectional R2 of 93% and a t-statistic of 4.0 for the market price of risk. One may be tempted to conclude that elephants are the new outstanding factor in empirical asset pricing. We argue that one has to be careful with such conclusions. Standard GMM cross-sectional asset pricing tests can generate spurious explanatory power for factor models when the weight on certain moment conditions is set inappropriately. In fact, by shifting the weights in the GMM, any desired level of cross-sectional fit can be attained at the price of not matching the factor means. We run placebo tests with factors that by construction do not explain the cross-section of expected returns and obtain spuriously high cross-sectional R2's. Finally, we document some examples of factor models proposed in the literature that suffer from this bias.

Lee, T.-H. and Seregina, E. (2022). "Optimal Portfolio Using Factor Graphical Lasso." In: *arXiv e-Print*.

Graphical models are a powerful tool to estimate a high-dimensional inverse covariance (precision) matrix, which has been applied for a portfolio allocation problem. The assumption made by these models is a sparsity of the precision matrix. However, when stock returns are driven by common factors, such assumption does not hold. We address this limitation and develop a framework, Factor Graphical Lasso (FGL), which integrates graphical models with the factor structure in the context of portfolio allocation by decomposing a precision matrix into low-rank and sparse components. Our theoretical results and simulations show that FGL consistently estimates the portfolio weights and risk exposure and also that FGL is robust to heavy-tailed distributions which makes our method suitable for financial applications. FGL-based portfolios are shown to exhibit superior performance over several prominent competitors including equal-weighted and Index portfolios in the empirical application for the S&P500 constituents.

Lemenkova, P. (2020). "R Libraries (dendextend and magrittr) and Clustering Package scipy.cluster of Python For Modelling Diagrams of Dendrogram Trees." In: *Carpathian Journal of Electronic and Computer Engineering* 13(3), pp. 5–12.

The paper presents a comparison of the two languages Python and R related to the classification tools and demonstrates the differences in their syntax and graphicaloutput. It indicates the functionality of R and Python packages dendextend and scipy.cluster as effective tools for the dendrogram modelling by the algorithms of sorting and ranking datasets. R and Python programming languages have been tested on a sample dataset including marine geological measurements. The work aims to detect how bathymetric data change along the 25 bathymetric profiles digitized across the Mariana Trench. The methodology includes performed hierarchical cluster analysis with dendrograms and plotted clustermap with marginal dendrograms. The statistical libraries include Matplotlib, SciPy, NumPy, Pandas by Python and dendextend, pvclust, magrittr by R. The dendrograms were compared by the model-simulated clusters of the bathymetric ranges. The results show three distinct groups of the profiles sorted by the elevation ranges with maximal depths detected in a group of profiles 19-21. The dendrogram visualization in a cluster analysis demonstrates the effective representation of the data sorting, grouping and classifying by the machine learning algorithms. The programming codes presented in this study enable to sort a dataset in a similar research aimed to group data based on the similarity of attributes. Effective visualization by dendrograms is a useful modelling tool for the geospatial management where data ranking is required. Plotting dendrograms by R, comparing to Python, presented functional and sophisticated algorithms, refined design control and fine graphical data output. The interdisciplinary nature of this work consists in application of the coding algorithms for spatial data analysis.

Leon, D., Aragon, A., Sandoval, J., Hernandez, G., Arevalo, A., and Nino, J. (2017). "Clustering algorithms for Risk-Adjusted Portfolio Construction." In: *Procedia Computer Science* 108, pp. 1334–1343.

This paper presents the performance of seven portfolios created using clustering analysis techniques to sort out assets into categories and then applying classical optimization inside every cluster to select best assets inside each asset category. The proposed clustering algorithms are tested constructing portfolios and measuring

their performances over a two month dataset of 1-minute asset returns from a sample of 175 assets of the Russell 1000 index. A three-week sliding window is used for model calibration, leaving an out of sample period of five weeks for testing. Model calibration is done weekly. Three different rebalancing periods are tested: every 1, 2 and 4 hours. The results show that all clustering algorithms produce more stable portfolios with similar volatility. In this sense, the portfolios volatilities generated by the clustering algorithms are smaller when compare to the portfolio obtained using classical Mean-Variance Optimization (MVO) over all the dataset. Hierarchical clustering algorithms achieve the best financial performance obtaining an adequate trade-off between accumulated financial returns and the risk-adjusted measure, Omega Ratio, during the out of sample testing period.

Leopold, N. and Rose, O. (2020). "UNIC: A fast nonparametric clustering." In: *Pattern Recognition* 100, p. 107117.
Clustering is among the tools for exploring, analyzing, and deriving information from data. In the case of large data sets, the real burden to the application of clustering algorithms can be their complexity and demand of control parameters. We present a new fast nonparametric clustering algorithm, UNIC, to address these challenges. To identify clusters, the algorithm evaluates the distances between selected points and other points in the set. While assessing these distances, it employs methods of robust statistics to identify the cluster borders. The performance of the proposed algorithm is assessed in an experimental study and compared with several existing clustering methods over a variety of benchmark data sets.

Li, H. and Liu, Z. (2021). "Multivariate time series clustering based on complex network." In: *Pattern Recognition* 115, p. 107919.
Recent years have seen an increase in research on time series data mining (especially time-series clustering) owing to the widespread existence of time series in various fields. Techniques such as clustering can extract valuable information and potential patterns from time-series data. In this regard, the clustering analysis of multivariate time series is challenging because of the high dimensionality. Our study led us to develop a novel method based on complex networks for multivariate time series clustering (BCNC). BCNC includes a new method for mapping multivariate time series into complex networks and a new method to visualize multivariate time series. The solution is innovatively based on a relationship network and relies on the use of community detection technology to achieve complete multivariate time series clustering. The detailed algorithm and the simulation experiments of the proposed BCNC method are reported. The experimental results on various datasets show that BCNC is superior to traditional multivariate time series clustering methods.

Li, T., Levina, E., and Zhu, J. (2020). "Network cross-validation by edge sampling." In: *Biometrika* 107(2), pp. 257–276.
While many statistical models and methods are now available for network analysis, resampling of network data remains a challenging problem. Cross-validation is a useful general tool for model selection and parameter tuning, but it is not directly applicable to networks since splitting network nodes into groups requires deleting edges and destroys some of the network structure. In this paper we propose a new network resampling strategy, based on splitting node pairs rather than nodes, that is applicable to cross-validation for a wide range of network model selection tasks. We provide theoretical justification for our method in a general setting and examples of how the method can be used in specific network model selection and parameter tuning tasks. Numerical results on simulated networks and on a statisticians' citation network show that the proposed cross-validation approach works well for model selection.

Lim, T. and Ong, C. S. (2021). "Portfolio Diversification Using Shape-Based Clustering." In: *The Journal of Financial Data Science* 3(1), pp. 111–126.
Portfolio diversification involves lowering the correlation between portfolio assets to achieve improved risk-return exposure. It is reasonable to infer from the classic Anscombe quartet that relying on descriptive statistics, and specifically, correlation, to achieve portfolio diversification may not derive the most optimal multiperiod portfolio risk-adjusted return because stocks in a portfolio can exhibit different price trends over time, even with the same computed pairwise correlation. This research applied a shape-based time-series clustering technique of agglomerative hierarchical clustering using dynamic time-series warping as a distance measure to aggregate stocks into like-trending clusters across time as a portfolio diversification tool. Results support the use of the shape-based clustering technique for (1) portfolio allocation and rebalancing, (2) dynamic predictive portfolio construction, and (3) individual stock selection through outlier identification. The findings will be a useful addition to the existing literature in portfolio management by providing shape-based clustering as an alternative tool for portfolio construction and security selection.

Lipor, J. and Balzano, L. (2020). "Clustering quality metrics for subspace clustering." In: *Pattern Recognition* 104, p. 107328.
We study the problem of clustering validation, i.e., clustering evaluation without knowledge of ground-truth labels, for the increasingly-popular framework known as subspace clustering. Existing clustering quality metrics (CQMs) rely heavily on a notion of distance between points, but common metrics fail to capture the geometry of subspace clustering. We propose a novel point-to-point pseudometric for points lying on a union of subspaces and show how this allows for the application of existing CQMs to the subspace clustering problem. We provide theoretical and empirical justification for the proposed point-to-point distance, and then demonstrate on a number of common benchmark datasets that our proposed methods generally outperform existing graph-based CQMs in terms of choosing the best clustering and the number of clusters.

Lohre, H., Rother, C., and Schafer, K. A. (2020). "Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations." In: *Machine Learning for Asset Management: New Developments and Financial Applications*. Ed. by E. Jurczenko. Wiley, pp. 329–368.
This chapter examines the use and merits of hierarchical clustering techniques in the context of multi-asset multi-factor investing. In particular, it contrasts these techniques with several competing risk-based allocation paradigms, such as 1/N, minimum-variance, standard risk parity and diversified risk parity. The chapter introduces hierarchical risk parity (HRP) strategies based on the Pearson correlation coefficient and also introduces hierarchical clustering based on the lower tail dependence coefficient. The chapter provides an overview of traditional risk-based allocation strategies and outlines a framework to measure and manage portfolio diversification. It examines the performance of the introduced HRP strategies relative to the traditional alternatives. The chapter discusses Meucci's approach to managing diversification, which serves to construct a diversified risk parity strategy based on economic factors.

Loistl, O. and Konstantinov, G. S. (2020). "Interactions and Interconnectedness Shape Financial Market Research." In: *The Journal of Financial Data Science* (2), 2, pp. 51–63.
In this article the authors investigate two fields that might be relevant for financial data sciences. The first issue covers the entire production chain from orders to prices by realistically modeling stock exchange microstructure (e.g., NASDAQ and Xetra). Specifically, the authors show how data-driven research can model decisions to place orders and to generate prices by matching orders accordingly. The other issue is price interconnectedness at markets by networks. The authors show that interactions shape a market performance. Emergence comprises the interactions at markets; as such, the collective may not be equal to the sum of individual activities. As a consequence, the assumption that markets are in equilibrium and that arbitrage opportunities do not exist can be replaced by more realistic working hypotheses. The authors show with the two examples that market participants interact, learn, and trade. These individual interactions can be described as organized complexity. Whereas calculus may not support explicit modeling of interactions, the age of big data permits their modeling and application of innovative concepts, such as network solutions for asset allocation, which can be modeled using machine learning. This article illustrates that assertion with concrete examples.

Lopez de Prado, M. (2016). "Building Diversified Portfolios that Outperform Out of Sample." In: *The Journal of Portfolio Management* 42(4), pp. 59–69.
In this article, the author introduces the Hierarchical Risk Parity (HRP) approach to address three major concerns of quadratic optimizers, in general, and Markowitz's critical line algorithm (CLA), in particular: instability, concentration, and underperformance. HRP applies modern mathematics (graph theory and machine-learning techniques) to build a diversified portfolio based on the information contained in the covariance matrix. However, unlike quadratic optimizers, HRP does not require the invertibility of the covariance matrix. In fact, HRP can compute a portfolio on an ill-degenerated or even a singular covariance matrix an impossible feat for quadratic optimizers. Monte Carlo experiments show that HRP delivers lower out-ofsample variance than CLA, even though minimum variance is CLA's optimization objective. HRP also produces less risky portfolios out of sample compared to traditional risk parity methods.

Lopez de Prado, M. (2019a). "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." In: *The Journal of Financial Data Science* 1(1), pp. 99–110.
Most discoveries in empirical finance are false, as a consequence of selection bias under multiple testing. Although many researchers are aware of this problem, the solutions proposed in the literature tend to be complex and hard to implement. In this article, the author reduces the problem of selection bias in the context of investment strategy development to two sub-problems: determining the number of essentially independent trials and determining the variance across those trials. The author explains what data researchers need to report to allow others to

evaluate the effect that multiple testing has had on reported performance. He applies his method to a real case of strategy development and estimates the probability that a discovered strategy is false.

Lopez de Prado, M. (2019b). "Estimation of Theory-Implied Correlation Matrices." In: *SSRN e-Print*.
Correlation matrices are ubiquitous in finance. Some key applications include portfolio construction, risk management, and factor/style analysis. Correlation matrices are usually estimated from historical empirical observations or derived from historically estimated factors. It is widely acknowledged that empirical correlation matrices: (a) have poor numerical properties that lead to unreliable estimators; and (b) have poor predictive power. Additionally, factor-based correlation matrices have their own caveats. In particular, estimated factors are typically non-hierarchical and do not allow for interactions at different levels. This contravenes the fact that financial instruments typically exhibit a nested cluster structure (e.g., MSCI GICS levels 1-4).This paper introduces a machine learning (ML) algorithm to estimate forward-looking correlation matrices implied by economic theory. Given a particular theoretical representation of the hierarchical structure that governs a universe of securities, the method fits the correlation matrix that complies with that theoretical representation of the future. This particular use case demonstrates how, contrary to popular perception, ML solutions are not black-boxes, and can be applied effectively to develop and test economic theories.

Lopez de Prado, M. (2020a). "Clustering." In: *SSRN e-Print*.
Many problems in finance require the clustering of variables or observations. Despite its usefulness, clustering is almost never taught in Econometrics courses. In this seminar we review two general clustering approaches: partitional and hierarchical.

Lopez de Prado, M. (2020b). *Machine learning for asset managers*. Cambridge University Press. 190 pp.
Successful investment strategies are specific implementations of general theories. An investment strategy that lacks a theoretical justification is likely to be false. Hence, an asset manager should concentrate her efforts on developing a theory rather than on backtesting potential trading rules. The purpose of this Element is to introduce machine learning (ML) tools that can help asset managers discover economic and financial theories. ML is not a black box, and it does not necessarily overfit. ML tools complement rather than replace the classical statistical methods. Some of ML's strengths include (1) a focus on out-of-sample predictability over variance adjudication; (2) the use of computational methods to avoid relying on (potentially unrealistic) assumptions; (3) the ability to learn complex specifications, including nonlinear, hierarchical, and noncontinuous interaction effects in a high-dimensional space; and (4) the ability to disentangle the variable search from the specification search, robust to multicollinearity and other substitution effects.

Lopez de Prado, M. and Lewis, M. J. (2019). "Detection of false investment strategies using unsupervised learning methods." In: *Quantitative Finance* 19(9), pp. 1555–1565.
In this paper we address the problem of selection bias under multiple testing in the context of investment strategies. We introduce an unsupervised learning algorithm that determines the number of effectively uncorrelated trials carried out in the context of a discovery. This estimate is critical for computing the familywise false positive probability, and for filtering out false investment strategies.

Louiset, R., Gori, P., Dufumier, B., Houenou, J., Grigis, A., and Duchesnay, E. (2021). "UCSL : A Machine Learning Expectation-Maximization framework for Unsupervised Clustering driven by Supervised Learning." In: *arXiv e-Print*.
Subtype Discovery consists in finding interpretable and consistent sub-parts of a dataset, which are also relevant to a certain supervised task. From a mathematical point of view, this can be defined as a clustering task driven by supervised learning in order to uncover subgroups in line with the supervised prediction. In this paper, we propose a general Expectation-Maximization ensemble framework entitled UCSL (Unsupervised Clustering driven by Supervised Learning). Our method is generic, it can integrate any clustering method and can be driven by both binary classification and regression. We propose to construct a non-linear model by merging multiple linear estimators, one per cluster. Each hyperplane is estimated so that it correctly discriminates - or predict - only one cluster. We use SVC or Logistic Regression for classification and SVR for regression. Furthermore, to perform cluster analysis within a more suitable space, we also propose a dimension-reduction algorithm that projects the data onto an orthonormal space relevant to the supervised task. We analyze the robustness and generalization capability of our algorithm using synthetic and experimental datasets. In particular, we validate its ability to identify suitable consistent sub-types by conducting a psychiatric-diseases cluster analysis with known ground-truth labels. The gain of the proposed method over previous state-of-the-art techniques is about +1.9 points in terms of balanced accuracy. Finally, we make codes and examples available in a scikit-learn-compatible Python package at `https://github.com/neurospin-projects/2021_rlouiset_ucsl`.

Lu, Y., Li, M., Tang, X., and Wang, H. (2021). "A Cluster Representative Selection Method for Stock Portfolio Based on Efficient Frontier." In: *IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE.
Portfolio is a financial concept to combine several stocks to reduce the risks and improve the profits. To choose the basic members of portfolio, we can group similar stocks into one cluster and then choose representative stock from each cluster. In this paper, we focus on the method of choosing representative stocks in clusters. The ordinary representative of a cluster is often the center of that cluster. We propose a new cluster representative method MDR (maximum distance representatives). In our method MDR, we choose the stocks which has maximum distance with other representatives. MDR can construct a more diverse portfolio than center method. The effectiveness of cluster representative selection methods can be evaluated by an index IBEF based on the concept of efficient frontier. Our experiments show that MDR can effectively improve the efficient frontier, which means MDR can bring more profits than center representative method at the same risk level.

Ma, E. (2020). *Network Analysis Made Simple*. URL: https://ericmjl.github.io/Network-Analysis-Made-Simple/.
Network Analysis Made Simple is a collection of Jupyter notebooks designed to help you get up and running with the NetworkX package in the Python programming langauge. It's written by programmers for programmers, and will give you a basic introduction to graph theory, applied network science, and advanced topics to help kickstart your learning journey. There's even case studies to help those of you for whom example narratives help a ton!

Ma, E. and Seth, M. (2020). *Network Analysis Made Simple*. 278 pp.
An introduction to network analysis and applied graph theory using Python and NetworkX.

Magner, N. S., Lavin, J. F., Valle, M. A., and Hardy, N. (2021). "The Volatility Forecasting Power of Financial Network Analysis." In: *Complexity* 2020 (7051402).
This investigation connects two crucial economic and financial fields, financial networks, and forecasting. From the financial network's perspective, it is possible to enhance forecasting tools, since econometrics does not incorporate into standard economic models, second-order effects, nonlinearities, and systemic structural factors. Using daily returns from July 2001 to September 2019, we used minimum spanning tree and planar maximally filtered graph techniques to forecast the stock market realized volatility of 26 countries. We test the predictive power of our core models versus forecasting benchmarks models in and out of the sample. Our results show that the length of the minimum spanning tree is relevant to forecast volatility in European and Asian stock markets, improving forecasting models' performance. As a new contribution, the evidence from this work establishes a road map to deepening the understanding of how financial networks can improve the quality of prediction of financial variables, being the latter, a crucial factor during financial shocks, where uncertainty and volatility skyrocket.

Maharaj, E. A., D'Urso, P., and Caiado, J. (2019). *Time Series Clustering and Classification*. CRC Press. 244 pp.
Time Series Clustering and Classification includes relevant developments on observation-based, feature-based and model-based traditional and fuzzy clustering methods, feature-based and model-based classification methods, and machine learning methods. It presents a broad and self-contained overview of techniques for both researchers and students.

Mahfuz, N. M., Yusoff, M., and Ahmad, Z. (2019). "Review of single clustering methods." In: *IAES International Journal of Artificial Intelligence (IJ-AI)* 8(3), p. 221.
Clustering provides a prime important role as an unsupervised learning method in data analytics to assist many real-world problems such as image segmentation, object recognition or information retrieval. It is often an issue of difficulty for traditional clustering technique due to non-optimal result exist because of the presence of outliers and noise data. This review paper provides a review of single clustering methods that were applied in various domains. The aim is to see the potential suitable applications and aspect of improvement of the methods. Three categories of single clustering methods were suggested, and it would be beneficial to the researcher to see the clustering aspects as well as to determine the requirement for clustering method for an employment based on the state of the art of the previous research findings.

Malavasi, M., Lozza, S. O., and Truck, S. (2021). "Second order of stochastic dominance efficiency vs mean variance efficiency." In: *European Journal of Operational Research* 290(3), pp. 1192–1206.
In this paper, we compare two of the main paradigms of portfolio theory: mean variance analysis and expected utility. In particular, we show empirically that mean variance efficient portfolios are typically sub-optimal for non satiable and risk averse investors. We illustrate that the second order stochastic dominance (SSD) efficient

set is the solution of a multi-objective optimization problem. We further show that the market portfolio is not necessarily a solution to this optimization problem. We also conduct an empirical analysis, examining the ex ante and ex post performance of SSD and mean variance efficient portfolios, using a bootstrap approach. In an ex ante analysis, we compare empirical moments, the level of diversification and set distances of mean variance and SSD efficient sets. We also show that the global minimum variance (GMV) portfolio and the part of the mean variance efficient frontier (MVEF) composed of highly diversified portfolios is second order stochastically dominated. This result also provides a possible alternative explanation for the diversification puzzle. Conducting an ex post analysis, we construct second order stochastic dominating strategies that outperform the GMV portfolio in terms of wealth and various other performance measures, producing a positive ex post opportunity cost.

Marti, G., Nielsen, F., Bihkowski, M., and Donnat, P. (2021). "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.

Massahi, M., Mahootchi, M., and Khamseh, A. A. (2020). "Development of an efficient cluster-based portfolio optimization model under realistic market conditions." In: *Empirical Economics*.
Modern portfolio theory introduced by Markowitz in 1952 is the most popular portfolio optimization framework established based on the trade-off between risk and return as an operation research model. The main shortcoming of applying Markowitz portfolio optimization in practice is that the obtained optimal weights are really sensitive to the embedded uncertainty in return series of stocks. In this paper, it is demonstrated how using a new methodology of time series clustering as a remedy can lead to a more robust and accurate portfolio in terms of the gap between mean variance efficient frontier obtained from the optimization model and the one observed in reality. In this regard, two similarity measures, the autocorrelation coefficients and the weighted dynamic time warping, are used in an innovative way to construct the desired portfolio optimization model. Moreover, the effectiveness of proposed approach is investigated in two different market conditions: semi-realistic and full-realistic. In the first one, it is assumed that the forecasted and realized stocks mean returns are the same; however, these returns are not necessarily equal in the second market conditions. Finally, a database of stock prices from the literature is utilized to show the robustness and accuracy of the proposed approach in empirical results in comparison with applied similarity measures in previous researches.

Mattera, R., Giacalone, M., and Gibert, K. (2021). "Distribution-Based Entropy Weighting Clustering of Skewed and Heavy Tailed Time Series." In: *Symmetry* 13(6), p. 959.
The goal of clustering is to identify common structures in a data set by forming groups of homogeneous objects. The observed characteristics of many economic time series motivated the development of classes of distributions that can accommodate properties, such as heavy tails and skewness. Thanks to its flexibility, the skewed exponential power distribution (also called skewed generalized error distribution) ensures a unified and general framework for clustering possibly skewed and heavy tailed time series. This paper develops a clustering procedure of model-based type, assuming that the time series are generated by the same underlying probability distribution but with different parameters. Moreover, we propose to optimally combine the estimated parameters to form the clusters with an entropy weighing k-means approach. The usefulness of the proposal is shown by means of application to financial time series, demonstrating also how the obtained clusters can be used to form portfolio of stocks.

McCabe, S., Torres, L., LaRock, T., Haque, S. A., Yang, C.-H., Hartle, H., and Klein, B. (2020). "netrd: A library for network reconstruction and graph distances." In: *arXiv e-Print*.
Over the last two decades, alongside the increased availability of large network datasets, we have witnessed the rapid rise of network science. For many systems, however, the data we have access to is not a direct description of the underlying network. More and more, we see the drive to study networks that have been inferred or reconstructed from non-network data—in particular, using time series data from the nodes in a system to infer likely connections between them. Selecting the most appropriate technique for this task is a challenging problem in network science. Different reconstruction techniques usually have different assumptions, and their

performance varies from system to system in the real world. One way around this problem could be to use several different reconstruction techniques and compare the resulting networks. However, network comparison is also not an easy problem, as it is not obvious how best to quantify the differences between two networks, in part because of the diversity of tools for doing so. The netrd Python package seeks to address these two parallel problems in network science by providing, to our knowledge, the most extensive collection of both network reconstruction techniques and network comparison techniques (often referred to as graph distances) in a single library (`https://github.com/netsiphd/netrd`). In this article, we detail the two main functionalities of the netrd package. Along the way, we describe some of its other useful features. This package builds on commonly used Python packages and is already a widely used resource for network scientists and other multidisciplinary researchers. With ongoing open-source development, we see this as a tool that will continue to be used by all sorts of researchers to come.

Mehta, V., Bawa, S., and Singh, J. (2020). "Analytical review of clustering techniques and proximity measures." In: *Artificial Intelligence Review* 53(8), pp. 5995–6023.
One of the most fundamental approaches to learn and understand from any type of data is by organizing it into meaningful groups (or clusters) and then analyzing them, which is a process known as cluster analysis. During this process of grouping, proximity measures play a significant role in deciding the similarity level of two objects. Moreover, before applying any learning algorithm on a dataset, different aspects related to preprocessing such as dealing with the sparsity of data, leveraging the correlation among features and normalizing the scales of different features are required to be considered. In this study, various proximity measures have been discussed and analyzed from the aforementioned aspects. In addition, a theoretical procedure for selecting a proximity measure for clustering purpose is proposed. This procedure can also be used in the process of designing a new proximity measure. Second, clustering algorithms of different categories have been overviewed and experimentally compared for various datasets of different domains. The datasets have been chosen in such a way that they range from a very low number of dimensions to a very high number of dimensions. Finally, the effect of using different proximity measures is analyzed in partitional and hierarchical clustering techniques based on experiments.

Millington, T. and Niranjan, M. (2020a). "Construction of Minimum Spanning Trees from Financial Returns using Rank Correlation." In: *arXiv e-Print*.
The construction of minimum spanning trees (MSTs) from correlation matrices is an often used method to study relationships in the financial markets. However most of the work on this topic tends to use the Pearson correlation coefficient, which relies on the assumption of normality and can be brittle to the presence of outliers, neither of which is ideal for the study of financial returns. In this paper we study the inference of MSTs from daily US financial returns using Pearson and two rank correlation methods, Spearman and Kendall's tau. We find that the trees constructed using these rank methods tend to be more stable and maintain more edges over the dataset than those constructed using Pearson correlation, that there are significant differences in the agreement of the centrality of various sectors and that despite these, the trees tend to have similar topologies.

Millington, T. and Niranjan, M. (2020b). "Partial correlation financial networks." In: *Applied Network Science* 5(1) (11).
Correlation networks have been a popular way of inferring a financial network due to the simplicity of construction and the ease of interpretability. However two variables which share a common cause can be correlated, leading to the inference of spurious relationships. To solve this we can use partial correlation. In this paper we construct both correlation and partial correlation networks from S&P500 returns and compare and contrast the two. Firstly we show that the partial correlation networks have a smaller and much less variable intensity than the correlation networks, but in fact are less stable. We look at the centrality of the various sectors in the graph using degree centrality and eigenvector centrality, finding that sector centralities move together during the 2009 market crash and that the financial sector generally has a higher mean centrality over most of the dataset. Exploring the use of these centrality measures for portfolio construction, we shown there is mild correlation between the in-sample centrality and the out of sample Sharpe ratio but there is negative correlation between the in-sample centrality and out of sample risk. Finally we use a community detection method to study how the networks reflect the underlying sector structure and study how stable these communities are over time.

Millington, T. and Niranjan, M. (2021). "Stability and similarity in financial networks – How do they change in times of turbulence?" In: *Physica A: Statistical Mechanics and its Applications* 574, p. 126016.
Diversified portfolios are a key component of modern portfolio theory, based on the idea of choosing uncorrelated or unrelated stocks to minimize risk. With this in mind, we use networks to study the correlations between stocks and how this varies over time, using daily returns from the S&P500 (US), FTSE100 (UK) and DAX30 (Germany).

We study both the full correlation networks and those filtered using the PMFG method. We conclude that stocks tend to become more similar in the full correlation networks during times of market disruption for the US and UK markets - implying that nodes that were once dissimilar (and therefore a good choice for a low risk portfolio) are no longer so, demonstrating the difficulties of choosing a diversified portfolio. Furthermore, these full networks are also more stable by certain measures during these periods of disruption, contrary to expectations. However, these apply less to the PMFGs and the German market.

Miranda, F. M., Koehnecke, N., and Renard, B. Y. (2022). "HiClass: a Python library for local hierarchical classification compatible with scikit-learn." In: *arXiv e-Print*.
HiClass is an open-source Python package for local hierarchical classification fully compatible with scikit-learn. It provides implementations of the most popular machine learning models for local hierarchical classification, including Local Classifier Per Node, Local Classifier Per Parent Node and Local Classifier Per Level. In addition, the library includes tools to evaluate model performance on hierarchical data. The documentation contains installation instructions, interactive notebooks, and a complete description of the API. HiClass is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings. Source code and documentation are available at `https://gitlab.com/dacs-hpi/hiclass`.

Molyboga, M. (2020). "A Modified Hierarchical Risk Parity Framework for Portfolio Management." In: *The Journal of Financial Data Science* 2(3), pp. 128–139.
This article introduces a modified hierarchical risk parity (MHRP) approach that extends the HRP approach by incorporating three intuitive elements commonly used by practitioners. The new approach (1) replaces the sample covariance matrix with an exponentially weighted covariance matrix with Ledoit-Wolf shrinkage; (2) improves diversification across portfolio constituents both within and across clusters by relying on an equal volatility, rather than an inverse variance, allocation approach; and (3) improves diversification across time by applying volatility targeting to portfolios. The author examines the impact of the enhancements on portfolios of commodity trading advisors within a large-scale Monte Carlo simulation framework that accounts for the realistic constraints of institutional investors. The author finds a striking improvement in the out-of-sample Sharpe ratio of 50%, on average, along with a reduction in downside risk.

Montero, P. and Vilar, J. A. (2015). "TSclust: An R Package for Time Series Clustering." In: *Journal of Statistical Software* 62.
Time series clustering is an active research area with applications in a wide range of fields. One key component in cluster analysis is determining a proper dissimilarity measure between two data objects, and many criteria have been proposed in the literature to assess dissimilarity between two time series. The R package TSclust is aimed to implement a large set of well-established peer-reviewed time series dissimilarity measures, including measures based on raw data, extracted features, underlying parametric models, complexity levels, and forecast behaviors. Computation of these measures allows the user to perform clustering by using conventional clustering algorithms. TSclust also includes a clustering procedure based on p values from checking the equality of generating models, and some utilities to evaluate cluster solutions. The implemented dissimilarity functions are accessible individually for an easier extension and possible use out of the clustering context. The main features of TSclust are described and examples of its use are presented.

Mooney, T., Rapaka, R., and Vera, T. (2020). "Dynamic Regime Strategy for Stress Testing and Optimizing Institutional Investor Portfolios." In: *SSRN e-Print*.
Our work aims to develop a stand-alone trading system to construct portfolios that show the benefits of value and momentum style integration and presents the effectiveness of alternative integration methods for long-only absolute return funds, which seeks absolute returns that are not highly correlated to traditional assets such as stocks and bonds. Our approach uses the CRoss Industry Standard Process for Data Mining (CRISP-DM) model to guide the necessary steps, processes, and workflows for executing our project.

Mori, U., Mendiburu, A., and Lozano, J. A. (2016). "Distance Measures for Time Series in R: The TSdist Package." In: *The R Journal*.
The definition of a distance measure between time series is crucial for many time series data mining tasks, such as clustering and classification. For this reason, a vast portfolio of time series distance measures has been published in the past few years. In this paper, the TSdist package is presented, a complete tool which provides a unified framework to calculate the largest variety of time series dissimilarity measures available in R at the moment, to the best of our knowledge. The package implements some popular distance measures which were not previously available in R, and moreover, it also provides wrappers for measures already included in other R packages. Additionally, the application of these distance measures to clustering and classification tasks is also

supported in TSdist, directly enabling the evaluation and comparison of their performance within these two frameworks.

Murialdo, P., Ponta, L., and Carbone, A. (2021). "Inferring multi-period optimal portfolios via detrending moving average cluster entropy." In: *EPL (Europhysics Letters)* 133(6), p. 60004.
Despite half a century of research, there is still no general agreement about the optimal approach to build a robust multi-period portfolio. We address this question by proposing the detrended cluster entropy approach to estimate the weights of a portfolio of high-frequency market indices. The information measure gathered from the markets produces reliable estimates of the weights at varying temporal horizons. The portfolio exhibits a high level of diversity, robustness and stability as not affected by the drawbacks of traditional mean-variance approaches.

Nanakorn, N. and Palmgren, E. (2021). "Hierarchical Clustering in Risk-Based Portfolio Construction." MA thesis. KTH.
Following the global financial crisis, both risk-based and heuristic portfolio construction methods have received much attention from both academics and practitioners since these methods do not rely on the estimation of expected returns and as such are assumed to be more stable than Markowitz's traditional mean-variance portfolio. In 2016, Lopez de Prado presented the Hierarchical Risk Parity (HRP), a new approach to portfolio construction which combines hierarchical clustering of assets with a heuristic risk-based allocation strategy in order to increase stability and improve out-of-sample performance. Using Monte Carlo simulations, Lopez de Prado was able to demonstrate promising results.
This thesis attempts to evaluate HRP using walk-forward analysis and historical data from equity index and bond futures, against more realistic benchmark methods and using additional performance measures relevant to practitioners. The main conclusion is that applying hierarchical clustering to risk-based portfolio construction does indeed improve the out-of-sample return and Sharpe ratio. However, the resulting portfolio is also associated with a remarkably high turnover, which may indicate numerical instability and sensitivity to estimation errors. It is also identified that Lopez de Prado's original HRP approach has an undesirable property and alternative approaches to HRP have consequently been developed. Compared to Lopez de Prado's original HRP approach, these alternative approaches increase the Sharpe ratio with 10% and reduce the turnover with 60-65%. However, it should be noted that compared to more mainstream portfolios the turnover is still rather high, indicating that these alternative approaches to HRP are still somewhat unstable and sensitive to estimation errors.

Naraoka, M., Hayashi, T., Yoshino, T., Sugie, T., Takano, K., and Ohsawa, Y. (2020). "Detecting and explaining changes in various assets' relationships in financial markets." In: *arXiv e-Print*.
We study the method for detecting relationship changes in financial markets and providing human-interpretable network visualization to support the decision-making of fund managers dealing with multi-assets. First, we construct co-occurrence networks with each asset as a node and a pair with a strong relationship in price change as an edge at each time step. Second, we calculate Graph-Based Entropy to represent the variety of price changes based on the network. Third, we apply the Differential Network to finance, which is traditionally used in the field of bioinformatics. By the method described above, we can visualize when and what kind of changes are occurring in the financial market, and which assets play a central role in changes in financial markets. Experiments with multi-asset time-series data showed results that were well fit with actual events while maintaining high interpretability. It is suggested that this approach is useful for fund managers to use as a new option for decision making.

Olmo, J. (2021). "Optimal portfolio allocation and asset centrality revisited." In: *Quantitative Finance* 21(9), pp. 1475–1490.
This paper revisits the relationship between eigenvector asset centrality and optimal asset allocation in a minimum variance portfolio. We show that the standard definition of eigenvector centrality is misleading when the adjacency matrix in a network can take negative values. This is, for example, the case when the network topology is induced by the correlation matrix between assets in a portfolio. To correct for this, we introduce the concept of positive and negative eigenvector centrality. Our results show that the loss function associated to the minimum variance portfolio is positively/negatively related to the positive and negative eigenvector centrality under short-selling constraints but cannot be generalized beyond that. Furthermore, in contrast to what is claimed in the related literature, this relationship does not imply any monotonic relationship between the centrality of an asset and its optimal portfolio allocation. These theoretical insights are illustrated empirically in a portfolio allocation exercise with assets from U.S. and U.K. financial markets.

Page, S. and Panariello, R. A. (2018). "When Diversification Fails." In: *Financial Analysts Journal* 74(3), pp. 19–32.

One of the most vexing problems in investment management is that diversification seems to disappear when investors need it the most. We surmise that many investors still do not fully appreciate the impact of extreme correlations on portfolio efficiency particular, on exposure to loss. We take an in-depth look at what drives the stock-to-credit, stock-to-hedge fund, stock-to-private asset, stock-to-risk factors, and stock-to-bond correlations during tail events. We introduce a data-augmentation technique to improve the robustness of tail correlation estimates. Finally, we discuss implications for multi-asset investing.

Pang, R. K.-K., Granados, O. M., Chhajer, H., and Legara, E. F. T. (2021). "An analysis of network filtering methods to sovereign bond yields during COVID-19." In: *Physica A: Statistical Mechanics and its Applications* 574, p. 125995.

In this work, we investigate the impact of the COVID-19 pandemic on sovereign bond yields. We consider the temporal changes from financial correlations using network filtering methods. These methods consider a subset of links within the correlation matrix, which gives rise to a network structure. We use sovereign bond yield data from 17 European countries between the 2010 and 2020 period. We find the mean correlation to decrease across all filtering methods during the COVID-19 period. We also observe a distinctive trend between filtering methods under multiple network centrality measures. We then relate the significance of economic and health variables towards filtered networks within the COVID-19 period. Under an exponential random graph model, we are able to identify key relations between economic groups across different filtering methods.

Papenbrock, J., Schwendner, P., Jaeger, M., and Krugel, S. (2021a). "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.

Papenbrock, J., Schwendner, P., and Sandner, P. (2021b). "Can Adaptive Seriational Risk Parity Tame Crypto Portfolios?" In: *SSRN e-Print*.

As cryptocoins are not tied to fundamental values or to investor protection regulation, their price dynamics is unhinged in both directions. In institutional asset management of conventional asset classes, target volatility concepts and dynamic allocation heuristics are popular to improve the robustness of portfolio. Can similar techniques also be used to construct delevered and diversified portfolios of crypto assets? A robust candidate approach for allocation is Hierarchical Risk Parity (HRP), as it incorporates a filtered correlation structure and is less sensitive to noise than quadratic optimization, as shown in several studies. Recent publications have extended the concept of HRP in several directions. We compare some of these extensions to determine which variant is most useful for constructing crypto baskets. We find that a particular type of adaptive HRP strategy outperforms other extensions on a risk-adjusted basis, leading us to a deeper investigation of the changing nature of correlation structures between cryptos - both quantitatively and visually. We find that structural breaks in crypto correlations are prevalent and that the best-fitting hierarchical cluster representations change over time, which is only captured by distance matrix-based adaptive HRP approaches.

Park, J. (2020). "Clustering Approaches for Global Minimum Variance Portfolio." In: *arXiv e-Print*.

The only input to attain the portfolio weights of global minimum variance portfolio (GMVP) is the covariance matrix of returns of assets being considered for investment. Since the population covariance matrix is not known, investors use historical data to estimate it. Even though sample covariance matrix is an unbiased estimator of the population covariance matrix, it includes a great amount of estimation error especially when the number of observed data is not much bigger than number of assets. As it is difficult to estimate the covariance matrix with high dimensionality all at once, clustering stocks is proposed to come up with covariance matrix in two steps: firstly, within a cluster and secondly, between clusters. It decreases the estimation error by reducing the number of features in the data matrix. The motivation of this dissertation is that the estimation error can still remain high even after clustering, if a large amount of stocks is clustered together in a single group. This research proposes to utilize a bounded clustering method in order to limit the maximum cluster size. The result

of experiments shows that not only the gap between in-sample volatility and out-of-sample volatility decreases, but also the out-of-sample volatility gets reduced. It implies that we need a bounded clustering algorithm so that maximum clustering size can be precisely controlled to find the best portfolio performance.

Parmentier, L. (2018). "Measures of Portfolio Diversification." MA thesis. Louvain School of Management.
Diversification is one the main and most important concept in the financial world. It is often said that diversification is the only free lunch in finance. From a qualitative point of view, the concept of diversification is quite clear: a portfolio is well-diversified if shocks in the individual components do not heavily impact on the overall portfolio. Relatively simple to understand then but profoundly difficult to define. Indeed, there is no broadly accepted precise and quantitative definition of diversification. Over the years, many different measures of diversification have been developed in the literature, each with its pros and cons. In the framework of this thesis, we have chosen to analyze six of them. Because we wanted to confront the weights concentration criterion with the risk minimization criterion, we decided to select measures that are based on the entropy of the weights and others that are based on the sources of risk. Those six different measures are the Shannon's Entropy, the Diversification Delta, the Diversification Ratio, the MarginalRisk Contributions, the Portfolio Diversification Index and the Effective Number of Bets.

Peng, H., Wang, H., Hu, Y., Zhou, W., and Cai, H. (2022). "Multi-dimensional clustering through fusion of high-order similarities." In: *Pattern Recognition* 121, p. 108108.
Clustering objects with heterogeneous attributes captured from different dimensions remains challenging in integrating the multiple dimensional information. Most of the current multi-dimensional clustering models pin on direct sample-wised similarity and fail to exploit hidden mutual affinity among different sampling spaces. Thus, it is hard to capture a legible cluster structure. To tackle this issue, we propose a High-order multi-dimensional Spectral Clustering method (HSC). The proposed HSC aims to learn a high-order similarity to characterize the intrinsic relationship among different dimensional spaces instead of the ordinary similarity. It then performs a clustering task within a latent space by jointly learning the high-order similarity and ordinary similarity. Extensive experiments over synthetic and real-world data sets show that the proposed HSC outperforms benchmark multi-dimensional methods in most scenarios and is capable of revealing a reliable structure concealed across multi-dimensional spaces.

Peralta, G. and Zareei, A. (2016). "A network approach to portfolio selection." In: *Journal of Empirical Finance* 38, pp. 157–180.
Low-central stocks receive higher weights in optimal allocation. Financial and market variables are major drivers of stocks' centrality. We construct a network-based investment strategy that performs well out-of-sample. Our network-based strategy results in positive and significant Carhart's alphas. In this study, a financial market is conceived as a network where the securities are nodes and the links account for returns' correlations. We theoretically prove the negative relationship between the centrality of assets in this financial market network and their optimal weights under the Markowitz framework. Therefore, optimal portfolios overweight low-central securities to avoid the large variances that result when highly influential stocks are included in the investor's opportunity set. Next, we empirically investigate the major financial and market determinants of stock's centralities. The evidence indicates that highly central nodes tend to coincide with older, larger-cap, cheaper and financially riskier securities. Finally, we explore by means of in-sample and out-of-sample analysis the extent to which the structure of the stock market network can be employed to improve the portfolio selection process. We propose a network-based investment strategy that outperforms well-known benchmarks while presenting positive and significant Carhart alphas. The major contribution of the paper is to employ the financial market network as a useful device to improve the portfolio selection process by targeting a group of assets according to their centrality.

Pharasi, H. K., Sadhukhan, S., Majari, P., Chakraborti, A., and Seligman, T. H. (2021). "Dynamics of the market states in the space of correlation matrices with applications to financial markets." In: *arXiv e-Print*.
The concept of states of financial markets based on correlations has gained increasing attention during the last 10 years. We propose to retrace some important steps up to 2018, and then give a more detailed view of recent developments that attempt to make the use of this more practical. Finally, we try to give a glimpse to the future proposing the analysis of trajectories in correlation matrix space directly or in terms of symbolic dynamics as well as attempts to analyze the clusters that make up the states in a random matrix context.

Pimentel, B. A. and de Carvalho, A. C. (2020). "A Meta-learning approach for recommending the number of clusters for clustering algorithms." In: *Knowledge-Based Systems* 195, p. 105682.

One of the main challenges in Clustering Analysis is choosing the optimal number of clusters. A typical methodology is to evaluate a validity index over the data and to optimize it as a function of the number of clusters. However, this process can have a high computational cost. In this work, we introduce a new approach for recommending the number of clusters for a particular dataset by using Meta-learning. As the predictive performance of the meta-models induced by Meta-learning is affected by how datasets are described by meta-features, we propose a new set of meta-features able to improve the predictive performance of meta-models used for recommending the number of clusters. Experimental results show that the proposed approach provides a good recommendation of the number of clusters. Additionally, the proposed meta-feature obtains better results than meta-features for clustering tasks found in the literature.

Platanakis, E., Sutcliffe, C. M., and Ye, X. (2021). "Horses for Courses: Mean-Variance for Asset Allocation and 1/N for Stock Selection." In: *European Journal of Operational Research* 288(1), pp. 302–317.
For various organizational reasons, large investors typically split their portfolio decision into two stages - asset allocation and stock selection. We hypothesise that mean-variance models are superior to equal weighting for asset allocation, while the reverse applies for stock selection, as estimation errors are less of a problem for mean-variance models when used for asset allocation than for stock selection. We confirm this hypothesis in separate analyses of US and international equities using four different types of mean-variance model (Bayes-Stein, Black-Litterman, Bayesian diffuse prior and Markowitz), a range of parameter settings, and a simulation analysis calibrated to US data.

Poletaev, A. Y. and Spiridonova, E. M. (2020). "Hierarchical Clustering as a Dimension Reduction Technique for Markowitz Portfolio Optimization." In: *Modeling and Analysis of Information Systems* 27(1), pp. 62–71.
Optimal portfolio selection is a common and important application of an optimization problem. Practical applications of an existing optimal portfolio selection methods is often difficult due to high data dimensionality (as a consequence of the large number of securities available for investment). In this paper, a method of dimension reduction based on hierarchical clustering is proposed. Clustering is widely used in computer science, a lot of algorithms and computational methods have been developed for it. As a measure of securities proximity for hierarchical clustering Pearson pair correlation coefficient is used. Further, the proposed method's influence on the quality of the optimal solution is investigated on several examples of optimal portfolio selection according to the Markowitz Model. The influence of hierarchical clustering parameters (intercluster distance metrics and clustering threshold) on the quality of the obtained optimal solution is also investigated. The dependence between the target return of the portfolio and the possibility of reducing the dimension using the proposed method is investigated too. For each considered example in the paper graphs and tables with the main results of the proposed method - application which are the decrease of the dimension and the drop of the yield (the decrease of the quality of the optimal solution) - for a portfolio constructed using the proposed method compared to a portfolio constructed without the proposed method are given. For the experiments the Python programming language and its libraries: scipy for clustering and cvxpy for solving the optimization problem (building an optimal portfolio) are used.

Policastro, V., Righelli, D., Carissimo, A., Cutillo, L., and Feis, I. D. (2021). "ROBustness In Network (robin): an R package for Comparison and Validation of communities." In: *arXiv e-Print*.
In network analysis, many community detection algorithms have been developed, however, their implementation leaves unaddressed the question of the statistical validation of the results. Here we present robin(ROBustness In Network), an R package to assess the robustness of the community structure of a network found by one or more methods to give indications about their reliability. The procedure initially detects if the community structure found by a set of algorithms is statistically significant and then compares two selected detection algorithms on the same graph to choose the one that better fits the network of interest. We demonstrate the use of our package on the American College Football benchmark dataset.

Puerto, J., Rodriguez-Madrena, M., and Scozzari, A. (2020). "Clustering and portfolio selection problems: A unified framework." In: *Computers & Operations Research* 117, p. 104891.
Given a set of assets and an investment capital, the classical portfolio selection problem consists in determining the amount of capital to be invested in each asset in order to build the most profitable portfolio. The portfolio optimization problem is naturally modeled as a mean-risk bi-criteria optimization problem where the mean rate of return of the portfolio must be maximized whereas a given risk measure must be minimized. Several mathematical programming models and techniques have been presented in the literature in order to efficiently solve the portfolio problem. A relatively recent promising line of research is to exploit clustering information of an assets network in order to develop new portfolio optimization paradigms. In this paper we endow the assets

network with a metric based on correlation coefficients between assets' returns, and show how classical location problems on networks can be used for clustering assets. In particular, by adding a new criterion to the portfolio selection problem based on an objective function of a classical location problem, we are able to measure the effect of clustering on the selected assets with respect to the non-selected ones. Most papers dealing with clustering and portfolio selection models solve these problems in two distinct steps: cluster first and then selection. The innovative contribution of this paper is that we propose a Mixed-Integer Linear Programming formulation for dealing with this problem in a unified phase. The effectiveness of our approach is validated reporting some computational experiments on some real financial datasets.

Putra, Y. E., Saepudin, D., and Aditsania, A. (2021). "Portfolio Selection of KOMPAS-100 Stocks Index Using B-Spline Based Clustering." In: *Procedia Computer Science* 179, pp. 375–382.
Investment risk in stocks is one of the things that need to be considered by investors. Therefore investors need to develop strategies to manage portfolios. One way to manage risk in stock investments is to diversify the portfolio by selecting stocks. In this paper, stocks in a portfolio are chosen based on the similarity of the price movement data through the clustering using K-means. Since stocks in the same cluster have a higher similarity compared to shares in different clusters, the portfolio consists of stocks selected in each different cluster. Stock price movements are high dimensional data, requiring computation costs during clustering, so the dimension reduction is applied by conducting an interpolation using B-Spline. Based on the weekly dataset in 10 years (01/01/2009 - 12/31/2018), the Mean-Variance and the Equal-Weight portfolio consists of the selected stocks using this approach offer less volatility, higher Sharpe Index, and better cumulative performance.

Radovanov, B. and Marcikic, A. (2014). "Testing The Performance Of The Investment Portfolio Using Block Bootstrap Method." In: *Economic Themes* 52(2).
The aim of this paper is to create a stable model of investment portfolio optimization through a high degree of diversification and reduction of sudden changes in the allocation with monitoring of the dynamics of the impact factor. In this sense, there is bootstrap application procedure, which, without an excessive number of constraints involved in the optimization process provides solutions based on uncertain information. Thus defined, the optimization method has been patented by Michaud (1999) entitled re-sampled efficiency. Accordingly, this paper offers a comparison of the performance block bootstrap optimization models and traditional Markowitz's model inside and outside of the sample by applying the most frequently traded stocks on the BSE. The results show a better performance out of the sample and the presence of a larger number of shares forming the portfolio through bootstrap methodology. However, only through the traditional optimization process could be attained optimum according to the required limits. Such effects can be observed by comparing the limits of efficiency obtained through these optimization models. However, optimization-based methods bootstrap finds its place in reducing errors of assessment resulting from the limited sample size.

Raffinot, T. (2017). "Hierarchical Clustering-Based Asset Allocation." In: *The Journal of Portfolio Management* 44(2), pp. 89–99.
This article proposes a hierarchical clustering-based asset allocation method, which uses graph theory and machine learning techniques. Hierarchical clustering refers to the formation of a recursive clustering, suggested by the data, not defined a priori. Several hierarchical clustering methods are presented and tested. Once the assets are hierarchically clustered, the authors compute a simple and efficient capital allocation within and across clusters of assets, so that many correlated assets receive the same total allocation as a single uncorrelated one. The out-of-sample performances of hierarchical clustering-based portfolios and more traditional risk-based portfolios are evaluated across three disparate datasets, which differ in term of the number of assets and the assets' composition. To avoid data snooping, the authors assess the comparison of profit measures using the bootstrap-based model confidence set procedure. Their empirical results indicate that hierarchical clustering-based portfolios are robust and truly diversified and achieve statistically better risk-adjusted performances than commonly used portfolio optimization techniques.

Raffinot, T. (2018). "The Hierarchical Equal Risk Contribution Portfolio." In: *SSRN e-Print*.
Building upon the fundamental notion of hierarchy, the "Hierarchical Risk Parity" (HRP) and the "Hierarchical Clustering based Asset Allocation" (HCAA), the Hierarchical Equal Risk Contribution Portfolio (HERC) aims at diversifying capital allocation and risk allocation. HERC merges and enhances the machine learning approach of HCAA and the Top-Down recursive bisection of HRP. In more detail, the modified Top-Down recursive division is based on the shape of dendrogram, follows an Equal Risk Contribution allocation and is extended to downside risk measures such as conditional value at risk (CVaR) and Conditional Drawdown at Risk (CDaR). The out-of-sample performances of hierarchical clustering based portfolios are evaluated across two empirical

datasets, which differ in terms of number of assets and composition of the universe (multi-assets and individual stocks). Empirical results highlight that HERC Portfolios based on downside risk measures achieve statistically better risk-adjusted performances, especially those based on the CDaR.

Rahgoshay, M. and Salavatipour, M. R. (2021). "Hierarchical Clustering: New Bounds and Objective." In: *arXiv e-Print*.

Hierarchical Clustering has been studied and used extensively as a method for analysis of data. More recently, Dasgupta [2016] defined a precise objective function. Given a set of n data points with a weight function $w_{i,j}$ for each two items i and j denoting their similarity/dis-similarity, the goal is to build a recursive (tree like) partitioning of the data points (items) into successively smaller clusters. He defined a cost function for a tree T to be $Cost(T) = \sum_{i,j \in [n]} \big(w_{i,j} \times |T_{i,j}|\big)$ where $T_{i,j}$ is the subtree rooted at the least common ancestor of i and j and presented the first approximation algorithm for such clustering. Then Moseley and Wang [2017] considered the dual of Dasgupta's objective function for similarity-based weights and showed that both random partitioning and average linkage have approximation ratio 1/3 which has been improved in a series of works to 0.585 [Alon et al. 2020]. Later Cohen-Addad et al. [2019] considered the same objective function as Dasgupta's but for dissimilarity-based metrics, called $Rev(T)$. It is shown that both random partitioning and average linkage have ratio 2/3 which has been only slightly improved to 0.667078 [Charikar et al. SODA2020]. Our first main result is to consider $Rev(T)$ and present a more delicate algorithm and careful analysis that achieves approximation 0.71604. We also introduce a new objective function for dissimilarity-based clustering. For any tree T, let $H_{i,j}$ be the number of i and j's common ancestors. Intuitively, items that are similar are expected to remain within the same cluster as deep as possible. So, for dissimilarity-based metrics, we suggest the cost of each tree T, which we want to minimize, to be $Cost_H(T) = \sum_{i,j \in [n]} \big(w_{i,j} \times H_{i,j}\big)$. We present a 1.3977-approximation for this objective.

Rebonato, R. (2019). "A financially justifiable and practically implementable approach to coherent stress testing." In: *Quantitative Finance* 19(5), pp. 827–842.

We present an approach to stress testing that is both practically implementable and solidly rooted in well-established financial theory. We present our results in a Bayesian-net context, but the approach can be extended to different settings. We show (i) how the consistency and continuity conditions are satisfied; (ii) how the result of a scenario can be consistently cascaded from a small number of macrofinancial variables to the constituents of a granular portfolio; and (iii) how an approximate but robust estimate of the likelihood of a given scenario can be estimated. This is particularly important for regulatory and capital-adequacy applications.

Rehman, A. U. and Belhaouari, S. B. (2022). "Divide well to merge better: A novel clustering algorithm." In: *Pattern Recognition* 122, p. 108305.

In this paper, a novel non-parametric clustering algorithm which is based on the concept of divide-and-merge is proposed. The proposed algorithm is based on two primary phases, after data cleaning: (i) the Division phase and (ii) the Merging phase. In the initial phase of division, the data is divided into an optimized number of small sub-clusters utilizing all the dimensions of the data. In the second phase of merging, the small sub-clusters obtained as a result of division are merged according to an advanced statistical metric to form the actual clusters in the data. The proposed algorithm has the following merits: (i) ability to discover both convex and non-convex shaped clusters, (ii) ability to discover clusters different in densities, (iii) ability to detect and remove outliers/noise in the data (iv) easily tunable or fixed hyperparameters (v) and its usability for high dimensional data. The proposed algorithm is extensively tested on 20 benchmark datasets including both, the synthetic and the real datasets and is found better/competing to the existing state-of-the-art parametric and non-parametric clustering algorithms.

Romashchenko, A. (2021). "Clustering with Respect to the Information Distance." In: *arXiv e-Print*.

We discuss the notion of a dense cluster with respect to the information distance and prove that all such clusters have an extractable core that represents the mutual information shared by the objects in the cluster.

Roncalli, T. (2021). "Advanced Course in Asset Management." In: *SSRN e-Print*.

These presentation slides have been written for the Advanced Course in Asset Management (theory and applications) given at the University of Paris-Saclay. They contain 15 tutorial exercises and 5 main lectures:

1) Portfolio Optimization

2) Risk Budgeting

3) Smart Beta, Factor Investing and Alternative Risk Premia

4) Green and Sustainable Finance, ESG Investing and Climate Risk

5) Machine Learning in Asset Management

The Table of contents is the following:
Part 1. Portfolio Optimization 1. Theory of portfolio optimization 1.a. The Markowitz framework 1.b. Capital asset pricing model (CAPM) 1.c. Portfolio optimization in the presence of a benchmark 1.d. Black-Litterman model 2. Practice of portfolio optimization 2.a. Covariance matrix 2.b. Expected returns 2.c. Regularization of optimized portfolios 2.d. Adding constraints 3. Tutorial exercises 3.a. Variations on the efficient frontier 3.b. Beta coefficient 3.c. Black-Litterman model
Part 2. Risk Budgeting 1. The ERC portfolio 1.a. Definition 1.b. Special cases 1.c. Properties 1.d. Numerical solution 2. Extensions to risk budgeting portfolios 2.a. Definition of RB portfolios 2.b. Properties of RB portfolios 2.c. Diversification measures 2.d. Using risk factors instead of assets 3. Risk budgeting, risk premia and the risk parity strategy 3.a. Diversified funds 3.b. Risk premium 3.c. Risk parity strategies 3.d. Performance budgeting portfolios 4. Tutorial exercises 4.a. Variation on the ERC portfolio 4.b. Weight concentration of a portfolio 4.c. The optimization problem of the ERC portfolio 4.d. Risk parity funds
Part 3. Smart Beta, Factor Investing and Alternative Risk Premia 1. Risk-based indexation 1.a. Capitalization-weighted indexation 1.b. Risk-based portfolios 1.c. Comparison of the four risk-based portfolios 1.d. The case of bonds 2. Factor investing 2.a. Factor investing in equities 2.b. How many risk factors? 2.c. Construction of risk factors 2.d. Risk factors in other asset classes 3. Alternative risk premia 3.a. Definition 3.b. Carry, value, momentum and liquidity 3.c. Portfolio allocation with ARP 4. Tutorial exercises 4.a. Equally-weighted portfolio 4.b. Most diversified portfolio 4.c. Computation of risk-based portfolios 4.d. Building a carry trade exposure
Part 4. Green and Sustainable Finance, ESG Investing and Climate Risk 1. ESG investing 1.a. Introduction to sustainable finance 1.b. ESG scoring 1.c. Performance in the stock market 1.d. Performance in the corporate bond market 2. Climate risk 2.a. Introduction to climate risk 2.b. Climate risk modeling 2.c. Regulation of climate risk 2.d. Portfolio management with climate risk 3. Sustainable financing products 3.a. SRI Investment funds 3.b. Green bonds 3.c. Social bonds 3.d. Other sustainability-linked strategies 4. Impact investing 4.a. Definition 4.b. Sustainable development goals (SDG) 4.c. Voting policy, shareholder activism and engagement 4.d. The challenge of reporting 5. Tutorial exercises 5.a. Probability distribution of an ESG score 5.b. Enhanced ESG score and tracking error control
Part 5. Machine Learning in Asset Management 1. Portfolio optimization 1.a. Standard optimization algorithms 1.b. Machine learning optimization algorithms 1.c. Application to portfolio allocation 2. Pattern learning and self-automated strategies 3. Market generators 4. Tutorial exercises 4.a. Portfolio optimization with CCD and ADMM algorithms 4.b. Regularized portfolio optimization.

Rusch, T., Mair, P., and Hornik, K. (2021). "Cluster Optimized Proximity Scaling." In: *Journal of Computational and Graphical Statistics*, pp. 1–12.
Proximity scaling methods such as multidimensional scaling represent objects in a low-dimensional configuration so that fitted object distances optimally approximate object proximities. Besides finding the optimal configuration, an additional goal may be to make statements about the cluster arrangement of objects. This fails if the configuration lacks appreciable clusteredness. We present cluster optimized proximity scaling (COPS), which attempts to find a configuration that exhibits clusteredness. In COPS, a flexible parameterized scaling loss function that may emphasize differentiation information in the proximities is augmented with an index (OPTICS Cordillera) that penalizes lack of clusteredness of the configuration. We present two variants of this, one for finding a configuration directly and one for hyperparameter selection for parametric stresses. We apply both to a functional magnetic resonance imaging dataset on neural representations of mental states in a social cognition task and show that COPS improves clusteredness of the configuration, enabling visual identification of clusters of mental states. Online supplementary materials are available including an R package and a document with additional details.

Ruta, N., Sawada, N., McKeough, K., Behrisch, M., and Beyer, J. (2020). "SAX Navigator: Time Series Exploration through Hierarchical Clustering." In: *arXiv e-Print*.
Comparing many long time series is challenging to do by hand. Clustering time series enables data analysts to discover relevance between and anomalies among multiple time series. However, even after reasonable clustering, analysts have to scrutinize correlations between clusters or similarities within a cluster. We developed SAX Navigator, an interactive visualization tool, that allows users to hierarchically explore global patterns as well

as individual observations across large collections of time series data. Our visualization provides a unique way to navigate time series that involves a "vocabulary of patterns" developed by using a dimensionality reduction technique,Symbolic Aggregate approXimation(SAX). With SAX, the time series data clusters efficiently and is quicker to query at scale. We demonstrate the ability of SAX Navigator to analyze patterns in large time series data based on three case studies for an astronomy data set. We verify the usability of our system through a think-aloud study with an astronomy domain scientist.

Sakurai, Y., Yuki, Y., Katsuki, R., Yazane, T., and Ishizaki, F. (2021). "Correlation diversified passive portfolio strategy based on permutation of assets." In: *The Journal of Investment Strategies*.

In this paper we develop a passive strategy to improve index investing, which we call the correlation diversified portfolio strategy. The proposed method adjusts the weight vector of the original index based on the permutation of the assets belonging to the original index. We seek the permutation of these assets such that those assets with a strong correlation to many other assets are placed in the center of the permutation. By reducing the weights of such central assets, we can construct portfolios that are more diversified and have better risk-return characteristics than the original index. We solve this asset-permutation problem by adopting a quantum-inspired approach. Concretely, we convert this permutation problem into a quadratic unconstrained binary optimization problem and use simulated annealing on a personal computer or annealing machine to find a near-optimal solution in a reasonable time. To examine the usefulness and computational feasibility of the proposed method, we apply it to three major indexes of the United States and Japan, and we provide numerical experiments that show portfolios constructed by the proposed method can achieve a higher return with lower volatility compared with the original indexes, while their behaviors are still similar to those of the original indexes.

Samal, A., Kumar, S., Yadav, Y., and Chakraborti, A. (2021). "Network-centric indicators for fragility in global financial indices." In: *arXiv e-Print*.

Over the last two decades, financial systems have been studied and analysed from the perspective of complex networks, where the nodes and edges in the network represent the various financial components and the strengths of correlations between them. Here, we adopt a similar network-based approach to analyse the daily closing prices of 69 global financial market indices across 65 countries over a period of 2000-2014. We study the correlations among the indices by constructing threshold networks superimposed over minimum spanning trees at different time frames. We investigate the effect of critical events in financial markets (crashes and bubbles) on the interactions among the indices by performing both static and dynamic analyses of the correlations. We compare and contrast the structures of these networks during periods of crashes and bubbles, with respect to the normal periods in the market. In addition, we study the temporal evolution of traditional market indicators, various global network measures and the recently developed edge-based curvature measures. We show that network-centric measures can be extremely useful in monitoring the fragility in the global financial market indices.

Sarda-Espinosa, A. (2019a). "Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package." In: *SSRN e-Print*.

Most clustering strategies have not changed considerably since their initial definition. Most of the improvements are either related to the distance measure used to assess dissimilarity, or the function used to calculate prototypes or centroids. Time-series clustering is no exception, with the Dynamic Time Warping distance being particularly popular in that context. This distance is computationally expensive, so many related optimizations have been developed over the years. Since no single clustering algorithm can be said to perform best on all datasets, different strategies must be tested and compared, so a common infrastructure can be advantageous. In this manuscript, a general overview of time-series clustering is given, including many specifics related to Dynamic Time Warping and other recently proposed techniques. At the same time, a description of the dtwclust package for the R statistical software is provided, showcasing how it can be used to evaluate many different time-series clustering procedures.

Sarda-Espinosa, A. (2019b). "Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package." In: *The R Journal*.

Most clustering strategies have not changed considerably since their initial definition. Most of the improvements are either related to the distance measure used to assess dissimilarity, or the function used to calculate prototypes or centroids. Time-series clustering is no exception, with the Dynamic Time Warping distance being particularly popular in that context. This distance is computationally expensive, so many related optimizations have been developed over the years. Since no single clustering algorithm can be said to perform best on all datasets, different strategies must be tested and compared, so a common infrastructure can be advantageous. In this manuscript, a general overview of time-series clustering is given, including many specifics related to Dynamic

Time Warping and other recently proposed techniques. At the same time, a description of the dtwclust package for the R statistical software is provided, showcasing how it can be used to evaluate many different time-series clustering procedures.

Sass, J. and Thos, A.-K. (2022). "Risk reduction and portfolio optimization using clustering methods." In: *Econometrics and Statistics*.
Diversification is one of the main pillars of investment strategies. The prominent equal weight or one-over-N portfolio, which puts equal weight on each asset, is apart from its simplicity a strategy which is hard to outperform in realistic settings. But depending on the number of considered assets it can lead to very large portfolios. An approach to reduce the number of chosen assets based on clustering is proposed and its advantages and disadvantages are investigated. Using clustering techniques the possible assets are separated into non-overlapping clusters and the assets within a cluster are ordered by their Sharpe ratio. Then the best asset of each portfolio is chosen to be a member of the new portfolio with equal weights, the cluster portfolio. It is shown that this portfolio inherits the advantages of the equal weight portfolio and that it can even outperform it empirically. To this end different performance measures are used to compare the portfolios on simulated and real data. To explain the observations on real data, explanatory results are derived in an extreme model setting and analyzed in several simulation studies.

Sato-Ilic, M. (2021). "Cluster-scaled principal component analysis." In: *WIREs Computational Statistics*.
Cluster-scaled analysis means exploiting the cluster-based scaling to conventional data analysis to obtain more accurate results or results that we cannot obtain by using ordinary analysis. Our target data is complex and large amounts of data. For this type of data, it is well known that ordinary statistical methods do not always work well, or theoretically, we know that we cannot obtain a correct result. As a tool of this implementation, we utilize fuzzy clustering, which is well known as a robust clustering to a complex and large amount of data. That is, we use the fuzzy clustering result as a scale of data and apply the rescaled data by the cluster-scale to another target analysis. Our target analysis in this article is principal component analysis, which is a well-known dimensional reduction method. A numerical example shows a better performance of the cluster-scaled principal component analysis.

Scherer, B. (2021). "Adding alternative assets: return enhancement, diversification or hedging?" In: *Journal of Asset Management* 22, pp. 437–442.
Adding assets (so-called extensions) to an already existing portfolio is a reoccurring question in times of rapidly expanding investment opportunity sets. Examples for this "how much" question are the incorporation of liquid alternative assets in the form of hedge funds or alternative risk premia in a global balanced portfolio, the addition of global equities to a domestic equity portfolios or simply the optimal allocation of corporate credit within a government debt portfolio. While this is hardly a new question and a variety of tools have already been established, we suggest a new framework to decompose the demand for risky assets in economically meaningful components. This allows us to identify whether a particular allocation is driven by demand created from noisy return estimates or by more predictable hedging and diversification demand.

Schumann, E. (2019). "Backtesting." In: *SSRN e-Print*.
We discuss the backtesting of investment and trading strategies. We start with the challenges and pitfalls: overfitting, data preparation, and the effects of randomness. Then, we introduce and describe R software for backtesting. We demonstrate how to use the software for univariate and multivariate strategies (i.e. portfolio strategies) for two equity data sets. Specifically, we discuss the implementation and testing of momentum and portfolio optimization models. Throughout, we stress the analysis of sensitivity and robustness checks. Since such analyses require to run many backtests, we also discuss how backtests can be run in parallel.

Schwendner, P., Papenbrock, J., Jaeger, M., and Krugel, S. (2021). "Adaptive Seriational Risk Parity and Other Extensions for Heuristic Portfolio Construction Using Machine Learning and Graph Theory." In: *The Journal of Financial Data Science* 3(4), pp. 65–83.
In this article, the authors present a conceptual framework named adaptive seriational risk parity (ASRP) to extend hierarchical risk parity (HRP) as an asset allocation heuristic. The first step of HRP (quasi-diagonalization), determining the hierarchy of assets, is required for the actual allocation done in the second step (recursive bisectioning). In the original HRP scheme, this hierarchy is found using single-linkage hierarchical clustering of the correlation matrix, which is a static tree-based method. The authors compare the performance of the standard HRP with other static and adaptive tree-based methods, as well as seriation-based methods that do not rely on trees. Seriation is a broader concept allowing reordering of the rows or columns of a matrix to best express similarities between the elements. Each discussed variation leads to a different time series reflecting portfolio

performance using a 20-year backtest of a multi-asset futures universe. Unsupervised learningbased on these time-series creates a taxonomy that groups the strategies in high correspondence to the construction hierarchy of the various types of ASRP. Performance analysis of the variations shows that most of the static tree-based alternatives to HRP outperform the single-linkage clustering used in HRP on a risk-adjusted basis. Adaptive tree methods show mixed results, and most generic seriation-based approaches underperform.

Seabrook, I. E., Barucca, P., and Caccioli, F. (2021). "Evaluating structural edge importance in temporal networks." In: *EPJ Data Science* 10(1).
To monitor risk in temporal financial networks, we need to understand how individual behaviours affect the global evolution of networks. Here we define a structural importance metric – which we denote as l_e– for the edges of a network. The metric is based on perturbing the adjacency matrix and observing the resultant change in its largest eigenvalues. We then propose a model of network evolution where this metric controls the probabilities of subsequent edge changes. We show using synthetic data how the parameters of the model are related to the capability of predicting whether an edge will change from its value of l_e. We then estimate the model parameters associated with five real financial and social networks, and we study their predictability. These methods have applications in financial regulation whereby it is important to understand how individual changes to financial networks will impact their global behaviour. It also provides fundamental insights into spectral predictability in networks, and it demonstrates how spectral perturbations can be a useful tool in understanding the interplay between micro and macro features of networks.

Sekula, M., Datta, S., and Datta, S. (2017). "optCluster: An R Package for Determining the Optimal Clustering Algorithm." In: *Bioinformation* 13(03), pp. 101–103.
There exist numerous programs and packages that perform validation for a given clustering solution; however, clustering algorithms fare differently as judged by different validation measures. If more than one performance measure is used to evaluate multiple clustering partitions, an optimal result is often difficult to determine by visual inspection alone. This paper introduces optCluster, an R package that uses a single function to simultaneously compare numerous clustering partitions (created by different algorithms and/or numbers of clusters) and obtain a "best" option for a given dataset. The method of weighted rank aggregation is utilized by this package to objectively aggregate various performance measure scores, thereby taking away the guesswork that often follows a visual inspection of cluster results. The optCluster package contains biological validation measures as well as clustering algorithms developed specifically for RNA sequencing data, making it a useful tool for clustering genomic data.

Serur, J. A. and Avellaneda, M. (2021). "Hierarchical PCA and Modeling Asset Correlations." In: *SSRN e-Print*.
Modeling cross-sectional correlations between thousands of stocks, acrosscountries and industries, can be challenging. In this paper, we demonstratethe advantages of using Hierarchical Principal Component Analysis (HPCA)over the classic PCA. We also introduce a statistical clustering algorithmto identify homogeneous clusters of stocks or "synthetic sectors". We apply these methods to study cross-sectional correlations in the US, Europe, China,and Emerging Markets.

Seymour, A., Flint, E. J., and Chikurunhe, F. (2018). "Dynamic portfolio management strategies: A framework for historical analysis." In: *SSRN e-Print*.
The performance of dynamic trading and investment strategies can be difficult to predict. Although not without its problems, analysis of the historical performance of a strategy can provide valuable insight into its general risk and return properties. Furthermore, historical analysis allows one to compare variations of a strategy and examine the impact of various parameter choices and implementation rules. Dynamic strategy applications in three areas are considered, namely derivatives, asset allocation and equity factor portfolios. Firstly, the analysis of a strategy involving single-stock derivatives is examined in which call options on certain constituents of an index portfolio are sold as an alternative method of under-weighting the underlying. Secondly, the historical performance of an optimization-based asset allocation strategy is considered. The assumed aim of the strategy is to outperform a benchmark of CPI 5 via dynamic trading in a portfolio of domestic equities, bonds, property and cash, as well as international equities and bonds. Finally, the effects of portfolio construction on factor performance are studied via an historical analysis in which portfolios corresponding to a selection of fundamental factors are managed according to a range of weighting schemes, rebalance frequencies and portfolio sizes.

Shirota, Y. and Murakami, A. (2021). "Long-term Time Series Data Clustering of Stock Prices for Portfolio Selection." In: *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*. IEEE.
In this paper, clustering for stock data is conducted with two clustering methods, k-Shape and k-means with DTW distance measure and the results are compared. The data is the top 129 global electronics manufactures'

stock prices from 2018 to 2020 which included the worst Christmas in 2018 and the beginning of COVID-19 outbreak. The involved countries are US, China, Taiwan, Korea, Japan and some others. The clustering results by k-Shape indicate distinctively different effects on those countries' stock markets due to the COVID-19 turmoil. The patterns of the clusters can be visualized to identify the differences among the clusters. We found that each of eight clusters comprises of the same country companies. From that, we could guess that investors or their algorithms tend to invest in companies according to its country rather than the individual company's performance.

Silva, V. F., Silva, M. E., Ribeiro, P., and Silva, F. (2021a). "Novel Features for Time Series Analysis: A Complex Networks Approach." In: *arXiv e-Print*.

Time series data are ubiquitous in several domains as climate, economics and health care. Mining features from these time series is a crucial task with a multidisciplinary impact. Usually, these features are obtained from structural characteristics of time series, such as trend, seasonality and autocorrelation, sometimes requiring data transformations and parametric models. A recent conceptual approach relies on time series mapping to complex networks, where the network science methodologies can help characterize time series. In this paper, we consider two mapping concepts, visibility and transition probability and propose network topological measures as a new set of time series features. To evaluate the usefulness of the proposed features, we address the problem of time series clustering. More specifically, we propose a clustering method that consists in mapping the time series into visibility graphs and quantile graphs, calculating global topological metrics of the resulting networks, and using data mining techniques to form clusters. We apply this method to a data sets of synthetic and empirical time series. The results indicate that network-based features capture the information encoded in each of the time series models, resulting in high accuracy in a clustering task. Our results are promising and show that network analysis can be used to characterize different types of time series and that different mapping methods capture different characteristics of the time series.

Silva, V. F., Silva, M. E., Ribeiro, P., and Silva, F. (2021b). "Time series analysis via network science: Concepts and algorithms." In: *WIREs Data Mining and Knowledge Discovery* 11(3).

There is nowadays a constant flux of data being generated and collected in all types of real world systems. These data sets are often indexed by time, space, or both requiring appropriate approaches to analyze the data. In univariate settings, time series analysis is a mature field. However, in multivariate contexts, time series analysis still presents many limitations. In order to address these issues, the last decade has brought approaches based on network science. These methods involve transforming an initial time series data set into one or more networks, which can be analyzed in depth to provide insight into the original time series. This review provides a comprehensive overview of existing mapping methods for transforming time series into networks for a wide audience of researchers and practitioners in machine learning, data mining, and time series. Our main contribution is a structured review of existing methodologies, identifying their main characteristics, and their differences. We describe the main conceptual approaches, provide authoritative references and give insight into their advantages and limitations in a unified way and language. We first describe the case of univariate time series, which can be mapped to single layer networks, and we divide the current mappings based on the underlying concept: visibility, transition, and proximity. We then proceed with multivariate time series discussing both single layer and multiple layer approaches. Although still very recent, this research area has much potential and with this survey we intend to pave the way for future research on the topic.

Sjostrand, D. and Behnejad, N. (2020). "Exploration of Hierarchical Clustering in Long-only Risk-based Portfolio Optimization." MA thesis. Copenhagen Business School.

Modern portfolio optimization methods have introduced new ways of allocating capital and have drawn the attention of scholars, practitioners, and the general public alike. The thesis aims to add to the empirical evidence on the impact and risk-based performance of hierarchical clustering portfolios in long-only risk-based portfolio optimization. This is achieved by analyzing and investigating the Hierarchical Risk Parity, Hierarchical Equal Risk Contribution, and Nested Clustered Optimization methods, and compare these from a risk-based perspective to several traditional optimization methods. The relative risk-based performance is assessed through Monte Carlo simulations using synthetic data as well as through a walk-forward backtest applied on historical S&P 500 data. Together, the methodology provides a broad view of the general performance, but also more focused insights into potential estimation error reduction and the impact of different clustering parameters. The combined empirical results do not provide conclusive support for any general performance gains from hierarchical clustering in portfolio optimization. The initial positive effects found in earlier studies for Nested Clustered Optimization are hypothesized to stem from the highly stylized and simplified assumptions applied. The re-

sults given in this thesis suggest that these initial positive effects diminish when applied to more realistic data. Furthermore, the results for Hierarchical Risk Parity and Hierarchical Equal Risk Contribution show results in line with previous studies by Raffinot (2018). It is concluded that they are performing reasonably well but underperform in comparison to several of the traditional portfolios on most risk-based performance dimensions included. The findings do not indicate any general increase in risk-based performance, but do, however, show promise in providing more control over the weight concentration. In conclusion, the authors find that clustering indicates some promising aspects, but that these are limited given the applied hierarchical methodology, and further research is warranted to reach more conclusive answers.

Snow, D. (2020). "Machine Learning in Asset Management - Part 2: Portfolio Construction - Weight Optimization." In: *The Journal of Financial Data Science* 2 (2), pp. 17–24.
This is the second in a series of articles dealing with machine learning in asset management. This article focuses on portfolio weighting using machine learning. Following from the previous article (Snow 2020), which looked at trading strategies, this article identifies different weight optimization methods for supervised, unsupervised, and reinforcement learning frameworks. In total, seven submethods are summarized, with the code made available for further exploration.

Sobczyk, P., Wilczynski, S., Bogdan, M., Graczyk, P., Josse, J., Panloup, F., Seegers, V., and Staniak, M. (2020). "VARCLUST: clustering variables using dimensionality reduction." In: *arXiv e-Print*.
VARCLUST algorithm is proposed for clustering variables under the assumption that variables in a given cluster are linear combinations of a small number of hidden latent variables, corrupted by the random noise. The entire clustering task is viewed as the problem of selection of the statistical model, which is defined by the number of clusters, the partition of variables into these clusters and the 'cluster dimensions', i.e. the vector of dimensions of linear subspaces spanning each of the clusters. The optimal model is selected using the approximate Bayesian criterion based on the Laplace approximations and using a non-informative uniform prior on the number of clusters. To solve the problem of the search over a huge space of possible models we propose an extension of the ClustOfVar algorithm which was dedicated to subspaces of dimension only 1, and which is similar in structure to the K-centroid algorithm. We provide a complete methodology with theoretical guarantees, extensive numerical experimentations, complete data analyses and implementation. Our algorithm assigns variables to appropriate clusterse based on the consistent Bayesian Information Criterion (BIC), and estimates the dimensionality of each cluster by the PEnalized SEmi-integrated Likelihood Criterion (PESEL), whose consistency we prove. Additionally, we prove that each iteration of our algorithm leads to an increase of the Laplace approximation to the model posterior probability and provide the criterion for the estimation of the number of clusters. Numerical comparisons with other algorithms show that VARCLUST may outperform some popular machine learning tools for sparse subspace clustering. We also report the results of real data analysis including TCGA breast cancer data and meteorological data. The proposed method is implemented in the publicly available R package varclust.

Son, B. and Lee, J. (2022). "Graph-based multi-factor asset pricing model." In: *Finance Research Letters* 44 (102032).
We propose a latent multi-factor asset pricing model that estimates risk exposure based on firm characteristics and connectivity between assets. To handle connected high-dimensional characteristics, we adopted a graph convolutional network while estimating the connectivity between assets from the correlation of asset returns. Unlike recent literature involving the deep-learning-based latent factor model, we propose a forward stagewise additive factor modeling architecture that constructs latent factors sequentially to maintain the previous stage's factors. Our empirical results on individual U.S. equities show that the proposed graph factor model outperforms other benchmark models in terms of explanatory power and the Sharpe ratio of the factor tangency portfolio.

Stankovic, L., Mandic, D., Dakovic, M., Brajovic, M., Scalzo, B., and Constantinides, T. (2020a). "Data Analytics on Graphs Part I: Graphs, Graph Spectra, and Spectral Clustering." In: *Foundations and Trends in Machine Learning* 13 (1).
The area of Data Analytics on graphs promises a paradigm shift as we approach information processing of classes of data, which are typically acquired on irregular but structured domains (social networks, various ad-hoc sensor networks). Yet, despite its long history, current approaches mostly focus on the optimization of graphs themselves, rather than on directly inferring learning strategies, such as detection, estimation, statistical and probabilistic inference, clustering and separation from signals and data acquired on graphs. To fill this void, we first revisit graph topologies from a Data Analytics point of view, and establish a taxonomy of graph networks through a linear algebraic formalism of graph topology (vertices, connections, directivity). This serves as a basis for spectral analysis of graphs, whereby the eigenvalues and eigenvectors of graph Laplacian and adjacency matrices are shown to convey physical meaning related to both graph topology and higher-order graph

properties, such as cuts, walks, paths, and neighborhoods. Next, to illustrate estimation strategies performed on graph signals, spectral analysis of graphs is introduced through eigenanalysis of mathematical descriptors of graphs and in a generic way. Finally, a framework for vertex clustering and graph segmentation is established based on graph spectral representation (eigenanalysis) which illustrates the power of graphs in various data association tasks. The supporting examples demonstrate the promise of Graph Data Analytics in modeling structural and functional/semantic inferences. At the same time, Part I serves as a basis for Part II and Part III which deal with theory, methods and applications of processing Data on Graphs and Graph Topology Learning from data.

Stankovic, L., Mandic, D., Dakovic, M., Brajovic, M., Scalzo, B., Li, S., and Constantinides, A. G. (2020b). "Data Analytics on Graphs Part II: Signals on Graphs." In: *Foundations and Trends in Machine Learning* 13 (2-3).
The area of Data Analytics on graphs deals with information processing of data acquired on irregular but structured graph domains. The focus of Part I of this monograph has been on both the fundamental and higher-order graph properties, graph topologies, and spectral representations of graphs. Part I also establishes rigorous frameworks for vertex clustering and graph segmentation, and illustrates the power of graphs in various data association tasks. Part II embarks on these concepts to address the algorithmic and practical issues related to data/signal processing on graphs, with the focus on the analysis and estimation of both deterministic and random data on graphs. The fundamental ideas related to graph signals are introduced through a simple and intuitive, yet general enough case study of multisensor temperature field estimation. The concept of systems on graph is defined using graph signal shift operators, which generalize the corresponding principles from traditional learning systems. At the core of the spectral domain representation of graph signals and systems is the Graph Fourier Transform (GFT), defined based on the eigendecomposition of both the adjacency matrix and the graph Laplacian. Spectral domain representations are then used as the basis to introduce graph signal filtering concepts and address their design, including Chebyshev series polynomial approximation. Ideas related to the sampling of graph signals, and in particular the challenging topic of data dimensionality reduction through graph subsampling, are presented and further linked with compressive sensing. The principles of time-varying signals on graphs and basic definitions related to random graph signals are next reviewed. Localized graph signal analysis in the joint vertex-spectral domain is referred to as the vertex-frequency analysis, since it can be considered as an extension of classical time-frequency analysis to the graph serving as signal domain. Important aspects of the local graph Fourier transform (LGFT) are covered, together with its various forms including the graph spectral and vertex domain windows and the inversion conditions and relations. A link between the LGFT with a varying spectral window and the spectral graph wavelet transform (SGWT) is also established. Realizations of the LGFT and SGWT using polynomial (Chebyshev) approximations of the spectral functions are further considered and supported by examples. Finally, energy versions of the vertex-frequency representations are introduced, along with their relations with classical timefrequency analysis, including a vertex-frequency distribution that can satisfy the marginal properties. The material is supported by illustrative examples.

Stankovic, L., Mandic, D., Dakovic, M., Brajovic, M., Scalzo, B., Li, S., and Constantinides, A. G. (2020c). "Data Analytics on Graphs Part III: Machine Learning on Graphs, from Graph Topology to Applications." In: *Foundations and Trends in Machine Learning*.
Many modern data analytics applications on graphs operate on domains where graph topology is not known a priori, and hence its determination becomes part of the problem definition, rather than serving as prior knowledge which aids the problem solution. Part III of this monograph starts by addressing ways to learn graph topology, from the case where the physics of the problem already suggest a possible topology, through to most general cases where the graph topology is learned from the data. A particular emphasis is on graph topology definition based on the correlation and precision matrices of the observed data, combined with additional prior knowledge and structural conditions, such as the smoothness or sparsity of graph connections. For learning sparse graphs (with small number of edges), the least absolute shrinkage and selection operator, known as LASSO is employed, along with its graph specific variant, graphical LASSO. For completeness, both variants of LASSO are derived in an intuitive way, and explained. An in-depth elaboration of the graph topology learning paradigm is provided through several examples on physically well defined graphs, such as electric circuits, linear heat transfer, social and computer networks, and spring-mass systems. As many graph neural networks (GNN) and convolutional graph networks (GCN) are emerging, we have also reviewed the main trends in GNNs and GCNs, from the perspective of graph signal filtering. Tensor representation of lattice-structured graphs is next considered, and it is shown that tensors (multidimensional data arrays) are a special class of graph signals, whereby the graph

vertices reside on a high-dimensional regular lattice structure. This part of monograph concludes with two emerging applications in financial data processing and underground transportation networks modeling.

Stavroglou, S. (2020). "Finding Hidden Structures in Financial Networks." PhD thesis. University of Liverpool.
In this thesis we delve into the dynamic evolution of financial networks seeking real world meaning imprinted in their hidden structures. The hypothesis that permeates our research is that despite the stochastic behavior of single assets, when studied collectively there should be some emergent and persistent patterns which signal highly important information for scientists and policymakers alike. Notwithstanding the insights from industry-standard methods, the hidden nature of causality remains a puzzling yet critical notion for effective decision-making. Financial markets are characterized by fluctuating interdependencies that can give rise to emergent phenomena such as bubbles or crashes. Motivated by these uncertainties, we designed a novel causality framework based on symbolic dynamics that probes beneath the surface of abstract causality and unveils the nature of causal interactions. We named our framework "pattern causality". This novel algorithm allows for a distinction between positive and negative interdependencies as well as a hybrid form that we refer to as "dark causality". We benchmark this method on asset pairs and on a network of sovereign credit default swaps, where the dominant form of interaction is that of dark causality. Our results are critical to financial advisors who have a fiduciary duty to their clients and retail investors. Further contemplating upon the operational laws and concepts from complex systems, we composed a second algorithm out of the pattern causality framework with the purpose of capturing important aspects and interactions beyond stock markets. In an abstract complex network, it is an enigmatic and inspiring challenge to predict the actual interdependencies that comprise the structure of such systems, be it financial markets, ecosystems, or even the human brain. Particularly considering that the vital interdependencies underlying disparate real-world phenomena might be persistently hidden, the task of creating one algorithm to tackle them all seems daunting. Yet, our second algorithm is excellent at detecting the latent and elusive structures of complex systems. Our treatment utilizes short-term predictions from information embedded in reconstructed state space. Using a broad class of real-world applications, we are able to demonstrate our method's power to reconstruct the backbone of complex systems and simultaneously highlight their most fundamental operations. This last algorithm can serve as a tool for decision-makers and policymakers alike, and the demonstrated effectiveness establishes its potential for capturing hidden interactions in a much broader area of applications.

Suhonen, A., Lennkh, M., and Perez, F. (2017). "Quantifying Backtest Overfitting in Alternative Beta Strategies." In: *The Journal of Portfolio Management* 43 (2), pp. 90–104.
The authors investigate the biases in the backtested performance of "alternative beta"strategies using a unique sample of 215 trading strategies developed and promoted by global investment banks. Their results lend support to the cautions in the recent literature regarding backtest overfitting and lack of robustness in trading strategy performance during the "live"period (out of sample). The authors report a median 73 percent deterioration in Sharpe ratios between backtested and live performance periods for the strategies, and they establish a link between performance deterioration and strategy complexity, with the realized reduction in live versus back-tested Sharpe ratios of the most complex strategies exceeding those of the simplest ones by over 30 percentage points. The robustness of strategy exposure to risk factors varies between asset classes and strategies; it appears reasonable in equity volatility and FX carry strategies but quite weak in the equity value strategy in particular.

Swedroe, L. (2020). "The Importance of Diversification in Achieving Long-Term Goals." In: *Advisor Perspectives*.
My 2007 book, Wise Investing Made Simple: Larry Swedroe's Tales to Enrich Your Future, contained 27 tales to educate investors about important investment concepts and strategies. This article is in the spirit of those tales. The examples are hypothetical.

Taljaard, B. H. and Maré, E. (2021). "Why has the equal weight portfolio underperformed and what can we do about it?" In: *Quantitative Finance* 21(11), pp. 1855–1868.
It is widely noted that market capitalisation weighted portfolios are inefficient and underperform an equal weighted portfolio over the long-term. However, at least since 2016, an equal weighted portfolio of stocks in the S&P500 has significantly underperformed the market capitalisation weighted portfolio. In this paper, we analyse this underperformance using stochastic portfolio theory. We show that the equal weighted portfolio does appear to outperform the market capitalisation weighted portfolio over the long-term but with periods of significant short-term underperformance. In addition, we find that concentration in the market capitalisation weighted portfolio has increased in recent years and has contributed to the recent underperformance together with a significantly lower level of diversification benefits. Furthermore, we highlight an approach to improve the

performance of a portfolio by dynamically selecting a market cap or an equal weighting using a rudimentary linear regression model.

Tang, W., Xu, X., and Zhou, X. Y. (2021). "Asset Selection via Correlation Blockmodel Clustering." In: *arXiv e-Print*.
We aim to cluster financial assets in order to identify a small set of stocks to approximate the level of diversification of the whole universe of stocks. We develop a data-driven approach to clustering based on a correlation blockmodel in which assets in the same cluster have the same correlations with all other assets. We devise an algorithm to detect the clusters, with a theoretical analysis and a practical guidance. Finally, we conduct an empirical analysis to attest the performance of the algorithm.

Tayali, S. T. (2020). "A novel backtesting methodology for clustering in mean–variance portfolio optimization." In: *Knowledge-Based Systems* 209, p. 106454.
The decisions of asset selection and allocation lie at the heart of financial portfolio management. For these challenging tasks, the mathematical programming model of the mean-variance optimization problem proposes to use the concept of diversification. The novel methodology in this article is a representation of the accumulated knowledge of this model from the modern portfolio theory. It is a practical application for portfolio managers to help synthesize the available historical data and to infer rational decisions. The state-of-the-art backtesting methodology integrates the unsupervised machine learning method of clustering analysis into the mean-variance portfolio optimization model. The test results from the proposed novel methodology show that clustering with Euclidean distance measures outperform the results of the benchmark and other specified clustering methods for different datasets, backtesting periods, and temporal scales of major stock indices.

Tellaroli, P., Bazzi, M., Donato, M., Brazzale, A. R., and Draghici, S. (2016). "Cross-Clustering: A Partial Clustering Algorithm with Automatic Estimation of the Number of Clusters." In: *PLOS ONE* 11(3), e0152333+.
Four of the most common limitations of the many available clustering methods are: i) the lack of a proper strategy to deal with outliers; ii) the need for a good a priori estimate of the number of clusters to obtain reasonable results; iii) the lack of a method able to detect when partitioning of a specific data set is not appropriate; and iv) the dependence of the result on the initialization. Here we propose Cross-clustering (CC), a partial clustering algorithm that overcomes these four limitations by combining the principles of two well established hierarchical clustering algorithms: Ward's minimum variance and Complete-linkage. We validated CC by comparing it with a number of existing clustering methods, including Ward's and Complete-linkage. We show on both simulated and real datasets, that CC performs better than the other methods in terms of: the identification of the correct number of clusters, the identification of outliers, and the determination of real cluster memberships. We used CC to cluster samples in order to identify disease subtypes, and on gene profiles, in order to determine groups of genes with the same behavior. Results obtained on a non-biological dataset show that the method is general enough to be successfully used in such diverse applications. The algorithm has been implemented in the statistical language R and is freely available from the CRAN contributed packages repository.

Thiagarajan, R., Han, J., Hurd, A., Im, H., and Mallik, G. (2021). "Financial Globalization and Its Implications for Diversification of Portfolio Risk." In: *The Journal of Investing* 30(6), pp. 22–33.
Trade disputes and the impact of the COVID-19 pandemic on global supply chains have drawn much attention to the notion of "deglobalization." The common concern is that the steady trend of globalization and its many benefits may reverse. But the globalization trend is not a monolith. In this article, we show that although trade globalization has stalled since the Global Financial Crisis (GFC), financial globalization has continued to increase. We further show that financial globalization has a much more significant impact on portfolios than trade globalization. The primary mechanism of this impact, US dollar hegemony, impacts portfolios primarily through increased spillover of US monetary policy shocks. The two implications for investors are: (1) global equity markets have become increasingly correlated and are likely to stay that way, and (2) this increased correlation reduces the benefits of portfolio diversification and leads to a more concentrated exposure to US monetary policy shocks.

Thrun, M. C. (2021). "The Exploitation of Distance Distributions for Clustering." In: *International Journal of Computational Intelligence and Applications* 20(03).
Although distance measures are used in many machine learning algorithms, the literature on the context-independent selection and evaluation of distance measures is limited in the sense that prior knowledge is used. In cluster analysis, current studies evaluate the choice of distance measure after applying unsupervised methods based on error probabilities, implicitly setting the goal of reproducing predefined partitions in data. Such studies use clusters of data that are often based on the context of the data as well as the custom goal of the specific

study. Depending on the data context, different properties for distance distributions are judged to be relevant for appropriate distance selection. However, if cluster analysis is based on the task of finding similar partitions of data, then the intrapartition distances should be smaller than the interpartition distances. By systematically investigating this specification using distribution analysis through the mirrored-density (MD plot), it is shown that multimodal distance distributions are preferable in cluster analysis. As a consequence, it is advantageous to model distance distributions with Gaussian mixtures prior to the evaluation phase of unsupervised methods. Experiments are performed on several artificial datasets and natural datasets for the task of clustering.

Thrun, M. C. and Stier, Q. (2021). "Fundamental clustering algorithms suite." In: *SoftwareX* 13, p. 100642.
The article presents immediate access to over fifty fundamental clustering algorithms. Additionally, access to clustering benchmark datasets published priorly as "Fundamental Clustering Problems Suite" (FCPS) is provided. The software library is named "FCPS", available in R on CRAN and accessible within Python. The input and output of clustering algorithms are standardized to enable users a swift execution of cluster analysis. By combining mirrored-density plots (MD plots) with statistical testing, FCPS provides a tool to investigate the cluster-tendency quickly before the cluster analysis itself. Common clustering challenges can be generated with an arbitrary sample size. Additionally, FCPS sums up 26 indicators intending to estimate the number of clusters and provides an appropriate implementation of the clustering accuracy for more than two clusters.

Tola, V., Lillo, F., Gallegati, M., and Mantegna, R. N. (2008). "Cluster analysis for portfolio optimization." In: *Journal of Economic Dynamics and Control* 32(1), pp. 235–258.
We consider the problem of the statistical uncertainty of the correlation matrix in the optimization of a financial portfolio. By assuming idealized conditions of perfect forecast ability for the future return and volatility of stocks and short selling, we show that the use of clustering algorithms can improve the reliability of the portfolio in terms of the ratio between predicted and realized risk. Bootstrap analysis indicates that this improvement is obtained in a wide range of the parameters N (number of assets) and T (investment horizon). The predicted and realized risk level and the relative portfolio composition of the selected portfolio for a given value of the portfolio return are also investigated for each considered filtering method. We also show that several of the results obtained by assuming idealized conditions are still observed under the more realistic assumptions of no short selling and mean return and volatility forecasting based on historical data.

Tong, W., Liu, S., and Gao, X.-Z. (2021). "A density-peak-based clustering algorithm of automatically determining the number of clusters." In: *Neurocomputing* 458, pp. 655–666.
Clustering is a typical and important method to discover new structures and knowledge from data sets. Most existing clustering methods need to know the number of clusters in advance, which is difficult. Some algorithms claim they do not need to know the number of clusters in advance. Among these algorithms, however, some need to manually determine the cluster centers in a decision graph, which is not easy; some assume that the number of initial cluster centers given is greater than the actual number of classes, but in fact the true number of clusters is not known. In order to tackle this issue, we propose a density-peak-based clustering algorithm of automatically determining the number of clusters. First, we design a density metric by using a continuous function which can well distinguish the densities of different data points. Then, we design a pre-clustering method which can get the initial cluster centers and the corresponding clusters. Furthermore, we propose an automatic clustering method which can automatically determine the final cluster centers and the corresponding clusters. Experiments are conducted on widely used data sets, and the results show the effectiveness of the proposed method.

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.

Turner, E. (2021). "Graph Auto-Encoders for Financial Clustering." In: *arXiv e-Print*.
Deep learning has shown remarkable results on Euclidean data (e.g. audio, images, text) however this type of data is limited in the amount of relational information it can hold. In mathematics we can model more general relational data in a graph structure while retaining Euclidean data as associated node or edge features. Due to the ubiquity of graph data, and its ability to hold multiple dimensions of information, graph deep learning has become a fast emerging field. We look at applying and optimising graph deep learning on a finance graph to produce more informed clusters of companies. Having clusters produced from multiple streams of data can be highly useful in quantitative finance; not only does it allow clusters to be tailored to the specific task but

the culmination of multiple streams allows for cross source pattern recognition that would have otherwise gone unnoticed. This can provide financial institutions with an edge over competitors which is crucial in the heavily optimised world of trading. In this paper we use news co-occurrence and stock price for our data combination. We optimise our model to achieve an average testing precision of 78% and find a clear improvement in clustering capabilities when dual data sources are used; cluster purity rises from 32% for just vertex data and 42% for just edge data to 64% when both are used in comparisons to ground-truth Bloomberg clusters. The framework we provide utilises unsupervised learning which we view as key for future work due to the volume of unlabelled data in financial markets.

Underwood, W. G., Elliott, A., and Cucuringu, M. (2020). "Motif-based spectral clustering of weighted directed networks." In: *Applied Network Science* 5(1) (62).
Clustering is an essential technique for network analysis, with applications in a diverse range of fields. Although spectral clustering is a popular and effective method, it fails to consider higher-order structure and can perform poorly on directed networks. One approach is to capture and cluster higher-order structures using motif adjacency matrices. However, current formulations fail to take edge weights into account, and thus are somewhat limited when weight is a key component of the network under study.We address these shortcomings by exploring motif-based weighted spectral clustering methods. We present new and computationally useful matrix formulae for motif adjacency matrices on weighted networks, which can be used to construct efficient algorithms for any anchored or non-anchored motif on three nodes. In a very sparse regime, our proposed method can handle graphs with a million nodes and tens of millions of edges. We further use our framework to construct a motif-based approach for clustering bipartite networks.We provide comprehensive experimental results, demonstrating (i) the scalability of our approach, (ii) advantages of higher-order clustering on synthetic examples, and (iii) the effectiveness of our techniques on a variety of real world data sets; and compare against several techniques from the literature. We conclude that motif-based spectral clustering is a valuable tool for analysis of directed and bipartite weighted networks, which is also scalable and easy to implement.

Valentine, K. D., Buchanan, E. M., Scofield, J. E., and Beauchamp, M. T. (2019). "Beyond p values: utilizing multiple methods to evaluate evidence." In: *Behaviormetrika* 46(1), pp. 121–144.
Null hypothesis significance testing is cited as a threat to validity and reproducibility. While many individuals suggest that we focus on altering the p value at which we deem an effect significant, we believe this suggestion is short-sighted. Alternative procedures (i.e., Bayesian analyses and observation-oriented modeling: OOM) can be more powerful and meaningful to our discipline. However, these methodologies are less frequently utilized and are rarely discussed in combination with NHST. Herein, we discuss three methodologies (NHST, Bayesian Model comparison, and OOM), then compare the possible interpretations of three analyses (ANOVA, Bayes Factor, and an Ordinal Pattern Analysis) in various data environments using a frequentist simulation study. We found that changing significance thresholds had little effect on conclusions. Furthermore, we suggest that evaluating multiple estimates as evidence of an effect allows for more robust and nuanced interpretations of results and implies the need to redefine evidentiary value and reporting practices. Recent events in psychological science have prompted concerns within the discipline regarding research practices and ultimately, the validity and reproducibility of published reports (Etz and Vandekerckhove 2016; Lindsay 2015, Open Science Collaboration 2015; van Elk et al. 2015). One often discussed matter is over-reliance, abuse, and potential hacking of p values produced by frequentist null hypothesis significance testing (NHST), as well as misinterpretations of NHST results (Gigerenzer 2004; Ioannidis 2005; Simmons et al. 2011). We agree with these concerns and believe that many before us have voiced sound, generally accepted opinions on potential remedies, such as an increased focus on effect sizes (Cumming 2008; Lakens 2013; Maxwell et al. 2015; Nosek et al. 2012). However, other suggestions have been met with less enthusiasm, including an article by Benjamin et al. (2018) advocating that researchers should begin thinking only of p values less than .005 as "statistically significant", thus changing alpha levels to control Type I error rates. Alternatively, Pericchi and Pereira (2016) promote the use of fluctuating alpha levels as a function of sample size to assist with these errors. Trafimow et al. (2018) critiques this suggestion to broadly lower the alpha level to .005 and suggested that findings should be weighted on the basis of evidence accumulation from multiple studies. We argue that alpha should not be the sole focus of our attention, but rather, we should wonder if a p value should be utilized at all, and, if so, what that p value can tell us in relation with other indicators. While NHST and p values may have merit, researchers have a wealth of other statistical tools available to them. We believe that improvements may be made to the sciences as a whole when individuals become aware of these tools and how these methods may be used, either alone or in combination, to strengthen understanding of data and conclusions. These sentiments have been shared by the

American Statistical Association who recently held a conference focusing on going beyond NHST, expanding their previous stance on p values (Wasserstein and Lazar 2016). Therefore, the main goal of this project was to show researchers how two alternative paradigms compare to NHST in terms of their methodological design, statistical interpretations, and comparative robustness. Herein, we will discuss the following methodologies: NHST, Bayes factor comparisons, and observation-oriented modeling. To compare their methodological designs, we first provide historical backgrounds, procedural steps, and limitations for each paradigm. We then simulated data using a three timepoint repeated measures design with a Likert-type scale as the outcome variable to be able to compare the statistical interpretations and comparative robustness. By simulating possible data sets and analyzing them with each of the three paradigms, we will be able to discuss the conclusions these three methods reach given the same data and to compare how often these methodologies agree within different data environments (i.e., given varying sample sizes and effect sizes). Beyond simply comparing methodologies, we also sought to identify how changing the alpha criteria within the NHST framework may alter conclusions. Although previous work has already compared Frequentist NHST to Bayesian approaches (Goodman 1999; Rouder et al. 2012; Wetzels et al. 2011), this manuscript adds a novel contribution: observation-oriented modeling. By introducing social scientists to observation-oriented modeling (OOM), a relatively new paradigm that is readily interpretable, we will show both how useful this paradigm can be in these contexts, and how it compares to two well-known methods. We hope that by discussing these methodologies in terms of a simple statistical analysis researchers will be able to easily compare and contrast methodologies.

Valk, M. and Cybis, G. B. (2021). "U-Statistical Inference for Hierarchical Clustering." In: *Journal of Computational and Graphical Statistics* 30(1), pp. 133–143.

Clustering methods are valuable tools for the identification of patterns in high-dimensional data with applications in many scientific fields. However, quantifying uncertainty in clustering is a challenging problem, particularly when dealing with high-dimension low sample size (HDLSS) data. We develop a U-statistics based clustering approach that assesses statistical significance in clustering and is specifically tailored to HDLSS scenarios. These nonparametric methods rely on very few assumptions about the data, and thus can be applied to a wide range of datasets for which the Euclidean distance captures relevant features. Our main result is the development of a hierarchical significance clustering method. To do so, we first introduce an extension of a relevant U-statistic and develop its asymptotic theory. Additionally, as a preliminary step, we propose a binary nonnested significance clustering method and show its optimality in terms of expected values. Our approach is tested through multiple simulations and found to have more statistical power than competing alternatives in all scenarios considered. Our methods are further showcased in three applications ranging from genetics to image recognition problems. Code for these methods is available in R-package uclust. Supplementary materials for this article are available online.

Vankwikelberge, X., Kang, B., Heiter, E., and Lijffijt, J. (2021). "ExClus: Explainable Clustering on Low-dimensional Data Representations." In: *arXiv e-Print*.

Dimensionality reduction and clustering techniques are frequently used to analyze complex data sets, but their results are often not easy to interpret. We consider how to support users in interpreting apparent cluster structure on scatter plots where the axes are not directly interpretable, such as when the data is projected onto a two-dimensional space using a dimensionality-reduction method. Specifically, we propose a new method to compute an interpretable clustering automatically, where the explanation is in the original high-dimensional space and the clustering is coherent in the low-dimensional projection. It provides a tunable balance between the complexity and the amount of information provided, through the use of information theory. We study the computational complexity of this problem and introduce restrictions on the search space of solutions to arrive at an efficient, tunable, greedy optimization algorithm. This algorithm is furthermore implemented in an interactive tool called ExClus. Experiments on several data sets highlight that ExClus can provide informative and easy-to-understand patterns, and they expose where the algorithm is efficient and where there is room for improvement considering tunability and scalability.

Vázquez, I., Villar, J. R., Sedano, J., Simić, S., and Cal, E. de la (2021). "An ensemble solution for multivariate time series clustering." In: *Neurocomputing* 457, pp. 182–192.

Technologies such as Big Data and IoT have shown the need for intelligent unsupervised processing of Multivariate Time Series (MTS), MTS clustering among them. The challenges in MTS clustering includes not only the selection of the algorithm but also the MTS representation and the similarity measurement among the instances. This study proposes an ensemble of MTS clustering methods that merges different MTS representations and distance functions, aggregating them to obtain a similarity measurement. Furthermore, a proposal for prior

knowledge representation is propose to balance the aggregation of the distances. The final clustering is performed either using k-means or hierarchical clustering. The experimentation set up includes the implementation of the ensemble with either 4 or 5 different methods, including an MTS extension of k-Shape. The results show that the ensemble is biased towards the best methods, which helps the clustering practitioner in the selection of the most suitable prototypes. Moreover, the evaluation of the ensemble with the number of clusters set to the number of labels shows that metrics, such as the sensitivity and specificity, must drive the rule of the elbow; alternatively, this value represents the most interesting prior knowledge bit in MTS clustering. Further work includes the study of digital markers to compare MTS representations and distance functions and the use of external metrics to balance the aggregation of the methods.

Vigen, T. (2019). *Spurious Correlations*. URL: https://www.tylervigen.com/spurious-correlations.
Military intelligence analyst and Harvard Law student Tyler Vigen illustrates the golden rule that "correlation does not equal causation" through hilarious graphs. Is there a correlation between Nic Cage films and swimming pool accidents? What about beef consumption and people getting struck by lightning? Absolutely not. But that hasn't stopped millions of people from going to tylervigen.com and asking, "Wait, what?" Vigen has designed software that scours enormous data sets to find unlikely statistical correlations. He began pulling the funniest ones for his website and has since gained millions of views, hundreds of thousands of likes, and tons of media coverage. Subversive and clever, Spurious Correlations is geek humor at its finest, nailing our obsession with data and conspiracy theory.

Vincent, K., Hsu, Y.-C., and Lin, H.-W. (2018). "Analyzing the Performance of Multifactor Investment Strategies under a Multiple Testing Framework." In: *The Journal of Portfolio Management* 44(4), pp. 113–126.
Evaluating portfolios based on numerous combinations of factors using the individual backtesting method could suffer from serious data mining bias and lead to spurious significant findings. Accordingly, the authors employ a multiple hypothesis testing method to examine the multifactor portfolio performance. Their empirical results show that even after they adjust for the multiple comparisons bias, stock-picking strategies with certain combined firm characteristics could generate significantly better liquidity risk-adjusted returns. In addition, the outperforming multifactor strategies that the authors report are robust to alternative definitions of factors. However, they observe that the number of significantly profitable multifactor portfolios has decreased substantially in the era of increased liquidity and trading activity in the U.S. stock market.

Vojtko, R. and Cisár, D. (2021). "An Analysis of Volatility Clustering of Equity Factor Strategies." In: *SSRN e-Print*.
Volatility clustering is a well-known effect in equity markets. In simple meaning, volatility clustering refers to a tendency of large changes in asset prices to follow large changes and small changes in asset prices to follow small changes. We tested two hypotheses: (1) firstly, if there is a volatility clustering present in equity factor strategies, (2) secondly, whether past factor volatility predicts future factor performance. We were able to confirm the first hypothesis. However, a factor allocation trading strategy based on volatility predictability doesn't perform well.

Vovk, V. and Wang, R. (2020). "True and false discoveries with e-values." In: *arXiv e-Print*.
The topic of this paper is multiple hypothesis testing based on e-values, which are Bayes factors stripped of their Bayesian content. Using e-values instead of p-values, which are standard in this area, leads to simple and efficient procedures that control the number of false discoveries under arbitrary dependence of the base e-values. We prove an optimality result for our main procedure and demonstrate advantages of our methods over standard methods using simulated and real-world datasets.

Vovk, V. and Wang, R. (2021). "E-values: Calibration, combination, and applications." In: *Annals of Statistics* 49(3), pp. 1736–1753.
Multiple testing of a single hypothesis and testing multiple hypotheses are usually done in terms of p-values. In this paper we replace p-values with their natural competitor, e-values, which are closely related to betting, Bayes factors, and likelihood ratios. We demonstrate that e-values are often mathematically more tractable; in particular, in multiple testing of a single hypothesis, e-values can be merged simply by averaging them. This allows us to develop efficient procedures using e-values for testing multiple hypotheses.

Vyrost, T., Lyocsa, S., and Baumohl, E. (2019). "Network-based asset allocation strategies." In: *The North American Journal of Economics and Finance* 47, pp. 516–536.
In this study, we construct financial networks in which nodes are represented by assets and where edges are based on long-run correlations. We construct four networks (complete graph, a minimum spanning tree, a planar maximally filtered graph, and a threshold significance graph) and use three centrality measures (betweenness, eigenvalue centrality, and the expected force). To improve risk return characteristics of well-known return maximization and risk minimization benchmark portfolios, we propose simple adjustments to portfolio selection

strategies that utilize centralization measures from financial networks. From a sample of 45 assets (stock market indices, bond and money market instruments, commodities, and foreign exchange rates) and from data for 1999 to 2015, we show that irrespective of the network and centrality employed, the proposed network-based asset allocation strategies improve key portfolio return characteristics in an out of sample framework, most notably, risk and left tail risk adjusted returns. Resolving portfolio model selection uncertainties further improves risk return characteristics. Improvements made to portfolio strategies based on risk minimization are also robust to transaction costs.

Wadhwa, R. R. and Scott, J. G. (2020). "Exploring complex networks with the ICON R package." In: *arXiv e-Print*.
We introduce ICON, an R package that contains 1075 complex network datasets in a standard edgelist format. All provided datasets have associated citations and have been indexed by the Colorado Index of Complex Networks - also referred to as ICON. In addition to supplying a large and diverse corpus of useful real-world networks, ICON also implements an S3 generic to work with the network and ggnetwork R packages for network analysis and visualization, respectively. Sample code in this report also demonstrates how ICON can be used in conjunction with the igraph package. Currently, the Comprehensive R Archive Network hosts ICON v0.4.0. We hope that ICON will serve as a standard corpus for complex network research and prevent redundant work that would be otherwise necessary by individual research groups. The open source code for ICON and for this reproducible report can be found at `https://github.com/rrrlw/ICON`.

Wang, M., Abrams, Z. B., Kornblau, S. M., and Coombes, K. R. (2018). "Thresher: determining the number of clusters while removing outliers." In: *BMC Bioinformatics* 19(1), p. 9.
BACKGROUND: Cluster analysis is the most common unsupervised method for finding hidden groups in data. Clustering presents two main challenges: (1) finding the optimal number of clusters, and (2) removing "outliers" among the objects being clustered. Few clustering algorithms currently deal directly with the outlier problem. Furthermore, existing methods for identifying the number of clusters still have some drawbacks. Thus, there is a need for a better algorithm to tackle both challenges. RESULTS: We present a new approach, implemented in an R package called Thresher, to cluster objects in general datasets. Thresher combines ideas from principal component analysis, outlier filtering, and von Mises-Fisher mixture models in order to select the optimal number of clusters. We performed a large Monte Carlo simulation study to compare Thresher with other methods for detecting outliers and determining the number of clusters. We found that Thresher had good sensitivity and specificity for detecting and removing outliers. We also found that Thresher is the best method for estimating the optimal number of clusters when the number of objects being clustered is smaller than the number of variables used for clustering. Finally, we applied Thresher and eleven other methods to 25 sets of breast cancer data downloaded from the Gene Expression Omnibus; only Thresher consistently estimated the number of clusters to lie in the range of 4-7 that is consistent with the literature. CONCLUSIONS: Thresher is effective at automatically detecting and removing outliers. By thus cleaning the data, it produces better estimates of the optimal number of clusters when there are more variables than objects. When we applied Thresher to a variety of breast cancer datasets, it produced estimates that were both self-consistent and consistent with the literature. We expect Thresher to be useful for studying a wide variety of biological datasets.

Wang, S., Sun, Y., and Bao, Z. (2020). "On the Efficiency of K-Means Clustering: Evaluation, Optimization, and Algorithm Selection." In: *arXiv e-Print*.
This paper presents a thorough evaluation of the existing methods that accelerate Lloyd's algorithm for fast k-means clustering. To do so, we analyze the pruning mechanisms of existing methods, and summarize their common pipeline into a unified evaluation framework UniK. UniK embraces a class of well-known methods and enables a fine-grained performance breakdown. Within UniK, we thoroughly evaluate the pros and cons of existing methods using multiple performance metrics on a number of datasets. Furthermore, we derive an optimized algorithm over UniK, which effectively hybridizes multiple existing methods for more aggressive pruning. To take this further, we investigate whether the most efficient method for a given clustering task can be automatically selected by machine learning, to benefit practitioners and researchers.

Wang, Y. and Tsay, R. S. (2019). "Clustering Multiple Time Series with Structural Breaks." In: *Journal of Time Series Analysis* 40(2), pp. 182–202.
Time series clustering pattern could change over time. In this article we develop a new Bayesian approach to handle clustering analysis of multiple time series with structural breaks. The number of breaks is treated as a random variable, with group membership and groupspecific parameters allowed to change on these breaks. Groupspecific parameters in each regime can be integrated analytically, so we only have a small number of parameters to be handled by posterior simulation. We further discuss prediction, identification, clustering, and

detection of the number of groups. Using Monte Carlo simulation, we document the performance of the proposed approach in statistical efficiency, forecasting, and detection of the structural breaks. An application on quarterly industrial production growth rates of 21 countries links regimes to historical business cycles. Prediction performance and economic gains are illustrated based on the proposed method.

Wang, Y. and Aste, T. (2022). "Dynamic Portfolio Optimization with Inverse Covariance Clustering." In: *arXiv e-Print*.

Market conditions change continuously. However, in portfolio's investment strategies, it is hard to account for this intrinsic non-stationarity. In this paper, we propose to address this issue by using the Inverse Covariance Clustering (ICC) method to identify inherent market states and then integrate such states into a dynamic portfolio optimization process. Extensive experiments across three different markets, NASDAQ, FTSE and HS300, over a period of ten years, demonstrate the advantages of our proposed algorithm, termed Inverse Covariance Clustering-Portfolio Optimization (ICC-PO). The core of the ICC-PO methodology concerns the identification and clustering of market states from the analytics of past data and the forecasting of the future market state. It is therefore agnostic to the specific portfolio optimization method of choice. By applying the same portfolio optimization technique on a ICC temporal cluster, instead of the whole train period, we show that one can generate portfolios with substantially higher Sharpe Ratios, which are statistically more robust and resilient with great reductions in maximum loss in extreme situations. This is shown to be consistent across markets, periods, optimization methods and selection of portfolio assets.

Weylandt, M., Nagorski, J., and Allen, G. I. (2019). "Dynamic Visualization and Fast Computation for Convex Clustering via Algorithmic Regularization." In: *arXiv e-Print*.

Convex clustering is a promising new approach to the classical problem of clustering, combining strong performance in empirical studies with rigorous theoretical foundations. Despite these advantages, convex clustering has not been widely adopted, due to its computationally intensive nature and its lack of compelling visualizations. To address these impediments, we introduce Algorithmic Regularization, an innovative technique for obtaining high-quality estimates of regularization paths using an iterative one-step approximation scheme. We justify our approach with a novel theoretical result, guaranteeing global convergence of the approximate path to the exact solution under easily-checked non-data-dependent assumptions. The application of algorithmic regularization to convex clustering yields the Convex Clustering via Algorithmic Regularization Paths (CARP) algorithm for computing the clustering solution path. On example data sets from genomics and text analysis, CARP delivers over a 100-fold speed-up over existing methods, while attaining a finer approximation grid than standard methods. Furthermore, CARP enables improved visualization of clustering solutions: the fine solution grid returned by CARP can be used to construct a convex clustering-based dendrogram, as well as forming the basis of a dynamic path-wise visualization based on modern web technologies. Our methods are implemented in the open-source R package clustRviz, available at this https URL.

Wiecki, T., Campbell, A., Lent, J., and Stauth, J. (2016). "All That Glitters Is Not Gold: Comparing Backtest and Out-of-Sample Performance on a Large Cohort of Trading Algorithms." In: *The Journal of Investing* 25(3), pp. 69–80.

When automated trading strategies are developed and evaluated using backtests on historical pricing data, there exists a tendency to overfit to the past. Using a unique dataset of 888 algorithmic trading strategies developed and backtested on the Quantopian platform, with at least six months of out-of-sample performance, this article studies the prevalence and impact of backtest overfitting. Specifically, the authors find that commonly reported backtest evaluation metrics, such as the Sharpe ratio, offer little value in predicting out-of-sample performance ($R2 < 0.025$). In contrast, higher-order moments, such as volatility and maximum drawdown, as well as portfolio construction features (e.g., hedging), show significant predictive value of relevance to quantitative finance practitioners. Moreover, in line with prior theoretical considerations, the authors find empirical evidence of overfitting-the more backtesting a quant has done for a strategy, the larger the discrepancy between backtest and out-of-sample performance. Finally, they show that by training nonlinear, machine-learning classifiers on a variety of features that describe backtest behavior, out-of-sample performance can be predicted with much greater accuracy ($R2 = 0.17$) on hold-out data than when using linear, univariate features. A portfolio constructed by using predictions on hold-out data performed significantly better out-of-sample than one constructed from algorithms with the highest backtest Sharpe ratios.

Wu, C., Peng, Q., Lee, J., Leibnitz, K., and Xia, Y. (2021). "Effective hierarchical clustering based on structural similarities in nearest neighbor graphs." In: *Knowledge-Based Systems* 228, p. 107295.

Hierarchical clustering allows better performance in grouping heterogeneous and non-spherical datasets than the center-based clustering, at the expense of increased time complexity. Meanwhile, the bottom-up approach of hierarchical clustering methods often tend to be sensitive or vulnerable to datasets containing obscure cluster boundaries. This paper presents an effective method for hierarchical clustering, called HCNN, which utilizes two types of structural similarities in nearest neighbor graph of a dataset to group similar data into clusters. In particular, the first metric is used to identify those pairs of data with maximal similarity in their nearest neighborhoods that can serve as local centers of every initial cluster. This can contribute to observing the boundaries and detecting clusters, hubs and outliers, thereby alleviating remarkably the influence of obscure boundaries between clusters. The initial clusters will be merged recursively in accordance with the second similarity metric measured in terms of the connectivity between clusters in the nearest neighbor graph. In this case, rather than combining two clusters with highest similarity during each iteration, we consider the maximum similarity as a transitive and closure relation between clusters, i.e., an equivalence relation, which enables more effective and efficient merging of clusters via application of advanced data structures. Experiments based on synthetic and real datasets demonstrate that the proposed HCNN algorithms can possibly outperform the state-of-the-art clustering methods evaluated in terms of the clustering accuracy, normalized mutual information and adjust rand index. Moreover, experiments on unlabeled real datasets show that our method enables effective identification of the quantity of clusters based on Davies-Bouldin index and average silhouette coefficient.

Wu, X., Wu, J., Zou, J., and Zhang, Q. (2020). "Analyses and applications of optimization methods for complex network reconstruction." In: *Knowledge-Based Systems* 193, p. 105406.

Inferring the topology of a network from observable dynamics is a key topic in the research of complex network. With the observation error considered, the topology inferring is formulated as a connectivity reconstruction problem that can be solved through optimization estimation. It is found that the different optimization methods should be selected to deal with the different degrees of noise, different scales of observable time series and such other situations when it comes to the problem of connectivity reconstruction, which has not been analyzed and discussed before yet. In this paper, four regression methods, namely least squares, ridge, lasso and elastic net, are used to solve the problem of network reconstruction in different situations. In particular, a further analysis is made of the effects of each regression method on the network reconstruction problem in detail. Through simulation of a variety of artificial and real networks, as it has turned out, the four regression methods are effective in respect to network reconstruction when certain conditions are respectively satisfied. Based on the experimental results, it is possible to reach some interesting conclusions that can guide our readers to know the internal mechanisms for network reconstruction and choose the appropriate regression method in accordance with the actual situation and existing knowledge.

Yang, L., Zhao, L., and Wang, C. (2019). "Portfolio optimization based on empirical mode decomposition." In: *Physica A: Statistical Mechanics and its Applications* 531, p. 121813.

The investigation about the cross-correlation among financial assets has drawn broad attention recently. Due to the nonlinear and non-stationary identities of the financial time series, e.g., stock return time series, the cross-correlation for different level of fluctuations are quite important for both academia and financial practitioners. Here we use the empirical mode decomposition (EMD) method to analyze the cross-correlation structure among different level of fluctuations for financial assets. The correlation-based networks are then employed to determine the clustering property of stock market. We then propose several portfolio optimization strategies based on the EMD correlation-based networks. Using the topological information of the networks, we can construct some portfolios with high return and low risk. Under two portfolio evaluation frameworks, we prove that these portfolios have consistently good performance.

Yang, Y., Zhao, L., Chen, L., Wang, C., and Han, J. (2021). "Portfolio optimization with idiosyncratic and systemic risks for financial networks." In: *arXiv e-Print*.

In this study, we propose a new multi-objective portfolio optimization with idiosyncratic and systemic risks for financial networks. The two risks are measured by the idiosyncratic variance and the network clustering coefficient derived from the asset correlation networks, respectively. We construct three types of financial networks in which nodes indicate assets and edges are based on three correlation measures. Starting from the multi-objective model, we formulate and solve the asset allocation problem. We find that the optimal portfolios obtained through the multi-objective with networked approach have a significant over-performance in terms of return measures in an out-of-sample framework. This is further supported by the less drawdown during the periods of the stock market fluctuating downward. According to analyzing different datasets, we also show that improvements made to portfolio strategies are robust.

Yelibi, L. and Gebbie, T. (2021). "Agglomerative Likelihood Clustering." In: *arXiv e-Print*.
We consider the problem of fast time-series data clustering. Building on previous work modeling the correlation-based Hamiltonian of spin variables we present an updated fast non-expensive Agglomerative Likelihood Clustering algorithm (ALC). The method replaces the optimized genetic algorithm based approach (f-SPC) with an agglomerative recursive merging framework inspired by previous work in Econophysics and Community Detection. The method is tested on noisy synthetic correlated time-series data-sets with built-in cluster structure to demonstrate that the algorithm produces meaningful non-trivial results. We apply it to time-series data-sets as large as 20,000 assets and we argue that ALC can reduce compute time costs and resource usage cost for large scale clustering for time-series applications while being serialized, and hence has no obvious parallelization requirement. The algorithm can be an effective choice for state-detection for online learning in a fast non-linear data environment because the algorithm requires no prior information about the number of clusters.

Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M., and Blair, R. H. (2018). "Bootstrapping estimates of stability for clusters, observations and model selection." In: *Computational statistics* 34(1), pp. 349–372.
Clustering is a challenging problem in unsupervised learning. In lieu of a gold standard, stability has become a valuable surrogate to performance and robustness. In this work, we propose a non-parametric bootstrapping approach to estimating the stability of a clustering method, which also captures stability of the individual clusters and observations. This flexible framework enables different types of comparisons between clusterings and can be used in connection with two possible bootstrap approaches for stability. The first approach, scheme 1, can be used to assess confidence (stability) around clustering from the original dataset based on bootstrap replications. A second approach, scheme 2, searches over the bootstrap clusterings for an optimally stable partitioning of the data. The two schemes accommodate different model assumptions that can be motivated by an investigator trust (or lack thereof) in the original data and additional computational considerations. We propose a hierarchical visualization extrapolated from the stability profiles that give insights into the separation of groups, and projected visualizations for the inspection of the stability of individual operations. Our approaches show good performance in simulation and on real data. These approaches can be implemented using the R package bootcluster that is available on the Comprehensive R Archive Network (CRAN).

Yu, L. (2021). "Comparing Classical Portfolio Optimization and Robust Portfolio Optimization on Black Swan Events." MA thesis. University of Waterloo.
Black swan events, such as natural catastrophes and manmade market crashes, historically have a drastic negative influence on investments; and there is a discrepancy on losses caused by these two types of disasters. In general, there is a recovery and it is of interest to understand what type of investment strategies lead to better performance for investors. In this thesis we study classical portfolio optimization, robust portfolio optimization and some historical black swan events. We compare two main strategies: mean variance optimization vs robust portfolio optimization on two types of black swan events: natural vs anthropogenic. The comparison illustrates that robust portfolio optimization is much more conservative, and has a shorter recovery time than classical portfolio optimization. Moreover, the losses in the stock investment resulted from a natural disaster are very minor compared to the losses resulted from an anthropogenic market crash.

Yuan, M. and Zhou, G. (2022). "Why Naive 1/N Diversification Is Not So Naive, and How to Beat It?" In: *SSRN e-Print*.
In this paper, we study portfolio choice problem under estimation risk and show why the 1/N rule is very difficult to beat in applications and studies. First, as long as the dimensionality is high relative to sample size, we show that the usual estimated investment strategies are biased even asymptotically. Second, we show that the 1/N rule is optimal in a one-factor model with diversifiable risks as dimensionality increases, irrespectively of the sample size, making investment theory-based rules inadequate as they suffer from estimation errors. Third, we provide strategies that can outperform the 1/N under suitable conditions.

Zaimovic, A., Omanovic, A., and Arnaut-Berilo, A. (2021). "How Many Stocks Are Sufficient for Equity Portfolio Diversification? A Review of the Literature." In: *Journal of Risk and Financial Management* 14(11), p. 551.
Using extensive and comprehensive databases to select a subset of research papers, we aim to critically analyze previous empirical studies to identify certain patterns in determining the optimal number of stocks in well-diversified portfolios in different markets, and to compare how the optimal number of stocks has changed over different periods and how it has been affected by market turmoil such as the Global Financial Crisis (GFC) and the current COVID-19 pandemic. The main methods used are bibliometric analysis and systematic literature review. Evaluating the number of assets which lead to optimal diversification is not an easy task as it is impacted

by a huge number of different factors: the way systematic risk is measured, the investment universe (size, asset classes and features of the asset classes), the investor's characteristics, the change over time of the asset features, the model adopted to measure diversification (i.e., equally weighted versus optimal allocation), the frequency of the data that is being used, together with the time horizon, conditions in the market that the study refers to, etc. Our paper provides additional support for the fact that (1) a generalized optimal number of stocks that constitute a well-diversified portfolio does not exist for whichever market, period or investor. Recent studies further suggest that (2) the size of a well-diversified portfolio is larger today than in the past, (3) this number is lower in emerging markets compared to developed financial markets, (4) the higher the stock correlations with the market, the lower the number of stocks required for a well-diversified portfolio for individual investors, and (5) machine learning methods could potentially improve the investment decision process. Our results could be helpful to private and institutional investors in constructing and managing their portfolios and provide a framework for future research.

Zambelli, A. (2021). "Ensemble Method for Cluster Number Determination and Algorithm Selection in Unsupervised Learning." In: *arXiv e-Print*.

Unsupervised learning, and more specifically clustering, suffers from the need for expertise in the field to be of use. Researchers must make careful and informed decisions on which algorithm to use with which set of hyperparameters for a given dataset. Additionally, researchers may need to determine the number of clusters in the dataset, which is unfortunately itself an input to most clustering algorithms. All of this before embarking on their actual subject matter work. After quantifying the impact of algorithm and hyperparameter selection, we propose an ensemble clustering framework which can be leveraged with minimal input. It can be used to determine both the number of clusters in the dataset and a suitable choice of algorithm to use for a given dataset. A code library is included in the Conclusion for ease of integration.

Zhan, N., Sun, Y., Jakhar, A., and Liu, H. (2021). "Graphical Models for Financial Time Series and Portfolio Selection." In: *arXiv e-Print*.

We examine a variety of graphical models to construct optimal portfolios. Graphical models such as PCA-KMeans, autoencoders, dynamic clustering, and structural learning can capture the time varying patterns in the covariance matrix and allow the creation of an optimal and robust portfolio. We compared the resulting portfolios from the different models with baseline methods. In many cases our graphical strategies generated steadily increasing returns with low risk and outgrew the S&P 500 index. This work suggests that graphical models can effectively learn the temporal dependencies in time series data and are proved useful in asset management.

Zhang, C., Li, Y., Chen, X., Jin, Y., Tang, P., and Li, J. (2020a). "DoubleEnsemble: A New Ensemble Method Based on Sample Reweighting and Feature Selection for Financial Data Analysis." In: *IEEE International Conference on Data Mining (ICDM)*. IEEE.

Modern machine learning models (such as deep neural networks and boosting decision tree models) have become increasingly popular in financial market prediction, due to their superior capacity to extract complex non-linear patterns. However, since financial datasets have very low signal-to-noise ratio and are non-stationary, complex models are often very prone to overfitting and suffer from instability issues. Moreover, as various machine learning and data mining tools become more widely used in quantitative trading, many trading firms have been producing an increasing number of features (aka factors). Therefore, how to automatically select effective features becomes an imminent problem. To address these issues, we propose DoubleEnsemble, an ensemble framework leveraging learning trajectory based sample reweighting and shuffling based feature selection. Specifically, we identify the key samples based on the training dynamics on each sample and elicit key features based on the ablation impact of each feature via shuffling. Our model is applicable to a wide range of base models, capable of extracting complex patterns, while mitigating the overfitting and instability issues for financial market prediction. We conduct extensive experiments, including price prediction for cryptocurrencies and stock trading, using both DNN and gradient boosting decision tree as base models. Our experiment results demonstrate that DoubleEnsemble achieves a superior performance compared with several baseline methods.

Zhang, F., Guo, R., and Cao, H. (2020b). "Information Coefficient as a Performance Measure of Stock Selection Models." In: *arXiv e-Print*.

Information coefficient (IC) is a widely used metric for measuring investment managers' skills in selecting stocks. However, its adequacy and effectiveness for evaluating stock selection models has not been clearly understood, as IC from a realistic stock selection model can hardly be materially different from zero and is often accompanies with high volatility. In this paper, we investigate the behavior of IC as a performance measure of stick selection models. Through simulation and simple statistical modeling, we examine the IC behavior both statically and

dynamically. The examination helps us propose two practical procedures that one may use for IC-based ongoing performance monitoring of stock selection models.

Zhang, M. (2021). "Weighted Clustering Ensemble: A Review." In: *arXiv e-Print*.
Clustering ensemble, or consensus clustering, has emerged as a powerful tool for improving both the robustness and the stability of results from individual clustering methods. Weighted clustering ensemble arises naturally from clustering ensemble. One of the arguments for weighted clustering ensemble is that elements (clusterings or clusters) in a clustering ensemble are of different quality, or that objects or features are of varying significance. However, it is not possible to directly apply the weighting mechanisms from classification (supervised) domain to clustering (unsupervised) domain, also because clustering is inherently an ill-posed problem. This paper provides an overview of weighted clustering ensemble by discussing different types of weights, major approaches to determining weight values, and applications of weighted clustering ensemble to complex data. The unifying framework presented in this paper will help clustering practitioners select the most appropriate weighting mechanisms for their own problems.

Zhang, Z., Zohren, S., and Roberts, S. (2020c). "Deep Learning for Portfolio Optimization." In: *The Journal of Financial Data Science* 22(4), pp. 8–20.
In this article, the authors adopt deep learning models to directly optimize the portfolio Sharpe ratio. The framework they present circumvents the requirements for forecasting expected returns and allows them to directly optimize portfolio weights by updating model parameters. Instead of selecting individual assets, they trade exchange-traded funds of market indexes to form a portfolio. Indexes of different asset classes show robust correlations, and trading them substantially reduces the spectrum of available assets from which to choose. The authors compare their method with a wide range of algorithms, with results showing that the model obtains the best performance over the testing period of 2011 to the end of April 2020, including the financial instabilities of the first quarter of 2020. A sensitivity analysis is included to clarify the relevance of input features, and the authors further study the performance of their approach under different cost rates and different risk levels via volatility scaling.

Zhao, L., Wang, C., Wang, G.-J., Stanley, H. E., and Chen, L. (2021a). "Community detection and portfolio optimization." In: *arXiv e-Print*.
Community detection methods can be used to explore the structure of complex systems. The well-known modular configurations in complex financial systems indicate the existence of community structures. Here we analyze the community properties of correlation-based networks in worldwide stock markets and use community information to construct portfolios. Portfolios constructed using community detection methods perform well. Our results can be used as new portfolio optimization and risk management tools.

Zhao, L., Wang, G.-J., Wang, M., Bao, W., Li, W., and Stanley, H. E. (2018). "Stock market as temporal network." In: *Physica A: Statistical Mechanics and its Applications* 506, pp. 1104–1112.
Financial networks have become extremely useful in characterizing the structure of complex financial systems. Meanwhile, the time evolution property of the stock markets can be described by temporal networks. We utilize the temporal network framework to characterize the time-evolving correlation-based networks of stock markets. The market instability can be detected by the evolution of the topology structure of the financial networks. We employ the temporal centrality as a portfolio selection tool. Those portfolios, which are composed of peripheral stocks with low temporal centrality scores, have consistently better performance under different portfolio optimization schemes, suggesting that the temporal centrality measure can be used as new portfolio optimization and risk management tools. Our results reveal the importance of the temporal attributes of the stock markets, which should be taken serious consideration in real life applications.

Zhao, Z., Xu, F., Du, D., and Meihua, W. (2021b). "Robust portfolio rebalancing with cardinality and diversification constraints." In: *Quantitative Finance* 21(10), pp. 1707–1721.
In this paper, we develop a robust conditional value at risk (CVaR) optimal portfolio rebalancing model under various financial constraints to construct sparse and diversified rebalancing portfolios. Our model includes transaction costs and double cardinality constraints in order to capture the trade-off between the limit of investment scale and the diversified industry coverage requirement. We first derive a closed-form solution for the robust CVaR portfolio rebalancing model with only transaction costs. This allows us to conduct an industry risk analysis for sparse portfolio rebalancing in the absence of diversification constraints. Then, we attempt to remedy the hidden industry risk by establishing a new robust portfolio rebalancing model with both sparse and diversified constraints. This is followed by the development of a distributed-version of the Alternating Direction Method of Multipliers (ADMM) algorithm, where each subproblem admits a closed-form solution. Finally,

we conduct empirical tests to compare our proposed strategy with the standard sparse rebalancing and no-rebalancing strategies. The computational results demonstrate that our rebalancing approach produces sparse and diversified portfolios with better industry coverage. Additionally, to measure out-of-sample performance, two superiority indices are created based on worst-case CVaR and annualized return, respectively. Our ADMM strategy outperforms the sparse rebalancing and no-rebalancing strategies in terms of these indices.

Zheng, Q., Zhu, J., Ma, Y., Li, Z., and Tian, Z. (2021). "Multi-view subspace clustering networks with local and global graph information." In: *Neurocomputing* 449, pp. 15–23.
This study investigates the problem of multi-view subspace clustering, the goal of which is to explore the underlying grouping structure of data collected from different fields or measurements. Since data do not always comply with the linear subspace models in many real-world applications, most existing multi-view subspace clustering methods based on the shallow linear subspace models may fail in practice. Furthermore, the underlying graph information of multi-view data is usually ignored in most existing multi-view subspace clustering methods. To address the aforementioned limitations, we proposed the novel multi-view subspace clustering networks with local and global graph information, termed MSCNLG, in this paper. Specifically, autoencoder networks are employed on multiple views to achieve latent smooth representations that are suitable for the linear assumption. Simultaneously, by integrating fused multi-view graph information into self-expressive layers, the proposed MSCNLG obtains the common shared multi-view subspace representation, which can be used to get clustering results by employing the standard spectral clustering algorithm. As an end-to-end trainable framework, the proposed method fully investigates the valuable information of multiple views. Comprehensive experiments on six benchmark datasets validate the effectiveness and superiority of the proposed MSCNLG.

Zhong, C., Hu, L., Yue, X., Luo, T., Fu, Q., and Xu, H. (2019). "Ensemble clustering based on evidence extracted from the co-association matrix." In: *Pattern Recognition* 92, pp. 93–106.
The evidence accumulation model is an approach for collecting the information of base partitions in a clustering ensemble method, and can be viewed as a kernel transformation from the original data space to a co-association matrix. However, cluster structure information may be partially lost in this transformation; hence, some methods proposed in the literature try to find the lost information and return it to the ensemble process. In this paper, an interesting phenomenon is introduced: remove some evidences from the co-association matrix, which can result in more accurate clustering results. The intuitive explanation for this is that some evidences in the original co-association matrix could be noise, with negative effects on the final clustering. However, it is difficult to detect those evidences practically, let alone remove them from the matrix. To remedy this problem, we remove multiple level evidences having low occurrence frequencies, because negative evidences do not normally occur regularly in the base partitions. Subsequently, we use normalized cut to achieve multiple clustering results. To discriminate the optimal ensemble result, an internal validity index, which uses only the co-association matrix, is specially designed for the clustering ensemble. The experimental results on 16 datasets demonstrate that the proposed scheme outperforms some state-of-the-art clustering ensemble approaches.

Zhong, G. and Pun, C.-M. (2020). "Subspace clustering by simultaneously feature selection and similarity learning." In: *Knowledge-Based Systems* 193, p. 105512.
Learning a reliable affinity matrix is the key to achieving good performance for graph-based clustering methods. However, most of the current work usually directly constructs the affinity matrix from the raw data. It may seriously affect the clustering performance since the original data usually contain noises, even redundant features. On the other hand, although integrating manifold regularization into the framework of clustering algorithms can improve clustering results, some entries of the pre-computed affinity matrix on the original data may not reflect the true similarities between data points. To address the above issues, we propose a novel subspace clustering method to simultaneously learn the similarities between data points and conduct feature selection in a unified optimization framework. Specifically, we learn a high-quality graph under the guidance of a low-dimensional space of the original data such that the obtained affinity matrix can reflect the true similarities between data points as much as possible. A new algorithm based on augmented Lagrangian multiplier is designed to find the optimal solution to the problem effectively. Extensive experiments are conducted on benchmark datasets to demonstrate that our proposed method performs better against the state-of-the-art clustering methods.

Zhou, P., Chen, J., Fan, M., Du, L., Shen, Y.-D., and Li, X. (2020). "Unsupervised feature selection for balanced clustering." In: *Knowledge-Based Systems* 193, p. 105417.
In many real-world applications of data mining, such as energy load balance of wireless sensor networks, given data points with balanced distribution, i.e., each class contains approximately the same number of instances, we often need to obtain a clustering result to reflect such balance. In many data, especially the high-dimensional

data, such balanced structure is not obvious in the original feature space, due to the noisy and redundant features. Therefore we need to apply feature selection methods to pick several informative features to reveal such balanced structure of data. Feature selection is a fundamental problem in machine learning tasks and has attracted considerable attentions in recent years. However, conventional feature selection methods often focus on how to select the most discriminative features, whereas ignoring the balance property of the data. To tackle this problem, we propose a novel unsupervised feature selection method for balanced clustering which can reveal the intrinsic balanced structure of data. In our method, a balanced regularization term is introduced to select the features which can help to produce balanced clusters. Then, we provide an Alternating Direction Method of Multipliers (ADMM) to optimize the introduced objective function. At last, the experiments are conducted on six benchmark data sets, including Yale and 20NG data sets and so on, by comparing with other state-of-the-art unsupervised feature selection methods published in the literature. The experimental results show that our method not only has better clustering performance but also leads to more balanced clustering structure.