Estimation of latent factors from high-dimensional financial time series based on unsupervised learning.

Begušić, Stjepan

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Stjepan Begušić

ESTIMATION OF LATENT FACTORS FROM HIGH-DIMENSIONAL FINANCIAL TIME SERIES BASED ON UNSUPERVISED LEARNING

DOCTORAL THESIS

Zagreb, 2020.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Stjepan Begušić

ESTIMATION OF LATENT FACTORS FROM HIGH-DIMENSIONAL FINANCIAL TIME SERIES BASED ON UNSUPERVISED LEARNING

DOCTORAL THESIS

Supervisor: Associate Professor Zvonko Kostanjčar, PhD

Zagreb, 2020.

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Stjepan Begušić

PROCJENA LATENTNIH FAKTORA IZ VISOKODIMENZIONALNIH FINANCIJSKIH VREMENSKIH NIZOVA PRIMJENOM NENADZIRANOGA UČENJA

DOKTORSKI RAD

Mentor: Izv. prof. dr. sc. Zvonko Kostanjčar

Zagreb, 2020.

The dissertation was made at the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Information Processing.

Supervisor: Associate Professor Zvonko Kostanjčar, PhD

The dissertation has: 105 pages

Dissertation number: _____

About the Supervisor

Zvonko Kostanjčar received his Dipl.-Ing. degree from the University of Zagreb, Faculty of Electrical Engineering and Computing in Zagreb, Croatia, in 2002, he received his Dipl.-Ing. degree from the University of Zagreb Faculty of Science, Financial mathematics in 2008 and he received his Ph. D. degree from University of Zagreb in 2010.

He is an Associate Professor at the University of Zagreb, Faculty of Electrical Engineering and Computing, and the head and founder of the Laboratory for Financial and Risk Analytics. He received the Roberto Giannini Teaching Award from the Faculty of Electrical Engineering and Computing, and has served as the President of the IEEE Signal Processing Chapter, Croatia Section. He has participated in and led multiple international research projects funded by the Croatian Science Foundation, EU funds, and partners from the industry. He is the author of multiple scientific papers in high-ranking journals and conferences on the topics of statistical and machine learning methods for finance.

O mentoru

Zvonko Kostanjčar diplomirao je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva 2002. godine, zatim je diplomirano na Sveučilištu u Zagrebu Prirodoslovno matematičkom fakultetu smjer Financijska matematika 2008. godine, te je doktorirao na Sveučilištu u Zagrebu 2010. godine.

Izvanredni je profesor na Sveulištu u Zagrebu, Fakultetu elektrotehnike i računarstva, te je voditelj i osnivač Laboratorija za analitiku rizika i financija. Primio je nagradu Roberto Giannini za izvrsnost u nastavi i rad sa studentima od Fakulteta elektrotehnike i računarstva, te je bio predsjednik IEEE Odjela za obradu signala, Hrvatska sekcija. Vodio je i sudjelovao u više međunarodnih istraživačkih projekata financiranima od strane Hrvatske zaklade za znanost, EU fondova i partnera iz industrije. Autor je više znanstvenih radova u visokorangiranim časopisima i konferencijama na teme statističkih metoda i metoda strojnog učenja za financije.

Acknowledgement

Since I started descending through this rabbit hole of doctoral research, I was unsure at times if I would ever emerge from it – however, I was always certain that it would only be possible by the good grace of those around me. Now, as I stand at the other end of the tunnel, still pondering on whether I deserve to be here, I can only express my gratitude to the people who shared this journey with me.

First, to my supervisor Zvonko Kostanjčar, thank you for shaping me as a researcher and professional while never forgetting to be kind and warm-hearted – you've taught me so much more than just the work we do.

A special thanks goes to all the professors and colleagues who supported me throughout the past few years in different ways, especially prof. Bojana Dalbelo Bašić, prof. Branko Jeren, prof. Sven Lončarić, and prof. Boris Podobnik. I am also deeply grateful to a whole army of other doctoral and postdoctoral researchers whom I've had the privilege to work and be friends with, especially the D158 (U+1F4A9) crew: Vanessa Keranović, Josip Marić, Andro Merćep, and Robert Vaser.

I also wish to express my gratitude to my friends at the PK gang – your support and belief in me cannot be overstated, nor can my deepest appreciation for having you in my life.

To my family: parents Julija and Dinko, brothers Tomislav and Domagoj, and grandmothers Mirjana and Ljuba – for better or worse, you shaped me into who I am today and I would not be here without your love and support, for which I am forever indebted.

Finally, and most importantly, to my wife Jasmina: As a quantitative guy, I cannot resist the need to quantify your support in all the cups of coffee you made each morning, the all-night working sessions you put up with, each moment of sanity you provided...but I remain at a loss of words (and numbers) when it comes to the feeling of love, devotion, and security you provided which I so often needed in the past few years. For this and so much more, I am eternally yours (or at least until you divorce me after I apply for a postdoc).

Throughout the journey of doctoral research, it is hard not to associate yourself and your worth with your work. All the above mentioned, and many more unmentioned, all helped me remember I was much more than just my work, but also that my work was much more than I often perceived it to be. In their own way, they all contributed to this balance which proved to be crucial for me to stand where I am today. One can only be so lucky to have such people in life, thank you all.

Abstract

Unsupervised learning methods have been increasingly used for detecting latent factors in high-dimensional time series, with many applications, especially in financial risk modelling. Most latent factor models assume that the factors are pervasive and affect all of the time series. However, some factors may affect only certain assets in financial markets, due to their clustering within countries, asset classes, or sector classifications. In this thesis, high-dimensional financial time series with pervasive and cluster-specific latent factors are considered. For the assumed latent factor model, an iterative method for clustering and latent factor estimation is proposed. A model selection algorithm is also developed, based on the spectral properties of asset correlation matrices and asset graphs. Based on the estimated latent factor structures, a covariance matrix estimator is also proposed, decomposing the security return covariance into the pervasive latent factor component, cluster-specific latent factor component, and a sparse idiosyncratic risk component. The covariance matrix estimates are used in a portfolio optimization scenario, focusing on risk-based portfolios. Moreover, a new portfolio optimization method based on the risk contributions of the identified latent factors and security clusters is also developed. A simulation study with known data generating processes demonstrates that the proposed latent factor estimation and clustering method outperforms other clustering methods and provides estimates with a high degree of accuracy. Moreover, the model selection procedure is also shown to provide stable and accurate estimates for the number of clusters and latent factors. In addition, risk-based portfolios using the estimated latent factor structures are tested on datasets of asset returns from global financial markets using a backtesting approach. The results demonstrate that the clustering approach and estimated latent factors yield relevant information, improve risk modelling and reduce volatility in the out-of-sample portfolio returns.

Keywords: Latent factor models, High-dimensional data analysis, Financial risk modelling, Portfolio optimization.

Procjena latentnih faktora iz visokodimenzionalnih financijskih vremenskih nizova koristeći nenadzirano učenje

S napretkom financijske tehnologije i globalizacijom financijskih tržišta, broj financijskih instrumenata i vrijednosnica dostupnih investitorima diljem svijeta je veći nego ikad. Investitori su danas izloženi tisućama vrijednosnih papira iz različitih tržišta, država i klasa imovine oko kojih grade svoje investicijske odluke. Pri modeliranju financijskog rizika najčešće korišteni tradicionalni modeli pretpostavljaju da na cijene velikog broja financijskih imovina utječe manji broj latentnih faktora. Upravo ta pretpostavka se kroz povijest pokazala ključnom za mnoge rezultate u vrednovanju imovine i optimizaciji portfelja koji čine okosnicu modernih financija. Metode za identifikaciju zajedničkih faktora rizika u financijskim tržištima su tema iscrpnih istraživanja, pogotovo u zadnjem desetljeću nakon globalne financijske krize. Jedan od značajnih pristupa u financijskim istraživanjima su tzv. fundamentalni faktorski modela u kojima su faktori predstavljeni sintetičkim portfeljima povezanima s fundamentalnim pokazateljima vrijednosnica, kao što je knjigovodstvena vrijednost tvrtke. Dok nedavna istraživanja na širokom skupu međunarodnih tržišta i dalje pronalaze snažne statističke dokaze o mogućnostima ovog pristupa, postojanje i ponašanje nekih specifičnih faktora i danas uzrokuju rasprave među znanstvenicima i profesionalcima iz financijske industrije. No, u mnogim slučajevima kad se razmatraju skupovi imovina različitih klasa, kao što su obveznice, robe, ili različiti investicijski fondovi ili nova tržišta kao što su nerazvijena tržišta i tržišta u nastajanju (npr. razna mala tržišta slabije razvijenih država, tržišta različitih kompleksnijih proizvoda poput ETF-ova (eng. *exchange-traded fund*), ili tržište kriptovaluta), procjena fundamentalnih faktorskih modela nije uvijek moguća. S druge strane, latentni faktorski modeli koriste metode multivarijatne statistike kako bi procijenili faktorske strukture iz vremenskih nizova povrata vrijednosnica, što omogućuje procjenu u bilo kojem skupu imovina. Upravo su latentni faktorski modeli u nedavnoj prošlosti sve više u fokusu istraživača iz različitih područja, od ekonomije i financija do računarstva i znanosti o podatcima, te su sve češće korišteni u upravljanju imovinom i donošenju odluka u financijskoj industriji.

U ovoj disertaciji razmatraju se latentni faktorski modeli i nenadzirane metode strojnog učenja za procjenu latentnih faktora iz visokodimenzionalnih financijskih vremenskih nizova. Budući da u financijskim podatcima nema oznaka o pravim vrijednostima varijabli ili klasa, nenadzirane metode strojnog učenja se koriste kako bi se procijenio latentni prostor koji objašnjava značajan dio osmotrene dinamike, najčešće koristeći povijesne vremenske nizove povrata različitih financijskih vrijednosnica. Konkretno, ova disertacija se fokusira na latentni faktorski modele sa širokim faktorima koji utječu na sve imovine i faktorima specifičnima za pojedine grupe imovina (npr. pojedine vrste vrijednosnica ili vrijednosnice iz specifičnih država i tržišta). Upravo takve latentne faktorske strukture se mogu koristiti kao adekvatan model za visokodi-

menzionalne skupove vrijednosnica iz različitih tržišta, država ili klasa imovine, koje će zbog toga stvarati određene grupe unutar kojih su izložene specifičnim faktorima, ali istovremeno biti izložene širokim zajedničkim faktorima kao što su globalni makroekonomski šokovi. Budući da se tržišni uvjeti i strukture povezanosti među imovinama mogu naglo promijeniti, relativno kratki vremenski nizovi se koriste kako bi se procijenile faktorske structure, s pretpostavkom da će procijenjeni model vrijediti i u budućnosti. U takvim situacijama je potrebno razviti algoritme za procjenu latentnih faktora koji dobro rade u visokodimenzionalnim podatcima, pogotovo kad je broj vremenskih nizova veći od njihove duljine.

Razmatrani model s latentnim širokim faktorima i faktorima specifičnima za grupe vrijednosnica potrebno je procijeniti iz stvarnih podataka u kojima nisu poznate oznake grupa, kao ni realizacije faktora. U disertaciji je predložena iterativna metoda za procjenu latentnih faktora s nepoznatom grupnom strukturom iz visokodimenzionalnih financijskih vremenskih nizova. Predloženi algoritam provodi grupiranje vremenskih nizova i procjenjuje latentne faktore tako da se procijenjeni latentni faktori koriste u pridjeljivanju pojedinih vremenskih nizova onim grupama čiji latentni faktori najbolje objašnjavaju osmotrenu dinamiku. No, budući da će one grupe koje imaju više latentnih faktora u pravilu moći objasniti više varijabilnosti u podatcima od grupa s manje faktora, grupe s više faktora će kroz proceduru privlačiti više pojedinih vremenskih nizova iako im one nužno ne bi trebale pripadati. Kako bi se izbjegla pristranost u veličini grupa s obzirom na broj latentnih faktora, predložena metoda sadrži dvije faze. Prvo se procjenjuju latentni faktori i grupe pojedinih vremenskih nizova koristeći konstantan broj latentnih faktora u svakoj grupi. Potom se procijenjene grupne pripadnosti koriste za procjenu broja faktora u svakoj grupi, kao i za konačnu procjenu latentnih širokih faktora i faktora specifičnih za grupe. Također je razvijena i procedura za procjenu broja latentnih širokih faktora i grupa u podatcima, zasnovana na spektralnim svojstvima korelacijske matrice vremenskih nizova i grafova sastavljenih koristeći sličnost vremenskih nizova. Iz sortiranih svojstvenih vrijednosti korelacijske matrice podataka definiraju se omjeri susjednih svojstvenih vrijednosti, te se kao kandidate za broj širokih faktora biraju oni koji odgovaraju najvišim omjerima svojstvenih vrijednosti. Analogno ovoj proceduru, za svaki kandidat za broj širokih faktora procjenjuje se graf vrijednosnica iz čije se Laplaceove matrice računaju omjeri svojstvenih vrijednosti i definiraju kandidati za broj grupa. Za svaki par kandidata za broj širokih faktora i grupa se procjenjuju latentni faktori (broj faktor specifičnih za grupe dolazi iz same procedure za procjenu), te se od svih kandidata konačno biraju oni brojevi širokih faktora i grupa koji daju takav model koji minimizira određen informacijski kriterij. Predloženi informacijski kriterij kažnjava grešku modela, ali i dodatne latentne faktore i grupe u podatcima, budući da oni povećavaju konačni broj parametara modela. Predložene metode stoga daju procjenu broja latentnih faktora i grupa u podatcima, kao i same procjene latentnih faktorskih struktura i grupnih pripadnosti.

S obzirom na procijenjene latentne faktorske strukture, u ovom radu se također razmatra i

procjena kovarijance podataka. Naime, kovarijanca podataka u predloženom faktorskom modelu se može rastaviti na komponentu niskog ranga koju objašnjavaju latentni široki faktori, rijetku komponentu niskog ranga koju objašnjavaju faktori specifični za grupe, te kovarijancu vlastitih rizika pojedinih vrijednosnica. Prve dvije komponente mogu se izračunati direktno iz procjena latentnih faktora, a kovarijanca vlastitih rizika, budući da dopušta određene rijetke elemente izvan dijagonale (dakle pojedine vlastite komponente mogu biti korelirane), procjenjuje se metodom adaptivnog praga. Konačna procjena kovarijance podataka, zasnovana na procjenama latentnih širokih faktora i faktora specifičnih za grupe, dana je kao suma ove tri komponente.

Jedna od važnijih primjena latentnih faktorskih modela je upravo u modeliranju rizika za upravljanje imovinom i optimizaciju portfelja. U ovoj se disertaciji stoga kao primjena razvijenih metoda razmatra optimizacija portfelja velikih skupova vrijednosnica. Povijesno jedan od najvažnijih rezultata u financijama, moderna teorija portfelja predstavlja elegantan matematički okvir za optimizaciju portfelja u kojem je cilj maksimizacija očekivanja povrata portfelja uz minimalni rizik, mjeren varijancom povrata. No, pokazano je da dobiveni portfelji mogu značajno lošija svojstva imati izvan uzorka procjene ako su greške u procjenama očekivanih povrata i kovarijanci velike. S obzirom na to da je poznato kako je teško ili gotovo nemoguće predvidjeti buduće očekivane povrate, u zadnjem desetljeću je fokus prebačen na tzv. portfelje zasnovane na riziku, koji se oslanjaju samo na procjenu kovarijance. U toj klasi portfelja, u ovom radu se razmatraju portfelji minimalne varijance (gdje je optimalan portfelj upravo onaj koji ima najmanju varijancu, odnosno rizik) i portfelji maksimalne diverzifikacije (koji maksimiziraju tzv. diversifikacijski omjer). U obje ove optimizacijske metode se mogu uključiti različite procjene kovarijance - od empirijskog procjenitelja do procjenitelja zasnovanih na latentnim faktorima. Tako je također moguće ocijeniti koji procjenitelj daje pouzdanije procjene rizika i kovarijance povrata vrijednosnica. Osim ove dvije metode optimizacije portfelja, također je predložena i nova metoda optimizacije portfelja zasnovana na rizicima procijenjenih grupa i latentnih faktora u podatcima. Predložena metoda radi u dva koraka: u prvom koraku se unutar svake grupe formira portfelj maksimalne diverzifikacije koji uključuje samo vrijednosnice unutar pojedine grupe; potom se formira portfelj koji sadrži sve identificirane grupne portfelje, također koristeći metodu maksimalne diverzifikacije. Tako se fokus optimizacije prebacuje na identificirane grupe, te odnose vrijednosnica unutar njih i između njih, umjesto na cijelu matricu kovarijance.

Budući da tržišni povijesni podatci nemaju oznake o tome koje grupe postoje, niti kojim grupama su pojedine imovine propadale, razvijen je i simulacijski okvir u kojem su poznate grupne oznake i moguće je ocijeniti kvalitetu grupiranja i procjene latentnih faktora. Rezultati na simulacijskim podatcima ukazuju na to da, čak i u slučajevima kad su vremenski nizovi vrlo visoko-dimenzionalni i njihove distribucije imaju jako teške repove, predložene metode vrlo uspješno procjenjuju latentne faktore s grupnom strukturom, te su preciznije od ostalih metoda

za procjenu latentnih faktora i grupiranje vremenskih nizova. Također, predloženi algoritam uspješno procjenjuje grupe vremenskih nizova čak i kad postoje velike razlike u broju faktora specifičnih za pojedine grupe. Rezultati ukazuju i na to da predložena metoda profitira od visoke dimenzionalnosti podataka, što znači da se s povećanjem broja vremenskih nizova zapravo smanjuje greška procjene zajedničkih komponenti niskog ranga u podatcima. Općenito, predložena metoda se pokazala robusnijom i preciznijom za procjenu grupa i latentnih faktora od drugih razmatranih metoda za procjenu latentnih faktora i metoda grupiranja.

Predložene metode su također ispitane i na povijesnim tržišnim podatcima. U tu svrhu korištena su dva skupa podataka: (i) skup od 982 NASDAQ indeksa iz različitih tržišta cijelog svijeta, (ii) skup od 1480 dionica različitih tvrtki sadržanih u globalnim MSCI indeksima. Oba skupa podataka sadrže povijesne cijene od 2005. do 2020. godine, te se u oba slučaja razmatraju tjedni povrati (budući da se tim vrijednosnicama trguje na burzama u različitim vremenskim zonama, dnevni podatci ne odražavaju ispravne korelacijske strukture). Kako bi se ispitala predložena metoda za procjenu latentnih faktora i grupa vrijednosnica iz tržišnih podataka, korišten je pristup u kojem se modeli procjenjuju na vremenskim prozorima povijesnih podataka, te se potom koriste podatci na sljedećim vremenskim prozorima kao određena "budućnost" na kojoj se kvaliteta procijenjenih modela ispituje. Budući da je latentne faktorske modele moguće prikazati kao modele koji rekonstruiraju originalne podatke koristeći latentne faktorske reprezentacije nižeg ranga od originalne dimenzionalnosti prostora uzorka, moguće je dobiti rekonstrukciju budućih realizacija vremenskih nizova povrata (na kojima model nije procijenjen ili treniran). Na taj način je ispitana greška rekonstrukcije modela unutar i izvan uzorka procjene, te je uspoređena s dobro poznatim i najčešće korištenim procjeniteljem za latentne faktore u visokodimenzionalnim vremenskim nizovima. Također je mjereno i pogoršanje greške rekonstrukcije izvan uzorka procjene u odnosu na grešku unutar uzorka procjene modela. Rezultati pokazuju da u oba korištena skupa podataka predloženi model ima veću grešku rekonstrukcije unutar uzorka procjene, što je i očekivano budući da model sadrži faktore specifične za grupe, koji unutar uzorka procjene ne mogu objašnjavati jednake količine varijabilnosti kao i modeli koji sadrže isključivo široke faktore. No, greška rekonstrukcije izvan uzorka procjene je manja za sve razmatrane duljine vremenskih prozora i oba skupa podataka, a samim time je i pogoršanje rekonstrukcijske greške puno manje za predloženi model. Ti rezultati upućuju na to da su grupe i latentne faktorske strukture procijenjene u podatcima pouzdanije i robusnije izvan uzorka procjene od latentnih faktorskih struktura koje koriste isključivo široke faktore. Također su ispitane i procjene matrice kovarijance koje daje predloženi model, korištenjem portfelja zasnovanih na riziku (portfelji minimalne varijance i portfelji maksimalne diverzifikacije). Za oba razmatrana portfelja, u oba skupa podataka, te za više razmatranih duljina vremenskih prozora, rezultati upućuju na to da procijenjene matrice kovarijance smanjuju rizik u optimalnim portfeljima, u odnosu na druge razmatrane procjenitelje. Osim toga, ispitana je i predložena metoda

optimizacije portfelja zasnovana na doprinosu riziku pojedinih grupa vrijednosnica. Ti portfelji u rezultatima pokazuju najbolja svojstva u smislu Sharpeovih omjera, zadržavajući pritom relativno niske razine rizika i malene obrtaje koji garantiraju niske transakcijske troškove. Ovi rezultati potvrđuju i simulacijske rezultate te upućuju na to da predloženi model daje pouzdane procjene latentnih faktorskih struktura, te da predložena metoda za optimizaciju portfelja rezultira portfeljima koji ostvaruju visoke povrate uz niske razine rizika, te su bolji od drugih razmatranih metoda u smislu omjera povrata i rizika. Osim toga, pokazano je da je korištenjem modela s latentnim širokim faktorima i faktorima specifičnima za grupe moguće ostvariti pouzdanije i točnije procjene latentnih faktorskih struktura u visokodimenzionalnim financijskim vremenskim nizovima. One u konačnici smanjuju rizik optimalnih portfelja i općenito mogu poboljšati upravljanje rizicima i optimizaciju portfeljima u velikim skupovima financijskih vrijednosnica.

Ključne riječi: Modeli s latentnim faktorima, analiza visokodimenzionalnih podataka, modeliranje financijskog rizika, optimizacija portfelja.

Contents

Chapter 1

Introduction

1.1 Motivation

With the rise of data driven decision making in risk management, statistical and machine learning methods are becoming increasingly important as their ability to uncover meaningful information and perform well out-of-sample is put to the test in real world scenarios. This field has recently attracted a fair amount of interdisciplinary research, bringing together mathematical, physical, econometric and computer science approaches [1, 2, 3]. These methods are of critical importance in financial risk modelling, where the dynamics of asset return time series are driven by underlying risk factors [4]. To estimate the effects that these underlying factors have on observed asset returns, traditional modelling approaches use observable macroeconomic time series (such as GDP growth, interest rates, or market returns) as model inputs [5], while others focus on finding proxies for unobservable factors (known as size, value, or momentum) using economic firm-level data [6, 7]. However, this information is not always available for every security (i.e. derivatives or certain ETFs or indices), meaning that these standard approaches may not be universally applicable [8]. Moreover, recent empirical results have been challenging some of these models, giving advantage to more agnostic statistical approaches [9].

Today, with the advances in financial technology and the globalization of financial markets, the number of investable securities and their diversity in terms of asset classes and country of origin is larger than ever. Throughout the past decades, these developments motivated the increased focus on statistical and unsupervised learning techniques for uncovering latent risk factors in asset return data [10, 11]. However, even though the number of assets continues to grow, the observable time period used to estimate these models must remain short. This is due to the fact that financial markets are known to exhibit sudden changes in dynamics and stationarity can not be assumed over long time periods – asset return volatilities and correlations change over time, especially in the presence of financial bubbles and crashes [12, 13, 14]. Most commonly, inferences are made and models are trained on historical data, with the assumption that the

knowledge found in the data (conclusions or trained model parameters) hold in the future, as depicted in Figure 1.1. Therefore, these unsupervised learning methods must be able to perform well on high-dimensional datasets when the number of time series N is commensurate or even larger than their length T [15].

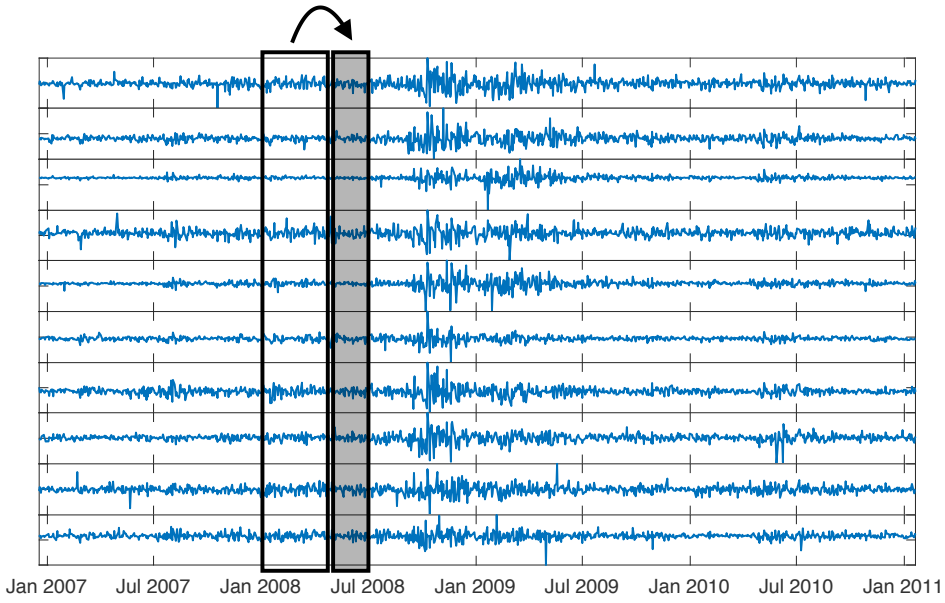


Figure 1.1: A setting with several financial return time series around the financial crisis of 2007-2009. Due to the dynamic properties of the time series, stationarity can only be assumed over short time periods, and thus relatively short time windows (transparent rectangle) may be used to make inferences about the near future (shaded rectangle).

The out-of-sample performance of these estimates is crucial for many portfolio optimization or risk management applications [1, 16]. Machine learning (and especially unsupervised learning) approaches have been increasingly applied to modelling financial risk [17, 18], and, as a natural expansion, more complex and nonlinear models have been studies, under the hypothesis that they can detect certain intricate relationships in financial data [19]. However, since these complex approaches (such as deep learning techniques) generally come with a large number of parameters, they require large amounts of data for training and thus perform poorly in high-dimensional settings. Instead, more restricted and parsimonious methods are required [20]. In this search for tractable and plausible latent factor estimation methods, it is crucial to take advantage of the structural specifics and statistical stylized facts of financial markets [21].

1.2 Latent factors and clusters in financial data

Dimensionality reduction techniques are commonly applied to obtain lower-dimensional representations of high-dimensional data, such that these representations maintain some key properties of the original data [22, 23]. This is a crucial step in coping with the so-called *curse of*

dimensionality which manifests itself through computational issues, such as sparse samples in high dimensions [24] or the rank deficiency of sample covariance estimates and the difficulties in estimating sample distributions [15]. Feature selection algorithms primarily focus on finding a function $z = g(x)$ transforming the original high-dimensional data x (which may have irrelevant or redundant information) to a lower-dimensional set of variables z which aggregate the relevant information for a certain modelling task [25]. The function g is found by optimization of certain properties, which may be assisted by the class labels or target variables, depending on the modelling task. When the class labels or target variables are not available, unsupervised feature selection techniques focus on finding features which best preserve the clusters in the data [26], remove redundancy [27, 28] or optimize certain spectral properties of the underlying data graphs [29, 30] – either in the original data space or new subspaces [31]. Generally, unsupervised feature selection approaches have been found to yield relevant results in many machine learning tasks, including sequence analysis in bioinformatics [32], text classification [33], and other applications [34].

As opposed to feature selection, the latent factor model approach is focused primarily on finding a function $x = h(f)$ which explains the high-dimensional observed data x by a lower-dimensional set of factors f. The task of estimating factor models in high-dimensional data may be reduced to a regression task when these factors are known and observed – such cases may be common in biometric, psychometric or economic applications, where factor models are used to investigate the driving factors underlying the dynamics of some phenomena or processes * [36]. However, these factors can often be unknown and unobserved, meaning that they must be estimated as latent variables from the data [37], requiring an unsupervised learning approach. The primary task is still to estimate the function $x = h(f)$, but now the factors need to be estimated from the data $f = g(x)$. Evidently, autoencoder-type approaches can be used to estimate the encoder ($f = g(x)$) and decoder ($x = h(f)$) parts of the model, offering a large range of architectures and the ability to model non-linear relationships [38, 39]. However, in the presence of high-dimensional data with the number of samples being small in comparison to the number of features/variables, nonlinear models often fail to generalize due to the large number of parameters – this turns the attention of recent research to high-dimensional latent factor estimation based on robust and regularized statistical methods [20, 40].

In this thesis, high-dimensional financial time series of asset returns are considered, with the goal of modelling the asset return time series by associating the assets with a lower-dimensional set of underlying factors. Since risk in finance is most commonly proxied by the variability of asset returns, the goal is to explain the variability of asset return time series by their exposure to

*In financial asset pricing models, these factors can sometimes be known but are not observable – for instance, factor asset pricing models identify factors such as *size*, *value*, or *momentum*, and resort to finding proxies for their realizations using observable firm-level fundamental data and market prices [6, 7, 35]. The factor model is then obtained through a regression on the estimated observable factors.

latent factors. In addition to explaining risk, the estimated latent factor models are often used to obtain better estimates of the high-dimensional covariance matrices, which are ultimately a key component in portfolio optimization [15]. Traditionally, latent factor models in finance assume that the factors are pervasive (they affect all assets) and thus can be found as common components in high-dimensional asset return time series [20, 41]. On the other hand, some recent results suggest that assets indeed tend to form clusters and communities which can be observed in their dependence network structures (modelled either by correlation or other measures of connectedness) [42, 43]. Assuming a strict hierarchical clustering structure, Tumminello et al. [44] form a hierarchical latent factor model and propose an estimation method based on the minimum spanning tree of the underlying assets. Clusters of assets are also known to emerge in stocks of single equity markets (for instance, clusters of stocks belonging to the same sectors) - Kakushadze et al. [45] consider clustering techniques for estimating these groups from the asset return time series. Verma et al. [46] proposed a cluster-specific factor model for the log-volatility with the goal of studying the heteroskedastic properties of volatility in financial assets returns. Other clustering approaches were also shown to improve high-dimensional covariance matrix estimates, which ultimately reduces risk in optimized portfolios [17, 18, 47, 48]. However, while the evidence on the existence of asset clusters is compelling, certain latent factors may still be pervasive and affect all assets - for instance, global macroeconomic shocks or the market factor [49, 50]. These may not be omitted in the search for asset clusters. To fully exploit the structural properties and obtain better latent factor models, both the asset clustering as well as latent pervasive and clusters-specific factors need to be estimated from the data. For instance, in a global set of financial assets, pervasive global factors may affect all time series (such as the global macroeconomic and market shocks), and cluster-specific factor related to certain countries will affect only specific clusters of assets (for instance, European stocks will be affected by their own set of factors and may not be affected by some Asian market factors, after controlling for the common global component).

In this thesis a clustering and latent factor estimation method is proposed which simultaneously estimates the unknown cluster structures with the pervasive and cluster-specific latent factors. An *approximate factor model*[†] is considered, which belongs to a class of models proposed by Ando and Bai [51, 52], who consider panel data with observable pervasive and unobserved pervasive and cluster-specific factors. The variability of asset returns is decomposed into the variability explained by pervasive factors, cluster-specific factors and idiosyncratic components. The pervasive factors affect all asset return time series, and these assets are divided into clusters in which a certain number of cluster-specific latent factors (the number of which may vary between clusters) affect the assets within that cluster. Since the clustering procedure may

[†]Approximate factor models, as opposed to strict factor modes, allow for correlated residuals, thus relaxing the strict assumption of a diagonal residual covariance and allowing for off-diagonal non-zero covariance elements, providing a more realistic assumption on the data.

be biased towards the clusters with a larger number of cluster-specific factors (due to the fact that more factors will always be able to explain more variability in the data), the algorithm is divided into two main phases: the clustering phase which uses a fixed number of cluster-specific factors for all clusters, and the latent factor estimation phase based on the estimated asset clusters. A computational approach to model selection is also proposed, which detects the number of pervasive factors, the number of clusters and the number of cluster-specific factors in each cluster.

Since there is no "ground truth" in financial data (the number of factors, the factors themselves, as well as the clusters are all unknown), a simulation framework is developed based on data generating processes (DGPs) which feature heavy-tailed distributed returns and correlated residuals (thus replicating statistical properties of asset returns), in which the ground truth is known - allowing to measure the performance of the estimation procedure and the model selection method. Furthermore, datasets covering global financial markets are considered, and the proposed methods are applied to the security return time series. Based on the estimated latent factors and clusters, a portfolio optimization method is proposed and tested using a back-testing approach. The results demonstrate the value of the proposed approach and the ability of the method to reduce risk in portfolio optimization scenarios.

The rest of the thesis is organized as follows. In Chapter 2 an overview of the basic principle of financial time series is given, and some of their most important statistical properties are studied. Chapter 3 first provides an introduction to the factor modelling approach and the factor models in finance. Latent factor models are also introduced, with some state-of-the-art methods for the estimation of latent factor structures in high-dimensional financial time series. A new model with pervasive and cluster-specific factors is defined, and a novel latent factor estimation algorithm is proposed. In addition, a model selection procedure is also developed, based on the spectra of the empirical correlation matrices and security graphs. Moreover, an algorithm for the estimation of the model-implied covariance matrix with sparse idiosyncratic component correlations is also proposed. In Chapter 4 several portfolio optimization approaches are considered with the goal of applying the latent factor estimates to obtain improved portfolios. In addition to the minimum variance portfolio, a novel cluster-based portfolio optimization method is proposed, and some conditions for optimality in the mean-variance sense are discussed. Chapter 5 first introduces a simulation framework, which replicates the properties of the high-dimensional financial time series under the assumption of the latent factor model with pervasive and cluster-specific factors. The simulation results are discussed and a statistical analysis is performed, which indicates that the proposed method yields accurate estimates which outperform other benchmark methods, and is robust to high-dimensionality and heavy tails in the data distributions. Moreover, historical market data are used to gauge the out-of-sample explanatory power of the latent factor estimates given by the proposed method and test the considered portfolio

optimization methods. The results demonstrate that the proposed method yields relevant and robust estimates of latent pervasive and cluster-specific factors, which can be applied to market data for improved risk modelling and portfolio optimization. Finally, Chapter 6 ends with a conclusion.

Chapter 2

Financial time series

2.1 Introduction and basic principles

Financial securities represent assets traded on exchanges in which the price is determined by the market participants. These securities may represent equity ownership in publicly traded companies (stocks), government or corporate debt (bonds), or any of the other classes (index funds, ETFs, options, futures etc.). Financial exchanges serve as matching services in which the traders on the supply and demand side post bid (buy) and ask (sell) limit orders which are stored in the limit order book. As soon as any of the buyers/sellers posts a market order, accepting the best opposing ask/bid limit order, a trade is executed at the limit order price. The price of the traded security is thus determined by this market mechanism, also known as double auction (since the buyers and sellers compete, forming bid/ask queues in the limit order book). This process continues during the exchange trading hours, generating millions of trades and respective price quotes in a single trading day [53]. Therefore, the prices are an outcome of the consensus mechanism in which investors with different strategies and valuations of the underlying asset agree on the prices at which the security is traded. Since the prices reflect information held by the market participants, they are in the focus of the majority of financial and economic applications, from forecasting to risk modelling and asset allocation [54]. In addition to the security price $S(t)$ at time t, the variable of most interest is the δ-period rate of change, which can expressed either as periodic (linear) return $R_\delta(t)$ or continuous (logarithmic) return $r_\delta(t)$:

$$R_\delta(t) = \frac{S(t) - S(t-\delta)}{S(t-\delta)} = \frac{S(t)}{S(t-\delta)} - 1, \tag{2.1}$$

$$r_\delta(t) = \log S(t) - \log S(t-\delta) = \log \frac{S(t)}{S(t-\delta)}. \tag{2.2}$$

Linear returns are associated with the discrete (periodic) compounding method:

$$S(t) = S(t-\delta) \cdot (1 + R_\delta(t)), \tag{2.3}$$

and the log-returns are associated with the continuous compounding method:

$$S(t) = S(t-\delta) \cdot e^{r_\delta(t)}. \tag{2.4}$$

Generally, prices are nonstationary variables (with a drift depending on the mean rate of return), and are serially dependent, since the price $S(t)$ depends on the past price $S(t-\delta)$, as seen in both (2.3) and (2.4). In adition, security prices may take on values of different orders of magnitude, ranging from pennies to thousands, since the underlying assets may be divided into different numbers of shares (e.g. a company divided into 100 shares of 10$ is worth the same as a company divided into 10 shares of 100$).On the other hand, as seen in Figure 2.1, returns resemble a stationary variable (although not strict-sense stationary, as will be discussed in the following sections), and are usually used in studying financial risk. Moreover, the time step and the return periods δ are usually discrete (minutely, hourly, daily, weekly, or monthly prices and returns).

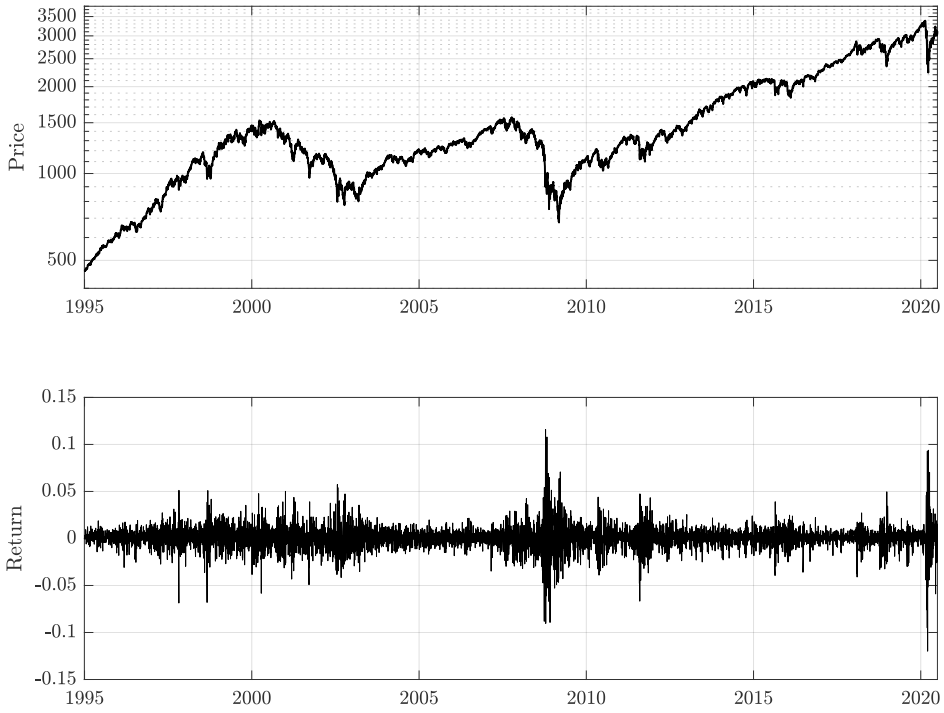


Figure 2.1: Daily prices (above, displayed in log-scale) and periodic returns (below) of the S&P 500 index, which contains the 500 largest U.S. publicly traded companies. The returns resemble a noise signal with a changing variance, as is best visible in the increased variance around the dot-com bubble of 2000-2002, the global financial crisis of 2007-2008 and the COVID-19 pandemic crisis of 2020.

Even though the difference in values between these two types of returns is notable only

for large magnitudes (for small price differences $R(t)$ and $r(t)$ are virtually the same), they do have some properties useful for different situations. Logarithmic returns can be aggregated over time, meaning that the total log-return over multiple periods $t, ..., t+\tau$ is simply the sum of log-returns: $r(t) + r(t+1) + ... + r(t+\tau)$. On the other hand, linear returns can be aggregated across securities, meaning that the return of a portfolio which has weights $w_1, ..., w_N$ across N securities is the sum: $w_1 R_1(t) + ... + w_N R_N(t)$. Across large sets of assets or large time frames, these differences can accumulate, and it is thus important to use the proper data depending on the application [55]. In the remainder of this thesis, linear returns will be considered, since the approach focuses on high-dimensional data across a large number of assets N. Nevertheless, the following section lists some of the most important statistical properties of financial returns which hold for both types of returns.

2.2 Statistical properties

Many empirical financial studies and decades worth of evidence suggest that price dynamics and financial returns exhibit certain statistical properties (also known as *stylized facts*) which universally hold across seemingly different markets [56] - ranging from international stock markets [57] to currency pairs and precious metal prices [58], and even novel and unregulated markets such as cryptocurrencies [59]. These phenomena include primarily:

(i) absence of autocorrelations (returns are serially uncorrelated),

(ii) volatility clustering (autocorrelation of absolute returns),

(iii) heavy tails (the tails of the distributions of financial returns decay following a power-law).

Other more specific effects have also been documented, such as the existence and the distributions of drawdowns and drawups [60], leverage effect [61], the volume-volatility correlation [62] etc. - however, these are not considered in detail in this theses (for a review of statistical properties of asset returns, see Cont [21]). In recent years, these phenomena gained a fair amount of attention as multidisciplinary research endeavors brought together methods from physics, statistics, computer science and game theory to develop models explaining the statistical properties of financial returns [63, 64, 65]. These properties are of critical importance in building risk models which attempt to explain the variability in high-dimensional return time series, since they must be able to work well in the presence of these stylized facts, rather than just in sterile i.i.d. Gaussian simulation environments.

2.2.1 Absence of autocorrelation

One of the most fundamental properties of financial returns is the absence of linear serial correlations. Consider the sample autocorrelation function of return time series:

$$f_R(\tau) = \frac{\sum_{t=1}^{T-\tau} [R(t) - \mu_R] [R(t+\tau) - \mu_R]}{\sum_{t=1}^{T} [R(t) - \mu_R]^2}, \tag{2.5}$$

where $R(t)$ is the return at time step t, μ_R is the sample mean of $R(t)$, and τ is the time lag at which the autocorrelation function is estimated. The autocorrelation $f_R(\tau)$ is known to be zero or very close to zero for all time lags $\tau > 0$ in daily, weekly, or monthly data [21]. Figure 2.2 displays the autocorrelation functions of returns for a number of U.S. stocks, demonstrating how this finding holds for a multitude of the considered stocks.

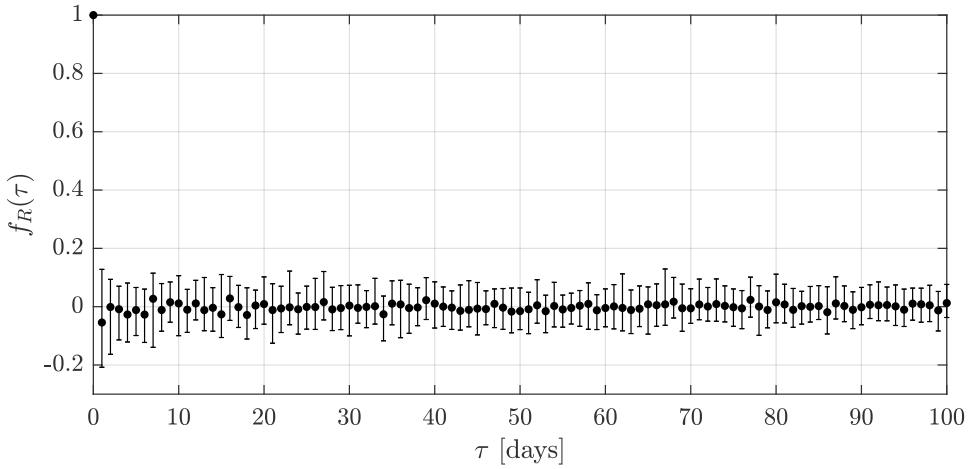


Figure 2.2: Autocorrelations of linear returns of 600 publicly traded U.S. companies between 2005 and 2020. The dots for each lag τ are the mean and the lines represent the minimum and maximum autocorrelation for that lag, among all stocks.

This property also holds for the mean return $\mu_i(t)$ estimated on a given time window $(t-T,t]$ and the future subsequent mean $\mu_i(t+T)$ on the next time window $(t,t+T]$. In other words, returns (and their mean values) are hard to predict using past return data, as is also demonstrated by the result in Figure 2.3, which displays the mean returns estimated using a dataset of 600 U.S. stocks between 2005 and 2020.

However, it is important to note that for high-frequency intraday returns there are some statistically significant autocorrelations below the $\tau = 5$ min. mark – this is due to market microstructure (order book and the bid-ask spread) and in some cases the reaction of markets to news and investor behavior [66]. According to some studies [65], certain negative autocorrelations will occur when investors overreact to news and the price rebounds – on the other hand, investor underreaction to news will cause the price to shift slowly over time, resulting in positive autocorrelations. The temporary existence of these autocorrelations is confirmed by emergence

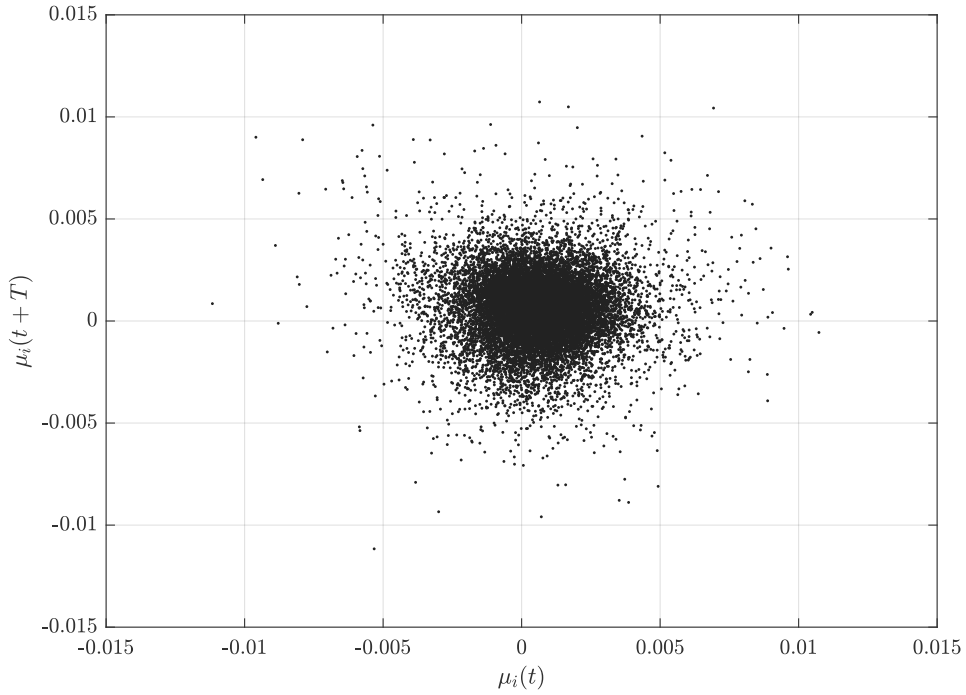


Figure 2.3: Mean return $\mu_i(t)$ estimated on a $T = 6$ month period and future 6-month mean returns $\mu_i(t+T)$ of each stock i in a dataset of 600 U.S. stocks. The past 6-month mean returns and future 6-month mean returns are evidently not correlated, with Corr $\left[\sigma_i^2(t), \sigma_i^2(t+T)\right] = -0.03$.

of momentum and mean-reversion in financial markets [67] – however they may shift abruptly and is generally more persistent on an intraday level [68].

Thus, excluding high-frequency trading data and temporary phenomena, for all other practical purposes the absence of autocorrelations is a well-established fact and is often used to support the *efficient market hypothesis* (EMH*) [69]. Moreover, the absence of return autocorrelations on the long-run is intuitively easy to understand: if price changes exhibit significant temporal correlations, they would be used to formulate simple trading strategies (so-called *statistical arbitrage* [70]) with positive expected returns, which will in turn reduce these correlations through trading activity in the market. Such strategies do exist – however, they are not profitable on return autocorrelations, but are rather applied to statistical market findings which are known to be more persistent and significant, such as cross-correlations between assets [71, 72].

2.2.2 Volatility clustering

Another important statistical property of financial asset returns is *volatility clustering* - large price fluctuations tend to cluster in time, thus exhibiting autocorrelations of the return amplitudes [73]. This is also seen in the return time series displayed in Figure 2.1 - periods of high-

*There are mutiple interpretations and stipulations of the EMH – the so-called "weak form" EMH states that past prices and returns do not influence future price movements.

volatility tend to persist – as well as periods of low volatility. Considering the autocorrelation of absolute returns:

$$f_{|R|}(\tau) = \frac{\sum_{t=1}^{T-\tau} \left[|R|(t) - \mu_{|R|}\right] \left[|R|(t+\tau) - \mu_{|R|}\right]}{\sum_{t=1}^{T} \left[|R|(t) - \mu_{|R|}\right]^2}, \tag{2.6}$$

as opposed to the return autocorrelation $f_R(\tau)$, the absolute return autocorrelation $f_{|R|}(\tau)$ will exhibit positive values over long ranges of lags τ. This is demonstrated in Figure 2.4, which displays the absolute return autocorrelation function for a set of U.S. stocks over daily time lags.

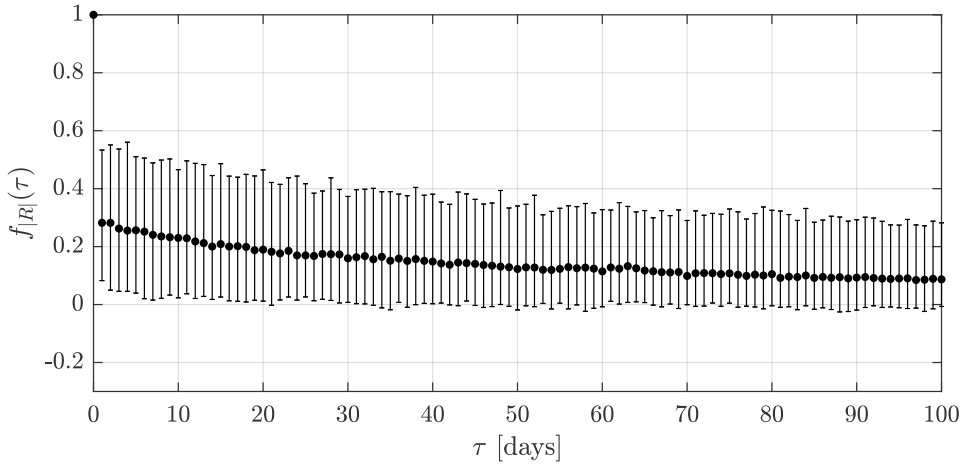


Figure 2.4: Autocorrelations of absolute linear returns of 600 publicly traded U.S. companies between 2005 and 2020. The dots for each lag τ are the mean and the lines represent the minimum and maximum autocorrelation for that lag, among all stocks.

The autocorrelation of absolute returns is just one of many ways to quantify and measure this phenomenon – instead of absolute returns, one could measure the autocorrelation of squared returns, or the correlations of variance or standard deviation estimates across subsequent time intervals. Consider the variance $\sigma_i^2(t)$ of each asset $i = 1, ..., N$, which can be estimated at time step t using look-back windows of length T and at subsequent time steps $t+T$, yielding the estimate $\sigma_i^2(t+T)$. Figure 2.5 shows these variances for time windows of length $T = 1$ year (which is approximately 252 trading days) for over 600 U.S. stock return time series. This result also confirms the positive association of past volatility with future volatility holds, thus demonstrating the existence of so-called -ARCH (autoregressive conditional heteroskedasticity) effects in financial time series [21]. Therefore, although there are no serial correlations, financial returns are obviously not independent.

This property can be traced to behavioral effects of market participants – when uncertainty is high, investors are not confident about their valuations and the spread between supply and demand (bid and ask) is larger [73]. In addition, effects such as herding behavior cause investors to overreact and drive prices to extremes, also increasing volatility. This phenomenon has been in the focus of many agent-based models and game-theoretic approaches attempting to provide

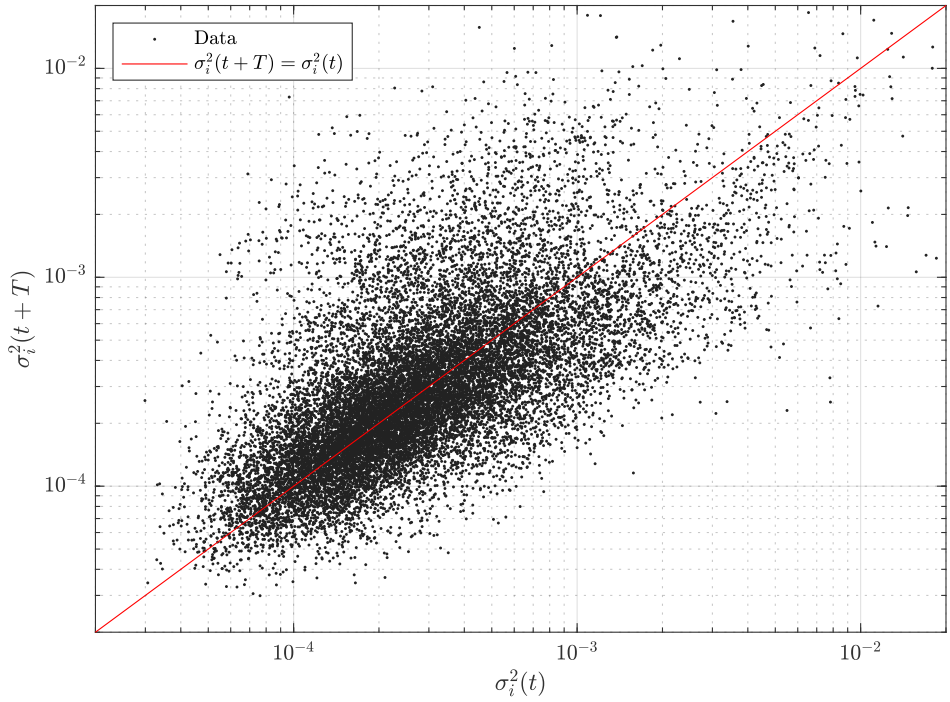


Figure 2.5: Variances $\sigma_i^2(t)$ (shown in log-scale) estimated on a $T = 6$ month period and future 6-month variances $\sigma_i^2(t+T)$ of return time series in a dataset of 600 U.S. stocks. The red line represents a simple model $\sigma_i^2(t+T) = \sigma_i^2(t)$ (although not necessarily the optimal linear fit). The past 6-month variances and future 6-month variances are evidently correlated, with Corr $\left[\sigma_i^2(t), \sigma_i^2(t+T)\right] = 0.48$.

insight into its origin [74, 75]. Moreover, some results suggest that the volatility dependence exhibits long-memory properties, meaning that the effects may last for weeks [61, 76].

In addition, not only do the volatilities exhibit long-range memory, but so do the covariances and cross-correlations between assets [12]. Figure 2.6 displays the subsequent $T = 1$ year estimates of crosscorrelations $\rho_{ij}(t)$ and $\rho_{ij}(t+T)$ (estimated as the sample Pearson correlation coefficient) for each pair i, j in the dataset of U.S. stock return time series.

These important statistical findings – the fact that variability and cross-correlations of financial asset return time series are autocorrelated and dependent in time – give rise to many covariance-based methods for risk modelling [77, 78]. Indeed, as seen in this section and the previous one - the first moment of financial return time series (or, more simply put, their direction) is hard to predict, but the second moment, especially in multivariate cases (which boil down to the covariance matrices of asset returns) does exhibit memory and can be modelled. Most financial risk models will therefore not attempt to predict whether stocks will rise or fall, but rather explain their volatility and cross-correlations, which will generally hold regardless of their directional changes [75, 79]. Regardless of their predictability, it is also a well-documented fact the correlation structures in financial markets may change, especially in the presence of bubbles and crashes [13, 43], and therefore historical windows used to estimate risk models may only be of limited lengths. This property of memory in correlations structures and the

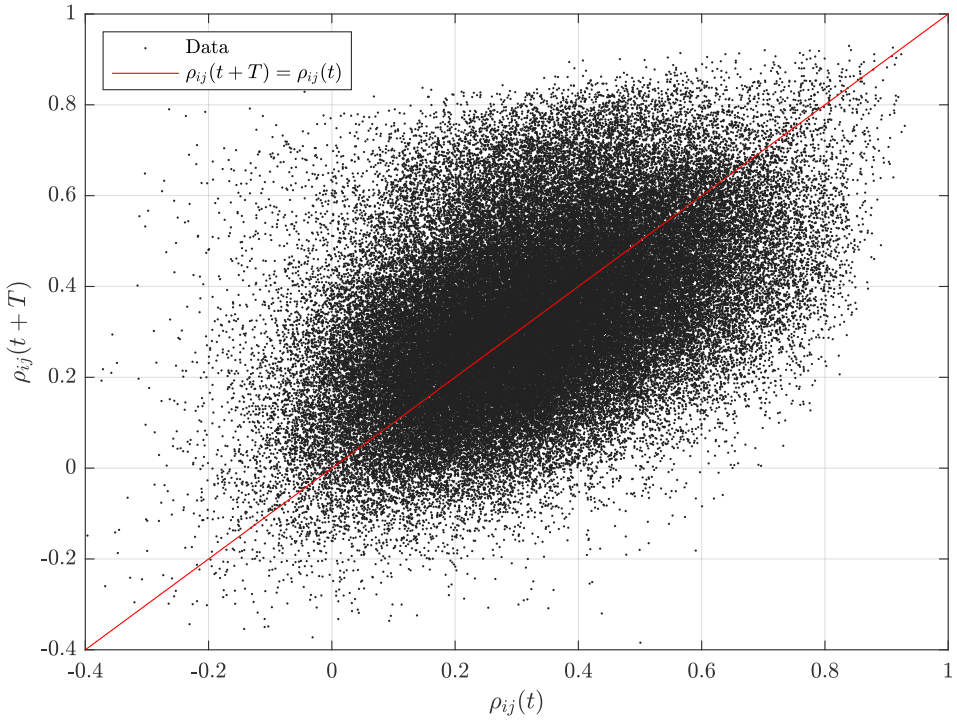


Figure 2.6: Cross-correlations $\rho_{ij}(t)$ estimated on a $T = 6$ month period and future 6-month cross-correlations $\rho_{ij}(t+T)$ of pairs i, j of return time series in a dataset of 600 U.S. stocks. The red line represents a simple model $\rho_{ij}(t+T) = \rho_{ij}(t)$ (although not necessarily the optimal linear fit). The past 6-month cross-correlations and future 6-month cross-correlations are evidently correlated, with $\mathrm{Corr}\,[\rho_{ij}(t), \rho_{ij}(t+T)] = 0.54$.

aforementioned finite sample issues are some of the essential concepts upon which this thesis is also built.

2.2.3 Distributions and extreme values

The properties of return distributions have been in the focus of decades of research, and are still an important topic in modelling financial risk. Some of the first models assume Gaussian returns, which in turn lead to some elegant mathematical properties and results, such as the mean-variance analysis [80] or the Black-Scholes pricing model [81]. However, compelling evidence from the past decades do not support the Gaussian return assumption – on the contrary, asset returns are known to have excess kurtosis, even when the ARCH effects are taken into account [21]. Table 2.1 shows the average skewness and kurtosis, as well as the median and the 5^{th} and 95^{th} percentiles for the daily returns of 600 U.S. stocks between 2005 and 2020. The statistics are calculated for both linear and log-returns. The skewness of a Gaussian normal is 0 and its kurtosis is 3, meaning that these results demonstrate a drastically different distribution in comparison. The skewness estimates of daily return time series evidently vary over different stocks, as well as over the return calculation method - this is expected since the linear returns are usually limited by -100% from below, as opposed to log-returns which are not limited

(a linear return of -100% corresponds to a log-return of $-\infty\%$). In log-returns, the negative skewness is indicative of the so-called loss/gain asymmetry, meaning that the losses (negative returns) may be more pronounced than gains, which are more stable [21]. Regarding kurtosis, it is very similar in both cases, clearly demonstrating how heavy the tails of these distributions are. Fundamentally, this means that the extreme events are much more likely to happen than a normal distribution would assume, implying that Gaussian assumptions underestimate these tail risks [82, 83].

Table 2.1: Skewness and kurtosis statistics estimated on daily linear and log returns of 600 U.S. stocks from 2005 to 2020.

	Linear returns		Log-returns	
	Skewness	Kurtosis	Skewness	Kurtosis
Mean	0.26	21.14	-0.41	21.62
Median	0.18	15.87	-0.30	16.36
5^{th} perc.	-0.51	9.06	-1.49	9.08
95^{th} perc.	1.41	45.22	0.35	45.56

These results demonstrate how heavier the tails of the empirical return distributions really are than the Gaussian normal. In fact, Mandelbrot suggested that returns follow a class of Levy alpha-stable distributions (of which the Gaussian normal is a special case), which exhibits power-law tails [84]. However, the *inverse cubic law* [85], states that the tail of the return distribution[†] follows a power-law:

$$P(X > x) \propto x^{-\alpha}, \tag{2.7}$$

where the exponent α is found to be around 3 (for the pdf this is equal to 4), which is outside the Levy regime ($0 < \alpha \leq 2$). This finding holds for a number of different asset classes, such as stocks [87] or financial market indices [88]. Although some results find different power-law exponents in other assets, such as $2 < \alpha < 2.5$ for the cryptocurrency market [89], the fact that these price fluctuations still exhibit power-law tails remains [90]. An example of this phenomenon is shown in Figure 2.7, for the Bitcoin-USD pair on several cryptocurrency exchanges [89], with the dashed lines correspond to the fitted power laws, all of them remarkably close.

This important empirical fact means that, even though the tails are much heavier than those of the Gaussian normal distribution, they have finite variances – a crucial assumption for many

[†]although the methodologies vary, the "tail" of these distributions is considered to begin after 2 or 3 standard deviations. For a comprehensive review on estimating power laws in empirical data , see [86].

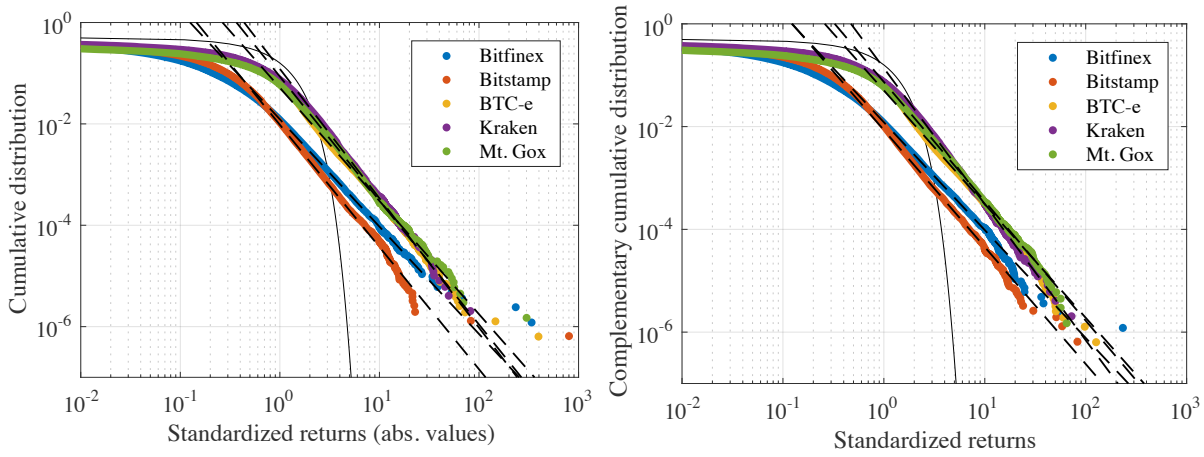


Figure 2.7: Negative (left) and Positive (right) tails of the cumulative distribution for the Bitcoin returns on several different exchanges (Bitfinex, Bitsamp, BTC-e, Kraken, and Mt. Gox) and the time scale of $\Delta t = 1$ min. The black dashed lines correspond to the fitted power-law distributions for the return tails of the considered exchanges and the black full line is the cumulative distribution function of the standard normal $\mathcal{N}(0,1)$.

volatility and covariance-based risk measures. The power-laws in the financial return distribution tails have been in the focus of many statistical and physical models, explaining this phenomenon by the distribution of wealth, market impact, or herding behavior of market participants [61, 91]. However, it is also important to note this behavior is more specific to the higher-frequency returns and that, as the time step δ is increased, the tail power-law may be less pronounced [87], and some other statistical properties may approach those of the Gaussian distirbution [21]. Nevertheless, the heavy tails and extreme values of financial returns are observed at all relevant time scales and remain a key component in modelling returns and financial risk [92].

2.3 Multivariate financial time series and risk modelling

Since risk is primarily proxied by uncertainty in the change of asset prices it is most commonly quantified using the dispersion of returns. This is primarily measured as the volatility (the standard deviation of financial returns) σ, but other common measures include Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR) and maximum drawdown [93]. The $X\%$ VaR is defined as the X^{th} percentile of the return distribution - in other words, the largest possible loss excluding the worst $X\%$ cases. The $X\%$ CVaR (also known as expected shortfall) is defined as the mean return within the X^{th} percentile - in other words, the expected loss in the worst $X\%$ cases. The maximum drawdown is simply the largest amount an investor could use on a given time frame, from peak to bottom. These measures focus more on the negative side and tail properties of the distributions, thus portraying a more accurate image of risk than volatility. However, all of

them require larger amounts of data and better knowledge of financial distributions, since they are estimated using only a fraction of the return realizations. For parametric and symmetric distributions, all of these are uniquely defined by the variance. In addition, return variance and covariance have very elegant statistical properties in terms of multivariate risk models, as will be seen in the following section. Therefore, in most practical applications, volatility and variance are commonly used as risk measures, and risk models are primarily tasked with explaining the variability of returns.

2.3.1 Return covariance and correlation

In risk modelling applications, especially for asset management and portfolio optimization, the focus is on a multivariate set of financial security return time series, representing the assets in a considered market or investable security universe. Understanding the co-movement of financial asset prices is crucial for modelling the potential downside movements and managing the risk of financial portfolios. For the remainder of this thesis, let $\boldsymbol{X} \in \mathbb{R}^{T\times N}$ denote the matrix of N time series of linear asset returns of length T. A central component in modelling the variablilty and the dependence structure of the considered assets is the covariance matrix:

$$\mathbf{Q} = (\sigma_{ij})_{N\times N}, \quad \sigma_{ij} = \mathrm{Cov}(X_i, X_j), \tag{2.8}$$

where $X_i \in \mathbb{R}^T$ is the vector of returns of asset i. The sample estimate of the covariance matrix is most commonly obtained using the unbiased estimator

$$\hat{\sigma}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}(X_{ti} - \hat{\mu}_i)(X_{tj} - \hat{\mu}_j), \tag{2.9}$$

where $\hat{\mu}_i$ is the sample mean of X_i. However, this may not necessarily be the most efficient estimator in the presence of stochastic volatility, ARCH effects, and considering that the financial time series may have some specific patterns in dependence structures, which will be discussed in the following sections and chapters of this thesis. In addition to the sample covariance matrix, another important tool in modelling the dependence structures of multivariate asset return time series is the correlation matrix

$$\mathbf{R} = (\rho_{ij})_{N\times N}, \quad \rho_{ij} = \mathrm{Corr}(X_i, X_j), \tag{2.10}$$

obtained from sample data using the unbiased estimator

$$\hat{\rho}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\frac{(X_{ti} - \hat{\mu}_i)(X_{tj} - \hat{\mu}_j)}{\hat{\sigma}_i\hat{\sigma}_j}, \tag{2.11}$$

where $\hat{\sigma}_i$ is the sample standard deviation of return time series of asset i. As opposed to the covariance, which has the individual asset variances $\sigma_{ii} = \sigma_i^2$ on the diagonal, the correlation matrix is normalized by the asset variances, which in turn makes all the diagonal elements equal to 1. The asset covariance and correlation matrices (which can be thought of as a special, standardized case of the covariance) have some important universal properties. Firstly, they are always symmetric, since $\sigma_{ij} = \sigma_{ji}$ and $\rho_{ij} = \rho_{ji}$. Due to this, they are always positive semi-definite, meaning that their eigenvalues are non-negative. In addition, a sample estimate of the covariance/correlation matrix from data $\mathbf{X} \in \mathbb{R}^{T \times N}$ will be of rank $\min(N,T)$. This means that the estimated $N \times N$ covariance matrices will be of full rank N when there is enough data, i.e. when $T \geq N$ – and in that case the estimates will be positive definite (all eigenvalues will be strictly positive). This is important, since many parametric modelling approaches such as maximum likelihood require the inverse of the covariance matrix, also known as the *precision matrix* – it will not universally exist if the matrices are not positive definite.

When modelling financial risk, the asset variances and cross-correlations are usually estimated on windows of historical data, assuming that the estimated models will hold in the near future, owing to the long-range memory properties. However, due to the dynamic nature of financial markets, the historical windows on which risk models are estimated cannot be very long. Even though volatility and cross-correlations are known to exhibit long-range dependencies, they are limited and do not last indefinitely [12]. To demonstrate this in the U.S. stock return time series, the results in Figures 2.5 and 2.6 can be extrapolated to different time lags than just two subsequent 6-month periods. Specifically, the variances $\sigma_i^2(t)$ for each asset i and cross-correlations $\rho_{ij}(t)$ for each pair i, j are estimated on rolling windows $t-T, ..., t$ of length $T = 1$ month. By doing so, for each time step t at which the correlations are calculated, the vectors $\boldsymbol{\sigma}(t) = [\sigma_1^2(t), ..., \sigma_N^2(t)]$ and $\boldsymbol{\rho}(t) = [\rho_{1,1}(t), ..., \rho_{1,N}(t), \rho_{2,1}(t), ..., \rho_{N,N}(t)]$ are defined. Then, for a given time lag τ, the average correlation $f(\tau)$ between $\boldsymbol{\sigma}(t)$ and $\boldsymbol{\sigma}(t+\tau)$ (as well as $\boldsymbol{\rho}(t)$ and $\boldsymbol{\rho}(t+\tau)$) can be calculated. Figure 2.8 displays these average correlations for the individual asset variances across a range of time lags τ up to 60 months (5 years).

It is evident that there is some pronouced autocorrelation (memory) in the subsequent estimates, as well as estimates for up to 2 years (24 months). However, after this period, the correlations fade and indicate that past variances are not as useful for modelling the future beyond several years. Moreover, this effect is seems to be even stronger in Figure 2.9, where the correlations of the cross-correlation estimate fade even quicker and seem to be very low after 1 year.

These results are in line with other recent research, suggesting that financial markets may exhibit dynamic changes to the underlying dependece and correlation structures between assets [13, 43]. These are intrinsically related to the multivariate models of financial risk, especially when they are estimated using historical return data. A way to model these dynamic properties

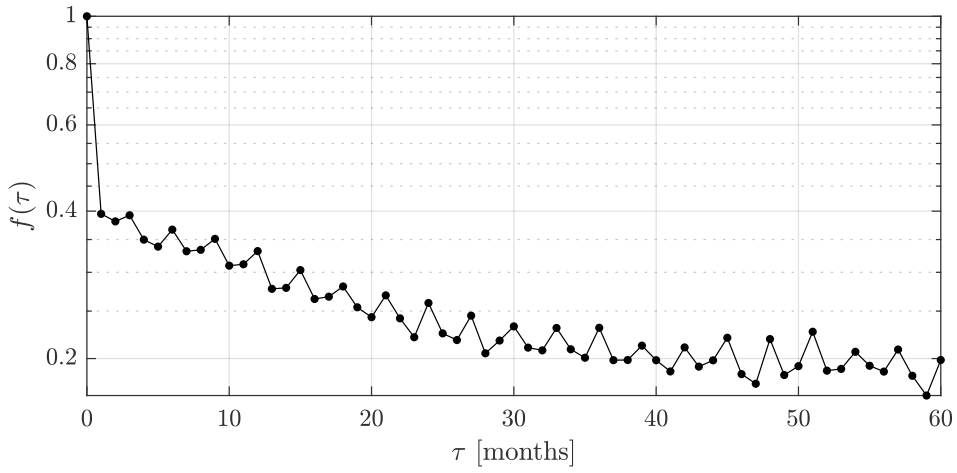


Figure 2.8: Correlations of the individual asset return variances estimated for 600 U.S. stocks on a rolling window of $T = 1$ month, for a range of time lags τ up to 60 months (5 years). The y-axis is shown in log-scale.

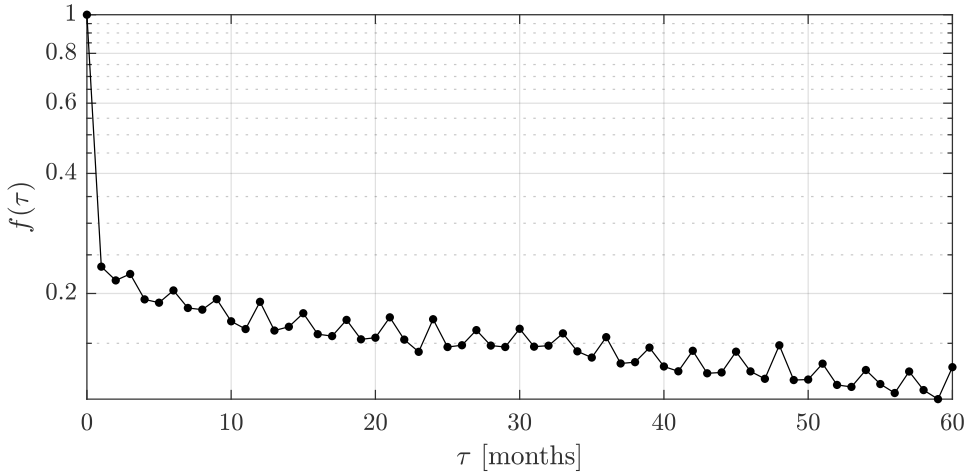


Figure 2.9: Correlations of the pairwise asset cross-correlations estimated for 600 U.S. stocks on a rolling window of $T = 1$ month, for a range of time lags τ up to 60 months (5 years). The y-axis is shown in log-scale.

are the multivariate ARCH-type models [94]. These models have many different forms [95, 96], all of which model the dynamic covariance matrices as autoregressive processes, while parameterizing the model and ensuring their positive semi-definiteness in different ways. However, most of these (unless they focus on low-rank representations, which will be in the focus of the next chapter of this thesis), come with an order of magnitude of N^2 parameters [94], and quickly become computationally infeasible when the number of time series N is large. For these reasons, rather than modelling their dynamic properties, high dimensional covariance and correlation matrices are most commonly estimated on a fixed estimation windows and are assumed to hold in the near future. Since the estimation windows may only be of limited length and since the number of assets in modern risk management applications is high, certain estimation issues arise, which will be discussed in the following section.

2.3.2 High-dimensional estimation issues

The development of financial technology, securitization, and globalization of financial markets have sparked an unprecedented growth in the number of investable assets. This drastically increases the dimensionality of many risk modelling problems, while the estimation windows remain of limited length, as seen in the previous section. Thus the number of time series N will in most cases be commensurate to their length T, or even higher. The estimation problems of such high-dimensional covariance matrices are most commonly observed by analyzing their spectra, mainly stemming from the field of random matrix theory [15, 97]. A cornerstone of this entire approach is the Marčenko-Pastur law [98], which provides a theoretical distribution of the spectra of high-dimensional covariance estimates. Consider a random vector $X_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_{N\times N})$. The true covariance is an identity matrix $\mathbf{I}_{N\times N}$, but its sample estimates will differ depending on the amount of data. In fact, the spectra of the covariance estimates will follow a specific distribution which is parameterized only by the ratio $q = N/T$ where N is the dimensionality of the data (number of assets) and T is the number of data points (the length of sample time series). In the limits $N \to +\infty$ and $T \to +\infty$, when $0 < q < 1$, the distribution density function reads:

$$f_{MP}(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi q \lambda}, \quad \lambda \in [\lambda_-, \lambda+], \tag{2.12}$$

where λ_- and λ_+ are the lower and upper bounds of the spectrum:

$$\lambda_\pm = (1 \pm \sqrt{q})^2. \tag{2.13}$$

This means that, depending on the length of time series T, the estimated covariance matrix spectra will be somewhat blurred as opposed to the true eigenvalues. This can be seen in Figure 2.10, where the Marčenko-Pastur distribution is displayed for different values of the dimensionality ratio q together with the true eigenvalues (which are all equal to 1). A striking fact, visible in these results, is that the spectrum is already very broad for $q = 0.25$ - meaning that even if $T = 4N$, the estimates are still not very accurate (estimated eigenvalues ranging form 0.25 to 2.25). Moreover, when the ratio q approaches 1, the estimated spectrum becomes very wide and there is a bulk of eigenvalues very close to 0, which may result in numerical instabilities.

An important consequence of the Marčenko-Pastur law is the fact that the sample eigenvalues are bounded. However, as seen in the previous Figure, these bounds diverge very quickly as q approaches 1. To illustrate this, Figure 2.11 displays the density function of the eigenvalue distribution according to the Marčenko-Pastur law, for values of q between 0 and 1. As the ratio q approaches 1 (i.e. as the number of time series N approaches their length T), the covariance estimate eigenvalues diverge and their lower limit quickly approaches 0. As stated above, for $N > T$, the covariance estimates will be of rank T, and exactly $N - T$ eigenvalues will be equal

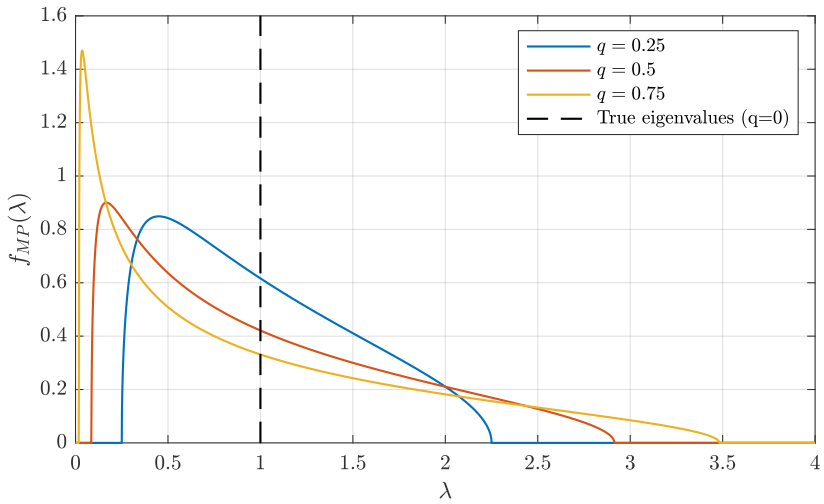


Figure 2.10: The theoretical Marčenko-Pastur distributions for different values of the dimensionality ratio $q = N/T$, together with the true eigenvalues, which are in this case all equal to 1.

to 0.

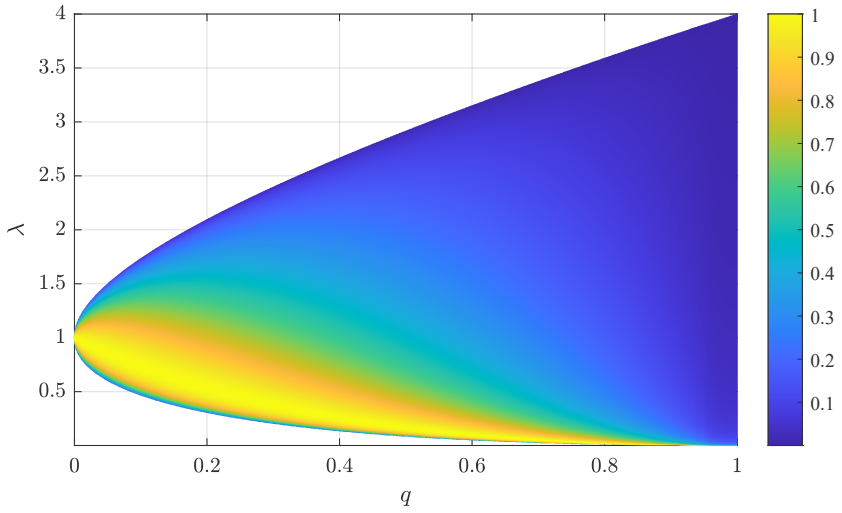


Figure 2.11: The normalized densities of sample eigenvalues following the Marčenko-Pastur law for different values of of the dimensionality ratio $q = N/T$. The area outside the $[\lambda_-, \lambda_+]$ range is not colored (white).

Similar results have also recently been obtained for the Student's t-distributed data [99] and for low-rank matrices with additive noise [100]. Evidently, when the ratio q is not vanishingly small (i.e. when N i commensurate to T), the sample covariance estimates are not trustworthy. This issue becomes even more pronounced when considering high-dimensional financial returns which are not independent variables, but have a specific correlation structure and common components. In this case, the theoretical bounds of the Marčenko-Pastur law are often used to discern between the eigenvalues (and their respective eigenvectors) representing meaningful data, and those attributed to the noise [42]. However, since financial return time series have distributions with heavy tails and extreme values, the spectra of empirical return correlation

and covariance matrices are even more dispersed in high-dimensional situations [99]. These issues make the usage of methods relying on sample distributions (such as maximum likelihood estimation) impractical and may have severe consequences on any risk management applications relying on these estimates. Thus, the focus of modern estimation methods and models is on increasing the robustness of high-dimensional covariance and correlation estimates, while reducing the number of parameters and regularizing the estimation procedures [101].

2.3.3 Shrinkage estimation

To combat the estimation issues in high-dimensional financial covariance matrices, a number of shrinkage (regularization) methods have been proposed. In a bias-variance tradeoff sense, the shrinkage estimation of covariance matrices reduces the sensitivity of the estimates to the high-dimensionality issues at the expense of introducing some oversimplifications or model bias [102]. Some basic shrinkage methods focus directly on the covariance matrix estimate, by altering the empirical estimate towards a shrinkage target. A linear shrinkage estimator is given by:

$$\mathbf{Q}^{(\text{shrinkage})} = \alpha\hat{\mathbf{Q}} + (1-\alpha)\mathbf{Q}^{(\text{target})}, \tag{2.14}$$

where $\hat{\mathbf{Q}}$ is the sample estimate and $\mathbf{Q}^{(\text{target})}$ is the shrinkage target. The *basic linear shrinkage* estimator has a diagonal matrix with individual sample variances on the diagonal as the shrinkage target

$$\mathbf{Q}^{(\text{target})} = \text{diag}(\hat{\sigma}_1^2, ..., \hat{\sigma}_N^2). \tag{2.15}$$

By reducing the off-diagonal estimates towards zero, the final covariance estimate is regularized and large off-diagonal elements are avoided.

A somewhat more advanced estimator uses the *constant correlation model* for the shrinkage target, where the off-diagonal elements are calculated assuming a fixed correlation for all pairs, calculated as the average pairwise correlation

$$\boldsymbol{\sigma}_{ij}^{(\text{target})} = \hat{\sigma}_i\hat{\sigma}_j\overline{\rho}, \quad \overline{\rho} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\rho_{ij}. \tag{2.16}$$

This estimator has been found to improve portfolio optimization based on high-dimensional covariance estimates, with resulting portfolios being more diversified and exhibiting less risk out-of-sample [103].

Another popular approach is to take advantage of the so-called *market mode* in financial data – the eigenvector $\mathbf{u}_1$ corresponding to the largest eigenvalue λ_1 of the sample covariance matrix [99, 104]. This eigenvector represents the market factor, to which all of the securities in a given market are exposed and is often the strongest factor in financial markets. An estimator

for the high-dimensional covariance matrix would then shrink the sample estimates towards the single-index model covariance [49]:

$$\mathbf{Q}^{(\text{target})} = \mathbf{u}_1 \lambda_1 \mathbf{u}_1^\intercal + \mathbf{\Psi}, \tag{2.17}$$

where $\mathbf{\Psi}$ is a diagonal matrix of the residual variances, such that the diagonal of $\mathbf{Q}^{(\text{target})}$ is the same as that of $\hat{\mathbf{Q}}$. This estimator has been found to perform very well when considering large portfolios of assets from a given market [105]. This is an elegant and simple way to statistically exploit the fact that all securities in a given market are known to be exposed to the market factor (one of the fundamental findings of pricing models [106]). However, it is also a known fact that other risk factors may explain significant amounts of variance in returns [4]. Moreover, when considering securities belonging to different asset classes and various markets, the single factor assumption does not hold.

In recent years, more advanced estimators have been proposed, following either model-free or model-based approaches. Model-free methods do not assume an underlying model for the observed time series realizations, and instead focus on the spectral properties of the high-dimensional sample covariance matrices, mostly building on the tools from random matrix theory (RMT) [15, 107]. A most notable method is the *rotational invariant estimator* by Bun et al. [108]. This estimator focuses on the bulk of small eigenvalues, since they tend to be underestimated by the sample estimator. This is verified by comparing the in-sample estimates of the eigenvalues λ with their respective out-of-sample (oracle) estimates ξ, as seen in Figure 2.12. The oracle estimates are obtained as the out-of-sample variances of the respective eigenvectors (which are estimated in-sample). The figure, displays a non-linear phenomenon in the relationship between the in-sample eigenvalue estimates and their oracle values, suggesting that the very smallest eigenvalues (between 0 and 0.5) may be overestimated, while the bulk of the small eigenvalues are underestimated (they are below the dashed line). The RIE estimator builds on tools from RMT and delivers an estimator which corrects these estimates, without changing the eigenvector estimates or assuming an underlying model [109].

Some methods focus on unsupervised learning for detecting more complex correlation structures in the data, but do not associate them with a specific model of the observed time series [42, 48, 110]. Among these, the eigenvalue clipping approaches, as opposed to the RIE estimator, focus on the largest K eigenvalues in the data, preserving them while altering (clipping) the rest:

$$\mathbf{Q}^{(clip.)} = \sum_{k=1}^{N} \xi_k^{(clip.)} \mathbf{u}_k \mathbf{u}_k^\intercal, \quad \xi_k^{(clip.)} = \begin{cases} \lambda_k, & \text{if } k \leq K \\ \gamma, & \text{otherwise} \end{cases}, \tag{2.18}$$

where $\mathbf{u}_k$ and λ_k are eigenvectors and eigenvalues of the sample covariance $\hat{\mathbf{Q}}$, sorted by de-

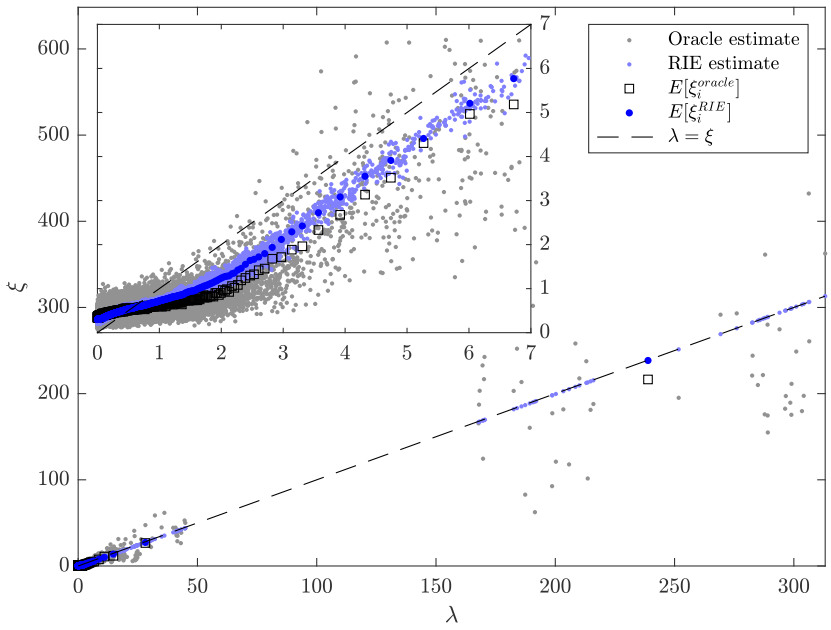
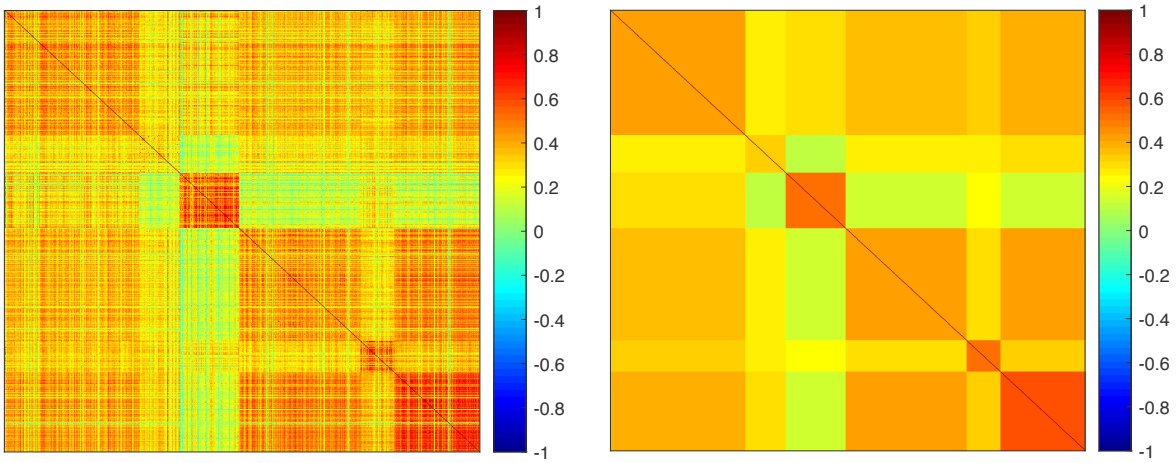


Figure 2.12: The eigenvalues of the correlation matrices of 600 U.S. stocks estimated on a rolling look-back window of $T = 3$ years and a look-ahead window of $T' = 2$ months for the oracle eigenvalues. The pale grey and blue dots represent the oracle and RIE estimates for the entire sample, and the averages for each eigenvalue are displayed in black squares (oracle) and blue circles (RIE).

scending eigenvalue magnitude, and γ is a constant calculated as the mean of the remaining eigenvalues ($k > K$), to preserve the matrix trace. The number of eigenvalues K to keep can be estimated using various estimators for the number of components in the data, including the Marčenko-Pastur (MP) threshold. Although it corrects the smaller eigenvalues, this method may yield varying results, depending on the noise present in the eigenvectors associated with the smallest eigenvalues. Nevertheless, estimators using eigenvalue clipping as one of the shrinkage targets have been shown to yield relevant estimates which help improve the properties of the optimal portfolios based on these matrices [110].

Another approach focuses on the clustering property of financial securities, which have been shown to form groups, depending on their industry classification, asset class or country of origin [42, 45]. Begušić and Kostanjčar [48] propose a cluster-based shrinkage target which combines a clustering procedure and the constant correlation method. Specifically, assets are clustered into a number of clusters K (also estimated using the MP threshold from the empirical correlation matrix). Then, the correlations within each cluster are estimated as the average pairwise correlation of all assets within the cluster and he correlations between two clusters as the average pairwise correlation of all assets in the two clusters. This target correlation matrix is shown in Figure 2.13, together with the original empirical estimate. By clustering the data, the cluster structures in the empirical matrix (Figure 2.13a) are uncovered in the block-structured target matrix (Figure 2.13b). The final shrinkage estimator is then defined as a linear combination of these two.



(a) Empirical estimate of the correlation matrix.

(b) Shrinkage target.

Figure 2.13: The empirical estimate and the shrinkage target for the 600 U.S. stock returns, using the cluster-based shrinkage approach.

Shrinkage estimators can be used to facilitate likelihood-based methods for estimation of other models, such as latent factor models. Alternatively, factor models may also be used to obtain better estimates of the covariance and correlation matrices by exploiting the knowledge of the underlying factors driving observed security returns. Thus, the estimation of latent factors in high-dimensional data may also be seen as a model-based approach to the estimation of high-dimensional covariance and correlation matrices. The following chapter contains a deeper look into the latent factors in security returns and the estimation of latent factor models from sample data.

Chapter 3

Latent factor estimation in financial time series

3.1 Factor models in finance

The existence of a number of risk factors which drive the observed security returns has been a well-documented fact in finance. Different approaches to factor modelling rely either on observable factors such as macroeconomic time series (inflation, productivity, GDP, etc.), the market return (the returns of a broad market index), or try to find and estimate these factors from the data (which can include fundamental firm-level characteristics or just price data). The estimation of risk factors and modelling their impact on returns of securities has been gaining increased attention in the financial academic community as well as in the financial industry, especially asset management. Some of the first approaches to factor modelling come from the area of asset pricing, focusing mostly on explaining the so-called *cross-section* of expected returns. Building on the seminal work of Markowitz [80] and his modern portfolio theory, the capital asset pricing model (CAPM) has been a cornerstone of financial risk modelling for decades [106]*. CAPM models the expected return of a security or a portfolio of securities by their exposure to the market rate of return: $\mathrm{E}[R_i] = R_f + \beta_i(\mathrm{E}[R_m] - R_f)$, where R_f is the risk-free rate of return and R_m is the market return. The risk model associated with CAPM is the Sharpe single-index model:

$$R_i - R_f = \alpha + \beta_i^{(M)}(R_m - R_f) + e_i. \tag{3.1}$$

The only factor modelling the returns R_i of asset i is the so-called market factor, expressed as market return R_m. The important stipulation of this model is that a long-only portfolio of securities can diversify the idiosyncratic sources of risk (individual asset risk), but can not diversify

*Harry Markowitz, William Sharpe, and Merton Miller shared the 1990 Nobel Memorial Prize in Economics for this contribution to financial economics.

the systematic risk (which all of the securities in the portfolio are exposed to). Regarding expected returns and the CAPM, the model implies that the exposure to the systematic market risk is rewarded by harvesting the market risk premium. The higher the β_i, the more systematic risk a security will assume, and the higher its rate of return should be. However, empirical evidence shows that this may not always be the case [9], ranging from the low-volatility premium [111] (the fact that low-volatility stocks do not exhibit lower returns as expected by the CAPM), or the existence of multiple factors in the returns of financial securities [4]. Since the introduction of CAPM, new factor models have been proposed, following one of three main approaches [36]: macroeconomic, fundamental, and statistical (latent).

Macroeconomic factor models focus on estimating the impact of observable economic factors such as inflation, interest rates or other macroeconomic data on the security returns [112]. The factors are known and measurable, and are assumed to represent the main sources of risk to which the securities are exposed. However, the macroeconomic factors may not be responsible for the risk in all securities and markets, and their view may be limited – nevertheless, the macroeconomic risk factors are often included or identified in other types of factor models [113]. The somewhat more prominent fundamental and latent factor models are described in more detail in the following sections.

3.1.1 Fundamental factor models

Fundamental factor models focus on some economic properties of financial assets (mostly applied to stock and bond returns) – for instance the book value, market capitalization or other company characteristics [4, 114]. These models often include the market factor from CAPM, while the other factors are known and defined upfront (for instance, *size*, *value*, *investment*, etc.) – however, they are not directly observable. To estimate their realizations, stocks are sorted by their firm-level characteristics (for the size factor this is the market capitalization of the company). Long-short portfolios are formed which have a long position in the top fraction (often around 30%) and a short position in the bottom fraction of the stocks (for instance, the size portfolio is long the smallest 30% companies and short the largest 30%). The returns of these theoretical portfolios are then used as proxies for the factor realizations. Since the factor portfolios are zero-investment (the net sum of weights is 0), they are usually uncorrelated to the market portfolio, which is a convenient statistical property for the estimation of factor loadings. Another important property of these models is that the factors are required to exhibit statistically significant positive mean returns, which implies that they deliver a premium for the risk they exhibit – this property is mainly important for asset pricing studies, which focus on the expected returns of securities on the long run [35, 115]. A very well-known fundamental factor

model is the Fama-French 3 factor model[†]:

$$R_i - R_f = \alpha + \beta_i^{(M)}(R_m - R_f) + \beta_i^{(S)} SMB + \beta_i^{(V)} HML + e_i, \tag{3.2}$$

which includes the idiosyncratic component e_i, market factor R_m, size factor SMB and the value factor HML (obtained as the return of the long-short portfolio of assets sorted according to the book-to-price ratio)[116]. A similar model has also been proposed for the bond returns, with maturity and and default risks instead of size and value [4]. Lately, more elaborate and complex fundamental factor models have been proposed, with the addition of new factors such as momentum, profitability, investment, or others [6, 117, 118, 119, 120]. Recent years have also seen debates on whether the positive expected risk premia of certain factors still exist, or even if they existed in the first place [7].

Another fundamental factor approach is the BARRA risk model [121], which uses firm-level characteristics directly as factor loadings. For instance, the BARRA industry factor model will have a number of factors equal to the number of underlying industry sectors in the market, and the individual stocks will have factor loadings set to 1 only in those sectors with which the stocks are associated, and 0 otherwise. These factor loadings are then used to estimate factor realizations in a regression setting – somewhat inverse of what the Fama-French approach (where the factor realizations are estimated first, and then used to estimate factor loadings).

The fundamental factor approach to modelling financial returns has also gained a lot of attention in the financial industry, in the light of so-called *smart beta* strategies, which focus on harvesting the factor risk premia while diversifying other sources of risk [122]. In addition, this approach has also been applied to other markets, such as the cryptocurrency market [123]. However, the debates on the many proposed factors are still unresolved [124], with new statistical evidence suggesting that many of them are either redundant or statistically insignificant [125]. In addition, the information required for estimating fundamental factors is not always available for every security (i.e. commodities, derivatives or certain ETFs or indices may not have the fundamental firm-level characteristics similar to those of stocks), meaning that these standard approaches may not be universally applicable.

3.1.2 Latent factor models

Strict factor models

In latent factor models, the factors are both unknown and unobservable, i.e. latent. This means that the both the factor realizations and the exposures of securities to these factors need to be estimated from the data. Latent factor models can be thought of as an unsupervised dimension-

[†]For this contribution to financial economics, Eugene Fama shared the 2013 Nobel Memorial Prize in Economic Sciences.

ality reduction technique for uncovering a set of variables $f_{t1},...f_{tk}$ explaining the variance in the observed security returns $x_{t1},...,x_{tN}$ at time step t, where the dimension K is lower than N. The most common form of factor models is a linear factor model:

$$x_{ti} = \sum_{k=1}^{K} f_{tk} b_{ik} + e_{ti}, \tag{3.3}$$

where b_{ik} is the factor loading of time series i to factor k. The residual term e_{ti} is also known as the idiosyncratic component and represents the individual source of risk, uncorrelated with the common factor risk. Note that this formulation does not include an explicit intercept term (sometimes also called alpha), similar to the CAPM model, but unlike the Fama-French model. In fact, while some latent factor models include an intercept term, others explicitly omit it, and model the expected returns only through the expected factor returns [126, 127]. Moreover, some latent factor models are estimated by de-meaning the data first, thus leaving no means in the cross section of returns to be explained by the factors. The model (3.3) can also be stated in matrix form:

$$\underset{T\times N}{\boldsymbol{X}} = \underset{T\times K}{\boldsymbol{F}}\ \underset{N\times K}{\boldsymbol{B}^{\mathsf{T}}} + \underset{T\times N}{\boldsymbol{e}}, \tag{3.4}$$

where T is the length of the observed return time series. Latent factor models belonging to the class of *strict factor models* assume the following [93]:

(i) The factor covariance $\boldsymbol{Q}^{(F)} = \mathrm{Cov}(\mathbf{F})$ is positive-definite,

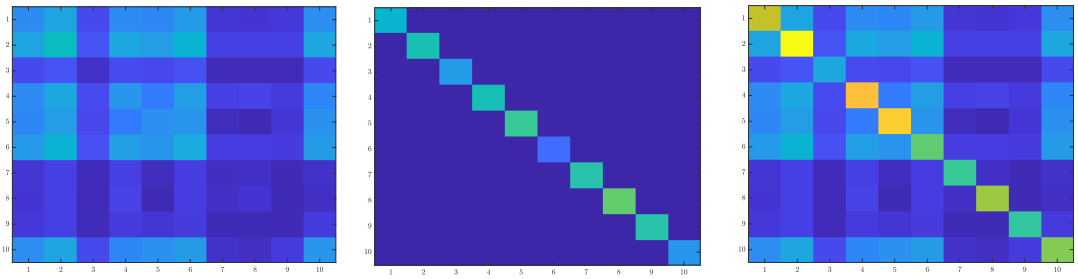
(ii) The idiosyncratic components $\boldsymbol{e} = [e_1, e_2, ..., e_N]$ are zero-mean and uncorrelated, meaning that the covariance matrix $\boldsymbol{\Psi} = \mathrm{Cov}(\boldsymbol{e})$ is a diagonal matrix with idiosyncratic risks on the diagonal,

(iii) $\mathrm{Cov}(\boldsymbol{F}, \boldsymbol{e}) = 0$.

The first assumption is necessary for the estimation of latent factor models – in case of (multi)colinearity in factors (i.e. a positive semi-definite factor covariance), the estimates of the loadings matrix would not be computationally stable. Although the orthogonality assumption on the idiosyncratic components does not necessarily mean independence, if their distribution were Gaussian, this would also imply independence. The final assumption is a consequence of the regression of $\boldsymbol{X}$ against $\boldsymbol{F}$ – in other words, no information which could be explained by $\boldsymbol{F}$ could be left over in the residuals $\boldsymbol{e}$. If the assumptions (i) and (ii) hold, the model-implied data covariance matrix can be expressed in the following form:

$$\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{Q} = \boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\mathsf{T}} + \boldsymbol{\Psi}. \tag{3.5}$$

The first component $\mathbf{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\mathsf{T}}$ is an $N \times N$ matrix of rank K (thus, a low rank component), while the idiosyncratic covariance $\boldsymbol{\Psi}$ is diagonal (and therefore of full rank), as stated in the assumptions. The resulting matrix $\boldsymbol{Q}$ is full rank and positive definite. This model is called strict



(a) $\boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\intercal}$ (b) $\boldsymbol{\Psi}$ (c) $\boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\intercal}+\boldsymbol{\Psi}$

Figure 3.1: The low-rank common component (a) and the idiosyncratic component (b) of the data covariance matrix (c), for a linear factor model.

because all cross-correlations in return time series are exclusively explained by the common factors, whereas the idiosyncratic components are uncorrelated [127]. This means that the variance of each time series is decomposed into two components: (i) the variance explained by common factors; (ii) idiosyncratic variance, as shown in Figure 3.1.

Moreover, whereas the common variance component $\boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\intercal}$ can be identified uniquely, the factor realizations and loadings can only be estimated up to an orthogonal rotation. Specifically, for any non-singular rotation matrix $\boldsymbol{H}$, the common variance component calculated from the rotated factors $\boldsymbol{F}\boldsymbol{H}$ is the same (since $\boldsymbol{H}^{\intercal}\boldsymbol{H}=\mathbf{I}$), but the factor realizations are evidently not. This property (also known as *rotational indeterminacy*) is simply stated as:

$$\boldsymbol{F}\boldsymbol{B}^{\intercal}=\tilde{\boldsymbol{F}}\tilde{\boldsymbol{B}}^{\intercal},\quad \tilde{\boldsymbol{F}}=\boldsymbol{F}\boldsymbol{H},\ \tilde{\boldsymbol{B}}=\boldsymbol{B}\boldsymbol{H}^{-1}. \tag{3.6}$$

Owing to this property, the factor covariance can often be set to: $\boldsymbol{Q}^{(F)}=\boldsymbol{I}_K$, but this may not necessarily be the case in all applications.

To estimate strict factor models, several approaches may be applied, the most common being principal axis, least squares, and maximum likelihood. The principal axis method focuses on the so-called *reduced correlation matrix* of the data [128]. Where the correlation matrix R has all diagonal elements equal to 1, each diagonal element of the reduced correlation matrix $R^{(RCM)}$ is equal to the percentage of the variance explained by the common factors. This matrix is initialized as:

$$r_{ij}^{(RCM)}=\begin{cases} r_{ij}, & i\neq j \\ 1-1/r'_{ii}, & i=j, \end{cases} \tag{3.7}$$

where $R'=R^{-1}$ is the inverse of the correlation matrix R. This initialization in used in order to remove the influence of the idiosyncratic components. The method then uses the eigendecomposition of the reduced correlation matrix to obtain the low-rank factor representation:

$$\boldsymbol{R}^{(RCM)}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\intercal}, \tag{3.8}$$

where the matrix $\boldsymbol{U}$ contains the eigenvectors of $\boldsymbol{R}^{(RCM)}$ and $\boldsymbol{D}$ is a diagonal matrix with the respective eigenvalues on the diagonal. The factor loadings matrix for the K factors is estimated as $\hat{\boldsymbol{B}} = \boldsymbol{U}^{(K)}\sqrt{\boldsymbol{D}^{(K)}}$, from the K eigenvectors corresponding to the largest eigenvalues. Although the reduced correlation matrix is not necessarily positive-semidefinite and all of its eigenvalues are not non-negative, the largest K eigenvalues should be positive if the factor model holds. Then the new reduced correlation matrix is calculated as $\boldsymbol{R}^{(RCM)} = \boldsymbol{U}^{(K)}\boldsymbol{D}^{(K)}\boldsymbol{U}^{(K)\intercal}$ and the process is repeated until convergence. The estimated factor loadings $\hat{\boldsymbol{B}}$ obtained this way can be used to calculate the factor realizations from the data:

$$\hat{\boldsymbol{F}} = \boldsymbol{X}\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^{\intercal}\hat{\boldsymbol{B}})^{-1}. \tag{3.9}$$

Since $K < N$, the $K \times K$ matrix $\hat{\boldsymbol{B}}^{\intercal}\hat{\boldsymbol{B}}$ is of full rank and its inverse exists.

The the maximum likelihood (ML) and ordinary least squares (OLS) approaches both focus on minimizing respective loss functions. First, without loss of generality, the factor realizations are assumed to have an identity covariance: an assumption is made on the factor realizations, namely: $\boldsymbol{Q}^{(F)} = \mathrm{Cov}(\boldsymbol{F}) = \boldsymbol{I}_K$. Owing to this, the covariance matrix of the data can be expressed as:

$$\boldsymbol{Q} = \boldsymbol{B}\boldsymbol{B}^{\intercal} + \boldsymbol{\Psi}. \tag{3.10}$$

The ML approach is most commonly used with a Gaussian i.i.d. (no autocorrelation in the time series) assumption:

$$f(X_t) = \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{Q}|}}\exp\left[-\frac{1}{2}(X_t - \boldsymbol{\mu})\boldsymbol{Q}^{-1}(X_t - \boldsymbol{\mu})\right]. \tag{3.11}$$

The ML estimates are obtained by maximizing the likelihood of the parameters in $\boldsymbol{B}$ and $\boldsymbol{\Psi}$ given the observed data $\boldsymbol{X}$. The reduced form of the log-likelihood loss function reads:

$$LL^{(ML)}(\boldsymbol{B}, \boldsymbol{\Psi}; \boldsymbol{X}) = -\frac{T}{2}\left[\ln(\boldsymbol{B}\boldsymbol{B}^{\intercal} + \boldsymbol{\Psi}) + \mathrm{tr}((\boldsymbol{B}\boldsymbol{B}^{\intercal} + \boldsymbol{\Psi})^{-1}\hat{\boldsymbol{Q}})\right], \tag{3.12}$$

where $\hat{\boldsymbol{Q}}$ is the sample covariance matrix. This log-likelihood function is usually maximized using iterative EM-type procedures [30].

On the other, hand, OLS estimates are obtained without the assumptions on the data distribution, but rather by minimizing the following loss function:

$$L^{(OLS)}(\boldsymbol{B}, \boldsymbol{\Psi}; \boldsymbol{X}) = \mathrm{tr}\left[(\boldsymbol{Q} - \boldsymbol{B}\boldsymbol{B}^{\intercal} - \boldsymbol{\Psi})^{\intercal}(\boldsymbol{Q} - \boldsymbol{B}\boldsymbol{B}^{\intercal} - \boldsymbol{\Psi})\right]. \tag{3.13}$$

This loss function amounts to the sum of all squared differences between the elements of the sample covariance matrix and the model-implied covariance matrix with a latent factor structure [129].

In certain cases, both OLS and ML are applied to the data covariance matrix, with the estimated factor loadings used on the non-scaled original data $\boldsymbol{X}$ to calculate the factor realizations, as in (3.9). It has been argued that the ML estimator will generally focus more on fitting larger correlations better, whereas the OLS estimator will attempt to fit all of the off-diagonal elements equally well [130]. This was also confirmed by the findings that the OLS estimator outperforms ML in recovering weak common factors (those that do not explain large amounts of variability in the data and their respective factor loadings are of a smaller magnitude) [130].

Approximate factor models

Strict factor models are a mathematically tractable and elegant way of explaining the commonalities in the data and decomposing the variance. However, the assumptions of the strict factor model may not hold in empirical data. Specifically, the assumption (ii) of uncorrelated residuals is very often violated, due to certain structural properties of the data which cannot be explained by common factors. For instance, in financial security return data, two securities may represent two classes of shares of the same company (for instance, BRK-A and BRK-B represent the A-class and B-class shares of the Berkshire Hathaway company). They are evidently connected through a commonality specific to them - this connection cannot be explained by common factors since none of the other assets are exposed to this specific factor, but the assumption of uncorrelated idiosyncratic components also does not hold. The reality is: they are both exposed to a very specific factor which only affect these two companies and emerges as a correlation of their residuals beyond the correlations explained by the common factors. These factors are sometimes labeled as *narrow*, since they affect few securities, as opposed to *broad* (pervasive) factors. A more suitable set of assumptions for modelling such realities comes in the form of *approximate factor models* [131].

An approximate factor model is still assumed to be a linear factor model of the form:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^{\mathsf{T}} + \boldsymbol{e}. \tag{3.14}$$

However, instead of the original strict factor model assumptions, the assumptions of the approximate factor model are somewhat relaxed. The following assumptions are made:

(i) **Factors and factor loadings:**
- Factors have finite variance and positive definite covariance: $Cov(\boldsymbol{F}) = \boldsymbol{Q}^{(F)} \succ 0$.
- Each factor has a non-trivial contribution to the variance of $\boldsymbol{X}$ – factors are *pervasive*, and $\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} \succ 0$.

(ii) **Idiosyncratic components:**
- (a) Zero-mean idiosyncratic components: $\mathrm{E}[e_{ti}] = 0$ with finite variance.
- (b) Weak temporal and cross-sectional dependence allowed:

- $\mathrm{E}\left[\frac{1}{N}\sum_i^N e_{t,i}e_{s,i}\right] = \gamma_N(s,t)$, $|\gamma_N(s,s)| \leq M, \forall s$ and $\frac{1}{T}\sum_s^T\sum_t^T |\gamma_N(s,t)| \leq M$
- $\mathrm{E}\left[e_{t,i}e_{t,j}\right] = \tau_{t,i,j}, \quad |\tau_{t,i,j}| \leq |\tau_{i,j}|, \quad \frac{1}{N}\sum_i^N\sum_j^N |\tau_{i,j}| \leq M$
- $\mathrm{E}\left[e_{t,i}e_{s,j}\right] = \tau_{t,s,i,j}, \quad \frac{1}{NT}\sum_i^N\sum_j^N\sum_t^T\sum_s^T |\tau_{t,s,i,j}| \leq M$
- $\mathrm{E}\left[\left|\frac{1}{\sqrt{N}}\sum_i^N (e_{t,i}e_{s,i} - \mathrm{E}[e_{t,i}e_{s,i}])\right|^4\right] \leq M, \quad \forall t,s$

(iii) **Weak dependence between factors and idiosyncratic components:**
$\mathrm{E}\left[\frac{1}{N}\sum_i^N ||\frac{1}{\sqrt{T}}\sum_t^T \boldsymbol{f}_t e_{t,i}||^2\right] \leq M$

This set of assumptions allows for cross-sectional dependence and heteroskedasticity of idiosyncratic components, which is much closer to the reality of financial time series. These off-diagonal elements of the idiosyncratic covariance matrix $\boldsymbol{\Psi}$ can not be a consequence of pervasive common factors, but rather sparse effects of very narrow (weak) factors. In terms of the data covariance matrix spectrum, the common factors emerge as diverging eigenvalues which grow with the growth of N. In approximate factor models, the eigenvalues belonging to the idiosyncratic part of the spectrum are allowed to grow, but will be bounded since they represent weak factors affecting only a limited number of time series, the number of which grows much slower than N. In addition, these conditions imply that idiosyncratic risks are diversifiable when considering a well-diversified portfolio $\boldsymbol{w} = [w_1, ..., w_N]$ of a large number of assets [127]:

$$\lim_{N\to\infty} \boldsymbol{w}^\intercal \boldsymbol{\Psi} \boldsymbol{w} = 0, \quad \lim_{N\to\infty} \boldsymbol{w}^\intercal \boldsymbol{w} = 0. \tag{3.15}$$

The well-diversified condition means that the portfolio must not be concentrated in a small number of securities, and is stated in the expression above as the condition: $\boldsymbol{w}^\intercal \boldsymbol{w} = 0$. Note that these results rely on the assumption that the number of assets N is sufficiently large - it is however fortunate that this assumption mostly holds when considering large datasets of globally traded financial securities.

To estimate approximate factor models, a principal components estimator[‡] is most commonly applied, which considers a squared Frobenius norm of the residuals as a loss function:

$$L^{(PC)} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (X_{it} - \hat{\boldsymbol{F}}_t\hat{\boldsymbol{B}}_i^\intercal)^2 = \frac{1}{NT}||\boldsymbol{X} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\intercal||_F^2. \tag{3.16}$$

Under the assumption that the factors are orthogonal $\frac{\boldsymbol{F}^\intercal\boldsymbol{F}}{T} = \boldsymbol{I}_K$, the problem becomes equivalent to the maximization of $\mathrm{tr}(\boldsymbol{F}^\intercal(\boldsymbol{X}\boldsymbol{X}^\intercal)\boldsymbol{F})$ under the above condition [20]. The solution to this problem, if the matrix $\boldsymbol{B}^\intercal\boldsymbol{B}$ is also diagonal, is given by the spectrum of the matrix $\frac{\boldsymbol{X}^\intercal\boldsymbol{X}}{T}$. The

[‡]Bai and Ng [20] define two estimators: PC (principal components) and APC (asymptotic principal components), but they are shown to be equivalent up to an orthogonal rotation and the difference is in computational characteristics.

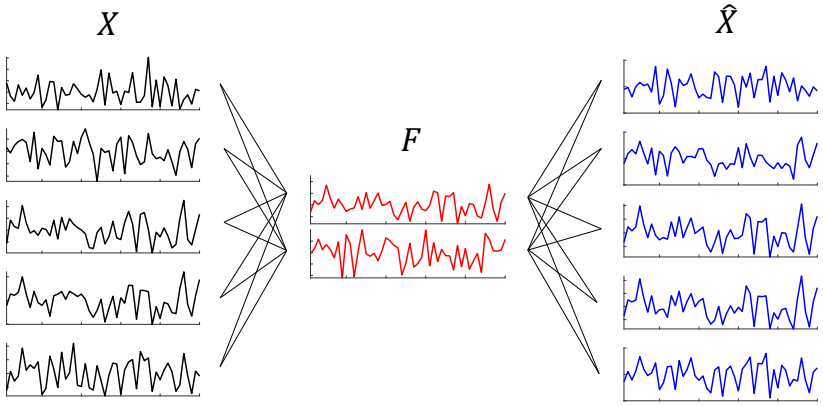


Figure 3.2: An autoencoder-type view of the latent factor model where the original higher-dimensional space of the data $\boldsymbol{X}$ is encoded into a lower-dimensional latent factor space $\hat{\boldsymbol{F}}$, from which the reconstruction $\hat{\boldsymbol{X}}$ can be obtained.

estimator for the factor loadings matrix and the factor realizations is:

$$\begin{aligned} \hat{\boldsymbol{B}} &= \boldsymbol{U}_K \sqrt{\boldsymbol{D}_K} \\ \hat{\boldsymbol{F}} &= \boldsymbol{X}\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^\intercal \hat{\boldsymbol{B}})^{-1} = \boldsymbol{X}\boldsymbol{U}_K \sqrt{\boldsymbol{D}_K}^{-1}, \end{aligned} \tag{3.17}$$

where $\boldsymbol{U}_K$ are the K eigenvectors corresponding to the largest K eigenvalues of the matrix $\frac{\boldsymbol{X}^\intercal \boldsymbol{X}}{T}$, which are also the diagonal elements of the diagonal matrix $\boldsymbol{D}_K$. Although there are many other formulations for the PC estimator [20, 132, 133, 134], in this thesis this specific formulation is chosen in order to have the factor loadings estimates depend on the in-sample covariance structure, and only the factor realizations estimates depend on the data realizations $\boldsymbol{X}$. By doing so, an autoencoder-type view on the latent factor model can be considered. The encoder function transforms the N-dimensional data space into a K-dimensional latent factor space $\hat{\boldsymbol{F}} = \boldsymbol{X}\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^\intercal \hat{\boldsymbol{B}})^{-1}$. The decoder function transforms the K-dimensional factor realizations into the original N-dimensional space: $\hat{\boldsymbol{X}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\intercal$. Such a setting is depicted in Figure 3.2.

The principal components estimator is known to be biased towards the time series with large individual variances – in an extreme case when a single individual variance is orders of magnitude larger than others, the estimates will be biased towards that time series, since it would reduce the loss function (3.16) most effectively. However, a fortunate circumstance in the high-dimensional case is the so-called *big data blessing*. Specifically, if the number of time series N is large, the effects of individual variances in the loss function become negligible in comparison to the sum of the other variances, and generally the cross-correlations in the data. Owing to this, both the estimation error and this bias are reduced with the increase in the number of times series N. To illustrate this, consider the mean squared error of the common factor component in the data:

$$MSE = \frac{1}{NT} ||\boldsymbol{F}\boldsymbol{B}^\intercal - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\intercal||_F^2. \tag{3.18}$$

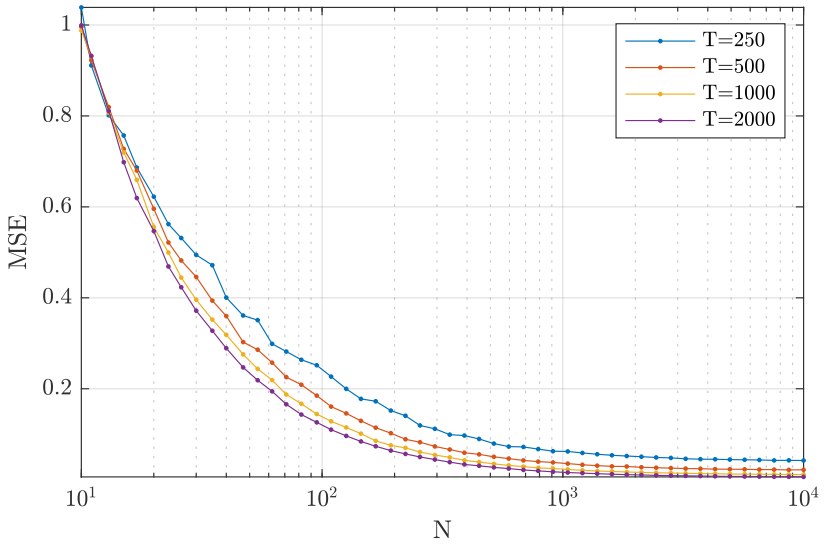


Figure 3.3: The *MSE* of the PC estimator with respect to the true factor realizations and loadings, for different values of the number of time series N and their length T. The x-axis is shown in logarithmic scale.

In a simulation setting, where the factor realizations and the model factor loadings are known, the PC estimator can be tested with respect to the dimensionality of the data. Figure 3.3 demonstrates the results for a simulation scenario where the factor loadings matrix and the factor realizations themselves are drawn from a standard normal distribution $\mathcal{N}(0,1)$. For a given time series length T the number of time series N is increased, and for each $N = 10, ..., 10000$ the PC estimator is applied and the MSE is measured.

These results demonstrate the so-called "big data blessing" exploited by the PC estimator [20]. Instead of deteriorating with increased dimensionality, the estimator effectively improves with respect to the true common component $\boldsymbol{FB}^\intercal$ as N grows, even beyond the length of the time series T. This is a very desirable property of the PC estimator, and the reason why it is used in so many high-dimensional latent factor estimation studies [20, 41, 132, 133, 135]. It is also important to note that the estimator will not be most efficient in low-dimensional settings, when the number of time series N is not large enough - the results in Figure 3.3 suggest that the minimum number of time series to reach these properties is between $N = 100$ and $N = 1000$. However, theoretical results suggest that this may deteriorate in cases with weak factors or excessive differences in individual time series variances [133]. In addition, some differences exist with respect to the time series length T – the cases with longer time series will, as expected, reduce the MSE quicker (in terms of rising N) than the shorter time series, as indicated by the results in Figure 3.3.

3.1.3 Estimation of the number of latent factors

The estimation of the number of latent factors or generally common components in the data has been an area of study which has increasingly gained attention with the rise of data science and data-driven decision making applications [136, 137, 138]. However, the problem of estimating the number of components is not new, and has been studied decades ago. Some of the oldest methods rely on statistical tools which are equally useful in modern estimators used today. Generally, the methods for estimation of the number of latent factors from data can be divided into three distinct approaches: (i) thresholding approaches based on the spectra of data correlation matrices, (ii) information criteria based on the estimated models, and (iii) statistical heuristic approaches based on different statistical properties of the data.

Thresholding approaches

The thresholding approaches focus on the spectral properties of the data correlation matrix, most commonly the eigenvalues of these matrices $\lambda_1, ..., \lambda_N$. The goal is to find a threshold value $\lambda^{(t)}$ for which a decision rule can be made:

$$\hat{K} = |\{i : \lambda_i > \lambda^{(t)}\}|. \tag{3.19}$$

This rule simply states that the estimated number of components $\hat{K}$ is equal to the number of eigenvalues of the data correlation matrix larger than the threshold value. One of the oldest such methods is the Kaiser-Guttman rule [139], which simply sets the threshold value to: $\lambda^{(t)} = 1$. This makes sense in the asymptotic case when the covariance estimates are reliable and close to the true covariance – in this case the eigenvalues of a diagonal correlation matrix will all be equal to 1 (see Figure 2.10). If the strict factor model holds, all the correlations in the data can be explained by the common components and thus the eigenvalues representing these components will stand out from the spectrum and will be larger than 1, while the other eigenvalues will be somewhat smaller than 1. However, in finite sample cases, as depicted in Figure 2.10, the spectrum will be somewhat "blurred" – meaning that some eigenvalues will be larger than 1 even if they do not represent common components in the data.

This is why a somewhat more advanced method is based on the Marčenko-Pastur distribution [98], described in Section 2.3.2. The elegant property of this distribution is that it defines a maximum possible empirical eigenvalue of the correlation matrix estimate, and this value can then be used as a threshold: $\lambda^{(t)} = \lambda_+ = (1 + \sqrt{\frac{N}{T}})^2$. The Marčenko-Pastur rule is an equally elegant method as the Kaiser-Guttman rule, but with better properties in limited sample sizes, and has been commonly used in factor modelling and clustering applications [42, 48]. However, it does not account for the fact that heavy-tailed data will also exhibit certain eigenvalues larger than those expected by the Marčenko-Pastur distribution [99, 140]. This means that even when

accounting for the finite sample sizes, the Marčenko-Pastur distribution will generally expect smaller eigenvalues than those which empirical correlation matrices of heavy-tailed data will exhibit.

A better way to incorporate the empirical data distribution for determining the right number of common components in the data are the permutation methods (the most common being Horn's parallel analysis) [141, 142]. In essence, these methods use the data matrix $\boldsymbol{X}$ and permute the points in each time series independently, so that the individual marginal distributions are kept unchanged, but the permuted matrix $\boldsymbol{X}'$ eliminates any correlations between the time series. By doing so, the spectrum of the permuted data correlation can be used to obtain eigenvalues $\lambda_i^{(t)}$ used as thresholds for each eigenvalue λ_i – these can be the mean or the 95% confidence bounds from the multiple permutations performed on the data. The number of components is determined by the number of empirical eigenvalues λ_i larger than their respective thresholds $\lambda_i^{(t)}$. In addition, instead of using permutations of empirical data, any sort of assumed data distribution can be used to generate simulations and obtain the thresholds for the correlation eigenvalues [143]. Although parallel analysis and permutation methods are gaining increased attention, they may overestimate the number of latent factors in case of approximate factor models – due to the fact that the correlations are completely eliminated when obtaining the threshold eigenvalues, all correlations are thus explained only by the common components. This, however, is not the case in approximate factor models and especially high-dimensional financial time series, where certain pairwise correlations or weak dependencies may not be explained by common components or latent factors.

Information criteria

Based on the estimated models, an *information criterion* can be defined as a measure of quality of the model, which can be used to determine the right number of model parameters. Such information criteria are often used as for feature selection in many machine learning applications. This approach has also been used for determining the number of factors in approximate factor models [137]. Most notably, Bai and Ng [144] propose several panel information criteria:

$$\begin{aligned}
PIC_1(k) &= \ln\left(\frac{1}{NT}||\boldsymbol{X}-\hat{\boldsymbol{F}}^{(k)}\hat{\boldsymbol{B}}^{(k)\intercal}||_F^2\right)+k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right),\\
PIC_2(k) &= \ln\left(\frac{1}{NT}||\boldsymbol{X}-\hat{\boldsymbol{F}}^{(k)}\hat{\boldsymbol{B}}^{(k)\intercal}||_F^2\right)+k\left(\frac{N+T}{NT}\right)\ln C_{NT}^2,\\
PIC_3(k) &= \ln\left(\frac{1}{NT}||\boldsymbol{X}-\hat{\boldsymbol{F}}^{(k)}\hat{\boldsymbol{B}}^{(k)\intercal}||_F^2\right)+k\left(\frac{\ln C_{NT}^2}{C_{NT}^2}\right),
\end{aligned} \tag{3.20}$$

where $\hat{\boldsymbol{F}}^{(k)}$ and $\hat{\boldsymbol{B}}^{(k)}$ are the estimated factor realizations and loadings for k factors. The constant C_{NT}^2 for the PC estimator is $C_{NT}^2 = \min\{N,T\}$. These information criteria all have an error

component and a penalty component, where the penalty components may differ – they are designed to reach their minimum when the number of factors k reaches the true underlying number of factors. These information criteria have been demonstrated that they work very well in high dimensions – however, they require some computational efforts since they depend on estimating the model for each candidate number of factors k. In certain cases this may become inefficient and some more practical solutions are needed. Moreover, the information criteria methods may produce unstable results on empirical data which do not always exhibit clear evidence on the number of latent factors, or the factors themselves may be weak.

Statistical heuristic approaches

Most commonly used approaches to model selection and determining the correct number of latent factors rely on the statistical properties of the multivariate data, and the spectra of the empirical correlation matrices. For instance, the scree plot and the elbow method are used to select a point at which additional factors or clusters do not increase the model performance as significantly and define this as the estimated number of factor [145, 146]. A somewhat more formalized version of this approach is the Onatski test [147], which can both estimate the number of factors and test it against the null hypothesis that $K = 0$:

$$\hat{K} = \max_k \frac{\lambda_k - \lambda_{k+1}}{\lambda_{k+1} - \lambda_{k+2}}, \tag{3.21}$$

where λ_i is the i-th largest eigenvalue of the data correlation matrix. This test takes the advantage of the fact that the first K eigenvalues in a latent factor model with K factors will explode with increasing N, but the $K+1$-st eigenvalue (and all subsequent eigenvalues, which represent the idiosyncratic part of the spectrum) will remain bounded. However, the Onatski test does require the difference of the idiosyncratic eigenvalues to converge to zero, which may not be the case when the limit between the systematic and idiosyncratic parts of the spectrum is not clear [148]. Nevertheless, the division of the spectrum into a systematic part (common factors) and idiosyncratic part has sparked other approaches relying on the empirical correlation eigenvalues [136]. A very straightforward method is the eigenvalue ratio test [149, 150]:

$$\hat{K} = \max_k \frac{\lambda_k}{\lambda_{k+1}}, \tag{3.22}$$

which simply finds the largest ratio between two successive eigenvalues as the limit between this common (systematic) part and the eigenvalues corresponding to the idiosyncratic part of the spectrum. Although it has been found to underestimate the number of factors in high-frequency data [148], this estimator generally produces reliable and stable results in simulations and empirical market data.

3.2 Latent factor model with pervasive and cluster-specific factors

Traditionally, latent factor models in finance assume that the factors are pervasive (they affect all assets) and thus can be found as common components in high-dimensional security return time series [20, 41]. On the other hand, some recent results suggest that assets indeed tend to form clusters and communities which can be observed in their dependence network structures (modelled either by correlation or other measures of connectedness) [42, 43]. Using a dataset of $N = 1123$ weekly stock returns from 2010 to 2020, a security graph can be constructed as a k nearest neighbors graph (in this case $k = 20$), as depicted in Figure 3.4. In this security graph, the clustering of stocks is evident – the asset classes which the ETFs represent are encoded in different colors, however, these clusters do not necessarily follow the asset classification. For instance, the large purple cluster of stocks in the bottom is the U.S. market, the teal, green, yellow, red and lightblue clusters around the central cluster are Japan, India, Brazil, Turkey, and China, respectively. The European countries are in the central (multicolored) cluster. This structure visibly affirms the existence of clusters in financial securities, either owing to their asset classes, countries of origin, or some other underlying factors to which the assets in a specific cluster are exposed. However, the mapping of financial securities to specific clusters, and their relationships cannot be always assumed from asset level information (such as as asset classes or countries), and in most cases need to be estimated from the market data.

The underlying clustering structures in financial return data have been increasingly gaining attention in different research approaches throughout the past decade. Assuming a strict hierarchical clustering structure, Tumminello et al. [44] form a hierarchical latent factor model and propose an estimation method based on the minimum spanning tree of the underlying assets. Clusters of assets are also known to emerge in stocks of single equity markets (for instance, clusters of stocks belonging to the same sectors) - Kakushadze et al. [45] consider clustering techniques for estimating these groups from the security return time series. Verma et al. [46] proposed a cluster-specific factor model for the log-volatility with the goal of studying the heteroskedastic properties of volatility in financial assets returns. Other clustering approaches were also shown to improve high-dimensional covariance matrix estimates, which ultimately reduces risk in optimized portfolios [17, 18, 47, 48]. However, while the evidence on the existence of asset clusters is compelling, certain latent factors may still be pervasive and affect all assets. For instance, in a global set of financial assets, pervasive global factors may affect all time series (such as the global macroeconomic and market shocks [49, 50]), and cluster-specific factor related to certain countries will affect only specific clusters of assets (for instance, European stocks will be affected by their own set of factors and may not be affected by some Asian market factors, after controlling for the common global component). The majority of modelling

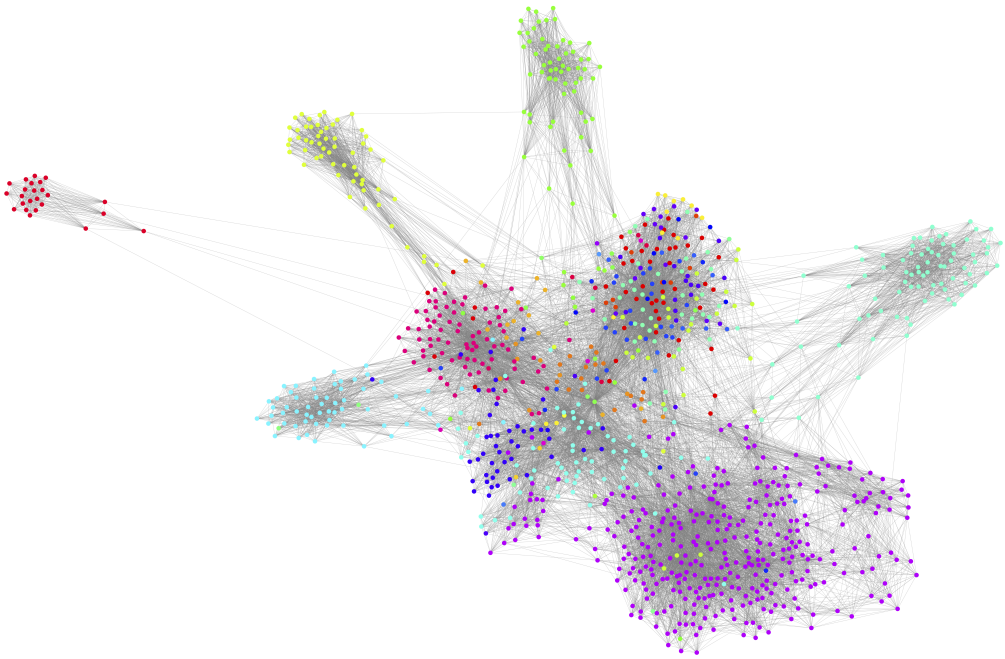


Figure 3.4: A k nearest neighbors graph obtained from absolute correlations between the weekly return time series of $N = 1123$ international stocks from 2010 to 2020, with $k = 20$. The countries of the stocks are encoded with colors (no legend is provided since there are 50 different countries).

approaches consider only pervasive latent factors, decomposing the security return variability into the variability explained by pervasive factors (affecting all assets) and idiosyncratic components (individual asset risk) [5, 41].

In this dissertation, a latent factor model with pervasive and cluster-specific factors is considered. The pervasive factors affect all security return time series, and these assets are divided into clusters in which a certain number of cluster-specific latent factors (the number of which may vary between clusters) affect the assets within that cluster. Since the clustering procedure may be biased towards the clusters with a larger number of cluster-specific factors (due to the fact that more factors will always be able to explain more variability in the data), the estimation algorithm is divided into two main phases: the clustering phase which uses a fixed number of cluster-specific factors for all clusters, and the latent factor estimation phase based on the estimated asset clusters. A computational approach to model selection is also proposed, which detects the number of pervasive factors, the number of clusters and the number of cluster-specific factors in each cluster.

3.2.1 Model definition

Let X_{ti} denote the return of asset i at time step t. Each asset i is associated with one of K clusters where $g_i \in \{1,...,K\}$ denotes the cluster index for asset i. A latent factor model is assumed in which security returns depend on the realizations of pervasive (common) factors f_{tp}

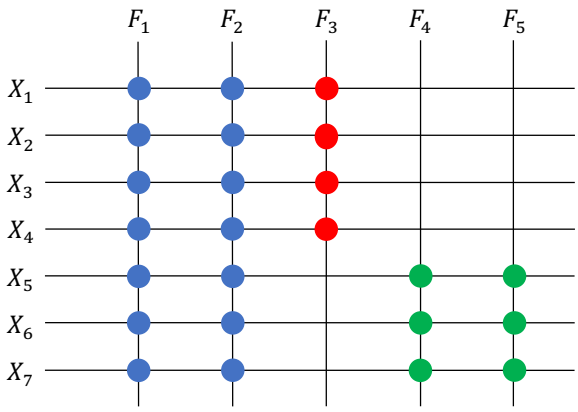


Figure 3.5: A grid view of a setting with time series $X_1,...,X_7$ affected by factors $F_1,...,F_5$, such that F_1 and F_2 are pervasive factors, F_3 is specific to time series $X_1,...,X_4$, while F_4 and F_5 are specific to time series $X_5,...,X_7$.

and cluster-specific factors ϕ_{tq}:

$$X_{ti} = \sum_{p=1}^{P} f_{tp} b_{ip} + \sum_{q=1}^{C_k} \phi_{tq} \lambda_{iq}^{(k)} + e_{ti}, \quad g_i = k, \tag{3.23}$$

where $t = 1,...,T$ is the time index, $i = 1,...,N$ is the asset index, $p = 1,...,P$ is the pervasive factor index, and $q = 1,...,C_k$ is the cluster-specific factor index for cluster $k = 1,...,K$. The pervasive factors are assumed to affect all individual time series, while the cluster-specific factors only affect assets which belong to the respective cluster. Each of the K clusters is allowed a different number of factors C_k - thus, the total number of cluster-specific factors is $Q = \sum_k C_k$. The residual term e_{ti} represents the idiosyncratic sources of risk, not explained by the common factors. Such a setting is shown in Figure 3.5, where the assets $X_1,...,X_7$ are exposed to pervasive factors F_1 and F_2 and only certain clusters of assets are exposed to the cluster specific factors F_3 (affecting cluster of assets $X_1,...,X_4$) and F_4, F_5 (affecting cluster of assets $X_5,...,X_7$).

The model (3.23) can also be written in matrix notation as:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^\intercal + \boldsymbol{\Phi}\boldsymbol{\Lambda}^\intercal + \boldsymbol{e}, \tag{3.24}$$

where $\boldsymbol{X} \in \mathbb{R}^{T \times N}$ contains N security return time series of length T, $\boldsymbol{F} \in \mathbb{R}^{T \times P}$ are the realizations and $\boldsymbol{B} \in \mathbb{R}^{N \times P}$ are the loadings for P pervasive factors. The realizations of Q cluster-specific factors for all K clusters are $\boldsymbol{\Phi} = \left[\boldsymbol{\Phi}^{(1)},...,\boldsymbol{\Phi}^{(K)}\right]$ and the cluster-specific factor loadings are $\boldsymbol{\Lambda} = \left[\boldsymbol{\Lambda}^{(1)},...,\boldsymbol{\Lambda}^{(K)}\right]$, where $\boldsymbol{\Phi}^{(k)} \in \mathbb{R}^{T \times C_k}$ and $\boldsymbol{\Lambda}^{(k)} \in \mathbb{R}^{N \times C_k}$ denote the C_k columns of $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$ corresponding to factor realizations and loadings associated with cluster k. The term $\boldsymbol{e} \in \mathbb{R}^{T \times N}$ contains all of the N individual idiosyncratic components.

Since the pervasive factors affect all time series, the pervasive factor loading matrix $\boldsymbol{B}$ is full, whereas the cluster-specific loading matrix $\boldsymbol{\Lambda}$ is non-zero only for the elements which

correspond to assets and factors associated with the same cluster:

$$\Lambda_i^{(k)} = 0, \quad g_i \neq k. \tag{3.25}$$

The assumed factor model is approximate, meaning that the idiosyncratic components $\boldsymbol{e}$ are zero-mean but are allowed cross-sectional correlations and heteroskedasticity. Other assumptions on the factors and factor loadings stand as in Section 3.1.2 – with some exceptions. Firstly, the pervasive factors and the cluster-specific factors are allowed to have some bounded correlations within themselves (i.e. the covariances $\boldsymbol{Q}^{(F)}$ and $\boldsymbol{Q}^{(\Phi)}$ are not necessarily diagonal). However, they cannot be correlated between each other (i.e. $\text{Cov}(\boldsymbol{f}_i, \boldsymbol{\lambda}_j) = 0$, $\forall i = 1, ..., P, j = 1, ..., Q$.). Moreover, for the purpose of covariance matrix estimation, the idiosyncratic components may have some correlations between themselves (the idiosyncratic covariance $\boldsymbol{\Psi}$ may contain some off-diagonal elements), but they are assumed orthogonal to the pervasive and cluster-specific factors.

The factors are latent (unobservable), the clustering is unknown, as well as the numbers of factors, clusters, and cluster-specific factors - all of these need to be estimated from the data. Given the model (3.24) and the assumptions, in the following sections new methods for factor estimation and model selection are presented. First an iterative method clusters the data assuming a fixed number of cluster-specific factors in each cluster. Then the numbers of cluster-specific factors inferred from the data using the estimated clusters. To estimate the number of pervasive factors and clusters, we propose a model selection method based on the spectral propertied of the asset correlation matrix and the security graph estimated from the return time series.

3.2.2 Estimation procedure

Let $||\mathbf{A}||_F^2 = \sum_i \sum_j A_{ij}^2$ denote the Frobenius norm of a matrix $\mathbf{A}$. Given a data matrix $\mathbf{X}$, and assuming a known number of pervasive factors P, number of clusters K and number of cluster-specific factors in each cluster C_k, consider the following loss function:

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{F}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{\Lambda}) = \frac{1}{NT} ||\boldsymbol{X} - \boldsymbol{F}\boldsymbol{B}^\intercal - \boldsymbol{\Phi}\boldsymbol{\Lambda}^\intercal||_F^2. \tag{3.26}$$

The loss function is the error of unexplained variation in the data. If all factors are pervasive the optimal low-rank approximation is given by the principal components (PC) estimator [5, 20, 145], based on the eigenvalue decomposition of the matrix $\frac{1}{T}\boldsymbol{X}^\intercal\boldsymbol{X}$, as described in Section 3.1.2 and expression 3.17. The PC estimator would still be able to estimate the factor loadings matrix $[\boldsymbol{B}, \boldsymbol{\Phi}]$ up to an orthogonal rotation in the case of enough data points (long time series). However, since it would effectively estimate $(P+Q) \times N$ factor loadings instead of

$P \times N + \sum_k^K C_k N_k$, the estimated model would contain much more parameters than the assumed underlying model, thus introducing more estimation error and unnecessary complexity. A better analytical estimation for the assumed loss function is not obtainable, since the loss function (3.26) needs to be optimized subject to the cluster-specific factor condition (3.25), given the clustering $G = [g_1, ..., g_N]$. The estimates of the pervasive factors, cluster memberships, and cluster-specific factors all depend on each other, and thus require an iterative approach [151] - in which the PC estimator will prove useful.

Cluster assignment

If the pervasive factors $\boldsymbol{F}$ with loadings $\boldsymbol{B}$ and cluster specific factors $\boldsymbol{\Phi}$ are known, each asset can be assigned to the cluster which minimizes its value of the loss function (3.26). To do so, we define $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{F}\boldsymbol{B}^\intercal$ and find the candidate cluster-specific loadings for cluster k as:

$$\tilde{\boldsymbol{\Lambda}}^{(k)} = \boldsymbol{Y}^\intercal \boldsymbol{\Phi}^{(k)} (\boldsymbol{\Phi}^{(k)\intercal} \boldsymbol{\Phi}^{(k)})^{-1}, \tag{3.27}$$

where $\boldsymbol{\Phi}^{(k)}$ are the cluster-specific factor realizations for cluster k, as defined previously. Using the estimates we calculate the loss matrix $\mathcal{L}_{ik} = \mathcal{L}(X_i; \boldsymbol{F}, \boldsymbol{B}, \boldsymbol{\Phi}^{(k)}, \tilde{\boldsymbol{\Lambda}}^{(k)})$ for each combination of assets $i = 1, ..., N$ and clusters $k = 1, ..., K$. The clusters are then directly assigned as:

$$\hat{g}_i = \underset{k}{\operatorname{argmin}}\, \mathcal{L}_{ik}, \tag{3.28}$$

meaning that each asset belongs to the cluster whose factors minimize the loss function (3.26) for that asset. This step can also be interpreted as a generalization of the assignment step in Lloyd's algorithm for k-means clustering, with C_k cluster-specific factors instead of centroids, and the loss function (3.26) instead of the Euclidean distance.

Estimation of cluster-specific factors

For a given clustering $\boldsymbol{g} = [g_1, ..., g_N]$ and known pervasive factors $\boldsymbol{F}$ with loadings $\boldsymbol{B}$, all assets within cluster k are exposed to the cluster-specific factors $\boldsymbol{\Phi}^{(k)}$ – for that subset of assets, these factors can be considered pervasive. This enables the estimation of the factors using the subset of security return time series $\boldsymbol{Y}^{(k)} \in \mathbb{R}^{T \times N_k}$ containing only the N_k time series in cluster k. Following the logic in (3.17), the factor loadings $\hat{\boldsymbol{\Lambda}}^{(k)}$ for cluster k are then estimated from the eigenvectors of the largest C_k eigenvalues of the $N_k \times N_k$ matrix $\frac{1}{T}\boldsymbol{Y}^{(k)\intercal}\boldsymbol{Y}^{(k)} = \boldsymbol{U}^{(Y)}\boldsymbol{D}^{(Y)}\boldsymbol{U}^{(Y)\intercal}$:

$$\begin{aligned} \hat{\boldsymbol{\Lambda}}^{(k)} &= \boldsymbol{U}^{(Y)}_{C_k} \sqrt{\boldsymbol{D}^{(Y)}_{C_k}} \\ \hat{\boldsymbol{\Phi}}^{(k)} &= \boldsymbol{Y}^{(k)} \hat{\boldsymbol{\Lambda}}^{(k)} (\hat{\boldsymbol{\Lambda}}^{(k)\intercal} \hat{\boldsymbol{\Lambda}}^{(k)})^{-1}. \end{aligned} \tag{3.29}$$

This approach simply applies the PC estimator to the time series of securities in cluster k, and repeats this for estimating the cluster-specific factor realizations and loadings for all clusters $k = 1, .., K$.

Estimation of pervasive factors

Given the clustering $\boldsymbol{g}$ and cluster-specific factors $\boldsymbol{\Phi}$ with loadings $\boldsymbol{\Lambda}$, define the residual term from the cluster-specific factors: $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{\Phi}\boldsymbol{\Lambda}^\intercal$. Given the residuals, the pervasive factor loadings $\hat{\boldsymbol{B}}$ are estimated from the eigenvectors of the largest P eigenvalues of $\frac{1}{T}\boldsymbol{Z}^\intercal\boldsymbol{Z} = \boldsymbol{U}^{(Z)}\boldsymbol{D}^{(Z)}\boldsymbol{U}^{(Z)\intercal}$:

$$\begin{aligned} \hat{\boldsymbol{B}} &= \boldsymbol{U}_P^{(Z)}\sqrt{\boldsymbol{D}_P^{(Z)}} \\ \hat{\boldsymbol{F}} &= \boldsymbol{Z}\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^\intercal\hat{\boldsymbol{B}})^{-1}. \end{aligned} \tag{3.30}$$

Since the cluster-specific common components in the data are accounted for and eliminated in the term $\boldsymbol{Z}$, it only contains the P pervasive factors which are in this way estimated by the PC estimator.

The three steps given above are iterated over, but the entire estimation procedure also depends on the model selection – the estimation of the number of pervasive factors, number of clusters and the number of cluster-specific factors in each cluster. The following section describes the model selection approach and after that an overview of the entire procedure is given.

3.2.3 Model selection

Estimating the number of pervasive factors and clusters

To estimate the number of pervasive factors P and the number of clusters K from the data, the principles of the eigenvalue ratio (ER) test [150] are applied, and a method for estimating the number of clusters using a graph (network) of assets is also proposed. Since the estimates depend on each other, P and K are estimated choosing from several candidate pairs $\tilde{P}, \tilde{K}$, based on a joint criterion.

The ER approach sorts the eigenvalues of the data correlation matrix in a descending order and defines the eigenvalue ratio:

$$\eta_i^{(p)} = \xi_i / \xi_{i+1}, \tag{3.31}$$

where ξ_i is the i-th largest eigenvalue. In the case of the assumed model with pervasive and cluster-specific factors, the ER test will detect the shift in the eigenvalues between the pervasive factor part and the cluster-specific factor part (since the cluster-specific factors affect less assets, the eigenvalues corresponding to them will be lower than those representing pervasive factors). This is displayed in Figure 3.6, for a simulation example of the assumed model.

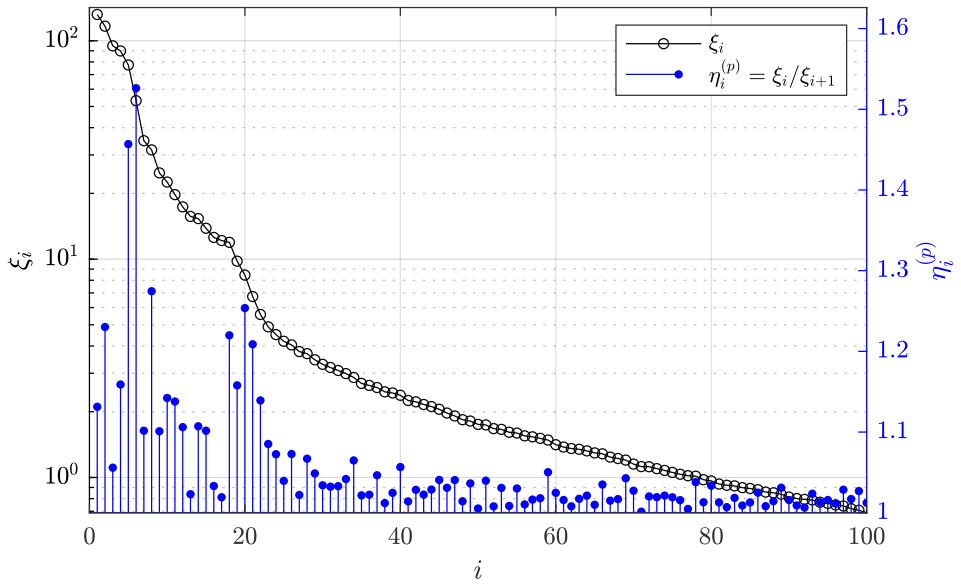


Figure 3.6: The first 100 eigenvalues and eigenvalue ratios of a sample correlation matrix. The best candidates for P in this case are 5 and 6, as seen in the eigenvalue ratios.

The larger the ER ratio $\eta_i^{(p)}$, the more evidence in favor of i being the correct number of pervasive factors. In the original ER test, this means that the simple rule for the estimated number of factors is $\hat{P} = \text{argmax}_i \eta_i^{(p)}$. However, since the ER test may not always detect the exact shift between the pervasive and cluster-specific factors (due to certain pervasive factors being weak, or cluster-specific factors being strong within their cluster), a more robust approach is taken. Instead of just picking the maximum value of ER, to avoid discarding potentially better solutions, a several candidates are considered for the the number of pervasive factors $\tilde{P}_1, ..., \tilde{P}_n$, corresponding to the n largest ratios $\tilde{\eta}_1^{(p)}, ..., \tilde{\eta}_n^{(p)}$. The selection of the exact $\hat{P}$ from these candidates depends on the clusters, as described in the following paragraph.

To detect the clusters of data, for each $\tilde{P}_i$ a security graph is formed from the time series $\boldsymbol{Y} = \boldsymbol{X} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\intercal$, where $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{B}}$ are estimated from the data using the PC estimator. Each of the N nodes in the graph represents a security and the edges between them depend on a similarity measure. In this case, since the securities themselves may be negatively correlated through exposure to the same factors but with different signed loadings, a similarity measure based on absolute correlations is considered:

$$w_{ij} = |\rho(Y_i, Y_j)| \tag{3.32}$$

The goal is to construct a security graph, which is represented by a sparse matrix (instead of the full matrix $\boldsymbol{W}$), and thus transform the dependency structure of the data into a new space in which the clusters may be more detectable. These techniques are in the core of spectral clustering approaches, which rely on the spectra of similarity graphs of the original data, from

which the clusters can be estimated [152, 153].

In order to obtain accurate and robust estimates, the security graph needs to reflect the following properties:

(i) time series pairs which are very close (having a high w_{ij}) should be connected,

(ii) time series in the same cluster should have a short path between them (high connectivity clusters),

(iii) the spectral properties of the graph need to be stable, since the estimation depends on the Laplacian spectrum.

The first property is found in the ε-neighborhood (εN) graph, constructed simply by keeping only the edges $w_{ij} > \varepsilon$ which are above a certain threshold ε. The adjacency matrix for the ε-N graph is simply calculated as:

$$w_{ij}^{(\varepsilon N)} = \begin{cases} w_{ij}, & w_{ij} \geq \varepsilon \\ 0, & w_{ij} < \varepsilon. \end{cases} \tag{3.33}$$

The properties of the graph depend on the threshold ε – for a very small ε the graph will be very dense and not much information will be extractable, for a larger ε the graph will have multiple components of densely connected time series. An example of the εN graph is displayed in Figure 3.7 – the emergence of 5 clusters is visible, but many nodes remain unconnected. However, increasing ε to obtain a connected graph is not an effective strategy since the graph would become too dense to extract meaningful information.

To reduce the dependence on the choice of ε and to achieve the second property (reduce the path length between securities in the same cluster), the k-nearest neighbors (kNN) graph is also considered. The graph is constructed by keeping the k edges with highest values of w_{ij} for each node i – essentially each node i is only connected to the k nodes in its neighborhood ω_i:

$$w_{ij}^{(kNN)} = \begin{cases} w_{ij}, j \in \omega_i \\ 0, \text{otherwise.} \end{cases} \tag{3.34}$$

Evidently, the neighborhoods of different nodes will occasionally overlap – it is in these cases that clusters of securities will be visible in the kNN graph. Moreover, even if the securities from the same clusters may not be connected in the ε-N graph, they will most likely be either connected or have mutual connections in the kNN graph. An example of the kNN graph is shown in Figure 3.8 – here the cluster structure is visible (and coincidentally, the graph is connected, which may not always be the case), but the division between clusters is not very good.

Finally, since the εN and kNN graphs may not always be connected graphs (they may contain multiple connected components), their spectral properties may differ depending on the

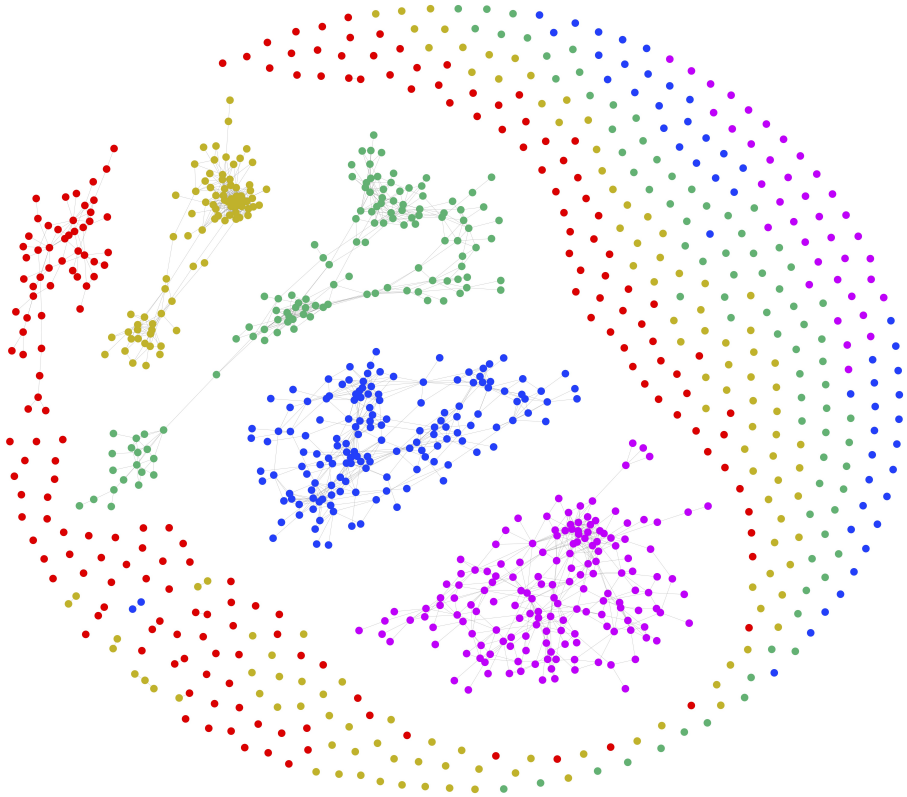


Figure 3.7: An example of the ε-neighborhood graph containing $N = 1000$ nodes for a simulation with $K = 5$ clusters (encoded in different colors) and $\varepsilon = 0.6$.

number of connected components – unfortunately, these connected components may not always correspond to the underlying clusters in the data. In order to ensure that the security graph is always connected, we also consider the *maximum spanning tree* (MST) graph, which always consists of one connected component. To obtain the maximum spanning tree, the edges w_{ij} are multiplied by -1 and Kruskal's algorithm for minimum spanning tree construction is applied. The MST is directly associated with the single-linkage hierarchical clustering method [44], and retains the cluster structures in the data. Therefore, in addition to ensuring that the graph is connected, the MST also adds to the cluster visibility in the estimated security graph. An example of the MST graph is shown in Figure 3.9, where the cluster structures are evidently visible, but the division between clusters is not perfect.

The final security graph is a combination of the three approaches: the MST provides a backbone to the network, ensuring that the graph is connected and reflecting the basic clustering structures in the data; the kNN graph joins together communities of securities, reducing the average path length between pairs nodes which should belong to the same clusters; the εN graph can be thought of a strengthening addition to the communities of securities which belong to the same cluster. The adjacency matrix for the security graph is given as the union of the

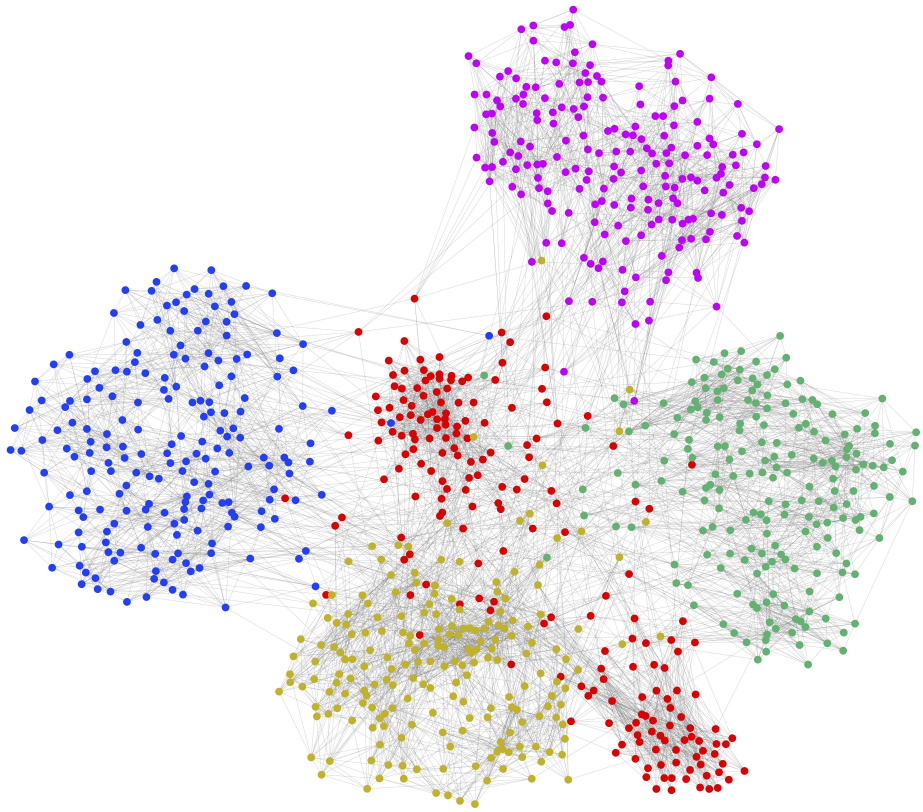


Figure 3.8: An example of the k nearest neighbors graph containing $N = 1000$ nodes for a simulation with $K = 5$ clusters (encoded in different colors) and $k = 10$.

three graphs:

$$\boldsymbol{W} = \boldsymbol{W}^{(\varepsilon N)} \cup \boldsymbol{W}^{(kNN)} \cup \boldsymbol{W}^{(MST)}, \tag{3.35}$$

with $\boldsymbol{W}^{(\varepsilon N)}$, $\boldsymbol{W}^{(kNN)}$, and $\boldsymbol{W}^{(MST)}$ being the adjacency matrices of the εN, kNN and MST graphs. Each element $\boldsymbol{W}_{ij}$ in the union of two security graph adjacency matrices $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$ is defined as: $\boldsymbol{W}_{ij} = \max\{\boldsymbol{W}_{ij}^{(1)}, \boldsymbol{W}_{ij}^{(2)}\}$. The security graph has favorable properties from all three methods combined, resulting in a structure as shown in Figure 3.10. The clusters are clearly visible, and the division between them is fairly clear – the proposed graph evidently has the best of all three components used to build it.

Given the graph adjacency matrix $\boldsymbol{W}$, the number of clusters can be estimated based on the spectral properties of the security graph Laplacian:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}, \tag{3.36}$$

where $\boldsymbol{D}$ is the diagonal node degree matrix $\boldsymbol{D} = \text{diag}(d_1, ..., d_N)$. The number of zero valued eigenvalues in the spectrum of the Laplacian matrix is equal to the number of connected components in the graph. Since the proposed graph $\boldsymbol{W}$ contains the MST, it will always have one connected component – thus the Laplacian $\boldsymbol{L}$ will have exactly one eigenvalue equal to zero.

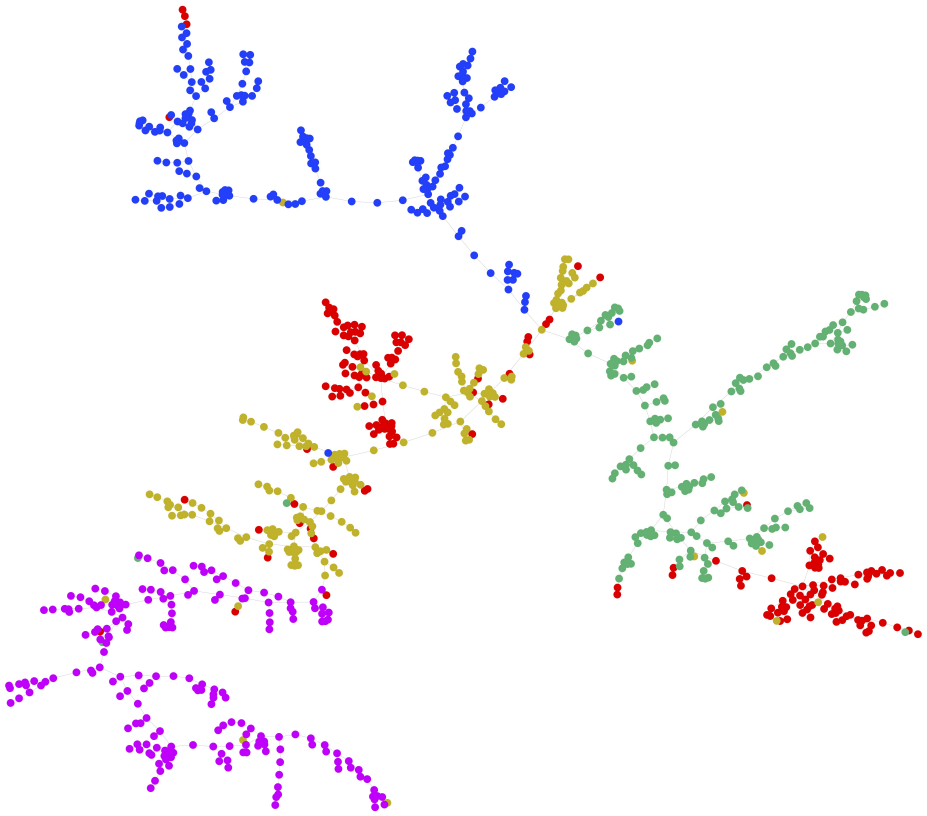


Figure 3.9: An example of the maximum spanning tree graph containing $N = 1000$ nodes for a simulation with $K = 5$ clusters (encoded in different colors).

The $K-1$ eigenvalues $\xi_2^{(L)}, ..., \xi_K^{(L)}$ will be close to zero for a graph containing K clusters (the end case being a graph divided into K connected components which will have exactly K eigenvalues equal to zero)[§]. To find the number of clusters in the graph, eigenvalues $\xi_i^{(L)}$ of the Laplacian are sorted in an ascending order and, in analogy with (3.31), define the the Laplacian eigenvalue ratio (LER):

$$\eta_i^{(c)} = \xi_{i+1}^{(L)} / \xi_i^{(L)}, \tag{3.37}$$

as the ratio between two subsequent graph Laplacian eigenvalues. Similar to the ER test, the number of clusters is estimated as the i which maximizes the LER: $\hat{K} = \text{argmax}_i \eta_i^{(L)}$. An example of the Laplacian eigenvalue ratios for the given graph above is displayed in Figure 3.11.

Combining the two approaches, for each $\tilde{P}_i$, a number n of $\tilde{K}_i$ which have the largest LER are considered. Between the candidate numbers of pervasive factors and clusters the selected

[§]In spectral graph theory, the second smallest eigenvalue of the Laplacian (also called the Fiedler eigenvalue), and the corresponding eigenvector (also called the algebraic connectivity) are in the focus of research on the bisection of graphs – here instead of bisecting graphs into two components, we consider dividing graphs into a number of clusters, and thus consider the $K-1$ smallest non-zero eigenvalues.

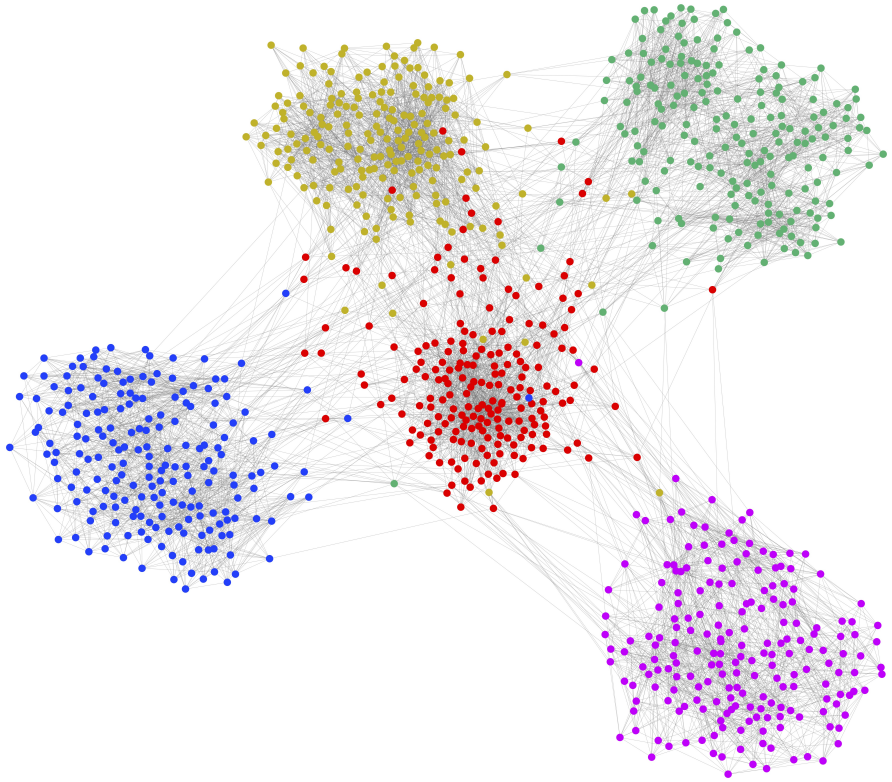


Figure 3.10: An example of the security graph containing $N = 1000$ nodes, estimated from a sample with $K = 5$ clusters – all of which are clearly visible in the graph structure.

values are $\hat{P} = \tilde{P}_j$ and $\hat{K} = \tilde{K}_j$, which have the smallest value of an information criterion:

$$\begin{aligned}\mathcal{I}(P,K) = &\frac{1}{NT}||\boldsymbol{X} - \hat{\boldsymbol{F}}^{(P)}\hat{\boldsymbol{B}}^{(P)\intercal} - \hat{\boldsymbol{\Phi}}^{(K)}\hat{\boldsymbol{\Lambda}}^{(K)\intercal}||_F^2 + \\ &Ps^2\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right) + \\ &\sum_{k=1}^{K} s^2\left(\frac{N_k+T}{N_kT}\right)\ln\left(\frac{N_kT}{N_k+T}\right) + \\ &\sum_{q=1}^{Q} s^2\left(\frac{N_q+T}{N_qT}\right)\ln\left(\frac{N_qT}{N_q+T}\right),\end{aligned} \tag{3.38}$$

where N_k denotes the number of securities in cluster k, N_q the number of securities exposed to cluster-specific factor q, and s^2 is a consistent estimate of $\frac{1}{NT}\sum\sum \mathrm{E}[e_{it}^2]$ given by fitting the model with the largest considered $\tilde{P}_j$ and $\tilde{K}_j$. The factor loadings and realizations are obtained by estimating the model with the given combination of P and K. The proposed criterion is an extension of the Bai-Ng criteria [144] which incorporates the error term (first component) and the penalty components, which penalize the number of pervasive factors, the number of clusters, and the number of cluster-specific factors, respectively. The term s^2 provides the proper

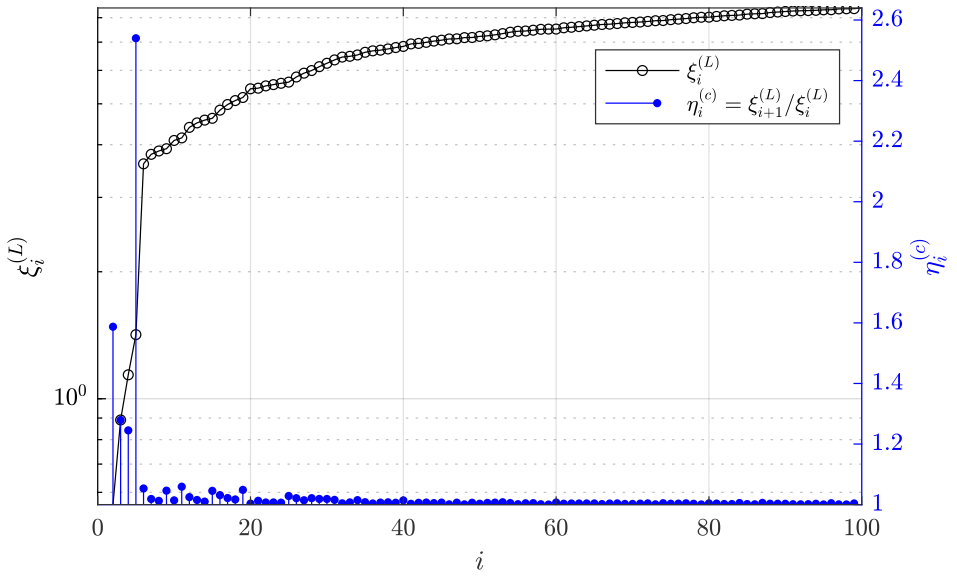


Figure 3.11: The first 100 eigenvalues and Laplacian eigenvalue ratios (LER) of the Laplacian matrix of a sample security graph. The first eigenvalue and LER are omitted since the first eigenvalue is zero (the graph has one connected component). The best candidate for K in this case is 5, as seen by the LER.

scaling for the penalty components, given the error, and the penalty components grow with increasing numbers of factors P, clusters K and the total number of cluster-specific factors Q. Moreover, the information criterion $\mathcal{I}$ is similar to the Ando-Bai information criteria for panel data models with grouped structures [51, 52] – however, the model considered here does not include observable factors and the clusters are penalized separately from the cluster-specific factors (i.e. between two models with the same number of cluster-specific factors affecting the same number of securities, the one with less clusters is preferred). The overview of the proposed model selection algorithm is given in Algorithm 1.

Algorithm 1: Model selection procedure

estimate candidates $\tilde{P}$ and $\tilde{\eta}^{(p)}$ from $\boldsymbol{X}$
foreach $\tilde{P}_i$ **do**
  estimate $\tilde{P}_i$ factors $\hat{\boldsymbol{F}}$ and loadings $\hat{\boldsymbol{B}}$ from $\boldsymbol{X}$
  construct the security graph from $\boldsymbol{Y} = \boldsymbol{X} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^{\mathsf{T}}$
  estimate candidates $\tilde{K}$ given $\tilde{\eta}^{(c)}$ from the graph Laplacian
  foreach $\tilde{K}_j$ **do**
    estimate latent factor model given$\tilde{P}_i$ and $\tilde{K}_j$
    calculate $\mathcal{I}(\tilde{P}_i, \tilde{K}_j)$ given the data $\boldsymbol{X}$
  end
end
select $\hat{P} = P_i$ and $\hat{K} = K_j$ which minimize $\mathcal{I}(P_i, K_j)$

Estimating the number of cluster-specific factors

During the cluster assignment step, the clusters with a larger number of cluster-specific factors C_k will naturally attract more assets (since the time series in clusters with more cluster-specific factors will tend to have a lower value of $\mathcal{L}_{ik}$), and the cluster membership estimates will be biased towards them. Even knowing the right number of cluster-specific factors in each cluster will not guarantee that the assets will be associated with the correct clusters. The proposed algorithm resolves this issue by having the number of clusters equal for all clusters $C_k = C_0, \forall k$ during the entire iterative clustering procedure. Given the estimated clustering $\hat{\boldsymbol{g}}$, the N_k time series $\boldsymbol{Y}^{(k)} = \boldsymbol{X}^{(k)} - \boldsymbol{F}\boldsymbol{B}^{(k)\mathsf{T}}$ will have a pure factor structure, containing C_k factors, and C_k can be estimated using the ER estimator. After C_k is estimated for each cluster, another phase of the iterative procedure is run, containing only the update step for the cluster-specific factor estimates and the pervasive factor estimates. An overview of the entire procedure, including clustering, factor estimation and the estimation of the number of cluster-specific factors is given in Algorithm 2.

Algorithm 2: Clustering and estimation of pervasive and cluster-specific factors

initialize $\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Lambda}}, \hat{\mathbf{g}}$
set $C_k = C_0$ for all clusters $k = 1, ..., K$
while *clustering convergence criteria not met* **do**
  update cluster membership:
    given $\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Phi}}$, estimate $\tilde{\boldsymbol{\Lambda}}$ from $\boldsymbol{Y} = \boldsymbol{X} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^{\mathsf{T}}$
    calculate $L_{ik} = l(X_i; \hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Phi}}^{(k)}, \tilde{\boldsymbol{\Lambda}}^{(k)})$
    set $\hat{g}_i \leftarrow \underset{k}{\operatorname{argmin}}\, L_{ik}$
  update cluster-specific factors:
    given $\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{g}}$, calculate $\boldsymbol{Y}^{(k)} = \boldsymbol{X}^{(k)} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^{(k)\mathsf{T}}$
    estimate $\hat{\boldsymbol{\Phi}}^{(k)}, \hat{\boldsymbol{\Lambda}}^{(k)}$ for all clusters $k = 1, ..., K$
    set $\hat{\boldsymbol{\Phi}} \leftarrow [\hat{\boldsymbol{\Phi}}^{(1)}, ..., \hat{\boldsymbol{\Phi}}^{(K)}]$, $\hat{\boldsymbol{\Lambda}} \leftarrow [\hat{\boldsymbol{\Lambda}}^{(1)}, ..., \hat{\boldsymbol{\Lambda}}^{(K)}]$
  update pervasive factors:
    given $\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Lambda}}$, calculate $\mathbf{Z} = \boldsymbol{X} - \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}^{\mathsf{T}}$
    estimate and set $\hat{\mathbf{F}}, \hat{\boldsymbol{B}}$ from $\mathbf{Z}$
end
given $\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{g}}$ update C_k for all clusters $k = 1, ..., K$
while *error convergence criteria not met* **do**
  update cluster-specific factors:
    given $\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{g}}$, calculate $\boldsymbol{Y}^{(k)} = \boldsymbol{X}^{(k)} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^{(k)\mathsf{T}}$
    estimate $\hat{\boldsymbol{\Phi}}^{(k)}, \hat{\boldsymbol{\Lambda}}^{(k)}$ for all clusters $k = 1, ..., K$
    set $\hat{\boldsymbol{\Phi}} \leftarrow [\hat{\boldsymbol{\Phi}}^{(1)}, ..., \hat{\boldsymbol{\Phi}}^{(K)}]$, $\hat{\boldsymbol{\Lambda}} \leftarrow [\hat{\boldsymbol{\Lambda}}^{(1)}, ..., \hat{\boldsymbol{\Lambda}}^{(K)}]$
  update pervasive factors:
    given $\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Lambda}}$, calculate $\boldsymbol{Z} = \mathbf{X} - \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}^{\mathsf{T}}$
    estimate and set $\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}$ from $\boldsymbol{Z}$
end

3.2.4 Intialization and hyperparameter selection

For the initialization, the P pervasive factors $\boldsymbol{F}$ and loadings $\boldsymbol{B}$ are estimated from the data $\boldsymbol{X}$ first. Following the initial pervasive factor estimate, the security graph is constructed from $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{F}\boldsymbol{B}^\intercal$. Using the security graph, a spectral clustering method is used to obtain the initial clustering. Specifically, the spectral clustering method uses the Laplacian matrix of the graph $\boldsymbol{L}$ and its eigendecomposition: $\boldsymbol{L} = \boldsymbol{U}^{(L)}\boldsymbol{D}^{(L)}\boldsymbol{U}^{(L)\intercal}$. The K eigenvectors corresponding to the smallest non-zero eigenvalues of $\boldsymbol{L}$ (thus, not counting the first one, which is zero), are then used as a new space in which the k-means algorithm can be applied in order to obtain the K clusters. For the given clustering $\boldsymbol{g}$ and pervasive factors $\boldsymbol{F}$ with loadings $\boldsymbol{B}$, the cluster-specific factors can be estimated using the data $\boldsymbol{Y}^{(k)}$, for each cluster $k = 1, ..., K$. In both phases (the clustering and the cluster-specific factor estimation), the algorithm stops when there are no cluster changes and the reduction in the loss function $\mathcal{L}^{(i)} - \mathcal{L}^{(i-1)}$ is less than $10^{-5} \cdot \sigma_m^2$, where σ_m^2 is the median variance of all time series $\boldsymbol{X}$.

The proposed estimation method depends on a handful of hyperparameters: the fixed number of cluster-specific factors C_0 in the clustering phase, number of neighbors k in the kNN graph, and the neighborhood threshold ε in the ε-N graph. Although the algorithm is not too sensitive to small changes in these hyperparameters, some quick guidelines are provided here on how to select them. Firstly, the algorithm in its clustering phase will not depend too much on the selection of C_0 since the cluster-specific factors in clusters where $C_k < C_0$ will model the C_k latent factors and the rest will be noise, while for clusters where $C_k > C_0$, all C_0 latent factors will be relevant. Nevertheless, a larger C_0 seems to be somewhat better, since it can efficiently incorporate the clusters with the larger number of cluster-specific factors. Therefore an effective strategy would be to simply choose C_0 as the largest number of cluster-specific factors one would expect in a single cluster. In this thesis, the value $C_0 = 5$ is used in all of the simulations and results. Furthermore, the number of neighbors k in the kNN graph should primarily reflect the size of the clusters to detect in the data. These are naturally dependent on the number of time series N - as a rule of thumb, a good choice will be somewhere between $\log N$ and $\sqrt{N}$. In the simulations and results, the value used is $k = \lceil (\log N + \sqrt{N})/2 \rceil$. Finally, for the selection of the neighborhood threshold in the ε in the ε-N graph, both the length of the time series T and their number N are best taken into account. Since longer time series will provide smaller estimation error and more accurate correlations between asserts ρ_{ij}, the standard error in the estimates will be reduced and the threshold may be lower - however, the threshold still needs to be above a certain level ρ_0 above which we wish the pairs of assets to be connected in the graph. To account for the statistical uncertainty in the estimation, the threshold can be set to the critical value of the approximate Pearson correlation test for the null hypothesis $H_0 : \rho_{ij} = \rho_0$

with a two-sided alternative (the correlation being significantly different than zero):

$$\varepsilon = \frac{1+\rho_0}{1-\rho_0} \exp\left(\frac{2z}{\sqrt{T}-3}\right), \tag{3.39}$$

where $\rho_0 = 0.4$, T is the time window length, and z is the $1-\alpha$ quantile of the standard normal distribution $\mathcal{N}(0,1)$. To account for the fact that the test is applied to all pairwise coefficients ρ_{ij}, the Bonferroni correction is used, with $\alpha = 0.05/\binom{N}{2}$. These values are used in the simulations and results for all different lengths of time windows.

Finally, the number n of pervasive factors and clusters to be considered in the model selection algorithm is set to $n = 3$. In this way, a total of $n^2 = 9$ models are fitted, which is significantly less than a grid search method using only the information criterion as a decision metric. Nevertheless, the fact that several combinations are considered allows the procedure not to discard potentially better solutions (which might happen if $n = 1$). The model selection procedure is therefore a hybrid approach between the information criteria (which usually require grid search algorithms) and statistical heuristic approaches (which are more computationally efficient but may produce unreliable results).

3.3 Covariance estimation with pervasive and cluster-specific latent factors

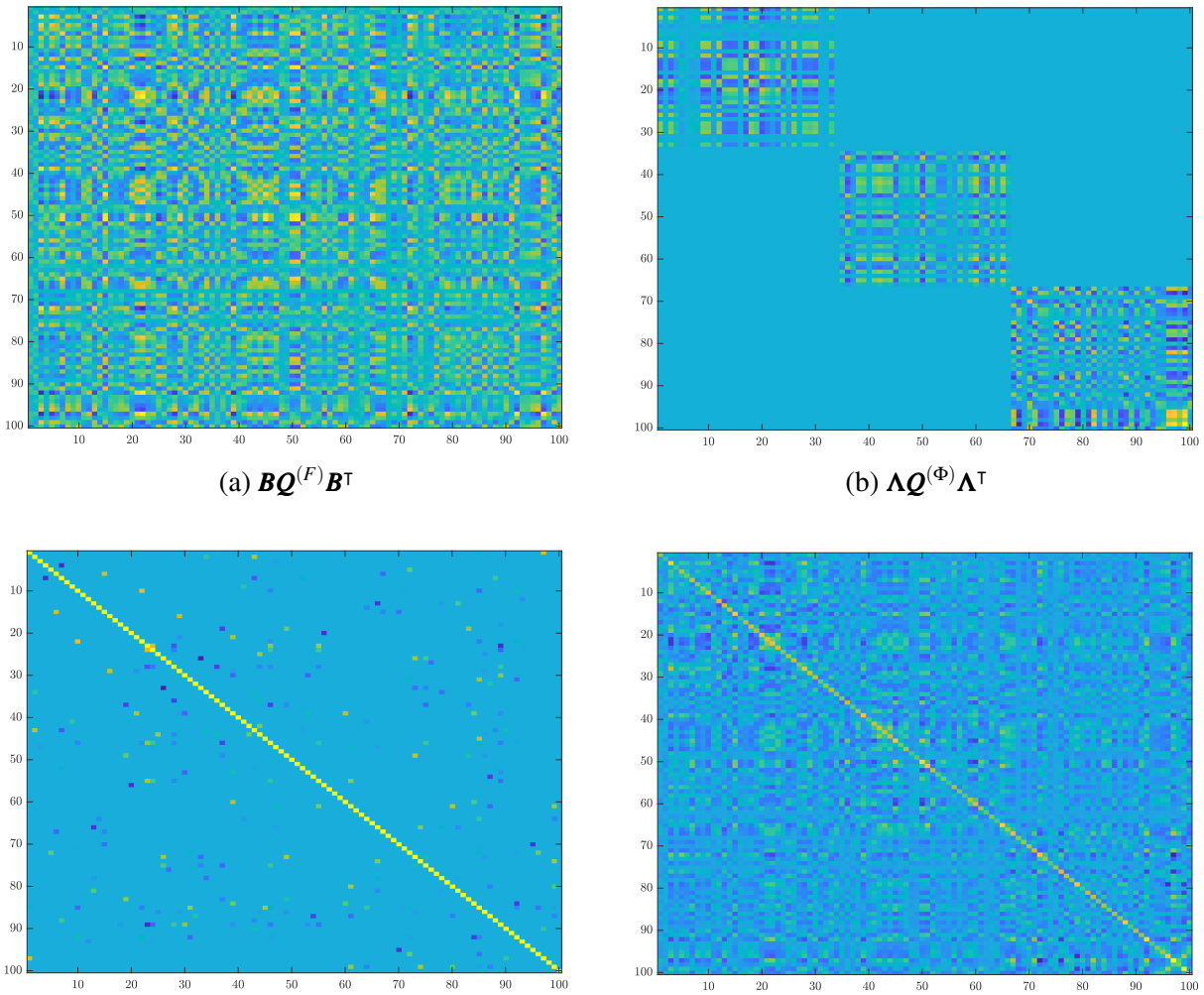
3.3.1 Covariance decomposition

Given the data $\boldsymbol{X}$ and the estimates of the latent factors, an improved estimate for the covariance matrix $\boldsymbol{Q}$ can be obtained. In the proposed factor model with pervasive and cluster-specific factors, the pervasive factor covariance $\boldsymbol{Q}^{(F)} = \mathrm{Cov}(\boldsymbol{F})$ and the cluster-specific factor covariance $\boldsymbol{Q}^{(\Phi)} = \mathrm{Cov}(\boldsymbol{\Phi})$ are assumed to be positive-definite, thus allowing for some correlation between factors. The idiosyncratic covariance $\boldsymbol{\Psi} = \mathrm{Cov}(\boldsymbol{e})$ is not necessarily diagonal, but it needs to be sparse (the cross-correlations in the idiosyncratic components can not be a consequence of common factors in the data) [20]. In this model, the variability in the data is decomposed into three components:

$$\boldsymbol{Q} = \boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\mathsf{T}} + \boldsymbol{\Lambda}\boldsymbol{Q}^{(\Phi)}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{\Psi}, \tag{3.40}$$

namely the covariance due to the pervasive factors, the covariance due to cluster-specific factors (which will have a block structure), and the idiosyncratic covariance – as displayed in Figure 3.12.

The pervasive factor part of the covariance $\hat{\boldsymbol{B}}\hat{\boldsymbol{Q}}^{(F)}\hat{\boldsymbol{B}}^{\mathsf{T}}$ can be directly calculated from the latent factor estimates: $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{F}}$, with $\hat{\boldsymbol{Q}}^{(F)} = \mathrm{Cov}(\hat{\boldsymbol{F}})$. The same applies to the cluster-specific factor part of the covariance $\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{Q}}^{(\Lambda)}\hat{\boldsymbol{\Phi}}^{\mathsf{T}}$ and the estimates: $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{\Lambda}}$, with $\hat{\boldsymbol{Q}}^{(\Lambda)} = \mathrm{Cov}(\hat{\boldsymbol{\Lambda}})$.



(a) $\boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^{\intercal}$ (b) $\boldsymbol{\Lambda}\boldsymbol{Q}^{(\Phi)}\boldsymbol{\Lambda}^{\intercal}$

(c) $\boldsymbol{\Psi}$ (d) $\boldsymbol{Q}$

Figure 3.12: The decomposition of the data covariance $\boldsymbol{Q}$ into the pervasive factor component, cluster-specific factor component, and the idiosyncratic component. The pervasive factor component is a low rank matrix of rank P, the cluster-specific factor component is a low rank and block matrix of rank Q, with a total of K blocks, and the idiosyncratic component is diagonal matrix of idiosyncratic variances and sparse idiosyncratic covariance elements.

3.3.2 Sparse idiosyncratic covariance estimation

To estimate the sparse idiosyncratic component covariance matrix $\hat{\boldsymbol{\Psi}}$ from the data $\boldsymbol{X}$, one can start from a full sample covariance estimate $\hat{\boldsymbol{\Psi}}^{(\text{full})}$ and apply an adaptive thresholding technique [154, 155]. A specific threshold is set for each element of the matrix $\hat{\boldsymbol{\Psi}}^{(\text{full})}$, so that the scale (the variance of each time series) is taken into account. The simplest way to do this is to consider the sample correlation matrix $\hat{\boldsymbol{R}}^{(\text{full})}$, and apply a fixed threshold ε_r to all elements:

$$\hat{\boldsymbol{R}}_{ij}^{(\text{sparse})} = \begin{cases} 0, & \text{if } |\hat{\boldsymbol{R}}_{ij}^{(\text{full})}| < e_r \\ \hat{\boldsymbol{R}}_{ij}^{(\text{full})}, & \text{if } |\hat{\boldsymbol{R}}_{ij}^{(\text{full})}| \geq \varepsilon_r. \end{cases} \tag{3.41}$$

The sparse correlation matrix $\hat{\boldsymbol{R}}^{(\text{sparse})}$ thus contains only elements larger than ε_r or smaller than $-\varepsilon_r$. However, this simple hard thresholding rule does not always produce positive-definite matrices $\hat{\boldsymbol{R}}^{(\text{sparse})}$, since certain elements $\hat{\boldsymbol{R}}_{ik}^{(\text{sparse})}$ and $\hat{\boldsymbol{R}}_{jk}^{(\text{sparse})}$ may be non-zero (pass above the threshold ε_r, but the element $\hat{\boldsymbol{R}}_{ij}^{(\text{sparse})}$ may be zero (fall under ε_r). This case may be generalized in the term of graphs - the sparse correlation matrix defines a graph where the edges are only those pairwise correlations which surpass the threshold value ε_r. This graph is actually a very sparse graph with a relatively large number of connected components - however each component may not necessarily be fully connected, and as long as they are not, the resulting correlation matrix will not necessarily be positive-definite. Thus, in order to correct this, it is possible to over all connected components defined by matrix $\hat{\boldsymbol{R}}^{(\text{sparse})}$, and assure that all links in those components are non-zero – thus adding additional non-zero elements $\hat{\boldsymbol{R}}_{ij}^{(\text{sparse})}$ (if $\hat{\boldsymbol{R}}_{ik}^{(\text{sparse})}$ and $\hat{\boldsymbol{R}}_{jk}^{(\text{sparse})}$ exist). The resulting new matrix $\hat{\boldsymbol{R}}^{(\text{sparse})}$ is still sparse, but will be positive-definite. Finally, the sparse covariance matrix is reconstructed from the sparse correlation matrix:

$$\hat{\boldsymbol{\Psi}} = \sqrt{\text{diag}(\hat{\boldsymbol{\Psi}}^{(\text{full})})}\,\hat{\boldsymbol{R}}^{(\text{sparse})}\sqrt{\text{diag}(\hat{\boldsymbol{\Psi}}^{(\text{full})})}. \tag{3.42}$$

By doing so, more elements which do not pass the threshold ε_r are included in the idiosyncratic correlation/covariance matrix, but the estimate is assured to be positive-definite (as long as certain securities are not identical and their correlation is not equal to 1).

Chapter 4

Portfolio optimization based on latent factors

4.1 Portfolio optimization framework

In this thesis, the mathematical framework of the Markowitz modern portfolio theory [80] is considered. The main aspect of this most commonly used approach to portfolio selection is the fact that risk is modeled as the variance of the portfolio and the portfolio mean return is considered a reward. The general portfolio selection problem (also called mean-variance analysis [156]) can be expressed as an optimization task

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^\top \boldsymbol{Q} \boldsymbol{w} \\ \text{s. t.} \quad & \boldsymbol{w}^\top \boldsymbol{\mu} \geq r, \\ & \boldsymbol{w}^\top \mathbf{1} = 1, \end{aligned} \tag{4.1}$$

where $\boldsymbol{\mu}$ are the expected returns and $\boldsymbol{Q}$ is the covariance matrix of N security returns, r is a required level of return, and $\boldsymbol{w} \in \mathbb{R}^N$ is the vector of portfolio weights for N securities. The problem can simply be interpreted as finding the portfolio with minimal variance for a given expected rate of return r.

Many other similar formulations also exist (for instance, maximizing the rate of return for a given portfolio variance or maximizing a utility function which is proportional to the portfolio return and inversely proportional to the variance). In addition, many applications of mean-variance analysis for portfolio optimization in real-world scenarios also include the non-

negativity constraint:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^\top \boldsymbol{Q} \boldsymbol{w} \\ \text{s. t.} \quad & \boldsymbol{w}^\top \boldsymbol{\mu} \geq r, \\ & \boldsymbol{w}^\top \mathbf{1} = 1, \\ & w_i \geq 0, \ \forall i = 1, ..., N. \end{aligned} \tag{4.2}$$

This is due to the fact that, in practice, short positions (negative weights) are often more complicated (and expensive) to enter and maintain. Although other characterizations of risk and reward have been proposed and used through the past decades (for instance, downside risk measures such as VaR, CVaR, maximum drawdown etc.), they have been found to produce less stable results and often come with estimation issues – mean-variance analysis has proven to be a robust and elegant approach [157].

The optimization problem (4.2) has an optimal portfolio solution for each required rate of return r – however, for a given vector of expected security returns $\boldsymbol{\mu}$, the minimal and maximal values of r for which the problem has a solution are defined by the minimum and maximum element of $\boldsymbol{\mu}$ (i.e. the smallest and largest of the expected security returns). All of the optimal portfolios form the *efficient frontier* – a continuous line of portfolios in the risk-return coordinate system, as displayed in Figure 4.1. The Figure demonstrates the Markowitz approach and its main benefit – through diversification of individual security risks it is possible to obtain less risky portfolios with equal or higher rates of returns.

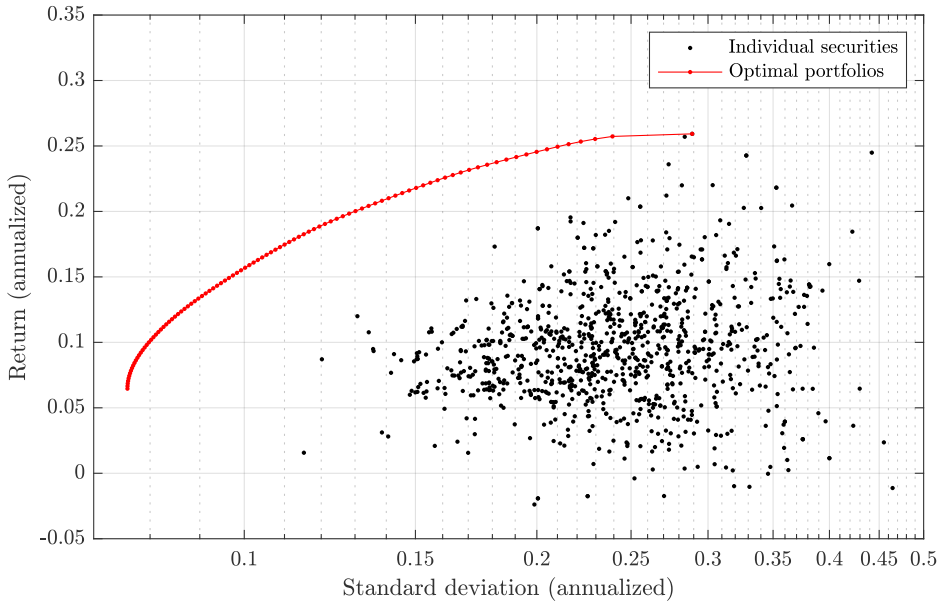


Figure 4.1: The risk and return profile of individual securities and the efficient frontier of the optimal mean-variance long-only portfolios. The x-axis is displayed in log-scale.

On this efficient frontier lie several specific portfolios. The minimum variance portfolio is the point on the efficient frontier furthest to the left, and represents the portfolio with the

minimal variance an investor can obtain given the considered securities. The maximum Sharpe ratio portfolio is the portfolio which maximizes:

$$S = \frac{\mu_p - r_{rf}}{\sigma_p}, \tag{4.3}$$

which is a simple ratio of the mean portfolio return μ_p (in excess of the risk-free rate r_{rf}) and the portfolio variance σ_p. The Sharpe ratio is the most commonly used risk-adjusted portfolio performance measures, useful in comparing portfolios of different rates of return and risk – according to that measure, the maximum Sharpe ratio (MSR) portfolio is the one with the best risk-adjusted performance. It is also known as the tangency portfolio, since it also lies on the tangent connecting the point $(0, r_{rf})$ and the efficient frontier (therefore it also depends on the risk-free rate r_{rf}).

In the most general sense, the mean-variance portfolio optimization approach requires the knowledge of expected security returns $\boldsymbol{\mu}$ and their covariance $\boldsymbol{Q}$, or rather – their prediction for the future portfolio holding period. As shown previously (see the results in Section 2.2.1), the security return covariance is generally predictable and does exhibit some degree of memory – however, the expected returns are generally hard to predict. This ultimately means that the efficient frontier of optimal portfolios calculated using in-sample estimations of mean returns will be prone to these estimation errors and the optimal portfolios will most definitely not be optimal out-of-sample. Due to this, the focus of academic research and financial industry has recently shifted to methods which do not require the prediction of future returns. Many different portfolio selection methods have been proposed and are in use - among the most simple portfolios is the *equal-weights* portfolio, which assigns equal fractions as the weight of each security:

$$\boldsymbol{w}_i^{(EW)} = \frac{1}{N}, \quad \forall i = 1, ..., N. \tag{4.4}$$

The EW portfolio is optimal in the mean-variance sense when all pairwise security correlations are equal, the security volatilities are equal, and their expected returns are equal – a very reductive set of assumptions which rarely hold. Nevertheless, since it does not require any estimation or optimization, it generally does not exhibit estimation error risk which more complicated methods often come with. The EW portfolio is often the benchmark used in most portfolio optimization scenarios and has been found to be fairly difficult to outperform [158]. However, since some inference can be made on the security covariance, it can be used to obtain better portfolios – this family of so-called *risk-based* portfolios is considered in the following section.

4.2 Risk-based portfolio optimization

Risk-based portfolio optimization methods are generally based only on the estimated security covariance, and have recently been gaining increased attention in the academic community and financial industry [159, 160]. The general idea is to avoid making inferences on future returns, and instead focus on diversifying or managing the risks associated with the considered security universe. They have been shown to provide robust portfolios which can often outperform other portfolio optimization methods, mostly due to the fact that they avoid the risk induced by the expected return estimation errors. Naturally, it is then expected that these portfolios are sensitive to the estimation of security covariance matrices [161] – but can also be improved by better estimates of the covariance, which can be provided by the considered latent factor models. Among the different methods proposed within this approach, two specific formulations for portfolio optimization problems are considered in this thesis – the *minimum variance* and *maximum diversification* portfolios.

4.2.1 Minimum variance portfolio

The simplest way to avoid expected return estimation in portfolio optimization is to simply omit the terms containing the portfolio returns. This leads to the global minimum variance (GMV) portfolio, which is obtained by optimizing a simpler optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^\top \boldsymbol{Q} \boldsymbol{w} \\ \text{s. t.} \quad & \boldsymbol{w}^\top \mathbf{1} = 1. \end{aligned} \tag{4.5}$$

This formulation also has an analytical solution:

$$\boldsymbol{w}^{(\mathrm{GMV})} = \frac{\boldsymbol{Q}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{Q}^{-1}\mathbf{1}}. \tag{4.6}$$

However, this solution requires the existence of the covariance matrix inverse $\boldsymbol{Q}^{-1}$, i.e. that the covariance is positive definite. Since this may not always be the case, especially in high-dimensional settings, optimization procedures are often employed to obtain better solutions.

The absence of the requirement for a given rate of return of the portfolio may be interpreted in two ways: (i) no knowledge on the expected returns is necessary so the portfolio is somewhat agnostic to expected returns; (ii) it can be shown the GMV portfolio is in fact equal to the MSR portfolio when all expected returns of individual securities are equal [50]. This also holds for

the constrained minimum variance portfolios, which are more commonly used in practice:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^\top \boldsymbol{Q} \boldsymbol{w} \\ \text{s. t.} \quad & \boldsymbol{w}^\top \mathbf{1} = 1, \\ & w_i \geq 0, \, \forall i = 1, ..., N. \end{aligned} \tag{4.7}$$

Minimum variance portfolios have been found to perform very well out of sample [159, 162] – mostly due to the absence of prediction errors for expected returns which substantially deteriorate the performance of mean-variance portfolios [163]. Another important element in the empirical performance of these portfolios in market environments is the so-called *low-volatility anomaly*, documented in some financial markets [111]. This phenomenon is essentially the tendency of securities with low volatility to exhibit higher mean returns (on a long period of historical data), contrary to the stipulations of the CAPM model. In such environments, the minimum volatility portfolios will often outperform the EW portfolios and some other optimal portfolios, not only in terms of lower risk, but also in terms of returns.

Since the estimation procedure only depends on the covariance, it is a suitable way to demonstrate the ability of the various risk models to provide reliable estimates, and is often used to benchmark covariance estimation methods [41, 108]. Previous studies have indeed found that improved estimators of the correlation and covariance matrix produce more diversified portfolios with lower out-of-sample risk [164]. Thus, latent factor models can be used to obtain covariance estimates, as described in Section 3.3 for the latent factor model with pervasive and cluster-specific factors.

However, the low-volatility anomaly has only been reported for some financial markets and may not be a persistent assumption across global universes of securities. Moreover, the minimum variance portfolios often allocate very high weights to lower volatility and lower return securities (such as Treasury Bills or money market funds, if such exist in the asset universe), which often have undesirably low returns. With such concentration in less volatile securities, the optimal minimum variance portfolios do not necessarily provide a diversified investment in terms of exposure to different asset classes or global securities.

4.2.2 Maximum diversification portfolio

Another perspective on what makes a portfolio portfolio optimal comes from looking into the diversification properties of the portfolio. Consider the *diversification ratio* of a portfolio:

$$d = \frac{\sum_{i=1}^{N} w_i \sigma_i}{\sqrt{\boldsymbol{w}^\top \boldsymbol{Q} \boldsymbol{w}}}, \tag{4.8}$$

in other words, a ratio of the weighted average of individual security volatilities divided by the portfolio volatility [165]. For instance, of the two portfolios with equal volatilities, the one with a larger weighted average of weights will be more diversified. Perhaps surprisingly, equal-weighted portfolios will often come up as very well diversified as opposed to the minimum variance portfolios, since the latter often have concentrated weights in a few low-volatility securities (thus reducing the the numerator of the ratio). The ratio is largest for portfolios with weights leaning towards more volatile securities, but with low overall volatility – such a portfolio evidently exploits the cross-correlations to reduce the volatility and thus provides maximum diversification.

As opposed to the mean-variance analysis, here the optimal portfolio is the portfolio which maximizes the diversification ratio - the maximum diversification portfolio (also known as the most diversified portfolio). For the unconstrained MD portfolio, an analytical solution is available:

$$\boldsymbol{w}^{(\mathrm{MD})} = \frac{\boldsymbol{Q}^{-1}\boldsymbol{\sigma}}{\boldsymbol{\sigma}^{\intercal}\boldsymbol{Q}^{-1}\boldsymbol{\sigma}}, \tag{4.9}$$

where σ is the vector of individual security volatilities. However, in the constrained case (i.e. no short selling allowed), and when the covariance matrix estimate is positive semi-definite, an optimization procedure is applied to solve the following problem:

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \frac{\sum_{i=1}^{N} w_i \sigma_i}{\sqrt{\boldsymbol{w}^{\intercal}\boldsymbol{Q}\boldsymbol{w}}} \\ \text{s. t.} \quad & \boldsymbol{w}^{\intercal}\mathbf{1} = 1, \\ & w_i \geq 0,\ \forall i = 1,...,N. \end{aligned} \tag{4.10}$$

The MD portfolio has been shown to be equivalent to the maximum Sharpe ratio portfolio when the individual security Sharpe ratios are equal – meaning that their expected returns are proportional to their volatilities [165]. Thus, the optimization problem (4.10) can be converted to the mean-variance problem in (4.2) with the expected returns equal to the individual security volatilities. The equal Sharpe ratio condition may be true if the markets are efficient to the extent that investor expect higher returns for securities associated with higher volatilities and that all other forecasts except risk are either inaccurate or already priced in [165]. The empirical results suggest that the MD portfolio has very favorable properties and provides stable risk-adjusted returns, avoiding the trap of over-concentrating into low-volatility assets. Even though it can be thought of as an MSR portfolio where the expected returns are proportional to individual security volatilities, it nevertheless depends only on the risk estimates which are more robust than expected return predictions. The equal Sharpe ratio condition for the mean-variance optimality of the MD portfolio may not always hold – in addition, they assume that the risk estimates are accurate. Thus, the latent factor models can help obtain improved covariance matrices and

better risk estimates, which will in turn improve the performance of the MD portfolios.

4.2.3 Cluster-based portfolio diversification

Rather than improving portfolios by including latent factor models in the estimation of risk and covariance matrices in high-dimensional financial time series, another approach is to rely on the estimated latent factor structures to devise a new portfolio optimization procedure. In a global universe of securities which are exposed to certain pervasive an cluster-specific factors, the latent factors and asset clusters will most commonly have a clear interpretation. The pervasive factors are global market shocks, and the cluster-specific factors either correspond to the geographical classification (country or region) or the asset classes of the considered securities. The risks can be estimated in a structured way – using the intra-cluster covariance for the risks of assets within the same clusters, and inter-cluster covariance for the different clusters. Instead of optimizing the diversification of the entire portfolio of N assets, here a two-step long-only portfolio optimization method is proposed.

Firstly, within each cluster $k = 1, ..., K$, an intra-cluster long-only MD portfolio $\boldsymbol{w}^{(k)} \in \mathbb{R}^{N_k}$ is calculated, containing only the N_k securities in the cluster, by solving the optimization problem in (4.10). The covariance matrix $\boldsymbol{Q}^{(k)} \in \mathbb{R}^{N_k \times N_k}$ of the securities in cluster k used for optimizing the intra-cluster portfolio is calculated using the latent factor model estimate in (3.40), using only the elements $\boldsymbol{Q}_{ij}$ for which both $\boldsymbol{g}_i = k$ and $\boldsymbol{g}_j = k$. The resulting portfolios are K vectors $\boldsymbol{w}^{(k)}$, with nonzero weights which all sum to 1.

From the K intra-cluster portfolios, an MD inter-cluster portfolio $\tilde{\boldsymbol{w}} \in \mathbb{R}^{\mathbb{K}}$ is formed. To this end, the inter-cluster covariance matrix is calculated as follows:

$$\tilde{\boldsymbol{Q}} = \boldsymbol{w}^{(\text{clust.})\intercal} \boldsymbol{Q} \boldsymbol{w}^{(\text{clust.})}, \tag{4.11}$$

where $\boldsymbol{w}^{(\text{clust.})} = [\boldsymbol{w}^{(1)}, ..., \boldsymbol{w}^{(K)}]$ is the $N \times K$ matrix of K intra-cluster MD portfolios. The resulting $K \times K$ inter-cluster covariance matrix is used to find the inter-cluster long only portfolio $\tilde{\boldsymbol{w}}$. The final two-step portfolio is calculated as:

$$\boldsymbol{w} = \sum_{k=1}^{K} \boldsymbol{w}^{(k)} \tilde{w}_k = \boldsymbol{w}^{(\text{clust.})} \tilde{\boldsymbol{w}}. \tag{4.12}$$

This portfolio reduces the possibility of allocating large weights to certain securities uncorrelated to the majority others (which the MD portfolio will generally do), by allocating all securities with respect to their cluster first, and then assigning inter-cluster weights. From a top-down approach, the proposed portfolio optimization method generally attempts to maximize the diversification of the portfolio with respect to the clusters of securities (rather than individual securities themselves). This is done by the inter-cluster portfolio $\tilde{\boldsymbol{w}}$ which is a MD portfolio of

the individual intra-cluster portfolios. In order to represent the clusters of securities in a single intra-cluster portfolio for each cluster, the MD portfolio of securities belonging to each cluster is used, thus maximizing the diversification within and between clusters. The resulting portfolio will outperform the MD portfolio of individual securities if the security covariance estimates are more reliable within the clusters and between the cluster MD portfolios, than the entire security universe covariance. Moreover, the portfolio will generally be expected to outperform in markets where the Sharpe ratios can be considered more homogeneous between the clusters (represented by the intra-cluster portfolios), rather than between all individual securities. Even if this may not strictly hold in empirical data, the resulting portfolios are very robust to estimation errors, as will be shown in the following chapter.

Chapter 5

Results

5.1 Simulations

5.1.1 Simulation framework

To verify the validity of the proposed approach and test the empirical properties of the estimation algorithm, several data-generating processes (DGP) are defined, which correspond to the assumed factor model structures. To obtain a model in the form of 3.24, random clusters and factor loadings are generated. The elements of the pervasive factor loadings matrix $\boldsymbol{B}$ are drawn from a uniform random distribution with mean 0 and variance 1. For the cluster-specific loadings matrix $\boldsymbol{\Lambda}$, the elements $\Lambda_i^{(k)}$ are random (also uniform with mean 0 and variance 1) if asset i belongs to cluster k, and are zero otherwise. The clusters are defined so that their sizes are all equal $N_k = N/K$. Since the approximate factor model allows for some off-diagonal elements in the covariance of residuals, random sparse covariance matrices are also generated, with a given idiosyncratic variance σ_e^2 on the diagonal. Given the factor loadings and the idiosyncratic components, the security return mean and covariance can then be calculated as

$$\begin{aligned} \boldsymbol{\mu} &= \boldsymbol{\mu}^{(F)}\boldsymbol{B}^\intercal + \boldsymbol{\mu}^{(\Phi)}\boldsymbol{\Lambda}^\intercal, \\ \boldsymbol{Q} &= \boldsymbol{B}\boldsymbol{Q}^{(F)}\boldsymbol{B}^\intercal + \boldsymbol{\Lambda}\boldsymbol{Q}^{(\Phi)}\boldsymbol{\Lambda}^\intercal + \boldsymbol{\Psi}, \end{aligned} \tag{5.1}$$

where $\boldsymbol{\mu}_F$ are the means of P pervasive factors, while $\boldsymbol{\mu}_\Phi$ are the means of Q cluster-specific factors. In the simulations, the means are all zero, and the covariances are both diagonal matrices with equal variances σ_F^2 and σ_Φ^2 on the diagonal. The full set of simulation parameters is given in Table 5.1. The clusters selected have a relatively diverse number of cluster-specific factors, which will present an additional problem to the clustering algorithm in the aspect of cluster size bias (the fact that sizes of clusters with a larger number of cluster-specific factor may be overestimated).

To simulate security returns, means and covariances are used to simulate realizations of N-

Table 5.1: Simulation parameters.

Parameter	Symbol	Value
Number of assets	N	1000
Number of pervasive factors	P	5
Number of clusters	K	5
Number of cluster-specific factors	C	$[1,2,3,4,5]$
Pervasive factor variance	σ_F^2	0.1
Cluster-specific factor variance	σ_Φ^2	0.1
Idiosyncratic variance	σ_e^2	0.5

dimensional returns, drawing from the i.i.d. multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q})$, with the probability density:

$$f_{\mathcal{N}}(\boldsymbol{X}_t) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{Q}|}} \exp\left[-\frac{1}{2}(\boldsymbol{X}_t - \boldsymbol{\mu})^\intercal \boldsymbol{Q}^{-1}(\boldsymbol{X}_t - \boldsymbol{\mu})\right]. \tag{5.2}$$

Since the Gaussian normal distribution does not replicate the heavy-tailed properties of security returns (as discussed in Section 2.2.3), the Student's t-distribution is also considered:

$$f_t(\boldsymbol{X}_t) = \frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu^N \pi^N |N|}}\left[1 + \frac{1}{\nu}(\boldsymbol{X}_t - \boldsymbol{\mu})^\intercal \boldsymbol{Q}^{-1}(\boldsymbol{X}_t - \boldsymbol{\mu})\right]^{\frac{\nu+N}{2}}, \tag{5.3}$$

where $\Gamma(x)$ is the gamma function and ν is the degrees-of-freedom parameter. For $\nu < 2$, the distribution does not have a finite variance, and for $\nu \to \infty$ the distribution is equivalent to the Gaussian normal. In this simulation scenario, the Student's t-distributions with 6 degrees of freedom $t_6(\boldsymbol{\mu}, \boldsymbol{Q})$ and 4 degrees of freedom $t_4(\boldsymbol{\mu}, \boldsymbol{Q})$ are considered. Although many latent factor models in finance and the corresponding estimation methods are often tested using simulations of normally distributed data [20, 51], the Student's t-distributions are additionally used here, since they replicate the heavy-tailed property of financial returns, as seen in Figure 5.1.

Therefore, the three DGPs defined for this simulation study all have the same latent factor structure (with pervasive and cluster-specific factors), and are simulated with the three considered distributions with varying levels of tail heaviness: $\mathcal{N}$, t_6, and t_4.

5.1.2 Estimator properties in high dimensions

First, the empirical properties of the proposed latent factor estimation method are tested. To this end, consider the mean squared error of the latent factor component in the data, as defined in

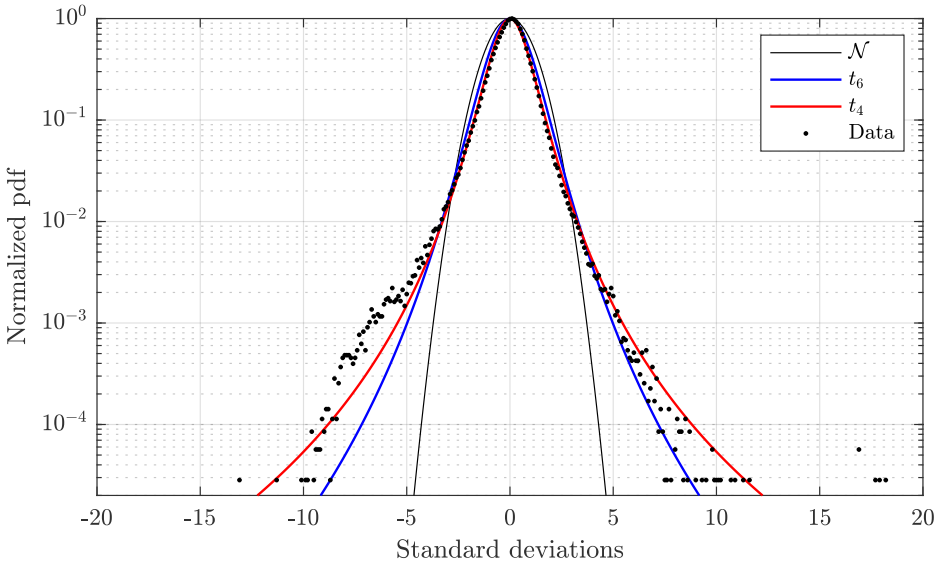


Figure 5.1: The normalized pdfs of the three considered theoretical distributions, together with the empirical histogram of the weekly returns of NASDAQ global equity indices between 2005 and 2020.

(3.18), but in this case with a pervasive and cluster-specific component:

$$MSE = \frac{1}{NT}||\boldsymbol{F}\boldsymbol{B}^\intercal + \boldsymbol{\Phi}\boldsymbol{\Lambda}^\intercal - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\intercal - \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}^\intercal||_F^2, \tag{5.4}$$

where the true factor loadings are $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$, and their estimates are $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Lambda}}$ (the same holds for the factor realizations $\boldsymbol{F}$ and $\boldsymbol{\Phi}$ and their estimates $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{\Phi}}$). The error defined above should converge towards zero as the time series length T and their dimension N grow (on the other hand, the error of the data realizations $\boldsymbol{X}$ and the estimate $\hat{\boldsymbol{X}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\intercal - \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}^\intercal$ will converge towards the average idiosyncratic variance. To compare the efficiency of the proposed estimator with the PC estimator (which can ultimately estimate any high-dimensional approximate factor model), both are applied to a randomly generated factor model as described in the Section 5.1.1 given a maximal number of securities $N_{\max}$. The number of securities N was then selectively increased from N_1 to $N_{\max}$ so that the securities selected equally represent all clusters. The results are displayed in Figure 5.2.

The results indicate that, similar to the PC estimator on latent factor models with only pervasive factors (as in Figure 3.3), the estimator all reduce the MSE with increasing N – the "big data blessing" holds in the model with pervasive and cluster-specific latent factors as well. However, the PC estimator is not able to reduce the error enough, even for very high N – on the other hand, the proposed model-based estimator quickly converges to a very low error. In addition, the presence of heavy tails in the data deteriorates the performance of the PC estimator considerably, while the error of model-based estimator remains fairly low for both Gaussian and Student's t distributed data. This remarkable property of the model-based estimator is most likely due to the fact that heavy tails in high-dimensional time series may extremely affect the

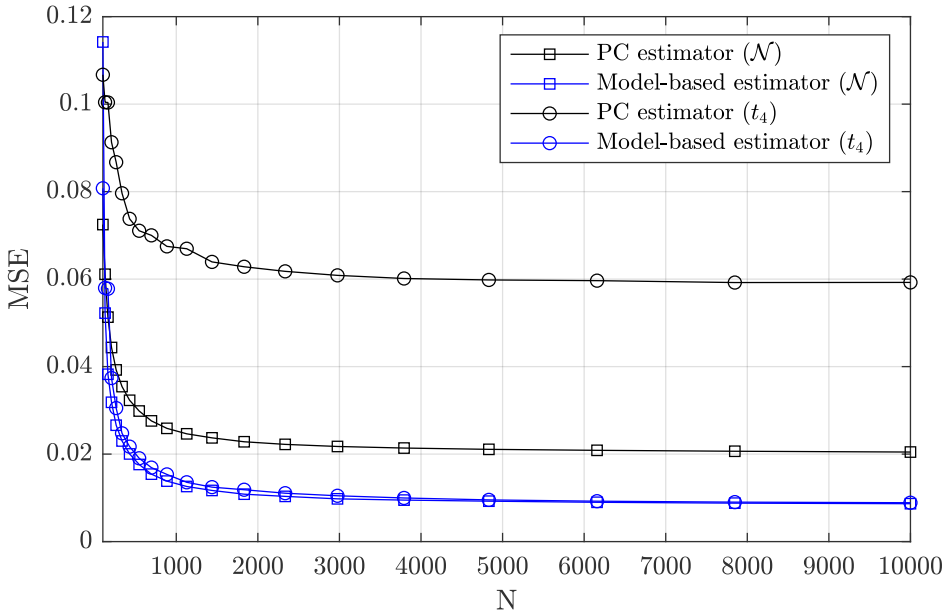


Figure 5.2: The *MSE* of the PC and model-based latent factor estimators, for the number of assets between 100 and 10000, and the Gaussian normal $\mathcal{N}$ and the Student' t-distribution with $\nu = 4$ degrees of freedom (t_4). The simulated time series length is $T = 250$.

empirical correlation estimates – when the latent factors are estimated for the entire set of N time series instead of the cluster N_k, both the loadings and estimated factor realizations may be severely influenced.

5.1.3 Clustering performance

The proposed algorithm is also verified in terms of clustering accuracy. The estimator is applied to the simulation data given the correct P and K, and the quality of clustering is measured by comparing the estimated clustering $\hat{\boldsymbol{g}}$ and ground truth clustering $\boldsymbol{g}$ using the Rand statistic and Jaccard coefficient [166], both of which are commonly used techniques to measure the agreement between different partitions of the same set and can be used even when there are no class labels available [166]. Given the estimated clustering $\hat{\boldsymbol{g}}$ and the ground truth clustering $\boldsymbol{g}$, define the following variables:

$$\begin{aligned} SS &= \sum_{i}^{N} \sum_{j=i+1}^{N} 1[(\hat{g}_i = \hat{g}_j) \wedge (g_i = g_j)], \\ SD &= \sum_{i}^{N} \sum_{j=i+1}^{N} 1[(\hat{g}_i = \hat{g}_j) \wedge (g_i \neq g_j)], \\ DS &= \sum_{i}^{N} \sum_{j=i+1}^{N} 1[(\hat{g}_i \neq \hat{g}_j) \wedge (g_i = g_j)], \\ DD &= \sum_{i}^{N} \sum_{j=i+1}^{N} 1[(\hat{g}_i \neq \hat{g}_j) \wedge (g_i \neq g_j)], \end{aligned} \tag{5.5}$$

where $1[c]$ is an indicator function with value 1 if the condition c in the brackets holds, and 0 otherwise. The variable SS simply counts the number of pairs of assets which belong to the same cluster in both clusterings $\hat{\boldsymbol{g}}$ and $\boldsymbol{g}$; SD counts the number of pairs belonging to the same cluster in $\hat{\boldsymbol{g}}$ and different clusters in $\boldsymbol{g}$; DS counts the number of pairs belonging to different clusters in $\hat{\boldsymbol{g}}$ and the same cluster in $\boldsymbol{g}$; DD counts the number of pairs belonging to different clusters in both clusterings $\hat{\boldsymbol{g}}$ and $\boldsymbol{g}$. Given these variables, the Rand statistic and the Jaccard coefficient can be calculated:

$$\begin{aligned} \text{Rand} &= \frac{SS+DD}{SS+SD+DS+DD}, \\ \text{Jaccard} &= \frac{SS}{SS+SD+DS}. \end{aligned} \tag{5.6}$$

Following the above expression, in this case the Rand statistic simply measures the proportion of pairs which are correctly clustered together or apart, and the Jaccard coefficient measures the intersection of the correctly clustered pairs in proportion to the union of all the pairs of securities. Both of these can be interpreted as focusing on the sets of pairs, rather than the original set of securities, and look into whether the pairwise clustering properties match in the two given clusterings.

Moreover, for both of these cluster validation measures and any pair of clustering methods, a paired statistical test procedure is defined* in order to test the hypothesis:

$$\begin{aligned} &H_0 : \text{There is no difference between two clustering methods,} \\ &H_1 : \text{Method 2 outperforms Method 1.} \end{aligned} \tag{5.7}$$

For each randomly generated model $m = 1, ..., m_{\max}$, the considered clustering methods are applied and the cluster validation measure is calculated for both results (for instance $\text{Rand}_1(m)$ and $\text{Rand}_2(m)$), then the p-value is calculated as the fraction of pairs for which Method 2 outperforms Method 1 (in this example, the fraction of samples for which $\text{Rand}_2(m) > \text{Rand}_1(m)$).This procedure is repeated for the both cluster validation measures, pairing the proposed model-based method with several commonly used clustering approaches (k-means algorithm, spectral clustering [152], and the Ando-Bai estimation procedure [52]). The k-means method uses $1 - |\rho_{ij}|$ as a distance measure, and the spectral clustering method employs the proposed asset graph estimated directly from $\boldsymbol{X}$. The Ando-Bai procedure iteratively estimates clusters and latent factors, but using a procedure which does not account for the bias in clusters with different numbers of cluster-specific factors.

A number of $m_{\max} = 1000$ models are randomly generated, for each model the time series realizations of length $T = 1000, 500, 250$ are simulated, and the considered clustering methods

*Since the models are randomly generated, each model realization presents different conditions for the considered clustering methods, which need to be taken into account in a paired fashion.

and tests are applied. The average Rand and Jaccard statistics, as well as the p-values of the paired resampling tests (comparing the proposed model-based method with each of the other considered clustering methods) are shown in Tables 5.2 and 5.3.

Table 5.2: Rand statistics on simulation data for the proposed method and other considered clustering techniques, using different simulation time window lengths and data distributions. The brackets below each value contain the p-value of the paired resampling test of the considered method compared to the proposed model-based algorithm. All of the values are obtained using simulation parameters given in Table 5.1.

	$T = 1000$		
	$\mathcal{N}$	t_6	t_4
k-means	68.03% (< 0.001)	68.03% (< 0.001)	67.98% (< 0.001)
Spectral clust.	77.02% (< 0.001)	74.10% (< 0.001)	74.10% (< 0.001)
Ando-Bai	89.99% (< 0.001)	88.90% (< 0.001)	88.34% (< 0.001)
Model-based	**99.03%**	**98.83%**	**98.53%**
	$T = 500$		
	$\mathcal{N}$	t_6	t_4
k-means	68.03% (< 0.001)	68.01% (< 0.001)	67.94% (< 0.001)
Spectral clust.	75.94% (< 0.001)	71.91% (< 0.001)	71.90% (< 0.001)
Ando-Bai	89.31% (0.003)	88.40% (0.003)	87.73% (< 0.001)
Model-based	**98.75%**	**98.39%**	**98.01%**
	$T = 250$		
	$\mathcal{N}$	t_6	t_4
k-means	68.02% (< 0.001)	67.98% (< 0.001)	67.90% (< 0.001)
Spectral clust.	73.68% (< 0.001)	69.75% (< 0.001)	69.74% (< 0.001)
Ando-Bai	88.09% (0.001)	87.26% (0.001)	86.44% (< 0.001)
Model-based	**98.13%**	**97.58%**	**97.15%**

These results demonstrate the advantage of the proposed model-based approach, as well as the fact that the existence of pervasive factors may severely hinder clustering accuracy when

Table 5.3: Jaccard coefficients on simulation data for the proposed method and other considered clustering techniques, using different simulation time window lengths and data distributions. The brackets below each value contain the p-value of the paired resampling test of the considered method compared to the proposed model-based algorithm. All of the values are obtained using simulation parameters given in Table 5.1.

	$T = 1000$		
	$\mathcal{N}$	t_6	t_4
k-means	11.09%	11.10%	11.12%
	(< 0.001)	(< 0.001)	(< 0.001)
Spectral clust.	29.34%	23.78%	23.77%
	(< 0.001)	(< 0.001)	(< 0.001)
Ando-Bai	61.16%	59.40%	
	(< 0.001)	(< 0.001)	(< 0.001)
Model-based	**95.27%**	**94.35%**	**92.95%**
	$T = 500$		
	$\mathcal{N}$	t_6	t_4
k-means	11.10%	11.11%	11.15%
	(< 0.001)	(< 0.001)	(< 0.001)
Spectral clust.	27.07%	20.00%	19.99%
	(< 0.001)	(< 0.001)	(< 0.001)
Ando-Bai	62.36%	59.57%	57.80%
	(0.003)	(0.003)	(< 0.001)
Model-based	**93.92%**	**92.31%**	**90.59%**
	$T = 250$		
	$\mathcal{N}$	t_6	t_4
k-means	11.11%	11.13%	11.17%
	(< 0.001)	(< 0.001)	(< 0.001)
Spectral clust.	22.69%	16.26%	16.26%
	(< 0.001)	(< 0.001)	(< 0.001)
Ando-Bai	58.86%	56.43%	54.47%
	(0.001)	(0.001)	(< 0.001)
Model-based	**91.10%**	**88.62%**	**86.75%**

they are not taken into account. Moreover, in the paired tests, the proposed method outperformed the considered methods for virtually all of the 1000 resampled model realizations (the p-values of < 0.001 mean that in the $m_{\max} = 1000$ simulated models, none were found for which the considered benchmark methods outperformed the proposed model-based estimator). To better visualize the paired comparison for these two methods across the simulations, the

Rand statistic is shown for the Ando-Bai and the proposed model-based method across all 1000 simulations in Figure 5.3. This figure demonstrates that not only the proposed model-based method outperforms the Ando-Bai estimator in the great majority of simulated cases, but also that the variance of the Rand statistic of the model-based method is considerably lower – meaning that the proposed method yields both accurate and stable results over a large number of different randomly generated models.

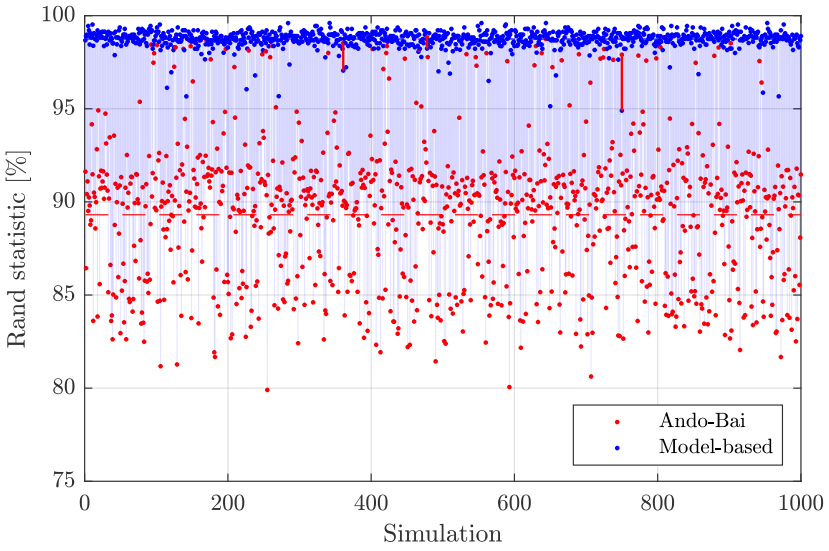


Figure 5.3: The Rand statistic for all the 1000 simulations and $T = 500$, given for the Ando-Bai and the proposed model-based estimation method. The two statistics for each simulation are connected with a transparent blue line if the model-based method outperforms the Ando-Bai method, and a red line otherwise (only 3 samples in this case). The dashed lines represent the average values of the statistics, corresponding to the values in Table 5.2.

Another important issue in the estimation of clusters of securities in models with cluster-specific factors is the cluster size bias. If the correct number of cluster-specific factors C_k for each cluster are used in the clustering procedure, the clusters with a larger C_k will contain more securities since the larger number of latent factors will necessarily explain more variability and thus attract more securities in the cluster through the iterative procedure. The proposed estimation algorithm in this thesis avoids the cluster size bias by holding an equal and fixed number of cluster-specific factors C_0 across all clusters during the clustering phase of the algorithm.

In order to look into the bias in clustering when the number of cluster-specific factors differ between clusters, in Figure 5.4 the average number of securities in clusters with different numbers of cluster-specific factors are also given, estimated by two different approaches. The first uses the real C_k as the number of cluster-specific factors in each cluster (corresponding to the Ando-Bai method [52]), while the second uses the fixed C_0 in each cluster during the clustering phase. The bias towards the clusters with a larger C_k is evident and might be a large source of inaccuracy in the clustering procedure, while the proposed model-based method with C_0 seems to provide accurate clustering without any evident bias in the cluster sizes. These results are

obtained for the $T = 500$ window and the t_4 distribution, but hold for all of the considered combinations.

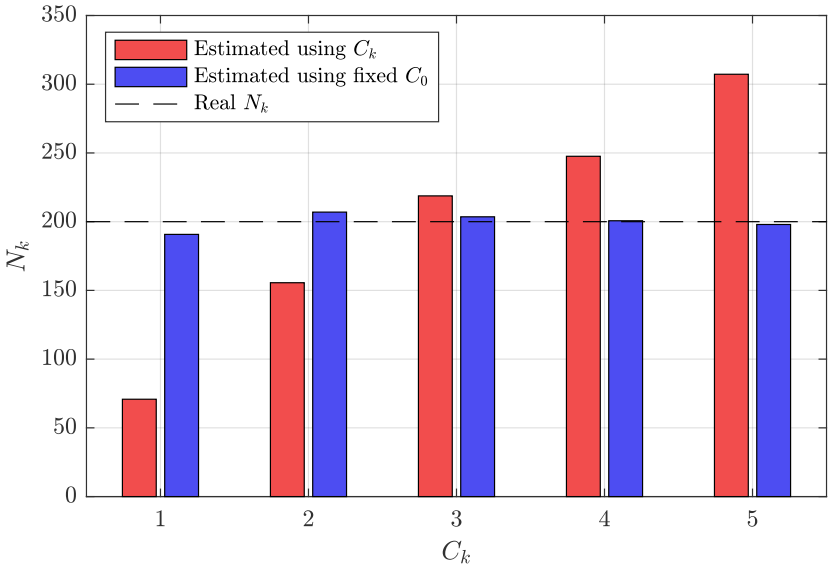


Figure 5.4: The sizes of clusters (number of assets N_k for different numbers of cluster-specific factors C_k, given by two estimation methods. The real number of assets in each cluster is known in the simulation and is equal to $N_k = 200$ for each k.

The above results all demonstrate that the proposed estimation algorithm works very well and outperforms other benchmark methods (the PC estimator and the benchmark clustering algorithms). The performance is shown to be robust to heavy tails in the data distributions and indeed profits from the high-dimensionality of the time series. Moreover, the proposed method avoids the cluster size bias which emerges when considering clusters with different numbers of cluster-specific factors – in the proposed estimation procedure, the numbers of cluster-specific factors are estimated after the clustering phase, given the estimated clusters and the data.

5.1.4 Model selection performance

The model selection method is also tested using the same simulation environment and the simulated time series lengths. In addition to measuring the percentage of correctly estimated number of pervasive factors, clusters and cluster-specific factors, the mean absolute deviation (*MAD*) for each of these is also calculated. The results are shown in Table 5.4.

The accuracy of the proposed model selection method is remarkably high, even when presented with heavy tailed data and short time window length. Only the number of pervasive factors seems to suffer a bit in case of the t_4 distribution and $T = 250$ – nevertheless, the accuracy for this case is 90%. The results suggest that the proposed model selection method may indeed be used with high-dimensional and heavy-tailed data to obtain reliable estimates for the number of clusters, pervasive factors and cluster-specific factors.

Table 5.4: Model selection accuracy on simulation data over different simulation time window lengths.

			$T = 1000$			
		Acc.			MAD	
	$\mathcal{N}$	t_6	t_4	$\mathcal{N}$	t_6	t_4
P	100%	100%	95%	0.00	0.00	0.05
K	100%	100%	100%	0.00	0.00	0.00
C_k	95.2%	100%	99.6%	0.10	0.00	0.01
			$T = 500$			
		Acc.			MAD	
	$\mathcal{N}$	t_6	t_4	$\mathcal{N}$	t_6	t_4
P	100%	100%	93%	0.00	0.00	0.08
K	99%	100%	100%	0.01	0.00	0.00
C_k	100%	99.4%	97.4%	0.00	0.01	0.03
			$T = 250$			
		Acc.			MAD	
	$\mathcal{N}$	t_6	t_4	$\mathcal{N}$	t_6	t_4
P	100%	98%	90%	0.00	0.02	0.14
K	100%	98%	98%	0.00	0.02	0.04
C_k	99.8%	99.8%	98.4%	0.01	0.01	0.02

5.2 Market data results

5.2.1 Historical market data

In addition to the simulations, two datasets containing weekly financial return time series are also used to obtain results on historical market data. Firstly, a dataset of NASDAQ global equity indices between 2005 and 2020 is considered [167]. The original dataset contains a large number of redundant time series, from which only the total return NASDAQ indices are considered, under the condition that they are available in the considered period, leaving the dataset with $N = 982$ securities. Additionally, a dataset of $N = 1480$ international stocks is considered, in the time period between 2005 and 2020. The data for this dataset is obtained by downloading historical return data using Yahoo Finance for the tickers of constituents of several MSCI broad internation market indices[†], and then again selecting only those time series which have price data for the entire considered period. Both of these datasets cover a wide range of exchanges, countries and specific sectors, and can be used to represent and study the latent risk

[†]Specifically, the combination of all stocks within the MSCI World, MSCI All Country World, MSCI Emerging Markets, and MSCI IXUS indices, since these represent the most of the international stocks in a large variety of markets and countries.

factors in global financial markets.

The simulation results confirm the ability of the proposed method to provide accurate estimates, even in the presence of correlated residuals, heavy tails, and high-dimensional sample data. However, in real financial market data, such as the NASDAQ global equity indices and MSCI data, the latent factors are unknown, as well as the clustering and the number of clusters and latent factors. The proposed method allows one to study and estimate these from the data directly. Firstly, two distinct periods in the NASDAQ dataset are considered: Figure 5.5 shows the asset graph for the period 2007-2009 around the global recession, and Figure 5.6 show the graph for the subsequent period 2010-2020 which corresponds to one of the strongest and longest bull markets in the history of financial markets. In both graphs, some common clusters emerge (shown in same colors on both graphs): European markets (pink), Brazil and Latin America (purple), North America and global developed market indices (blue), Asian emerging markets (teal), Middle East and Africa (darker green), Asian developed markets (light green). The 2007-2009 graph contains another cluster for India and New Zealand (yellow), and European emerging markets (red) - both of which are encapsulated within other clusters in the 2010-2020 graph. In addition to serving as a sanity check for the meaning behind the estimated clusters, these results suggest that the clusters and latent factor structures in the data may change through time. This is why, in the rest of the analysis, rolling time window estimates of the latent factors and clusters are considered, and out-of-sample data from subsequent future windows is used to measure the quality of the model estimates.

5.2.2 Reconstruction of out-of-sample returns

To validate the proposed approach on the available financial market data, a backtesting framework is considered. The model is estimated on return time series $\boldsymbol{X}$ on look-back windows of fixed length T. Using the estimated model (mainly, the factor loadings $\hat{\boldsymbol{A}} = [\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}]$), a reconstruction of any time series $\boldsymbol{X}'$ can be obtained using the $N \times N$ filtering matrix of rank $P+Q$:

$$\hat{\boldsymbol{M}} = \hat{\boldsymbol{A}}(\hat{\boldsymbol{A}}^\top \hat{\boldsymbol{A}})^{-1} \hat{\boldsymbol{A}}^\top. \tag{5.8}$$

This enables a reconstruction of any out-of-sample time series $\boldsymbol{X}'$ to be obtained, using the in-sample loadings estimates from which $\hat{\boldsymbol{M}}$ is calculated:

$$\hat{\boldsymbol{X}}' = \boldsymbol{X}' \hat{\boldsymbol{M}}. \tag{5.9}$$

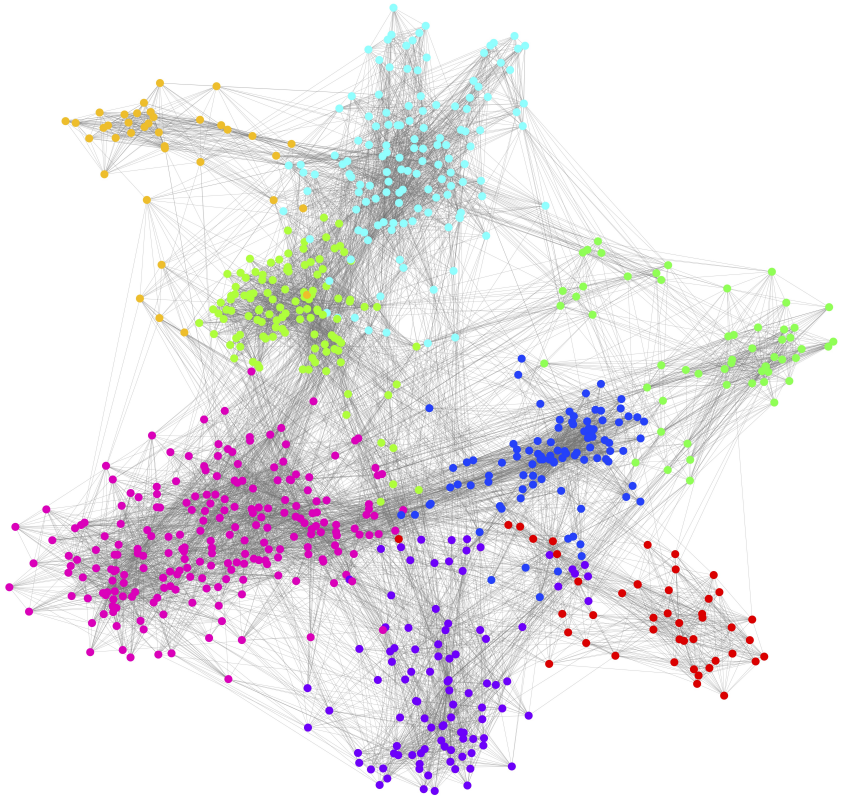


Figure 5.5: The asset graph for NASDAQ indices between 2007 and 2009.

Using the reconstructed time series $\hat{\boldsymbol{X}}$, the unexplained variance in each security is calculated (either for the in-sample or out-of-sample data):

$$V_i = \frac{\sum_t^T (X_{ti} - \hat{X}_{ti})^2}{\sum_t^T (X_{ti} - \overline{X}_i)^2}, \tag{5.10}$$

where X_{ti} is the realization of time series i at time t, $\hat{X}_{ti}$ is the model reconstruction given by (5.9), and $\overline{X}_i$ is the sample mean of time series i. This framework enables one to apply cross-validation principles for estimating the out-of-sample model performance. Specifically, the model estimates from a look-back window of length T are used to reconstruct the future out-of-sample returns on a look-ahead window of length T'. Using these, the average unexplained variance $V = \frac{1}{N}\sum_i^N V_i$ can be measured for both in-sample and out-of-sample data (denoted V_i and V_i', respectively). Thus, ultimately the out of sample unexplained variance V' is the measure to use to compare the estimated latent factor models. In the results presented below, the proposed model-based estimation method is compared to the PC estimator, where the number of components for the PC estimator is selected so that it explains at least the amount of variance explained by the proposed model (both measured in-sample). However, since more factors will always explain more variance in-sample, and often out-of-sample as well, to compliment the out-of-sample performance measure, the deterioration between the in-sample and out-of-

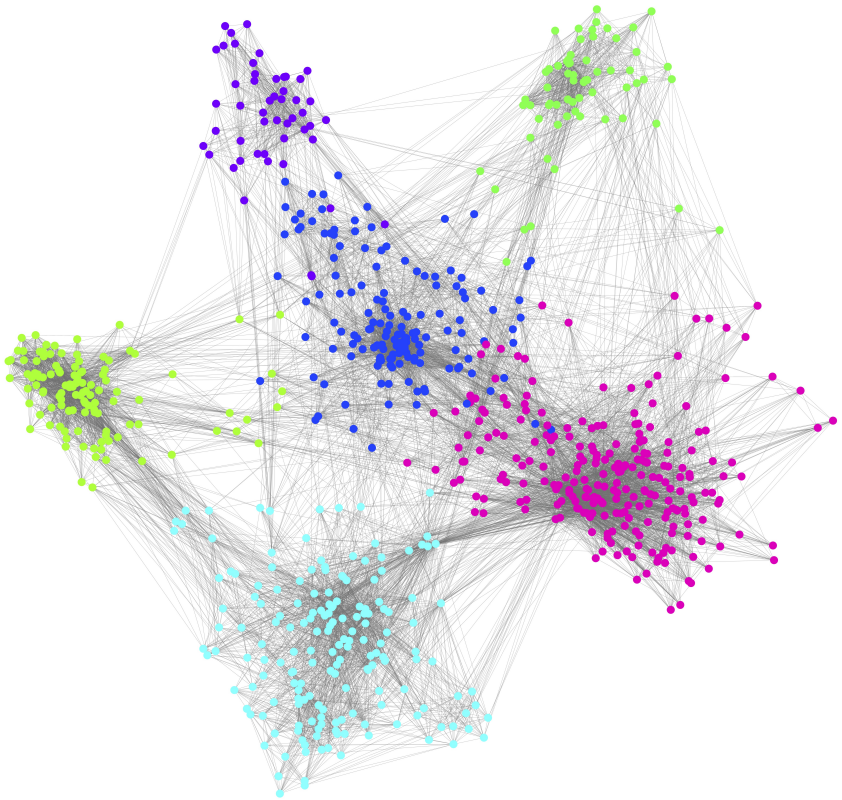


Figure 5.6: The asset graph for NASDAQ indices between 2010 and 2020.

sample unexplained variance is also calculated: $d = 1/N \sum_i^N V_i'/V_i$.

The results in Table 5.5 demonstrate that the proposed approach finds relevant estimates of latent factors in the data which outperform the PC estimates in ouf-of-sample data, for both considered datasets. Even though the PC estimator yields the lowest in-sample unexplained variance V, the PC estimates deteriorate much more than the proposed model, as seen in the out-of-sample unexplained variance V' and the average deterioration d. Moreover, all of these results are in line with other econometric and unsupervised learning studies which find that approximately 30-50% of variance in financial data corresponds to idiosyncratic components [36, 50]. In addition, the model performance in terms of unexplained variance deteriorates fairly less than the PC estimates, suggesting that the proposed estimation method finds more persistent and relevant latent factors in high-dimensional financial time series. In other words, the proposed model-based estimation method generalizes very well to out-of-sample data. This result is expected since the proposed method utilizes the assumed clustering structures within the markets to reduce the number of parameters, thus providing a type of structural regularization of the estimates.

Table 5.5: Unexplained variances of the model estimates compared to the PC estimator given different lengths of the look-back windows, on both considered datasets.

	$T = 4$ years, $T' = 1$ year					
	NASDAQ data			MSCI data		
	V	V'	d	V	V'	d
PC	**31.03%**	42.80%	39.33%	**60.94%**	72.80%	19.90%
Model	31.67%	**40.47%**	**29.00%**	61.90%	**70.99%**	**15.02%**
	$T = 3$ years, $T' = 1$ year					
	NASDAQ data			MSCI data		
	V	V'	d	V	V'	d
PC	**30.34%**	42.86%	44.06%	**60.41%**	72.66%	20.80%
Model	31.15%	**39.65%**	**29.48%**	61.00%	**70.05%**	**15.19%**
	$T = 2$ years, $T' = 1$ year					
	NASDAQ data			MSCI data		
	V	V'	d	V	V'	d
PC	**33.25%**	45.42%	40.06%	**61.82%**	73.81%	19.78%
Model	34.20%	**42.84%**	**28.03%**	63.63%	**71.57%**	**12.88%**

5.2.3 Portfolio backtests

The proposed approach is applied to the two considered datasets in order to test the properties of risk-based portfolios formed using the model estimates. Several look-back periods of $T = 4$, 3, and 2 years are used, with annual rebalancing. Since the two considered datasets contain different profiles of risk and diversification potential, the portfolio optimization method is compared to the EW portfolio of the securities in the considered dataset. Firstly, to look into the dynamics of the estimated number of latent factors, the estimates $\hat{P}$ and $\hat{K}$ for each rebalance time are displayed in Figure 5.7 for the NASDAQ dataset and in Figure 5.8 for the MSCI dataset.

The number of clusters for both datasets vary between 6 and 12, with the number of pervasive factors 0-3. However, the number of pervasive factors in the MSCI dataset is consistently estimated to be zero – this is mostly due to a large cluster of Japanese stocks which were mostly uncorrelated with the rest of the world stocks during the considered time period. Moreover, the total number of cluster-specific factors $\hat{Q}$ is often higher than the number of clusters for the NASDAQ dataset, meaning that certain clusters have multiple cluster-specific factors – on the other hand, the clusters in the MSCI dataset have one cluster-specific factor each.

Using the considered look-back window lengths the risk-based portfolios are calculated (for the different estimates of the covariance matrices), together with the proposed cluster-based portfolio, and compared to the EW portfolio. The annualized return and volatility are calculated

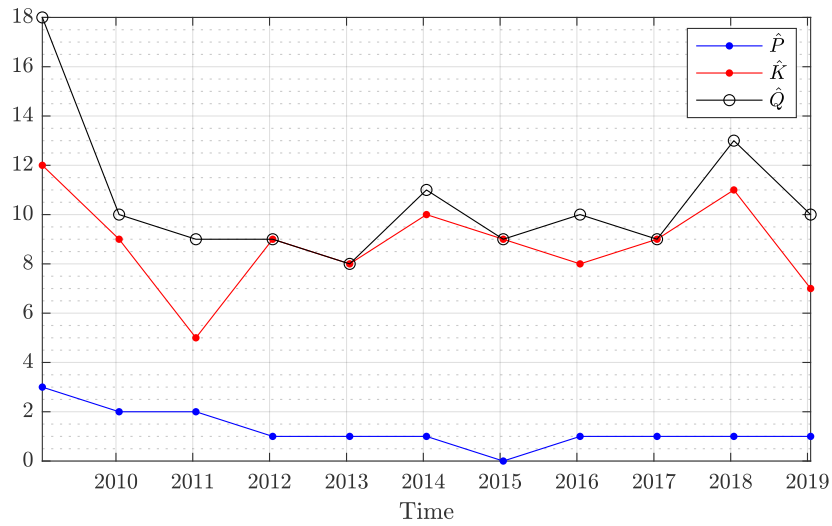


Figure 5.7: The estimated number of pervasive factors $\hat{P}$, clusters $\hat{K}$, and the total number of cluster-specific factors $Q = \sum \hat{C}_k$, for the NASDAQ indices dataset.

from the mean weekly return r_{w} and volatility σ_{w} as: $r_{\text{ann}} = r_{\text{w}} \cdot 52$ and $\sigma_{\text{ann}} = \sigma_{\text{w}} \cdot \sqrt{52}$ (since there are 52 weeks in a year). The (annualized) Sharpe ratio S is then calculated using the annualized returns and volatilities. The risk free rates used in the Sharpe ratio calculations are the 3-month T-Bill rates, obtained at the Federal Reserve Bank of St. Louis website [168]. The maximum drawdown of each portfolio is also calculated as the maximum percentage loss from peak to bottom in the portfolio value throughout the entire time period. The turnover of a portfolio is the average percentage of assets traded in order to rebalance the portfolio (to achieve the target weights given by the portfolio optimization algorithm) – generally large turnovers incur larger trading costs.

The portfolio performance measures for the considered portfolios are given in Table 5.6 for the NASDAQ dataset and in Table 5.7 for the MSCI dataset.

Firstly, the results indicate that all the optimal risk-based portfolios reduce the volatility in comparison to the EW portfolios. This means that the inferences made on the security covariance matrices can indeed help manage and reduce the portfolio risk. The same is also confirmed by the maximum drawdowns, which are evidently the largest in the EW portfolios. However, the reduced risk also comes with a reduction in portfolio returns – nevertheless, as indicated by the Sharpe ratios, nearly all the portfolios outperform the EW portfolio in risk-adjusted terms.

Within the minimum variance portfolio, in both datasets and all considered look-back window lengths the portfolios using the model-based estimates of the covariance matrix have the lowest variances and smallest maximum drawdowns. These additionally affirm the findings in Table 5.5 and suggest that the proposed model yields the most reliable estimates of the security covariance matrices which ultimately reduce the out-of-sample risk in minimum variance portfolios. Interestingly, the risk of MV portfolios estimated using empirical covariances seems to

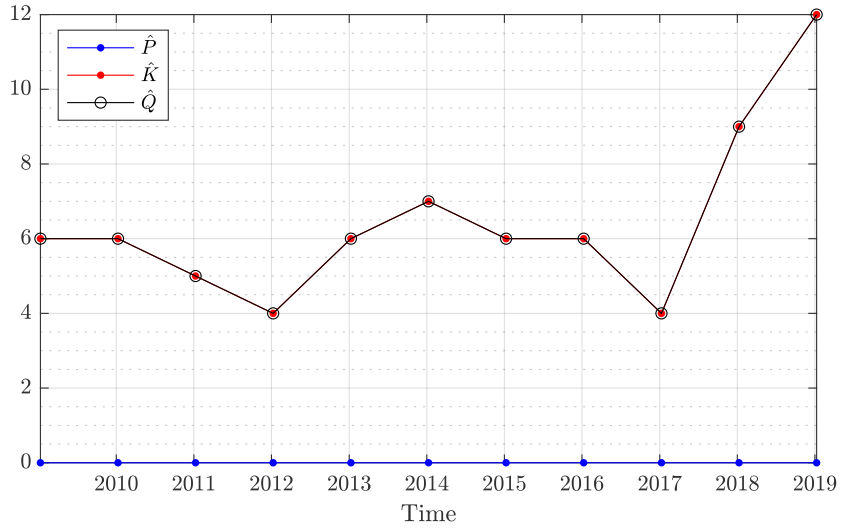


Figure 5.8: The estimated number of pervasive factors $\hat{P}$, clusters $\hat{K}$, and the total number of cluster-specific factors $Q = \sum \hat{C}_k$, for the MSCI stocks dataset.

somewhat increase with the shorter look-back windows. However, the proposed model-based estimates manage to keep the risk relatively low and stable, and thus enable the usage of short estimation windows in order to remain adaptive to the changes in market dynamics. Finally, the turnovers of all the MV turnovers of these portfolios remain very low and any trading costs would be negligible.

Regarding the maximum diversification portfolios and the proposed cluster-based portfolio, the results suggest that these exhibit improved risk-adjusted returns (as opposed to MV portfolios which do reduce risk the most, but also exhibit relatively low returns) – especially the cluster-based portfolio. The results indicate that the proposed cluster-based portfolios outperform the other considered methods in most of the considered cases (the only exception being the 4-year look-back period for the MSCI dataset, where the MDP with the model-based estimate outperforms). Generally, the cluster-based portfolio exhibits high returns with relatively low risk (as indicated by their volatility and max. drawdown), and very low turnovers, which makes the proposed method a very attractive risk-based portfolio optimization approach.

These results demonstrate the validity of the model and the ability of the proposed estimator to provide reliable estimates of latent factors in high-dimensional financial returns. The model-based covariances are shown to reduce the volatility and drawdowns of risk-based portfolios, as opposed to the empirical covariances and the PC estimator. Moreover, the proposed cluster-based portfolio relying on the model estimates and identified security clusters is demonstrated to outperform other risk-based portfolios in risk-adjusted terms, which additionally affirms the proposed approach, which can ultimately help manage and reduce the risk in large portfolios of financial securities.

Table 5.6: The portfolio performance for the considered long-only portfolios, for the proposed latent factor model with pervasive and cluster-specific factors, calculated on the NASDAQ indices dataset.

	$T = 2$ years, $T' = 1$ year				
	r	σ	S	Max. DD	T/o
EW	8.48%	13.84%	0.56	31.50%	3.57%
MV (emp.)	5.77%	7.58%	0.66	24.45%	6.20%
MV (PC)	6.12%	7.50%	0.72	23.63%	6.11%
MV (model)	6.15%	7.20%	0.75	23.20%	6.34%
MDP (emp.)	6.58%	9.38%	0.62	25.85%	6.57%
MDP (PC)	7.26%	9.03%	0.72	24.92%	6.07%
MDP (model)	7.29%	8.46%	0.78	22.70%	6.68%
Cluster-based	8.95%	9.24%	0.89	24.90%	8.94%
	$T = 3$ years, $T' = 1$ year				
	r	σ	S	Max. DD	T/o
EW	8.48%	13.84%	0.56	31.50%	3.57%
MV (emp.)	5.85%	7.28%	0.70	24.29%	6.18%
MV (PC)	5.90%	7.21%	0.72	23.86%	5.46%
MV (model)	5.89%	6.98%	0.74	23.42%	6.67%
MDP (emp.)	7.47%	9.33%	0.72	26.61%	7.44%
MDP (PC)	7.66%	9.02%	0.77	25.61%	6.31%
MDP (model)	7.00%	8.68%	0.72	24.66%	6.56%
Cluster-based	9.58%	9.19%	0.96	25.69%	8.90%
	$T = 4$ years, $T' = 1$ year				
	r	σ	S	Max. DD	T/o
EW	8.48%	13.84%	0.56	31.50%	3.57%
MV (emp.)	5.75%	7.06%	0.71	24.88%	5.66%
MV (PC)	6.04%	6.99%	0.76	24.08%	5.08%
MV (model)	6.54%	6.93%	0.84	23.71%	5.24%
MDP (emp.)	8.11%	9.29%	0.79	25.78%	6.59%
MDP (PC)	7.94%	9.13%	0.79	25.07%	6.69%
MDP (model)	7.85%	8.58%	0.83	22.62%	7.69%
Cluster-based	9.71%	8.66%	1.04	23.35%	8.62%

Table 5.7: The portfolio performance for the considered long-only portfolios, for the proposed latent factor model with pervasive and cluster-specific factors, calculated on the MSCI stocks dataset.

	$T = 2$ years, $T' = 1$ year				
	r	σ	S	Max. DD	T/o
EW	10.99%	14.38%	0.71	30.89%	7.40%
MV (emp.)	8.01%	10.36%	0.70	23.74%	10.09%
MV (PC)	6.82%	9.37%	0.65	20.32%	8.78%
MV (model)	7.48%	8.99%	0.75	17.93%	10.12%
MDP (emp.)	12.07%	12.41%	0.91	25.84%	12.42%
MDP (PC)	7.49%	10.68%	0.63	23.52%	10.12%
MDP (model)	12.20%	10.88%	1.05	18.93%	12.69%
Cluster-based	11.92%	10.87%	1.03	21.11%	12.00%
	$T = 3$ years, $T' = 1$ year				
	r	σ	S	Max. DD	T/o
EW	10.99%	14.38%	0.71	30.89%	7.40%
MV (emp.)	8.89%	9.51%	0.86	21.52%	8.22%
MV (PC)	9.14%	9.29%	0.90	19.52%	7.80%
MV (model)	8.32%	8.80%	0.86	16.82%	7.77%
MDP (emp.)	11.29%	12.47%	0.85	27.45%	11.83%
MDP (PC)	11.66%	11.38%	0.96	23.38%	11.11%
MDP (model)	12.06%	10.78%	1.05	18.26%	11.76%
Cluster-based	12.50%	10.71%	1.10	18.92%	11.58%
	$T = 4$ years, $T' = 1$ year				
	r	σ	S	Max. DD	T/o
EW	10.99%	14.38%	0.71	30.89%	7.40%
MV (emp.)	8.86%	9.30%	0.87	21.64%	7.04%
MV (PC)	8.18%	9.04%	0.82	18.63%	6.86%
MV (model)	8.01%	8.79%	0.83	18.37%	8.03%
MDP (emp.)	13.50%	12.45%	1.02	22.86%	12.11%
MDP (PC)	11.02%	11.09%	0.93	22.38%	10.45%
MDP (model)	14.11%	11.78%	1.14	20.58%	11.93%
Cluster-based	13.11%	11.52%	1.07	21.39%	12.15%

Chapter 6

Conclusion

The growing number in financial securities and the globalization of financial markets motivate an increasing amount of attention to high-dimensional risk modelling by the financial industry and the academic community. Due to the dynamic nature of financial markets, the estimation windows for risk models must remain limited, but the dimensionality of the models grows with the number of securities. Standard statistical tools often fall short due to the so-called *curse of dimensionality*, while complex nonlinear models come with large numbers of parameters which are near impossible to efficiently estimate in high-dimensional settings. Recent research efforts thus turn their attention to regularized methods which exploit certain structural characteristics of financial markets in order to obtain parsimonious and robust estimates.

This thesis focuses on high-dimensional financial time series with pervasive and cluster-specific factors. An estimation method is proposed which performs time series clustering and estimates latent pervasive and cluster-specific factors iteratively. In order to estimate the unknown number of clusters and latent pervasive and cluster-specific factors, a model selection method is developed based on the asset correlation matrices and security graphs. Using the estimated latent factor structures in high-dimensional time series of asset returns, a risk-based portfolio optimization method relying on latent factor and cluster estimates is also proposed.

The methods are tested using several data generating processes under the approximate factor model assumptions, featuring heavy tailed returns with some off-diagonal correlations of residuals. The simulation study shows that the proposed method yields very accurate clustering results, even for the most severe high-dimensional setting and heavy-tailed distributions. Moreover, the results demonstrate that the proposed two-phase model-based method estimates clusters which are not biased towards those clusters with a larger number of cluster-specific factors, as is the case with other clustering methods using cluster-specific factors. In addition, the simulation study results suggest that the proposed model selection method provides stable and accurate estimates of the number of clusters, latent pervasive, and latent cluster-specific factors.

The methods are also applied to datasets of return time series of NASDAQ indices and world

stocks in MSCI indices in a backtesting approach which allows the in-sample model estimates to be used for the reconstruction of out-of-sample return data. By doing so the unexplained variance can be cross-validated. The result of the out-of-sample unexplained variances suggest that the proposed model-based estimation method, although explaining less variance in-sample than the PC estimator, explains more variance out-of-sample, meaning that it generalizes better and provides more robust estimates. In addition, the proposed portfolio optimization method based on the estimated latent factors is also backtested on historical market returns. The results demonstrate the ability of the proposed method to reduce risk in the minimum variance portfolios, which outperform the portfolios built on empirical and PC estimates of the covariance matrix. Moreover, it is found that, whereas the empirical covariances deteriorate with the shorter look-back windows, the model-based estimates thrive in these high-dimensional situations, allowing one to use short look-back windows and thus being more adaptive to changing market conditions. Moreover, the considered maximum diversification portfolios affirm these findings – the model-based covariance estimates yield the best risk-adjusted performance of the MDP portfolios in both considered datasets and all look-back windows. Finally, the proposed cluster-based portfolios outperform other risk-based methods, while keeping the risk low and turnover negligible.

The results presented in this paper suggest that the clustering assumption in high-dimensional financial time series data holds, and that the model-based estimation method indeed extracts useful information about the latent factor structure. These findings affirm and refine asset pricing theories based on multi-factor models, providing evidence on the clustering structures of latent risk factors. This approach may help shed more light on the intricate latent factor structures in global financial markets, as is demonstrated in our results. Ultimately, the robust estimates of pervasive and cluster-specific factors may be used to improve risk assessment and enhance the out-of-sample performance of portfolios built on the estimated models.

Bibliography

[1] D. Johnston and P. Djurić, "The Science Behind Risk Management," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 26–36, sep 2011.

[2] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812–820, jun 2015.

[3] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Systems with Applications*, vol. 83, pp. 405–417, oct 2017.

[4] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, feb 1993.

[5] J. Bai and S. Ng, "Evaluating latent and observed factors in macroeconomics and finance," in *Journal of Econometrics*, vol. 131, no. 1-2. North-Holland, mar 2006, pp. 507–537.

[6] E. F. Fama and K. R. French, "International tests of a five-factor asset pricing model," *Journal of Financial Economics*, vol. 123, no. 3, pp. 441–463, mar 2017.

[7] C. Asness, A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen, "Size matters, if you control your junk," *Journal of Financial Economics*, 2018.

[8] M. Lettau and M. Pelger, "Factors That Fit the Time Series and Cross-Section of Stock Returns," *The Review of Financial Studies*, vol. 33, no. 5, 2020.

[9] M. Agrawal, D. Mohapatra, and I. Pollak, "Empirical evidence against CAPM: Relating alphas and returns to betas," *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 4, pp. 298–310, 2012.

[10] X.-P. S. Zhang and F. Wang, "Signal Processing for Finance, Economics, and Marketing: Concepts, framework, and big data applications," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 14–35, may 2017.

[11] M. M. Loépez de Prado, *Advances in financial machine learning*. Wiley, 2018.

[12] D. Wang, B. Podobnik, D. Horvatić, and H. E. Stanley, "Quantifying and modeling long-range cross correlations in multiple time series with applications to world stock indices," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 83, no. 4, 2011.

[13] Z. Zheng, B. Podobnik, L. Feng, and B. Li, "Changes in cross-correlations as an indicator for systemic risk," *Scientific Reports*, vol. 2, no. 1, pp. 1–8, nov 2012.

[14] Z. Kostanjčar, S. Begušić, H. E. Stanley, and B. Podobnik, "Estimating Tipping Points in Feedback-Driven Financial Networks," *IEEE Journal on Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1040–1052, sep 2016.

[15] J. Bun, J. P. Bouchaud, and M. Potters, "Cleaning large correlation matrices: Tools from Random Matrix Theory," *Physics Reports*, vol. 666, pp. 1–109, 2017.

[16] F. Rubio, X. Mestre, and D. P. Palomar, "Performance analysis and optimal selection of large minimum variance portfolios under estimation risk," *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 4, pp. 337–350, 2012.

[17] M. López de Prado, "Building Diversified Portfolios that Outperform Out of Sample," *The Journal of Portfolio Management*, vol. 42, no. 4, pp. 59–69, jul 2016.

[18] F. G. Duarte and L. N. De Castro, "A Framework to Perform Asset Allocation Based on Partitional Clustering," *IEEE Access*, vol. 8, pp. 110 775–110 788, jun 2020.

[19] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Systems with Applications*, vol. 83, pp. 187–205, oct 2017.

[20] J. Bai and S. Ng, "Rank regularized estimation of approximate factor models," *Journal of Econometrics*, apr 2019.

[21] R. Cont, "Empirical properties of asset returns: Stylized facts and statistical issues," *Quantitative Finance*, vol. 1, no. 2, pp. 223–236, feb 2001.

[22] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, dec 2000.

[23] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, jan 2007.

[24] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[25] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[26] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for Multi-Cluster data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, New York, USA: ACM Press, 2010, pp. 333–342.

[27] N. Kambhatla and T. K. Leen, "Dimension Reduction by Local Principal Component Analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, oct 1997.

[28] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, mar 2002.

[29] X. He, D. Cai, and P. Niyogi, "Laplacian Score for feature selection," in *Advances in Neural Information Processing Systems*, 2005, pp. 507–514.

[30] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *ACM International Conference Proceeding Series*, vol. 227. New York, New York, USA: ACM Press, 2007, pp. 1151–1157.

[31] D. Huang, X. Cai, and C. D. Wang, "Unsupervised feature selection with multi-subspace randomization and collaboration," *Knowledge-Based Systems*, vol. 182, p. 104856, oct 2019.

[32] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," pp. 2507–2517, oct 2007.

[33] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[34] M. Gong, M. Zhang, and Y. Yuan, "Unsupervised Band Selection Based on Evolutionary Multiobjective Optimization for Hyperspectral Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 544–557, jan 2016.

[35] J. H. Cochrane, *Asset pricing*. Princeton University Press, 2005.

[36] G. Connor, "The Three Types of Factor Models: A Comparison of Their Explanatory Power," *Financial Analysts Journal*, vol. 51, no. 3, pp. 42–46, may 1995.

[37] C. Lam, Q. Yao, and N. Bathia, "Estimation of latent factors for high-dimensional time series," *Biometrika*, vol. 98, no. 4, pp. 901–918, dec 2011.

[38] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 2018.

[39] S. Gu, B. Kelly, and D. Xiu, "Autoencoder asset pricing models," *Journal of Econometrics*, jul 2020.

[40] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Annals of Statistics*, 2015.

[41] Y. Aït-Sahalia and D. Xiu, "Using principal component analysis to estimate a high dimensional factor model with high-frequency data," *Journal of Econometrics*, vol. 201, no. 2, pp. 384–399, dec 2017.

[42] M. MacMahon and D. Garlaschelli, "Community Detection for Correlation Matrices," *Physical Review X*, vol. 5, no. 2, p. 021006, apr 2015.

[43] S. Begušić, Z. Kostanjčar, D. Kovač, H. E. Stanley, and B. Podobnik, "Information Feedback in Temporal Networks as a Predictor of Market Crashes," *Complexity*, vol. 2018, pp. 1–13, sep 2018.

[44] M. Tumminello, F. Lillo, and R. N. Mantegna, "Correlation, hierarchies, and networks in financial markets," *Journal of Economic Behavior and Organization*, 2010.

[45] Z. Kakushadze and W. Yu, "Statistical Industry Classification," *Journal of Risk & Control*, vol. 3, no. 1, pp. 17–65, jun 2017.

[46] A. Verma, R. J. Buonocore, and T. Di Matteo, "A cluster driven log-volatility factor model: a deepening on the source of the volatility clustering," *Quantitative Finance*, pp. 1–16, nov 2018.

[47] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna, "Cluster analysis for portfolio optimization," *Journal of Economic Dynamics and Control*, 2008.

[48] S. Begušić and Z. Kostanjčar, "Cluster-Based Shrinkage of Correlation Matrices for Portfolio Optimization," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, sep 2019, pp. 301–305.

[49] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, dec 2003.

[50] R. Clarke, H. De Silva, and S. Thorley, "Minimum-variance portfolio composition," *Journal of Portfolio Management*, vol. 37, no. 2, pp. 31–45, 2011.

[51] T. Ando and J. Bai, "Panel Data Models with Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics*, vol. 31, no. 1, pp. 163–191, jan 2016.

[52] ——, "Clustering Huge Number of Financial Time Series: A Panel Data Approach With High-Dimensional Predictors and Factor Structures," *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 1182–1198, 2017.

[53] R. Cont, "Statistical modeling of high-frequency financial data," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 16–25, sep 2011.

[54] E. F. Fama, *Foundations of finance: portfolio decisions and securities prices*. Basic Books, 1976.

[55] A. Meucci, "Quant Nugget 2: Linear vs. Compounded Returns – Common Pitfalls in Portfolio Management," *GARP Risk Professional*, pp. 49–51, may 2010.

[56] A. Pagan, "The econometrics of financial markets," *Journal of Empirical Finance*, vol. 3, no. 1, pp. 15–102, may 1996.

[57] A. Corhay and A. T. Rad, "Statistical Properties of Daily Returns: Evidence from European Stock Markets," *Journal of Business Finance & Accounting*, vol. 21, no. 2, pp. 271–282, mar 1994.

[58] D. A. Hsieh, "The statistical properties of daily foreign exchange rates: 1974-1983," *Journal of International Economics*, vol. 24, no. 1-2, pp. 129–145, feb 1988.

[59] A. F. Bariviera, M. J. Basgall, W. Hasperué, and M. Naiouf, "Some stylized facts of the Bitcoin market," *Physica A: Statistical Mechanics and its Applications*, vol. 484, pp. 82–90, 2017.

[60] D. Sornette, "Critical market crashes," *Physics Reports*, vol. 378, no. 1, pp. 1–98, 2003.

[61] J. P. Bouchaud, "Power laws in economics and finance: Some ideas from physics," *Quantitative Finance*, vol. 1, no. 1, pp. 105–112, 2001.

[62] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley, "Cross-correlations between volume change and price change." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 52, pp. 22 079–84, dec 2009.

[63] S. Maslov, "Simple model of a limit order-driven market," *Physica A: Statistical Mechanics and its Applications*, vol. 278, no. 3-4, pp. 571–578, apr 2000.

[64] A. Chakraborti, I. M. Toke, M. Patriarca, and F. Abergel, "Econophysics review: I. Empirical facts," pp. 991–1012, jul 2011.

[65] D. Sornette, "Physics and financial economics (1776-2014): puzzles, Ising and agent-based models." *Reports on progress in physics. Physical Society (Great Britain)*, vol. 77, no. 6, p. 062001, jun 2014.

[66] B. LeBaron, "Some Relations Between Volatility and Serial Correlations in Stock Market Returns," *The Journal of Business*, vol. 65, no. 2, p. 199, jan 1992.

[67] D. Avramov, S. Cheng, and A. Hameed, "Time-Varying Liquidity and Momentum Profits," *Journal of Financial and Quantitative Analysis*, vol. 51, no. 06, pp. 1897–1923, dec 2016.

[68] S. Bianco, F. Corsi, and R. Renò, "Intraday LeBaron effects," *Proceedings of the National Academy of Sciences*, vol. 106, no. 28, pp. 11 439–11 443, jul 2009.

[69] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, p. 383, may 1970.

[70] A. Pole, *Statistical arbitrage : algorithmic trading insights and techniques*. J. Wiley & Sons, 2007.

[71] M. Avellaneda and J.-H. Lee, "Statistical arbitrage in the US equities market," *Quantitative Finance*, vol. 10, no. 7, pp. 761–782, aug 2010.

[72] L. Mrčela, A. Merćep, S. Begušić, and Z. Kostanjčar, "Portfolio optimization using preference relation based on statistical arbitrage," *2017 International Conference on Smart Systems and Technologies (SST)*, pp. 161–165, 2017.

[73] R. Cont, "Volatility clustering in financial markets: Empirical facts and agent-based models," in *Long Memory in Economics*. Springer Berlin Heidelberg, 2007, pp. 289–309.

[74] T. Lux and M. Marchesi, "Volatility Clustering in Financial Markets: a Microsimulation of Interacting Agents," *International Journal of Theoretical and Applied Finance*, vol. 03, no. 04, pp. 675–702, oct 2000.

[75] L. Feng, B. Li, B. Podobnik, T. Preis, and H. E. Stanley, "Linking agent-based models and stochastic models of financial markets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 22, pp. 8388–8393, 2012.

[76] A. W. Lo, "Long-Term Memory in Stock Market Prices," *Econometrica*, vol. 59, no. 5, p. 1279, sep 1991.

[77] J. Bai and S. Shi, "Estimating high dimensional covariance matrices and its applications," pp. 199–215, nov 2011.

[78] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of Financial Economics*, vol. 104, no. 3, pp. 535–559, 2012.

[79] S. Chib, F. Nardari, and N. Shephard, "Analysis of high dimensional multivariate stochastic volatility models," *Journal of Econometrics*, vol. 134, no. 2, pp. 341–371, oct 2006.

[80] H. M. Markowitz, "Portfolio Selection," *Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[81] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, 1973.

[82] T. Bollerslev and V. Todorov, "Tails, Fears, and Risk Premia," *Journal of Finance*, vol. 66, no. 6, pp. 2165–2211, dec 2011.

[83] B. Kelly and H. Jiang, "Tail risk and asset prices," *Review of Financial Studies*, vol. 27, no. 10, 2014.

[84] B. Mandelbrot, A. Fisher, and L. Calvet, "A Multifractal Model of Asset Returns," *Working Papers – Yale School of Management's Economics Research Network*, no. October 1996, 1997.

[85] P. Gopikrishnan, M. Meyer, L. A. Amaral, and H. E. Stanley, "Inverse cubic law for the distribution of stock price variations," *European Physical Journal B*, vol. 3, no. 2, pp. 139–140, 1998.

[86] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[87] V. Plerou, P. Gopikrishnan, L. A. N. Amaral, M. Meyer, and H. E. Stanley, "Scaling of the distribution of price fluctuations of individual companies," *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, vol. 60, no. 6 Pt A, pp. 6519–29, dec 1999.

[88] P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, "Scaling of the distribution of fluctuations of financial market indices," *Physical Review E*, vol. 60, no. 5, pp. 5305–5316, nov 1999.

[89] S. Begušić, Z. Kostanjčar, H. Eugene Stanley, and B. Podobnik, "Scaling properties of extreme price fluctuations in Bitcoin markets," *Physica A: Statistical Mechanics and its Applications*, vol. 510, pp. 400–406, jul 2018.

[90] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, "A theory of power-law distributions in financial market fluctuations," *Nature*, vol. 423, no. 6937, pp. 267–270, may 2003.

[91] X. Gabaix, "Power Laws in Economics and Finance," *Annual Review of Economics*, vol. 1, no. 1, pp. 255–294, 2009.

[92] J. D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen, "What really causes large price changes?" *Quantitative Finance*, vol. 4, no. 4, pp. 383–397, 2004.

[93] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: Concepts, techniques, and tools*. Princeton University Press, 2005.

[94] L. Bauwens, S. Laurent, and J. V. Rombouts, "Multivariate GARCH models: A survey," pp. 79–109, jan 2006.

[95] R. Engle, "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business and Economic Statistics*, vol. 20, no. 3, pp. 339–350, 2002.

[96] R. F. Engle, O. Ledoit, and M. Wolf, "Large Dynamic Covariance Matrices," 2017.

[97] V. Plerou, P. Gopikrishnan, B. Rosenow, L. a. N. Amaral, T. Guhr, and H. E. Stanley, "A Random Matrix Approach to Cross-Correlations in Financial Data," *Physical Review E*, 2001.

[98] V. A. Marčenko and L. A. Pastur, "Distribution of Eigenvalues for Some Sets of Random Matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, apr 1967.

[99] G. Biroli, J. P. Bouchaud, and M. Potters, "On the top eigenvalue of heavy-tailed random matrices," *EPL*, vol. 78, no. 1, p. 10001, mar 2007.

[100] A. M. Sengupta and P. P. Mitra, "Distributions of singular values for some random matrices," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 60, no. 3, pp. 3389–3392, sep 1999.

[101] J. Fan, J. Lv, and L. Qi, "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, vol. 3, no. 1, pp. 291–317, 2011.

[102] Mohsen Pourahmadi, *High-Dimensional Covariance Estimation*, ser. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., jun 2013.

[103] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 2004.

[104] G. Akemann, J. Baik, and P. Di Francesco, *The Oxford handbook of random matrix theory*. Oxford University Press, 2011.

[105] O. Ledoit and M. Wolf, "Honey, I shrunk the sample covariance matrix," pp. 110–119+7, jul 2004.

[106] W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *The Journal of Finance*, vol. 19, no. 3, p. 425, sep 1964.

[107] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, "Noise Dressing of Financial Correlation Matrices," *Physical Review Letters*, 1999.

[108] J. Bun, J.-P. Bouchaud, and M. Potters, "Cleaning correlation matrices," *Risk Magazine*, vol. 2015, no. April, 2015.

[109] J. Bun, R. Allez, J. P. Bouchaud, and M. Potters, "Rotational invariant estimator for general noisy matrices," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7475–7490, 2016.

[110] S. Deshmukh and A. Dubey, "Improved Covariance Matrix Estimation with an Application in Portfolio Optimization," *IEEE Signal Processing Letters*, vol. 27, pp. 985–989, 2020.

[111] A. Ang, R. J. Hodrick, Y. Xing, and X. Zhang, "The cross-section of volatility and expected returns," *Journal of Finance*, vol. 61, no. 1, pp. 259–299, feb 2006.

[112] N.-F. Chen, R. Roll, and S. A. Ross, "Economic Forces and the Stock Market," *Journal of Business*, vol. 59, no. 3, pp. 383–403, 1986.

[113] L. X. Liu and L. Zhang, "Momentum profits, factor pricing, and macroeconomic risk," *Review of Financial Studies*, vol. 21, no. 6, pp. 2417–2448, nov 2008.

[114] E. F. Fama and K. R. French, "Multifactor explanations of asset pricing anomalies," *Journal of Finance*, vol. 51, no. 1, pp. 55–84, mar 1996.

[115] S. A. Ross, "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, vol. 13, no. 3, pp. 341–360, dec 1976.

[116] E. F. Fama and K. R. French, "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, vol. 47, no. 2, pp. 427–465, jun 1992.

[117] N. Jegadeesh and S. Titman, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *The Journal of Finance*, vol. 48, no. 1, p. 65, mar 1993.

[118] M. M. Carhart, "On persistence in mutual fund performance," *Journal of Finance*, vol. 52, no. 1, pp. 57–82, mar 1997.

[119] E. F. Fama and K. R. French, "Size, value, and momentum in international stock returns," *Journal of Financial Economics*, vol. 105, no. 3, pp. 457–472, sep 2012.

[120] ——, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1–22, apr 2015.

[121] BARRA Inc., "United States Equity," 1998.

[122] D. C. Blitz and P. Van Vliet, "Global tactical cross-asset allocation: Applying value and momentum across asset classes," *Journal of Portfolio Management*, vol. 35, no. 1, pp. 23–38, sep 2008.

[123] D. Shen, A. Urquhart, and P. Wang, "A three-factor pricing model for cryptocurrencies," *Finance Research Letters*, jul 2019.

[124] D. Blitz, M. X. Hanauer, M. Vidojevic, and P. Van Vliet, "Five concerns with the five-factor model," *Journal of Portfolio Management*, vol. 44, no. 4, pp. 71–78, mar 2018.

[125] G. Feng, S. Giglio, and D. Xiu, "Taming the Factor Zoo: A Test of New Factors," *Journal of Finance*, vol. 75, no. 3, pp. 1327–1370, jun 2020.

[126] V. V. Acharya and L. H. Pedersen, "Asset pricing with liquidity risk," *Journal of Financial Economics*, vol. 77, no. 2, pp. 375–410, aug 2005.

[127] G. Connor and R. A. Korajczyk, "Factor Models in Portfolio and Asset Pricing Theory," in *Handbook of Portfolio Construction*. Springer US, 2010, pp. 401–418.

[128] E. E. Cureton and R. B. D'Agostino, *Factor analysis: An applied approach*. Psychology press, 1983.

[129] K. Coughlin, "An Analysis of Factor Extraction Strategies: A Comparison of the Relative Strengths of Principal Axis, Ordinary Least Squares, and Maximum Likelihood in Research Contexts that Include both Categorical and Continuous Variables," Ph.D. dissertation, University of South Florida, jan 2013.

[130] N. E. Briggs and R. C. MacCallum, "Recovery of weak common factors by maximum likelihood and ordinary least squares estimation," *Multivariate Behavioral Research*, vol. 38, no. 1, pp. 25–56, 2003.

[131] G. Chamberlain and M. Rothschild, "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, vol. 51, no. 5, p. 1281, sep 1983.

[132] J. H. Stock and M. W. Watson, "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, vol. 97, no. 460, 2002.

[133] A. Onatski, "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, vol. 168, no. 2, pp. 244–258, jun 2012.

[134] M. Lettau and M. Pelger, "Estimating latent asset-pricing factors," *Journal of Econometrics*, feb 2020.

[135] J. Bai and S. Ng, "Large Dimensional Factor Analysis," *Foundations and Trends in Econometrics*, vol. 3, no. 2, pp. 89–163, 2008.

[136] A. Onatski, "Determining the number of factors from empirical distribution of eigenvalues," *Review of Economics and Statistics*, vol. 92, no. 4, pp. 1004–1016, nov 2010.

[137] I. Choi and H. Jeong, "Model selection for factor analysis: Some new criteria and performance comparisons," *Econometric Reviews*, vol. 38, no. 6, pp. 577–596, jul 2019.

[138] V. Keranović, S. Begušić, and Z. Kostanjčar, "Estimating the Number of Latent Factors in High-Dimensional Financial Time Series," in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, sep 2020, pp. 1–5.

[139] H. F. Kaiser, "On Cliff's Formula, the Kaiser-Guttman Rule, and the Number of Factors," *Perceptual and Motor Skills*, vol. 74, no. 2, pp. 595–598, apr 1992.

[140] G. Biroli, J. P. Bouchaud, and M. Potters, "The student ensemble of correlation matrices: Eigenvalue spectrum and Kullback-leibler entropy," in *Acta Physica Polonica B*, 2007.

[141] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, jun 1965.

[142] A. Buja and N. Eyuboglu, "Remarks on Parallel Analysis," *Multivariate Behavioral Research*, vol. 27, no. 4, pp. 509–540, oct 1992.

[143] J. C. Hayton, D. G. Allen, and V. Scarpello, "Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis," *Organizational Research Methods*, vol. 7, no. 2, pp. 191–205, apr 2004.

[144] J. Bai and S. Ng, "Determining the number of factors in approximate factor models," *Econometrica*, vol. 70, no. 1, pp. 191–221, jan 2002.

[145] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag New York, 2002.

[146] J. Han, M. Kamber, and J. Pei, *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann, 2012.

[147] A. Onatski, "Testing Hypotheses About the Number of Factors in Large Factor Models," *Econometrica*, vol. 77, no. 5, pp. 1447–1479, 2009.

[148] M. Pelger, "Large-dimensional factor modeling based on high-frequency observations," *Journal of Econometrics*, vol. 208, no. 1, pp. 23–42, jan 2019.

[149] G. Connor and R. A. Korajczyk, "A Test for the Number of Factors in an Approximate Factor Model," *The Journal of Finance*, vol. 48, no. 4, pp. 1263–1291, sep 1993.

[150] S. C. Ahn and A. R. Horenstein, "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, vol. 81, no. 3, pp. 1203–1227, may 2013.

[151] S. Begušić and Z. Kostanjčar, "Cluster-Specific Latent Factor Estimation in High-Dimensional Financial Time Series," *IEEE Access*, vol. 8, pp. 164 365–164 379, sep 2020.

[152] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, dec 2007.

[153] P. Veenstra, C. Cooper, and S. Phelps, "Spectral clustering using the kNN-MST similarity graph," in *2016 8th Computer Science and Electronic Engineering Conference, CEEC 2016 - Conference Proceedings*. Institute of Electrical and Electronics Engineers Inc., jan 2017, pp. 222–227.

[154] T. Cai and W. Liu, "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 2011.

[155] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, feb 2016.

[156] F. J. Fabozzi, H. M. Markowitz, P. N. Kolm, and F. Gupta, "Mean-Variance Model for Portfolio Selection," in *Encyclopedia of Financial Models*. Hoboken, NJ, USA: John Wiley & Sons, Inc., dec 2012.

[157] A. Rigamonti, "Mean-Variance Optimization Is a Good Choice, But for Other Reasons than You Might Think," *Risks*, vol. 8, no. 1, p. 29, mar 2020.

[158] R. Malladi and F. J. Fabozzi, "Equal-weighted strategy: Why it outperforms value-weighted strategies? Theory and evidence," *Journal of Asset Management*, vol. 18, no. 3, pp. 188–208, may 2017.

[159] R. Clarke, H. de Silva, and S. Thorley, "Risk Parity, Maximum Diversification, and Minimum Variance: An Analytic Perspective," *The Journal of Portfolio Management*, vol. 39, no. 3, pp. 39–53, apr 2013.

[160] T. Roncalli, *Introduction to Risk Parity and Budgeting*. Chapman and Hall, 2014.

[161] H. du Plessis and P. van Rensburg, "Risk-based portfolio sensitivity to covariance estimation," *Investment Analysts Journal*, pp. 1–26, oct 2020.

[162] R. G. Clarke, H. de Silva, and S. Thorley, "Minimum-Variance Portfolios in the U.S. Equity Market," *The Journal of Portfolio Management*, vol. 33, no. 1, pp. 10–24, oct 2009.

[163] R. Jagannathan and T. Ma, "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps," *Journal of Finance*, vol. 58, no. 4, pp. 1651–1683, 2003.

[164] E. Pantaleo, M. Tumminello, F. Lillo, and R. N. Mantegna, "When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators," *Quantitative Finance*, vol. 11, no. 7, pp. 1067–1080, jul 2011.

[165] Y. Choueifaty and Y. Coignard, "Toward maximum diversification," *Journal of Portfolio Management*, vol. 35, no. 1, pp. 40–51, sep 2008.

[166] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, dec 2001.

[167] Quandl, "NASDAQ OMX Global Index Data," 2020.

[168] Board of Governors of the Federal Reserve System (US), "3-Month Treasury Bill: Secondary Market Rate [TB3MS]," 2020.

List of Figures

List of Tables

Biography

Stjepan Begušić received his M.Sc. degree in Information and Communication Technology in 2014 from the University of Zagreb,Faculty of Electrical Engineering and Computing. He is a Research Associate at the Laboratory for Financial and Risk Analytics, with the University of Zagreb, Faculty of Electrical Engineering and Computing, where he works on projects funded by the Croatian Science Foundation, EU funds and partners from the industry. His main research interests include statistical and machine learning methods for high-dimensional financial data, risk modelling and portfolio optimization. He is also an assistant for the courses "Statistical Data Analysis" and "Distributed Ledgers and Cryptocurrencies" at the University of Zagreb, Faculty of Electrical Engineering and Computing. He is an IEEE student member since 2012. He has been the reviewer for the *IEEE Journal of Selected Topics in Signal Processing*, *Physica A: Statistical Mechanics and its Applications*, *PLoS One*, and *IEEE Access*, and has published several papers in international journals and conferences.

Publications

Journal papers

1. S. Begušić and Z. Kostanjčar, "Cluster-Specific Latent Factor Estimation in High-Dimensional Financial Time Series," IEEE Access, vol. 8, pp. 164365-164379, Sep. 2020, doi: 10.1109/access.2020.3021898
2. S. Begušić, Z. Kostanjčar, D. Kovač, H. E. Stanley, and B. Podobnik, "Information Feedback in Temporal Networks as a Predictor of Market Crashes," Complexity, vol. 2018, pp. 1–13, Sep. 2018, doi: 10.1155/2018/2834680
3. S. Begušić, Z. Kostanjčar, H. Eugene Stanley, and B. Podobnik, "Scaling properties of extreme price fluctuations in Bitcoin markets," Physica A: Statistical Mechanics and its Applications, vol. 510, pp. 400–406, Jul. 2018, doi: 10.1016/j.physa.2018.06.131
4. Z. Kostanjčar, S. Begušić, H. E. Stanley, and B. Podobnik, "Estimating Tipping Points in Feedback-Driven Financial Networks," IEEE Journal on Selected Topics in Signal Processing, vol. 10, no. 6, pp. 1040–1052, Sep. 2016, doi: 10.1109/JSTSP.2016.2593099

Conference papers

1. V. Keranović, S. Begušić, and Z. Kostanjčar, “Estimating the Number of Latent Factors in High-Dimensional Financial Time Series,” in 2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2020, pp. 1–5, doi: 10.23919/SoftCOM50211.2020.9238229
2. S. Begušić and Z. Kostanjčar, “Cluster-Based Shrinkage of Correlation Matrices for Portfolio Optimization,” in 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), 2019, pp. 301–305, doi: 10.1109/ISPA.2019.8868482
3. M. Puljiz, S. Begušić, and Z. Kostanjčar, “Market Microstructure and Order Book Dynamics in Cryptocurrency Exchanges,” in Crypto Valley Conference on Blockchain Technology, 2018.
4. L. Mrčela, A. Merćep, S. Begušić, and Z. Kostanjčar, “Portfolio optimization using preference relation based on statistical arbitrage,” 2017 International Conference on Smart Systems and Technologies (SST), pp. 161–165, 2017, doi: 10.1109/SST.2017.8188688

Životopis

Stjepan Begušić je diplomirao na studiju Informacijske i komunikacijske tehnologije 2014. godine na Sveučilištu u Zagrebu, Fakultetu elektrotehnike i računarstva. On je znanstveni suradnik u Laboratoriju za analitiku financija i rizika pri Sveučilištu u Zagrebu, Fakultetu elektrotehnike i računarstva, gdje radi na projektima financiranima od strane Hrvatske zaklade za znanost, EU fondova i partnera iz industrije. Njegovi glavni istraživački interesi su statističke metode i metode strojnog učenja za visokodimenzionalne financijske podatke, modeliranje rizika i optimizaciju portfelja. Također je i asistent na predmetima "Statistička analiza podataka" i "Raspodijeljene glavne knjige i kriptovalute" pri Sveučilištu u Zagrebu, Fakultetu elektrotehnike i računarstva. Studentski je član IEEE od 2012. godine. Recenzirao je za časopise *IEEE Journal of Selected Topics in Signal Processing*, *Physica A: Statistical Mechanics and its Applications*, *PLoS One* i *IEEE Access*, te je objavio više radova u međunarodnim časopisima i konferencijama.