Essays on Modern Econometrics and Machine Learning

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Georg Keilbar

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Daniel Klapper

Gutachter:
1. Prof. Dr. Weining Wang, Ph.D.
2. Prof. Dr. Wolfgang Karl Härdle, Ph.D.

Tag des Kolloquiums: 20. August 2021

Acknowledgments

I want to express my deepest gratitude to my principal advisor Professor Weining Wang, who encouraged me to pursue a PhD in statistics. I benefited a lot from her encouragement, support and patience. I am equally indebted to my second advisor, Professor Wolfgang Karl Härdle, for his encouragement to continuously expand my knowledge and his advice that goes beyond research.

During the three years of my PhD studies, I had the privilege to collaborate with many excellent researchers. I want to thank Professor Wei Biao Wu, Professor Likai Chen, Professor Juan Manuel Rodriduez-Poo, Dr. Alexandra Soberon and Yanfen Zhang for sharing their wisdom.

It was a great joy to be a part of the team at the IRTG 1792, where I had the opportunity to work with many great and interesting people. I want to thank my colleagues Daniel Jacob, Junjie Hu, Bingling Wang, Elizaveta Zinovyeva and all the others for the continuous support. Special thanks go to Raphael Reule for the tremendous assistance with all the administrative work.

Very special thanks belong to my family and friends, who supported me all the time during my PhD. Most importantly, I want to thank my girlfriend Binfang.

Finally, financial support from the DFG via the IRTG 1792 "High Dimensional Nonstationary Time Series" is gratefully acknowledged.

Abstract

This thesis focuses on different aspects of the union of modern econometrics and machine learning. Chapter 2 considers a new estimator of the regression parameters in a panel data model with unobservable interactive fixed effects. A distinctive feature of the proposed approach is to model the factor loadings as a nonparametric function. We show that our estimator is $\sqrt{NT}$-consistent and asymptotically normal, as well that it reaches the semiparametric efficiency bound under the assumption of i.i.d. errors. Chapter 3 is concerned with the recursive estimation of quantiles using the stochastic gradient descent (SGD) algorithm with Polyak-Ruppert averaging. The algorithm offers a computationally and memory efficient alternative to the usual empirical estimator. Our focus is on studying the nonasymptotic behavior by providing exponentially decreasing tail probability bounds under minimal assumptions. In Chapter 4 we propose a novel approach to calibrate the conditional value-at-risk (CoVaR) of financial institutions based on neural network quantile regression. We model systemic risk spillover effects in a network context across banks by considering the marginal effects of the quantile regression procedure. An out-of-sample analysis shows great performance compared to a linear baseline specification, signifying the importance that nonlinearity plays for modelling systemic risk. A comparison to existing network-based risk measures reveals that our approach offers a new perspective on systemic risk. In Chapter 5 we aim to model the joint dynamics of cryptocurrencies in a nonstationary setting. In particular, we analyze the role of cointegration relationships within a large system of cryptocurrencies in a vector error correction model (VECM) framework. To enable analysis in a dynamic setting, we propose the *COINtensity* VECM, a nonlinear VECM specification accounting for a varying system-wide cointegration exposure.

Keywords: modern econometrics, machine learning, nonparametric statistics, quantile regression.

Zusammenfassung

Diese Dissertation behandelt verschiedene Aspekte moderner Ökonometrie und Machine Learnings. Kapitel 2 stellt einen neuen Schätzer für die Regressionsparameter in einem Paneldatenmodell mit interaktiven festen Effekten vor. Eine Besonderheit unserer Methode ist die Modellierung der factor loadings durch nichtparametrische Funktionen. Wir zeigen die $\sqrt{NT}$-Konvergenz sowie die asymptotische Normalverteilung unseres Schätzers. Kapitel 3 betrachtet die rekursive Schätzung von Quantilen mit Hilfe des stochastic gradient descent (SGD) Algorithmus mit Polyak-Ruppert Mittelwertbildung. Der Algorithmus ist rechnerisch und Speicher-effizient verglichen mit herkömmlichen Schätzmethoden. Unser Fokus ist die Untersuchung des nichtasymptotischen Verhaltens, indem wir eine exponentielle Wahrscheinlichkeitsungleichung zeigen. In Kapitel 4 stellen wir eine neue Methode zur Kalibrierung von conditional Value-at-Risk (CoVaR) basierend auf Quantilregression mittels Neural Networks vor. Wir modellieren systemische Spillovereffekte in einem Netzwerk von systemrelevanten Finanzinstituten. Eine Out-of-Sample Analyse zeigt eine klare Verbesserung im Vergleich zu einer linearen Grundspezifikation. Im Vergleich mit bestehenden Risikomaßen eröffnet unsere Methode eine neue Perspektive auf systemisches Risiko. In Kapitel 5 modellieren wir die gemeinsame Dynamik von Kryptowährungen in einem nicht-stationären Kontext. Um eine Analyse in einem dynamischen Rahmen zu ermöglichen, stellen wir eine neue vector error correction model (VECM) Spezifikation vor, die wir *COINtensity* VECM nennen.

Schlagworte: moderne Ökonometrie, Machine Learning, nichtparametrische Statistik, Quantilregression.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is an ongoing discussion about the relation between traditional statistical and econometric tools and the recent development of machine learning methods. Breiman et al. (2001) argued about the existence of a clear dichotomy between a "data modeling culture" on the one hand and an "algorithmic modeling culture" on the other hand. The former relies on an explicit assumption of a stochastic model while the latter takes a black box approach. In recent years, the boundaries between these two paradigms seem to have vanished. It has proved to be useful to analyze machine learning methods from a statistical and econometric perspective. Barron (1993) and Chen and White (1999) studied the rate of convergence of single-layer neural networks under the nonparametric framework of nonlinear sieve estimation (see Chen, 2007). Similarly, Scornet et al. (2015) and Scornet (2016) showed that random forests can be viewed as a special case of kernel regression with an adaptive bandwidth. Furthermore, machine learning methods have been applied successfully to the semiparametric estimation of treatment effects (Athey et al., 2019; Chernozhukov et al., 2018). This dissertation is written in the same spirit, namely that econometrics and machine learning methods can be viewed as two sides of the same coin. The remainder of the thesis consists of four chapters that focus on various aspects of the union of modern econometrics and machine learning. The contributions are both of a theoretical and applied nature.

The second Chapter considers a new estimator of the regression parameters in a panel data model with unobservable interactive fixed effects, which are allowed to be correlated with the regressors. A distinctive feature of the proposed approach is the projection of the (smoothed) data matrix onto the orthogonal linear sieve space spanned by the covariates, instead of projecting on the orthogonal spaces of factors. Therefore, the new estimator adopts the well-known form of partial least squares estimation. Further, this approach facilitates us to have a direct estimator for regression parameters without the need of estimating factors. In addition, we show that our estimator is $\sqrt{NT}$-consistent and asymptotically normal, as well that it reaches the semiparametric efficiency bound under the assumption of i.i.d. errors. A Monte Carlo study indicates a great performance in terms of mean squared error. We apply our methodology to analyze the determinants of growth rates in OECD countries.

Chapter 3 is concerned with the recursive estimation of quantiles using the stochastic gradient descent (SGD) algorithm with Polyak-Ruppert averaging. The algorithm offers a computationally and memory efficient alternative to the usual empirical estimator. Our focus is on studying the nonasymptotic behavior by providing exponentially decreasing tail probability bounds under minimal assumptions. This novel result is based on a bound of the moment generating function of the SGD estimate. We apply our result to the problem of best arm identification in a multi-armed stochastic bandit setting under quantile preferences.

In Chapter 4 we propose a novel approach to calibrate the conditional value-at-risk (CoVaR) of financial institutions based on neural network quantile regression. Building on the estimation results, we model systemic risk spillover effects in a network context across banks by considering the marginal effects of the quantile regression procedure. An out-of-sample analysis shows great performance compared to a linear baseline specification, signifying the importance that nonlinearity plays for modelling systemic risk. We then propose three network-based measures from our fitted results. First, we use the *Systemic Network Risk Index (SNRI)* as a measure for total systemic risk. A comparison to existing network-based risk measures reveals that our approach offers a new perspective on systemic risk due to the focus on the lower tail and to the allowance for nonlinear effects. We also introduce the *Systemic Fragility Index (SFI)* and the *Systemic Hazard Index (SHI)* as firm-specific measures, which allow us to identify systemically relevant firms during the financial crisis.

In Chapter 5 we aim to model the joint dynamics of cryptocurrencies in a nonstationary setting. In particular, we analyze the role of cointegration relationships within a large system of cryptocurrencies in a vector error correction model (VECM) framework. To enable analysis in a dynamic setting, we propose the *COINtensity* VECM, a nonlinear VECM specification accounting for a varying system-wide cointegration exposure. Our results show that cryptocurrencies are indeed cointegrated with a cointegration rank of four. We also find that all currencies are affected by these long term equilibrium relations. The nonlinearity in the error adjustment turned out to be stronger during the height of the cryptocurrency bubble. A simple statistical arbitrage trading strategy is proposed showing a great in-sample performance, whereas an out-of-sample analysis gives reason to treat the strategy with caution.

All codes of this dissertation are available on quantlet.de.

Bibliography

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, *39*(3), 930–945.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Chen, X. (2007). Handbook of econometrics. North Holland.

Chen, X., & White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, *45*(2), 682–691.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, *146*, 72–83.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4), 1716–1741.

Chapter 2

A Projection-Based Approach to Interactive Fixed Effects

2.1 Introduction

Panel data models have proved to be useful tools for the estimation of regression parameters under the presence of unobserved heterogeneity in the data (see Arellano (2003) for an overview). A large proportion of the literature focused on the special case of additive individual effects. However, in many empirical applications this assumption of a single time-invariant component might be excessively restrictive. Instead, a more realistic setting should allow for multiple effects which can change over time. For instance, in the analysis of worker's wages different unobserved talents and skills might be evaluated differently across time. As standard econometric tools such as the fixed effects estimator are not suitable in this situation, recent studies have been devoted to the analysis of interactive effects models, where the unobserved heterogeneity is modelled via a latent factor structure. Factor models have been studied extensively in the literature with applications in asset pricing, macroeconomics and empirical labor economics (Bai, 2003; Bai & Ng, 2013; Stock & Watson, 2002). In the seminal works of Connor and Linton (2007) and Connor et al. (2012), the authors model the individual-specific factor loadings as a nonparametric function of time-invariant regressors. Fan et al. (2016) extended this framework and proposed an projected principal component algorithm. The estimation of regression parameters in the interactive fixed effects model is studied in Ahn et al. (2013), who extended the quasi-differencing method of Holtz-Eakin et al. (1988) to the case of multiple factors. Another estimator, which is based on filtering out individual regressors by taking cross-sectional averages, was proposed by Pesaran (2006). Bai (2009) proposed a principal component-based algorithm to estimate the regression parameters in a model with multiplicative fixed effects. Moon and Weidner (2015) proposed a bias-corrected least squares estimator and discussed inference-related issues.

In this paper, we propose a new projection-based estimator for the regression parameters in the interactive fixed effects model. A distinctive feature of the approach is the

projection of the data matrix onto the orthogonal linear subspace spanned by the covariates, instead of projecting on the orthogonal spaces of factors. We extend the framework of Fan et al. (2016) to the regression case and to the case of time-varying covariates. In particular, we assume that the individual-specific factor loadings are smooth functions of the time-averages of covariates, perturbed by an error term that is independent of the regressors. The regression model takes the form of a partial linear model (see e.g. Härdle et al., 2012). Therefore, the new estimator adopts the well-known form of partial least squares estimation. An advantage of our method is that we do not need to require knowledge about the number of factors. We show the $\sqrt{NT}$-consistency and asymptotic normality of our estimator. In the i.i.d. situation, and if the loadings can be explained completely by the nonparametric functions, our estimator reaches the semiparametric efficiency bound. While the main focus of the paper is the estimation of the regression parameters, we also obtain consistent estimators for the latent factors and corresponding factor loadings.

We validate the theoretical results in a simulation study. In the case, where the time averages of regressors have a non-vanishing explaining power on the factor loadings, our estimator outperforms alternative estimators which do not account for the relation. We apply our method to the identification of the determinants of economic growth. We obtain growth rates and country-specific variables from the Penn World Table and from the World Bank Development Indicators. Lu and Su (2016) argued that the GDP growth rates per capita might not only be determined by observed factors, but might also be influenced by latent factors or shocks. Our projection-based interacted fixed effects estimator is well suited for such a setting. Indeed, our empirical findings suggest an important role of these latent effects. Especially, when only concentrating on the subset of OECD countries, the factor loadings can be well explained by the time averages of regressors.

The main contribution of this paper lies in the synthesis of the projection-based principal components approach proposed by Fan et al. (2016) and the regression framework with interactive effects of Pesaran (2006) and Bai (2009). Another key contribution is the extension of Fan et al. (2016) to the case of time-varying regressors. Our estimator takes the very simple and intuitive structure of a partial least squares estimator. An important advantage of the estimator is that it is unaffected by and thus does not require knowledge on the number of factors.

The paper is organized as follows. In Section 2.2, we present the model setup and derive our projection-based interactive fixed effects estimator. Section 2.3 provides our assumptions and studies the asymptotic properties. In Section 2.4, we examine the performance of our estimator in a Monte Carlo study. We apply our method to

analyze the determinants of economic growth in Section 2.5. Section 2.6 discusses future research directions and concludes.

All codes of this paper are available on quantlet.de.

2.2 Model and Estimation Procedure

2.2.1 Panel Data with Interactive Fixed Effects

We consider the following panel data model with interactive fixed effects, see Bai (2009),

$$\begin{aligned} y_{it} &= x_{it}^{\top}\beta + \varepsilon_{it}, \\ \varepsilon_{it} &= \lambda_i^{\top} f_t + u_{it} \end{aligned} \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{2.1}$$

where y_{it} denotes the response variable of the individual i in the period t, x_{it} is the Q-dimensional covariate vector and β is the Q-dimensional vector of parameters to be estimated. The relationship between x_{it} and y_{it} described in (2.1) contains K unobserved common factors, $f_t = (f_{t1}, \ldots, f_{tK})^{\top}$, and the corresponding factor loadings for individual i, $\lambda_i^{\top} = (\lambda_{i1}, \ldots, \lambda_{iK})$. Further, these quantities are perturbed by the idiosyncratic error term, u_{it}. Throughout the paper, λ_i, f_t, and u_{it} are all unobserved. In addition, it is assumed that u_{it} is uncorrelated with λ_i and f_s for all i, t and s.

A key feature of this regression model is that the explanatory variables x_{it} are allowed to be dependent on the interactive effect components. Consequently, this model is more realistic and can be applied in a great variety of economic studies and other social sciences in which some of the regressors are decision variables that can be influenced by unobserved individual heterogeneities, see Bai and Li (2014). In labor economics, for example, y_{it} can be the wage of individual i at period t, x_{it} is a vector of observed covariates such as education, experience, gender or marital status, λ_i is a vector of unmeasured features of the individuals such as ability, perseverance, motivation, and dedication, f_t is the vector of prices for the unobserved skills, and u_{it} captures the idiosyncratic variation in the wages. In macroeconomic studies, the interactive effects represent unobserved common shocks and their heterogeneous impacts on the cross section. Thus, y_{it} can be the output growth rate for country i in year t, x_{it} is the vector that contains the production inputs (i.e., labor and capital), f_t are the common shocks such as technological shocks or financial crises, and u_{it} is the country-specific error term of the output growth rate. In finance, y_{it} can be the return of the stock i in period t, x_{it} is the vector of observed covariates such as dividend yields, dividend payout ratio, and consumption gap, f_t represents unobserved common factors such as systematic risks, λ_i is the exposure to the risks, and u_{it} is the idiosyncratic part of

returns.

Our interest is centered on the estimation of the slope coefficient β. As it is well known in the literature, any attempt to estimate β directly through standard panel data estimation techniques will lead to inconsistent estimators since $\mathrm{E}(\lambda_{ik}|x_{it}) \neq 0$. Alternatively, standard panel data transformation techniques, like the within transformation, are used to remove the heterogeneity term from the statistical model of interest. However, in this case such transformations fail as a result of the multiplicative form of the fixed effects and the resulting estimators are inconsistent.

2.2.2 Semiparametric Interactive Fixed Effects Model

In order to overcome this situation, we propose to incorporate these unknown relationships into the model of interest in order to obtain a consistent estimator for β. Specifically, as suggested in Bai (2009), the correlated random effects framework of Mundlak (1978) can be extended to the case of interactive effects. A natural way of expressing these relationships is projecting λ_i over $\bar{x}_{i\cdot}$, where $\bar{x}_{i\cdot}$ is the time average of x_{it}. Thus, the factor loadings can be expressed as

$$\lambda_i = \theta \bar{x}_{i\cdot} + \gamma_i, \tag{2.2}$$

where θ is a $K \times Q$ matrix of unknown parameters and γ_i is a $K \times 1$ component of the loading coefficients that cannot be explained by the covariates $\bar{x}_{i\cdot}$. We assume that γ_i has zero mean and is independent of $\bar{x}_{i\cdot}$ and u_{it}.

Note that the above relationship enables us to consider much more general situations than those in Fan et al. (2016). Specifically, instead of focusing only on observed time-invariant covariates, we also consider those regressors that vary across the individuals and across time. However, this linear specification can be rather restrictive in many cases, therefore we instead follow a nonparametric approach,

$$\lambda_i = g(\bar{x}_{i\cdot}) + \gamma_i, \tag{2.3}$$

where $g(\cdot)$ is a $K \times 1$ vector of unknown smooth functions.

Then, plugging (2.3) into (2.1) and rearranging terms we get

$$y_{it} = x_{it}^{\top}\beta + g(\bar{x}_{i\cdot})^{\top} f_t + \gamma_i^{\top} f_t + u_{it}, \tag{2.4}$$

which reduces to model (2.1) when $g(\cdot) = 0$ and to the model analyzed in Connor and Linton (2007) and Connor et al. (2012) when $\gamma_i = 0$.

The above regression model still has a factor error structure. However, assuming that f_t is uncorrelated with the explanatory variables x_{it}, it is possible to define the following composed error term,

$$v_{it} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \gamma_{ki} f_{kt} + u_{it}.$$

Given that v_{it} is uncorrelated with x_{it}, consistent estimators for β in (2.4) can be obtained by using standard estimation techniques for panel data models.

Let y_t and v_t be $N \times 1$ vectors of y_{it} and v_{it}, respectively, X_t be the $N \times Q$ matrix of regressors, β be the $Q \times 1$ unknown vector to be estimated, f_t be the $K \times 1$ vector of f_{tk}, and $G(\bar{X})$ be the $N \times K$ matrix of nonparametric functions, $g_k(\bar{x}_{i\cdot})$. Then, the model (2.4) can be written in matrix form as

$$y_t = X_t\beta + G(\bar{X})f_t + v_t, \qquad t = 1, \ldots, T. \tag{2.5}$$

In this situation, instead of projecting on the space of factors f_t, one can consider a sieve estimation for $G(\bar{X})$ to estimate β and project on the space expanded by the sieve basis. Before introducing the estimation procedure that we propose in this paper, we recall simply the polynomial spline function.

Let $\mathcal{X}$ be an interval with end points $\zeta_0 < \zeta_{M+1}$. A polynomial spline of degree $d \geq 0$ on $\mathcal{X}$ with knot sequence $\zeta_0 < \zeta_1 < \ldots < \zeta_{M+1}$ is a function that is a polynomial of degree d on each interval $[\zeta_0, \zeta_1), \ldots, [\zeta_{M-1}, \zeta_M), [\zeta_M, \zeta_{M+1})$, and globally has continuous $d-1$ derivatives for $d \geq 1$. The collection of spline functions of a particular degree and knot sequence form a linear space. Specifically, a piecewise constant function, linear spline, quadratic spline, and cubic spline corresponds to $d = 0, 1, 2, 3$, respectively. We refer to de Boor (1978) and Schumaker (1980) as a good overview for spline functions.

Given that $\bar{x}_{i\cdot}$ is Q-variate, in order to avoid the curse of dimensionality in the nonparametric estimation of $g_k(\cdot)$, we will assume that for each k, $g_k(\cdot)$ is an additive function, that is,

$$g_k(\bar{x}_{i\cdot}) = \sum_{q=1}^{Q} g_{kq}(\bar{x}_{i\cdot,q}). \tag{2.6}$$

Suppose that, for each k and q, the function $g_{kq}(\cdot)$ can be approximated by some spline

function, that is,

$$g_{kq}(\bar{x}_{i\cdot,q}) = \sum_{\ell=1}^{J_g} b_{\ell,kq}\phi_\ell(\bar{x}_{i\cdot,q}) + R_{kq}(\bar{x}_{i\cdot,q}), \quad k = 1,\dots,K, \quad q = 1,\dots,Q. \tag{2.7}$$

where $\phi_\ell(\cdot)$'s are the spline basis functions. For $l = 1,\dots,J_g$, $b_{\ell,kq}$'s are the sieve coefficients of the qth additive component of $g_k(\bar{x}_{i\cdot})$ corresponding to the kth factor loading, and R_{kq} is a "remainder function" that represents the approximation error. Also, J_g denotes the number of sieve terms which grows slowly as $N \to \infty$. As it is well-known in the literature, the basic assumption for sieve approximation is that the approximation error approaches zero, $\sup_x |R_{kq}(\bar{x}_{i\cdot,q})| \to 0$, as $J_g \to \infty$.

For the sake of simplicity, we take the same basis functions in (2.7). For each $k \le K$, $q \le Q$ and $i \le N$, let us define

$$\begin{aligned} b_k^\top &= (b_{1,k1},\dots,b_{J_g,k1},\dots,b_{1,kQ},\dots,b_{J_g,KQ}) \in \mathbb{R}^{J_gQ}, \\ \phi(\bar{x}_{i\cdot})^\top &= (\phi_1(\bar{x}_{i\cdot,1}),\dots,\phi_J(\bar{x}_{i\cdot,1}),\dots,\phi_1(\bar{x}_{i\cdot,Q}),\dots,\phi_J(\bar{x}_{i\cdot,Q})) \in \mathbb{R}^{J_gQ}. \end{aligned}$$

Thus, the above equation can be rewritten as

$$g_k(\bar{x}_{i\cdot}) = \phi(\bar{x}_{i\cdot})^\top b_k + \sum_{q=1}^{Q} R_{kq}(\bar{x}_{i\cdot,q}). \tag{2.8}$$

By considering (2.8) in matrix form we obtain

$$G(\bar{X}) = \Phi(\bar{X})B + R(\bar{X}), \tag{2.9}$$

where $\Phi(\bar{X}) = (\phi(\bar{x}_{1\cdot}),\dots,\phi(\bar{x}_{N\cdot}))^\top$ is a $N \times J_gQ$ matrix of basis functions, $B = (b_1,\dots,b_K)$ is a $J_gQ \times K$ matrix of sieve coefficients, and $R(\bar{X})$ is a $N \times K$ matrix with the (i,k)th element $\sum_{q=1}^{Q} R_{kq}(\bar{x}_{i\cdot,q})$.

Then, substituting (2.9) into (2.5) leads to

$$y_t = X_t\beta + \Phi(\bar{X})Bf_t + R(\bar{X})f_t + v_t, \qquad t = 1,\dots,T. \tag{2.10}$$

We want to point out that the residual term of this regression model consists of two parts, the sieve approximation error $R(\bar{X})f_t$ and the idiosyncratic error v_t that is of the form $v_t = \Gamma f_t + u_t$, where Γ is a $N \times K$ matrix of unknown loading coefficients.

With the aim of estimating β and taking as benchmark the idea in Fan et al. (2016),

we define P_Φ as the projection matrix onto $\mathcal{X}$, where $\mathcal{X}$ is the sieve spaced spanned by the basis functions of $\bar{X}$. More precisely, P_Φ is the $N \times N$ projection matrix of the form

$$P_\Phi(\bar{X}) = \Phi(\bar{X}) \left\{ \Phi(\bar{X})^\top \Phi(\bar{X}) \right\}^{-1} \Phi(\bar{X})^\top. \tag{2.11}$$

Therefore, one can obtain the estimator of β by partialling out the effect of factors f_t, i.e.

$$\widehat{\beta} = \left[\sum_{t=1}^{T} X_t^\top \left\{ I_N - P_\Phi(\bar{X}) \right\} X_t \right]^{-1} \sum_{t=1}^{T} X_t^\top \left\{ I_N - P_\Phi(\bar{X}) \right\} y_t, \tag{2.12}$$

where $X_t^\top \left\{ I_N - P_\Phi(\bar{X}) \right\} X_t$ is assumed to be asymptotically nonsingular.

As the reader can remark, the resulting estimator of β appears as the solution of a partially linear model (see Härdle et al. (2012) for a comprehensive review of the literature), where the nonparametric part is "partialled out". Although the asymptotic properties of this estimator have been already studied under many alternative sets of assumptions, it is worthwhile to establish these conditions in our context and obtain its asymptotic distribution.

2.2.3 Estimation of Interactive Fixed Effects Components

The latent factors and corresponding factor loadings can be estimated from the regression residuals,

$$\widetilde{y}_t = y_t - X_t \widehat{\beta}, \tag{2.13}$$

and let $\widetilde{Y} = (\widetilde{y}_1, \ldots, \widetilde{y}_T)$. We can estimate the matrix of factors, $F = (f_1, \ldots, f_T)^\top$, by following the approach of Fan et al. (2016). In particular, we can estimate $\sqrt{T}\widehat{F}$ by the eigenvectors associated with the largest K eigenvalues of the matrix $\widetilde{Y}^\top P_\Phi(\bar{X}) \widetilde{Y}$. Using the estimated matrix of factors, we estimate the corresponding matrix of factor loadings by

$$\widehat{\Lambda} = \widetilde{Y} \widehat{F} / T. \tag{2.14}$$

The part of the factor loadings that can be explained by $\bar{X}$ and the idiosyncratic part can be estimated by $\widehat{G}(\bar{X}) = \frac{1}{T} P_\Phi(\bar{X}) \widetilde{Y} \widehat{F}$ and $\widehat{\Gamma} = \frac{1}{T} \{ I_N - P_\Phi(\bar{X}) \} \widetilde{Y} \widehat{F}$, respectively. In order to estimate the functions g_k, we can obtain an estimator for the sieve coefficients by least squares,

$$\widehat{B} = (\widehat{b}_1, \ldots, \widehat{b}_K) = \frac{1}{T} \left\{ \Phi(\bar{X})^\top \Phi(\bar{X}) \right\}^{-1} \Phi(\bar{X})^\top \widetilde{Y} \widehat{F}. \tag{2.15}$$

Finally, we can construct an estimator for the nonparametric functions,

$$\widehat{g}_k(x) = \phi(x)^\top \widehat{b}_k, \quad k = 1, \ldots, K. \tag{2.16}$$

In practice, the number of latent factors, K, is unknown and needs to be estimated. We follow the approach of Fan et al. (2016) to select $\widehat{K}$ according to the largest ratio of eigenvalues of the matrix $\widetilde{Y}^\top P_\Phi(\bar{X})\widetilde{Y}$,

$$\widehat{K} = \arg \max_{0<k<J_g Q/2} \frac{\lambda_k \left\{\widetilde{Y}^\top P_\Phi(\bar{X})\widetilde{Y}\right\}}{\lambda_{k+1} \left\{\widetilde{Y}^\top P_\Phi(\bar{X})\widetilde{Y}\right\}}. \tag{2.17}$$

The condition that the true dimension of the factors is smaller than $J_g Q/2$ is fulfilled naturally since the sieve dimension J_g grows slowly with the sample size. Interestingly, our estimator for the regression parameters $\widehat{\beta}$ does not require any knowledge of K. However, the number of factors is crucial to the estimation of the factor components as well as to the estimation of standard errors for the regression parameters.

2.3 Asymptotic Properties

In this section we analyze the main asymptotic properties of $\widehat{\beta}$. With this aim, the following assumptions are considered. Also, some additional notation is necessary. For a real matrix A, let $\|A\|_F = \{tr(A^\top A)\}^{1/2}$ denote its Frobenius norm and $\|A\|_2 = \{\lambda_{\max}(A^\top A)\}^{1/2}$ denotes its spectral norm where $\lambda_{\max}(\cdot)$ is the largest eigenvalue of "$\cdot$".

2.3.1 Assumptions

We start by introducing some assumptions related to the data generating process in (2.5). Specifically, about the vector of explanatory variables x_{it}, we assume that, as common in the partially linear regression models, x_{it} and $\bar{x}_{i.}$ are related.

Assumption 2.3.1.

$$x_{itq} = \sum_{q'} h_{qq'}\left(\bar{x}_{i.,q'}\right) + \pi_{itq}, \quad i = 1, \ldots, N;\ t = 1, \ldots, T;\ q = 1, \ldots, Q,$$

where the $h_{qq'}(\cdot)$'s are unknown functions and the π_{itq}'s are random variables with zero mean.

We also need to characterize the asymptotic behavior of $\bar{x}_{i.}$. We will assume that,

Assumption 2.3.2.

$$\frac{1}{T}\sum_t x_{itq} = \mu_{iq} + \mathcal{O}_p\left(T^{-1/2}\right), \quad i = 1, \ldots, N;\ q = 1, \ldots, Q,$$

as T tends to infinity. Furthermore, the deterministic sequence of design points, μ_{iq}, has bounded support $\mathcal{M}$ and is generated by a design density $f(\mu)$ that is bounded below and above in $\mathcal{M}$.

On the nonparametric functions $h_{qq'}(\cdot)$ and $g_{kq}(\cdot)$ we impose the following smoothness assumption.

Assumption 2.3.3.

(i) *For all $k = 1, \ldots, K$ and $q, q' = 1, \ldots, Q$, the functions $h_{qq'}(\cdot)$ and $g_{kq}(\cdot)$ belong to a Hölder class $\mathcal{G}$ with Hölder coefficient $0 < \alpha \leq 1$,*

$$\mathcal{G} = \left\{ g : |g^{(r)}(s) - g^{(r)}(t)| \leq L|s-t|^{\alpha} \right\}$$

for some $L > 0$.

(ii) *For the sieve approximation error we assume, for $\kappa = 2(r+\alpha) \geq 4$,*

$$\rho_{g,N} = \sup_{\mu \in \mathcal{M}} \left| g_{kq} - \sum_{\ell=1}^{J_g} b_{\ell,kq} \phi_l(\mu_{i,q}) \right|^2 = \mathcal{O}\left(J_g^{-\kappa}\right)$$
$$\rho_{h,N} = \sup_{\mu \in \mathcal{M}} \left| h_{qq'} - \sum_{\ell=1}^{J_h} c_{\ell,qq'} \phi_l(\mu_{i,q'}) \right|^2 = \mathcal{O}\left(J_h^{-\kappa}\right).$$

(iii) $\max_{l,k,q} b_{l,kq}^2 < \infty$ *and* $\max_{l,q,q'} c_{l,qq'}^2 < \infty$.

As it is remarked in Fan et al. (2016), Assumption 2.3.3 (ii) is satisfied by the use of common basis functions such as polynomial basis or B-splines. Lorentz (1986) and Chen (2007) show that (i) implies (ii) in this case. We impose the following assumptions on the random matrix π_t.

Assumption 2.3.4. *Let $\pi_t = (\pi_{1t}, \ldots, \pi_{Nt})^{\top}$, for $t = 1, \ldots, T$, be independent random matrices with zero mean and $\mathrm{E}\|\pi_t\|_F^4 \leq a_0 < \infty$, where $\pi_{it} = (\pi_{it1}, \ldots, \pi_{itq})^{\top}$ for $i = 1, \ldots, N$.*

Let $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras generated by $\{(f_t, u_t) \,:\, t \leq 0\}$ and $\{(f_t, u_t) \,:\, t \geq T\}$ respectively. Define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{T}^{\infty}} |P(A)P(B) - P(AB)|.$$

Assumption 2.3.5.

(i) *Let* $\{u_t, f_t\}$ *be a strictly stationary process. In addition,* $\mathrm{E}(u_{it}) = 0$ *and* $\{u_t\}$ *is independent of* $\{\pi_t, f_t\}$.

(ii) *There exist* $\alpha_1, C_1 > 0$ *such that*

$$\alpha(T) < \exp\left(-C_1 T^{\alpha_1}\right).$$

(iii) *There exists* $C_2 > 0$, *so that*

$$\max_{j \le N} \sum_{i=1}^{N} |\mathrm{E}(u_{it}u_{jt})| < C_2,$$
$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} |\mathrm{E}(u_{it}u_{js})| < C_2,$$
$$\max_{i \le N} \frac{1}{NT} \sum_{k=1}^{N} \sum_{m=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} |\mathrm{Cov}\left(u_{it}u_{kt}, u_{is}u_{ms}\right)| < C_2.$$

(iv) *There exists* $\alpha_2, \alpha_3 > 0$, $\alpha_1^{-1} + \alpha_2^{-1} + \alpha_3^{-1} > 1$ *and* $b_1, b_2 > 0$, *such that for any* $x > 0$, $i \le N$ *and* $k \le K$,

$$P\left(|u_{it}| > x\right) \le \exp\left\{-\left(x/b_1\right)^{\alpha_2}\right\},$$
$$P\left(|f_{kt}| > x\right) \le \exp\left\{-\left(x/b_2\right)^{\alpha_3}\right\}.$$

Assumption 2.3.5 is standard in factor analysis (Bai, 2003; Fan et al., 2016; Stock & Watson, 2002). Part (ii) is a strong mixing condition, whereas (iii) imposes weak cross-sectional dependence. Finally, (iv) ensures that the tails of u_{it} and f_{kt} are sub-exponential and thus sufficiently light. Now, let ν_N be

$$\nu_N = \max_{k \le K} \frac{1}{N} \sum_{i \le N} \mathrm{Var}\left(\gamma_{ik}\right).$$

On the random part of the factor loadings, we impose the following conditions.

Assumption 2.3.6.

(i) $\mathrm{E}\left(\gamma_{ik}\right) = 0$, $\nu_N < \infty$ *and*

$$\max_{k \le K, j \le N} \sum_{i \le N} |\mathrm{E}\,\gamma_{ik}\gamma_{jk}| = \mathcal{O}\left(\nu_N\right).$$

(ii) π_{it} *is independent of* γ_{ik}, *for* $i = 1, \ldots, N$, $t = 1, \ldots, T$ *and* $k = 1, \ldots, K$.

For the identification of the factors F and the part of the loadings explained by the covariates, $G(\cdot)$, we need the following assumptions.

Assumption 2.3.7.

(i) *Almost surely,* $T^{-1}F^{\top}F = I_K$ *and* $G(\mu)^{\top}G(\mu)$ *is a* $K \times K$ *diagonal matrix with distinct entries.*

(ii) *There are two constants,* c_{min} *and* $c_{max} > 0$*, so that, as* N *tends to infinity,*

$$c_{min} < \lambda_{min}\left\{\frac{1}{N}G(\mu)^{\top}G(\mu)\right\} < \lambda_{max}\left\{\frac{1}{N}G(\mu)^{\top}G(\mu)\right\} < c_{max}.$$

Part (ii) of Assumption 2.3.7 ensures that the covariates have a nonvanishing explanatory power on the loadings. Finally, for the basis functions, following Fan et al. (2016) we assume the following.

Assumption 2.3.8.

(i) *There are* c'_{min} *and* $c'_{max} > 0$*, as* N *tends to infinity,*

$$c'_{min} < \lambda_{min}\left\{\frac{1}{N}\Phi(\mu)^{\top}\Phi(\mu\}\right) < \lambda_{max}\left\{\frac{1}{N}\Phi(\mu)^{\top}\Phi(\mu)\right\} < c'_{max}.$$

(ii) $\max_{\ell,i,q}\phi_{\ell}(\mu_{iq}) < \infty$.

2.3.2 Limit Theory

In this section, we present the main theoretical results of this paper. The following theorem provides the $\sqrt{NT}$-consistency and asymptotic normality of the projection-based interactive fixed effects estimator $\widehat{\beta}$.

Theorem 2.3.1. *Under assumptions 2.3.1 to 2.3.8 and if* $J_h \sim N^{1/2}$*,* $J_g \sim N^{1/2}$*,* $T/N \to 0$ *we have that, as both* N *and* T *tend to infinity,*

$$\sqrt{NT}\left(\widehat{\beta} - \beta\right) \xrightarrow{\mathcal{L}} N\left(0, \widetilde{V}\right),$$

where

$$\widetilde{V} = \widetilde{V}_{\pi}^{-1}\left(\widetilde{V}_{\Gamma} + \widetilde{V}_{u}\right)\widetilde{V}_{\pi}^{-1}, \tag{2.18}$$

with

$$\widetilde{V}_\pi = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{E}\left(\pi_{it}\pi_{it}^\top\right),$$
$$\widetilde{V}_\Gamma = \lim_{NT\to\infty} \frac{1}{NT} \sum_{t=1}^{T} \mathrm{E}\left(\pi_t^\top \Gamma\Gamma^\top \pi_t\right),$$
$$\widetilde{V}_u = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{t=1}^{T} \mathrm{E}\left(\pi_t^\top u_t u_t^\top \pi_t\right).$$

The proof of Theorem 2.3.1 is provided in Appendix 2.A. The key component of the proof is the Frisch-Waugh idea to partial out the effect of the latent factors and corresponding loadings. A second key component of the proof is using results from approximation theory for linear sieves to approximate the nonparametric functions. The asymptotic covariance matrix of our estimator has the typical sandwich structure. Note that there are two kinds of error terms entering the inner term of the covariance sandwich, the idiosyncratic part of the factor loadings, γ_i, and the idiosyncratic error term, u_{it}. In the case that both of these terms are i.i.d., with $\gamma_{ik} \sim N(0, \sigma_\gamma^2)$ and $u_{it} \sim N(0, \sigma_u^2)$, the asymptotic covariance matrix simplifies to $\widetilde{V} = \widetilde{V}_\pi^{-1}(\sigma_\gamma^2 + \sigma_u^2)$.

If we further assume that the latent factor loadings can be completely explained by the nonparametric functions, i.e. $\gamma = 0$, our estimator reaches the semiparametric efficiency bound (see Newey (1990) for an overview on semiparametric efficiency bounds). Denote $h_q(\bar{x}_{i\cdot}) = \sum_{q'=1}^{Q} h_{qq'}(\bar{x}_{i\cdot,q'})$ and let $H(\bar{x}_{i\cdot}) = (h_1(\bar{x}_{i\cdot,1}), \ldots, h_Q(\bar{x}_{i\cdot,Q}))^\top$. Then, using the results of Chamberlain (1992) and Li (2000), the efficiency bound of the regression model (2.5) is given by

$$J_0 = \mathrm{E}\left\{\left(x_{it} - H(\bar{x}_{i\cdot})\right) \mathrm{Var}\left(u_{it}\right)^{-1} \left(x_{it} - H(\bar{x}_{i\cdot})\right)^\top\right\}. \tag{2.19}$$

Using Assumption 2.3.1 and under i.i.d. errors the bound simplifies to

$$J_0 = \frac{1}{\sigma_u^2} \mathrm{E}\left(\pi_{it}\pi_{it}^\top\right). \tag{2.20}$$

Since the inverse of our asymptotic covariance matrix attends this bound, our estimator $\widehat{\beta}$ is asymptotically semiparametric efficient in the case of homoscedastic errors.

While the primary focus of this paper is the estimation of the regression parameters, we can also consistently estimate the interactive fixed effects components. The consistency result is established in the following theorem, which adapts Theorem 4.1 of Fan et al. (2016) to our case of time-varying covariates.

Theorem 2.3.2. *Under assumptions 2.3.1 to 2.3.8 and if $J_h \sim N^{1/2}$, $J_g \sim N^{1/2}$, $T/N \to 0$ we have that, as both N and T tend to infinity,*

$$\frac{1}{T}\|\widehat{F} - F\|_F^2 = \mathcal{O}_p\left(\frac{1}{N} + \frac{1}{J_g^\kappa}\right)$$

$$\frac{1}{N}\|\widehat{G}(\bar{X}) - G(\bar{X})\|_F^2 = \mathcal{O}_p\left(\frac{J_g}{N^2} + \frac{J_g}{N^T} + \frac{J_g}{J_g^\kappa} + \frac{J_g \nu_N}{N}\right)$$

$$\max_{k=1,\dots,K} \sup_{x\in\mathcal{M}} |\widehat{g}_k(x) - g_k(x)| = \mathcal{O}_p\left(\frac{J_g}{N} + \frac{J_g}{\sqrt{NT}} + \frac{J_g}{J_g^{\kappa/2}} + J_g\sqrt{\frac{\nu_N}{N}}\right).$$

The convergence rates are not affected by the need to estimate $\widehat{\beta}$. That is, they are identical to the pure factor model case of Fan et al. (2016). This follows from the $\sqrt{NT}$ rate of $\widehat{\beta}$.

2.3.3 Consistent Estimation of Standard Errors

To conduct valid inference on the estimated parameters, we present a consistent estimator for the asymptotic covariance matrix $\widetilde{V}$. We restrict our attention to the case of heteroskedasticity, assuming that the error terms are cross-sectionally and serially independent. An extension to the case of serial dependence could be easily achieved by following the approach of Newey and West (1986). In order to get a consistent estimator, $\widehat{V}$, we are required to have consistent estimators for all components of the covariance sandwich. To estimate $\widetilde{V}_\pi$, we define

$$\widehat{V}_\pi = \frac{1}{NT}\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi\left(\bar{X}\right)\right\} X_t. \tag{2.21}$$

Analogously, we define the estimators for $\widetilde{V}_\Gamma$ and $\widetilde{V}_u$ as

$$\widehat{V}_\Gamma = \frac{1}{NT}\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\bar{X})\right\}\widehat{\Gamma}\widehat{\Gamma}^\top \left\{I_N - P_\Phi(\bar{X})\right\} X_t, \tag{2.22}$$

$$\widehat{V}_u = \frac{1}{NT}\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\bar{X})\right\} \operatorname{diag}\left\{\widehat{u}_{1t}^2, \dots, \widehat{u}_{Nt}^2\right\} \left\{I_N - P_\Phi(\bar{X})\right\} X_t, \tag{2.23}$$

where $\widehat{u}_{it} = y_{it} - X_{it}^\top\widehat{\beta} - \widehat{\lambda}_i\widehat{f}_t$ are the fitted residuals of the projected interactive fixed effects estimator. Finally, we have the final estimator for the asymptotic covariance matrix, $\widetilde{V}$,

$$\widehat{V} = \widehat{V}_\pi^{-1}\left(\widehat{V}_\Gamma + \widehat{V}_u\right)\widehat{V}_\pi^{-1}.$$

The following Proposition shows the consistency of $\widehat{V}$.

Proposition 2.3.1. *Assume that the conditions of Theorem 2.3.1 hold. Then* $\widehat{V}_\pi \xrightarrow{p} \widetilde{V}_\pi$. *In addition, if* u_{it} *are serially and cross-sectionally uncorrelated, we have* $\widehat{V}_\Gamma \xrightarrow{p} \widetilde{V}_\Gamma$ *and* $\widehat{V}_u \xrightarrow{p} \widetilde{V}_u$.

The proof is provided in Appendix 2.A.3.

2.4 Numerical Studies

In this section, we evaluate the finite-sample performance of our estimator in a simulation study. We are interested both in the estimation of the regression parameter, β, and the interactive fixed effects parameters F and G. Throughout the study, we fix the number of factors, $K = 3$, and the dimension of covariates, $Q = 3$. The true regression coefficients are set to $\beta = (2, 1, -1)^\top$. The covariates are generated by setting $x_{itq} = \bar{x}_{iq} + \pi_{itq}$, where $\bar{x}_{iq} \sim N(1, 0.5)$ and $\pi_{itq} \sim N(0, 0.5)$. For the factors, we assume $f_{kt} \sim N(0, 1)$.

The factor loadings are set to $\lambda_{ik} = g_k(\bar{x}_i) + \gamma_i$, where $g_k(\bar{x}_i) = a_k \bar{x}_{i1}^2 + b_k \bar{x}_{i2}$, with $a_k, b_k \sim U[-1, 1]$. In order to satisfy the identification condition on the interactive effects components (Assumption 2.3.7 (i)), we further have to transform the factors and loadings. We set F_0 to $\sqrt{T}$ times the K eigenvectors of the matrix $FG^\top GF^\top$. We proceed with setting $G_0 = \frac{1}{T} GF^\top F_0$.

Finally, for the idiosyncratic terms we consider the case of normally distributed errors, $u_{it} \sim N(0, 1)$ and $\gamma_i \sim N(0, 0.05)$. As a robustness check, we consider a second setting in which $u_{it} \sim t_{10}$ and $\gamma_i \sim t_{10}/20$. Our estimator for the regression parameters is obtained by projecting the data onto the sieve space spanned by the basis functions of $\bar{x}_i$ and using an ordinary least squares estimator on the transformed data. In this numerical study, we rely on polynomial basis functions and $J = 2$. For each setting, $S = 1000$ simulations are conducted.

As a performance measure, we consider the root mean square error ($RMSE$) and the bias,

$$RMSE = \sqrt{\frac{1}{SQ} \sum_{s=1}^{S} \sum_{q=1}^{Q} (\widehat{\beta}_{q,s} - \beta_q)^2}$$

$$Bias = \frac{1}{Q} \sum_{q=1}^{Q} \left| \sum_{s=1}^{S} \widehat{\beta}_{q,s} - \beta_q \right|.$$

We compare the performance of our projection based interactive fixed effects (P-IFE) estimator for β with the pooled OLS (POLS) estimator and with the principal component-based interactive fixed effects (PC-IFE) estimator of Bai (2009). The

simulation results under normally distributed disturbances for different values of N and T are reported in Table 2.1.

The $RMSE$ and the bias of the P-IFE estimator can be effectively reduced with increasing N and T. In comparison to the alternative estimators, our estimator achieves the lowest $RMSE$ in all the settings we consider. The efficiency gain relative to the PC-IFE estimator is the largest in cases with small sample sizes. The advantage seems to vanish in settings of large N and T. In terms of bias, the P-IFE estimator performs slightly better in situations where T is small and slightly worse than the PC-IFE estimator in intermediate sample size. Again, the performance of the estimators converges if both N and T are large. The simulation results for the t-distributed error terms are reported in table 2.2. As in the case of normal errors, the P-IFE estimator performs best in terms of $RMSE$ among all the estimators we consider. The same holds for the bias, while as before the performance converges with increasing N and T.

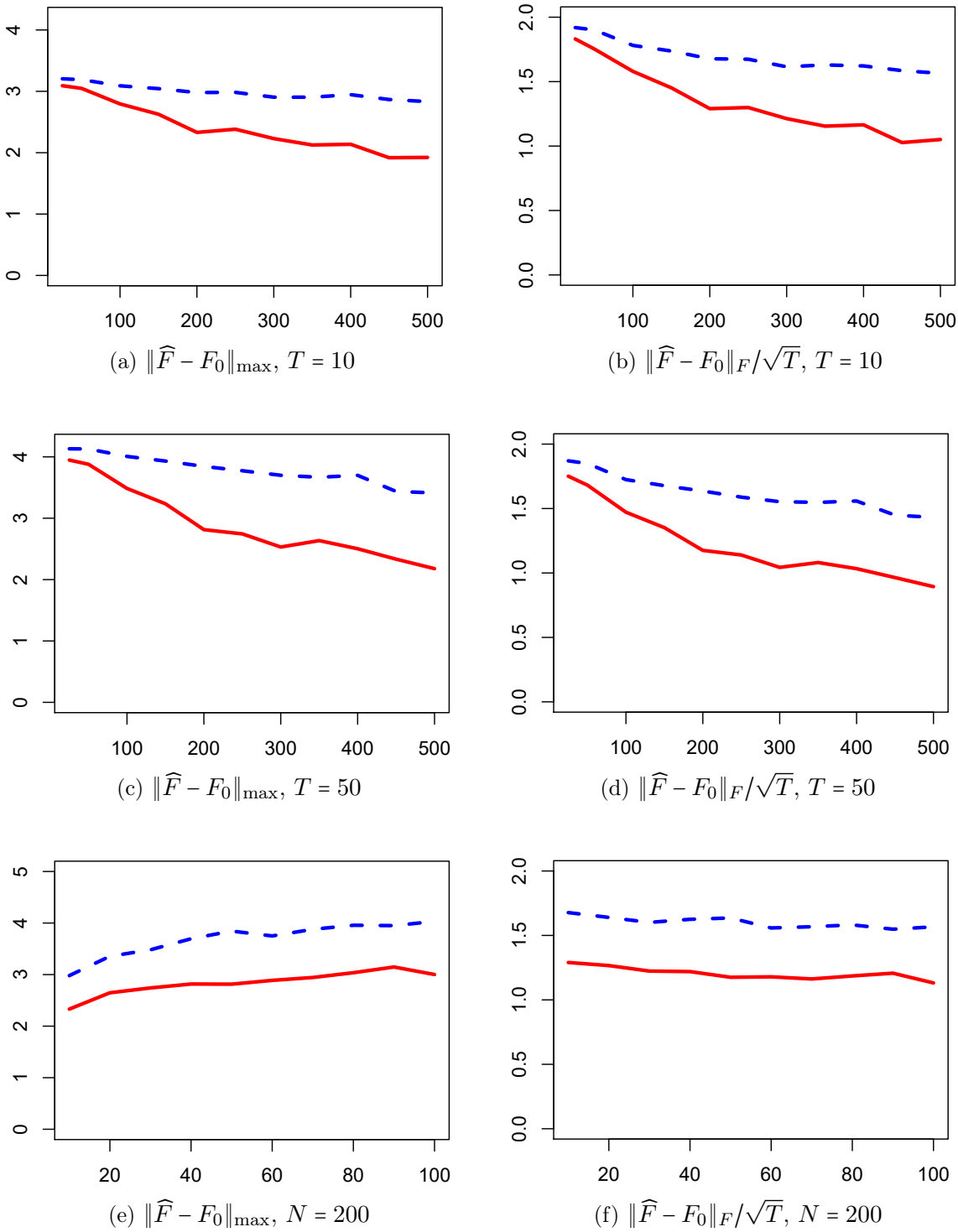
		RMSE			*Bias*		
N	T	P-IFE	POLS	PC-IFE	P-IFE	POLS	PC-IFE
20	10	**0.1637**	0.2214	0.2427	0.0050	**0.0038**	0.0052
50	10	**0.0962**	0.2147	0.1317	**0.0051**	0.0084	0.0057
100	10	**0.0659**	0.2089	0.0894	**0.0012**	0.0029	0.0016
20	50	**0.0765**	0.1099	0.0782	0.0024	0.0024	**0.0015**
50	50	**0.0425**	0.0920	0.0436	0.0019	0.0021	**0.0016**
100	50	**0.0289**	0.0946	0.0303	0.0013	0.0013	**0.0007**
100	100	**0.0206**	0.0686	0.0213	0.0004	0.0017	**0.0003**
200	100	**0.0147**	0.0665	0.0150	**0.0004**	0.0015	**0.0004**
500	100	**0.0090**	0.0636	0.0091	**0.0001**	0.0010	**0.0001**

Table 2.1: Simulation results for the projection-based IFE estimator, the pooled OLS estimator and the PC-based IFE estimator under Gaussian error terms.

		RMSE			*Bias*		
N	T	P-IFE	POLS	PC-IFE	P-IFE	POLS	PC-IFE
20	10	**0.1886**	0.2328	0.2607	**0.0031**	0.0038	0.0064
50	10	**0.1097**	0.2189	0.1442	**0.0040**	0.0082	0.0041
100	10	**0.0741**	0.2114	0.0987	**0.0012**	0.0040	0.0013
20	50	**0.0851**	0.1120	0.0873	**0.0008**	0.0028	0.0018
50	50	**0.0489**	0.0964	0.0502	**0.0011**	0.0018	0.0012
100	50	**0.0333**	0.0924	0.0345	**0.0002**	0.0018	0.0006
100	100	**0.0229**	0.0647	0.0231	**0.0002**	0.0028	**0.0002**
200	100	**0.0165**	0.0661	0.0167	**0.0002**	0.0020	**0.0002**
500	100	**0.0102**	0.0661	0.0104	**0.0004**	0.0013	**0.0004**

Table 2.2: Simulation results for the projection-based IFE estimator, the pooled OLS estimator and the PC-based IFE estimator under t-distributed error terms.

We also evaluate the estimation performance for the interactive fixed effects components F and G. Following the simulation design of Fan et al. (2016), we set $\Gamma = 0$. Our estimators are obtained by applying the projected principal component method (PPCA) on the residuals of our P-IFE estimator. We compare the performance to that of the standard principal component (PCA) method without projection of the data. For the performance measure we choose the max norm and the Frobenius norm. The idiosyncratic error terms are again normally distributed, $u_{it} \sim N(0, 1)$. We consider the case of $T = 10$ or 50 with N varying from 25 to 500. Additionally, we consider the case of $N = 200$ being fixed and T ranging from 25 to 500. The simulation results for the factors are reported in Figure 2.1. We also report the results for the factor loadings in Figure 2.2. In any of our settings, the PPCA method performs better than the standard PCA method. Also, the estimation error for the factors and the loadings can be reduced with increasing T.



(a) $\|\widehat{F} - F_0\|_{\max}$, $T = 10$

(b) $\|\widehat{F} - F_0\|_F/\sqrt{T}$, $T = 10$

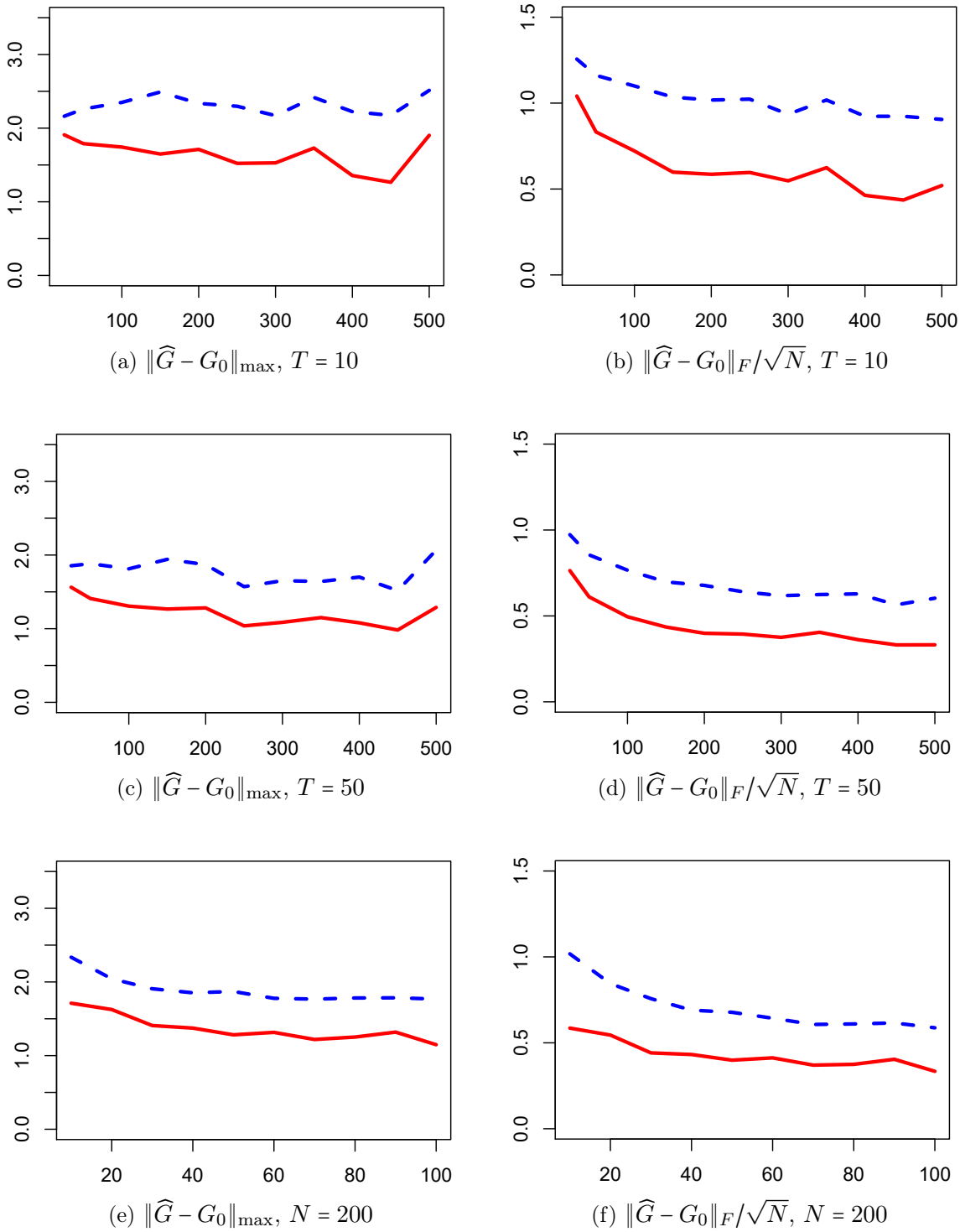
(c) $\|\widehat{F} - F_0\|_{\max}$, $T = 50$

(d) $\|\widehat{F} - F_0\|_F/\sqrt{T}$, $T = 50$

(e) $\|\widehat{F} - F_0\|_{\max}$, $N = 200$

(f) $\|\widehat{F} - F_0\|_F/\sqrt{T}$, $N = 200$

Figure 2.1: Average estimation error of factors F, estimated via Projected PCA (solid red line) and PCA (dashed blue line). Upper two panels: T fixed, N grows, bottom panels: N fixed, T grows.



(a) $\|\widehat{G} - G_0\|_{\max}$, $T = 10$ (b) $\|\widehat{G} - G_0\|_F/\sqrt{N}$, $T = 10$

(c) $\|\widehat{G} - G_0\|_{\max}$, $T = 50$ (d) $\|\widehat{G} - G_0\|_F/\sqrt{N}$, $T = 50$

(e) $\|\widehat{G} - G_0\|_{\max}$, $N = 200$ (f) $\|\widehat{G} - G_0\|_F/\sqrt{N}$, $N = 200$

Figure 2.2: Average estimation error of factor loadings G, estimated via Projected PCA (solid red line) and PCA (dashed blue line). Upper two panels: T fixed, N grows, bottom panels: N fixed, T grows.

2.5 Application: Determinants of Economic Growth

2.5.1 Data and Descriptive Statistics

As an empirical application for our interactive fixed effects estimator we study the determinants of economic growth. We refer to Durlauf et al. (2005) for a comprehensive review of the growth literature. While many studies focus on a cross-sectional analysis (see for instance Barro, 1991), there are also numerous studies employing a panel data approach with country-specific fixed effects (Acemoglu et al., 2019; Islam, 1995). However, Lu and Su (2016) argued that growth might not be solely determined by observable regressors, but could also be influenced by latent factors or shocks. Our projection-based interactive fixed effect estimator is well suited as it is flexible enough to model such latent factors.

The data on GDP growth rates and the country-specific characteristics are obtained from Penn World Table (PWT) and from World Bank World Development Indicators (WDI). Our sample contains 129 countries in a time period from 1991–2019, $N = 129$ and $T = 29$. Countries with an incomplete data availability or which did not exist yet in 1991 are excluded from our analysis. Our dependent variable is the real GDP growth rate per capita. The set of regressors is identical to Lu and Su (2016). Summary statistics of all dependent and independent variables can be found in Table 2.3. Figure 2.3 shows the time series of the mean growth rates, averaged over all countries in our sample. We also visualize the time series of the cross-sectional 5% and 95%-quantiles of the growth rates in the same figure.

Variable	Description	Mean	Median	Min	Max	Data Sources
Growth	Annual GDP growth per capita	2.96	2.54	-67.29	141.63	Penn Table
Young	Age dependency ratio	54.13	49.92	14.92	107.40	WDI
Fert	Fertility rate	3.23	2.69	1.09	7.7	WDI
Life	Life expectancy	68.30	71.21	26.17	84.36	WDI
Pop	Population growth	1.70	1.51	-6.54	19.14	Penn Table
Invpri	Price level of investment	0.54	0.50	0.01	7.98	Penn Table
Con	Consumption share	0.64	0.65	0.09	1.56	Penn Table
Gov	Government consumption share	0.17	0.17	0.01	0.75	Penn Table
Inv	Investment share	0.22	0.22	0.00	0.92	Penn Table

Table 2.3: Summary statistics and data sources of dependent and independent variables.

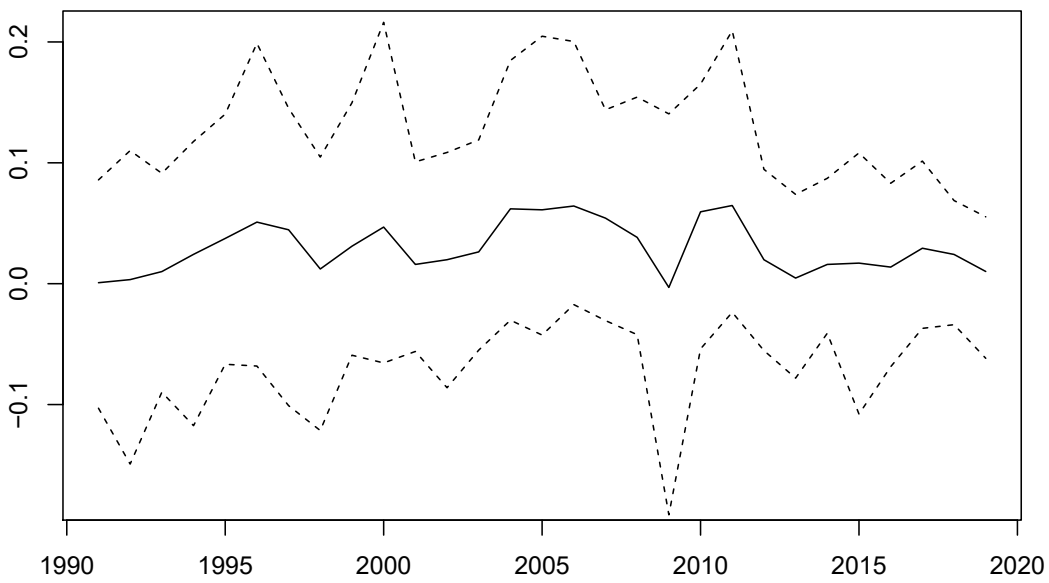


Figure 2.3: Time series of average annual real GDP growth rate per capita (solid line) and time series of 5% and 95%-quantiles (dashed lines).

2.5.2 Estimation Results

We first fit our projection-based interactive fixed effects model using the complete sample of $N = 129$ countries and the complete list of regressors. The estimation results can be found in Table 2.4. Consumption share, government consumption share and fertility rate have a significant negative impact on the growth rates, while the age dependency ratio has a significant positive influence. We also estimate a restricted model which only considers the significant variables from the full model. The parameter estimates and standard errors do not change substantially compared to the full model. Our list of significant variables overlaps with those identified by Lu and Su (2016), however they also include the investment share but do not include the fertility rate.

	Con	Gov	Inv	Invpri	Young	Fert	Life	Pop
Estimate	-0.0583	-0.0855	0.0194	-0.0008	0.0008	-0.0126	-0.0003	0.1107
t-statistic	-4.3501	-3.0777	0.8872	-0.1221	3.2312	-3.4471	-0.8056	0.5996
Estimate	-0.0625	-0.0939	-	-	0.0008	-0.0114	-	-
t-statistic	-5.0201	-3.5758	-	-	3.6343	-3.0265	-	-

Table 2.4: Estimation results for the projected IFE estimator based on the whole sample.

In contrast to standard panel models such as a country-specific fixed effects model, we are also able to estimate the latent factors and corresponding factor loadings. We select $K = 3$ as the number of factors, according to the procedure based on the ratio of eigenvalues described in equation (2.17). See Figure 2.4 for a visualization of the

eigenvalues. The three estimated latent factors can be found in Figure 2.6. They can be interpreted as unobserved macro risk factors and the loadings measure the exposure of a given country to these risk factors.

We now restrict our analysis to the subset of countries which are members of the OECD (Organisation for Economic Cooperation and Development). The estimation results can be found in Table 2.5. The signs of the estimated parameters are consistent with the previous regression based on the full sample, however the population growth now has a negative sign. The list of significant variables additionally includes the investment share, whereas the government consumption share becomes insignificant. The restricted model, for which we only include the significant variables, again does not deviate strongly from the full model.

	Con	Gov	Inv	Invpri	Young	Fert	Life	Pop
Estimate	-0.0997	-0.0485	0.1460	-0.0113	0.0018	-0.0262	-0.0015	-0.7994
t-statistic	-3.6030	-1.3886	4.2925	-1.1768	3.2114	-2.8924	-2.0314	-2.0876
Estimate	-0.0845	-	0.1155	-	0.0028	-0.0515	-	-
t-statistic	-3.3320	-	3.8613	-	6.0755	-6.3310	-	-

Table 2.5: Estimation results for the projected IFE estimator based on the sample OECD countries.

The number of factors for the OECD sample is $K = 4$, based on the ratio of eigenvalues (see Figure 2.5). Figure 2.7 shows the estimated first and second latent factors of the OECD sample. The first factor clearly represents a risk factor for the overall market condition. It increases after the bust of the dot-com bubble in 2000 and it has another sharp peak in the aftermath of the financial crisis in 2009. For all 30 OECD countries in our sample, the sign of the loading parameters associated with the first factor is negative. This implies that a positive-valued shock to the first factor leads to a reduction in the GDP growth rate for all OECD countries. The interpretation of the second factor requires a little more attention, as the factor loadings have different signs for different countries. Interestingly, four of the five countries with the largest estimated loading parameters belong to the list of PIIGS states, which suffered the most during the Euro crisis. Accordingly, we can observe that the second factor takes a negative value in the beginning of the Euro crisis in 2010.

The fundamental tenet of our model is that the covariates are assumed to have sufficient explanatory power on the latent factor loadings. It is thus a crucial task to check whether this is indeed the case. For this purpose, we take a closer look at the two

components of the matrix of estimated loadings. Recall the following decomposition,

$$\widehat{\Lambda} = \widehat{G}(\bar{X}) + \widehat{\Gamma},$$

where $\widehat{G}(\bar{X})$ represents the estimated systematic part of the loadings, i.e. the part which is explained by the covariates, and Γ represents the estimated random part. In Table 2.6 we calculate the Frobenius norm and the sup norm for both matrices as measures of their relative importance. We calculate the norms both for the full sample and the sample of OECD countries. It is evident that the systematic part dominates the random part. In the OECD sample, the loadings are almost completely explained by the $G(\bar{X})$, i.e. all γ_{ik} are very close to zero. This provides evidence for the validity of our projection-based approach to interactive fixed effects.

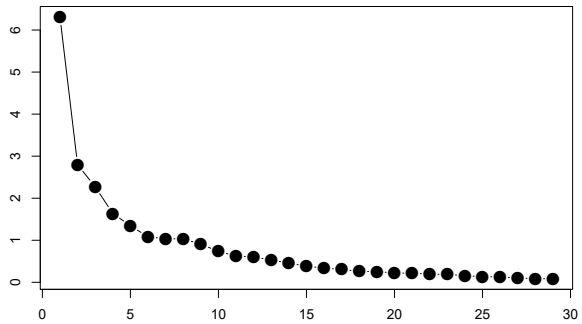


Figure 2.4: Eigenvalues of the projected PCA algorithm for the whole sample.

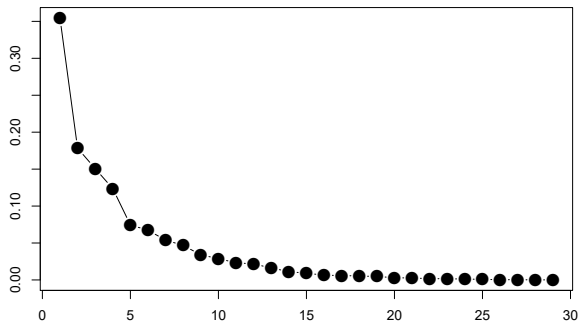


Figure 2.5: Eigenvalues of the projected PCA algorithm for the subsample of OECD countries.

	All countries		OECD countries	
	$\lVert\cdot\rVert_F$	$\lVert\cdot\rVert_{\max}$	$\lVert\cdot\rVert_F$	$\lVert\cdot\rVert_{\max}$
$\widehat{G}$	0.6686	0.3111	0.1168	0.0461
$\widehat{\Gamma}$	0.0650	0.0128	0.0000	0.0000

Table 2.6: Estimation results for the two components of the factor loadings, the systemic part $\widehat{G}$ and the random part $\widehat{\Gamma}$.

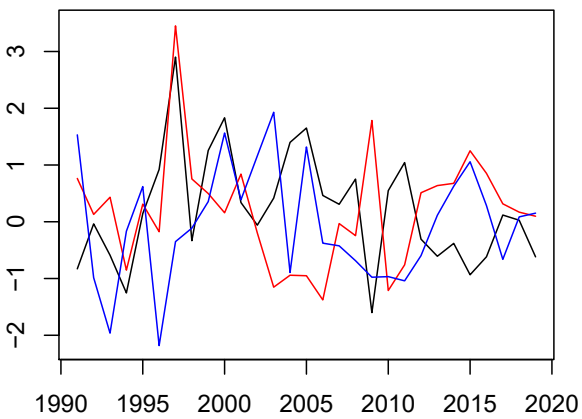


Figure 2.6: Estimated three factors for the whole sample, f_1, f_2 and f_3.

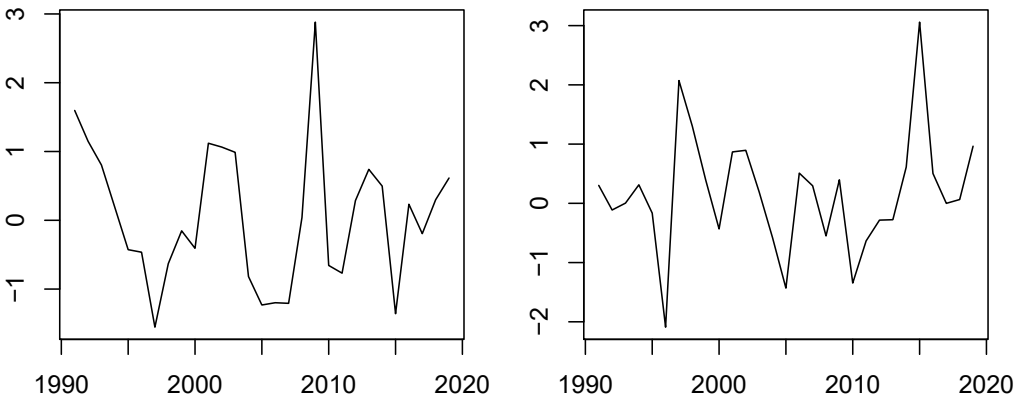


Figure 2.7: Estimated first (left panel) and second factor (right panel) based on the OECD subsample.

2.6 Conclusion

We propose a new estimator for the regression parameters in a panel data model with interactive fixed effects. The key idea of our estimator is the projection of the data onto the linear sieve space spanned by the covariates. Following the Frisch-Waugh approach, the estimator takes the form of an partial least squares estimator, which partials out the effect of the latent factors. We show that our estimator is $\sqrt{NT}$-consistent with an asymptotic normal distribution. In the special case of heteroskedasticity and if the loadings can be completely explained by the covariates, our estimator reaches the semiparametric efficiency bound. An important advantage of our estimator is that it does not require the estimation of the number of factors in advance.

There are several open topics for further research. First, specification tests for the different components of the factor loadings could be considered. In particular, one

could test hypotheses of the form $H_0 : \Gamma = 0$ vs. $H_1 : \Gamma \neq 0$. Alternatively, one could also consider hypotheses of the form $H_0 : G(\bar{X}) = 0$ vs. $H_1 : G(\bar{X}) \neq 0$. Such tests can provide information about the extent to which the factor loadings can be modeled by the covariates and ultimately, whether our estimator is suitable.

Another possible extension would be to relax the assumption on the additivity of the nonparametric functions. Instead, one could consider to move to the double/debiased machine learning framework of Chernozhukov et al. (2018). While in the additive case, we were able to rely on linear sieves such as splines and polynomials, it might be beneficial to use machine learning methods such as random forests or neural networks to partial out the effect of the covariates on the factor loadings.

2.A Proofs for Section 2.3

In order to show Theorem 2.3.1 we first present a Lemma.

Lemma 2.A.1. *Let*

$$\widetilde{\beta} = \left[\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} X_t\right]^{-1} \sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} y_t, \tag{2.24}$$

where

$$\begin{aligned} P_\Phi(\mu) &= \Phi(\mu)\left\{\Phi(\mu)^\top \Phi(\mu)\right\}^{-1} \Phi(\mu)^\top, \\ \Phi(\mu) &= (\phi(\mu_1), \ldots, \phi(\mu_N))^\top, \\ \phi(\mu_i)^\top &= (\phi_1(\mu_{i1}), \ldots, \phi_{J_g}(\mu_{i1}), \ldots, \phi_1(\mu_{iQ}), \ldots, \phi_{J_g}(\mu_{iQ})). \end{aligned}$$

Then, under assumptions 2.3.1 to 2.3.8 and if $J_h \sim N^{1/2}$, $J_g \sim N^{1/2}$, $T/N \to 0$ we have that, as both N and T tend to infinity,

$$\sqrt{NT}\left(\widetilde{\beta} - \beta\right) \xrightarrow{\mathcal{L}} N\left(0, \widetilde{V}\right),$$

where

$$\widetilde{V} = \widetilde{V}_\pi^{-1}\left(\widetilde{V}_\Gamma + \widetilde{V}_u\right)\widetilde{V}_\pi^{-1}, \tag{2.25}$$

with

$$\widetilde{V}_\pi = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{E}\left(\pi_{it}\pi_{it}^\top\right),$$
$$\widetilde{V}_\Gamma = \lim_{NT\to\infty} \frac{1}{NT} \sum_{t=1}^{T} \mathrm{E}\left(\pi_t^\top \Gamma\Gamma^\top \pi_t\right),$$
$$\widetilde{V}_u = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{t=1}^{T} \mathrm{E}\left(\pi_t^\top \operatorname{diag}\left\{u_t u_t^\top\right\} \pi_t\right).$$

2.A.1 Proof of Lemma 2.A.1

If we substitute (2.5) into (2.24) and we rearrange terms we obtain:

$$\begin{aligned}\widetilde{\beta} - \beta = &\left[\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} X_t\right]^{-1} \sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} G\left(\bar{X}\right) f_t \\ &+ \left[\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} X_t\right]^{-1} \sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} \Gamma f_t \\ &+ \left[\sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} X_t\right]^{-1} \sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} u_t.\end{aligned}$$

We first show that

$$\frac{1}{NT} \sum_{t=1}^{T} X_t^\top \left\{I_N - P_\Phi(\mu)\right\} X_t = \frac{1}{NT} \sum_t \pi_t^\top \pi_t + o_p(1). \tag{2.26}$$

Note that, by Assumption 2.3.1,

$$X_t = H\left(\bar{X}\right) + \pi_t, \quad t = 1, \ldots, T, \tag{2.27}$$

and an element (i, q) of $H\left(\bar{X}\right)$ is $h_q(\bar{x}_{i.}) = \sum_{q'} h_{qq'}(\bar{x}_{i.,q'})$. Now, by Assumption 2.3.3 and a Taylor expansion we have that

$$X_t = H(\mu) + \pi_t + \mathcal{O}_p\left(T^{-1/2}\right), \quad t = 1, \ldots, T, \tag{2.28}$$

uniformly in t, being $h_q(\mu_i) = \sum_{q'} h_{qq'}(\mu_{iq'})$. Hence, substituting (2.28) into the left hand side of (2.26) and making $M_\Phi(\mu) = I_N - P_\Phi(\mu)$ we obtain that

$$\begin{aligned}\frac{1}{NT} \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t = &\frac{1}{N} H(\mu)^\top M_\Phi(\mu) H(\mu) + \frac{1}{NT} \sum_t \pi_t^\top M_\Phi(\mu) \pi_t \\ &+ 2H(\mu)^\top M_\Phi(\mu) \frac{1}{NT} \sum_t \pi_t + \mathcal{O}_p\left(N^{-1}T^{-1/2}\right) + o_p\left(N^{-1}T^{-1/2}\right).\end{aligned} \tag{2.29}$$

Using Assumption 2.3.3, $h_{qq'}(\cdot)$ can be approximated by some spline function, that is,

$$h_{qq'}(\mu_{iq'}) = \sum_{\ell=1}^{J_h} c_{\ell,qq'}\phi_\ell(\mu_{iq'}) + R_{qq'}(\mu_{iq'}), \quad q, q' = 1, \ldots, Q. \tag{2.30}$$

where $\phi_\ell(\cdot)$'s are the spline basis functions. The $c_{\ell,kq}$'s are the sieve coefficients of the q-th additive component of $h_{qq'}(\cdot)$, and $R_{qq'}(\cdot)$ is the remainder term that represents the approximation error. Also, J_h denotes the number of sieve terms which grows slowly as $N \to \infty$.

For the sake of simplicity, we take the same basis functions in (2.7). For each $q, q' \leq Q$ and $i \leq N$, let us define

$$\begin{aligned} c_q^\top &= (c_{1,q1}, \ldots, c_{J_h,q1}, \ldots, c_{1,qQ}, \ldots, c_{J_h,qQ}) \in \mathbb{R}^{J_h Q}, \\ \phi(\mu_i)^\top &= (\phi_1(\mu_{i1}), \ldots, \phi_{J_h}(\mu_{i1}), \ldots, \phi_1(\mu_{iQ}), \ldots, \phi_{J_h}(\mu_{iQ})) \in \mathbb{R}^{J_h Q}. \end{aligned}$$

Thus, equation (2.30) can be rewritten as

$$h_q(\mu_i) = \phi(\mu_i)^\top c_q + \sum_{q'=1}^{Q} R_{qq'}(\mu_{iq'}). \tag{2.31}$$

By considering (2.31) in matrix form we obtain

$$H(\mu) = \Phi(\mu)C + R(\mu), \tag{2.32}$$

where $\Phi(\mu) = (\phi(\mu_1), \ldots, \phi(\mu_N))^\top$ is a $N \times J_h Q$ matrix of basis functions, $C = (c_1, \ldots, c_Q)$ is a $J_h Q \times Q$ matrix of sieve coefficients, and $R(\mu)$ is a $N \times Q$ matrix with the (i, q)-th element $\sum_{q'=1}^{Q} R_{qq'}(\mu_{iq'})$.

Using (2.32), the first term of the right hand side of (2.29) is then

$$\frac{1}{N} H(\mu)^\top M_\Phi(\mu) H(\mu) = \frac{1}{N} \{H(\mu) - \Phi(\mu)C\}^\top M_\Phi(\mu) \{H(\mu) - \Phi(\mu)C\}.$$

Let $H(\mu) = (H_1, \ldots, H_Q)$. It is easy to see that the q-th element of the right hand side

of the above term can be written as

$$\begin{aligned}
&\frac{1}{N}\left\{H_q(\mu)-\Phi(\mu)c_q\right\}^\top M_\Phi(\mu)\left\{H_q(\mu)-\Phi(\mu)c_q\right\}\\
&\leq \lambda_{max}\left\{M_\Phi(\mu)\right\}\frac{1}{N}\sum_{i=1}^{N}\left\{h_q(\mu_i)-\phi(\mu_i)^\top c_q\right\}^2\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left\{h_q(\mu_i)-\phi(\mu_i)^\top c_q\right\}^2\\
&\leq \max_i\left\{h_q(\mu_i)-\phi(\mu_i)^\top c_q\right\}^2=\mathcal{O}\left(\rho_{h,N}^2\right),
\end{aligned}$$

where $\rho_{h,N}=\sup_{\mu\in\mathcal{M}}\left|h_q(\mu)-\phi(\mu)^\top c_q\right|$. Note that under Assumption 2.3.3 then $\rho_{h,N}=\mathcal{O}\left(J_h^{-2}\right)$ (see Schumaker (1980), Theorem 6.27) and therefore,

$$\frac{1}{N}H(\mu)^\top M_\Phi(\mu)H(\mu)=\mathcal{O}\left(J_h^{-4}\right).$$

Furthermore, following the same line as in the proof before and assuming that $\frac{J_g}{N}\to 0$ and $TJ_g\to\infty$ we have that

$$\frac{1}{NT}\sum_t \pi_t^\top M_\Phi(\mu)\pi_t=\frac{1}{NT}\sum_t \pi_t^\top\pi_t+o_p(1). \tag{2.33}$$

This follows from

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\pi_t^\top P_\Phi(\mu)\pi_t &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}\pi_{it}^\top\phi(\mu_i)\left\{\frac{1}{N}\Phi(\mu)^\top\Phi(\mu)\right\}^{-1}\frac{1}{N}\sum_{i=1}^{N}\phi(\mu_i)^\top\pi_{it}\\
&= mathcalO_p\left(\frac{J_h}{N}\right).
\end{aligned}$$

Finally, noting that for the q-th element of the r.h.s. of equation (2.29)

$$\mathrm{Var}\left[\left\{H_q(\mu)-\Phi(\mu)c_q\right\}^\top M_\Phi(\mu)\frac{1}{NT}\sum_t\pi_t\right]=\mathcal{O}\left(\frac{1}{NT}\rho_{h,N}^2\right),$$

this implies that

$$H(\mu)^\top M_\Phi(\mu)\frac{1}{NT}\sum_t\pi_t=\mathcal{O}\left(\frac{1}{\sqrt{NT}}\rho_{h,N}\right). \tag{2.34}$$

This closes the proof of (2.26). Following the same line we can show that,

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}X_t^\top\left\{I_N-P_\Phi(\mu)\right\}G\left(\bar{X}\right)f_t=o_p(1). \tag{2.35}$$

Note that, by Assumptions 2.3.2 and 2.3.3 and Taylor expansion

$$\begin{aligned}&\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}X_t^\top M_\Phi(\mu)G\left(\bar{X}\right)f_t\\&=\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}X_t^\top M_\Phi(\mu)G\left(\mu\right)f_t+\mathcal{O}_p\left(N^{-1}T^{-1/2}\right)+o_p\left(N^{-1}T^{-1/2}\right).\end{aligned}$$

Furthermore, by (2.28) we have that

$$\begin{aligned}&\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}X_t^\top M_\Phi(\mu)G\left(\bar{X}\right)f_t\\&=\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\left\{H\left(\mu\right)-\Phi\left(\mu\right)C\right\}^\top M_\Phi(\mu)\left\{G\left(\mu\right)-\Phi\left(\mu\right)B\right\}f_t\\&+\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\pi_t^\top M_\Phi(\mu)\left\{G\left(\mu\right)-\Phi\left(\mu\right)B\right\}f_t+o_p\left(1\right).\end{aligned}$$

It is easy to show that

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\left\{H\left(\mu\right)-\Phi\left(\mu\right)C\right\}^\top M_\Phi(\mu)\left\{G\left(\mu\right)-\Phi\left(\mu\right)B\right\}f_t=\mathcal{O}\left(\rho_{h,N}\rho_{g,N}\right).$$

where $\rho_{g,N}=\sup_{\mathcal{M}}\left|g_k\left(\mu_i\right)-\phi\left(\mu_i\right)^\top b_k\right|$, and

$$\begin{aligned}b_k^\top&=\left(b_{1,k1},\ldots,b_{J_g,k1},\ldots,b_{1,kQ},\ldots,b_{J_g,kQ}\right)\in\mathbb{R}^{J_gQ},\\\phi(\mu_i)^\top&=\left(\phi_1(\mu_{i1}),\ldots,\phi_{J_g}(\mu_{i1}),\ldots,\phi_1(\mu_{iQ}),\ldots,\phi_{J_g}(\mu_{iQ})\right)\in\mathbb{R}^{J_gQ}.\end{aligned}$$

Furthermore, the q-th element of $\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\pi_t^\top M_\Phi(\mu)\left\{G\left(\mu\right)-\Phi\left(\mu\right)B\right\}f_t$ is

$$\mathrm{Var}\left[\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\pi_{qt}^\top M_\Phi(\mu)\left\{G\left(\mu\right)-\Phi\left(\mu\right)B\right\}f_t\right]=\mathcal{O}\left(\rho_{g,N}^2\right),$$

and therefore,

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\pi_{qt}^\top M_\Phi(\mu)\left\{G\left(\mu\right)-\Phi\left(\mu\right)B\right\}f_t=\mathcal{O}\left(\rho_{g,N}\right).$$

Finally, using (2.32) we show that

$$\begin{aligned}&\frac{1}{\sqrt{NT}}\sum_{t}X_t^\top M_\Phi\left(\mu\right)\Gamma f_t\\&=\frac{1}{\sqrt{NT}}\sum_{t}\left\{H\left(\mu\right)-\Phi\left(\mu\right)C\right\}^\top M_\Phi\left(\mu\right)\Gamma f_t\\&+\frac{1}{\sqrt{NT}}\sum_{t}\pi_t^\top M_\Phi\left(\mu\right)\Gamma f_t,\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{\sqrt{NT}}\sum_t X_t^\top M_\Phi(\mu)\, u_t \\
&= \frac{1}{\sqrt{NT}}\sum_t \{H(\mu) - \Phi(\mu)\, C\}^\top M_\Phi(\mu)\, u_t \\
&+ \frac{1}{\sqrt{NT}}\sum_t \pi_t^\top M_\Phi(\mu)\, u_t.
\end{aligned}$$

Following the same lines as before, by assumptions 2.3.5 and 2.3.6,

$$\frac{1}{\sqrt{NT}}\sum_t \{H(\mu) - \Phi(\mu)\, C\}^\top M_\Phi(\mu)\, \Gamma f_t = \mathcal{O}(\rho_{h,N}\nu_N), \tag{2.36}$$

and

$$\frac{1}{\sqrt{NT}}\sum_t \{H(\mu) - \Phi(\mu)\, C\}^\top M_\Phi(\mu)\, u_t = \mathcal{O}(\rho_{h,N}). \tag{2.37}$$

Finally,

$$\frac{1}{\sqrt{NT}}\sum_t X_t^\top M_\Phi(\mu)\{\Gamma f_t + u_t\} = \frac{1}{\sqrt{NT}}\sum_t \pi_t^\top\{\Gamma f_t + u_t\} + o_p(1) \tag{2.38}$$

and applying the central limit theorem of Bradley Jr (1981) we have that

$$\frac{1}{\sqrt{NT}}\sum_t \pi_t^\top\{\Gamma f_t + u_t\} \xrightarrow{\mathcal{L}} N\left(0, \widetilde{V}_\Gamma + \widetilde{V}_u\right),$$

where

$$\begin{aligned}
\widetilde{V}_\Gamma &= \lim_{N\to\infty}\frac{1}{N}E\left(\pi_t^\top\Gamma\Gamma^\top\pi_t\right), \\
\widetilde{V}_u &= \lim_{N,T\to\infty}\frac{1}{NT}\sum_{t=1}^{T}\mathrm{E}\left(\pi_t^\top u_t u_t^\top \pi_t\right).
\end{aligned}$$

∎

2.A.2 Proof of Theorem 2.3.1

According to the definitions of $\widehat{\beta}$ and $\widetilde{\beta}$, together with the equality

$$a_1a_2 - b_1b_2 = (a_1 - b_1)(a_2 - b_2) + (a_1 - b_1)b_2 + b_1(a_2 - b_2),$$

we obtain

$$\begin{aligned}
\widehat{\beta} - \beta = \widetilde{\beta} - \beta &+ \left[\left(\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) X_t\right)^{-1} - \left(\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right)^{-1}\right] \\
&\times \left[\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) G\left(\bar{X}\right) f_t - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) G\left(\bar{X}\right) f_t\right] \\
&+ \left[\left(\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) X_t\right)^{-1} - \left(\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right)^{-1}\right] \times \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) G\left(\bar{X}\right) f_t \\
&+ \left(\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right)^{-1} \times \left[\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) G\left(\bar{X}\right) f_t - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) G\left(\bar{X}\right) f_t\right] \\
&+ \left[\left(\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) X_t\right)^{-1} - \left(\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right)^{-1}\right] \\
&\times \left[\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) \left\{\Gamma f_t + u_t\right\} - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) \left\{\Gamma f_t + u_t\right\}\right] \\
&+ \left[\left(\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) X_t\right)^{-1} - \left(\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right)^{-1}\right] \times \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) \left\{\Gamma f_t + u_t\right\} \\
&+ \left(\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right)^{-1} \times \left[\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X}) \left\{\Gamma f_t + u_t\right\} - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu) \left\{\Gamma f_t + u_t\right\}\right].
\end{aligned}$$

We will have shown the desired result if we prove that

$$\widehat{\beta} - \beta = \widetilde{\beta} - \beta + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{2.39}$$

We already showed in the proof of Lemma 2.A.1,

$$\begin{aligned}
\frac{1}{NT}\left\{\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) X_t\right\} &= \mathcal{O}_p(1), \\
\frac{1}{NT}\left\{\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) G\left(\bar{X}\right) f_t\right\} &= o_p\left(1/\sqrt{NT}\right), \\
\frac{1}{NT}\left[\sum_{t=1}^{T} X_t^\top M_\Phi(\mu) \left\{\Gamma f_t + u_t\right\}\right] &= \mathcal{O}_p\left(1/\sqrt{NT}\right).
\end{aligned}$$

It remains to show the following results.

$$\frac{1}{NT}\left\{\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X})X_t - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu)X_t\right\} = \mathcal{O}_p(1),$$
$$\frac{1}{NT}\left\{\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X})G\left(\bar{X}\right)f_t - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu)G\left(\bar{X}\right)f_t\right\} = o_p\left(1/\sqrt{NT}\right),$$
$$\frac{1}{NT}\left[\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X})\left\{\Gamma f_t + u_t\right\} - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu)\left\{\Gamma f_t + u_t\right\}\right] = o_p\left(1/\sqrt{NT}\right).$$

The first results follows from Assumption 2.3.2 and the continuous mapping theorem,

$$\frac{1}{NT}\left\{\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X})X_t - \sum_{t=1}^{T} X_t^\top M_\Phi(\mu)X_t\right\} = \mathcal{O}_p\left(1/\sqrt{T}\right).$$

For the second result it is sufficient to show that

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X})G\left(\bar{X}\right)f_t = o_p(1).$$

We can again write,

$$\begin{aligned}
&\frac{1}{\sqrt{NT}}\sum_{t=1}^{T} X_t^\top M_\Phi(\bar{X})G\left(\bar{X}\right)f_t \\
&\quad = \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\left\{H(\bar{X}) - \Phi(\bar{X})C\right\}^\top M_\Phi(\bar{X})\left\{G(\bar{X}) - \Phi(\bar{X})B\right\}f_t \\
&\quad + \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\pi_t M_\Phi(\bar{X})\left\{G(\bar{X} - \Phi(\bar{X})B)\right\}f_t.
\end{aligned}$$

The first term on the right hand side is of order $o_p(\rho_{g,N}\rho_{h,N})$, the second term is of order $o_p(\rho_{g,N})$. Finally, for the third result, we have to consider the variance of the q-th element,

$$\mathrm{Var}\left[\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\pi_{tq}\left\{M(\bar{X}) - M(\mu)\right\}(\Gamma f_t + u_t)\right] = \mathcal{O}_p(1/T).$$

By Cauchy-Schwarz the term is of order $\mathcal{O}_p(1/\sqrt{T})$.

∎

2.A.3 Proof of Theorem 2.3.2

The proof follows from Theorem 4.1 of Fan et al. (2016) and from the $\sqrt{NT}$-consistency result of Theorem 2.3.1.

2.A.4 Proof of Proposition 2.3.1

By proof the Proof of Lemma 2.A.1, and Assumption 2.3.4 we have

$$\widehat{V}_\pi = \frac{1}{NT}\sum_{t=1}^{T}\pi_t^\top \pi_t + o_p(1) \xrightarrow{p} \widetilde{V}_p.$$

Similarly,

$$\widehat{V}_u = \frac{1}{N}\sum_{t=1}^{T}\pi_t^\top \operatorname{diag}\left\{\widehat{u}_{1t}^2, \ldots, \widehat{u}_{Nt}^2\right\}\pi_t + o_p(1).$$

Without loss of generality, assume that $Q = K = 1$. Then, plugging in $\widehat{u}_{it}$,

$$\begin{aligned}
\widehat{V}_u &= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\pi_{it}^2\widehat{u}_{it}^2 \\
&= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\pi_{it}^2 u_{it}^2 + \pi_{it}^4(\beta-\widehat{\beta})^2 + 2\pi_{it}^3(\beta-\widehat{\beta})u_{it} \\
&+ \pi_{it}^2(\lambda_i f_t - \widehat{\lambda}_i\widehat{f}_t) + 2\pi_{it}^3(\beta-\widehat{\beta})(\lambda_i f_t - \widehat{\lambda}_i\widehat{f}_t) + 2\pi_{it}^2(\lambda_i f_t - \widehat{\lambda}_i\widehat{f}_t)u_{it}.
\end{aligned}$$

By Theorem 2.3.1, $\widehat{\beta} \xrightarrow{p} \beta$, by Theorem 2.3.2, $\widehat{\lambda}_i \xrightarrow{p} \lambda_i$ and $\widehat{f}_t \xrightarrow{p} f_t$. The desired result follows from Assumption 2.3.4 and the continuous mapping theorem.

$$\widehat{V}_u = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\pi_{it}^2 u_{it}^2 + o_p(1) \xrightarrow{p} \widetilde{V}_u.$$

Finally, we have

$$\widehat{V}_\Gamma = \frac{1}{NT}\sum_{t=1}^{T}\pi_t^\top \operatorname{diag}\left\{\widehat{\Gamma}\widehat{\Gamma}^\top\right\}\pi_t + o_p(1).$$

By Theorem 2.3.2 we have $\widehat{\Gamma} \xrightarrow{p} \Gamma$, and by Assumption 2.3.4 we have

$$\frac{1}{NT}\sum_{t=1}^{T}\pi_t^\top \operatorname{diag}\left\{\widehat{\Gamma}\widehat{\Gamma}^\top - \Gamma\Gamma^\top\right\}\pi_t \xrightarrow{p} 0.$$

Therefore $\widehat{V}_\Gamma \xrightarrow{p} \widetilde{V}_\Gamma$.

∎

Bibliography

Acemoglu, D., Naidu, S., Restrepo, P., & Robinson, J. A. (2019). Democracy does cause growth. *Journal of Political Economy*, *127*(1), 47–100.

Ahn, S. C., Lee, Y. H., & Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of econometrics*, *174*(1), 1–14.

Arellano, M. (2003). *Panel data econometrics*. Oxford university press.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*, 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279.

Bai, J., & Li, K. (2014). Theory and methods of panel data models with interactive effects. *The Annals of Statistics*, *42*(1), 142–170.

Bai, J., & Ng, S. (2013). Principal components estimaton and identification of static factors. *Journal of Econometrics*, *176*, 18–29.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The quarterly journal of economics*, *106*(2), 407–443.

Bradley Jr, R. C. (1981). Central limit theorems under weak dependence. *Journal of Multivariate Analysis*, *11*(1), 1–16.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, 567–596.

Chen, X. (2007). Handbook of econometrics. North Holland.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Connor, G., Hagmann, M., & Linton, O. (2012). Efficient semiparametric estimation of the fama-french model and extensions. *Journal of the American Statistical Association*, *107*, 135–151.

Connor, G., & Linton, O. (2007). Semiparametric estimation of a characteristic-based factor model of stock returns. *Journal of Empirical Finance*, *14*, 694–717.

de Boor, C. (1978). *A practical guide to splines* (C. de Boor, Ed.). Springer.

Durlauf, S. N., Johnson, P. A., & Temple, J. R. (2005). Growth econometrics. *Handbook of economic growth*, *1*, 555–677.

Fan, J., Liao, Y., & Wang, W. (2016). Projected principal component analysis in factor models. *The Annals of Statistics*, *44*(1), 2019–254.

Härdle, W., Liang, H., & Gao, J. (2012). *Partially linear models*. Springer Science & Business Media.

Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica: Journal of the econometric society*, 1371–1395.

Islam, N. (1995). Growth empirics: A panel data approach. *The quarterly journal of economics*, *110*(4), 1127–1170.

Li, Q. (2000). Efficient estimation of additive partially linear models. *International Economic Review*, *41*(4), 1073–1092.

Lorentz, G. (1986). *Approximation of functions* (G. Lorentz, Ed.; 2nd). Chelsea Publishing.

Lu, X., & Su, L. (2016). Shrinkage estimation of dynamic panel data models with interactive fixed effects. *Journal of Econometrics*, *190*(1), 148–175.

Moon, H. R., & Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, *83*(4), 1543–1579.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69–85.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, *5*(2), 99–135.

Newey, W. K., & West, K. D. (1986). *A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix* (tech. rep.). National Bureau of Economic Research.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, *74*(4), 967–1012.

Schumaker, L. (1980). *Spline functions: Basic theory* (L. Schumaker, Ed.). Pure; applied mathematics. A Wiley-Interscience Publication, Wiley.

Stock, J., & Watson, M. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.

Chapter 3

Recursive Quantile Estimation: Non-Asymptotic Confidence Bounds

3.1 Introduction

The emergence of big data has brought serious challenges to traditional deterministic optimization methods. In many applications, the data arrives sequentially and the sample size is so large that a storage of the entire dataset is infeasible. In these situations, the stochastic gradient descent (SGD) algorithm (Kiefer, Wolfowitz, et al., 1952; Robbins & Monro, 1951) provides a scalable alternative for estimation. The algorithm updates estimates recursively according to the gradient of the objective function. This recursive nature of the SGD algorithm makes it computationally and memory efficient. Thus, SGD is naturally suited for online learning problems. Notable applications include anomaly detection (Ahmad et al., 2017) and matrix factorization (Mairal et al., 2010). The large sample properties of SGD are well established. For an averaged version of the algorithm (Polyak & Juditsky, 1992; Ruppert, 1988), it can be shown that the estimator converges with the optimal parametric rate to a Gaussian limit. To conduct inference, Chen et al. (2020) and Zhu et al. (2020) propose methods to estimate the covariance matrix of the parameter estimates. Fang et al. (2018) and Fang (2019) propose bootstrap procedures to measure the uncertainty of SGD estimates.

We consider the recursive estimation of quantiles. This classical problem is of great importance in a variety of applications ranging from finance (Engle & Manganelli, 2004), health care (Wang et al., 2018) and survival studies (Peng & Huang, 2008). The large sample properties of the traditional estimator for the quantile, which is based on order statistics, were studied in Bahadur (1966) and Kiefer (1967). The downside of the empirical estimator is that it is not memory-efficient in the presence of large and sequentially arriving datasets. Moreover, asymptotic normality does not provide any insights on the performance of the estimator in finite samples. The study of the finite-sample behavior is an important task since in practical problems the sample size is always finite. Usually, obtaining such results requires more mathematical effort

than merely obtaining asymptotic results and typically this involves more restrictive assumptions on the tail behavior and the existence of moments.

The aim of this paper is to study the tail probability of the averaged version of the SGD algorithm for estimating quantiles in finite samples. As our main result, we derive an exponential bound on the tail probability, while only imposing weak assumption on the smoothness of the distribution function. The proof relies on the decomposition of the gradient in the SGD algorithm into a martingale difference part, a shift part and a remainder part. Another key component of the proof is a bound on the moment generating function of the SGD estimate.

The non-asymptotic behavior of the SGD estimate of quantiles with Polyak-Ruppert averaging was studied in Costa and Gadat (2020). They derived finite sample bounds for the L^p loss. Another closely related paper is Cardot et al. (2013), who proposed a SGD estimation procedure for the geometric median (Haldane, 1948), which is a multi-dimensional generalization of the median. The SGD solution has the same asymptotic behavior as the empirical estimator of the geometric median. The result can be easily generalized to the geometric quantiles proposed by Chaudhuri (1996). In a subsequent paper, Cardot et al. (2017) studied the finite sample performance of the SGD algorithm. In particular, they derived non-asymptotic confidence balls for the averaged version of the algorithm. While the geometric median is a generalization of the classic median, their main result does not apply to this univariate case. One contribution of our paper is the extension of the result of Cardot et al. (2017) to the univariate median. It should be noted that this bound is only valid for a sample size exceeding a certain rank. In contrast, our new non-asymptotic bound is valid for each finite sample size. The reason is that the bound derived in this paper is based on a bound of the moment generating function, while previous results only relied on a finite-sample bound for the L^2 risk.

We apply our novel finite sample bound to the problem of best arm identification in the context of stochastic multi-armed bandit models. We refer to the monograph of Lattimore and Szepesvári (2020) for an overview on bandit algorithms. While the majority of the research on multi-armed bandits focused on the mean case, there are important arguments in favor of looking at the quantiles. First, quantiles are more robust location parameters compared to the mean and second, depending on the context of the application, the focus can be on different parts of the distribution. Previously, the problem of best arm identification in a quantile bandit settings were studied in Szorenyi et al. (2015), Howard and Ramdas (2019) and in Nikolakakis et al. (2021). We consider a quantile version of the successive rejects algorithm of Audibert et al. (2010) and the sequential elimination algorithm of Karnin et al. (2013).

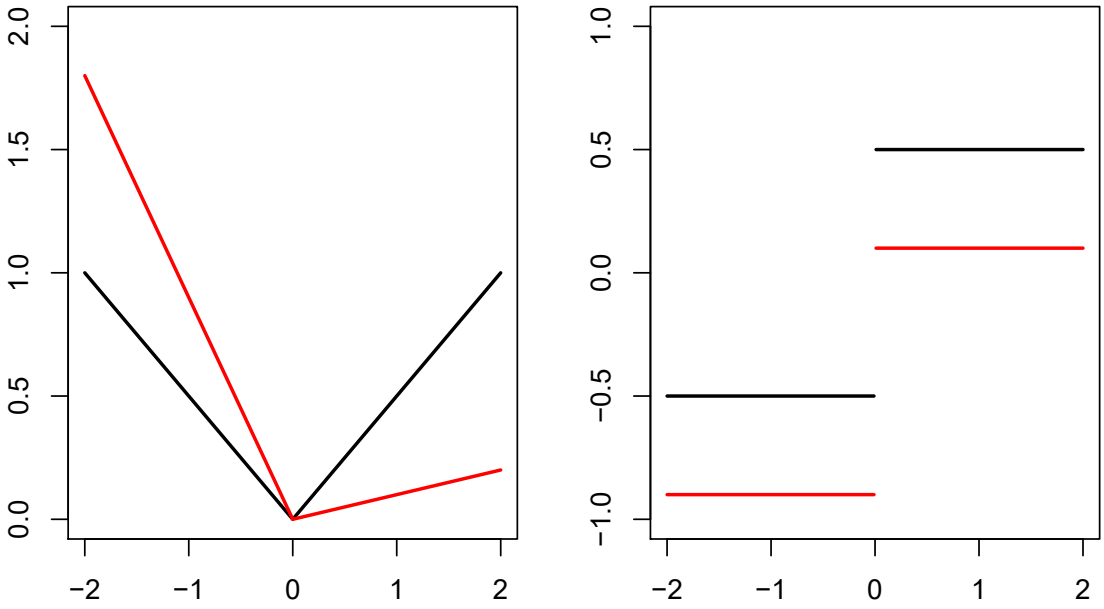


Figure 3.1: Quantile loss function (left panel) and score function (right panel) for quantile level $\tau = 0.5$ (black line) and for $\tau = 0.1$ (red line).)

The remainder of the paper is organized as follows. Section 3.2 provides an overview of the problem and introduces the SGD algorithm and its averaged version. The main theoretical results are presented in Section 3.3. In Section 3.4, we apply our probability bound to the problem of best arm identification. Section 3.5 concludes and Section 3.6 provides the proofs.

3.2 Overview of the Problem

In this paper, we are interested in estimating quantiles for high dimensional data. For a random vector $X = (X_1, X_2, \ldots, X_p)^\top \in \mathbb{R}^p$, the τ-th quantile of coordinate X_i is defined as the minimizer of the pinball loss function (see Figure 3.1),

$$Q_i(\tau) \stackrel{\text{def}}{=} \operatorname{argmin}_{x\in\mathbb{R}} \mathbb{E}\left\{(X_i - x)\left(\tau - \mathbf{1}_{x \geq X_i}\right)\right\}. \tag{3.1}$$

Denote the distribution function of X_i as $F_i(x)$, under the assumption of F_i being continuous, we have $F_i\{Q_i(\tau)\} = \tau$.

In financial applications, X_i could be the stock return of firm i and $Q_i(\tau)$ the corresponding value-at-risk (VaR) at confidence level τ.

Let $X_{i,1}, \ldots, X_{i,n}$ denote i.i.d. copies of the coordinates X_i, $1 \leq i \leq p$. Then a natural

empirical estimator of $Q_i(\tau)$ takes the form

$$\widehat{Q}_i(\tau) \stackrel{\text{def}}{=} \operatorname{argmin}_{x\in\mathbb{R}} \sum_{k=1}^{n} \left\{ (X_{i,k} - x) \left(\tau - \mathbf{1}_{X_{i,k}\geq x} \right) \right\}. \tag{3.2}$$

Asymptotic properties of the empirical estimator are extensively studied (Bahadur, 1966; Kiefer, 1967). It is well known that the estimator is strongly consistent and has an asymptotic normal distribution.

One of the problems of the empirical estimator is that it is not memory efficient in the case of streaming data, we follow a different estimation procedure. Munro and Paterson (1980) showed that any algorithm exactly calculating quantiles in p passes requires $\Omega(1/p)$ memory. Recent developments on estimating quantiles in the case of streaming data are discussed in Luo et al. (2016). Following Robbins and Monro (1951) we have, starting from a constant initial value $Y_{i,0}(\tau) = y_i$, $y_i \in \mathbb{R}$, and let

$$Y_{i,k+1}(\tau) = Y_{i,k}(\tau) + \gamma_k \left\{ \tau \mathbf{1}_{X_{i,k+1} > Y_{i,k}(\tau)} - (1-\tau) \mathbf{1}_{X_{i,k+1} \leq Y_{i,k}(\tau)} \right\}, \tag{3.3}$$

where the sequence of learning rates, (γ_k), determines the convergence of the algorithm. In particular, the following assumptions need to be fulfilled,

$$\sum_{k=1}^{\infty} \gamma_k^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma_k = \infty.$$

The first condition ensures the convergence to some point in $\mathbb{R}$, while the second condition ensures the convergence to a unique minimizer $Q_i(\tau)$. We consider sequences of step sizes in the form of $\gamma_k = (k+1)^{-\beta}$, with $1/2 < \beta \leq 1$.

Due to favorable asymptotic properties, we consider an averaged version of the algorithm, which takes the form

$$\bar{Y}_i(\tau) = \sum_{k=1}^{n} Y_{i,k}(\tau)/n,$$

where $\bar{Y}_{i,0} = 0$. Such an averaging step is known as Polyak-Ruppert averaging (Polyak & Juditsky, 1992; Ruppert, 1988). Estimators based on the averaged SGD algorithm converge almost surely to the true parameter and have the same Gaussian limit distribution as the empirical estimator. For the estimation of quantile, asymptotic normality of the solution of the averaged algorithm was shown by Bardou et al. (2009). In particular, it holds

$$\sqrt{n} \left\{ \bar{Y}_i(\tau) - Q_i(\tau) \right\} \xrightarrow{\mathcal{L}} N\left(0, \frac{\tau(1-\tau)}{f_i \left\{ Q_i(\tau) \right\}^2} \right).$$

However, these asymptotic results do not provide information on how well the estimator will perform in finite samples. Gadat and Panloup (2017) derived non-asymptotic bounds on the L^2-loss for the recursive quantile estimator based on Polyak-Ruppert averaging. It is shown that for each $n \geq 1$ it holds that, given the optimal choice of β,

$$\mathbb{E}\left\{\bar{Y}_i(\tau) - Q_i(\tau)\right\}^2 \leq \frac{\tau(1-\tau)}{f_i\left\{Q_i(\tau)\right\}^2 n} + \mathcal{O}\left(n^{-5/4}\right).$$

Recently, Cardot et al. (2017) analyzed the finite sample tail behavior of the Polyak-Ruppert algorithm for estimating the geometric median. The geometric median is a multivariate generalization of the univariate median (Haldane, 1948; Minsker et al., 2015), which can easily be generalized to geometric quantiles (Chaudhuri, 1996). The geometric median is defined by

$$m \stackrel{\text{def}}{=} \operatorname{argmin}_{x \in H} \mathbb{E}\left(\|X - x\| - \|X\|\right),$$

where X is a random variable taking values in a separable Hilbert space H with norm $\|\cdot\|$. The algorithm and its asymptotic properties are studied in Cardot et al. (2013). In particular, it was shown that the algorithm is strongly consistent and asymptotically normal. Cardot et al. (2017) further studied the non-asymptotic properties. Let $\bar{Z}_n$ denote the averaged SGD solution for the geometric median. In the main theorem of the paper, the authors derived non-asymptotic confidence balls for the averaged algorithm. Theorem 4.2 states that there exists a rank n_δ such that for all $n \geq n_\delta$ it holds that for all $\delta \in (0,1)$,

$$\mathbb{P}\left\{\|\bar{Z}_n - m\| \leq \frac{4}{\lambda_{min}\left(\frac{2}{3n} + \frac{1}{\sqrt{n}}\right)\log\left(\frac{4}{\delta}\right)}\right\} \geq 1 - \delta, \tag{3.4}$$

where λ_{min} is the lim inf of the eigenvalues of the Hessian of the loss function. Although the geometric median generalizes the univariate median, the asymptotic normality as well as the result on non-asymptotic confidence bounds only hold for the case of the dimensionality of the data being larger than 2, thus excluding the univariate median. The reason for this is condition (A3) in Cardot et al. (2017), which requires the existence of a constant C such that for all $x \in H$,

$$\mathbb{E}\left(\|X - x\|^{-2}\right) \leq C.$$

This condition does not hold for $H = \mathbb{R}^d$ with $d < 3$. As a second drawback of the tail bound in (3.4) the sample size is required to exceed a certain rank n_δ, which might be prohibitively large. The order of the rank is $\mathcal{O}((\frac{1}{\delta \log \delta})^6)$, which increases with

decreasing confidence level δ.

3.3 Theoretical Results

3.3.1 A Bound on the Moment Generating Function

In this section, we shall derive the main theoretical results of this paper. All proofs are deferred to Section 3.6. We are interested in the the tail probability

$$\mathbb{P}\left\{\max_{1\le i\le p}|\bar{Y}_i(\tau)-Q_i(\tau)|\ge x\right\}.$$

At first, we derive a non-asymptotic probability bound for the averaged SGD solution for a single coordinate X_i. A simple first result can be obtained by the finite sample bound on the L^2 risk of Gadat and Panloup (2017) and Markov's inequality. However, the resulting probability bound is only algebraically decreasing in n and x. In the following we will derive a sharper bound. The only assumption we impose is the following smoothness condition on the density of X_i, which is standard in the quantile literature.

Assumption 3.3.1. *Assume the random variable X_i has a differentiable density function $f_i(x)$, with $c_\tau \overset{def}{=} \min_{1\le i\le p}\inf_{\tau\in[\tau_0,\tau_1]} f_i\{Q_i(\tau)\}>0$ and $\max_{1\le i\le p}|f_i'|_\infty \le c_f<\infty$.*

This assumption ensures the existence of a unique theoretical quantile. We require no assumptions on the tail behavior nor on the existence of any moments of X_i. In order to derive the tail probability bound, we directly bound the moment generating function of the SGD solution without averaging, $Y_i(\tau)$, in the following Lemma.

Lemma 3.3.1. *Under Assumption 3.3.1, for $t\le cn^{(1-\beta)\beta}$ and $\tau_0\le\tau\le\tau_1$, we have*

$$\mathbb{E}\left[e^{t\{Y_{i,n}(\tau)-Q_i(\tau)\}}\right]\le c'n^\beta,$$

where $c,c'>0$ are some constants independent of i,n,τ.

A bound on the tail probability of the SGD solution without averaging follows immediately from the previous Lemma and Markov's inequality.

Proposition 1. *Under Assumption 3.3.1, we have*

$$\mathbb{P}\left\{|Y_{i,n}(\tau)-Q_i(\tau)|>x\right\}\le c'n^\beta\exp\left\{-cn^{\beta(1-\beta)}x\right\},$$

where $c,c'>0$ are some constants independent of i,n,τ.

3.3.2 Confidence Bounds for the Averaged SGD Algorithm

We now consider the averaged version of the SGD algorithm. The following theorem gives the non-asymptotic confidence bounds for the SGD algorithm with Polyak-Ruppert averaging.

Theorem 3.3.1. *Under Assumption 3.3.1, we have for $n \geq 1$,*

$$\begin{aligned}&\mathbb{P}\left\{|\bar{Y}_i(\tau) - Q_i(\tau)| > x\right\}\\ \lesssim & n^{1+\beta}\exp\left\{-cn^{(1-\beta^2)}x\right\} + n^{1+\beta}\exp\left\{-c'n^{\beta(1-\beta)}x^{1/2}\right\} + \exp\left\{-c''nx^2\right\},\end{aligned} \tag{3.5}$$

where the constants in $\lesssim$, c, c', c'' are all independent of i, τ, n, p.

The proof relies on a decomposition of the SGD algorithm into a martingale difference part, a shift part and a remainder part. Define a random variable for the gradient, $Z_{i,k+1}(\tau) \stackrel{\text{def}}{=} \tau - \mathbf{1}_{X_{i,k+1} \leq Y_(i,k)(\tau)}$. Let $\mathcal{F}_{i,k} \stackrel{\text{def}}{=} (X_{i,1}, \ldots, X_{i,k})$ be a sequence of σ-algebras. We can rewrite the SGD algorithm introduced in (3.3),

$$Y_{i,k+1}(\tau) = Y_{i,k}(\tau) + \gamma_k \left\{\xi_{i,k+1} + \mathbb{E}\left(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k}\right)\right\},$$

where $\xi_{i,k+1}(\tau) \stackrel{\text{def}}{=} Z_{i,k+1}(\tau) - \mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k})$. Note that the sequence $(\xi_{i,n})$ is a martingale difference sequence with respect to $\mathcal{F}_{i,n}$, with bounded increments, $|\xi_{i,k}| \leq 1$. Define $G_{i,\tau}(x) \stackrel{\text{def}}{=} \tau - F_i(x)$. Then we have that $G_{i,\tau}\{Y_{i,k+1}(\tau)\} = \mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k})$ and $G_{i,\tau}\{Q_i(\tau)\} = 0$. We can further decompose the SGD algorithm,

$$Y_{i,k+1}(\tau) = Y_{i,k}(\tau) + \gamma_k \left[\xi_{i,k+1} + G'_{i,\tau}\left\{Q_i(\tau)\right\} Y_{i,k}(\tau) + \rho_{i,k}\right],$$

where $\rho_{i,k} \stackrel{\text{def}}{=} G_{i,\tau}\{Y_{i,k}(\tau)\} - G'_{i,\tau}\{Q_i(\tau)\}Y_{i,k}(\tau)$. Summing up over the n iterations, using Abel's summation formula, we obtain

$$\begin{aligned}G'_{i,\tau}\left\{Q_i(\tau)\right\}\bar{Y}_{i,n}(\tau) =& n^{-1}\gamma_n^{-1}\left\{Y_{i,n+1}(\tau) - Y_{i,1}(\tau)\right\} - n^{-1}\sum_{k=1}^{n-1} Y_{i,k+1}(\tau)(\gamma_{k+1}^{-1} - \gamma_k^{-1})\\ &- \bar{\xi}_{i,n} - \bar{\rho}_{i,n},\end{aligned} \tag{3.6}$$

where $\bar{\rho}_{i,n} = n^{-1}\sum_{k=1}^{n}\rho_{i,k}$ and $\bar{\xi}_{i,n+1} = n^{-1}\sum_{k=1}^{n}\xi_{i,k+1}$. The proof relies on this decomposition of the SGD algorithm, the bound on the moment generating function obtained in Lemma 3.3.1 and Azuma's concentration inequality for martingale sequences with bounded increments.

In the following, we have a closer look at the three terms on the right hand side of the probability bound in (3.5).

Remark 1. *When $x \geq c_1 n^{-1/2}$ for some constant $c_1 > 0$, then the last term in* (3.5) *will be dominated by the second one, that is*

$$\mathbb{P}\left\{|\bar{Y}_i(\tau) - Q_i(\tau)| > x\right\} \lesssim n^{1+\beta}\exp\left\{-cn^{(1-\beta^2)}x\right\} + n^{1+\beta}\exp\left\{-c'n^{\beta(1-\beta)}x^{1/2}\right\}. \tag{3.7}$$

In addition, if $x \gg n^{2\beta-2}$, then the second term $n^{1+\beta}\exp\{-c'n^{\beta(1-\beta)}x^{1/2}\}$ would dominate, otherwise, the first term $n^{1+\beta}\exp\{-cn^{(1-\beta^2)}x\}$ would be the leading one.

Building on Theorem 3.3.1 and by the union bound, we also get an exponential probability bound holding uniformly over all coordinates, $i = 1, \ldots, p$.

Proposition 2. *Under Assumption 3.3.1, we have for $n \geq 1$,*

$$\begin{aligned}&\mathbb{P}\left\{\max_{1\leq i\leq p}|\bar{Y}_i(\tau) - Q_i(\tau)| > x\right\}\\ \lesssim &px\left[n^{1+\beta}\exp\left\{-cn^{1-\beta^2}x\right\} + n^{1+\beta}\exp\left\{-c'n^{\beta(1-\beta)}x^{1/2}\right\} + \exp\left\{-c''nx^2\right\}\right],\end{aligned}$$

where the constants in $\lesssim$, c, c', c'' are independent of n, p.

In the following, we want to compare our novel bound to the one from Cardot et al. (2017). For this purpose, we extend their Theorem 4.2 to the univariate case. Although their result is concerned with the geometric median, which is a multi-dimensional extension of the median, it is not valid in dimension one. The following theorem fills this gap.

Theorem 3.3.2. *Assume Conditions 3.3.1. For all $\delta \in (0,1)$, there exists a rank n_δ such that for all $n \geq n_\delta$,*

$$\mathbb{P}\left\{|\bar{Y}_i(\tau) - Q_i(\tau)| \leq \left(\frac{8}{3n} + \frac{4}{\sqrt{n}}\right)\log\left(\frac{4}{\delta}\right)\right\} \geq 1 - \delta.$$

The proof of Theorem 3.3.2, similar to the proof of Theorem 3.3.1, relies on the decomposition of the SGD algorithm defined in (3.6) into three parts. However, instead of using a bound on the moment generating function of $Y_{i,n}(\tau)$, the proof is based on a finite-sample bound on the L^2 error derived by Costa and Gadat (2020). As a consequence, the non-martingale terms are only negligible for the sample size n exceeding a certain rank n_δ. If we compare the bound obtained in Theorem 3.3.2 to our new bound obtained in Theorem 3.3.1, we observe that the latter is valid for each finite sample size n, while the former requires a sufficiently large sample size.

Analogous to Cardot et al. (2017), we provide a precise expression of the rank n_δ,

$$n_\delta = \max\left\{\left(\frac{6(C_1+C_3)}{\delta\log(\frac{4}{\delta})}\right)^{\frac{1}{1/2-\beta/2}}, \left(\frac{6C_2}{\delta\log(\frac{4}{\delta})}\right)^{\frac{1}{\beta-1/2}}, \left(\frac{6C_4}{\delta\log(\frac{4}{\delta})}\right)^{\frac{1}{2}}\right\}.$$

3.4 Application to Best Arm Identification

3.4.1 Stochastic Quantile Bandits

As an application of our novel tail probability bound, we consider the problem of best arm identification in a multi-armed bandit setting. A p-armed stochastic bandit is a collection of probability distributions, $\nu = (F_i : i \in [p])$. For each round $t = 1, \ldots, n$, the agent chooses an action $A_t \in [p]$ and observes the reward X_{tA_t} drawn independently from the distribution of the chosen arm, F_{A_t}. We refer to the monograph of Lattimore and Szepesvári (2020) for a comprehensive overview on bandit algorithms. We consider the setting of pure exploration (Bubeck et al., 2009). The learner is endowed with a fixed budget n and has to commit to one arm after the exploration phase in period $n+1$, according to a policy π. The goal is to select in action A_{n+1} the unique arm with the highest τ-quantile,

$$i^* = \text{argmax}_{i\in[p]} Q_i(\tau).$$

In the following, we write $Q_{i^*}(\tau) = Q^*(\tau)$. Most of the literature on best arm identification is concerned with the selection of arms with the highest expected value. Focusing instead on quantiles has at least two advantages. First, quantiles are more robust location parameters than the expected value. Many existing results on best arm identification in the mean case rely on the assumption of sub-Gaussian distributions. And second, the agent might be interested in different regions of the distribution of rewards, depending on her risk attitudes and preferences. Quantile preferences were first studied by Manski (1988) and formalized as a choice-theoretic model in Rostek (2010). A recent extension of quantile preferences to the dynamic setting was proposed by de Castro and Galvao (2019). We define the suboptimality gap of a given arm i relative to the optimal arm, for a fixed quantile level τ,

$$\Delta_i^\tau = Q^*(\tau) - Q_i(\tau).$$

The agent's goal is to minimize the regret R_n^τ, which is defined as the expected difference in quantiles of her policy in comparison to playing the optimal arm i^*,

$$R_n^\tau(\pi, \nu) = \mathbb{E}\left\{Q^*(\tau) - Q_{A_{n+1}}(\tau)\right\} = \mathbb{E}\left(\Delta_{A_{n+1}}^\tau\right). \tag{3.8}$$

The expectation is taken with respect to the interaction of the bandit environment ν and the policy of the learner π. In addition, we define the probability of selecting a suboptimal arm after the exploration period,

$$e_n = \mathbb{P}\left(A_{t+1} \neq i^*\right). \tag{3.9}$$

Our goal is to find a policy π which minimizes e_n and thus R_n^τ. Other accounts on best arm identification in a quantile bandit settings were discussed in Nikolakakis et al. (2021), Szorenyi et al. (2015) and in Howard and Ramdas (2019).

3.4.2 Algorithms and Bounds for Regret

A naive policy would be to play each arm uniformly during the exploration phase and then commit to the arm with the largest estimated quantile. The pseudo code for this uniform exploration algorithm is provided in Algorithm 1.

Algorithm 1: Uniform Exploration Algorithm

```
for t = 1 to n do
    Choose A_t = 1 + t mod (k)
end
Choose A_{n+1} = argmax_{i∈[p]} Ȳ_{i,n}(τ)
```

Although there is no trade-off between exploration and exploitation in our best arm identification setting, we can improve this strategy by allocating more actions to promising arms. For this we adapt both the successive elimination algorithm of Audibert et al. (2010) and the sequential halving algorithm of Karnin et al. (2013) to the quantile case.

The successive elimination algorithm divides the exploration period into $p-1$ phases. After each phase, the arm with the lowest estimated quantile is eliminated. Within the phase, the arms are played uniformly. The recommended arm A_{n+1} is the single remaining arm after n rounds. Algorithm 3 provides the pseudo code for the successive rejects algorithm.

Note that the arm eliminated in the first round is played $n_1 = \lceil \frac{1}{\overline{\log}(p)} \frac{n-p}{p} \rceil$ times, the one eliminated in the second round is played $n_2 = \lceil \frac{1}{\overline{\log}(p)} \frac{n-p}{p-2} \rceil$ times and the two remaining arms in round $p-1$ are played $n_{p-1} = \lceil \frac{1}{\overline{\log}(p)} \frac{n-p}{2} \rceil$ times. It can be easily verified that the

Algorithm 2: Successive Rejects Algorithm

Let $S_1 = \{1, \dots, p\}$, $\overline{\log}(p) = \frac{1}{2} + \sum_{i=2}^{p} \frac{1}{i}$, and

$$n_r = \lceil \frac{1}{\overline{\log}(p)} \frac{n-p}{p+1-r} \rceil, \quad r = 1, \dots, p-1$$

for $r = 1$ **to** $p-1$ **do**
  For each $i \in S_r$, select i for $n_r - n_{r-1}$ rounds
  Set $S_{r+1} = S_r \backslash \operatorname{argmin}_{i \in S_{r+1}} \bar{Y}_{i,n_r}(\tau)$
end
Choose $A_{n+1} = \operatorname{argmax}_{i \in S_{p-1}} \bar{Y}_{i,n_{p-1}}(\tau)$

budget constraint is satisfied. In the following, we denote with $(i) \in \{1, \dots, p\}$ the i-th best arm, which implies $\Delta^{\tau}_{(1)} \leq \Delta^{\tau}_{(2)} \leq \dots \leq \Delta^{\tau}_{(p)}$. Further, we write the tail probability bound from Theorem 3.3.1 as a function of the sample size and the suboptimality gap,

$$B(n, \Delta) = n^{1+\beta} \exp\Big\{ -cn^{(1-\beta^2)}\Delta \Big\} + n^{1+\beta} \exp\Big\{ -c' n^{\beta(1-\beta)} \Delta^{1/2} \Big\} + \exp\Big\{ -c'' n \Delta^2 \Big\}.$$

Let $\Delta^{\tau}_{max} = \max_{i \in [p]} \Delta^{\tau}_i$. We bound the regret using the policy of the successive rejects algorithm in the following theorem.

Theorem 3.4.1. *Let ν denote a p-armed stochastic bandit with F_i satisfying Assumption 3.3.1 and let π be the policy of the successive rejects algorithm. Then,*

$$R^{\tau}_n(\nu, \pi) \lesssim p(p-1) \Delta^{\tau}_{max} \max_{r \in \{1, \dots, p-1\}} B\left(\lceil \frac{1}{\overline{\log}(p)} \frac{n-p}{p+1-r} \rceil, \Delta^{\tau}_{(p+1-r)} \right)$$

Proof. Note that $R^{\tau}_n = \mathbb{E}(Q^*\{\tau) - Q_{A_{t+1}}(\tau)\} \leq \Delta^{\tau}_{max} e_n$. In the following, we bound e_n.

$$\begin{aligned}
e_n &\leq \sum_{r=1}^{p-1} \sum_{i=p+1-r}^{p} \mathbb{P}\left\{ \bar{Y}_{i^*,n_r}(\tau) \leq \bar{Y}_{(i),n_r}(\tau) \right\} \\
&= \sum_{r=1}^{p-1} \sum_{i=p+1-r}^{p} \mathbb{P}\left\{ \bar{Y}_{(i),n_r}(\tau) - Q_{(i)}(\tau) + Q^*(\tau) - \bar{Y}_{i^*,n_r}(\tau) \geq \Delta^{\tau}_{(i)} \right\} \\
&\lesssim \sum_{r=1}^{p-1} \sum_{i=p+1-r}^{p} B(n_r, \Delta^{\tau}_{(i)}) \\
&\leq \sum_{r=1}^{p-1} r B(n_r, \Delta^{\tau}_{(p+1-r)}) \\
&\leq p(p-1) \max_{r \in \{1, \dots, p-1\}} B(n_r, \Delta^{\tau}_{(p+1-r)}).
\end{aligned}$$

Plugging in the definition of n_r gives the desired result. □

An alternative to the successive rejects algorithm is the sequential halving algorithm

proposed by Karnin et al. (2013). Instead of eliminating only one arm per phase, the algorithm dismisses half of the arms with the lowest estimated τ-quantile. The number of rounds is reduced to $\lfloor \log_2 p \rfloor - 1$.

Algorithm 3: Sequential Halving Algorithm

```
Let S_1 = {1, ..., p},
for r = 1 to ⌈log_2 p⌉ do
    For each i ∈ S_r, select i for n_r = ⌊n / (|S_r| ⌈log_2 p⌉)⌋ times
    Let S_{r+1} denote the set of ⌊|S_r|/2⌋ arms in S_r with the highest value of
      Ȳ_{i,n_r}(τ).
end
Choose A_{n+1} by selecting the remaining arm in S_{⌈log_2 p⌉}.
```

The following theorem bounds the regret of the policy following the sequential elimination algorithm.

Theorem 3.4.2. *Let ν denote a p-armed stochastic bandit with F_i satisfying Assumption 3.3.1 and let π be the policy of the sequential halving algorithm. Then,*

$$R_n^\tau(\nu, \pi) \lesssim 2\log_2 p \Delta_{max}^\tau \max_{r \in \{1, \ldots, \log_2 p\}} B\left(\frac{2^r n}{p \log_2 p}, \Delta_{(p/(2^r))}^\tau\right).$$

Proof. WLOG, assume that p is a power of 2. Note that we can bound the probability that an arbitrary arm i has an higher estimated quantile than the optimal arm in a given round r,

$$\mathbb{P}\left\{\bar{Y}_{i^*,n_r}(\tau) < \bar{Y}_{i,n_r}(\tau)\right\} \leq B(n_r, \Delta_i^\tau).$$

If the best arm i^* is eliminated in round r, at least $1/2|S_r|$ arms need to have a larger estimated quantile. Denote by N_r the number of arms with a larger estimated quantile than the optimal arm. Then we have,

$$\begin{aligned}\mathbb{E}(N_r) &= \sum_{i \in S_r} \mathbb{P}\left\{\bar{Y}_{i^*,n_r}(\tau) < \bar{Y}_{i,n_r}(\tau)\right\} \\ &\leq |S_r| \max_{i \in S_r} B(n_r, \Delta_i^\tau).\end{aligned}$$

Using Markov's inequality, we can bound the probability of eliminating the optimal arm in round r,

$$\begin{aligned}\mathbb{P}(i^* \notin S_{r+1}) &= \mathbb{P}\left(N_r > \frac{1}{2}|S_r|\right) \\ &\leq \frac{2\mathbb{E}(N_r)}{|S_r|} \\ &\lesssim 2B\left(n_r, \Delta_{(p/(2^r))}\right).\end{aligned}$$

Then we can use the union bound to bound e_n,

$$\begin{aligned} e_n &\leq 2 \sum_{r=1}^{\log_2 p} B(n_r, \Delta^{\tau}_{(p/(2^r))}) \\ &\leq 2\log_2 p \max_{r \in \{1,\ldots,\log_2 p\}} B\left(\frac{2^r n}{p\log_2 p}, \Delta^{\tau}_{(p/(2^r))}\right) \end{aligned}$$

□

For a given bandit instance, ν, the probability of selecting a suboptimal arm, as defined in equation (3.8), is decreasing exponentially fast in the budget n. As a consequence, the simple regret in (3.9) is also decreasing exponentially fast in n. While most existing results on the mean bandit case rely on the assumption of sub-Gaussian distributions, we only need to impose a smoothness on the distribution.

3.5 Discussion

This paper studies the finite sample behavior of the SGD algorithm with Polyak-Ruppert averaging for the recursive estimation of quantiles. The main contribution is a new non-asymptotic tail probability bound which decreases exponentially fast in both x and n, while only imposing a smoothness assumption on the density. The proof relies on a decomposition of the averaged algorithm into a martingale difference part, a shift part and a remainder part, as well as on a new bound on the moment generating function of the SGD solution.

There are many promising directions for future research. As a first direction, one could move from the unconditional case to the quantile regression framework of Koenker and Bassett Jr (1978). Let y_t be a dependent variable and X_t a d-dimensional vector of regressors. Consider the following regression equation,

$$y_t = X_t^{\top}\beta(\tau) + \varepsilon_t(\tau),$$

where the conditional τ-quantile of there error satisfies $F^{-1}_{\varepsilon_t|X_t}(\tau) = 0$. An online quantile regression estimator based on the SGD algorithm could take the form

$$\widehat{\beta}_{t+1}(\tau) = \widehat{\beta}_t(\tau) + \gamma_t \left(\tau - \mathbf{1}_{y_{t+1} \leq X_{t+1}^{\top}\widehat{\beta}_t(\tau)}\right) X_{t+1}.$$

Another open research question is the derivation of a Bahadur type representation and a functional central limit theorem. For this purpose, consider a modified algorithm,

which guarantees the monotonicity of the SGD solution in the quantile level τ,

$$Y_{k+1}(\tau) = Y_k(\tau) + \gamma_k \left\{\tau - g_k\left(Y_k(\tau) - X_{k+1}\right)\right\}, \tag{3.10}$$

where

$$g_k(x) = \begin{cases} 1 & \text{if } x \geq \gamma_k/2, \\ \gamma_k^{-1}(x + \gamma_k/2) & \text{if } -\gamma_k/2 \leq x < \gamma_k/2, \\ 0 & \text{if } x < -\gamma_k/2. \end{cases}$$

This modified algorithm relies on the gradient of the Huber loss function (Huber, 1973). However, the gradient function is adaptively getting closer to the score function of the quantile loss, since the learning rate γ_k decreases with increasing sample size.

Finally, the i.i.d. assumption is often not satisfied in many empirical applications. It might be interesting to study whether the results of this paper remain true under certain dependence conditions.

3.6 Proofs for Section 3.3

3.6.1 Proof of Lemma 3.3.1

The key idea is to obtain a recursive equation for $Y_{i,k}(\tau)$. More specifically, we shall show

$$\mathbb{E}\left\{e^{tY_{i,k+1}(\tau)}\right\} \leq a_{k,t}\mathbb{E}(e^{tY_{i,k}(\tau)}) + c_0,$$

where $a_{k,t} > 0$ is some constant. This inequality can be obtained by the SGD generating scheme and Taylor's expansion. Then recursively applying above inequality leads to

$$\mathbb{E}\left\{e^{tY_{i,n}(\tau)}\right\} \leq c_0\left[1 + \sum_{k=k_t+1}^{n} \rho_k + \rho_{k_t}\mathbb{E}\left\{e^{tY_{i,k_t}(\tau)}\right\}\right] \text{ and } \rho_k = \prod_{l=k}^{n} a_{l,t},$$

where $k_t > 0$ is a selected starting point. Some elementary calculation shows that $\sum_{k=k_t+1}^{n} \rho_k$ and the starting term $\rho_{k_t}\mathbb{E}\{e^{tY_{i,k_t}(\tau)}\}$ is of order n^{β} when $t \leq n^{\beta(1-\beta)}$. The constrain on t is due to the fact that $a_{k,t} = 1 - c_1\gamma_k + c_2 t^2\gamma_k^2$. Thus t cannot be too large in order to make sure ρ_k does not explode.

Proof. WLOG, we can assume $Q_i(\tau) = 0$. Recall the distribution function of X_i is F_i.

Then the moment generating function takes the form

$$\begin{aligned}\mathbb{E}\left\{e^{tY_{i,k+1}(\tau)}\right\} &= \mathbb{E}\left[e^{t(Y_{i,k}(\tau)+\gamma_k\tau)}\mathbf{1}_{X_{i,k+1}>Y_{i,k}(\tau)} + e^{t(Y_{i,k}(\tau)-\gamma_k(1-\tau))}\mathbf{1}_{X_{i,k+1}\le Y_{i,k}(\tau)}\right]\\ &= \mathbb{E}\left[e^{tY_{i,k}(\tau)}\left\{e^{t\gamma_k\tau}(1-F_i(Y_{i,k}(\tau))) + e^{-t\gamma_k(1-\tau)}F_i(Y_{i,k}(\tau))\right\}\right].\end{aligned}$$

Take $L = c_\tau/(2c_f t)$, for any $k \ge t^{1/\beta}$ we have that $t\gamma_k \le 1$ and

$$\begin{aligned}\mathbb{E}\left\{e^{tY_{i,k+1}(\tau)}\right\} &\le \mathbb{E}\left[e^{tY_{i,k}(\tau)}\left\{e^{t\gamma_k\tau}(1-F_i(L)) + e^{-t\gamma_k(1-\tau)}F_i(L)\right\}\mathbf{1}_{Y_{i,k}(\tau)\ge L}\right.\\ &\qquad \left. + e^{tL}(1+e^{t\gamma_k\tau})\mathbf{1}_{Y_{i,k}(\tau)<L}\right]\\ &\le \mathbb{E}\left[e^{tY_{i,k}(\tau)}\left\{e^{t\gamma_k\tau}(1-F_i(L)) + e^{-t\gamma_k(1-\tau)}F_i(L)\right\}\right] + c_0, \qquad (3.11)\end{aligned}$$

where $c_0 = e^{c_\tau/(2c_f)}(1+e)$. Since $t\gamma_k\tau \le \tau_1 < 1$, by Taylor's expansion, $e^{t\gamma_k\tau} \le 1 + t\gamma_k\tau + 2t^2\gamma_k^2\tau^2$. Then

$$\begin{aligned}&e^{t\gamma_k\tau}\left\{1-F_i(L)\right\} + e^{-t\gamma_k(1-\tau)}F_i(L)\\ \le&(1+t\gamma_k\tau+2t^2\gamma_k^2)\left\{1-F_i(L)\right\} + \left\{1-t\gamma_k(1-\tau)+2t^2\gamma_k^2\right\}F_i(L)\\ \le&1 - t\gamma_k\left\{F_i(L)-\tau\right\} + 2t^2\gamma_k^2\\ \le&1 - t\gamma_k\left\{f_i(0)-c_f L\right\}L + 2t^2\gamma_k^2.\end{aligned}$$

Recall that $L = c_\tau/(2c_f t)$, inserting above into (3.11) leads to

$$\mathbb{E}\left\{e^{tY_{i,k+1}(\tau)}\right\} \le \mathbb{E}\left\{e^{tY_{i,k}(\tau)}\right\}(1 - c_1\gamma_k + 2t^2\gamma_k^2) + c_0,$$

where c_1 and $c_l, l > 1$, in the rest of the proof are positive constants independent of i, n, k, t, τ. Recursively applying above inequality, for $k_t = t^{1/\beta}$,

$$\mathbb{E}\left\{e^{tY_{i,n}(\tau)}\right\} \le c_0\left[1 + \sum_{k=k_t+1}^{n}\rho_k + \rho_{k_t}\mathbb{E}\left\{e^{tY_{i,k_t}(\tau)}\right\}\right] \text{ and } \rho_k = \prod_{l=k}^{n}(1 - c_1\gamma_l + 4t^2\gamma_l^2). \quad (3.12)$$

In the following, we shall bound terms $\sum_{k=k_t+1}^{n}\rho_k$ and $\rho_{k_t}\mathbb{E}(e^{tY_{i,k_t}(\tau)})$ separately. Firstly for $\sum_{k=k_t+1}^{n}\rho_k$ part, note that $1 + x \le e^x$, hence

$$\begin{aligned}\rho_k &\le \exp\Big\{-\sum_{l=k}^{n}(c_1\gamma_l - 4t^2\gamma_l^2)\Big\}\\ &\le \exp\Big\{-c_1(n^{1-\beta}-k^{1-\beta})/(1-\beta) + 4t^2(k^{-2\beta+1}-n^{-2\beta+1})/(2\beta-1)\Big\}.\end{aligned}$$

Assume $t \le n^{(1-\beta)\beta}$. To calculate $\sum_{k=k_t}^{n}\rho_k$, we shall deal with k close to n and far away from n separately. Firstly for any $k_t \le k \le an$, some constant $0 < a < 1$, and $t \ge t_0$,

where $t_0 > 0$ is a sufficiently large constant,

$$\rho_k \le \exp\Big\{ - c_2 n^{1-\beta} + c_3 t^2 k_t^{-2\beta+1} \Big\} \le \exp\Big\{ - c_2 n^{1-\beta}/2 \Big\}.$$

Secondly, when $k > an$, by the mean value theorem,

$$\rho_k \le \exp\Big\{ - c_4 n^{-\beta}(n-k) + c_5 t^2 n^{-2\beta}(n-k) \Big\} = \exp\Big\{ - c_4 n^{-\beta}(n-k)(1+o(1)) \Big\}.$$

Combining two parts, we obtain that

$$\sum_{k=k_t+1}^{n} \rho_k = \sum_{k=k_t+1}^{an-1} \rho_k + \sum_{k=an}^{n-n^\beta-1} \rho_k + \sum_{k=n-n^\beta}^{n} \rho_k \le c_5 n^\beta.$$

Secondly, for $\rho_{k_t}\mathbb{E}(e^{tY_{i,k_t}(\tau)})$ part, note that $|Y_{i,k}(\tau)| \le |y_{i,\tau}| + \sum_{j=1}^{k} |\gamma_j| = O(k^{1-\beta})$. Hence

$$\rho_{k_t}\mathbb{E}(e^{tY_{i,k_t}(\tau)}) \le \exp\Big\{ - c_2 n^{1-\beta}/2 + c_6 t k_t^{1-\beta} \Big\} \le \exp\Big\{ - c_2 n^{1-\beta}/4 \Big\}.$$

Therefore by (3.12), for $n \ge k_t$ and $t \ge t_0$ we have

$$\mathbb{E}\left\{ e^{tY_{i,n}(\tau)} \right\} \le c_0 \left\{ 1 + c_5 n^\beta + \exp\left(-c_2 n^{1-\beta}/4 \right) \right\} \lesssim n^\beta.$$

When $t < t_0$, $\mathbb{E}\{e^{tY_{i,n}(\tau)}\} \le \left[\mathbb{E}\{e^{tY_{i,n}(\tau)t_0/t}\}\right]^{t/t_0} \lesssim n^\beta$, and thus we complete the proof. □

3.6.2 Proof of Theorem 3.3.1

The main step is to decompose $Y_{i,k}(\tau)$ into a martingale difference part, a shift part and the remainder. WLOG assume $Q_i(\tau) = 0$. Then (3.3) can be rewritten into

$$Y_{i,k+1}(\tau) = Y_{i,k}(\tau) + \gamma_k Z_{i,k+1}(\tau), \tag{3.13}$$

where

$$Z_{i,k+1}(\tau) := \tau \mathbf{1}_{X_{i,k+1} > Y_{i,k}(\tau)} - (1-\tau)\mathbf{1}_{X_{i,k+1} \le Y_{i,k}(\tau)} = \tau - \mathbf{1}_{X_{i,k+1} \le Y_{i,k}(\tau)}. \tag{3.14}$$

To obtain the martingale difference part, we need to further decompose the term $Z_{i,k+1}(\tau)$ into $\xi_{i,k+1} = Z_{i,k+1}(\tau) - \mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k})$ and $\mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k})$, where $\mathcal{F}_{i,k} = (X_{i,k}, X_{i,k-1}, \ldots)$. Hence (3.13) can be rewritten into

$$\gamma_k^{-1}\left\{ Y_{i,k+1}(\tau) - Y_{i,k}(\tau) \right\} = \xi_{i,k+1} + \mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k}).$$

Further more, we have $\mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_k) = cY_{i,k}(\tau)+\rho_{i,k}$, here $\rho_{i,k}$ represents the remainder term and c is some constant. Hence we have the desired decomposition

$$cY_{i,k}(\tau) = \gamma_k^{-1}\left\{Y_{i,k+1}(\tau) - Y_{i,k}(\tau)\right\} - \xi_{i,k+1} - \rho_{i,k}.$$

To bound the first and third terms, we can adopt Lemma 3.3.1, and for the martingale difference part, we shall apply the Azuma's concentration inequality.

Proof. Firstly, we shall decompose $Z_{i,k+1}(\tau)$ into three parts and then deal with them separately. To this aim, and let $G_{i,\tau}(x) \stackrel{\text{def}}{=} \tau - F_i(x)$. Notice that $G_{i,\tau}\{Y_{i,k}(\tau)\} = \mathbb{E}(Z_{i,k+1}(\tau)|\mathcal{F}_{i,k})$ and $G_{i,\tau}\{Q_i(\tau)\} = 0$. Hence $Z_{i,k+1}(\tau)$ can be written as

$$Z_{i,k+1}(\tau) = \xi_{i,k+1} + G'_{i,\tau}\left\{Q_i(\tau)\right\} Y_{i,k}(\tau) + \rho_{i,k}, \tag{3.15}$$

where $\xi_{i,k+1}$ is the martingale difference part with respect to the filtration $\mathcal{F}_{i,k}$ and $\rho_{i,k}$ is the reminder with form:

$$\xi_{i,k+1} = Z_{i,k+1}(\tau) - G_{i,\tau}\left\{Y_{i,k}(\tau)\right\}, \quad \text{and} \quad \rho_{i,k} = G_{i,\tau}\left\{Y_{i,k}(\tau)\right\} - G'_{i,\tau}\left\{Q_i(\tau)\right\} Y_{i,k}(\tau).$$

Then by (3.15), the SGD equation (3.13) can be written as

$$Y_{i,k+1}(\tau) = Y_{i,k}(\tau) + \gamma_k\left\{\xi_{i,k+1} + G'_{i,\tau}(0)Y_{i,k}(\tau) + \rho_{i,k}\right\}. \tag{3.16}$$

Averaging (3.16) for k from 1 to n leads to

$$n^{-1}\sum_{k=1}^{n}\gamma_k^{-1}\left\{Y_{i,k+1}(\tau) - Y_{i,k}(\tau)\right\} = \bar{\xi}_{i,n} + G'_{i,\tau}(0)\bar{Y}_{i,n}(\tau) + \bar{\rho}_{i,n}.$$

where $\bar{Y}_{i,n}(\tau) = n^{-1}\sum_{k=1}^{n} Y_{i,k}(\tau)$, $\bar{\xi}_{i,n} = n^{-1}\sum_{k=1}^{n}\xi_{i,k+1}$ and $\bar{\rho}_{i,n} = n^{-1}\sum_{k=1}^{n}\rho_{i,k}$. By Abel's summation formula, we further obtain

$$\begin{aligned} G'_{i,\tau}(0)\bar{Y}_{i,n}(\tau) =& n^{-1}\gamma_n^{-1}(Y_{i,n+1}(\tau) - Y_{i,1}(\tau)) - n^{-1}\sum_{k=1}^{n-1} Y_{i,k+1}(\tau)(\gamma_{k+1}^{-1} - \gamma_k^{-1}) \\ & - \bar{\xi}_{i,n} - \bar{\rho}_{i,n} \stackrel{\text{def}}{=} \mathrm{I}_1 - \mathrm{I}_2 - \mathrm{I}_3 - \mathrm{I}_4. \end{aligned} \tag{3.17}$$

For I_1 part, by Lemma 3.3.1 and Markov's inequality, we have

$$\mathbb{P}(|\mathrm{I}_1| > x) \le c_1 n^{\beta}\exp\Big\{-c_2 n^{(1+\beta)(1-\beta)}x\Big\},$$

where c_i here and in the rest of this proof are some positive constants independent of i, n, p, τ. For I_2 part, take $a_k = c_\beta k^{-(1+\beta)(1-\beta)}n^{-\beta^2}$, where c_β only depends on β such

that $\sum_{k=1}^{n} a_k \le 1$. By Lemma 3.3.1 we have

$$\begin{aligned}\mathbb{P}\left\{|Y_{i,k+1}(\tau)(\gamma_k^{-1}-\gamma_{k+1}^{-1})| > na_k x\right\} &\le c_1 k^{\beta}\exp\Big\{-c_3 n a_k k^{(1+\beta)(1-\beta)}x\Big\}\\ &\le c_1 n^{\beta}\exp\Big\{-c_4 n^{1-\beta^2}x\Big\}.\end{aligned}$$

Therefore we have

$$\mathbb{P}(|\mathrm{I}_2| > x) \lesssim n^{1+\beta}\exp\Big\{-c_4 n^{1-\beta^2}x\Big\}.$$

For I_3 part, since $(\xi_{i,k})_k$ are martingale differences with respect to filtration $\mathcal{F}_{i,k}$, and $|\xi_{i,k}| \le 1$, thus by Azuma's concentration inequality, we have

$$\mathbb{P}(|\mathrm{I}_3| > x) \le 2\exp\Big\{-nx^2/2\Big\}.$$

For I_4 part, by Assumption 3.3.1, $|\rho_{i,k}| \le c_f Y_{i,k}^2(\tau)$, take $b_k = c_b k^{-2\beta(1-\beta)} n^{-1+2\beta(1-\beta)}$, where c_b is some constant only depends on β such that $\sum_{k=1}^{n} b_k \le 1$. Then by Lemma 3.3.1,

$$\begin{aligned}\mathbb{P}(|\mathrm{I}_4| > x) &\le \sum_{k=1}^{n}\mathbb{P}(c_f Y_{i,k}^2(\tau) > nb_k x)\\ &\lesssim \sum_{k=1}^{n} k^{\beta}\exp\{-c_5 k^{\beta(1-\beta)}(nb_k x)^{1/2}\}\\ &\lesssim n^{1+\beta}\exp\Big\{-c_6 n^{\beta(1-\beta)}x^{1/2}\Big\}.\end{aligned}$$

Desired result follows by combining all the above. □

3.6.3 Proof of Theorem 3.3.2

Proof. WLOG assume that $Q_i(\tau) = 0$. Recall the following decomposition.

$$\begin{aligned}G'_{i,\tau}(0)\bar{Y}_{i,n}(\tau) =& n^{-1}\gamma_n^{-1}\left\{Y_{i,1}-Y_{i,n}\right\}(\tau) - n^{-1}\sum_{k=1}^{n-1} Y_{i,k+1}(\tau)(\gamma_{k+1}^{-1}-\gamma_k^{-1})\\ &-\bar{\xi}_{i,n}-\bar{\rho}_{i,n} \overset{\text{def}}{=} \mathrm{I}_1-\mathrm{I}_2-\mathrm{I}_3-\mathrm{I}_4.\end{aligned}$$

We again bound each term on the right-hand side. By condition 3.3.1, we can apply Theorem 2.2 of Costa and Gadat (2020). For each $n \ge 1$, there is a constant $C > 0$ such that $\mathbb{E}(|Y_{i,n}|^2) \le C\frac{1}{n^{2-\beta}}$. For the I_1 part, we have

$$\mathbb{E}\left\{\left|\frac{Y_{i,n}(\tau)}{n\gamma_n}\right|^2\right\} \le n^{2\beta-2}\mathbb{E}\left\{|Y_{i,n}(\tau)|^2\right\} \le C\frac{1}{n^{2-\beta}}.$$

Applying Cauchy-Schwarz, we have for a constant $C_1 > 0$,

$$\mathbb{E}\left\{\left|\frac{Y_{i,n}(\tau)}{n\gamma_n}\right|\right\} \leq C_1 \frac{1}{n^{1-\beta/2}}.$$

Further, there is a constant $C_2 > 0$ such that

$$\mathbb{E}\left\{\left|\frac{Y_{i,1}(\tau)}{n\gamma_1}\right|\right\} \leq \frac{C_2}{n}.$$

For the I_2 part, since $\gamma_{k+1}^{-1} - \gamma_k^{-1} \leq 2\beta k^{\beta-1}$, there exists a constant C_3 such that

$$\begin{aligned}\mathbb{E}\left\{\left|\frac{1}{n}\sum_{k=1}^{n-1} Y_{i,k+1}(\tau)\left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right)\right|\right\} &\leq \frac{2\beta}{n}\sum_{k=1}^{n-1}\mathbb{E}\left(|Y_{i,k+1}(\tau)|\right)k^{\beta-1} \\ &\leq \frac{C_3}{n^{1-\beta/2}}.\end{aligned}$$

For the I_4 part, by Assumption 3.3.1, we have that $|\rho_{i,n}| \leq c_f Y_{i,n}(\tau)^2$. Consequently, we have for a constant C_4,

$$\mathbb{E}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\rho_{i,k}\right|\right) \leq \frac{c_f}{n}\sum_{k=1}^{n}\mathbb{E}\left\{|Y_{i,k}(\tau)|^2\right\} \leq \frac{c_f C}{n}\sum_{k=1}^{n}k^{-\beta} \leq C_4\frac{1}{n^\beta}.$$

Finally, we look at the Martingale term I_3. Since $\sup_k |\xi_k| \leq 1$ and $\sum_{k=1}^n \mathbb{E}(|\xi_k|^2|\mathcal{F}_{k-1}) \leq n$, we have by Pinelis-Bernstein's lemma, for all $x > 0$,

$$\begin{aligned}\mathbb{P}\left(\frac{\left|\sum_{k=1}^n \bar{\xi}_{k+1} > x\right|}{n}\right) &\leq \mathbb{P}\left(\sup_{1\leq k\leq n}\left|\sum_{j=1}^{k}\xi_{j+1}\right| > xn\right) \\ &\leq 2\exp\left\{-\frac{x^2}{2(\frac{1}{n}+\frac{x}{3n})}\right\}.\end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned}&\mathbb{P}\left\{\left|G'_{i,\tau}(0)\bar{Y}_{i,n}(\tau)\right| > x\right\} \\ &\leq 2\exp\left\{-\frac{x^2}{2(\frac{1}{n}+\frac{x}{3n})}\right\} + \frac{C_1+C_3}{xn^{1-\beta/2}} + \frac{C_2}{xn} + \frac{C_4}{xn^\beta} \\ &\stackrel{\text{def}}{=} g(x,n).\end{aligned}$$

We search for values of x satisfying $g(x,n) \leq \delta$. In particular, the following inequalities need to hold, for a given confidence level δ,

$$\begin{aligned} 2\exp\left\{-\frac{x^2}{2(\frac{1}{n}+\frac{x}{3n})}\right\} &\leq \delta/2 \\ \frac{C_1+C_3}{xn^{1-\beta/2}} &\leq \delta/6 \\ \frac{C_2}{xn} &\leq \delta/6 \\ \frac{C_4}{xn^{\beta}} &\leq \delta/6. \end{aligned}$$

We get the following conditions for x,

$$\begin{aligned} x &\geq 4\left(\frac{1}{3n}+\frac{1}{\sqrt{n}}\right)\log\left(\frac{4}{\delta}\right) \\ x &\geq \frac{6(C_1+C_3)}{\delta}\frac{1}{n^{1-\beta/2}} \\ x &\geq \frac{6C_2}{\delta}\frac{1}{n} \\ x &\geq \frac{6C_4}{\delta}\frac{1}{n^{\beta}}. \end{aligned}$$

Since $\beta \in (1/2, 1]$, the last three terms are of small order for large enough n. Finally, we choose

$$n_\delta = \max\left\{\left(\frac{6(C_1+C_3)}{\delta\log(\frac{4}{\delta})}\right)^{\frac{1}{1/2-\beta/2}}, \left(\frac{6C_2}{\delta\log(\frac{4}{\delta})}\right)^{\frac{1}{\beta-1/2}}, \left(\frac{6C_4}{\delta\log(\frac{4}{\delta})}\right)^{\frac{1}{2}}\right\}.$$

□

Bibliography

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, *262*, 134–147.

Audibert, J.-Y., Bubeck, S., & Munos, R. (2010). Best arm identification in multi-armed bandits. *COLT*, 41–53.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, *37*(3), 577–580.

Bardou, O., Frikha, N., & Pages, G. (2009). Computing var and cvar using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods and Applications*, *15*(3), 173–210.

Bubeck, S., Munos, R., & Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. *International conference on Algorithmic learning theory*, 23–37.

Cardot, H., Cénac, P., Godichon-Baggioni, A., et al. (2017). Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, *45*(2), 591–614.

Cardot, H., Cénac, P., Zitt, P.-A., et al. (2013). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, *19*(1), 18–43.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, *91*(434), 862–872.

Chen, X., Lee, J. D., Tong, X. T., Zhang, Y., et al. (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, *48*(1), 251–273.

Costa, M., & Gadat, S. (2020). Non asymptotic controls on a recursive superquantile approximation. *arXiv preprint arXiv:2009.13174*.

de Castro, L., & Galvao, A. F. (2019). Dynamic quantile models of rational behavior. *Econometrica*, *87*(6), 1893–1939.

Engle, R. F., & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

Fang, Y. (2019). Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics*, *46*(4), 987–1002.

Fang, Y., Xu, J., & Yang, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, *19*(1), 3053–3073.

Gadat, S., & Panloup, F. (2017). Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity, tse working paper. *arXiv preprint arXiv:1709.03342*.

Haldane, J. (1948). Note on the median of a multivariate distribution. *Biometrika*, *35*(3-4), 414–417.

Howard, S. R., & Ramdas, A. (2019). Sequential estimation of quantiles with applications to a/b-testing and best-arm identification. *arXiv preprint arXiv:1906.09712*.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The annals of statistics*, 799–821.

Karnin, Z., Koren, T., & Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. *International Conference on Machine Learning*, 1238–1246.

Kiefer, J. (1967). On bahadur's representation of sample quantiles. *The Annals of Mathematical Statistics*, *38*(5), 1323–1342.

Kiefer, J., Wolfowitz, J. et al. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, *23*(3), 462–466.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Luo, G., Wang, L., Yi, K., & Cormode, G. (2016). Quantiles over data streams: Experimental comparisons, new analyses, and further improvements. *The VLDB Journal*, *25*(4), 449–472.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*(1).

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. *Theory and Decision*, *25*(1), 79–104.

Minsker, S. et al. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, *21*(4), 2308–2335.

Munro, J. I., & Paterson, M. S. (1980). Selection and sorting with limited storage. *Theoretical computer science*, *12*(3), 315–323.

Nikolakakis, K. E., Kalogerias, D. S., Sheffet, O., & Sarwate, A. D. (2021). Quantile multi-armed bandits: Optimal best-arm identification and a differentially private scheme. *IEEE Journal on Selected Areas in Information Theory*.

Peng, L., & Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, *103*(482), 637–649.

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, *30*(4), 838–855.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400–407.

Rostek, M. (2010). Quantile maximization in decision theory. *The Review of Economic Studies*, *77*(1), 339–371.

Ruppert, D. (1988). *Efficient estimations from a slowly convergent robbins-monro process* (tech. rep.). Cornell University Operations Research and Industrial Engineering.

Szorenyi, B., Busa-Fekete, R., Weng, P., & Hüllermeier, E. (2015). Qualitative multi-armed bandits: A quantile-based approach. *International Conference on Machine Learning*, 1660–1668.

Wang, L., Zhou, Y., Song, R., & Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, *113*(523), 1243–1254.

Zhu, W., Chen, X., & Wu, W. B. (2020). A fully online approach for covariance matrices estimation of stochastic gradient descent solutions. *arXiv preprint arXiv:2002.03979*.

Chapter 4

Modelling Systemic Risk Using Neural Network Quantile Regression

PUBLICATION
Keilbar, G. and Wang, W. (2021). Modelling Systemic Risk Using Neural Network Quantile Regression, *Empirical Economics*.

4.1 Introduction

The issue of systemic risk attracts a lot of attention from academics as well as from regulators in the aftermath of the financial crisis of 2007-2009. Systemic risk refers to banks and other economic agents with substantial importance to the financial system due to their size (*too big to fail*) or their centrality within the financial network (*too interconnected to fail*). A bankruptcy of a systemically important financial institution can lead to the malfunctioning of the financial system or central banks and governments might be under pressure to interfere by bailing out respective firm. Due to these negative externalities, it is a crucial task for central banks and supervising agencies to identify systemically relevant firms.

A conventional quantitative risk measure is value-at-risk (VaR), which measures maximum losses at a certain confidence level. The Basel II Accord introduced VaR as a preferred measure for market risk. However, VaR is not capturing systemic risk adequately, as it is not capable to analyze the interdependency among firms. Given the subprime mortgage crisis in 2008, the Basel Committee on Banking Supervision has revised its Accords to focus on strong governance and risk management. Basel III is thus set up to control the systemic risk of the whole financial system, and it enforces additional requirements for identifying systemic risk important banks and generates demands on evaluating the interdependency of risk among banks. Adrian and Brunnermeier (2016) came up with conditional value-at-risk (CoVaR), a systemic extension of VaR. However, their original approach is restricted to analyze systemic risk in a linear and bivariate context. Namely, they focus primarily on the risk contribution

of an individual financial firm to the entire system, controlling for variables indicating general macroeconomic conditions.

This paper provides a new perspective for estimating CoVaR using neural networks. Nonlinearity is an important issue for the prediction performance of risk measures due to the complex dependency channels of financial institutions (Chao et al. (2015)). Neural networks have proved to be a suitable method for fitting nonlinear functions. Over the last years, neural networks have become state of the art models for prediction. They have been applied extensively and successfully to various fields, including image classification (Simonyan and Zisserman (2014)) as well as speech recognition problems (Graves et al. (2013)). Gu et al. (2020) and Bianchi et al. (2020) apply neural networks and other machine learning methods to asset pricing with promising results. We take the off-shelf neural network methodology and apply it to quantify financial risk. Our findings show that the quantile neural network-based approach provides a unique angle compared to the linear model for calibrating the systemic risk due to its flexibility. In particular, we find better out-of-sample prediction with our fine-tuned nonlinear neural network relative to the baseline linear quantile model of Koenker and Bassett Jr (1978, 1982).

We briefly summarize the steps of calibrating the systemic risk using a quantile neural network procedure. In the first step, we estimate the VaR for each global systemically important financial institution (G-SIB) from the United States by regressing their stock returns on a set of risk factors using linear quantile regression. Next, we estimate the CoVaRs of the same firms using neural network quantile regression. To characterize the interdependency among banks, we regress the return of one asset on the remaining returns respectively and aggregate the results into a systemic fit. By approximating the conditional quantile with a neural network we aim for capturing possible nonlinear effects. To estimate risk spillover effects across banks we calculate the marginal effects by taking the derivative of the fitted quantile with respect to the other banks' stock returns, evaluated at their VaR. By doing so we come up with a network of spillover effects represented by an adjacency matrix. This adjacency matrix is time-varying, i.e. we estimate a network for each window in our moving window estimation procedure. In the final step, we propose three systemic risk measures building on the previous results. As a first measure, we propose the *Systemic Fragility Index*, which identifies the most vulnerable banks in a given financial risk network. The second measure is the *Systemic Hazard Index*, which identifies the financial institutions which potentially pose the largest risk to the financial system. These two measures characterize the firm-specific aspects of systemic risk. Thus, we propose a third measure which estimates the total level of systemic risk, the *Systemic Network Risk Index*.

Our empirical findings confirm that systemic risk increased sharply during the height of the financial crisis in 2008. We also observe a high level of systemic risk at the end of 2011 due to the uncertainty surrounding the European debt crisis. By comparing our systemic risk measure to existing approaches for network-based interconnectedness, we find that our method offers a novel perspective due to the focus on the lower tail of the return distribution and due to the allowance for nonlinear dependencies. An out-of-sample comparison shows the superiority of our approach over a baseline model based on linear quantile regression. This leads to the conclusion that nonlinear effects are crucial for the modelling of systemic risk. Finally, we identify systemically relevant financial institutions during the financial crisis using our *SFI* and *SHI* measures. An advantage of our approach is the ability to capture the asymmetries of systemic risk, by differentiating between firms that affect and firms that are affected by the financial system. We also discover a risk cluster of four banks, which corresponds to the list of banks that received the largest funding in the course of the bank bailout of 2008.

This paper is an addition to the existing literature on systemic risk. Hautsch et al. (2014) modified the estimation of CoVaR further to analyze systemic risk in a multiple equation setup using the LASSO. Härdle et al. (2016) followed up this setup, and extended it to a nonlinear regression setting. In the meanwhile, there are numerous other methods for calibrating systemic risk. Acharya et al. (2017) built an economic model of systemic risk and measured the systemic risk externality of a financial institution by the systemic expected shortfall. Brownlees and Engle (2017) developed a systemic risk measure capturing the capital shortage given its degree of leverage and marginal expected shortfall. Diebold and Yılmaz (2014) analyzed the connectedness of financial firms in a network context using forecast variance decompositions in a vector autoregressive framework. Bianchi et al. (2019) proposed a Markov-switching graphical SUR model to model systematic and systemic risk.

There is a growing literature on econometric analysis using neutral networks. White (1988) started to investigate the usefulness of adopting a neural network for economic prediction. Unfortunately, the message is that even with simple neural networks the prediction performance is not ideal due to the overfitting issues. Kuan and White (1994) provided a further overview of neural networks with some basic concepts and theory. White (1992) provided the theoretical foundations of a nonparametric quantile neural network approach allowing for cases of dependent data. In terms of economic risk prediction, Taylor (2000) is concerned with predicting conditional volatility by adopting a quantile neural network approach. Xu et al. (2016) considered a quantile neural network procedure for evaluating VaR in the stock market. Cannon (2011) focused on the computational perspective of a quantile neural network.

The remainder of this paper is organized as follows. Section 4.2 provides a brief introduction to neural networks in general and neural network quantile regression in particular. Section 4.3 describes in detail the methodology of this paper. After establishing the research framework step by step, we present the results in section 4.4. Section 4.5 discusses the results and concludes.

All codes of this paper are available on quantlet.de.

4.2 Neural Network Quantile Regression

4.2.1 Neural Network Sieve Estimation

Neural networks attract increasing attention due to their success in a variety of prediction problems. Often described as a black box, single hidden layer neural networks can be seen as a special case of the nonparametric sieve estimator, see Grenander (1981) and Chen (2007). With increasing sample size n the complexity of the estimator of h_θ is required to increase appropriately fast. The structure of the neural network sieve is as follows, with $t = 1, 2, \cdots, n$,

$$\begin{aligned} Y_t &= h_\theta(X_t) + \varepsilon_t \\ &= \sum_{m=1}^{M_n} w_m^o \psi\left(\sum_{k=1}^{K} w_{k,m}^h X_{k,t} + b_m^h\right) + b_o + \varepsilon_t \end{aligned} \tag{4.1}$$

where Y_t is the dependent variable, X_t is a K-dimensional vector of independent variables and ε_t is an error term. The nonlinear activation function $\psi(\cdot)$ is assumed to be fixed and known. Typical choices are sigmoid functions, e.g. $\psi(z) = \tanh(z)$ or the ReLU (rectifier linear unit) function, $\psi(z) = \max(z, 0)$. There are two types of parameters, hidden layer parameters $w_{k,m}^h$ and b_m^h and output layer parameters w_m^o and b^o. The sieve parameter space Θ_n expands with n. In particular, the number of basis functions (i.e. the number of hidden nodes) goes to infinity, $M_n \to \infty$ as $n \to \infty$. Single layer neural networks have proved to be universal function approximators, as shown by Cybenko (1989) for sigmoid activation functions and Hornik et al. (1989) for the general case of bounded, non-constant activation functions. Sonoda and Murata (2017) extend the universal approximation property to unbounded activation functions, which includes the popular ReLU function.

The large sample properties of neural networks have been studied extensively in the literature. Notably, Chen and White (1999) show consistency and asymptotic normality of the nonparametric neural network sieve estimator under certain regularity conditions. Given that the number of basis functions grows appropriately with increasing sample

size, the root mean square convergence rate to an unknown (suitably smooth) true function is of order $o_p(n^{-1/4})$. This rate is crucial to obtain root-n asymptotic normality for plug-in estimators (Chen and Shen (1998)).

All of the above results concern with neural networks with a single hidden layer. The approximation theory and the asymptotic results of deep neural networks, i.e. neural networks with more than one hidden layer, is less understood compared to the shallow neural network case. Johnson (2018) shows that deep neural networks with limited width are not universal function approximators. Rolnick and Tegmark (2017) prove that deep neural networks can learn polynomial functions more efficiently (in terms of number of nodes required) than shallow ones.

4.2.2 Neural Network Sieves and Quantile Regression

Predominantly, neural networks have been applied to classification and mean regression problems. However, an extension to a quantile regression setting is straightforward. Consider the linear quantile regression equation for a fixed quantile level τ, as formulated in Koenker and Bassett Jr (1978, 1982).

$$Y_t = X_t\beta + \varepsilon_t, \quad t = 1, \dots, n \tag{4.2}$$

with $Q^\tau(\varepsilon_t|X_t) = 0$. In this setting the dependent variable Y_t is modelled as a linear function of independent variables X_t. The linear quantile estimator is then the solution to the following minimization problem:

$$\min_\beta \sum_{t=1}^{n} \rho_\tau \left(Y_t - X_t\beta\right) \tag{4.3}$$

where $\rho_\tau(z) = |z| \cdot |\tau - \mathbf{I}(z < 0)|$ is the quantile loss function. This minimization problem can be formulated as a linear program and can thus be solved by simplex or interior point algorithms. Neural network quantile regression is a nonlinear generalization of this regression framework. Instead of using a linear function, the conditional quantile is approximated by a neural network sieve estimator as defined in 4.2.1. The resulting optimization problem is nonconvex and cannot be solved by linear programming methods:

$$\min_\theta \sum_{t=1}^{n} \rho_\tau \left\{Y_t - h_\theta(X_t)\right\} \tag{4.4}$$

A possible alternative is to use the gradient-based backpropagation algorithm of Rumelhart et al. (1988). The asymptotic properties of nonparametric neural network

estimators for the conditional quantile are analyzed in White (1992). Under certain regularity conditions the estimator is consistent, see Appendix A. This result holds both for i.i.d. and dependent data.

4.2.3 Regularization Methods

Neural networks are prone to overfitting due to their high capacity. An effective tool to counteract overfitting lies in the choice of the structure and the hyperparameters of the neural network. In our single hidden layer setting, the most important hyperparameter is the number of hidden nodes, M_n. Other relevant parameters are the number of epochs and the specification of the learning algorithm. Typically, hyperparameters are selected according to a cross-validation criterion. A different approach is to put an extra penalty term on the weight parameters, $w^h_{k,m}$ and w^o_m. We are considering both L_1 and L_2 penalties which we summarize under the term elastic net (Zou and Hastie (2005)). This penalization method leads to the following optimization problem:

$$\min_{h_\theta} \sum_{t=1}^{n} \rho_\tau \left\{ Y_t - h_\theta(X_t) \right\} + \lambda_1 \| (w^{h\top}_{k,m}, w^{o\top}_m)^\top \|_1 + \lambda_2 \| (w^{h\top}_{k,m}, w^{o\top}_m)^\top \|_2^2 \tag{4.5}$$

where $\|\cdot\|_1$ is the L_1-norm, $\|\cdot\|_2$ is the L_2-norm. λ_1 and λ_2 are regularization parameters. A different method to prevent overfitting is the dropout method, proposed by Hinton et al. (2012) and Srivastava et al. (2014). In each iteration of the backpropagation algorithm, a given node is only considered with a probability $1 - p$. Consequently, each node is excluded with a probability p which is defined as the dropout rate. The motivation for this is to counteract memorization of the data by preventing co-adaptation of the nodes. Dropout is referred to be an ensemble method, as the final model is a result of training multiple models with reduced capacity.

4.3 Methodology to Calibrate Systemic Risk

In this section, we explain the details of our systemic risk analysis. Our methodology involves four steps. The first step is concerned with the estimation of VaR based on a linear quantile regression using a set of risk factors as explanatory variables. The results are used in the next step to estimate the CoVaR for each financial institution using a quantile regression neural network. Next, we calculate marginal effects to model systemic risk spillover effects, resulting in a time-varying systemic risk network. In the final step, we propose three systemic risk measures based on this systemic risk network.

Step 1: Estimation of VaR

VaR is defined as the maximum loss over a fixed time horizon at a certain level of confidence. The Basel II Accord introduced VaR as the preferred measure for market risk. The calculation of VaR functions as the basis for capital requirements of financial institutions. Mathematically, it is the τ-quantile of the return distribution:

$$\mathrm{P}(X_{i,t} \leq \mathrm{VaR}_{i,t}^{\tau}) = \tau, \tag{4.6}$$

where $X_{i,t}$ is the return of a financial firm i at time t and $\tau \in (0,1)$ is the quantile level. There exist numerous ways to estimate VaR. We refer to Kuester et al. (2006) for an extensive overview. One example is to assume a parametric model, and the most popular formulation involves the estimation of the latent volatility process via the GARCH model. Other approaches are based on the direct estimation of the conditional quantiles. Chernozhukov and Umantsev (2001) combine linear quantile regression with extreme value theory (EVT) to estimate VaR for extreme quantile levels. Chao et al. (2015) and Härdle et al. (2016) estimate VaR by using linear quantile regression on a set of macro state variables.

In this study, we compare three different specifications. First, we consider the dynamic quantile approach of Engle and Manganelli (2004), which is called CAViaR. The VaR is modelled as a latent process. We consider the symmetric absolute value (SAV) specification,

$$\mathrm{VaR}_{i,t}^{SAV,\tau} = \beta_{i,1} + \beta_{i,2}\,\mathrm{VaR}_{i,t-1}^{SAV,\tau} + \beta_{i,3}|X_{i,t-1}|. \tag{4.7}$$

Here the current level of VaR is determined by its lagged value as well as by the absolute value of the lagged return. Second, we consider the asymmetric slope (AS) CAViaR specification,

$$\mathrm{VaR}_{i,t}^{AS,\tau} = \beta_{i,1} + \beta_{i,2}\,\mathrm{VaR}_{i,t-1}^{AS,\tau} + \beta_{i,3}(X_{i,t-1})^{+} + \beta_{i,4}(X_{i,t-1})^{-}. \tag{4.8}$$

This specification allows for different responses to negative and positive returns. Finally, we consider the approach of Härdle et al. (2016). The VaR of each firm i is estimated by linear quantile regression using a set of macro state variables M_{t-1}.

$$X_{i,t} = \alpha_i + \gamma_i M_{t-1} + \varepsilon_{i,t}, \tag{4.9}$$

where the conditional quantile of the error term $Q^{\tau}(\varepsilon_{i,t}|M_{t-1}) = 0$. The VaR estimate

is the fitted value of the quantile regression,

$$\mathrm{VaR}_{i,t}^{LQR,\tau} = \widehat{\alpha}_i + \widehat{\gamma}_i M_{t-1} \tag{4.10}$$

VaR is a frequently used measure for understanding the critical risk level for an individual financial institution. The drawback of VaR is that it cannot account for dependency in a systemic context. Estimating VaR as an individual risk measure is a necessary first step to prepare for calibrating conditional risk.

Step 2: Estimation of CoVaR with Neural Network Quantile Regression

CoVaR was introduced as a systemic extension of standard VaR by Adrian and Brunnermeier (2016). Similar to VaR, it is a risk measure defined as a conditional quantile of the return distribution. But deviating from the VaR concept, CoVaR is contingent on a specific financial distress scenario. The motivation for using CoVaR is the identification of systemically important banks. For the distress scenario, we assume that all other firms are at their VaR. By doing this we follow the reasoning of Hautsch et al. (2014) and Härdle et al. (2016).

$$\mathrm{P}(X_{j,t} \leq \mathrm{CoVaR}_{\mathrm{j,t}}^{\tau} \,|X_{-j,t} = \mathrm{VaR}_{-j,t}^{\tau}) = \tau, \tag{4.11}$$

where $X_{-j,t}$ is a vector of returns of all firms except j at time t and $\mathrm{VaR}_{-j,t}^{\tau}$ is the corresponding vector of VaRs.

CoVaR can be estimated as a fitted conditional quantile, building on the results for the VaRs obtained in step 1. Chao et al. (2015) and Härdle et al. (2016) find evidence for nonlinearity in the dependence between pairs of financial institutions. Hence, linear quantile regression might not be an appropriate procedure to estimate the risk spillovers, as the interdependencies are potentially different in a state of worsening market conditions. The conditional quantile function of one bank on another may react nonlinearly to the change of critical level of another firm. We therefore propose the use of neural network quantile regression. The flexibility of the approach allows detecting possible nonlinear dependencies in the data.

The conditional quantile of bank j's returns is regressed on the returns of all other

banks and using a neural network as defined in section 4.2.2:

$$\begin{aligned} X_{j,t} &= h_\theta(X_{-j,t}) + \varepsilon_{j,t}, \\ &= \sum_{m=1}^{M_n} w_m^o \psi\left(\sum_{k\neq j}^{K} w_{k,m}^h X_{k,t} + b_m^h\right) + b^o + \varepsilon_{j,t}, \end{aligned} \tag{4.12}$$

with the conditional quantile of error term $Q^\tau(\varepsilon_{j,t}|X_{-j,t}) = 0$. To calculate the CoVaR of firm j, the fitted neural network has to be evaluated at the distress scenario:

$$\mathrm{CoVaR}_{j,t}^\tau = \widehat{h}_\theta(\mathrm{VaR}_{-j,t}^\tau), \tag{4.13}$$

where $\widehat{h}_\theta$ is the estimated neural network. Nonlinearity is introduced by the use of the nonlinear activation function. CoVaR can be interpreted as the hypothetical τ-quantile of the loss distribution if we are in a hypothetical distress scenario. In our case, this distress scenario is all other firms being at their VaR.

Step 3: Calculation of Risk Spillover Effects

Based on the weights estimated by the neural network quantile regression procedure, it is now possible to obtain risk spillover effects between each directed pair of banks. We propose to estimate the spillover effects by taking the partial derivative of the conditional quantile of firm j's return with respect to the return of firm i.

$$\frac{\partial Q^\tau(X_{j,t}|X_{-j,t})}{\partial X_{i,t}} = \frac{\partial}{\partial X_{i,t}} \sum_{m=1}^{M_n} w_m^o \ \psi\left(\sum_{k\neq j}^{K} w_{k,m}^h X_{k,t} + b_m^h\right) + b^o \tag{4.14}$$

In the case of a sigmoid tangent activation function we have

$$\frac{\partial Q^\tau(X_{j,t}|X_{-j,t})}{\partial X_{i,t}} = \sum_{m=1}^{M_n} w_m^o w_{i,m}^h \psi'\left(\sum_{k\neq j}^{K} w_{k,m}^h X_{k,t} + b_m^h\right) \tag{4.15}$$

with

$$\psi'(z) = \frac{2}{(\exp^{-z/2} + \exp^{z/2})^2}. \tag{4.16}$$

In the case of a ReLu activation function we have

$$\frac{\partial Q^\tau(X_{j,t}|X_{-j,t})}{\partial X_{i,t}} = \sum_{m=1}^{M_n} w_m^o w_{i,m}^h \mathbf{I}\left(\sum_{k\neq j}^{K} w_{k,m}^h X_{k,t} + b_m^h > 0\right), \tag{4.17}$$

where $\mathbf{I}(\cdot)$ is the indicator function. Note that the non-differentiability of the ReLU function is not an issue in practice since the input of the function is zero with probability

zero. As we are interested in the lower tail dependence, we consider the marginal effect evaluated at the distress scenario as defined in the previous subsection:

$$\left.\frac{\partial Q^{\tau}(X_{j,t}|X_{-j,t})}{\partial X_{i,t}}\right|_{X_{-j,t}=\mathrm{VaR}^{\tau}_{-j,t}} = \sum_{m=1}^{M_n} w^o_m w^h_{i,m}\ \psi'\left(\sum_{k\neq j}^{K} w^h_{k,m}\,\mathrm{VaR}^{\tau}_{k,t} + b^h_m\right). \tag{4.18}$$

Calculating such a marginal effect for each directed pair of firms yields an off-diagonal adjacency matrix of risk spillover effects at time t:

$$A_t = \begin{pmatrix} 0 & a_{12,t} & \dots & a_{1K,t} \\ a_{21,t} & 0 & \dots & a_{2K,t} \\ \vdots & \dots & \ddots & \vdots \\ a_{K1,t} & a_{K2,t} & \dots & 0 \end{pmatrix}, \tag{4.19}$$

with elements defined as absolute values of marginal effects:

$$a_{ji,t} = \begin{cases} \left|\left.\frac{\partial Q^{\tau}(X_{j,t}|X_{-j,t})}{\partial X_{i,t}}\right|_{X_{-j,t}=\mathrm{VaR}^{\tau}_{-j,t}}\right|, & \text{if } j \neq i \\ 0, & \text{if } j = i \end{cases}. \tag{4.20}$$

Note that the risk spillover effects are not symmetric in general, thus $a_{ji,t} \neq a_{ij,t}$. This adjacency matrix specifies a weighted directed graph modelling the systemic risk in the financial system.

Step 4: Network Analysis of Spillover Effects

To further analyze the systemic relevance of the financial institutions we can calculate several network measures building on the work of Diebold and Yılmaz (2014). They measure the connectedness of financial firms in terms of variance decomposition in a vector autoregressive framework. Their methodology is thus limited to capturing linear spillover effects.

First, the total directional connectedness *to* firm j at time t is defined as the sum of absolute marginal effects of all other firms on j.

$$C_{j\leftarrow\cdot,t} = \sum_{i=1}^{K} a_{ji,t} \tag{4.21}$$

Analogously, one can define the total directional connectedness *from* firm i at time t as the sum of absolute marginal effects from i to all other firms.

$$C_{\cdot\leftarrow i,t} = \sum_{j=1}^{K} a_{ji,t} \tag{4.22}$$

Lastly, Diebold and Yılmaz (2014) define the total connectedness at time t as the sum of all absolute marginal effects.

$$C_t = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} a_{ji,t} \tag{4.23}$$

The total connectedness is a measure for the connectedness level of the entire system without differentiating the roles of individual nodes of the network. Building on this network analysis, we refine the approach by incorporating VaR and CoVaR in the measurement of the systemic risk relevance. In particular, we propose the *Systemic Fragility Index (SFI)* and the *Systemic Hazard Index (SHI)* to rank financial institutions according to their relevance.

$$SFI_{j,t} = \sum_{i=1}^{K} \left(1 + |\,\mathrm{VaR}_{i,t}^{\tau}\,|\right) \cdot a_{ji,t}, \tag{4.24}$$

$$SHI_{i,t} = \sum_{j=1}^{K} \left(1 + |\,\mathrm{CoVaR}_{j,t}^{\tau}\,|\right) \cdot a_{ji,t}. \tag{4.25}$$

The *SFI* is a measure for the risk exposure of a financial institution j. It increases if those adjacency weights pointing to j are large and also if the VaRs of firms i (i.e. the risk factors for j) increase. This implies that the *SFI* will increase in times of financial distress. The index can be used by regulators to identify banks which have a high exposure to the tail risk in the financial system.

The *SHI* is a measure for the risk contribution of firm i to the whole system. It depends on the out-going adjacency weights from i weighted by the other firms' CoVaRs. Thus, the *SHI* tend to be large if the other firms are already affected by whole system, weigted by their CoVaR. The *SFI* and the *SHI* are firm-specific. It should be noted that our approach allows to model asymmetries. For instance, a firm which has a high tail risk exposure does not need to have a large impact on the whole system and vice versa. In contrast to the original CoVaR approach of Adrian and Brunnermeier (2016), our approach of identifying systemically important financial institutions has two advantages. First, we are able to capture possible nonlinear relationships in the data. Second, our approach operates in a network context which goes beyond the pairwise analysis proposed in the original CoVaR methodology.

As a third measure, we propose the *Systemic Network Risk Index (SNRI)*, a measure for the total systemic risk in the financial system which depends on the marginal effects, the outgoing VaRs, and the incoming CoVaRs. It is a measure for tail connectedness focusing a lower quantile level.

$$SNRI_t = \sum_{i=1}^{K}\sum_{j=1}^{K}(1+|\mathrm{VaR}_{i,t}^{\tau}|)\cdot(1+|\mathrm{CoVaR}_{j,t}^{\tau}|)\cdot a_{ji,t}. \tag{4.26}$$

Lastly, we define the adjusted adjacency matrix,

$$\widetilde{A}_t = \begin{pmatrix} 0 & \widetilde{a}_{12,t} & \dots & \widetilde{a}_{1K,t} \\ \widetilde{a}_{21,t} & 0 & \dots & \widetilde{a}_{2K,t} \\ \vdots & \dots & \ddots & \vdots \\ \widetilde{a}_{K1,t} & \widetilde{a}_{K2,t} & \dots & 0 \end{pmatrix}. \tag{4.27}$$

with elements defined as:

$$\widetilde{a}_{ji,t} = \begin{cases} a_{ji,t}\cdot(1+|\mathrm{VaR}_{i,t}^{\tau}|)\cdot(1+|\mathrm{CoVaR}_{j,t}^{\tau}|), & \text{if } j \neq i \\ 0, & \text{if } j = i \end{cases}. \tag{4.28}$$

The adjusted adjacency matrix accounts for the level of outgoing VaRs and incoming CoVaRs and is an improved representation of risk spillover effects. Systemic spillover effects are thus determined by the marginal effects of the neural network quantile regression procedure as well as by the VaRs and CoVaRs of the considered banks.

4.4 Empirical Study: US G-SIBs

4.4.1 Data

For the empirical application of our systemic risk methodology we are focusing on the *global systemically important banks (G-SIBs)* from the United States selected by the Financial Stability Board (FSB), see Table 4.1. These eight banks constitute systemic risk relevance to the global financial system and are deemed to be too-big-to-fail. We consider daily log returns in a time period between January 4, 2007 and May 31, 2018. The data is obtained from Yahoo Finance.

In addition to these stock return data, we consider daily observations of the following set of macro state variables:

i) Implied Volatility Index (VIX), from Yahoo Finance;

ii) the weekly S&P500 index returns, from Yahoo Finance;

iii) Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity from Federal Reserve Bank of St. Louis;

Financial Institution	NYSE symbol
Wells Fargo & Company	WFC
JP Morgan Chase & co.	JPM
Bank of America Corporation	BAC
Citygroup	C
The Bank of New York Mellon Corporation	BK
State Street Corporation	STT
Goldman Sachs Group, Inc.	GS
Morgan Stanley	MS

Table 4.1: List of G-SIBs in the USA.

iv) 10-Year Treasury Constant Maturity Minus 3-Month Treasury Constant Maturity from Federal Reserve Bank of St. Louis.

These macro variables are the common risk factors for the estimation of VaR in the first step of our systemic risk methodology.

4.4.2 Model Selection and Out-of-Sample Performance

The estimation of CoVaR based on neural network quantile regression involves several tuning parameters. Most importantly, we have to make a choice about the activation function and determine the sizes and structure of the neural network. We recalibrate these tuning parameters at the start of each year in a data-driven way. We propose the following moving-window model selection and evaluation procedure.

Following the common approach in the literature, e.g. Gu et al. (2020), Bianchi et al. (2020), we repeatedly divide our sample into three disjoint subsamples. These subsamples are consequential to maintain the time series structure of the data. The first sample is called the training set, which is denoted by $\mathcal{T}_1$. The training set is used to estimate the weight and bias parameters of the neural network for each candidate model specification. The performance is then evaluated using the validation set, denoted by $\mathcal{T}_2$. The tuning parameters are optimized by choosing the model specification which minimizes the objective function. This division into training and validation sets is an effective way to counteract overfitting. However, the validation fit is not truly out-of-sample since it is used to select the tuning parameters. Therefore, we finally consider the last subsample as the test set, which is denoted by $\mathcal{T}_3$. The test set is used to get an unbiased estimate of the method's performance.

To evaluate the predictive performance of our method, we calculate the out-of-sample

average quantile loss, (AQL^{oos}),

$$AQL^{oos} = \frac{1}{|\mathcal{T}_3|} \sum_{t \in \mathcal{T}_3} \rho_\tau \left\{ X_{j,t} - \widehat{Q}^\tau \left(X_{j,t} | X_{-j,t} \right) \right\}. \tag{4.29}$$

The tuning parameters include: the number of nodes in the neural network, the L_1 and L_2 penalty terms on the weight parameters and the dropout probability p. We recalibrate the tuning parameters for each financial firm at the start of the year. We choose a sample size of 200 and 50 days for the training and validation datasets respectively. This corresponds to one year of daily data. We evaluate the performance on the subsequent 250 days in the test set. By recalibrating the tuning parameters annually, we end up with ten windows in total. A visualization of the sample splitting scheme can be found in Figure 4.1. In the following, we summarize the steps of our model selection and the evaluation procedure.

Step 1: Split the data into training ($\mathcal{T}_1$), validation ($\mathcal{T}_2$) and test set ($\mathcal{T}_3$) for each window.

Step 2: For each bank j and each window, fit the conditional quantile of X_j contingent on X_{-j} using $\mathcal{T}_1$.

Step 3: Choose the model specification which minimizes the average quantile loss based on $\mathcal{T}_2$.

Step 4: Calculate AQL^{oos} based on the tuned neural network using $\mathcal{T}_3$.

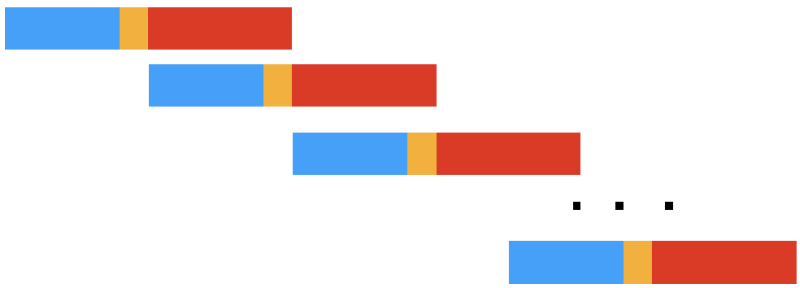


Figure 4.1: Visualization of the rolling window model selection scheme. Training data (blue), validation data (orange) and test data (red).

Finally, we compare the predictive performance of our neural network quantile regression procedure to a simple baseline model based on the linear quantile regression,

$$X_{j,t} = \beta_0 + \sum_{i \neq j}^{K} X_{i,t} \beta_i + \varepsilon_{j,t}, \tag{4.30}$$

with $Q^{\tau}(\varepsilon_t|X_{-j,t}) = 0$. The baseline model is estimated on training and validation data sets $\mathcal{T}_1$ and $\mathcal{T}_2$. The estimation does not involve any tuning parameters so we can make use of the combined data set. The out-of-sample forecast performance is then evaluated using the holdout data $\mathcal{T}_3$. We apply the test of Diebold and Mariano (2002) to compare the forecast performance. The test statistic is based on the quantile loss differentials between the neural network and the linear baseline model and has an asymptotic standard normal distribution. We choose a significance level of 1%. The test results are reported in Table 4.2.

For all of the financial institutions in our sample, the neural network fit performs better than the linear quantile regression fit. The outperformance of the neural network forecast is statistically significant for the majority of banks (seven out of eight). Only for Goldman Sachs the Diebold-Mariano fails to reject the null hypothesis of similar forecast performance. Overall, the use of a more complex model like a neural network appears to be recommendable. A plausible explanation for this is that a linear model is not capable to capture the complex interdependencies of financial firms under distress.

Firm	WCF	JPM	BAC	C	BK	STT	GS	MS
DM statistic	**-3.86**	**-2.44**	**-3.12**	**-3.27**	**-3.31**	**-2.76**	-1.56	**-2.88**
p-value	**0.000**	**0.008**	**0.001**	**0.001**	**0.001**	**0.003**	0.059	**0.002**

Table 4.2: The table reports the results of the Diebold-Mariano test comparing the neural network to the linear baseline model.

For the selection of the VaR approach used in the first step of our systemic risk analysis, we compare the predictive performance of the three candidate models introduced in section 3. We consider a sliding window of 250 days, which is used for estimation to predict the next day's conditional 5% quantile of the returns. The results are displayed in Table 4.3. For every bank in our sample, the linear quantile approach performs best. Results from the Diebold-Mariano test show that the difference is significant at the 1% confidence level after accounting for the multiple testing issue by using the Bonferroni correction for critical values. In the following, all VaR calculations are based on the linear quantile approach.

Firm	WCF	JPM	BAC	C	BK	STT	GS	MS
CaViaR SAV	2.42	2.47	3.39	3.40	2.48	3.05	2.49	3.26
CaViaR AS	2.63	2.60	3.50	3.46	2.73	3.10	2.58	3.45
LQR	**2.16**	**2.20**	**2.90**	**2.89**	**2.15**	**2.73**	**2.10**	**2.76**

Table 4.3: The table reports the out-of-sample average quantile loss of the candidate models for every financial institution ($\times 10^3$).

4.4.3 Estimation Results

VaR and CoVaR

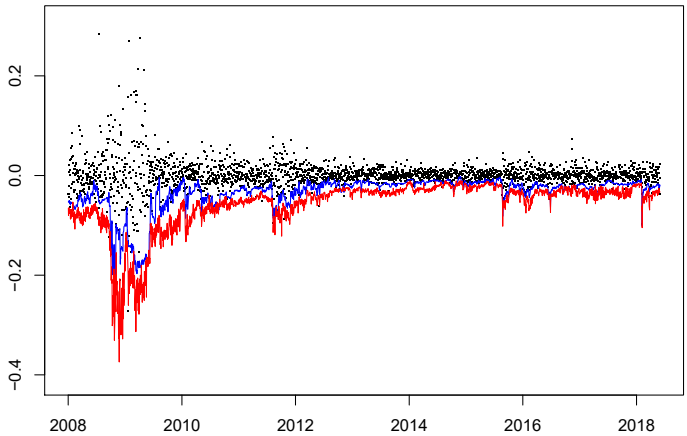


Figure 4.2: Plot of Returns (black dots), VaR (blue line) and CoVaR estimated by neural network quantile regression (red line) for Wells Fargo.

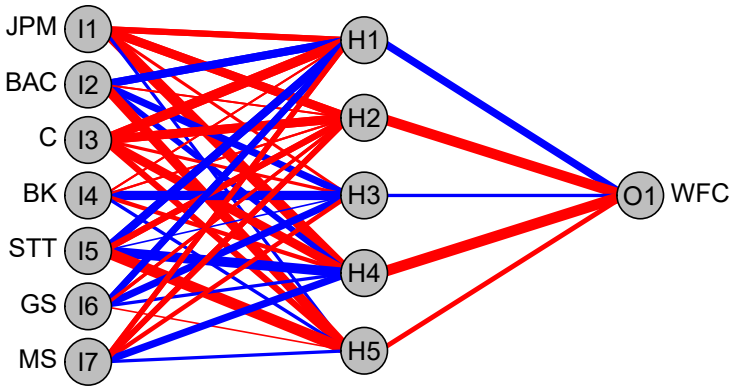
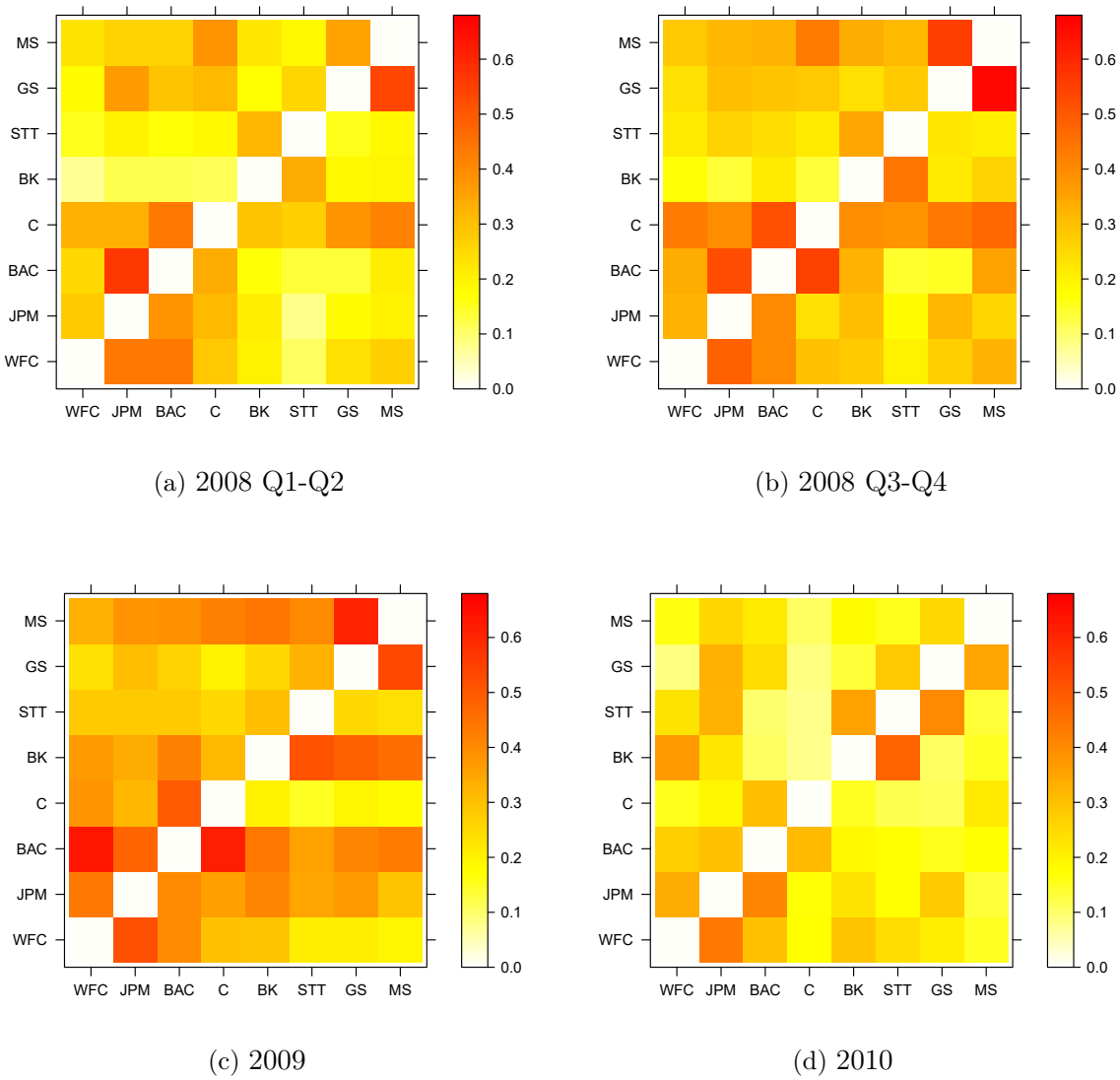


Figure 4.3: Fitted quantile regression neural network for Wells Fargo on March 13, 2008. Red connections indicate negative weights, blue connections indicate positive weights.

As explained in section 5.3, the analysis is carried out in four steps. In the first two steps, VaR and CoVaR are estimated for each firm, using linear quantile regression and neural network quantile regression, respectively. To account for potential non-stationarity, we employ a sliding window estimation framework for both measures. The window size is chosen to be 250 observations (representing one year of daily stock returns). We choose a quantile level of $\tau = 5\%$, which is the standard in the related literature, see Hautsch et al. (2014) and Härdle et al. (2016). A lower value for the quantile level leads to less reliable estimates, due to the inverse relation of the variance and the density of the error term. As a sensitivity analysis, we also report the results for $\tau = 1\%$, see Figure 4.11 and 4.12 in Appendix B. The results are robust with respect to the choice of the quantile level.

The estimation results for Wells Fargo are visualized in Figure 4.2. The estimated VaR and CoVaR follow a similar pattern. In the course of the financial crisis both risk measures explode, indicating an increase in systemic risk during this period. A second persistent spike appears in the second half of 2011 caused by the European debt crisis. In the following, both VaR and CoVaR stabilize with a few non-persistent spikes. Similar patterns can be found in the estimation results for the other financial institutions (see Figure 4.13 in Appendix B). An example of a fitted neural network is visualized in Figure 4.3.

Risk Spillover Network



(a) 2008 Q1-Q2

(b) 2008 Q3-Q4

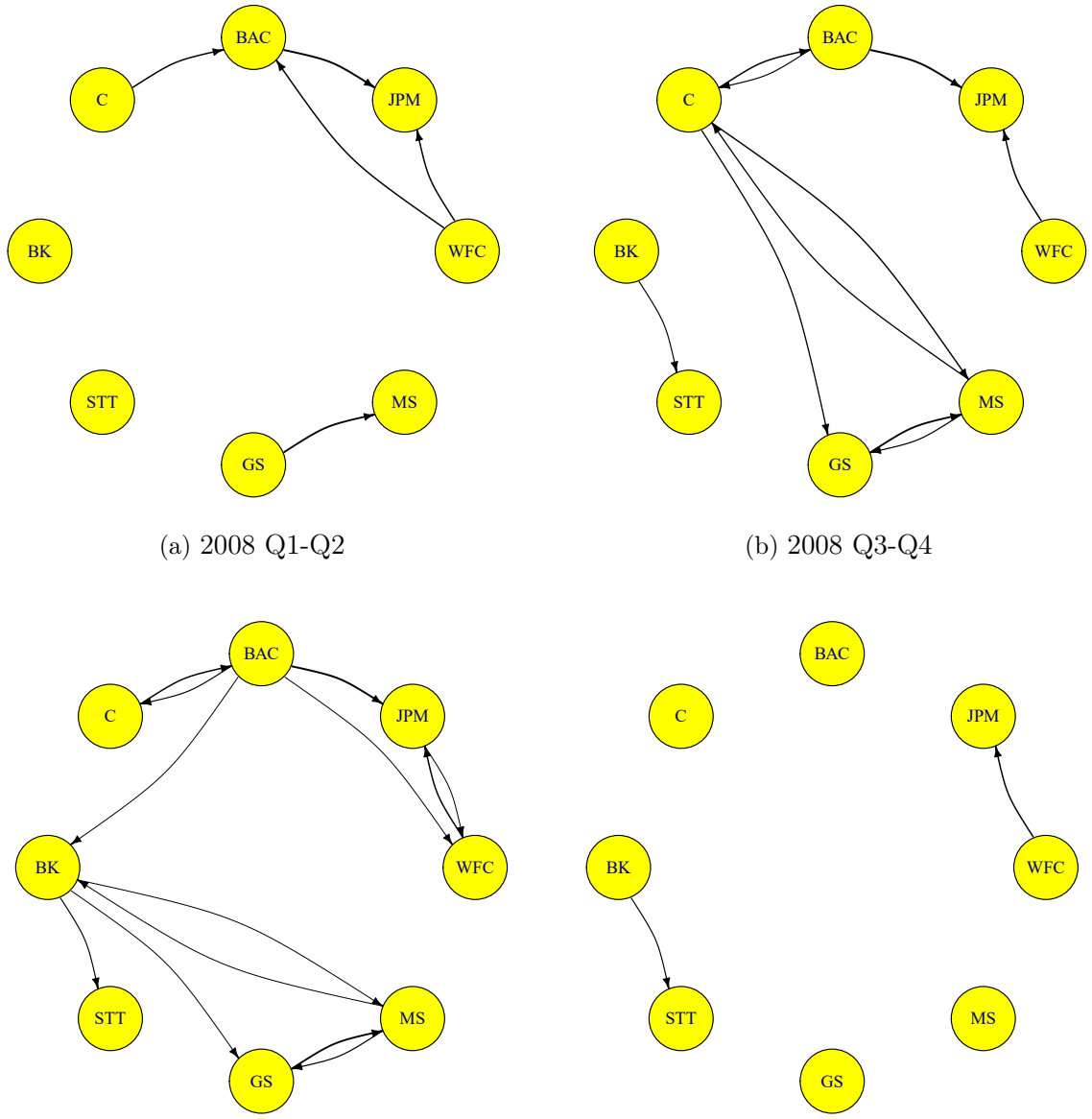
(c) 2009

(d) 2010

Figure 4.4: Time average of risk spillover effects across banks for different time periods.

Based on the estimation results of the neural network quantile regression procedure and on the fitted VaRs and CoVaRs, we calculate the directional spillover effects for

each pair of banks over our prediction horizon. The result is a time-varying weighted adjusted adjacency matrix (as defined in equation 4.27). This risk spillover network provides insights into the cross-section and the time dynamics of systemic risk. Figure 4.4 visualizes the evolution of the network in the course of the financial crisis. The first half of 2008 shows a moderate level of lower tail connectedness. This setting changes dramatically in the second half of 2008 with the bankruptcy of Lehman Brothers. As a consequence, the United States Department of the Treasury was compelled to bail out financial institutions to avoid a total collapse of the financial system. Also, the Federal Reserve Bank had to adjust its monetary policy. The time average of the adjacency matrix for 2009 shows a continuing state of financial distress. However, compared to the previous periods one can visually identify a risk cluster in the lower left part of the adjacency matrix. Finally, 2010 shows a decline in systemic risk spillover effects caused by a regained trust in the financial system. Figure 4.5 restricts the visualization to the largest edges of the financial risk network. As a first observation, spillover effects across banks tend to be symmetric. If bank i has a large impact on bank j, the converse is also very likely. A second observation is the identification of the risk cluster mentioned above. This cluster includes four financial institutions, Citigroup, Bank of America, JP Morgan and Wells Fargo. This cluster coincides with the list of the largest beneficiaries of the bailout program in 2008 and 2009.



(a) 2008 Q1-Q2

(b) 2008 Q3-Q4

(c) 2009

(d) 2010

Figure 4.5: Time average of risk spillover effects across banks after thresholding ($\widetilde{a}_{ji} > 0.4$) for different time periods.

Network Risk Measures

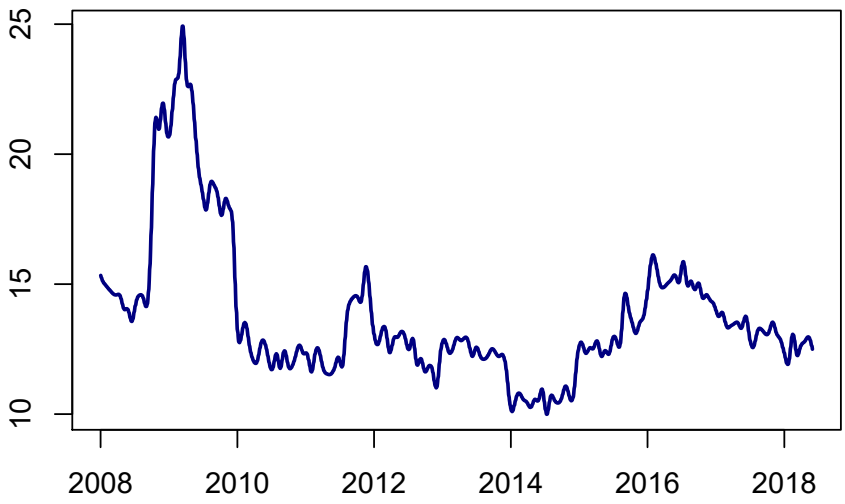


Figure 4.6: The figure shows the time series of the *SNRI*.

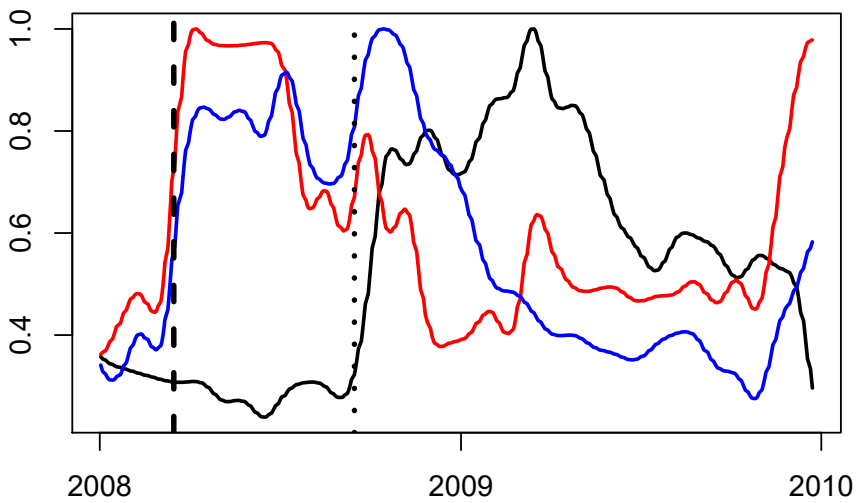


Figure 4.7: Plot of *SNRI* (black line), the Granger causality measure of Billio et al. (2012) (red line) and total connectedness of Diebold and Yılmaz (2014) (blue line). Dashed vertical line marks the bailout and acquisition of Bear Stearns by JP Morgan on March 14, 2008, the dotted vertical line indicates the bankruptcy of Lehman Brothers on September 15, 2008.

In this subsection, we estimate the systemic risk measures using the results from the previous steps. First, we consider the *Systemic Network Risk Index (SNRI)*, as a measure for total systemic risk in the financial system. Figure 4.6 shows the development over time. As expected, we see a sharp increase in systemic risk during the financial crisis in the second half of 2008. A second peak appears in the second half of 2011 as a result of the uncertainties associated with the European debt crisis. After a short period of stabilization, we see another rise in systemic risk from 2014 till

2016. In contrast to the previous peaks, this increase appears to be more gradual.

We now discuss the systemic risk measure calibration during the financial crisis in detail. We restrict our focus on the two-year period, i.e. from the start of 2008 to the end of 2009. We compare our *SNRI* to the Granger causality measure of Billio et al. (2012) and the total connectedness measure based on variance decomposition proposed by Diebold and Yılmaz (2014). Both measures are estimated using the same set of financial institutions and a rolling window of 250 days. The results are displayed in Figure 4.7. As reference dates, we have added the bailout of Bear Stearns and the resulting acquisition by JP Morgan on March 14, 2008, as well as the bankruptcy of Lehman Brothers on September 15, 2008. A few significant differences in the time series of the risk measures are apparent. While the Granger causality measure and the total connectedness increase sharply after the Bear Stearns event, the *SNRI* decreases slightly. In contrast to both alternatives, our measure is exclusively concerned with the lower quantile of the return distribution. We infer that the resulting intervention had a calming effect on the financial markets and thus prevented an increase in lower tail dependence. The Bear Stearns shock seemed to have a systematic but not necessarily a systemic effect. In contrast, we observe a simultaneous sharp increase of all three measures immediately after the Lehman Brothers bankruptcy. The increase in connectedness thus affected the mean as well as the lower tail of the distribution. We deduce that the shock from the Lehman bankruptcy had a truly systemic impact. In the aftermath of the collapse, the *SNRI* has its maximal point in March of 2009 and remains at a high level until the second half of the same year. The comparing measures have an earlier peak in the end of 2008 followed by a fast decrease. We conclude that the *SNRI* complements the network-based risk measures proposed by Billio et al. (2012) and Diebold and Yılmaz (2014) as it is more sensitive to shocks in the lower tail.

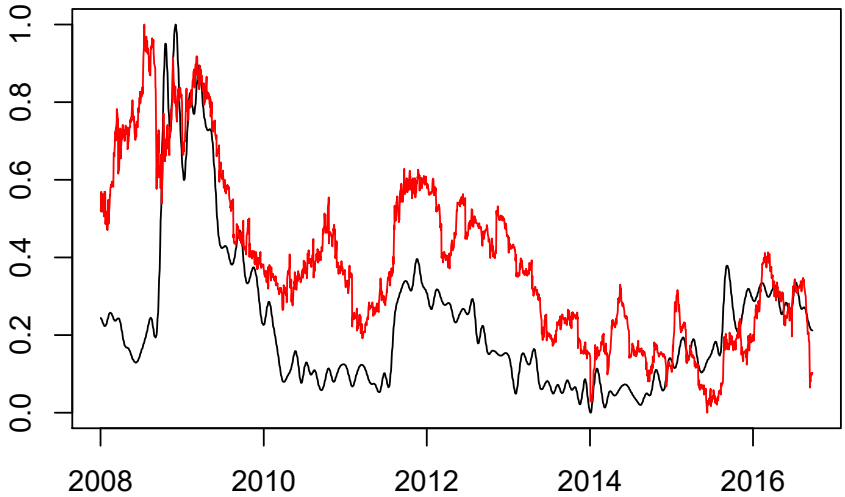


Figure 4.8: The figure shows the co-movement of the *SNRI* (black line) and the aggregate *SRISK* (**brownlees2017SRISK**, red line).

We also compare the *SNRI* to the aggregated *SRISK* of Brownlees and Engle (2017) in Figure 4.8. One can identify a co-movement of both indices. In particular, both the financial crisis and the European debt crisis lead to a sharp increase in both risk measures. However, we have to acknowledge that the aggregated *SRISK* already detects vulnerabilities in the financial system as early as the beginning of 2008. The reason for this is that the *SRISK* incorporates additional information on micro-prudential variables, namely the book value of debt and the quasi value of assets. An advantage of the *SNRI* is that it is entirely based on market data. Also, the *SRISK* requires assumptions on a number of structural parameters, such as the prudential capital ratio and the threshold loss, while our approach does not. Finally, another advantage of our approach is the estimation of spillover effects in a network context.

	2008 Q1-Q2		2008 Q3-Q4		2009		2010	
Rank	Ticker	SFI	Ticker	SFI	Ticker	SFI	Ticker	SFI
1	C	2.239	C	2.395	BAC	2.633	WCF	1.689
2	GS	1.962	MS	2.046	BK	2.426	JPM	1.640
3	WFC	1.822	BAC	1.983	MS	2.393	STT	1.541
4	MS	1.748	GS	1.970	JPM	2.222	BAC	1.472
5	BAC	1.709	WCF	1.907	GS	1.900	GS	1.442
6	JPM	1.546	JPM	1.752	WCF	1.847	BK	1.442
7	STT	1.300	STT	1.497	C	1.572	MS	1.260
8	BK	1.100	BK	1.365	STT	1.561	C	1.164

Table 4.4: The table reports the ranking of financial institutions according to their *SFI* averaged over different time intervals.

	2008 Q1-Q2		2008 Q3-Q4		2009		2010	
Rank	Ticker	SHI	Ticker	SHI	Ticker	SHI	Ticker	SHI
1	JPM	2.209	JPM	2.203	WCF	2.440	JPM	2.010
2	BAC	2.021	MS	2.149	JPM	2.438	BAC	1.616
3	MS	1.939	BAC	2.138	GS	2.377	STT	1.574
4	C	1.828	GS	1.981	BAC	2.349	WCF	1.555
5	GS	1.568	BK	1.976	BK	2.187	BK	1.488
6	BK	1.530	C	1.881	C	2.162	GS	1.475
7	WCF	1.426	WCF	1.820	MS	2.149	MS	1.254
8	STT	1.316	STT	1.721	STT	2.089	C	0.965

Table 4.5: The table reports the ranking of financial institutions according to their SHI averaged over different time intervals.

While the *SNRI* is an index for total systemic risk, we now consider firm-specific measures. Table 4.4 ranks financial firms according to their *Systemic Fragility Index (SFI)*. A large *SFI* indicates high systemic exposure to the financial system. Our findings suggest that Citigroup is among the most fragile banks during the height of the financial crisis, being top-ranked in the first and in the second half of 2008. Due to heavy exposure to troubled mortgages, the US government decided to bail out the bank in November 2008. In the periods following the bail-out, Citigroup's *SFI* rank dropped sharply. Figure 4.9 shows the time dynamics of the *SFI* of Citigroup. Another high-ranked financial institution is Bank of America, which is on position three in the second half of 2008 and the number one in 2009. In contrast, State Street Corporation is ranked at the bottom of the table throughout 2008 and 2009. This result is plausible since State Street was the first major financial institution to pay back its loans to the US Treasury in July 2009.

We conduct a similar ranking with respect to the *Systemic Hazard Index (SHI)*, which ranks the financial institutions according to the risk contributed to the financial system. In each of the time periods we consider, JP Morgan is listed in the top two of the ranking. Similar, Bank of America is ranked in the top four consistently, being the second highest ranked bank in the first half of 2008. Figure 4.10 visualizes the time dynamics of the *SHI* for Bank of America. In the aftermath of the crisis in 2009, Wells Fargo also emerges as a systemic risk factor to the financial system. An advantage of our approach is that we are able to differentiate between firms, which transmit systemic risk, and firms which are affected by systemic risk. By doing this we capture the asymmetric nature of the systemic risk. As an example, JP Morgan is ranked high according to the *SHI* in 2008 but relatively low in *SFI*. The opposite can be observed for Citigroup, which is ranked low in *SHI* and high in *SFI* during the same time periods. However, State Street is at the bottom of both rankings during the height of

the financial crisis, implying that it is neither a large risk factor nor strongly affected by the financial system.

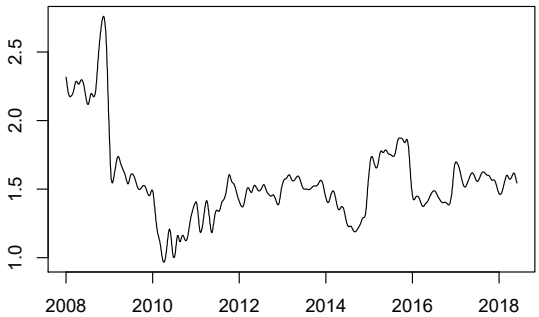


Figure 4.9: Time series of the *SFI* for Citigroup.

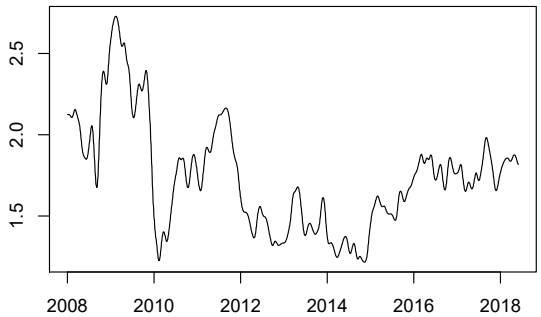


Figure 4.10: Time series of the *SHI* for Bank of America.

4.5 Conclusion

This paper proposes a novel approach to estimate the conditional value-at-risk (CoVaR) of financial institutions based on neural network quantile regression. Our methodology allows for the identification of risk spillover effects across banks in a nonlinear and multivariate context. We define three network-based measures for systemic risk, the *Systemic Fragility Index* and the *Systemic Hazard Index* as firm-specific measures and the *Systemic Network Risk Index* as a measure for the overall risk in the financial system. These measures quantify the connectedness of the financial system while restricting the analysis on the lower tail of the distribution. The neural network framework allows us to model systemic risk in a highly nonlinear setting. A comparison to a linear baseline model shows the predictive superiority of our neural network approach in terms of the out-of-sample performance.

We apply our methodology to global systemically important banks (G-SIBs) from the United States in the period 2007 - 2018. Consistent with previous findings in the literature, we observe the *Systemic Network Risk Index* increasing sharply during the financial crisis and during the European debt crisis. A comparison to the connectedness measures proposed in Billio et al. (2012) and Diebold and Yılmaz (2014) shows that our systemic risk measure captures different aspects of connectedness and offers therefore a new perspective on systemic risk. Furthermore, our approach allows to identify a risk cluster of banks which corresponds to the list of banks that receive the largest amount of funding from the US Department of Treasury. By ranking the financial firms according to their *Systemic Fragility Index* and their *Systemic Hazard Index* we are able to identify those firms which bear significant exposure to the financial system and those firms which impose the greatest risk to the financial system.

4.A Consistency of Neural Network Sieve Estimator for the Conditional Quantile

White (1992) shows the consistency of the neural network quantile regression estimator.

Assumption A.1: The data $Z_t = (X_t^\top, Y_t^\top)^\top$ is generated from a bounded stochastic process defined on a complete probability space $(\Omega, \mathcal{F}, P)$, X_t is a random $r \times 1$ vector, Y_t is a random scalar and

(i) Z_t is an i.i.d. process or

(ii) Z_t is a stationary $\phi-$ or $\alpha-$mixing process with such that the mixing coefficients $\phi(k) = \phi_0 \xi^k$ or $\alpha(k) = \alpha_0 \xi^k, 0 < \xi^k < 1, \phi_0, \alpha_0, k > 0$.

Without loss of generality, we may assume $Z_t : \Omega \to \mathbb{I}^{r+1} \stackrel{\text{def}}{=} [0,1]^{r+1}$.

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a bounded function and let (Θ, ρ) be a metric space, where ρ is the L_1-metric. For any $q \in \mathbb{N}$ and $\Delta \in \mathbb{R}^+$ define $T(\psi, q, \Delta) = \{\theta \in \Theta : \theta(x) = \beta_0 + \sum_{j=1}^q \beta_j \psi(x^\top \gamma_j)$ for all x in $\mathbb{I}^r$, $\sum_{j=0}^q |\beta_j| \le \Delta$, $\sum_{j=1}^q \sum_{i=1}^r |\gamma_{ji}| \le q\Delta\}$. Further let $Q_n(\theta) = n^{-1} \sum_{t=1}^n |Y_t - \theta(X_t)||\tau - \mathbf{I}(Y_t < \theta(X_t))|$.

Assumption A.2: $\Theta_n(\psi) = T(\psi, q_n, \Delta_n), n = 1, 2, \ldots$, where ψ is bounded, satisfies a Lipschitz condition and is either a cdf or is l-finite. q_n and Δ_n are such that $q_n \to \infty$ and $\Delta_n \to \infty$ as $n \to \infty$. $\Delta_n = o(n^{1/2})$ and either (i) $q_n \Delta_n^2 \log q_n \Delta_n = o(n)$ or (ii) $q_n \Delta_n \log q_n \Delta_n = o(n^{1/2})$.

Assumption A.3: For given quantile level $\tau \in (0,1)$, $\theta_\tau : \mathbb{I}^r \to \mathbb{I}$ is a measurable function such that $\mathrm{P}\{Y_t \le \theta_\tau(X_t) | X_t\} = \tau$ and for every $\theta \in \Theta$ and all $\epsilon > 0$ sufficiently small $\mathrm{E}\{\theta(X_t) - \theta_\tau(X_t)\} > \epsilon$ implies that for some $\delta_\epsilon > 0$,

$$\mathrm{E}\left[\mathbf{I}\left\{(\theta_\tau(X_t) + \theta(X_t))/2 \le Y_t < \theta_\tau(X_t)\right\} | \theta(X_t) < \theta_\tau(X_t)\right] > \delta_\epsilon$$

and

$$\mathrm{E}\left[\mathbf{I}\left\{\theta_\tau(X_t) \le Y_t < (\theta_\tau(X_t) + \theta(X_t))/2\right\} | \theta(X_t) \ge \theta_\tau(X_t\right] > \delta_\epsilon.$$

Theorem 2.5 White (1992): Given assumptions A.1(i), A.2(i) and A.3 or A.1(ii), A.2(ii) and A.3, there exists a measurable connectionist sieve estimator $\widehat{\theta}_n : \Omega \to \Theta$ such that $Q_n(\widehat{\theta}_n) \le Q_n(\theta)$, $\theta \in \Theta_n(\psi)$, $n = 1, 2, \ldots$. Further, $\rho(\widehat{\theta}_n, \theta_\tau) \stackrel{p}{\to} 0$.

4.B Estimation Results

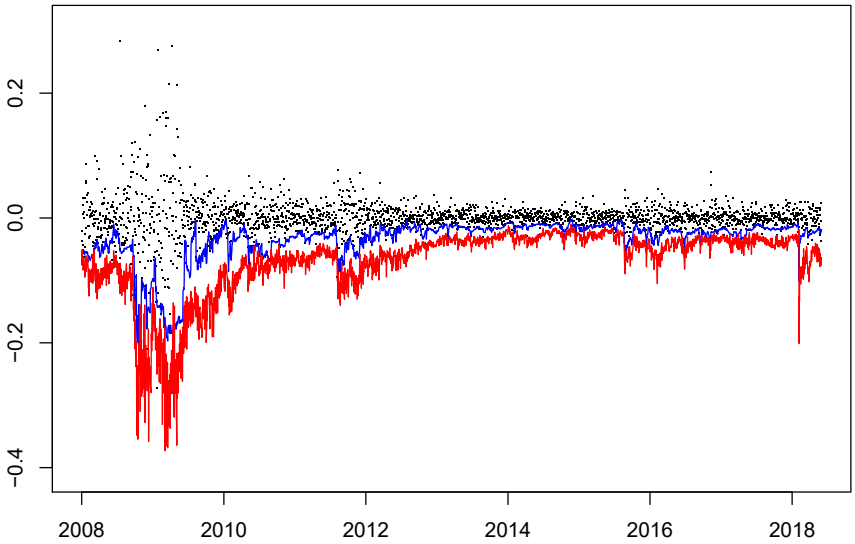


Figure 4.11: Plot of Returns (black dots), VaR (blue line) and CoVaR estimated by neural network quantile regression (red line) for Wells Fargo, $\tau = 1\%$.

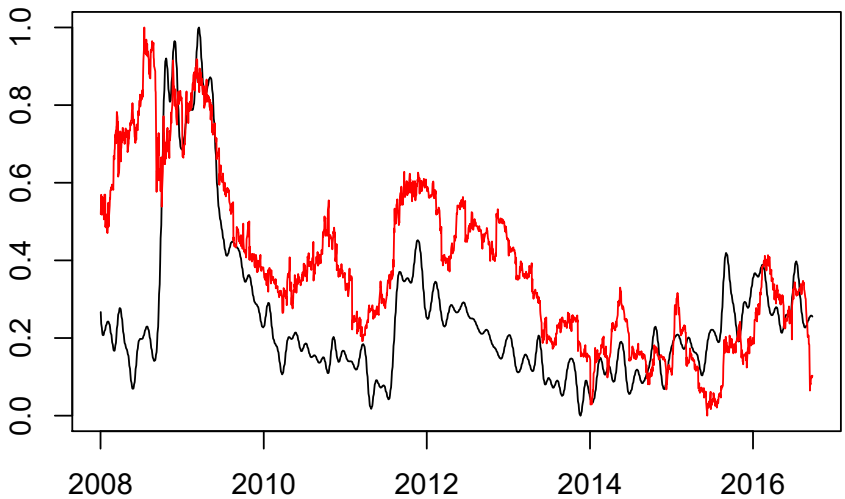


Figure 4.12: The figure shows the co-movement of the $SNRI$ (black line) and the $SRISK$ (**brownlees2017SRISK**, red line), $\tau = 1\%$.

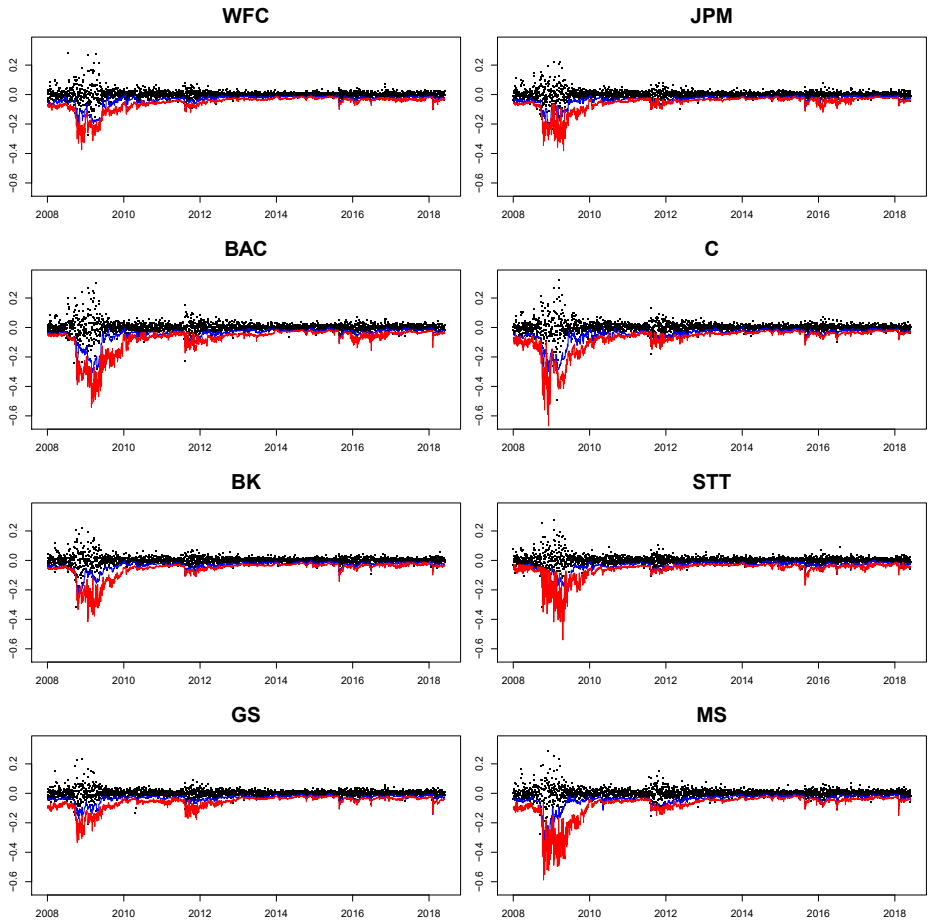


Figure 4.13: Plot of Returns (black dots), VaR (blue line) and CoVaR estimated by neural network quantile regression (red line), $\tau = 5\%$.

Bibliography

Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2017). Measuring systemic risk. *The Review of Financial Studies*, *30*(1), 2–47.

Adrian, T., & Brunnermeier, M. K. (2016). Covar. *The American Economic Review*, *106*(7), 1705.

Bianchi, D., Billio, M., Casarin, R., & Guidolin, M. (2019). Modeling systemic risk with markov switching graphical sur models. *Journal of Econometrics*, *210*(1), 58–74.

Bianchi, D., Büchner, M., & Tamoni, A. (2020). Bond risk premiums with machine learning. *The Review of Financial Studies*.

Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, *104*(3), 535–559.

Brownlees, C., & Engle, R. F. (2017). Srisk: A conditional capital shortfall measure of systemic risk. *The Review of Financial Studies*, *30*(1), 48–79.

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & geosciences*, *37*(9), 1277–1284.

Chao, S.-K., Härdle, W. K., & Wang, W. (2015). *Quantile regression in risk calibration.* Springer.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, *6*, 5549–5632.

Chen, X., & Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Chen, X., & White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, *45*(2), 682–691.

Chernozhukov, V., & Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, *26*(1), 271–292.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303–314.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Diebold, F. X., & Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, *182*(1), 119–134.

Engle, R. F., & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649.

Grenander, U. (1981). *Abstract inference* (tech. rep.).

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Härdle, W. K., Wang, W., & Yu, L. (2016). Tenet: Tail-event driven network risk. *Journal of Econometrics*, *192*(2), 499–513.

Hautsch, N., Schaumburg, J., & Schienle, M. (2014). Financial network systemic risk contributions. *Review of Finance*, *19*(2), 685–738.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580.*

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Johnson, J. (2018). Deep, skinny neural networks are not universal approximators. *arXiv preprint arXiv:1810.00393*.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R., & Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, 43–61.

Kuan, C.-M., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric reviews*, *13*(1), 1–91.

Kuester, K., Mittnik, S., & Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, *4*(1), 53–89.

Rolnick, D., & Tegmark, M. (2017). The power of deeper networks for expressing natural functions. *arXiv preprint arXiv:1705.05502*.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, *5*(3), 1.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sonoda, S., & Murata, N. (2017). Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, *43*(2), 233–268.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311.

White, H. (1988). Economic prediction using neural networks: The case of ibm daily stock returns. *ICNN*, *2*, 451–458.

White, H. (1992). Nonparametric estimation of conditional quantiles using neural networks. *Computing science and statistics* (pp. 190–199). Springer.

Xu, Q., Liu, X., Jiang, C., & Yu, K. (2016). Quantile autoregression neural network model with applications to evaluating value at risk. *Applied Soft Computing*, *49*, 1–12.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Chapter 5

On Cointegration and Cryptocurrency Dynamics

PUBLICATION
Keilbar, G. and Zhang, Y. (2021). On Cointegration and Cryptocurrency Dynamics, *Digital Finance*, 3 (1), 1–23.

5.1 Introduction

Cryptocurrencies have emerged as a new asset class over recent years. As of 2020, the crypto universe includes almost 5000 currencies with a total market capitalization close to 200 bn USD (coinmarketcap.com). We refer to Härdle et al. (2020) for a general overview on cryptocurrencies. While the market is still dominated by Bitcoin (BTC), the analysis of the interdependence of cryptocurrencies received a lot of attention from researchers as well as practitioners. For instance, Guo et al. (2018) analyzed latent communities from a network perspective. A large strand of literature is concerned with the relation of cryptocurrencies to other more traditional classes of assets (Shahzad et al. (2019), Corbet et al. (2018)). Yi et al. (2018) and Ji et al. (2019) analyzed directional volatility spillover effects using the variance decomposition method of Diebold and Yılmaz (2014). Sovbetov (2018) analyzed the cointegration of a VAR system of four cryptocurrencies. Leung and Nguyen (2019) proposed and discussed cointegration-based trading strategies.

While existing research contributions on cointegration restrict their focus to a small number of currencies, we argue that this only paints an incomplete picture. This paper aims to model the joint dynamics of cryptocurrencies in a nonstationary and high dimensional setting. In particular, we investigate the role of potential cointegration relationships among cryptocurrencies. In our empirical analysis we consider the ten largest currencies in terms of market capitalization in the period from July 2017 to February 2020.

Our methodology is based on the vector error correction model (VECM), developed by

Engle and Granger (1987), which augments the standard vector autoregressive (VAR) model with an additional role for deviations from long-run equilibria. To analyze the cointegration of cryptocurrencies in a dynamic setting, we propose a novel nonlinear VECM model, which we call *COINtensity* (cointegration intensity) VECM. The use of nonlinear specifications to model time series has a long tradition, see the monographs of Granger and Teräsvirta (1993) and Fan and Yao (2008)). Examples for nonlinear time series models include the smooth transition autoregressive (STAR) model (Luukkonen et al. (1988) and Teräsvirta (1994)) and neural networks (Kuan and White (1994) and Lee et al. (1993)). Nonlinear error correction models are discussed in Dijk et al. (2002) and extended to the vector case by Kristensen and Rahbek (2010). An advantage of nonlinear time series models is the increased flexibility compared to linear specifications. Usually, this flexibility comes at the expense of a large number of parameters to estimate. Our *COINtensity* VECM specification has the advantage that the number of additional parameters is equal to the cointegration rank, i.e. it is non increasing in the dimension of the VAR system. The nonlinear part of the model introduces a time-varying intensity effect for the error adjustment, which implies that the cyptocurrencies will return to the long-run equilibrium with varying speed. A crucial task is to select the number of those equilibria, also referred to as cointegration relations. Johansen (1988, 1991) proposed a likelihood ratio test, which is now commonly used. However, the testing procedure suffers from poor finite sample performance in systems of more than three variables (Johansen (2002) and Liang and Schienle (2019)). We therefore follow Onatski and Wang (2018), who proposed an alternative test for cointegration that is designed for a high-dimensional setting.

Our empirical results suggest that cointegration plays a crucial role for cryptocurrencies. In particular, we find four stationary long-run equilibria. We also find that all currencies are significantly affected by long-term stochastic trends, rejecting the hypothesis of weak exogeneity. The results of our dynamic *COINtensity* VECM show a time-varying dependence of cryptocurrencies on these stochastic trends. We find that the nonlinearity of error correction is stronger during the time of the cryptocurrency bubble, compared to a later time period. Based on our estimated cointegration vectors, we construct a simple trading rule, following and generalizing the strategy of Leung and Nguyen (2019). An in-sample analysis of our trading strategy indicates that trading on large deviations from the long-run equilibria can be profitable, while the out-of-sample analysis is more cautious. In particular, the success of such a statistical arbitrage strategy is dependent on the condition that the equilibrium relations will hold in the long-run.

The contributions of this paper are two-fold. First, it is the first attempt to model a system of cryptocurrencies in a large vector autoregression while accounting for nonstationary effects. Second, we propose a novel, nonlinear VECM specification which

increases the flexibility and also has a good interpretability even in large dimensions.

The remainder of the paper is organized as follows. Section 2 describes in detail the steps of our modelling and estimation procedure. To show the validity of our approach, we conduct a small simulation study in section 3. In section 4, we apply our methodology to a system of the largest ten cryptocurrencies. Section 5 introduces a simple cointegration-based trading strategy and section 6 concludes.

All codes of this paper are available on quantlet.de.

5.2 Modelling Framework

5.2.1 VECM and Testing for Cointegration

As a baseline model we consider the following p-dimensional vector autoregressive model with error correction term (VECM).

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{k} \Gamma_i \Delta X_{t-i} + \Phi D_t + \varepsilon_t, \tag{5.1}$$

where D_t are deterministic variables and ε_t are zero-mean, independent error terms. We assume that each univariate time series is integrated of order one, $X_{it} \sim I(1), i = 1, \ldots, p$. Under cointegration, there exists a linear combination which is stationary, i.e. $\beta^\top X_t \sim I(0)$. Thus, we can rewrite (1) in the following way,

$$\Delta X_t = \alpha\beta^\top X_{t-1} + \sum_{i=1}^{k} \Gamma_i \Delta X_{t-i} + \Phi D_t + \varepsilon_t, \tag{5.2}$$

where β is a $p \times r$ matrix of cointegration vectors and α is the $p \times r$ loading matrix. The order of cointegration is characterized by the rank r of β. Γ_i, $i = 1, \ldots, k$, are $p \times p$ parameter matrices associated with the impact of lagged values of ΔX_t.

Johansen (1988, 1991) developed a sequential likelihood testing procedure to determine the cointegration rank r. Under the null hypothesis there are at most r cointegration relationships.

$$H_0 : \text{rank}(\Pi) \leq r \quad \text{vs} \quad H_1 : \text{rank}(\Pi) > r \tag{5.3}$$

In the special case of $r = 0$, there is no cointegration and we have to proceed with a stationary VAR model in first differences. On the other hand, if $r = p$, we can use a stationary VAR model in levels without any error correction terms. In all other cases, $0 < r < p$, the series are cointegrated.

The test statistic LR is based on the squared canonical correlations between the residuals obtained by regressing ΔX_t and X_{t-1} on the lagged differences $(\Delta X_{t-1}, \ldots, \Delta X_{t-k})$ and the deterministic variables D_t, respectively. These correspond to the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ of the matrix $S_{01}S_{11}^{-1}S_{01}^{\top}S_{00}^{-1}$, with $S_{00} = \frac{1}{T}R_{0t}R_{0t}^{\top}$, $S_{01} = \frac{1}{T}R_{0t}R_{1t}^{\top}$ and $S_{11} = \frac{1}{T}R_{1t}R_{1t}^{\top}$. R_{0t} are the residuals of regressing ΔX_t and R_{1t} are the residuals of regressing X_{t-1} on $(\Delta X_{t-1}, \ldots, \Delta X_{t-k})$ and D_t.

$$LR = -T \sum_{i=r+1}^{p} \log(1 - \lambda_i). \tag{5.4}$$

Under the null hypothesis, the test statistic converges in distribution to a function of Brownian motions. The limiting distribution is different according to the specific form of D_t, see Proposition 8.2 in Lütkepohl (2005). The critical values of the Johansen test are obtained by simulations.

The test has been proved to have issues in small samples, in particular if the dimension of the VAR model, p, becomes large. This issue is addressed in Johansen (2002). Onatski and Wang (2018) therefore developed a different asymptotic setting. In particular, they consider the case where T and p go to infinity simultaneously such that $p/T \to c \in (0, 1]$. Consider a simplified representation of (1) without lagged differences.

$$\Delta X_t = \Pi X_{t-1} + \Phi D_t + \varepsilon_t \tag{5.5}$$

Under this asymptotic regime and under the null hypothesis of no cointegration, the empirical distribution function of the eigenvalues of the matrix $S_{01}S_{11}^{-1}S_{01}^{\top}S_{00}^{-1}$ converges weakly to the Wachter distribution.

$$F_p(\lambda) \Rightarrow W_c(\lambda) \stackrel{\text{def}}{=} W\left(\lambda; c/(1+c), 2c/(1+c)\right) \tag{5.6}$$

where $F_p(\lambda) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{I}(\lambda_i \leq \lambda)$ and $W(\lambda, \gamma_1, \gamma_2)$ denotes the Wachter distribution function with parameters $\gamma_1, \gamma_2 \in (0, 1)$ and density $f_W(\lambda, \gamma_1, \gamma_2) = \frac{1}{2\pi\gamma_1}\frac{\sqrt{(b_+ - \lambda)(\lambda - b_-)}}{\lambda(1-\lambda)}$ on $[b_-, b_+]$ with $b_\pm = \left(\sqrt{\gamma_1(1-\gamma_2)} \pm \sqrt{\gamma_2(1-\gamma_1)}\right)^2$ and atoms of size $\max(0, 1 - \gamma_2/\gamma_1)$ at zero and $\max(0, 1 - \frac{1-\gamma_2}{\gamma_1})$ at unity. The rank of cointegration can be determined graphically by comparing the empirical quantiles of the calculated eigenvalues with the theoretical quantiles of the Wachter distribution. Under the null hypothesis of no cointegration the empirical quantiles of eigenvalues should lie close to the theoretical quantiles of the Wachter. Onatski and Wang (2018) suggest to select the cointegration rank by the number of eigenvalues which deviate from the 45 degree line. We show the validity of this approach in a simulation study in section 3.

If the rank of the matrix of cointegration vectors is known, we can estimate cointegration

vectors β by reduced rank maximum likelihood estimation, corresponding to the r largest eigenvalues of the matrix $S_{01}S_{11}^{-1}S_{01}^{\top}S_{00}^{-1}$, which we defined in the previous subsection. Without normalization, this estimator is not unique. Therefore, we set the j-th element in the j-th cointegration vector to one Johansen (1995). Then, we can estimate the remaining parameters α and $\Gamma = (\Gamma_1 : \ldots : \Gamma_k)$ with equation-wise OLS by plugging in the estimator for β, and give their asymptotically normal distribution using standard arguments for stationary processes.

5.2.2 *COINtensity* VECM

As an extension to the baseline setting, we consider a nonlinear VECM specification. Such models originate from Granger and Teräsvirta (1993), who introduced the smooth transition error correction model (STECM). A vector version was proposed by Dijk et al. (2002). Kristensen and Rahbek (2010) considered the general setting of likelihood-based estimation with nonlinear error correction. Corresponding linearity tests and inference-related issues are discussed in Kristensen and Rahbek (2013). The general setting can be formulated as follows.

$$\Delta X_t = g\left(\beta^{\top} X_{t-1}; \theta\right) + \sum_{i=1}^{k} \Gamma_i \Delta X_{t-i} + \Phi D_t + \varepsilon_t, \tag{5.7}$$

where $g(\cdot)$ is a parametric error correction function with parameter vector θ. The error correction function can be nonlinear in the long term stochastic trends as well as in θ. In the baseline linear setting, $g(z;\theta) = \alpha z$ and $\theta = \mathrm{vec}(\alpha)$. In the vector version of the STECM we have $g(z;\theta) = \{\alpha + \widetilde{\alpha}\psi(z;\psi)\}$, where $\psi(z;\phi)$ is a fixed function satisfying $|\psi(z;\phi)| = \mathcal{O}(1)$ as $\|z\| \to \infty$, and $\theta = (\mathrm{vec}(\alpha)^{\top}, \mathrm{vec}(\widetilde{\alpha})^{\top}, \mathrm{vec}(\phi)^{\top})^{\top}$, where vec is the vector operator that transforms matrix $A_{m\times n}$ into an $(mn \times 1)$ vector by stacking the columns.

The advantage of using nonlinear models is an increased degree of flexibility. However, often this flexibility comes at the expense of worse interpretability and of overfitting the data. We therefore introduce a new class of vector error correction models, which we call *COINtensity* (cointegration intensity) VECM.

$$\Delta X_t = \alpha\beta^{\top} X_{t-1}\left\{1 + G\left(s_t;\gamma\right)\right\} + \sum_{i=1}^{k} \Gamma_i \Delta X_{t-i} + \Phi D_t + \varepsilon_t, \tag{5.8}$$

where s_t is a d-dimensional vector of transition variables and $G(\cdot) : \mathbb{R}^d \to (-1, 1)$ is a parametric function with parameter vector $\gamma \in \mathbb{R}^d$. We propose the following parameterisation, $G(s_t;\gamma) = \tanh(s_t^{\top}\gamma)$ and $s_t = \beta^{\top} X_{t-1}$, where tanh is the sigmoid tangent function. We denote $G(\cdot)$ as the *COINtensity* (cointegration intensity) function.

This function has a universal effect for all cryptocurrencies and measures the intensity of the impact of cointegration. $G(\cdot)$ takes values in $(-1, 1)$. In this model specification, we still have a loading matrix α which measures currency-specific marginal effects. Please note that our *COINtensity* VECM is a generalization of the baseline model, as model (5.8) reduces to model (5.2) if $\gamma = 0$.

Our model specification has two advantages. First, it has only a few additional parameters compared to the baseline specification. The overfitting problem of nonlinear error correction models can therefore be contained. Second, the modified model enables us to analyze cointegration and the exposure of cryptocurrencies to long-term equilibrium relationships in a dynamic context.

If the cointegration vectors β are estimated a priori, model parameters can be estimated by quasi maximum likelihood estimation (QMLE). For convenience, we write $\theta \stackrel{\text{def}}{=} (\text{vec}(\alpha)^\top, \text{vec}(\Gamma)^\top, \gamma^\top)^\top$. The QMLE, $\widehat{\theta}$ of θ, is defined as the minimizer of the following negative log-likelihood criterion,

$$L_T(\theta) = \sum_{t=1}^{T} \varepsilon_t^\top(\theta)\varepsilon_t(\theta). \tag{5.9}$$

We split the parameters into two parts and write $\theta = (\text{vec}(\theta_1)^\top, \theta_2^\top)^\top$, with $\theta_1 = (\alpha, \Gamma)^\top$ and $\theta_2 = \gamma$. Note that θ_1 is a $(r + pk) \times p$ parameter matrix. Further, we define

$$W_t(\theta_2) \stackrel{\text{def}}{=} \left(\left[\beta^\top X_{t-1} \left\{1 + \tanh\left(\theta_2^\top \beta^\top X_{t-1}\right)\right\}\right]^\top, \Delta X_{t-1}^\top, \ldots, \Delta X_{t-k}^\top \right)^\top, \tag{5.10}$$

where $W_t(\theta_2) \in \mathbb{R}^{r+pk}$. Now, we can rewrite model (5.8) as follows.

$$\Delta X_t = \theta_1^\top W_t(\theta_2) + \varepsilon_t \tag{5.11}$$

The profile estimator for $\theta_1(\theta_2)$ can be obtained by standard OLS.

$$\widehat{\theta}_1(\theta_2) = \left\{ \sum_{t=1}^{T} W_t(\theta_2) W_t^\top(\theta_2) \right\}^{-1} \sum_{t=1}^{T} W_t(\theta_2) \Delta X_t^\top \tag{5.12}$$

We proceed by obtaining the corresponding vector of residuals.

$$\widehat{\varepsilon}_t(\theta_2) = \Delta X_{t-1} - \widehat{\theta}_1^\top W_t(\theta_2) \tag{5.13}$$

Given the profile estimator, we can estimate θ_2 by

$$\widehat{\theta}_2 = \arg\min_{\theta_2 \in \Theta_2} L_T(\widehat{\theta}_1(\theta_2), \theta_2), \tag{5.14}$$

where Θ_2 is the parameter space of θ_2. The final estimator for θ_1 can be obtained

by plugging (5.14) into (5.12). The interpretation of the parameters α, β and Γ is almost completely analogous to the linear VECM. In particular, the cointegration vector has the same function as before, governing the long run equilibrium relations. The only difference in the interpretation of parameters is that the loading intensity is now time-varying. Regarding the selection of the cointegration rank r we cannot make a definitive statement whether the asymptotics of Onatski and Wang (2018) also hold in the *COINtensity* model. However, the procedure seems to work well in practice, as shown in our simulation study.

5.3 Simulation Study

In the first part of this simulation study, we examine the validity of the procedure of Onatski and Wang (2018) to test for cointegration. They suggest to determine the cointegration rank graphically by comparing the empirical quantiles of the eigenvalues with the theoretical quantiles of the Wachter distribution. The cointegration rank is chosen according to the number of eigenvalues deviating from the 45 degree line. Here, we calibrate the numerical example in Liang and Schienle (2019), which is an 8-dimensional VAR(2) process with four unit roots, i.e $p = 8$, $r = 4$, $k = 1$.

$$\Delta X_t = \alpha\beta^\top X_{t-1} + \Gamma_1 \Delta X_{t-1} + \varepsilon_t, \tag{5.15}$$

with full-rank matrices α, β of dimension $p \times r$ and iid-distributed ε_t generated from $N(0, I_8)$. We consider $T = 200$, matrices α, β and Γ_1 are listed in the appendix (setting 1.1). Figure 5.1 shows that there are exactly four eigenvalues that deviate from the 45 degree line, which also supports the simulation result of Onatski and Wang (2018), while the Johansen test rejects the null hypothesis of a cointegration rank smaller than or equal to four at 5% significance level, implying five cointegration relationship.

As a second setting, we consider the *COINtensity* VECM as our data generating process,

$$\Delta X_t = \alpha\beta^\top X_{t-1} \left\{1 + \tanh\left(\gamma^\top \beta^\top X_{t-1}\right)\right\} + \varepsilon_t. \tag{5.16}$$

We consider the high dimensional case of $p = 15$ and $r = 3$. The error term is iid and generated from $N(0, 0.05 \cdot I_{15})$. We choose a sample size of $T = 200$. The parameter matrices are listed in the appendix (setting 1.2). Figure 5.2 shows exactly three eigenvalues deviating from the 45 degree line. The testing procedure seems to also work well in the high dimensional and nonlinear setting. So, we apply the Wachter Q-Q plot to decide the number of cointegration in our large dimensional model.

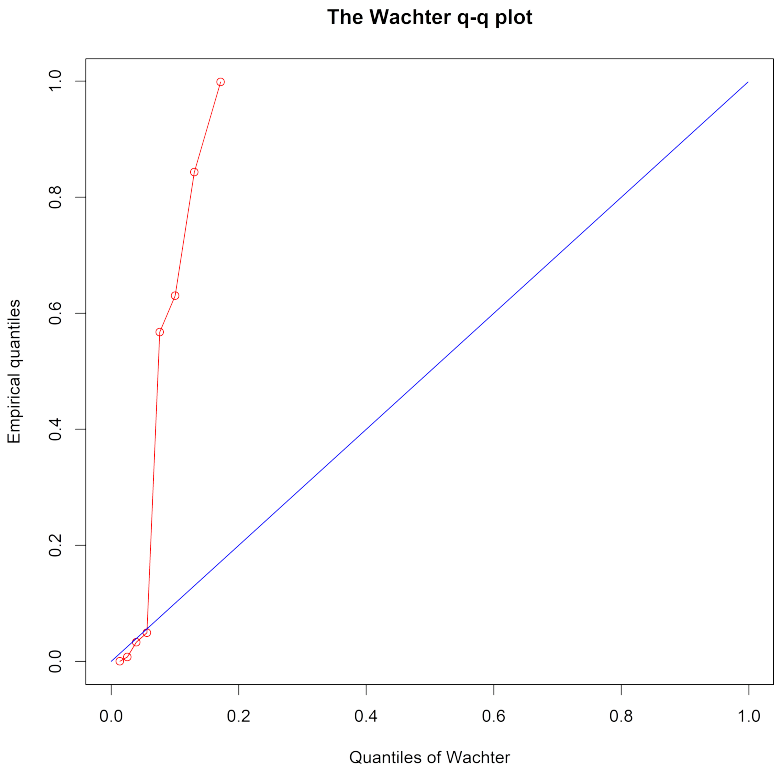


Figure 5.1: The Wachter Q-Q plot for the linear data generating process. The plot shows that the number of eigenvalues deviating from the 45 degree line is equal to the true cointegration rank, $r = 4$.

In the second part of the simulation study, we investigate the finite-sample properties of our estimator for the *COINtensity* VECM. We follow the study design of Kristensen and Rahbek (2010), focusing on the case where $p = 2$ and the number of cointegration relation is $r = 1$. Further, the number of lagged differences entering our model is $k = 1$. We consider four different sample sizes, $T \in \{250, 500, 1000, 2000\}$. The cointegration vector is assumed to be estimated in advance, $\beta = (1, -1)^\top$. The loading parameters are set to $\alpha_1 = 0.5$ and $\alpha_2 = -0.5$. The elements matrix of parameters associated with the lagged first differences are set to $\Gamma_{jk} = 0.05$ for $j, k = 1, 2$. Finally, we set $\gamma = 0.2$. For each sample size, we simulate 1000 sample paths of our VECM specification. We evaluate the performance of the estimator by the root mean square error (RMSE). The simulation results can be found in Table 5.1.

	$T = 250$	$T = 500$	$T = 1000$	$T = 2000$
α_1	0.1136	0.0910	0.0736	0.0581
α_2	0.0448	0.0322	0.0246	0.0187
γ	0.3584	0.2887	0.2218	0.1696

Table 5.1: RMSE for QMLE of individual parameters in our *COINtensity* VECM.
CryptoDynamics_Simulation

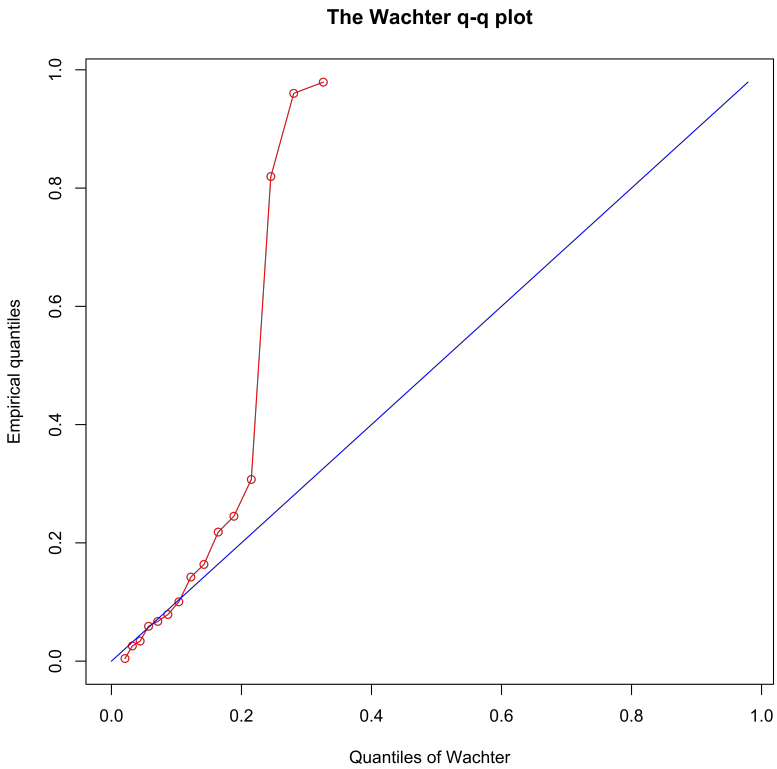


Figure 5.2: The Wachter Q-Q plot for the nonlinear data generating process. The plot shows that the number of eigenvalues deviating from the 45 degree line is equal to the true cointegration rank, $r = 3$.

For the individual-specific parameters, α_1 and α_2, we can observe a good estimation accuracy already in small and moderate samples. As expected, the estimates become more precise with increasing sample size T. This is also the case for γ, which governs the intensity by which the individual series are affected by deviations from the long-term equilibrium. However, the estimates for γ are not as precise as for the former parameters.

We consider a second setting for the evaluation of our estimation procedure. In this setting we have the relative large dimensional case of $p = 8$ and $r = 3$. The true values for α, β and γ are again listed in the appendix (setting 2.2). The simulation results, based on 1000 Monte Carlo iterations, can be found in Table 5.2. We report the average Frobenius error of estimating the loading matrix α, $\|\widehat{\alpha} - \alpha\|_F$, as well as the RMSE of γ_1,γ_2 and γ_3. The result confirm that the estimation error can be effectively reduced with increasing sample size, even if the dimensionality is comparably high.

	$T = 250$	$T = 500$	$T = 1000$	$T = 2000$
α	0.3965	0.3440	0.2948	0.2669
γ_1	0.0634	0.0457	0.0327	0.0298
γ_2	0.0627	0.0416	0.0466	0.0363
γ_3	0.0752	0.0714	0.0638	0.0447

Table 5.2: The first row shows the average Frobenius error, $\|\widehat{\alpha} - \alpha\|_F$, for the estimated parameter matrix $\widehat{\alpha}$. The remaining rows show the RMSE for QMLE of $\widehat{\gamma}$ in our *COINtensity* VECM.

CryptoDynamics_Simulation

5.4 Dynamics of Cryptocurrencies

5.4.1 Data and Descriptive Statistics

In the empirical part of the paper, we analyze the joint dynamics of the largest cryptocurrencies. In particular, we are interested in the following set of questions.

I. Do cointegration relations exist among cryptocurrencies?

II. Which cryptocurrencies affect and which are affected by long-term equilibrium effects?

III. How does the impact of the cointegration relationships change in a dynamic setting?

We use daily time series data of the largest ten cryptocurrencies, which we obtained from Coinmarketcap.com. Since some of the currencies have a very short trading history, we restrict our analysis to those with a time series dating back to at least July 2017. The reason for this decision is to include the boom and the bust of the crypto-bubble at the end of 2017 and start of 2018. To avoid pathological cases, we also remove stable coins such as Tether (USDT). Stable coins are characterized by a fixed exchange rate with the USD and are therefore expected to be stationary in levels. The list of currencies included in our analysis can be found in Table 5.3. In total, we have 945 daily price observations from July 25, 2017 until February 25, 2020.

The aggregated market capitalization of our sample is around 230 bn USD and captures more than 95% of the total market capitalization of cryptocurrencies. Our analysis therefore has a high degree of external validity. By looking at Table 5.3, it becomes apparent that the crypto market is still dominated by Bitcoin. However, also ETH and XRP occupy a dominant position in the market.

Currency	Symbol	Market Cap (10^6 USD)	Avg Return (%)	σ
Bitcoin	BTC	170,370	0.181	0.019
Ethereum	ETH	27,223	0.077	0.020
XRP	XRP	11,087	0.028	0.022
Bitcoin Cash	BCH	6,477	0.133	0.047
Litecoin	LTC	4,567	0.092	0.025
EOS	EOS	3,764	0.107	0.034
Binance Coin	BNB	3,164	0.338	0.053
Monero	XMR	1,978	0.204	0.031
Stellar	XLM	1,320	0.222	0.045
Ethereum Classic	ETC	1,076	0.066	0.032

Table 5.3: List of cryptocurrencies and descriptive statistics. Market capitalization as of February 25, 2020, obtained from Coinmarketcap.com.

CryptoDynamics_Scraping

Figure 5.3 shows the development of the log prices over time. The multivariate time series reveals a strong co-movement of cryptocurrencies. For instance, we can observe a sharp rise in prices for all currencies at the end of 2017, followed by a sharp decrease at the beginning of 2018 during burst of the cryptocurrency bubble. This empirical observation suggests a dependence of currencies in levels, not only in first differences. It is thus an essential task to account for cointegration, when analyzing the joint dynamics of cryptocurrencies. Failing to do so would only paint an incomplete picture.

Before any cointegration analysis can be done, one has to assure that all the currencies series are non-stationary and integrated of the same order. Performing the Augmented Dickey-Fuller (ADF) test with a constant and a time trend, the null hypothesis of a unit root cannot be rejected for the individual logged prices at 90% level. The lag length k for the ADF test has been selected by the Ng and Perron (1995) downtesting procedure starting with a maximum lag of 12. However, the results of the ADF test are not sensitive to the choice of k and the null cannot be rejected for any number of lagged terms in each of the series.

In the next step, we apply differences of the time series and compute the ADF test statistic on the differenced data. This time, the null of non-stationarity is rejected for all indices at the 99% level. This suggests that daily returns follow a stationary process. Since the original series must be differenced one time in order to achieve stationarity, we conclude that the cryptocurrency prices are integrated of order one, such that the vector X_t is $I(1)$. The results of the tests are summarized in Table 5.4. Having confirmed that all the series are integrated of the same order, this allows to test for cointegration.

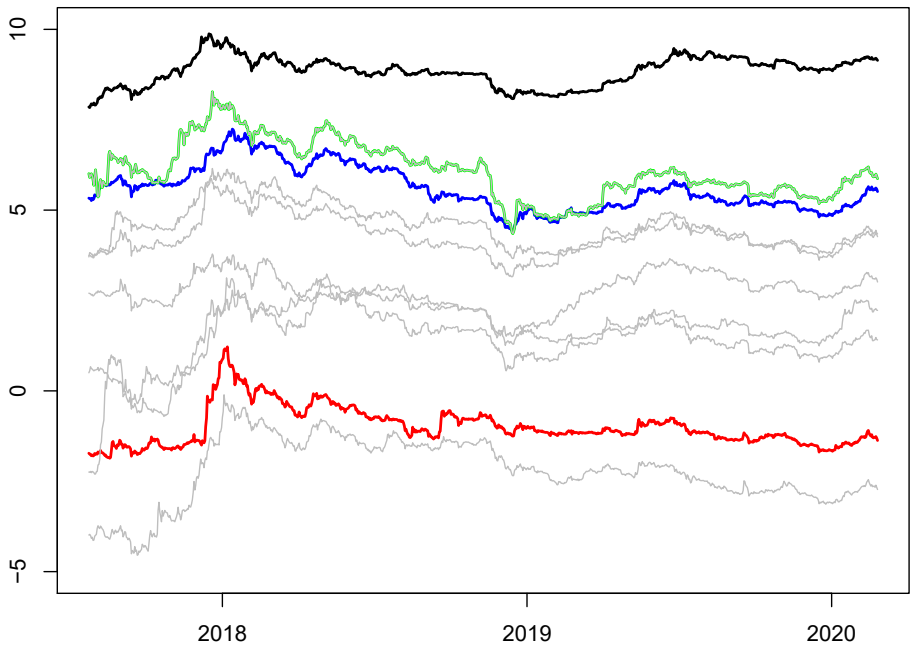


Figure 5.3: Time series of log prices from July 2017 - February 2020. BTC, ETH, XRP, BCH and all others.

CryptoDynamics_Series

	X_t		ΔX_t	
	ADF	KPSS	ADF	KPSS
BTC	0.76	< 0.01	< 0.01	> 0.1
ETH	0.56	< 0.01	< 0.01	> 0.1
XRP	0.21	< 0.01	< 0.01	> 0.1
BCH	0.59	< 0.01	< 0.01	> 0.1
LTC	0.60	< 0.01	< 0.01	> 0.1
EOS	0.41	< 0.01	< 0.01	> 0.1
BNB	0.40	< 0.01	< 0.01	0.04
XMR	0.62	< 0.01	< 0.01	> 0.1
XLM	0.28	< 0.01	< 0.01	0.07
ETC	0.39	< 0.01	< 0.01	> 0.1

Table 5.4: p-values of the stationary tests for the level and first difference data.

5.4.2 Estimation Results for Linear VECM

In the first step, we determine the cointegration rank graphically by using the Wachter Q-Q plot proposed by Onatski and Wang (2018). As explained in the last section, large deviations of the empirical quantiles of eigenvalues from the theoretical quantiles of the Wachter distribution indicate that the present matrix does not have full rank. We conclude from Figure 5.4 that there are four cointegration relations since we can observe four eigenvalues deviating from the 45 degree line.

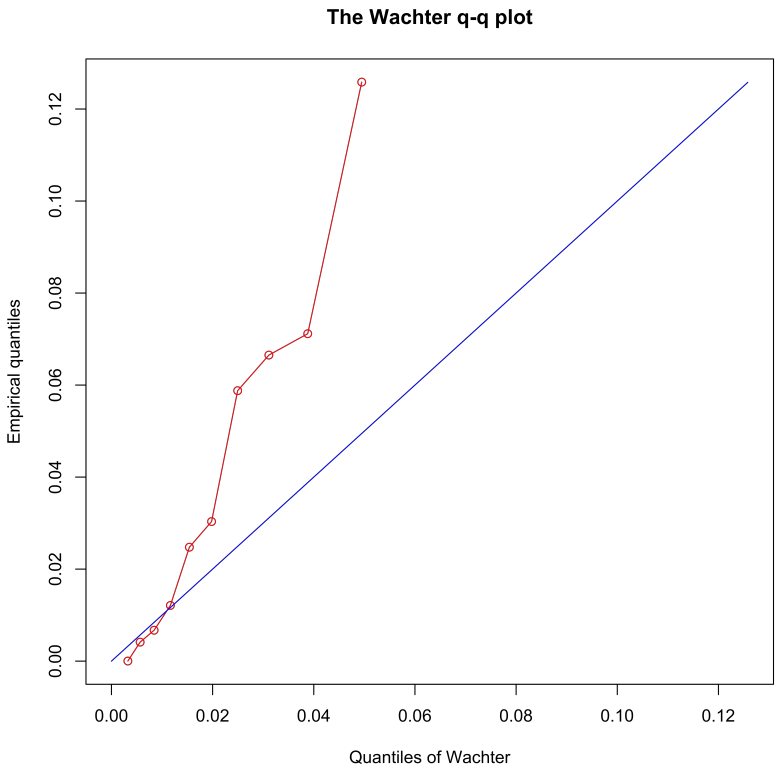


Figure 5.4: Wachter QQ plot to determine the cointegration rank r.
CryptoDynamics_Wachter

	BTC	ETH	XRP	BCH	LTC	EOS	BNB	XMR	XLM	ETC
β_1	1.00	0.00	0.00	0.00	1.98	0.13	-0.94	-3.42	0.57	0.70
β_2	0.00	1.00	0.00	0.00	-0.28	-0.27	0.24	-1.09	0.11	0.31
β_3	0.00	0.00	1.00	0.00	-0.97	0.39	0.20	0.54	-0.76	0.00
β_4	0.00	0.00	0.00	1.00	0.53	-0.43	-0.06	-1.27	0.37	-0.42

Table 5.5: Estimated cointegration vectors $\widehat{\beta}$.
CryptoDynamics_Estimation

Having fixed the cointegration rank, we can proceed with estimating the cointegration vectors. The estimated coefficients can be found in Table 5.5. To make the estimator unique, we normalize the j-th entry of the j-th cointegration vector to 1. Due to this normalization, we have one vector associated with each of the four largest currencies. For instance, we can observe for β_1 that the entry for BTC is one whereas the entries for ETH, XRP and BCH are all close to zero. Based on these estimation results, we plot the time series of our four stochastic trends in Figure 5.5. Apart from the beginning of our observation period and apart from the crypto bubble of 2017/2018, we can observe steady and mean-reverting stochastic trends. These observations can be confirmed statistically. Results from the ADF test reject the hypothesis that these trends have a unit root. We can continue to estimate the short-run parameters α and Γ. In the following, we select the lag order, $k = 1$, using the Bayesian information criterion (BIC).

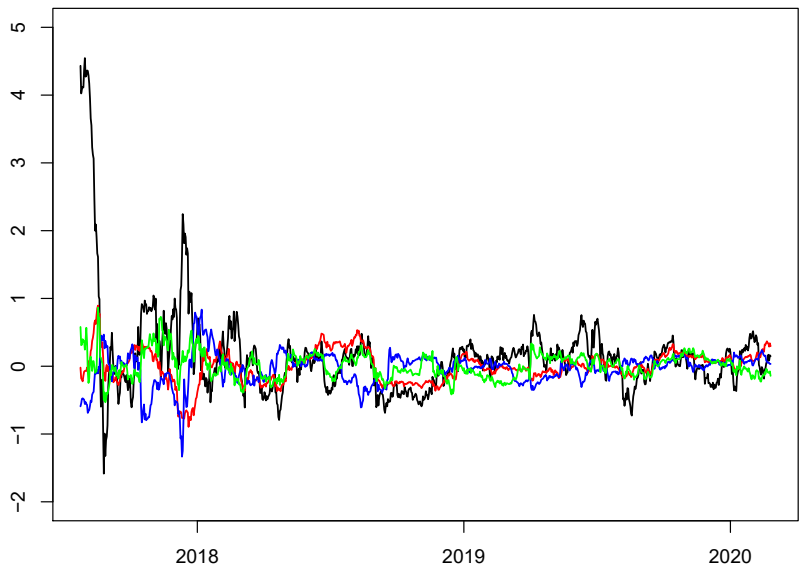


Figure 5.5: Time series of the long-term stochastic trends. $\widehat{\beta}_1^\intercal X_{t-1}$, $\widehat{\beta}_2^\intercal X_{t-1}$, $\widehat{\beta}_3^\intercal X_{t-1}$ and $\widehat{\beta}_4^\intercal X_{t-1}$.

CryptoDynamics_Estimation

The estimation results of our baseline VECM indicate that cointegration plays an important role for cryptocurrencies. See Table 5.6 for the estimation of the loading matrix α. The (j,i)-th entry of the table shows how currency j is affected by error correction term i, where $ECT_{j,t-1} \stackrel{\text{def}}{=} \widehat{\beta}_j^\intercal X_{t-1}$. Almost all currencies are significantly affected by at least one stochastic trend, with BTC and LTC being the only exceptions. We additionally test the hypothesis of weak exogeneity to examine whether a given currency is unaffected by all stochastic trends. The null and alternative hypotheses are:

$$H_0: \ \alpha_{j,1} = \ldots = \alpha_{j,r} = 0 \quad \text{vs.} \quad H_1: \ \exists \ \alpha_{j,k\leq r} \neq 0 \tag{5.17}$$

The test statistic is constructed as a classical Wald statistic. We reject the null hypothesis for all currencies at a significance level of 0.1%. Cointegration therefore has universal effects. The long-run linkages between the indices suggest that cryptocurrency prices are not independent, but predictable using information of others. The results also suggest that investors who seek to diversify their portfolios internationally should be aware that the ten cryptocurrency prices in the system follow a common stochastic trend. This means that these markets generate similar returns in the long-run. Therefore, diversification across the markets is limited and investors should include other markets with lower correlation to hedge their risk.

In the first error correction term, ETH and BNB do not tend to return to the long-run equilibrium as the coefficient on the error term is positive. In the second one, ETH, XRP, EOS and XLM all have the predicted negative sign, which indicates that the

disequilibrium given in the error correction term will be reduced period by period. However, the size of the estimates differs widely and is quite small compared to the other short-term adjustment parameters. These results suggest that distortions in the long-run equilibrium will be corrected slowly and unevenly among the ten cryptocurrencies. In the third one, XRP, BCH and EOS carry the burden of adjustment to return to the long-run relationship. In the fourth one, EOS, XMR, ETC are the leaders in the system and that BCH carries the burden of adjustment to return to the long-run relationship.

	ECT_1	ECT_2	ECT_3	ECT_4
BTC	0.0045	-0.0017	-0.0119	0.0005
ETH	0.0084	-0.0228	0.0098	0.0085
XRP	-0.0010	-0.0385	-0.0287	0.0164
BCH	0.0044	0.0024	-0.0348	-0.0403
LTC	0.0019	-0.0144	-0.0185	-0.0138
EOS	-0.0068	-0.0293	-0.0382	0.0438
BNB	0.0308	-0.0199	0.0142	-0.0047
XMR	0.0024	0.0026	-0.0123	0.0268
XLM	0.0067	-0.0276	0.0205	0.0205
ETC	-0.0009	-0.0096	0.0007	0.0282

Table 5.6: Estimated loading matrix $\widehat{\alpha}$. Red color indicates significance of negative coefficients, blue color indicates significance of positive coefficients, with significance at 5%, 1% and 0.1% level.

CryptoDynamics_Estimation

	BTC	ETH	XRP	BCH	LTC	EOS	BNB	XMR	XLM	ETC
BTC	0.08	-0.08	-0.03	-0.05	-0.06	0.07	0.00	0.06	0.02	-0.04
ETH	-0.07	0.05	-0.07	0.00	0.07	-0.02	0.01	0.02	0.02	-0.08
XRP	-0.17	0.06	0.11	-0.03	0.03	0.01	0.05	0.06	-0.08	-0.12
BCH	-0.28	0.13	-0.09	0.19	-0.05	-0.00	0.08	0.06	0.01	-0.14
LTC	0.01	-0.11	-0.03	0.02	0.09	-0.05	0.02	0.03	-0.01	-0.02
EOS	-0.07	-0.06	-0.07	-0.03	0.11	0.00	0.08	0.02	0.01	-0.01
BNB	0.15	0.01	0.02	0.03	-0.18	0.01	0.18	-0.04	-0.13	-0.06
XMR	-0.05	-0.01	-0.07	-0.01	0.02	0.03	0.07	-0.04	-0.01	-0.05
XLM	0.04	-0.08	-0.04	-0.07	0.09	0.03	0.00	-0.03	0.13	-0.11
ETC	0.05	-0.01	-0.09	0.05	-0.00	0.05	0.01	-0.08	0.02	-0.07

Table 5.7: Estimated coefficient matrix $\widehat{\Gamma}$. Red color indicates significance of negative coefficients, blue color indicates significance of positive coefficients, with significance at 5%, 1% and 0.1% level.

CryptoDynamics_Estimation

The estimation results for the lagged differences can be found in Table 5.7. Compared to the estimated coefficients for the error correction terms, the lagged differences seem to be less important. Some currencies, such as BCH and BNB, have highly significant coefficients associated with their own lagged value. Another interesting observation is that BTC and BCH both depend on each other negatively.

5.4.3 Estimation Results for *COINtensity* VECM

All the previous results are obtained in the baseline linear VECM setting. For a dynamic analysis we henceforth rely on our *COINtensity* VECM. We estimate the model by the profile likelihood estimation framework introduced in section 2.2. In the first step, we estimate the cointegration vectors β as before. In practice, we then estimate the nonlinear part of the model by random parameter search. We assume that the parameter vector $\theta_2 = \gamma$ lies in $\Theta_2 = [-1, 1]^r$. The candidate parameters are generated from the r-dimensional uniform distribution in the same range. Our number of simulations is 10,000.

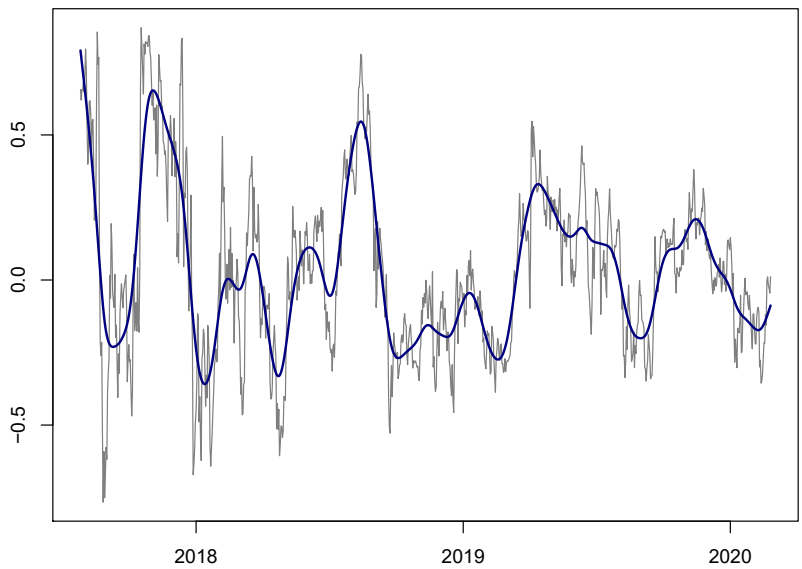


Figure 5.6: Time series of cointegration intensity $G(\widehat{\beta}^\top X_{t-1}; \widehat{\gamma})$ (grey) and spline interpolation (blue).

CryptoDynamics_Nonlinear

The time series of the estimated *COINtensity* function, $G(\widehat{\beta}^\top X_{t-1}; \widehat{\gamma})$, is visualized in Figure 5.6. We can observe a time-varying pattern of the intensity by which cryptocurrencies are affected by long run equilibrium effects. Prior to the building of the bubble at the end of 2017, cointegration intensity was low with values below zero. The following increase goes along with the strong increase in prices across all cryptocurrencies in the last quarter of the same year. The subsequent months can be characterized by a highly volatile cointegration intensity. Recently, from the second half of 2018, we can observe a period of stabilization with only few values exceeding the -0.5 and 0.5 thresholds. We conclude that nonlinearity was more prevalent in the turbulent period of the cryptocurrency bubble.

We also evaluate the out-of-sample predictive power of the *COINtensity* VECM compared to the linear baseline model. Even if prediction is not the main purpose of this

research, it can still provide insight into the usefulness of the nonlinear specification. For the out-of-sample analysis we consider the period from February 26 to October 13, 2020. The results can be found in Table 5.8. We report the root mean square error (RMSE) of prediction for both models and for each cryptocurrency separately. It becomes evident that the *COINtensity* specification outperforms the linear model. For nine out of ten currencies the RMSE is lower. We apply the test of Diebold and Mariano (2002) to test whether this outperformance is significant. We find that only for one currency (BNB) the forecast is significantly better.

	Linear	*COINtensity*	DM-test
BTC	0.0451	**0.0450**	0.1883
ETH	0.0614	**0.0579**	0.8132
XRP	0.0488	**0.0446**	0.8226
BCH	0.0580	**0.0579**	0.8991
LTC	**0.0504**	0.0506	0.8904
EOS	0.0575	**0.0541**	0.8004
BNB	0.0675	**0.0589**	**0.0413**
XMR	0.0538	**0.0527**	0.1571
XLM	0.0559	**0.0524**	0.7931
ETC	0.0539	**0.0539**	0.6541

Table 5.8: Out-of-sample predictive performance in terms of root mean squared error (RMSE) for the linear VECM and *COINtensity* VECM specification. Additionally, p-values from the Diebold-Mariano test are reported.

5.5 A Simple Statistical Arbitrage Trading Strategy

In this section, we apply a simple cointegration-based trading strategy for cryptocurrencies. We use the same data as in the previous section. Under the assumption of mean reversion of the long term stochastic trends, a large deviation from the equilibrium relationships should lead to profitable investment opportunities. In the following we define the cointegration spreads. For each cointegration relationship, $j = 1, \ldots, r$, we have

$$\begin{aligned} S_{j,t} &= \beta_j^\top X_t \\ &= \beta_{j,1} X_{1,t} + \ldots + \beta_{j,p} X_{p,t}. \end{aligned} \tag{5.18}$$

These spreads are nothing more than weighted averages of log prices of cryptocurrencies, where the weighting is done by the cointegration vectors. If the spread exceeds an upper threshold, we enter a short position, if the spread goes below the lower threshold, we enter a long position. The reasoning behind the strategy is very intuitive. A large positive spread is a signal that the portfolio is overpriced and it is profitable to sell it. On the other hand, if we encounter a large negative spread, the portfolio is underpriced

and we should buy it. We choose three different threshold levels, $\tau \in (\pm\sigma_j, \pm 1.5\sigma_j, \pm 2\sigma_j)$, which are chosen to be symmetric around the long term mean of the stochastic trend and σ_j is the estimated standard deviation. This investment decision is repeated for each estimated cointegration relationship and for each trading day. So each day, we have to make a decision to either buy, sell or hold our positions. The trading strategy follows Leung and Nguyen (2019), who consider a similar statistical arbitrage strategy. However, our strategy differs in two aspects. First, Leung and Nguyen (2018) use the approach of Engle and Granger (1987) to estimate the cointegration vector and second, our paper utilizes r cointegration relations while their paper is restricted to a single one. We backtest our strategy and compare the performance to the cryptocurrency index CRIX (Trimborn and Härdle (2018)).

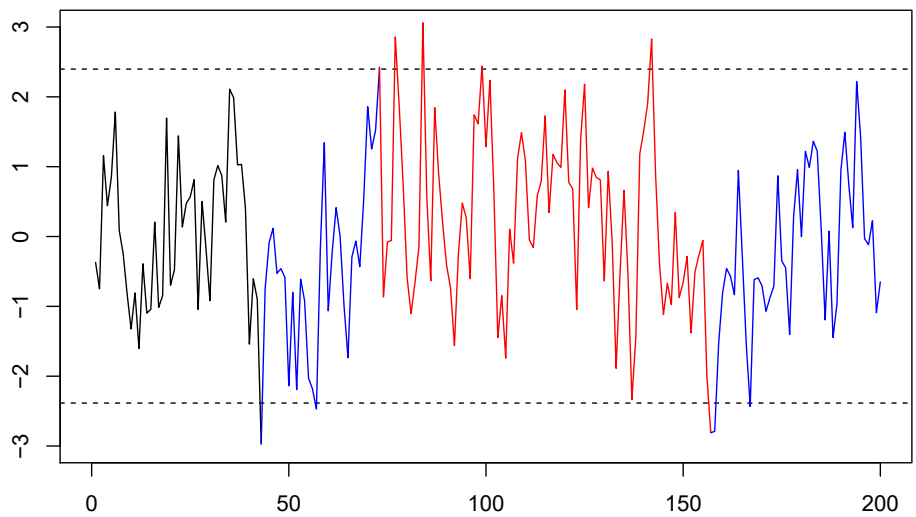


Figure 5.7: Visualization of the statistical arbitrage trading strategy for simulated data. Neutral position, short position and long position.

Threshold	$\pm\sigma_j$	$\pm 1.5\sigma_j$	$\pm 2\sigma_j$	CRIX
Number of Trades	40	21	12	-
Net Profits	28,474	24,876	21,247	15,330
Maximal Drawdown	3,078	2,989	2,893	55,297
Annual Sharpe Ratio	2.24	1.99	1.77	0.22

Table 5.9: In-sample performance statistics for different threshold levels.

CryptoDynamics_Trading

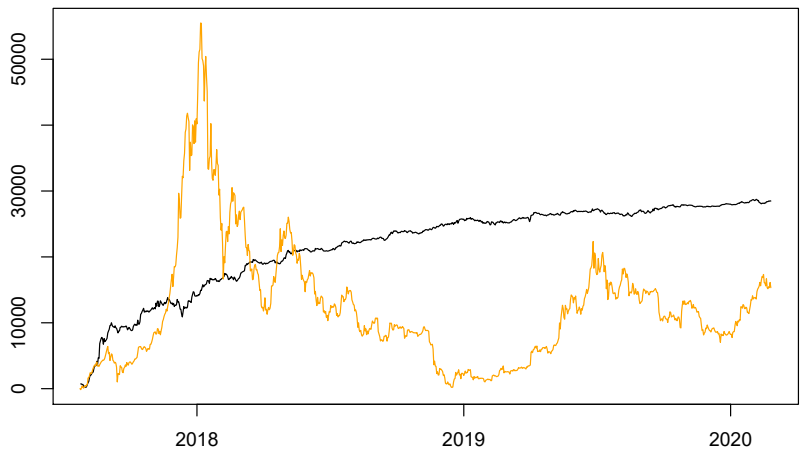


Figure 5.8: In-sample performance of the trading strategy with thresholds $\tau = \pm 1.5\sigma$ (black) vs. CRIX (yellow).

Table 5.9 summarizes the performance of our trading strategy for different threshold levels and compares it to the performance of the CRIX. The number of trades is decreasing with an increasing threshold level. For each of the candidate thresholds, we can make substantial profits. The optimal threshold in our analysis is $\tau = \pm\sigma_j$. It has the highest net profits, the largest Sharpe ratio and a similar maximal drawdown to the other threshold levels. While the net profits of the benchmark index portfolio (CRIX) are comparable to our arbitrage strategy, the risk is significantly higher. The maximal drawdown is more than ten times as large as for the optimal strategy. Also the Sharpe ratio, which relates expected returns to the standard deviation, is clearly smaller. Figure 5.8 visualizes the time series of the cumulative returns of our trading strategy and of the CRIX. As expected of an arbitrage strategy, there is almost no dependence of the cumulative returns to the market. An interesting observation is that the only substantial losses are made during the height of the crypto bubble at the end of 2017. The gains and losses are very volatile in this period. From the middle of 2018 until the beginning of 2020 we can observe small but steady profits.

While the backtesting results show a great performance of our trading strategy, a word of caution is needed. First, backtesting is an in-sample evaluation with limited external validity. There is no guarantee that long-term relationship will hold in the future, which is an implicit assumption in our cointegration analysis. This problem is particularly severe in the case of cryptocurrencies due to their very short history. Another caveat is that we assume perfect markets. In reality, investors face short selling restrictions and transaction costs, even if some exchanges as Bitfinex allow for short selling.

To further evaluate our trading strategy, we take a look at its out-of-sample performance. We consider the same set of cryptocurrencies in a time period from February 26 to

October 13, 2020. The results are reported in Table 5.10. It becomes evident that for none of the threshold levels we can make a positive profit. The reason for this is the divergence of the cointegration relation in the out-of-sample period, as shown in Figure 5.9. To analyze the trading performance in more detail, we consider each cointegration relation separately in Figure 5.10 for a threshold of $\pm 1.5\sigma$. Most of the losses originate from the first long-run relationship, which begins to deviate from its mean at the end of August. A similar phenomenon can be observed for the fourth cointegration relation. The remaining two stochastic trends do not diverge significantly, leading to a close-to-zero profit. It will be interesting to observe in the future whether our estimated cointegration relations are indeed mean-reversing, i.e. whether they will return to their equilibria, as predicted by our model. This would also provide more information on the profitability of our trading strategy. While the out-of-sample analysis can be seen as evidence against the possibility of statistical arbitrage, it is too soon to tell whether the long-run relations disappeared completely.

Threshold	$\pm\sigma_j$	$\pm 1.5\sigma_j$	$\pm 2\sigma_j$	CRIX
Number of Trades	5	4	3	-
Net Profits	-2,134	-2,363	-1,660	10,692
Maximal Drawdown	675	571	543	10,186
Annual Sharpe Ratio	-1.56	-2.20	-1.73	0.87

Table 5.10: Out-of-sample performance statistics for different threshold levels.
CryptoDynamics_Trading

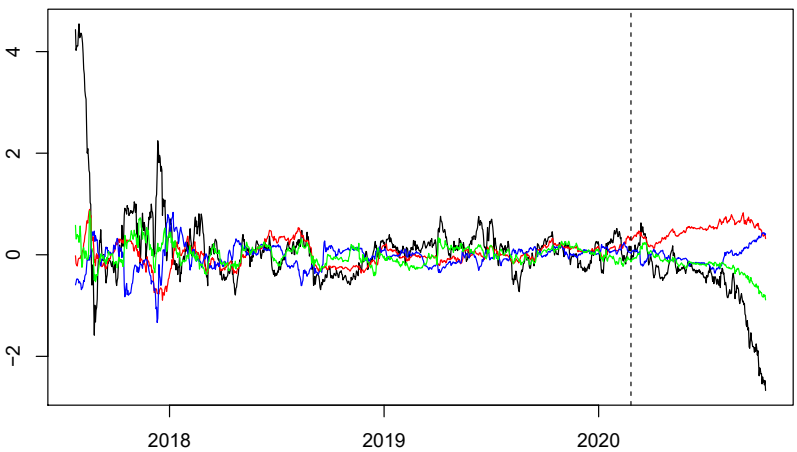


Figure 5.9: Time series of long-run stochastic trends. Dashed vertical line indicates the begin of the out-of-sample period.

5.6 Conclusion

This paper examined the joint behavior of cryptocurrencies in a non-stationary setting. We were in particular interested in three questions.

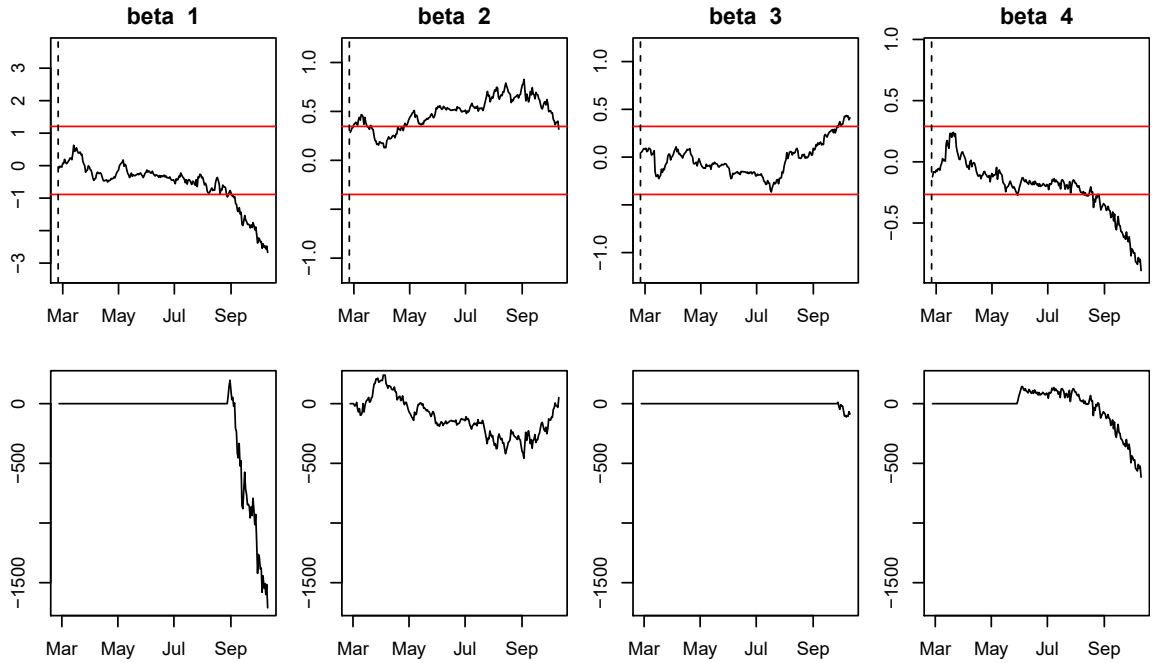


Figure 5.10: Out-of-sample analysis of long run equilibrium relationships and profits from corresponding trading strategies. The red horizontal lines visualize the thresholds $\tau = \pm 1.5\sigma$.

I. Do cointegration relations exist among cryptocurrencies?

II. Which cryptocurrencies affect and which are affected by long-term equilibrium effects?

III. How does the impact of the cointegration relationships change in a dynamic setting?

To address problem I. and II., we tested for cointegration using the approach of Onatski and Wang (2018) and estimated a linear VECM. We found that our sample of currencies are indeed cointegrated with rank four. By testing for weak exogeneity, we were able to show that all cryptocurrencies are significantly affected by long term stochastic trends. To address problem III., we proposed a new nonlinear VECM specification, which we call *COINtensity* VECM. The model has a good interpretability without the need of having to estimate many new parameters. The results of our dynamic VECM show a time-varying dependence of cryptocurrencies on deviations from long run equilibria. We find that the nonlinearity of error correction is stronger during the time of the cryptocurrency bubble, compared to a later time period.

Finally, we utilized the estimated cointegration relationships to construct a simple statistical arbitrage trading strategy, extending the one proposed in Leung and Nguyen (2019). Our strategy shows a great in-sample performance, beating the industry benchmark CRIX in terms of net profits, Sharpe ratio and maximal drawdown. A look

at the out-of-sample performance takes a more cautious perspective. In particular, the trading strategy can only be successful if the cointegration equilibrium relations hold in the long-run.

5.A Appendix: Simulation Design

Setting 1.1: Baseline VECM specification:

$$\Delta X_t = \alpha\beta^\top X_{t-1} + \Gamma_1 \Delta X_{t-1} + \varepsilon_t,$$

with parameter matrices

$$\alpha = \begin{pmatrix} -1.47 & -1.3 & 0 & -1.26 \\ 0 & 0.97 & 0 & 0 \\ 0 & 0 & -0.74 & 0 \\ -1.19 & 0.85 & 0 & 0 \\ 0.55 & 0.78 & -1 & -1.37 \\ 0.8 & 0.75 & 0 & 0 \\ 0 & -0.74 & -1.26 & 0.78 \\ 0 & -1.4 & 0 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0.8 \\ 0 & 0 & -1.29 & 1.49 \\ -0.87 & 0 & -0.53 & -0.82 \\ 1.45 & 1.48 & 0.9 & -0.69 \end{pmatrix},$$

$$\Gamma_1 = \mathrm{diag}\{0, 0.7979, 0, 0.7932, 0, 0.5377, 0, 0.7227\}.$$

Setting 1.2: *COINtensity* VECM specification:

$$\Delta X_t = \alpha\beta^\top X_{t-1} \left\{1 + \tanh\left(\gamma^\top \beta^\top X_{t-1}\right)\right\} + \varepsilon_t.$$

$$\alpha = \begin{pmatrix} -0.10 & 0.47 & -0.12 \\ 0.10 & -0.20 & -0.12 \\ 0.67 & 0.20 & -0.30 \\ 0.37 & 0.34 & 0.30 \\ -0.71 & 0.49 & 0.28 \\ -0.62 & -0.45 & 0.10 \\ -1.00 & 0.03 & 0.32 \\ -0.82 & 0.72 & 0.88 \\ -0.77 & 0.50 & -0.70 \\ -0.08 & 0.76 & 0.24 \\ -0.01 & -0.82 & -0.05 \\ -0.62 & 0.58 & 0.68 \\ 0.27 & 0.91 & -0.28 \\ 0.95 & -0.44 & 0.03 \\ -0.58 & 0.91 & 0.11 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\gamma = (0.2, 0.3, -0.4)^{\top}.$$

Setting 2.2:

$$\alpha = \begin{pmatrix} -0.10 & 0.27 & 0.37 \\ 0.10 & -0.20 & -0.23 \\ -0.14 & 0.20 & -0.30 \\ -0.08 & 0.10 & 0.30 \\ 0.04 & 0.08 & 0.16 \\ 0.24 & -0.26 & -0.00 \\ -0.18 & -0.18 & 0.13 \\ 0.24 & -0.19 & 0.30 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\gamma = (0.2, 0.3, -0.4)^{\top}.$$

Bibliography

Corbet, S., Meegan, A., Larkin, C., Lucey, B., & Yarovaya, L. (2018). Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Economics Letters*, *165*, 28–34.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Diebold, F. X., & Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, *182*(1), 119–134.

Dijk, D. v., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models—a survey of recent developments. *Econometric reviews*, *21*(1), 1–47.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Fan, J., & Yao, Q. (2008). *Nonlinear time series: Nonparametric and parametric methods*. Springer Science & Business Media.

Granger, C., & Teräsvirta, T. (1993). Modelling nonlinear economic relationships oxford university press. *New York*.

Guo, L., Tao, Y., & Härdle, W. K. (2018). A dynamic network perspective on the latent group structure of cryptocurrencies. *arXiv preprint arXiv:1802.03708*.

Härdle, W. K., Harvey, C. R., & Reule, R. C. G. (2020). Understanding Cryptocurrencies*. *Journal of Financial Econometrics*, *18*(2), 181–208. https://doi.org/10.1093/jjfinec/nbz033

Ji, Q., Bouri, E., Lau, C. K. M., & Roubaud, D. (2019). Dynamic connectedness and integration in cryptocurrency markets. *International Review of Financial Analysis*, *63*, 257–272.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, *12*(2-3), 231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, 1551–1580.

Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press on Demand.

Johansen, S. (2002). A small sample correction for the test of cointegrating rank in the vector autoregressive model. *Econometrica*, *70*(5), 1929–1961.

Kristensen, D., & Rahbek, A. (2010). Likelihood-based inference for cointegration with nonlinear error-correction. *Journal of Econometrics*, *158*(1), 78–94.

Kristensen, D., & Rahbek, A. (2013). Testing and inference in nonlinear cointegrating vector error correction models. *Econometric Theory*, *29*(6), 1238–1288.

Kuan, C.-M., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric reviews*, *13*(1), 1–91.

Lee, T.-H., White, H., & Granger, C. W. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, *56*(3), 269–290.

Leung, T., & Nguyen, H. (2019). Constructing cointegrated cryptocurrency portfolios for statistical arbitrage. *Studies in Economics and Finance*.

Liang, C., & Schienle, M. (2019). Determination of vector error correction models in high dimensions. *Journal of econometrics*, *208*(2), 418–441.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Luukkonen, R., Saikkonen, P., & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, *75*(3), 491–499.

Ng, S., & Perron, P. (1995). Unit root tests in arma models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, *90*(429), 268–281.

Onatski, A., & Wang, C. (2018). Alternative asymptotics for cointegration tests in large vars. *Econometrica*, *86*(4), 1465–1478.

Shahzad, S. J. H., Bouri, E., Roubaud, D., Kristoufek, L., & Lucey, B. (2019). Is bitcoin a better safe-haven investment than gold and commodities? *International Review of Financial Analysis*, *63*, 322–330.

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. *Journal of Economics and Financial Analysis*, *2*(2), 1–27.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the american Statistical association*, *89*(425), 208–218.

Trimborn, S., & Härdle, W. K. (2018). Crix an index for cryptocurrencies. *Journal of Empirical Finance*, *49*, 107–122.

Yi, S., Xu, Z., & Wang, G.-J. (2018). Volatility connectedness in the cryptocurrency market: Is bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, *60*, 98–114.

Declaration

I hereby declare that I completed this work without any improper help from a third party and without using any aids other than those cited. All ideas derived directly or indirectly from other sources are identified as such. The results of Chapter 2 are based on joint work with Juan Manuel Rodriguez-Poo, Alexandra Soberon and Weining Wang. The results of Chapter 3 are based on joint work with Likai Chen and Wei Biao Wu. Chapter 4 is based on a published paper with Weining Wang, which appeared in Empirical Economics in 2021. Finally, Chapter 5 is based on a collaboration with Yanfen Zhang, which appeared in Digital Finance in 2021.

I testify through my signature that all information that I have provided about resources used in the writing of my doctoral thesis, about the resources and support provided to me as well as in earlier assessments of my doctoral thesis correspond in every aspect to the truth.

Berlin, den 12.07.2021 Georg Keilbar