UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

Eduardo de Oliveira Horta

Essays in Nonparametric Econometrics and Infinite Dimensional Mathematical Statistics

Ensaios em Econometria Não-Paramétrica e Estatística Matemática em Dimensão Infinita

Porto Alegre
2015

Eduardo de Oliveira Horta

Essays in Nonparametric Econometrics and Infinite Dimensional Mathematical Statistics

Ensaios em Econometria Não-Paramétrica e Estatística Matemática em Dimensão Infinita

Tese submetida ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como quesito parcial para obtenção do título de Doutor em Economia, com ênfase em Economia Aplicada.

Orientador: Prof. Dr. Flávio Augusto Ziegelmann

Porto Alegre
2015

CIP - Catalogação na Publicação

de Oliveira Horta, Eduardo
Essays in Nonparametric Econometrics and Infinite Dimensional Mathematical Statistics / Eduardo de Oliveira Horta. -- 2015.
96 f.

Orientador: Flávio Augusto Ziegelmann.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Programa de Pós-Graduação em Economia, Porto Alegre, BR-RS, 2015.

1. Quantile regression. 2. Random measure. 3. Functional time series. 4. Covariance operator. I. Augusto Ziegelmann, Flávio, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Eduardo de Oliveira Horta

Essays in Nonparametric Econometrics and Infinite Dimensional Mathematical Statistics

Ensaios em Econometria Não-Paramétrica e Estatística Matemática em Dimensão Infinita

Tese submetida ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como quesito parcial para obtenção do título de Doutor em Economia, com ênfase em Economia Aplicada.

Aprovada em: São Paulo, 2 de outubro de 2015

Banca Examinadora

Prof. Dr. Flávio Augusto Ziegelmann - Orientador
PPGE-UFRGS

Prof. Dr. Nikolai Valtchev Kolev
IME-USP

Prof. Dr. Pedro Valls Pereira
EESP-FGV

Prof. Dr. Vladimir Belitsky
IME-USP

À minha mãe.

AGRADECIMENTOS

Uma Tese de Doutorado é, evidentemente, o fim de uma trajetória. Não tão evidente é em que ponto essa trajetória se iniciou; provavelmente, muito antes do que a maioria de nós suporia. Por essa razão, estes agradecimentos são incompletos e insuficientes: se estamos onde estamos, é por causa das pessoas que fazem ou fizeram parte de nossas vidas. Aqueles que, de fato, me trouxeram até aqui e, de alguma forma, tornaram possível esta Tese formam uma lista tão extensa que seria impossível mencioná-los todos. Espero não incorrer em nenhuma injustiça irreparável com as omissões que, inevitavelmente, aqui cometo.

Esta Tese é dedicada à minha mãe, Maria Fernanda. Meu maior agradecimento é a ela, que me deu a vida, o amor e os ensinamentos que me tornaram a pessoa que sou hoje. Não há um só dia em que sua ausência não seja sentida. Agradeço à minha irmã, Denise, pelo companheirismo de uma vida inteira, e ao Cristiano, que entrou para essa família há tanto tempo que seria difícil imaginar-nos sem ele. Ao meu pai, Vulpius, a quem devo o apreço que tenho pelo conhecimento, pela sabedoria e pela ciência.

À minha esposa, Maritê, cujo amor e companheirismo me sustentam. Sem ela, minha vida perderia o sentido. À família da minha esposa, especialmente à minha sogra Maria Terezinha, mas sem esquecer-me do meu sogro Luiz Carlos e dos meus cunhados Maurício e Fabíola, e Luciano e Anna Helena, e meus sobrinhos Henrique e Maria Antônia, e também dos familiares de Uruguaiana pela amizade de tantos anos.

À minha dinda Maria Elizabeth e à sua família, por terem sido sempre um porto seguro para mim – evidentemente, um porto seguro repleto de diversão! Obrigado Luiz, Paulo Ricardo e Kelen, Fernanda e Alex, Daniela e Saimon, Gabriel e pequenino Matheus.

Ao meu orientador, Flávio, por tudo que me ensinou. Sua inteligência tornou possível esta Tese. Para mim foi uma honra ter trabalhado com ele, e me alegra profundamente tê-lo, acima de tudo, como amigo. A Emmanuel Guerre, cuja sabedoria e conhecimento surpreendentes tanto admiro, e à sua esposa Stepana pelas inúmeras conversas, sempre caracterizadas por uma inteligência elegantíssima.

Aos meus colegas do IME–UFRGS, em especial aos colegas do DEST, que me acolheram no Departamento e tornaram-se grandes amigos. Eu não poderia deixar de agradecer muito especialmente à Gabriela, ao Guilherme, ao Hudson, à Marcia, e aos meus colegas de sala Jean e Débora. Agradeço profundamente à Adriana, ao Carlos Felipe, ao Eduardo Henrique, ao Luiz Fernando e ao Rafael, de quem tive a oportunidade de ser, por um tempo, também aluno.

Aos amigos do Campus do Vale com quem tive a honra de poder conviver: Carolina Graciolli, Fábio Casula, Matheus Bohrer, Otávio Menezes, Rangel Baldasso, Nanda Duarte, Pietro Duarte e, muito especialmente, Eduardo Longa e Luísa Borsato. A amizade diária de cada um foi imprescindível para tornar esses anos de Doutorado muito mais leves.

A todos os colegas, amigos e professores da Queen Mary University of London que, em meio a um inverno tão frio, tornaram Londres uma cidade calorosa; especialmente, à minha grande amiga Nathalie. Aos meus colegas de Mestrado Anna, Bruno, Márcio, Raphael e Rodrigo. Aos tantos amigos que ganhei ao longo da vida: Adriano e Flávia, Círtia, Clarissa e Guilherme Campana, Cristina e Diego, Daniela Ribeiro, Guilherme Rimoli, Tatiana e Gary, Bruno e Karen, Joana e Michael, Paulo e Indiara, Greta, Marcelo e Mari. Cada um, ao seu modo, contribuiu para que esta Tese se concretizasse. Agradeço também a todos os servidores desta Universidade, que com seu trabalho possibilitam que a vida acadêmica se concretize. Em especial, às secretárias Aninha, Fátima e Iara.

Àqueles que são minha segunda família, meus irmãos e irmãs, alguns desde há muito tempo, outros mais recentes (e aqui me permito chamá-los pelos apelidos): Leo e Mari, Cruz e Bruna, Bene e Babi, Dadá e Cris, Diegão e Júlia, Eduardinho e Jana, Dudu e Mo, Elisa e Antonio, Balvedi, Cofa e Lu, Pinty e Cássinha, Trentin, Gabriel e Nati, Gui e Lívia, Gustavinho e Fer, Jean e Pati, Matheus, Mauri e Ju, Pablão e Michele, Tio Paulo, Rica, Rodrigão e Maria, Ditongo e Lu, Gomes e Luana, Tiaguinho e Crisleni, Vitório e Kelly, e Tommaso.

E, por fim, à minha cachorrinha Chica Bela, que tantas vezes me chamou a atenção para o fato de que, sim, o mundo exterior continuava lá, ao me convidar para jogar bolinha em meio às intermináveis tardes de estudo.

“My fiftieth year had come and gone,
I sat, a solitary man,
In a crowded London shop,
An open book and empty cup
On the marble table-top.
While on the shop and street I gazed
My body of a sudden blazed;
And twenty minutes more or less
It seemed, so great my happiness,
That I was blessed and could bless.”
(William Butler Yeats, Vacillation)

RESUMO

A presente Tese de Doutorado é composta de quatro artigos científicos em duas áreas distintas. Em Horta, Guerre e Fernandes (2015), o qual constitui o Capítulo 2 desta Tese, é proposto um estimador suavizado no contexto de modelos de regressão quantílica linear (Koenker e Basset, 1978). Uma representação de Bahadur-Kiefer uniforme é obtida, a qual apresenta uma ordem assintótica que domina aquela correspondente ao estimador clássico. Em seguida, prova-se que o viés associado à suavização é negligenciável, no sentido de que o termo de viés é equivalente, em primeira ordem, ao verdadeiro parâmetro. A taxa precisa de convergência é dada, a qual pode ser controlada uniformemente pela escolha do parâmetro de suavização. Em seguida, são estudadas propriedades de segunda ordem do estimador proposto, em termos do seu erro quadrático médio assintótico, e mostra-se que o estimador suavizado apresenta uma melhoria em relação ao usual. Como corolário, tem-se que o estimador é assintoticamente normal e consistente à ordem $\sqrt{n}$. Em seguida, é proposto um estimador consistente para a matriz de covariância assintótica, o qual não depende de estimação de parâmetros auxiliares e a partir do qual pode-se obter diretamente intervalos de confiança assintóticos. A qualidade do método proposto é por fim ilustrada em um estudo de simulação. Os artigos Horta e Ziegelmann (2015a, 2015b, 2015c) se originam de um ímpeto inicial destinado a generalizar os resultados de Bathia et al. (2010). Em Horta e Ziegelmann (2015a), Capítulo 3 da presente Tese, é investigada a questão de existência de certos processos estocásticos, ditos processos conjugados, os quais são conduzidos por um segundo processo cujo espaço de estados tem como elementos medidas de probabilidade. Através dos conceitos de coerência e compatibilidade, obtém-se uma resposta afirmativa à questão anterior. Baseado nas noções de medida aleatória (Kallenberg, 1973) e desintegração (Chang e Pollard, 1997; Pollard, 2002), é proposto um método geral para construção de processos conjugados. A teoria permite um rico conjunto de exemplos, e inclui uma classe de modelos de mudança de regime. Em Horta e Ziegelmann (2015b), Capítulo 4 desta Tese, é proposto – em relação com a construção obtida em Horta e Ziegelmann (2015a) – o conceito de processo fracamente conjugado: um processo estocástico real a tempo contínuo, conduzido por uma sequência de funções de distribuição aleatórias, ambos conectados por uma condição de compatibilidade a qual impõe que aspectos da distribuição do primeiro processo são divisíveis em uma quantidade enumerável de ciclos, dentro dos quais este tem como marginais, precisamente, o segundo processo. Em seguida, mostra-se que a metodologia de Bathia et al. (2010) pode ser aplicada para se estudar a estrutura de dependência de processos fracamente conjugados, e com isso obtém-se resultados de consistência à ordem $\sqrt{n}$ para os estimadores que surgem naturalmente na teoria. Adicionalmente, a metodologia é ilustrada através de uma implementação a dados financeiros. Especificamente, o método proposto permite que características da dinâmica das distribuições de processos de retornos sejam traduzidas em termos de um processo escalar latente, a partir do qual podem ser obtidas previsões de quantidades associadas a essas distribuições. Em Horta e Ziegelmann (2015c), Capítulo 5 da presente Tese, são obtidos resultados de consistência à ordem $\sqrt{n}$ em relação à estimação de representações espectrais de operadores de autocovariância de séries de tempo Hilbertianas estacionárias, em um contexto de medições imperfeitas. Os resultados são uma generalização do método desenvolvido em Bathia et al. (2010), e baseiam-se no importante fato de que elementos aleatórios em um espaço de Hilbert separável são quase certamente ortogonais ao núcleo de seu respectivo operador de covariância. É dada uma prova direta deste fato.

Palavras-chave. Regressão Quantílica. Medidas aleatórias. Séries temporais funcionais. Operador de covariância.

Classificação JEL. C1, C14, C22

ABSTRACT

The present Thesis is composed of 4 research papers in two distinct areas. In Horta, Guerre, and Fernandes (2015), which constitutes Chapter 2 of this Thesis, we propose a smoothed estimator in the framework of the linear quantile regression model of Koenker and Bassett (1978). A uniform Bahadur-Kiefer representation is provided, with an asymptotic rate which dominates the standard quantile regression estimator. Next, we prove that the bias introduced by smoothing is negligible in the sense that the bias term is first-order equivalent to the true parameter. A precise rate of convergence, which is controlled uniformly by choice of bandwidth, is provided. We then study second-order properties of the smoothed estimator, in terms of its asymptotic mean squared error, and show that it improves on the usual estimator when an optimal bandwidth is used. As corollaries to the above, one obtains that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal. Next, we provide a consistent estimator of the asymptotic covariance matrix which does not depend on ancillary estimation of nuisance parameters, and from which asymptotic confidence intervals are straightforwardly computable. The quality of the method is then illustrated through a simulation study. The research papers Horta and Ziegelmann (2015a;b;c) are all related in the sense that they stem from an initial impetus of generalizing the results in Bathia et al. (2010). In Horta and Ziegelmann (2015a), Chapter 3 of this Thesis, we address the question of existence of certain stochastic processes, which we call conjugate processes, driven by a second, measure-valued stochastic process. We investigate primitive conditions ensuring existence and, through the concepts of coherence and compatibility, obtain an affirmative answer to the former question. Relying on the notions of random measure (Kallenberg (1973)) and disintegration (Chang and Pollard (1997), Pollard (2002)), we provide a general approach for construction of conjugate processes. The theory allows for a rich set of examples, and includes a class of Regime Switching models. In Horta and Ziegelmann (2015b), Chapter 4 of the present Thesis, we introduce, in relation with the construction in Horta and Ziegelmann (2015a), the concept of a weakly conjugate process: a continuous time, real valued stochastic process driven by a sequence of random distribution functions, the connection between the two being given by a compatibility condition which says that distributional aspects of the former process are divisible into countably many cycles during which it has precisely the latter as marginal distributions. We then show that the methodology of Bathia et al. (2010) can be applied to study the dependence structure of weakly conjugate processes, and therewith provide $\sqrt{n}$-consistency results for the natural estimators appearing in the theory. Additionally, we illustrate the methodology through an implementation to financial data. Specifically, our method permits us to translate the dynamic character of the distribution of an asset returns process into the dynamics of a latent scalar process, which in turn allows us to generate forecasts of quantities associated to distributional aspects of the returns process. In Horta and Ziegelmann (2015c), Chapter 5 of this Thesis, we obtain $\sqrt{n}$-consistency results regarding estimation of the spectral representation of the zero-lag autocovariance operator of stationary Hilbertian time series, in a setting with imperfect measurements. This is a generalization of the method developed in Bathia et al. (2010). The generalization relies on the important property that centered random elements of strong second order in a separable Hilbert space lie almost surely in the closed linear span of the associated covariance operator. We provide a straightforward proof to this fact.

Keywords. Quantile regression. Random measure. Functional time series. Covariance operator.

JEL Classification. C1, C14, C22

CONTENTS

1 INTRODUCTION

The present Thesis is composed of four research papers in two distinct areas. The first of these papers was written in collaboration with Professor Emmanuel Guerre and Professor Marcelo Fernandes and dwells upon the topic of quantile regression models. The remaining three papers were written in collaboration with Professor Flávio Ziegelmann and stem from an effort to generalize the results in Bathia et al. (2010).

Quantile regression has emerged in its modern formulation through the seminal paper by Koenker and Bassett (1978), and has since become both an object of theoretical interest and an important tool in applications. A complete account can be found in Koenker (2005). In recent years, quantile regression models have come to enjoy widespread application in many areas of research. See for instance Koenker (2000), Buchinsky (1998), Koenker and Hallock (2001), Koenker (2005) and references therein. Despite the undoubted generality of the linear conditional quantile model of Koenker and Bassett (1978), which led it to reach the aforementioned success, when it comes to inference there are a few drawbacks accompanying the standard approach. The standard quantile regression estimator minimizes an empirical counterpart to the population objective function, of which the true parameter is a minimizer. However, smoothness properties of the population objective function are not inherited by its sample analogue, and this lack of smoothness has implications on inferential procedures about the estimated parameter. As reviewed for instance in Koenker (1994), Buchinsky (1995), Koenker (2005), Fan and Liu (2013), Goh and Knight (2009), computation of asymptotic confidence intervals for components of the standard quantile regression estimator is not straightforward. This stems from the fact that there seems to be no canonical way of estimating the covariance matrix of the estimator, which in turn is a consequence of non-differentiability of the standard objective function. This issue has been widely investigated in the literature. See Koenker (2005) and Buchinsky (1995) for a review. In recent work, a wide variety of techniques has been proposed to tackle inferential aspects of quantile regression (Horowitz (1998), Machado and Parente (2005), Chernozhukov and Hong (2003), Otsu (2008), Whang (2006), Portnoy (2012), Goh and Knight (2009), Mammen et al. (2013), Fan and Liu (2013), to name a few), whereas a related literature investigates asymptotic distributional properties of the standard quantile regression estimator through Bahadur-Kiefer type representations (Koenker and Portnoy (1987), Chaudhuri et al. (1991), He and Shao (1996), Knight (2001), Guerre and Sabbah (2012), Portnoy (2012), Kong et al. (2013), Mammen et al. (2013)). For the standard quantile regression estimator, however, the remainder term in such expansions has poor rate, attaining at best the order $n^{-1/4}$ in many cases of interest like the iid scenario (see Koenker and Portnoy (1987), Knight (2001), Jurečková et al. (2012)).

In Horta, Guerre, and Fernandes (2015), Chapter 2 of this Thesis, we propose a convolution-type smoothing of the sample objective function, an approach which in the one-sample scenario corresponds to Nadaraya (1964). In the context of parametric and semi-parametric quantile regression models, kernel-type methods have mostly accompanied the literature tackling inferential matters, but surprisingly little attention has been given so far to propose estimators based on simple smoothing techniques as to generalize the quantile estimators of Nadaraya (1964) or Parzen (1979). Important exceptions are the smoothed least absolute deviations estimator of Horowitz (1998) and the smoothed estimating equations test of Kaplan and Sun (2012). The proposed estimator thus fills a gap in the semi-parametric quantile regression literature.

The main contributions to be found in Horta, Guerre, and Fernandes (2015) are as follows. It is first shown that the proposed smoothed estimator is 'more linear' than the standard quantile regression estimator: the stochastic order of the remainder term in its Bahadur-Kiefer representation is at least $n^{-1/2}$. Next we prove that the bias introduced by smoothing is negligible in the sense that the bias term is first-order equivalent to the true parameter. A precise rate of convergence, which is controlled uniformly by choice of bandwidth, is provided. We then study second-order properties of the smoothed estimator, in terms of its asymptotic mean squared error, and show that it improves on the usual estimator when an optimal bandwidth is used. As corollaries to the above, one obtains that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal. Next, we provide a consistent estimator of the asymptotic covariance matrix which does not depend on ancillary estimation of nuisance parameters, and from which asymptotic confidence intervals are straightforwardly computable. An aspect worth stressing is that our asymptotic results hold *uniformly both in the quantile level and in the bandwidth parameter*. Finally, we assess the quality of our method through a simulation study.

The research papers Horta and Ziegelmann (2015a;b;c), Chapters 3, 4 and 5 of this Thesis, respectively, are all related in the sense that they stem from an initial impetus of generalizing the results in Bathia et al. (2010). Following Hall and Vial (2006), who tackle an identification problem in noisy functional Principal Component Analysis, Bathia et al. (2010) propose a solution in the framework of functional time series which allows one to recover the underlying dynamic structure of the data, via a Law of Large Numbers for the estimator of an operator equivalent to the zero-lag covariance operator of the random curves. The methods that we shall consider in Chapters 3, 4 and 5 are thus intrinsically *functional*, in that we consider random elements in spaces of functions. Our theory lies somewhere in-between Functional Data Analysis and Probability in Banach spaces, which are two very important research fields in the statistics and probability literature respectively. Statistical inference on objects pertaining to function spaces has come to be known in the literature as Functional Data Analysis (FDA). In recent years, FDA has received growing attention from researchers of a wide spectrum of academic disciplines (see Dabo-Niang and Ferraty (2008) and the cornerstone monograph by Ramsay and Silverman (1998)). From a theoretical point of view, functional data are to be seen as realizations of function-valued random variables. The general approach is to consider random elements in a Banach space; classic texts include Ledoux and Talagrand (1991) and Vakhania et al. (1987). For stationary sequences and linear processes in Banach spaces, the monograph from Bosq (2000) is a complete account.

Initially the generalization of Bathia et al. (2010) was sought motivated by an application, namely to model time series of density functions – bivariate densities to be precise. At some point this objective shifted towards an inquiry of how exactly to interpret a stochastic process whose state space is a set of density functions. Concurrently, a remark in Bathia et al. (2010) reminded us that modeling kernel density estimators as 'true, random density' plus 'noise' is potentially misleading since in this setting the noise is not a centered random element (of whatever space it lies in – kernel density estimators are pointwise biased!). Instead, the model 'empirical distribution function' = 'true (random) distribution function' + 'noise' appeared to better capture the properties we had in mind. The adequacy of the latter approach is evinced in Lemma 4.1. Our effort to embed such model in a good theoretical framework eventually led to ramifications which culminated in the papers Horta and Ziegelmann (2015a;b). The key insight seems to have been equation (4.1), which we reproduce here:

$$\mathbb{P}[X_\tau \leq x \,|\, F_0, F_1, \dots] = F_t(x), \qquad \tau \in [t, t+1), \qquad t = 0, 1, \dots. \tag{4.1}$$

The early intuition which eventually led to the above condition appeared, as mentioned, when we were studying how to model the dynamics of distribution functions, more precisely the distribution of high frequency asset returns in financial data. Our original indagation can be posed as follows: assuming asset returns share the same marginal distribution inside each day, but allowing these marginals to vary from day to day (possibly in a stochastic manner), then how to give a reasonable formulation, in terms of stochastic processes, of these ideas? The answer 'is' the model of weakly conjugate processes: a continuous time, real valued stochastic process $(X_\tau : \tau \geq 0)$ driven by a sequence of distribution functions (which are random), the connection between them being equation (4.1), a condition that can be understood as saying that distributional aspects of (X_τ) are divisible, at least conditionally, into countably many cycles during which (X_τ) has some prescribed (random) marginal distribution.

This led to a fruitful theory which is the content of Horta and Ziegelmann (2015b), Chapter 4 of the present Thesis. There we show that processes satisfying equation (4.1) – which we call *weakly conjugate processes* – fall smoothly, via Lemma 4.1, into the methodology of Bathia et al. (2010), allowing us to derive $\sqrt{n}$-consistency results for the natural estimators that appear in the construction. This is the content of Proposition 4.2

and Theorems 4.1 and 4.2, the main contributions of the paper. Additionally, we illustrate the methodology through an implementation to financial data. Specifically, our method permits us to translate the dynamic character of the distribution of an asset returns process into the dynamics of a latent scalar process, which in turn allows us to generate forecasts of quantities associated to distributional aspects of the returns process.

In parallel, the probabilistic question of whether there exist, given some 'primitive' conditions, processes satisfying (4.1) appeared quite interesting to us. We therefore headed towards 'parsing' equation (4.1) into sufficient conditions, while keeping in mind the goal of being as general as possible. This is addressed in Horta and Ziegelmann (2015a), Chapter 3 of this Thesis. The natural approach turned out to depend on the concept of random measure (Kallenberg (1973; 1974)) together with the machinery of disintegration of measures (Pachl (1978), Faden (1985), Chang and Pollard (1997), Pollard (2002)), and eventually led to the notions of $\mathscr{L}$*-coherence* and *compatibility* which permit the elegant statement of Theorem 3.1 – an existence Theorem and the main contribution in Horta and Ziegelmann (2015a), although a more application-inclined reader will certainly appreciate the rich set of examples which come almost effortlessly side by side with the concept of conjugate process.

Last but not least, Horta and Ziegelmann (2015c), Chapter 5 of the present Thesis, is the byproduct of our initial effort of obtaining a generalization of the methodology of Bathia et al. (2010). In the beginning the idea was to translate their results to the general Banach space, but that sparkle was short-lived as it turns out the asymptotic theory relies strongly on the fact that the space of Hilbert-Schmidt operators is itself a Hilbert space – the lack of a natural concept of Hilbert-Schmidt operator in the general Banach space spoiled our intent. A short attempt was made then to relate their work to Reproducing Kernel Hilbert spaces but we later learned that this is restrictive: a stochastic process whose index set is infinite never has its sample paths lying in the corresponding RKHS (at least for Gaussian processes – see Driscoll (1973)). Nevertheless, it seemed to us that the L^2 setting was too restrictive since all the heuristics rely strongly on the Karhunen-Loève Theorem. The point that we make is that considerations of 'sample-path properties', which lie at the core of said Theorem, are dispensable. The correct way to interpret a generalization is given by the property that centered random elements of strong second order in a separable Hilbert space lie almost surely in the closed linear span of the associated covariance operator: this is Theorem 5.1, certainly not a new result (it appears for instance as an exercise in Vakhania et al. (1987) in a slightly different guise), but a rather overlooked one. In any case the proof that we give is, to our knowledge, new. Equipped with the latter result, we provide a reformulation of the theory of Bathia et al. (2010) in a Hilbert space setting, culminating in Theorem 5.2 and Corollary 5.2, which state $\sqrt{n}$-consistency of the proposed estimators and are the main contributions of the paper.

2 SMOOTHING QUANTILE REGRESSION

EDUARDO HORTA[1] EMMANUEL GUERRE[2]
MARCELO FERNANDES[3]

November, 2015

Abstract. We propose a smoothed estimator in the framework of the linear quantile regression model of Koenker and Bassett (1978). A uniform Bahadur-Kiefer representation is provided, with an asymptotic rate which dominates the standard quantile regression estimator. Second order improvements are obtained, generalizing a result of Azzalini (1981). In our setting, inference can be implemented in a canonical way. In particular, estimation of the asymptotic covariance matrix is intrinsic to the method. A simulation study illustrates the quality of the proposed estimator.

Keywords and phrases. Asymptotic expansion. Smoothing. Quantile regression.

JEL Classification. C1, C14

2.1 Introduction

Quantile regression has emerged in its modern formulation through the seminal paper by Koenker and Bassett (1978), and has since become both an object of theoretical interest and an important tool in applications (see Koenker (2005) and the discussions below for theoretical aspects; for applications refer to Koenker (2000), Buchinsky (1998), Koenker and Hallock (2001), Koenker (2005) and references therein). The standard quantile regression estimator minimizes an empirical counterpart to the population objective function, of which the true parameter is a minimizer. It is somewhat unfortunate that smoothness properties of the population objective function are not inherited by its sample analogue, and this lack of smoothness in turn has implications on inferential procedures about the estimated parameter. As reviewed for instance in Koenker (1994), Buchinsky (1995), Koenker (2005), Fan and Liu (2013), Goh and Knight (2009), computation of asymptotic confidence intervals for components of the standard quantile regression estimator is not straightforward. This stems from the fact that there is, to our knowledge, no canonical way of estimating the covariance matrix, which in turn is a consequence of non-differentiability of the standard objective function. Indeed the asymptotic covariance depends on the population conditional density evaluated at the true quantile. This issue has been widely investigated in the literature (see Koenker (2005) and Buchinsky (1995)

[1]Department of Statistics – Universidade Federal do Rio Grande do Sul. eduardo.horta@ufrgs.br
[2]School of Economics and Finance – Queen Mary, University of London. e.guerre@qmul.ac.uk
[3]São Paulo School of Economics – FGV and Queen Mary, University of London. marcelo.fernandes@fgv.br

for a survey on traditional approaches targeted at it) and is by no means settled. In recent work, a wide variety of techniques has been proposed to tackle inferential aspects of quantile regression, including bootstrap techniques (Horowitz (1998), Machado and Parente (2005)), MCMC methods (Chernozhukov and Hong (2003)), empirical likelihood (Otsu (2008), Whang (2006)), strong approximation methods (Portnoy (2012)), nonstandard inference (Goh and Knight (2009)), as well as nonparametric approaches (Mammen et al. (2013), Fan and Liu (2013)), to name a few. A related literature investigates asymptotic distributional properties of the standard quantile regression estimator through Bahadur-Kiefer type representations (Koenker and Portnoy (1987), Chaudhuri et al. (1991), He and Shao (1996), Knight (2001), Guerre and Sabbah (2012), Portnoy (2012), Kong et al. (2013), Mammen et al. (2013)). However, the remainder term in such expansions has poor rate of convergence since the estimator is altogether highly nonlinear. The precise order can be difficult to establish but in the iid error setting for instance it can be shown that the remainder is at best of order $n^{-1/4}$ (see Koenker and Portnoy (1987), Knight (2001), Jurečková et al. (2012)).

An aspect that permeates some of the aforementioned literature is consideration of smoothing techniques. The approach that we shall take here sails in this direction. We propose a convolution-type smoothing of the sample objective function which can be regarded as a generalization of the approach taken by Nadaraya (1964) in the one-sample scenario. In the latter setting, consideration of smoothing methods has made its way into the theoretical literature quite early. The most well-known smoothed quantile estimators were introduced by Nadaraya (1964) and Parzen (1979), and are obtained respectively by inverting a smoothed empirical cumulative distribution function (hereafter cdf) an by smoothing the sample quantile function. Azzalini (1981) has shown that the smoothing proposed by Nadaraya results in a quantile estimator which dominates the sample quantile at second order (see also Cheung and Lee (2010)), with a similar result proved by Sheather and Marron (1990) for Parzen's estimator. Falk (1984) discusses the relative deficiency of the sample quantiles with respect to kernel-type quantile estimators. See Kozek (2005) for a bibliography and a convincing simulation experiment. In parallel an important literature has established Bahadur-Kiefer type representations for smoothed quantile estimators. See Xiang (1994), Mack (1987) and Ralescu (1997) for instance.

Kernel-type approaches are quite common in the literature dealing with nonparametric estimation of conditional quantile functions (Mammen et al. (2013), Mehra et al. (1991), Samanta (1989), Stute (1986)). In the context of parametric and semi-parametric quantile regression models, smoothing methods have accompanied the literature tackling inferential matters, but surprisingly little attention has been given so far to propose estimators based on simple smoothing techniques as to generalize the quantile estimators of Nadaraya (1964) or Parzen (1979). Important exceptions are the smoothed least absolute deviations estimator of Horowitz (1998) and the smoothed estimating equations test of Kaplan and

Sun (2012). The approach put forth by Horowitz (1998) is to estimate the regression quantiles by minimizing a smoothed analogue of the median regression sample objective function, but a simple derivation will show that in the one-sample case this approach does not correspond to either Nadaraya (1964) or Parzen (1979). Therefore there is not too much hope to replicate the second-order improvements obtained by Azzalini (1981), and indeed it can be shown that the smoothing adopted by Horowitz (1998) cannot improve on the standard quantile regression estimator of the slope parameter. In contrast, the smoothing herein proposed is equivalent to the one adopted by Nadaraya (1964) in the absence of a covariate. Kaplan and Sun (2012) in turn propose a smoothed estimating equation test for significance or equality of the quantile regression slope coefficients in the instrumental variable setup, with a smoothing which is equivalent to ours.

The main contributions of the paper are as follows. We first propose a new smoothed version of the quantile regression estimator. It is shown that the proposed smoothed estimator is 'more linear' than standard quantile regression estimator, in the sense that the stochastic order of the remainder term in its Bahadur-Kiefer representation is at least $n^{-1/2}$. Next we prove that the bias introduced by smoothing is negligible in the sense that the bias term is first-order equivalent to the true parameter. A precise rate of convergence, which is controlled uniformly by choice of bandwidth, is provided. We then study second-order properties of the smoothed estimator, in terms of its asymptotic mean squared error, and show that it improves on the usual estimator when an optimal bandwidth is used. This generalizes a result proved by Azzalini (1981) in the one-sample set-up. As corollaries to the above one obtains that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal. Next, we provide a consistent estimator of the asymptotic covariance matrix which does not depend on ancillary estimation of nuisance parameters, and from which asymptotic confidence intervals are straightforwardly computable. Thus our method falls into a more palatable framework (as summarized in Newey and McFadden (1994) for instance), and Wald-type inference is easily implementable. It is worth stressing here that *our results are uniform both in the quantile level and in the bandwidth parameter*. Uniformity in the quantile level is important because it allows one to recover the conditional cdf of the response (except perhaps for the tails) at any given level of the covariate. Uniformity in the smoothing parameter, in turn, is crucial in that it encompasses data driven bandwidth choices, as well as bandwidths which depend on the quantile level, on the covariate level, etc. Finally, we assess the quality of our method through a simulation study.

2.2 The smoothed quantile regression estimator

Let $(Y_i, X_i),\ i = 1, \ldots, n$, be an iid sample drawn from the distribution of $(Y, X) \in \mathbb{R}\times\mathbb{R}^d$, where the conditional quantile of the response Y given the covariate $X = x$ satisfies the linear model

$$Q(\tau|x) = x'\beta(\tau), \quad \tau \in (0,1). \tag{2.1}$$

Here by definition $Q(\tau|x) := \inf\{q : F(q|x) \geq \tau\}$, and $F(\cdot|x)$ is the conditional cdf of Y given the covariate, i.e. $F(y|x) := \mathbb{P}[Y \leq y|X = x]$. Set $e_i(b) := Y_i - X_i'b$ and define

$$\widehat{R}(b;\tau) := \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(e_i(b)) = \int \rho_\tau(t)\, d\widehat{F}(t;b), \tag{2.2}$$

where $\widehat{F}(\cdot\,; b)$ denotes the discrete empirical distribution function of the error terms $e_i(b)$, and where $\rho_\tau(u) := u(\tau - \mathbb{I}[u < 0])$ is the usual check function. The standard quantile regression estimator $\widehat{\beta}(\tau)$ of Koenker and Bassett (1978) minimizes the map $b \mapsto \widehat{R}(b;\tau)$, a sample analogue of the population objective function

$$R(b;\tau) := \mathbb{E}[\rho_\tau(e(b))] =: \int \rho_\tau(t)\, dF(t;b) \tag{2.3}$$

of which the true parameter $\beta(\tau)$ is a minimizer. Here $e(b) := Y - X'b$, and $F(t;\, b) = \mathbb{P}[e(b) \leq t]$. Now the RHS of equation (2.3) suggests a class of quantile regression estimators taken by minimizing, with respect to $b \in \mathbb{R}^d$, objective functions that are integrals of ρ_τ where the integrating measure is an estimator of the distribution function of the error term $e(b)$. Under this interpretation, the standard quantile regression estimator corresponds to taking the discrete empirical distribution function as an estimator of $F(\cdot\,; b)$. The approach that we shall take here is to consider kernel-type estimators of this cdf. In the one-sample scenario this is equivalent to Nadaraya (1964).

Consider[1] a bandwidth $h > 0$ which goes to 0 when the sample size grows, and a smooth kernel function k satisfying $\int k(v)\, dv = 1$, and set $k_h(v) = k(v/h)/h$. Let $\widehat{f}_h(\cdot\,; b)$ and $\widehat{F}_h(\cdot\,; b)$ denote the usual kernel pdf and cdf estimators, given respectively by

$$\widehat{f}_h(v;b) = \frac{1}{n}\sum_{i=1}^{n} k_h(v - e_i(b))$$

and

$$\widehat{F}_h(t;b) = \int_{-\infty}^{t} \widehat{f}_h(v;b)\, dv.$$

The proposed objective function $\widehat{R}_h(b;\tau)$ is then computed as $\widehat{R}(b;\tau)$ in (2.2) but using the kernel pdf estimator $\widehat{f}_h(\,\cdot\,; b)$ instead of the discrete $d\widehat{F}(\,\cdot\,; b)$. That is to say,

$$\widehat{R}_h(b;\tau) := \int \rho_\tau(t)\, d\widehat{F}_h(t;b) = \int \rho_\tau(t)\widehat{f}_h(t;b)\, dt. \tag{2.4}$$

The resulting smoothed quantile regression estimator is then

$$\widehat{\beta}_h(\tau) := \arg\min_{b \in \mathbb{R}^d} \widehat{R}_h(b;\tau). \tag{2.5}$$

[1]Precise definitions, and conditions on the bandwidth parameter and the kernel function are given in the list of assumptions below.

An important consequence of our smoothing procedure is that it implies that the objective function $b \mapsto \widehat{R}_h(b;\tau)$ is twice continuously differentiable, contrasting with the lack of smoothness of the standard objective function $b \mapsto \widehat{R}(b;\tau)$. Whereas estimation of the covariance matrix of $\widehat{\beta}(\tau)$ through standard sandwich formulas is precluded, our approach circumvents this difficulties and ensures the covariance can be estimated in a natural fashion as discussed below. Moreover, differentiability of the smoothed objective function implies that $\widehat{\beta}_h(\tau)$ can be computed using standard Newton-Raphson algorithms, avoiding computational preoccupations that arise in the context of standard quantile regression.

Smoothness of $\widehat{R}_h(\cdot;\tau)$ can be established by observing that

$$\widehat{R}_h(b;\tau) = (1-\tau)\int_{-\infty}^{0} \widehat{F}_h(v;b)\,dv + \tau\int_{0}^{\infty}\Big(1-\widehat{F}_h(v;b)\Big)\,dv,$$

from which it follows through standard arguments that the first-order b-derivative of $\widehat{R}_h(b;\tau)$ is given by

$$\widehat{R}_h^{(1)}(b;\tau) = \frac{1}{n}\sum_{i=1}^{n} X_i\left[K\left(-\frac{e_i(b)}{h}\right)-\tau\right],$$

where $K(t) := \int_{-\infty}^{t} k(v)dv$. In the same fashion the second-order b-derivative of $\widehat{R}_h(b;\tau)$ is seen to be

$$\widehat{R}_h^{(2)}(b;\tau) = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' k_h\Big(-e_i(b)\Big).$$

As a consequence of differentiability, inference for $\widehat{\beta}_h(\tau)$ can be implemented in a standard manner. As seen from Proposition 2.1 and Corollary 2.2 below, $\sqrt{n}\Big(\widehat{\beta}_h(\tau)-\beta(\tau)\Big)$ is asymptotically centered normal with covariance matrix given by $\Sigma(\tau) := D(\tau)^{-1}V(\tau)D(\tau)^{-1}$ where $D(\tau)$ is the Hessian of the population objective function evaluated at the true parameter,

$$D(\tau) := R^{(2)}(\beta(\tau);\tau) \equiv \mathbb{E}XX'f(X'\beta(\tau)|X),$$

and $V(\tau) := \tau(1-\tau)\mathbb{E}XX'$. In our framework $\Sigma(\tau)$ can be consistently estimated by

$$\widehat{\Sigma}_h(\tau) := \widehat{D}_h(\tau)^{-1}\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1}$$

with $\widehat{D}_h(\tau) := \widehat{R}_h^{(2)}\Big(\widehat{\beta}_h(\tau);\tau\Big)$ and

$$\widehat{V}_h(\tau) := \frac{1}{n}\sum_{i=1}^{n} X_i X_i'\left[K\left(-\frac{e_i\Big(\widehat{\beta}_h(\tau)\Big)}{h}\right)-\tau\right]^2.$$

Thus for instance a $1-\alpha$ confidence interval for the kth entry $\beta^k(\tau)$ of the regression quantile can be computed straightforwardly via

$$\mathrm{CI}_{1-\alpha}\Big(\beta^k(\tau)\Big) := \widehat{\beta}_h^k(\tau) \pm \frac{z_{\alpha/2}\widehat{\sigma}_h^k(\tau)}{\sqrt{n}},$$

where $\widehat{\sigma}_h^k(\tau)$ is the square root of the kth diagonal entry of $\widehat{\Sigma}_h(\tau)$, $\widehat{\beta}_h^k(\tau)$ is the kth component of the smoothed quantile regression estimator, and z_α is the α quantile of a standard gaussian distribution.

To further illustrate the usefulness of our smoothing procedure let us consider a one-sample scenario, i.e. the case $d = 1$ and $X \equiv 1$. First observe that in this setting the first-order condition solved by the smoothed estimator, namely $\widehat{R}_h^{(1)}\Big(\widehat{\beta}_h(\tau); \tau\Big) = 0$, reduces to $\widehat{F}_h\Big(\widehat{\beta}_h(\tau)\Big) = \tau$, proving our prior claim that in the absence of a covariate our approach corresponds to Nadaraya (1964). It becomes clear as well that $\widehat{R}_h^{(1)}(b; \tau)$ is analogous to a smooth estimator of a cumulative distribution function. This is in contrast with the smoothing procedure put forth by Horowitz (1998), namely to substitute the indicator function in $\widehat{R}(b; \tau) = n^{-1}\sum_{i=1}^n e_i(b)(\tau - \mathbb{I}[e_i(b) < 0])$ by a smooth counterpart, from which one obtains the objective function

$$\widetilde{R}_h(b; \tau) := n^{-1}\sum_{i=1}^n e_i(b)(\tau - K(-e_i(b)/h)),$$

whose first-order b-derivative is seen to be equal to $\widehat{R}_h^{(1)}(b; \tau)$ plus an additional term which depends explicitly on the kernel k. Thus the first-order condition involves a quantity analogous to a kernel-based estimator of a probability density function. As a result, the second-order derivative of our smoothed quantile regression estimator, $\widehat{R}_h^{(2)}(b; \tau)$, is similar to a kernel density estimator, whereas Horowitz's involves terms that are similar to a kernel estimator of the derivative of a probability density function. This ensures that our estimator has better higher-order properties than Horowitz given that the kernel density estimator converges at a faster rate relative to the kernel derivative estimator. This argument can straightforwardly be carried further to show that in linear approximations the variance of Horowitz's estimator will be greater than that of our smoothed estimator.

2.3 Main results

Before establishing our main results, it will be convenient to introduce some further notation. The rationale will be discussed in the course of this section. Let us denote the conditional probability density function of Y given $X = x$ by $f(\cdot|x)$, and write $f^{(j)}(y|x)$ for the jth derivative $\partial^j/\partial y^j\, f(y|x)$. Similarly put $Q^{(1)}(\tau|x) := \partial/\partial\tau\, Q(\tau|x)$. Let also $\lfloor s \rfloor$ denote the lower integer part of any positive real number s, that is to say[2], the unique integer number satisfying $\lfloor s \rfloor < s \leq \lfloor s \rfloor + 1$.

As will be argued below, it is fruitful to interpret $\widehat{\beta}_h(\tau)$ as an estimator not of $\beta(\tau)$ but rather of $\beta_h(\tau)$ defined by

$$\beta_h(\tau) := \arg\min_{b \in \mathbb{R}^d} R_h(b; \tau),$$

[2]Pay attention to the definition as $\lfloor k \rfloor = k - 1$ if k is an integer.

where $R_h(b;\tau) := \mathbb{E}\widehat{R}_h(b;\tau)$. We shall refer to $\beta_h(\tau)$ as the smoothed parameter. It will also prove convenient to introduce the quantity $\widehat{S}_h(\tau) := \widehat{R}_h^{(1)}(\beta_h(\tau);\tau)$, and to set $D_h(\tau) := R_h^{(2)}\big(\beta_h(\tau);\tau\big)$.

In what follows we let $\|\cdot\|$ denote the Euclidean norm of a vector or a matrix, namely $\|A\| = \sqrt{\mathrm{tr}(AA')}$.

2.3.1 Assumptions

Our main assumptions are as follows

Assumption X The components of X are positive, bounded random variables, i.e. the support of X is a bounded subset of $\mathbb{R}^d_+$. The matrix $\mathbb{E}XX'$ is full rank.

Assumption Q The conditional quantile function $Q(\tau|x)$ and the conditional pdf $f(y|x)$ satisfy

Q1 The map $\tau \mapsto \beta(\tau)$ is continuously differentiable over $(0,1)$. The conditional densities $f(y|x)$ are continuous and strictly positive over $\mathbb{R} \times \mathrm{supp}(X)$.

Q2 There are some $s \geq 1$ and $L > 0$ such that $f^{(\lfloor s \rfloor)}(\cdot|x)$ exists, and

$$\sup_{x,y} \Big| f^{(j)}(y|x) \Big| \leq L$$

with $\lim_{y\to\pm\infty} f^{(j)}(y|x) = 0$, for all $j = 0, \dots, \lfloor s \rfloor$. Moreover, it holds that

$$\Big| f^{(\lfloor s \rfloor)}(y|x) - f^{(\lfloor s \rfloor)}(y + w|x) \Big| \leq L|w|^{s-\lfloor s \rfloor}$$

for all $x \in \mathrm{supp}(X)$ and all $y, w \in \mathbb{R}$.

Assumption K The kernel function k and the bandwidth h satisfy

K1 The kernel $k : \mathbb{R} \to \mathbb{R}$ is even, integrable, piecewise differentiable with a bounded derivative, and $\int k(z)\, dz = 1$. Moreover, it holds that

$$0 < \int_0^{+\infty} K(z)(1 - K(z))\, dz < \infty.$$

For s as in Assumption Q2, $\int |z^{s+1}k(z)|\, dz < \infty$, and k is orthogonal to all nonconstant monomials of degree up to $\lfloor s \rfloor + 1$:

$$\int z^j k(z)\, dz = 0.$$

for $j = 1, \dots, \lfloor s \rfloor + 1$.

K2 $h \in \big[\underline{h}, \bar{h}\big] \equiv \big[\underline{h}(n), \bar{h}(n)\big]$ with $1/\underline{h} = O\big(n/\log^3 n\big)$ and $\bar{h} = o(1)$.

Some remarks on the Assumptions come in handy. First, observe that Assumptions Q1 and X ensure that $R^{(2)}(b;\tau)$ is positive definite for all b and any τ. Indeed $R^{(2)}(b;\tau) = \mathbb{E}[XX'f(X'b|X)]$. In particular, $D(\tau)^{-1}$ as defined above exists for all τ. Assumption Q1 also ensures that $\tau \mapsto Q(\tau|x)$ is strictly increasing over $(0,1)$, with a strictly positive τ-derivative via the relation

$$Q^{(1)}(\tau|x) = \frac{1}{f(Q(\tau|x)|x)}.$$

Notice though that the assumption that Y is supported on the real line can be relaxed with minor adaptations. We adopt it for notational simplicity. One should also notice that for integer s in Q2 the assumed order of differentiability of $f(\cdot|x)$ is $\lfloor s \rfloor \equiv s-1$ and not s as may be thought, and that all the quantities that depend on h also depend implicitly on n through Assumption K2.

A last point that is worth remarking upon is that we have so far referred to $\widehat{f}_h(\cdot;b)$ as an estimator of the density of the error term $e(b)$ but strictly speaking when the kernel k is not a density then neither will $\widehat{f}_h(\cdot;b)$ be. Indeed except in the cases where $\lfloor s \rfloor = 0$ Assumption K1 precludes the possibility of $\widehat{f}_h(\cdot;b)$ being a density. It is however easily shown that our estimator satisfies $\widehat{R}_h(b;\tau) = (1/n)\sum_{i=1}^n \rho_\tau * k_h(e_i(b))$, where $*$ is the convolution operation. Thus technically speaking it is more correct to interpret our approach as a convolution-type smoothing, although some of the heuristic intuition is lost.

2.3.2 Bahadur-Kiefer representation

Our first theorem studies the remainder term of a Bahadur-Kiefer representation of the statistic $\sqrt{n}\big(\widehat{\beta}_h(\tau) - \beta_h(\tau)\big)$, that is, the approximation of this quantity by the standardized sum $-\sqrt{n}D_h(\tau)^{-1}\widehat{S}_h(\tau)$. A relevant aspect of Theorem 2.1 is that the stated representation holds uniformly both in the quantile level and in the bandwidth parameter. The proof of this fact relies on a powerful functional exponential inequality due to Massart (2007).

Theorem 2.1. *Let Assumptions X, Q and K hold. Then* $\widehat{\beta}_h(\tau)$ *is unique for* $(\tau,h) \in [\underline{\tau},\bar{\tau}] \times \left[\underline{h},\bar{h}\right]$ *with probability tending to 1, and satisfies the following representation uniformly with respect to* $(\tau,h) \in [\underline{\tau},\bar{\tau}] \times \left[\underline{h},\bar{h}\right]$,

$$\sqrt{n}\Big(\widehat{\beta}_h(\tau) - \beta_h(\tau)\Big) = -\sqrt{n}D_h(\tau)^{-1}\widehat{S}_h(\tau) + O_{\mathbb{P}}(\varrho_n(h)), \tag{2.6}$$

where

$$\varrho_n(h)^{-1} = \left(\frac{nh}{\log n}\right)^{1/2} + n^{1/2}.$$

An important consequence of the uniformity property in Theorem 2.1 is that if a stochastic process $\Big(\widehat{h}(\tau),\, \tau \in [\underline{\tau},\bar{\tau}]\Big)$ has its sample paths in $\left[\underline{h},\bar{h}\right]$ with a high probability, then the representation in (2.6) remains valid with h replaced by $\widehat{h}(\tau)$. This fact is specially important for it allows the statistician to plug in data-driven bandwidth choices, which may or may not depend on the quantile level, while retaining the properties implied by linearization. This is summarized in the following Corollary.

Corollary 2.1. *Let* $\left(\widehat{h}(\tau),\, \tau \in [\underline{\tau}, \bar{\tau}]\right)$ *satisfy* $\mathbb{P}\left(\widehat{h}(\tau) \in \left[\underline{h}, \bar{h}\right] \text{ for all } \tau\right) \to 1$. *Then* (2.6) *holds with* $\widehat{h}(\tau)$ *in place of* h*, uniformly in* τ.

Theorem 2.1 is first relevant because it shows that the smoothed estimator $\widehat{\beta}_h(\tau)$ is, in a sense, more linear than $\widehat{\beta}(\tau)$. Indeed, the well-known Bahadur-Kiefer representation of the standard quantile regression estimator, $\sqrt{n}\left(\widehat{\beta}(\tau) - \beta(\tau)\right) \approx -\sqrt{n}D(\tau)^{-1}\widehat{S}(\tau)$, where $\widehat{S}(\tau) = \widehat{R}^{(1)}(\beta(\tau);\tau)$ whenever this derivative exists, has a remainder term whose rate is $O_{\mathbb{P}}\left(n^{-1/4}\right)$ in many cases of interest (see Portnoy (2012), Jurečková et al. (2012), Knight (2001)). In contrast, the order of the remainder of (2.6) is at least $n^{-1/2}$. Theorem 2.1 also implies, together with Theorem 2.2 below, that $\widehat{\beta}_h(\cdot)$ gives a fair global picture of the slope coefficients $\beta(\cdot)$, in the sense that

$$\left\|\widehat{\beta}_h(\tau) - \beta(\tau)\right\| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + h^{s+1}\right), \tag{2.7}$$

uniformly for $\tau \in [\underline{\tau}, \bar{\tau}]$ and $h \in \left[\underline{h}, \bar{h}\right]$. In particular, if $h \leq O\left(n^{-1/(2(s+1))}\right)$, then the remainder in (2.7) is $O_{\mathbb{P}}\left(n^{-1/2}\right)$.

2.3.3 Asymptotic mean squared error

The bias term. Although this paper ought to illustrate the positive aspects of smoothing, any reader familiar with nonparametric approaches is probably already aware that the benefits expected from these techniques come with potential drawbacks. Indeed, the smoothed quantile regression estimator $\widehat{\beta}_h(\tau)$ should not be viewed as an estimator of $\beta(\tau)$, an approach which would amount to ignore the impact of smoothing. It is more suitable to interpret $\widehat{\beta}_h(\tau)$ as an estimator of $\beta_h(\tau)$ as defined above, a point of view that acknowledges that $\widehat{\beta}_h(\tau)$ can be a biased estimator of $\beta(\tau)$. The next result studies the order of the bias term $\beta_h(\tau) - \beta(\tau)$.

Theorem 2.2. *Given Assumptions X, Q, and K, and provided that* $\bar{h}$ *is small enough,* $\beta_h(\tau)$ *is uniquely defined for all* $\tau \in [\underline{\tau}, \bar{\tau}]$ *and satisfies, uniformly with respect to* $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$,

$$\beta_h(\tau) = \beta(\tau) + O\left(h^{s+1}\right). \tag{2.8}$$

Additionally, if s *is an integer number and* $y \mapsto f(y|x)$ *is* s *times continuously differentiable for all* x*, then the following expansion holds*

$$\beta_h(\tau) = \beta(\tau) - h^{s+1}\mathbf{B}(\tau) + o\left(h^{s+1}\right), \tag{2.9}$$

where

$$\mathbf{B}(\tau) = D(\tau)^{-1}\frac{\int z^{s+1}k(z)\,dz}{(s+1)!}\;\mathbb{E}\Big[Xf^{(s)}(X'\beta(\tau)|X)\Big].$$

Theorem 2.2 settles the issue of possible side effects of smoothing: $\beta_h(\tau)$ is eventually uniformly close to the true parameter $\beta(\tau)$. This fact also serves as a further justification for the standardization in (2.6).

Asymptotic variance. The following result shows that the large sample variance of $\widehat{\beta}_h(\tau)$ is smaller than that of $\widehat{\beta}(\tau)$ in regard to ordering of positive matrices.

Theorem 2.3. *Given Assumptions X, Q and K we have, uniformly with respect to* $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \big[\underline{h}, \bar{h}\big]$,

$$\operatorname{Var}\Big(\sqrt{n}D_h(\tau)^{-1}\widehat{S}_h(\tau)\Big) = \Sigma(\tau) - c_k h D(\tau)^{-1} + O\Big(h^{2\wedge s}\Big), \tag{2.10}$$

where $c_k = 2\int_0^{+\infty} K(y)(1 - K(y))\,dy$.

Asymptotic mean squared Error. The next result focuses on the optimal choice of the bandwidth h when estimating a linear combination $\lambda'\beta(\tau)$. This includes in particular estimation of each of the coefficients $\beta_j(\tau)$ by a proper choice of the vector $\lambda \in \mathbb{R}^d$. Define the Asymptotic Mean Squared Error of $\lambda'\widehat{\beta}_h(\tau)$ as

$$\text{AMSE}\Big(\lambda'\widehat{\beta}_h(\tau)\Big) = \mathbb{E}\Big\{\lambda'\Big(\beta_h(\tau) - D_h(\tau)^{-1}\widehat{S}_h(\tau) - \beta(\tau)\Big)\Big\}^2.$$

This quantity is a proxy for the mean squared error

$$\text{MSE}\Big(\lambda'\widehat{\beta}_h(\tau)\Big) = \mathbb{E}\Big\{\lambda'\Big(\widehat{\beta}_h(\tau) - \beta(\tau)\Big)\Big\}^2.$$

Studying the AMSE instead of the MSE amounts to neglecting the remainder term of (2.6), which shrinks to 0. The next result describes an optimal bandwidth choice with respect to the AMSE criterion.

Theorem 2.4. *Let Assumptions X, Q and K hold. If s is an integer number and $y \mapsto f(y|x)$ is s times continuously differentiable for all x, then, provided $\lambda'\mathbf{B}(\tau) \neq 0$, the* $\text{AMSE}\Big(\lambda'\widehat{\beta}_h(\tau)\Big)$ *is asymptotically minimal for the bandwidth h^* given by*

$$h^* = \left(\frac{c_k\lambda' D(\tau)^{-1}\lambda}{2n(s+1)(\lambda'\mathbf{B}(\tau))^2}\right)^{\frac{1}{2s+1}}.$$

In this case,

$$\text{AMSE}\Big(\lambda'\widehat{\beta}_h(\tau)\Big) = \frac{1}{n}\lambda'\Big(\Sigma(\tau) - c_k h^* D(\tau)^{-1}\Big)\lambda + o\Big(n^{-1}\Big).$$

Remark. Theorem 2.4 remains valid for bandwidths of the form $h = (1 + o(1))h^*$.

2.3.4 Inference

In this section inferential aspects of $\widehat{\beta}_h(\tau)$ are discussed. The next result deals with estimation of the asymptotic covariance matrix of $\widehat{\beta}_h(\tau)$. The proposed estimator is intrinsic to the method and falls into a canonical framework. An important consequence is that Wald-type inference is immediately applicable.

Proposition 2.1. *Under Assumptions X, Q and K it holds that*

$$\widehat{\Sigma}_h(\tau) = \Sigma(\tau) + O_{\mathbb{P}}\left(\frac{1}{h\sqrt{n}} + h\right)$$

uniformly in $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$.

Combining Theorem 2.6 and Proposition 2.1 yields the following Corollary.

Corollary 2.2 (Central Limit Theorem)**.** *Let Assumptions X, Q and K hold. If* $\sqrt{n}h^{s+1} \to 0$*, then*

$$\sqrt{n}\widehat{\Sigma}_h(\tau)^{-1/2}\left(\widehat{\beta}_h(\tau) - \beta(\tau)\right) \to N(0, Id).$$

Remark. Theorem 2.3 and Corollary 2.2 are to be contrasted with the asymptotic distribution of $\sqrt{n}\Sigma(\tau)^{-1/2}\left(\widehat{\beta}(\tau) - \beta(\tau)\right)$ which is also $N(0, \mathrm{Id})$. A key point in Theorem 2.3 is the fact that the variance expansion in (2.10) includes the negative term $-c_k h$: this implies that the large sample variance of the smoothed $\widehat{\beta}_h(\tau)$ is smaller than the one of the quantile regression estimator $\widehat{\beta}(\tau)$. As already noticed by Azzalini (1981) for the univariate quantile estimator, it then follows that the AMSE of $\lambda'\widehat{\beta}_h(\tau)$ can be made smaller than the one of $\lambda'\widehat{\beta}(\tau)$. In other words the smoothed quantile regression estimator improves the standard quantile regression estimator at second order.

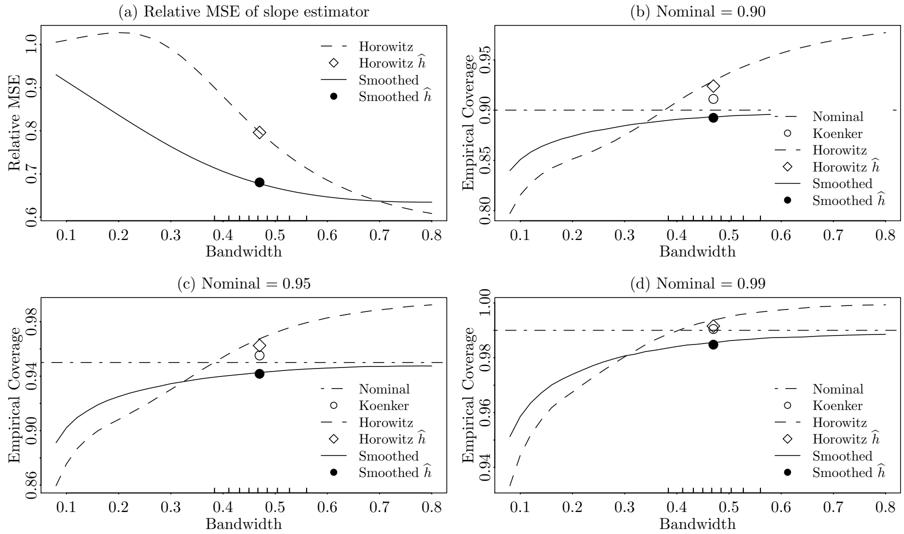
2.4 Simulation Study

In order to illustrate the obtained theoretical results, and to assess the quality of our method, we simulated data from the median regression model

$$Y = X'\beta + \epsilon. \tag{2.11}$$

The true parameter is set to $\beta = (1, 1)$, and the covariate defined as $X = \left(1, \tilde{X}\right)$, with $\tilde{X} \sim U[1, 5]$. Three different specifications for the distribution of the error term are considered. These are *(i)* an asymmetric model, with $\epsilon = Z - \log 2/\sqrt{2}$, where $Z \sim \text{Exponential}\left(1/\sqrt{2}\right)$; *(ii)* a heavy-tailed model, with $\epsilon = \sqrt{2/3} \times Z$, where $Z \sim t(3)$, and; *(iii)* a heteroskedastic model, with $\epsilon = 0.25\left(1 + \tilde{X}\right)Z$, where $Z \sim N(0, 1)$. DGP's *(ii)* and *(iii)* are considered for instance in Horowitz (1998), Whang (2006) and Kaplan and

Figure 2.1: Model = Exponential; $n = 100$. Relative MSE (panel (a)), and empirical coverage probabilities (panels (b)–(d)).

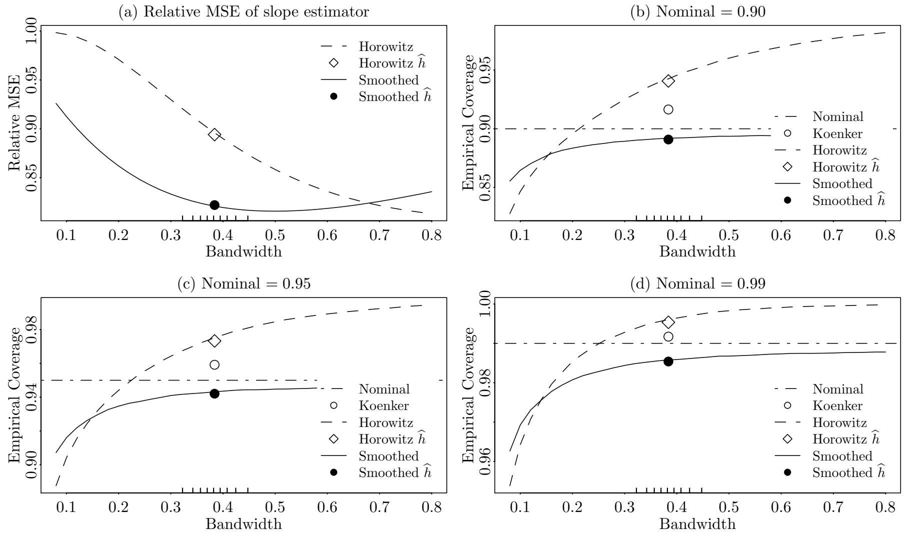


Sun (2012). In all three models the conditional quantile of ϵ given X is zero; in models *(i)* and *(ii)*, $\operatorname{Var}\epsilon = 2$.

The simulation experiment consists of, in each of the 10000 replications, sampling $n = 100$ observations from model (2.11), in each of the three error specifications above, and then computing the three estimates of β (i.e the standard quantile regression estimator of Koenker and Bassett (1978), the smoothed median regression estimator of Horowitz (1998), and the smoothed quantile regression estimator (2.5)). The corresponding t-statistics are also computed. The smoothed estimator (2.5) and the Horowitz (1998) median regression estimator were computed with the smoothing parameter varying in a grid of points ranging from 0.1 to 0.8 and using a Gaussian kernel k. We also computed these estimates using a (data dependent) rule of thumb bandwidth $\widehat{h} = 1.06\widehat{s}/n^{1/5}$, where $\widehat{s}$ measures the variability of the residuals from the standard quantile regression fit to the data (see Silverman (1986) and equations (3.28) and (3.30) therein). The standard errors for the t-statistics were calculated respectively as described in Koenker (2005, sections 3.4.2 and 4.10.1), in Horowitz (1998, section 2), and using the square root of diagonal entries of $\widehat{\Sigma}_h(0.5)$. All computations were carried out in the statistical package `R`.

Figures 2.1–2.3 display the simulation results. We shall explain in detail the contents of Figure 2.1, the other ones being entirely analogous. All Figures refer to the slope parameter β_2 and associated quantities. Figure 2.1 corresponds to the Exponential model. In Panel (a) one finds the relative Mean Squared Error (across replications) of Horowitz's estimator and of the smoothed estimator (2.5), that is the ratio of their MSEs to the MSE of standard quantile regression estimator. Panels (b)–(d) display the empirical coverage probabilities of the t-statistics of each of the three aforementioned estimators, that is the proportion of replications in which their absolute values lay below the thresholds of (b) 1.64; (c) 1.96, and; (d) 2.58, corresponding to the ranks at which the absolute value of a

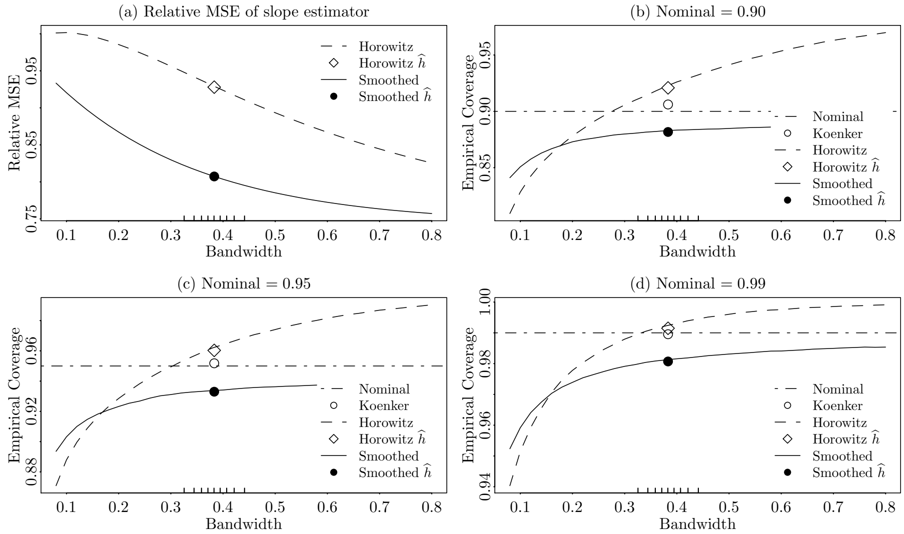
Figure 2.2: Model = t; $n = 100$. Relative MSE (panel (a)), and empirical coverage probabilities (panels (b)–(d)).



standard normal random variable attains the nominal coverage probabilities respectively of 0.90, 0.95 and 0.99. In all Panels, said values are plotted as a function of the smoothing parameter h (Horowitz's estimator and smoothed estimator) or as a point whose vertical position is set (arbitrarily) to be the mean of $\widehat{h}$ across replications. The vertical bars at the bottom of each plot correspond to the 10th to 90th percentile of the rule-of-thumb bandwidth $\widehat{h}$ across the 10000 replications.

Let us now analyze the results in further detail. In all three model specifications the relative MSE of $\widehat{\beta}_h$ dominated both the standard quantile regression estimator and Horowitz's smoothed median regression estimator, for a wide range of h values and with improvements as large as 30% for the exponential model. In particular the rule-of-thumb bandwidth induced a smoothed estimator with improvements of at least nearly 20% in all scenarios. In the exponential setting, coverage probabilities associated to our smoothed estimator were very precise for bandwidths above 0.45, and not too sensible to variations in the smoothing parameter, in contrast to Horowitz's estimator which attains the nominal coverage for h values near 0.36 but displays a poor performance as one moves h apart from this optimal value. When contrasted to the standard $\widehat{\beta}$, the smoothed estimator displayed comparable coverage for a wide range of bandwidth values. In DGP's *(ii)* and *(iii)* coverage probabilities associated to $\widehat{\beta}_h$ did not display the same degree of accuracy seen in the Exponential setting. Nevertheless there seems to be no significant compromise when compared to both the standard quantile regression estimator and Horowitz's smoothed estimator, even more when one takes into account the gains in MSE. Sensitivity to variations in h appear to remain an issue for Horowitz's estimator, whereas the empirical coverages of $\widehat{\beta}_h$ at large bandwidths (values above around 0.4) were nearly constant.

Figure 2.3: Model = Heteroskedastic; n = 100. Relative MSE (panel (a)), and empirical coverage probabilities (panels (b)–(d)).



2.5 Proofs

Notice that $\widehat{R}(b;\tau)$ is integrable if and only if Y and X are integrable. This points to the fact that the literature is not always completely rigorous when defining $\beta(\tau)$ as the minimizer of $R(b;\tau) \equiv \mathbb{E}\widehat{R}(b;\tau)$. It is convenient however to assume altogether that $\widehat{R}(b;\tau)$ and $\widehat{R}_h(b;\tau)$ are integrable, such that $R(b;\tau)$ and $R_h(b;\tau)$ are well defined. If not, $R_h(b;\tau)$ should be defined as $\mathbb{E}[\widehat{R}_h(b;\tau) - \widehat{R}_h(0;\tau)]$, and similarly for $R(b;\tau)$. These quantities are finite under Assumption X. Let $\mathcal{S}$ denote the set $\mathbb{R}^d \times [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$ to which (b,τ,h) belongs. Notice that $\mathcal{S}$ depends on n through $\left[\underline{h}, \bar{h}\right]$. In what follows, whenever convenient we write $\mathbb{E}^X$ (resp. $\mathbb{E}^x$) to denote conditional expectation given X (resp. given $X = x$). C is a constant which may vary from line to line.

2.5.1 The bias term

The proof of Theorem 2.2 makes use of the following Lemma.

Lemma 2.1. *Assumptions X, Q2 and K1 ensure that*

(i) $$\sup_{(b,\tau,h)\in\mathcal{S}} \left| \frac{R_h(b;\tau) - R(b;\tau)}{h^{\lfloor s \rfloor + 1}} \right| = O(1);$$

(ii) $$\sup_{(b,\tau,h)\in\mathcal{S}} \left\| \frac{R_h^{(1)}(b;\tau) - R^{(1)}(b;\tau)}{h^{s+1}} \right\| = O(1);$$

(iii) $$\sup_{(b,\tau,h)\in\mathcal{S}} \left\| \frac{R_h^{(2)}(b;\tau) - R^{(2)}(b;\tau)}{h^{s}} \right\| = O(1);$$

(iv) $$\sup_{(\delta,b,\tau,h)\in\mathbb{R}^d\times\mathcal{S}} \left\| \frac{R_h^{(2)}(b+\delta;\tau) - R_h^{(2)}(b;\tau)}{\|\delta\|} \right\| = O(1).$$

Proof of Lemma 2.1. Assume from now on that $\lfloor s \rfloor \geq 1$, as the case $\lfloor s \rfloor = 0$ can be dealt with in a similar fashion by invoking the Hölder condition for $f(\cdot|x)$ instead of the one on $f^{(\lfloor s \rfloor)}(\cdot|x)$ as used below. Under Assumption Q2, a Taylor expansion with integral remainder gives

$$f(v+hz|x) = \sum_{\ell=0}^{\lfloor s \rfloor -1} f^{(\ell)}(v|x)\frac{(hz)^{\ell}}{\ell!} + \frac{(hz)^{\lfloor s \rfloor}}{(\lfloor s \rfloor - 1)!}\int_0^1 f^{(\lfloor s \rfloor)}(v+whz|x)(1-w)^{\lfloor s \rfloor -1}\,dw.$$

The following identity will give items (*i*)–(*iii*). Using a change of variables $y = v + hz$ yields, under Assumption K1,

$$\begin{aligned}
&\mathbb{E}^x\{k_h(v-Y)\} - f(v|x)\\
&= \int k_h(v-y)f(y|x)\,dy - f(v|x)\\
&= \int k(z)(f(v+hz|x) - f(v|x))\,dz\\
&= \int_0^1 (1-w)^{\lfloor s \rfloor -1}\int \frac{(hz)^{\lfloor s \rfloor}}{(\lfloor s \rfloor - 1)!}k(z)f^{(\lfloor s \rfloor)}(v+whz|x)\,dz\,dw\\
&= \int_0^1 (1-w)^{\lfloor s \rfloor -1}\int \frac{(hz)^{\lfloor s \rfloor}}{(\lfloor s \rfloor - 1)!}k(z)\Big(f^{(\lfloor s \rfloor)}(v+whz|x) - f^{(\lfloor s \rfloor)}(v|x)\Big)\,dz\,dw.
\end{aligned} \tag{2.12}$$

Let us establish (*i*). Observe that if G is an arbitrary cdf then $\int \rho_\tau(v)\,dG(v) = (1-\tau)\int_{-\infty}^{0} G(v)\,dv + \tau\int_0^{+\infty}(1-G(v))\,dv$. Thus one has

$$R(b;\tau) = \int \left\{(1-\tau)\int_{-\infty}^{0}\int_{-\infty}^{t+x'b} f(v|x)\,dv\,dt + \tau\int_0^{\infty}\int_{t+x'b}^{\infty} f(v|x)\,dv\,dt\right\} dF_X(x),$$

and similarly,

$$\begin{aligned}
&R_h(b;\tau) =\\
&= \int \left\{(1-\tau)\int_{-\infty}^{0}\int_{-\infty}^{t+x'b} \mathbb{E}^x\{k_h(v-Y)\}\,dv\,dt + \tau\int_0^{\infty}\int_{t+x'b}^{\infty} \mathbb{E}^x\{k_h(v-Y)\}\,dv\,dt\right\} dF_X(x).
\end{aligned}$$

Since by hypothesis $\int \left|z^{\lfloor s \rfloor +1}k(z)\right| dz < \infty$, and noting that $f^{(\lfloor s \rfloor -2)}(\cdot|\cdot)$ is Lipschitz, one gets from equation (2.12)

$$\begin{aligned}
&\left|\int_{-\infty}^{0}\int_{-\infty}^{t+x'b} \mathbb{E}^x\{k_h(v-Y)\} - f(v|x)\,dv\,dt\right| =\\
&= \left|\int_0^1 (1-w)^{\lfloor s \rfloor -1}\int \frac{(hz)^{\lfloor s \rfloor}}{(\lfloor s \rfloor - 1)!}k(z)\int_{-\infty}^{0}\int_{-\infty}^{t+x'b}\Big(f^{(\lfloor s \rfloor)}(v+whz|x) - f^{(\lfloor s \rfloor)}(v|x)\Big)\,dv\,dt\,dz\,dw\right|\\
&= \left|\int_0^1 (1-w)^{\lfloor s \rfloor -1}\int \frac{(hz)^{\lfloor s \rfloor}}{(\lfloor s \rfloor - 1)!}k(z)\Big(f^{(\lfloor s \rfloor -2)}(x'b+whz|x) - f^{(\lfloor s \rfloor -2)}(x'b|x)\Big)\,dz\,dw\right|\\
&\leq Ch^{\lfloor s \rfloor +1}.
\end{aligned}$$

The bound for $\left|\int_0^\infty \int_{t+x'b}^\infty \mathbb{E}^x\{k_h(v-Y)\} - f(v|x)\,dv\,dt\right|$ is obtained analogously. Hence it holds that $|R_h(b;\tau) - R(b;\tau)| \le Ch^{\lfloor s\rfloor+1}$.

For (*ii*) notice that by the definition of $R(b;\tau)$ and $R_h(b;\tau)$, via the Lebesgue Dominated Convergence Theorem it holds that

$$R^{(1)}(b;\tau) = \mathbb{E}\{X[F(X'b|X) - \tau]\} \equiv \int x\left(\int_{-\infty}^{x'b} f(y|x)\,dv - \tau\right) dF_X(x),$$

whereas

(2.13)
$$R_h^{(1)}(b;\tau) = \mathbb{E}\left\{X\left[K\left(\frac{X'b - Y}{h}\right) - \tau\right]\right\} \equiv \int x\left(\int_{-\infty}^{x'b} \mathbb{E}^x\{k_h(v-Y)\}\,dv - \tau\right) dF_X(x).$$

Integrating (2.12) yields, since $\int z^{\lfloor s\rfloor}k(z)\,dz = \int z^{\lfloor s\rfloor+1}k(z)\,dz = 0$ and $\int\left|z^{\lfloor s\rfloor+1}k(z)\right|dz < \infty$,

$$\begin{aligned}
&\left|\int_{-\infty}^{x'b} \mathbb{E}^x\{k_h(v-Y)\} - f(v|x)\,dv\right| = \\
&= \left|\int_0^1 (1-w)^{\lfloor s\rfloor-1}\int \frac{(hz)^{\lfloor s\rfloor}k(z)}{(\lfloor s\rfloor-1)!}\int_{-\infty}^{x'b}\Big(f^{(\lfloor s\rfloor)}(v+whz|x) - f^{(\lfloor s\rfloor)}(v|x)\Big)\,dv\,dz\,dw\right| \\
&= \left|\int_0^1 (1-w)^{\lfloor s\rfloor-1}\int \frac{(hz)^{\lfloor s\rfloor}k(z)}{(\lfloor s\rfloor-1)!}\Big(f^{(\lfloor s\rfloor-1)}(x'b+whz|x) - f^{(\lfloor s\rfloor-1)}(x'b|x)\Big)\,dz\,dw\right| \\
\text{(2.14)}\quad &= \left|\int_0^1 w(1-w)^{\lfloor s\rfloor-1}\int \frac{(hz)^{\lfloor s\rfloor+1}k(z)}{(\lfloor s\rfloor-1)!}\int_0^1 f^{(\lfloor s\rfloor)}(x'b+twhz|x) - f^{(\lfloor s\rfloor)}(x'b|x)\,dt\,dz\,dw\right| \\
&\le Ch^{s+1},
\end{aligned}$$

by the Hölder condition on $f^{(\lfloor s\rfloor)}$. This implies, under Assumption X, that

$$\left\|R_h^{(1)}(b;\tau) - R^{(1)}(b;\tau)\right\| \le Ch^{s+1}.$$

In order to obtain (*iii*), differentiate $R^{(1)}(b;\tau)$ to get

$$R^{(2)}(b;\tau) = \mathbb{E}[XX'f(X'b|X)] \equiv \int xx'f(x'b|x)\,dF_X(x),$$

and likewise

$$R_h^{(2)}(b;\tau) = \mathbb{E}[XX'k_h(X'b-Y)] \equiv \int xx'\,\mathbb{E}^x\{k_h(x'b-Y)\}\,dF_X(x).$$

Setting $v = x'b$ in (2.12), one obtains from Assumptions X and Q2 that

$$\begin{aligned}\left\|R_h^{(2)}(b;\tau) - R^{(2)}(b;\tau)\right\| &\le C|\mathbb{E}^x\{k_h(v-Y)\} - f(v|x)| \\ &\le Ch^s.\end{aligned}$$

This establishes the stated result.

It remains to show the last item in the Lemma. Observe that

$$R_h^{(2)}(b;\tau) = \mathbb{E}[XX'k_h(X'b - Y)] = \int k(z)\int xx'f(x'b + hz|x)\, dF_X(x)\, dz.$$

From Assumption Q2 and noting that $f(\cdot|\cdot)$ is Lipschitz when $\lfloor s \rfloor \geq 1$, one obtains

$$\left\| R_h^{(2)}(b+\delta;\tau) - R_h^{(2)}(b;\tau)\right\| \leq C\int |k(z)| \int \|xx'\| |x'\delta|\, dF_X(x)\, dz \leq C\|\delta\|,$$

uniformly in b, h, δ and τ. This proves item (iv). □

Proof of Theorem 2.2. Notice that $\beta_h(\tau)$ is well defined, by convexity of $b \mapsto R_h(b;\tau)$. Let us first obtain (2.8). Set

$$A(\tau,h) := \int_0^1 R^{(2)}\Big(\beta(\tau) + w\big(\beta_h(\tau) - \beta(\tau)\big);\, \tau\Big)\, dw.$$

Claim: $A(\tau,h)$ has its eigenvalues bounded away from zero, uniformly in $(\tau,h) \in [\underline{\tau},\bar{\tau}] \times \left[\underline{h},\bar{h}\right]$.

Now since $R_h^{(1)}(\beta_h(\tau);\tau) = R^{(1)}(\beta(\tau);\tau) = 0$, we have via a Taylor expansion with integral remainder,

$$\begin{aligned} R^{(1)}(\beta_h(\tau);\tau) - R_h^{(1)}(\beta_h(\tau);\tau) &= R^{(1)}(\beta_h(\tau);\tau) - R^{(1)}(\beta(\tau);\tau) \\ &= A(\tau,h)\big(\beta_h(\tau) - \beta(\tau)\big), \end{aligned}$$

and therefore the claim implies, together with Lemma 2.1 (ii), that

$$\begin{aligned} \left\| \frac{\beta_h(\tau) - \beta(\tau)}{h^{s+1}} \right\| &\leq \sup\left\{ \left\| A(\tau,h)^{-1} \right\| \cdot \left\| \frac{R^{(1)}(\beta_h(\tau);\tau) - R_h^{(1)}(\beta_h(\tau);\tau)}{h^{s+1}} \right\| \right\} \\ &= O(1), \end{aligned}$$

where the supremum is taken for $(\tau,h) \in [\underline{\tau},\bar{\tau}] \times \left[\underline{h},\bar{h}\right]$. Hence (2.8) holds. Uniqueness of $\beta_h(\tau)$ for small $\bar{h}$ will be obtained in the proof of the above claim.

To prove the claim, put $c := \inf f\Big(x'\big[\beta(\tau) + w\big(\beta_h(\tau) - \beta(\tau)\big)\big]\,\Big|\, x\Big)$, the infimum being taken for (τ,h,w,x) over $[\underline{\tau},\bar{\tau}] \times \left[\underline{h},\bar{h}\right] \times [0,1] \times \operatorname{supp} X$. If the map $(\tau,h) \mapsto \beta_h(\tau)$ is continuous then by Assumption Q1 it holds that $c > 0$ and hence

$$\begin{aligned} v'A(\tau,h)v &= \int_0^1 \int (v'x)^2 f\Big(x'\big[\beta(\tau) + w\big(\beta_h(\tau) - \beta(\tau)\big)\big]\,\Big|\, x\Big)\, dF_X(x)\, dw \\ &\geq c\, v\mathbb{E}XX'v > 0 \end{aligned}$$

by Assumption X, which implies the claim.

It remains to establish continuity of $(\tau,h) \mapsto \beta_h(\tau)$. First observe that $b \mapsto R^{(2)}(b;\tau) \equiv \mathbb{E}XX'f(X'b|X)$ is continuous and does not depend on τ. Moreover, for any $b \in \mathbb{R}^d$,

$R^{(2)}(b;\tau)$ is an element of the open set M_d^+ of positive definite $d \times d$ matrices. Indeed, for $v \in \mathbb{R}^d$,

$$\begin{aligned} v' R^{(2)}(b;\tau) v &= \int (v'x)^2 f(x'b|x)\, dF_X(x) \\ &\geq v'\mathbb{E}XX'v \inf_x f(x'b|x) > 0, \end{aligned}$$

where the infimum is taken over the compact set $\operatorname{supp} X$ and hence is larger than zero by Assumption Q1. Now Lemma 2.1 (*iii*) ensures that for some $C > 0$ one has

$$\left\| R_h^{(2)}(\beta_h(\tau);\tau) - R^{(2)}(\beta_h(\tau);\tau) \right\| \leq C\bar{h}^s$$

uniformly in $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$. Hence for a proper choice of $\bar{h}$ it holds that $R_h^{(2)}(\beta_h(\tau);\tau) \in M_d^+$ uniformly in (τ, h). For such an $\bar{h}$, the first order condition $R_h^{(1)}(\beta_h(\tau);\tau) = 0$ and the Implicit Function Theorem then ensure that $(\tau, h) \mapsto \beta_h(\tau)$ is unique and continuous.

We now prove (2.9). Lemma 2.1 (*iii*) and a Taylor expansion give, since $R_h^{(1)}(\beta_h(\tau);\tau) = 0$,

$$\begin{aligned} -R_h^{(1)}(\beta(\tau);\tau) &= R_h^{(1)}(\beta_h(\tau);\tau) - R_h^{(1)}(\beta(\tau);\tau) \\ &= \Big(R^{(2)}(\beta(\tau);\tau) + O(h^s)\Big)(\beta_h(\tau) - \beta(\tau)) + o\Big(h^{s+1}\Big). \end{aligned}$$

Using

$$f^{(\lfloor s \rfloor)}(x'\beta(\tau) + twhz|x) - f^{(\lfloor s \rfloor)}(x'\beta(\tau)|x) = f^{(s)}(x'\beta(\tau)|x)twhz + o(h),$$

together with $s = \lfloor s \rfloor + 1$ and $\int w^2(1-w)^{\lfloor s \rfloor - 1}\, dw = 2/(\lfloor s \rfloor(\lfloor s \rfloor + 1)(\lfloor s \rfloor + 2))$, we get by (2.14)

$$\begin{aligned} &\int_{-\infty}^{x'\beta(\tau)} \mathbb{E}^x\{k_h(v - Y)\} - f(v|x)\, dv = \\ &= h^{s+1} \int_0^1 w^2(1-w)^{\lfloor s \rfloor - 1}\, dw \int \frac{z^{s+1}}{(\lfloor s \rfloor - 1)!} k(z)\, dz \int_0^1 t\, dt \cdot f^{(s)}(x'\beta(\tau)|x) + h^s o(h) \\ &= h^{s+1} \frac{\int z^{s+1} k(z)\, dz}{(s+1)!} f^{(s)}(x'\beta(\tau)|x) + o\Big(h^{s+1}\Big). \end{aligned}$$

Hence (2.13) gives

$$R_h^{(1)}(\beta(\tau);\tau) = h^{s+1} \frac{\int z^{s+1} k(z)\, dz}{(s+1)!} \int x f^{(s)}(x'\beta(\tau)|x)\, dF_X(x) + o\Big(h^{s+1}\Big),$$

so that (2.9) holds.

□

2.5.2 Bahadur-Kiefer representation

This section makes use of a powerful functional exponential inequality; see Massart (2007, Corollary 6.9). It is of technical interest on its own, so we state (a version of) it. Recall that for real valued functions $\underline{f}$ and $\bar{f}$ with $\underline{f} \leq \bar{f}$, the bracket $\left[\underline{f}, \bar{f}\right]$ is the set of all g such that $\underline{f} \leq g \leq \bar{f}$. For a set $\mathcal{F}$ and a family $\{F_i\}$ of subsets of $\mathcal{F}$, say that $\{F_i\}$ *covers* $\mathcal{F}$ if $\mathcal{F} \subset \bigcup_i F_i$.

Theorem 2.5 (Massart (2007, Corollary 6.9))**.** *Let Z_i be an iid sequence of random variables taking values in the measurable space $\mathcal{Z}$, and let $\mathcal{F}$ be a class of real valued, measurable functions on $\mathcal{Z}$. Assume that*

(i) there are some positive constants σ and M such that for all $f \in \mathcal{F}$, one has

$$\mathbb{E}\Big[|f(Z_i)|^2\Big] \leq \sigma^2 \text{ and } \sup_{z\in\mathcal{Z}}|f(z)| \leq M. \tag{2.15}$$

(ii) for each $\delta > 0$ there exist an integer $J(\delta) \geq 1$ and a set of brackets $\left\{\left[\underline{f}_j, \bar{f}_j\right];\ j = 1, \ldots, J(\delta)\right\}$ covering $\mathcal{F}$ such that for all $j = 1, \ldots, J(\delta)$,

$$\mathbb{E}\left[\left|\bar{f}_j(Z_i) - \underline{f}_j(Z_i)\right|^2\right] \leq \delta^2 \text{ and } \sup_{z\in\mathcal{Z}}\left|\bar{f}_j(z) - \underline{f}_j(z)\right| \leq M. \tag{2.16}$$

Then for any $r \geq 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n}(f(Z_i) - \mathbb{E}[f(Z_i)]) \geq \mathcal{H}_n + 7\sigma\sqrt{2r} + \frac{2Mr}{\sqrt{n}}\right) \leq \exp(-r). \tag{2.17}$$

where

$$\mathcal{H}_n := 27\left(\int_0^\sigma H^{1/2}(u)\,du + \frac{2(\sigma + M)H(\sigma)}{\sqrt{n}}\right),$$

and where H is any non-negative measurable function of $\delta > 0$ satisfying $H(\delta) \geq \log J(\delta)$.

Before proceeding to the proofs, let us introduce some additional definitions, and present a brief heuristic argument. In what follows, whenever we write $\sup_{(\tau,h)}$ we mean the supremum is being taken over $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$, and similarly for infima, unions, etc. Observe that this will depend on n via $\left[\underline{h}, \bar{h}\right]$.

Now recall that $\varrho_n(h)^{-1} := \left(\frac{nh}{\log n}\right)^{1/2} + n^{1/2}$, and let $\mathcal{E}_n(r)$ be the event,

$$\left\{\sqrt{n}\sup_{(\tau,h)} \varrho_n(h)^{-1}\left\|\widehat{\beta}_h(\tau) - \beta_h(\tau) + D_h(\tau)^{-1}\widehat{S}_h(\tau)\right\| \geq r^2\right\}.$$

Notice that $\mathcal{E}_n(r)$ depends on the sample size, n, and on a *tail parameter*, r, but not on τ or h. On the complementary set of $\mathcal{E}_n(r)$, it holds that

$$\sqrt{n}\Big(\widehat{\beta}_h(\tau) - \beta_h(\tau)\Big) = -\sqrt{n}D_h(\tau)^{-1}\widehat{S}_h(\tau) + \widehat{E}_h(\tau),$$

where the approximation error term satisfies $\left\|\widehat{E}_h(\tau)\right\| \leq \varrho_n(h)r^2$ uniformly in $\tau \in [\underline{\tau}, \bar{\tau}]$ and $h \in \left[\underline{h}, \bar{h}\right]$. In particular, if $\mathbb{P}\Big(\mathcal{E}_n(r)\Big)$ is small for large r as established below, then the representation (2.6) from Theorem 2.1 holds uniformly in τ and h.

Define

$$\mathcal{E}_n^1(r) = \left\{\sup_{(\tau,h)}\left\|\sqrt{n}\,\widehat{S}_h(\tau)\right\| \geq r\right\},$$
$$\mathcal{E}_n^2(r) = \left\{\sup_{(\tau,h)}\ \sup_{\{b\colon \|b-\beta_h(\tau)\|\leq 1\}}\left\|\sqrt{\frac{nh}{\log n}}\Big(\widehat{R}_h^{(2)}(b;\tau) - R_h^{(2)}(b;\tau)\Big)\right\| \geq r\right\}.$$

In $\mathcal{E}_n^1(r)$, the norm is the Euclidean norm, while in $\mathcal{E}_n^2(r)$ it is the matrix norm (the trace norm), so that for any matrix A and conformable vector v, $\|Av\| \leq \|A\|\,\|v\|$. E^c stands for the complementary event of E. We now state the functional exponential inequality.

Proposition 2.2. *Given Assumptions X, Q and K, it holds that $\widehat{\beta}_h(\tau)$ is unique for all $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$ with a probability tending to 1. Moreover there exist positive constants C_0 , C_1 and C_2 such that for all ϵ small enough, and all r and n large enough,*

(i) $\mathbb{P}\Big(\mathcal{E}_n(r) \cap \mathcal{E}_n^1(r)^c \cap \mathcal{E}_n^2(r)^c\Big) \leq C_0 \exp(-n\epsilon/C_0)$*;*

(ii) $\mathbb{P}\Big(\mathcal{E}_n^1(r)\Big) \leq C_1 \exp(-r^2/C_1)$*;*

(iii) $\mathbb{P}\Big(\mathcal{E}_n^2(r)\Big) \leq C_2 \exp(-r \log n/C_2)$*.*

The proof of Proposition 2.2, of which Theorem 2.1 is an immediate corollary, relies on Theorem 2.5 via the series of Lemmas below. It is convenient for some proofs to consider the auxiliary objective functions,

$$\widehat{\mathcal{R}}_h(b;\tau) = \widehat{R}_h(b;\tau) - \widehat{R}_h(\beta_h(\tau);\tau), \quad \mathcal{R}_h(b;\tau) := \mathbb{E}\,\widehat{\mathcal{R}}_h(b;\tau),$$

which are such that $\widehat{\beta}_h(\tau) = \arg\min_b \widehat{\mathcal{R}}_h(b;\tau)$ and $\beta_h(\tau) = \arg\min_b \mathcal{R}_h(b;\tau)$. Similarly, set $\widehat{\mathcal{R}}(b;\tau) = \widehat{R}(b;\tau) - \widehat{R}(\beta(\tau);\tau)$ and $\mathcal{R}(b;\tau) := \mathbb{E}\,\widehat{\mathcal{R}}(b;\tau)$. The next Lemma shows that $\widehat{\beta}_h(\tau)$ is close to $\beta_h(\tau)$ uniformly for $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$.

Lemma 2.2 (UCV)**.** *Suppose Assumptions X, Q, and K hold. Then there are some positive constants C_0 and C_1 such that for all n large enough and any $\eta \in [1/\log n, 1]$,*

$$\mathbb{P}\left(\sup_{(\tau,h)}\left\|\widehat{\beta}_h(\tau) - \beta_h(\tau)\right\| \geq \eta\right) \leq C_0 \exp\Big(-n\eta^4/C_1\Big).$$

Proof of Lemma 2.2. We have

$$\left\{\sup_{(\tau,h)}\left\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\right\|\geq 2\eta\right\}=\bigcup_{(\tau,h)}\left\{\left\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\right\|\geq 2\eta\right\}$$
$$\subset\bigcup_{(\tau,h)}\left\{\inf_{b;\|b-\beta_h(\tau)\|\geq 2\eta}\widehat{\mathcal{R}}_h(b;\tau)\leq\inf_{b;\|b-\beta_h(\tau)\|\leq 2\eta}\widehat{\mathcal{R}}_h(b;\tau)\right\}$$
$$\subset\bigcup_{(\tau,h)}\left\{\inf_{b;\|b-\beta_h(\tau)\|\geq 2\eta}\widehat{\mathcal{R}}_h(b;\tau)\leq\widehat{\mathcal{R}}_h(\beta_h(\tau);\tau)\right\}$$
$$=\bigcup_{(\tau,h)}\left\{\inf_{b;\|b-\beta_h(\tau)\|\geq 2\eta}\widehat{\mathcal{R}}_h(b;\tau)\leq 0\right\},$$

since $\widehat{\mathcal{R}}_h(\beta_h(\tau);\tau)=0$. Theorem 2.2 yields

$$\begin{aligned}\{b;\|b-\beta_h(\tau)\|\geq 2\eta\} &\subset \left\{b;\|b-\beta(\tau)\|+\sup_{(\tau,h)}\|\beta_h(\tau)-\beta(\tau)\|\geq 2\eta\right\}\\ &\subset \left\{b;\|b-\beta(\tau)\|+O\left(\overline{h}^{s+1}\right)\geq 2\eta\right\}\\ &\subset \{b;\|b-\beta(\tau)\|\geq\eta\}\end{aligned}$$

for all (τ,h) provided that n is large enough. Hence

$$\left\{\sup_{(\tau,h)}\left\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\right\|\geq 2\eta\right\}\subset\bigcup_{(\tau,h)}\left\{\inf_{b;\|b-\beta(\tau)\|\geq\eta}\widehat{\mathcal{R}}_h(b;\tau)\leq 0\right\}.$$

Now observe that

$$\begin{aligned}\widehat{R}_h(b;\tau)&=\frac{1}{nh}\sum_{i=1}^{n}\int\rho_\tau(t)k\left(\frac{t-(Y_i-X_i'b)}{h}\right)dt\\&=\frac{1}{n}\sum_{i=1}^{n}\int\rho_\tau(Y_i-X_i'b+hz)k(z)\,dz.\end{aligned}$$

Hence, since $\int k(z)\,dz=1$, $\int|zk(z)|\,dz<\infty$ under Assumption K1 and the fact that $t\mapsto\rho_\tau(t)$ is 1-Lipschitz, one has

$$\begin{aligned}\left|\widehat{R}_h(b;\tau)-\widehat{R}(b;\tau)\right|&=\left|\frac{1}{n}\sum_{i=1}^{n}\int[\rho_\tau(Y_i-X_i'b+hz)-\rho_\tau(Y_i-X_i'b)]k(z)\,dz\right|\\&\leq h\int|zk(z)|\,dz,\end{aligned}$$

for all b, τ and h. Hence from Theorem 2.2 and the Lipschitz property of $b\mapsto\widehat{R}(b;\tau)$ it follows that $\widehat{\mathcal{R}}_h(b;\tau)\geq\widehat{\mathcal{R}}(b;\tau)-Ch$ uniformly in b and τ, and therefore

$$\left\{\sup_{(\tau,h)}\left\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\right\|\geq 2\eta\right\}\subset\bigcup_{(\tau,h)}\left\{\inf_{b;\|b-\beta(\tau)\|\geq\eta}\widehat{\mathcal{R}}(b;\tau)\leq Ch\right\}.$$

The next step is a convexity argument which uses the change of variables $b=\beta(\tau)+\rho u$ with $\|u\|=1$ and $\rho\geq\eta$. Since $b\mapsto\widehat{\mathcal{R}}(b;\tau)$ is convex with $\widehat{\mathcal{R}}(\beta(\tau);\tau)=0$

$$\begin{aligned}\frac{\eta}{\rho}\widehat{\mathcal{R}}(\beta(\tau)+\rho u;\tau)&=\frac{\eta}{\rho}\widehat{\mathcal{R}}(\beta(\tau)+\rho u;\tau)+\left(1-\frac{\eta}{\rho}\right)\widehat{\mathcal{R}}(\beta(\tau);\tau)\\&\geq\widehat{\mathcal{R}}(\beta(\tau)+\eta u;\tau).\end{aligned}$$

Hence

$$\begin{aligned}
&\left\{\inf_{b;\|b-\beta(\tau)\|\geq\eta}\widehat{\mathcal{R}}(b;\tau)\leq Ch\right\}\\
&\quad\subset\bigcup_{\rho\in[\eta,\infty)}\left\{\inf_{u;\|u\|=1}\widehat{\mathcal{R}}(\beta(\tau)+\eta u;\tau)\leq C\frac{\eta}{\rho}h\right\}\\
&\quad\subset\left\{\inf_{b;\|b-\beta(\tau)\|=\eta}\widehat{\mathcal{R}}(b;\tau)\leq Ch\right\}.
\end{aligned}$$

and it then follows that

$$\begin{aligned}
&\bigcup_{(\tau,h)}\left\{\left\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\right\|\geq 2\eta\right\}\subset\bigcup_{\tau}\left\{\inf_{\{b:\|b-\beta(\tau)\|=\eta\}}\widehat{\mathcal{R}}(b;\tau)\leq C\bar{h}\right\}\\
&\qquad\subset\left\{\inf_{\tau}\inf_{\{b:\|b-\beta(\tau)\|=\eta\}}\left[\widehat{\mathcal{R}}(b;\tau)-\mathcal{R}(b;\tau)\right]\leq C\bar{h}-\inf_{\tau}\inf_{\{b:\|b-\beta(\tau)\|=\eta\}}\mathcal{R}(b;\tau)\right\}.
\end{aligned}$$

We first give an upper bound for $C\bar{h}-\inf_{\tau\in[\underline{\tau},\bar{\tau}]}\inf_{b:\|b-\beta(\tau)\|=\eta}\mathcal{R}(b;\tau)$. Since the eigenvalues of $\mathcal{R}^{(2)}(b;\tau)$ are bounded away from 0 uniformly with respect to b, with $\|b-\beta(\tau)\|\leq 1$ and $\tau\in[\underline{\tau},\bar{\tau}]$, a second-order Taylor expansion of $\mathcal{R}(b;\tau)=R(b;\tau)-R(\beta(\tau);\tau)$ gives for all b with $\|b-\beta(\tau)\|=\eta$,

$$\begin{aligned}
\mathcal{R}(b;\tau)&=0+\underbrace{R^{(1)}(\beta(\tau),\tau)'}_{=0}(b-\beta(\tau))\\
&\quad+\frac{1}{2}(b-\beta(\tau))'\left[\int_0^1(1-t)\mathcal{R}^{(2)}(\beta(\tau)+t(b-\beta(\tau));\tau)\,dt\right](b-\beta(\tau))\\
&\quad\geq C\eta^2.
\end{aligned}$$

It follows that for any $\eta_2<\eta$, where $\eta_2=\eta-\epsilon_2$ with conformable ϵ_2, and for any $\bar{h}$ small enough,

$$\bigcup_{(\tau,h)}\left\{\left\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\right\|\geq 2\eta\right\}\subset\left\{\sup_{\tau\in[\underline{\tau},\bar{\tau}]}\sup_{\{b:\|b-\beta(\tau)\|=\eta\}}\left|\widehat{\mathcal{R}}(b;\tau)-\mathcal{R}(b;\tau)\right|\geq C\eta_2^2\right\}.$$

Now let $Z_i=(Y_i,X_i')'$ and $\theta=(\tau,b')'$. Then by defining $f(Z_i,\theta)=\rho_\tau(Y_i-X_i'b)-\rho_\tau(Y_i-X_i'\beta(\tau))$ one obtains

$$\widehat{\mathcal{R}}(b;\tau)-\mathcal{R}(b;\tau)=\frac{1}{n}\sum_{i=1}^n\Big(f(Z_i,\theta)-\mathbb{E}f(Z_i,\theta)\Big).$$

Under Assumption X and since $\eta\leq 1$, it holds that for all b with $\|b-\beta(\tau)\|=\eta$ and $\tau\in[\underline{\tau},\bar{\tau}]$

$$|f(Z_i,\theta)|\leq\|X_i\|\,\|b-\beta(\tau)\|\leq C,$$

which also implies that $\mathrm{Var}(f(Z_i,\theta))\leq\sigma^2\leq C$. Observe also that Assumption X, together with the Lipschitz condition on $\tau\mapsto\beta(\tau)$ (Assumption Q1) and on $\tau\mapsto\rho_\tau(u)$, gives for all admissible z,

$$|f(z,\theta_1)-f(z,\theta_2)|\leq C\|\theta_1-\theta_2\|, \tag{2.18}$$

where $\|\theta\|^2 = \|b\|^2 + |\tau|^2$.

Now, for $\delta > 0$, let θ_j, $j = 1, \ldots, J(\delta) \leq C\delta^{-(d+1)}$ be such that

$$\Theta = \{\theta = (b, \tau) : \tau \in [\underline{\tau}, \bar{\tau}], \|b - \beta(\tau)\| = \eta_1\} \subset \bigcup_{j=1}^{J(\delta)} \mathcal{B}(\theta_j, \delta),$$

where $\mathcal{B}(\theta_j, \delta)$ is the $\|\cdot\|$-ball with center θ_j and radius δ. Define $\underline{f}_j(\cdot)$ and $\bar{f}_j(\cdot)$ respectively as

$$\underline{f}_j(z) := \inf_{\theta \in \mathcal{B}(\theta_j, \delta)} f(z, \theta), \qquad \bar{f}_j(z) = \sup_{\theta \in \mathcal{B}(\theta_j, \delta)} f(z, \theta),$$

so that $\{f(\cdot, \theta) : \theta \in \mathcal{B}(\theta_j, \delta)\} \subset \left[\underline{f}_j, \bar{f}_j\right]$ and then $\mathcal{F}_\Theta := \{f(\cdot, \theta) : \theta \in \Theta\} \subset \bigcup_{j=1}^{J(\delta)} \left[\underline{f}_j, \bar{f}_j\right]$. Observe also that (2.18) gives $\left|\bar{f}_j(z) - \underline{f}_j(z)\right| \leq C\delta \leq C$ and $\mathbb{E}\left[\left|\bar{f}_j(Z_i) - \underline{f}_j(Z_i)\right|^2\right] \leq C\delta^2$. Then, since (2.15) and (2.16) hold, setting $H(\delta) = -(d+1)\log\delta + C$ gives, by (2.17),

$$\mathbb{P}\left(\sup_{\theta \in \Theta}\left|\widehat{\mathcal{R}}(b; \tau) - \mathcal{R}(b; \tau)\right| \geq C\frac{1 + \sqrt{r} + {}^{r}/\sqrt{n}}{\sqrt{n}}\right) \leq \exp(-r),$$

and then, for n large enough with respect to η_2^2,

$$\mathbb{P}\left(\sup_{\tau} \sup_{\{b : \|b - \beta(\tau)\| = \eta_1\}} \left|\widehat{\mathcal{R}}(b; \tau) - \mathcal{R}(b; \tau)\right| \geq C\eta_2^2\right) \leq C\exp\left(-Cn\eta_2^4\right).$$

Therefore,

$$\mathbb{P}\left(\sup_{(\tau,h)}\left\|\widehat{\beta}_h(\tau) - \beta_h(\tau)\right\| \geq 2\eta\right) \leq C\exp\left(-Cn\eta_2^4\right),$$

from which the Lemma follows. □

Lemma 2.3. *Suppose Assumptions X, Q, and K hold. Consider $r > 0$ and $\eta \in (0, 1]$. Then, provided that n is large enough,*

$$\mathbb{P}\left(\sup_{(\tau,h)}\left\|\sqrt{n}\widehat{S}_h(\tau)\right\| \geq C_1(1 + r)\right) \leq C_0 \exp\left(-r^2\right),$$

$$\mathbb{P}\left(\sup_{(\tau,h)} \sup_{\{b : \|b - \beta_h(\tau)\| \leq \eta\}} \left\|\sqrt{\frac{nh}{\log n}}\left(\widehat{R}_h^{(2)}(b, \tau) - R_h^{(2)}(b, \tau)\right)\right\| \geq C_1(1 + r)\right) \leq C_0 \exp(-r\log n).$$

Proof of Lemma 2.3. Consider the first deviation probability in the Lemma. Since $R_h^{(1)}(\beta_h(\tau), \tau) = 0$,

$$\sup_{(\tau,h)}\left\|\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau), \tau)\right\| \leq \sup_{(\tau,h)} \sup_{\{b : \|b - \beta_h(\tau)\| \leq \eta\}} \left\|\sqrt{n}\left(\widehat{R}_h^{(1)}(b, \tau) - R_h^{(1)}(b, \tau)\right)\right\|,$$

and it is sufficient to consider the upper bound. Observe that

$$\begin{aligned}\widehat{R}_h^{(1)}(b, \tau) &= \frac{\partial}{\partial b}\left[\frac{1}{n}\sum_{i=1}^{n}\int \rho_\tau(Y_i - X_i'b + hz)k(z)\,dz\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\left[\int \mathbb{I}(Y_i - X_i'b + hz < 0)k(z)\,dz - \tau\right].\end{aligned}$$

Hence, for $\theta = (b', h, \tau)'$ and $Z_i = (Y_i, X_i')'$, one has $\widehat{R}_h^{(1)}(b,\tau) = \sum_{i=1}^n f(Z_i,\theta)/n$ with

$$f(Z_i,\theta) = X_i\left[\int \mathbb{I}(Y_i - X_i'b + hz < 0)k(z)\,dz - \tau\right],$$

for $\theta \in \Theta := \left\{\theta = (b', h, \tau);\quad (\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right],\quad \|b - \beta_h(\tau)\| \leq \eta\right\}$.

It is sufficient to prove the bound for each of the entries of $\widehat{R}_h^{(1)}(b,\tau)$ so that there is no loss of generality when assuming that X_i is of dimension 1. Note that $|f(Z_i,\theta)| \leq C$ and $\mathrm{Var}(f(Z_i,\theta)) \leq \sigma^2 \leq C$, and for all θ_1, θ_2, one has $|f(Z_i,\theta_2) - f(Z_i,\theta_1)| \leq C$. Let $\|\theta\|^2 = \|b\|^2 + |h|^2 + |\tau|^2$, and let $\mathcal{B}(\theta,\delta)$ be the $\|\cdot\|$ ball with center θ and radius δ. Now Assumption X gives, for any θ_1, θ_2 in $\mathcal{B}(\theta,\delta^2)$,

$$|f(Z_i,\theta_2) - f(Z_i,\theta_1)| \leq C\left[\int \mathbb{I}\Big(Y_i - X_i'b + hz \in \left[-C\delta^2, C\delta^2\right]\Big)|k(z)|\,dz + \delta^2\right]. \tag{2.19}$$

Consider a covering of Θ with $J(\delta^2) \leq C(\delta^2)^{-(d+1)}$ balls $\mathcal{B}(\theta_j,\delta^2)$ and define,

$$\underline{f}_j(z) = \inf_{\theta\in\mathcal{B}(\theta_j,\delta)} f(z,\theta), \quad \bar{f}_j(z) = \sup_{\theta\in\mathcal{B}(\theta_j,\delta)} f(z,\theta),$$

so that $\{f(\cdot,\theta) : \theta \in \mathcal{B}(\theta_j,\delta)\} \subset \left[\underline{f}_j, \bar{f}_j\right]$ and then $\mathcal{F}_\Theta := \{f(\cdot,\theta) :\ \theta \in \Theta\} \subset \bigcup_{j=1}^{J(\delta^2)}\left[\underline{f}_j, \bar{f}_j\right]$. (2.19) gives that, uniformly in j and $\delta^2 \leq \sigma^2$,

$$\mathbb{E}\left[\left|\bar{f}_j(Z_i) - \underline{f}_j(Z_i)\right|^2\right] \leq C\left\{\delta^4 + \mathbb{E}\left[\left(\int \mathbb{I}\Big(Y_i - X_i'b + hz \in \left[-C\delta^2, C\delta^2\right]\Big)|k(z)|\,dz\right)^2\right]\right\}.$$

Now, since $\sup f(y|x) < \infty$ by Assumption Q2, $\int|k(z)|\,dz < \infty$ and by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left(\int \mathbb{I}\Big(Y_i - X_i'b + hz \in \left[-C\delta^2, C\delta^2\right]\Big)k(z)\,dz\right)^2\right] \\
&\quad\leq \mathbb{E}\left[\int \mathbb{I}\Big(Y_i - X_i'b + hz \in \left[-C\delta^2, C\delta^2\right]\Big)|k(z)|\,dz\right] \times \int|k(z)|\,dz \\
&\quad\leq \int \mathbb{P}\Big(Y_i \in xb + hz + \left[-C\delta^2, C\delta^2\right]\Big|X_i = x\Big)|k(z)|\,dz \times \int|k(z)|\,dz \\
&\quad\leq C\delta^2.
\end{aligned}$$

It then follows that, uniformly in j and $\delta^2 \leq \sigma^2$,

$$\mathbb{E}\left[\left|\bar{f}_j(Z_i) - \underline{f}_j(Z_i)\right|^2\right] \leq C\left(\delta^4 + \delta^2\right) \leq C\delta^2.$$

It then follows that (2.15) and (2.16) hold with $\log H(\delta) = -2(d+1)\log\delta + C$, so that (2.17) gives

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\sqrt{n}\Big(\widehat{R}_h^{(1)}(b,\tau) - R_h^{(1)}(b,\tau)\Big)\right\| \geq C\left(\sqrt{r} + 1 + \frac{r}{\sqrt{n}}\right)\right) \leq 2\exp(-r),$$

which gives the first bound stated in the Lemma for n large enough.

For the second bound, there is no loss of generality to assume that X_i has dimension 1. Note that $(nh/\log n)^{1/2}\widehat{R}_h^{(2)}(b,\tau) = \sum_{i=1}^n g(Z_i,\theta)/n^{1/2}$, with

$$g(Z_i,\theta) = \left(\frac{1}{h\log n}\right)^{1/2} X_i^2 k\left(\frac{X_i'b - Y_i}{h}\right).$$

Assumptions K2, K1 and X give, uniformly in $\theta \in \Theta$,

$$|g(Z_i,\theta)| \leq C\left(\frac{1}{h\log n}\right)^{1/2} \leq C\frac{O\left(n^{1/2}\right)}{\log^2 n} = M_n/2.$$

It also follows from $\sup f(y|x) < \infty$ by Assumption Q2 that, uniformly in $\theta \in \Theta$,

$$\begin{aligned}
\operatorname{Var}(g(Z_i,\theta)) &\leq C\frac{1}{h\log n}\int\int k\left(\frac{x'b - y}{h}\right) f(y|x)\,dy\ \ dF_X(x) \\
&= \frac{C}{\log n} \times \int\int k(v) f(x'b + hv|x)\,dv\ \ dF_X(x) \\
&\leq \frac{C}{\log n} = \sigma_n^2.
\end{aligned}$$

Assumptions K1 and K2 give that $g(Z_i,\theta)$ is Lipschitz over Θ with a polynomial in n Lipschitz coefficient, that is, for any θ_1, θ_2 in Θ, $|g(Z_i,\theta_2) - g(Z_i,\theta_1)| \leq Cn^C\|\theta_2 - \theta_1\|$. Consider a covering of Θ with $J\left(\delta/n^C\right) \leq C\left(\delta/n^C\right)^{-(d+1)}$ balls $\mathcal{B}\left(\theta_j, \delta/n^C\right)$ and define

$$\underline{g}_j(z) := \inf_{\theta\in\mathcal{B}(\theta_j,\delta)} g(z,\theta), \quad \bar{g}_j(z) := \sup_{\theta\in\mathcal{B}(\theta_j,\delta)} g(z,\theta),$$

so that $\{g(z,\theta), \in \mathcal{B}(\theta_j,\delta)\} \subset \left[\underline{g}_j, \bar{g}_j\right]$ and then $\mathcal{G}_\Theta = \{g(\cdot,\theta) : \theta \in \Theta\} \subset \bigcup_{j=1}^{J\left(\delta/n^C\right)}\left[\underline{g}_j, \bar{g}_j\right]$, with
$\mathbb{E}\left[\left|\bar{g}(Z_i) - \underline{g}(Z_i)\right|^2\right] \leq C\delta^2$. It then follows that (2.15) and (2.16) hold with

$$\log H(\delta) = -2(d+1)(\log\delta - C\log n) + C,$$

so that (2.17) gives, for any $u > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\sqrt{\frac{nh}{\log n}}\left(\widehat{R}_h^{(2)}(b,\tau) - R_h^{(2)}(b,\tau)\right)\right\| \geq C\left(1 + \frac{\sqrt{u}}{\log^{1/2} n} + \frac{u}{\log n}\right)\right) \leq 2\exp(-u).$$

Setting $u^{1/2} = t^{1/2}\log^{1/2} n$ gives the desired result. □

Proof of Proposition 2.2. Let

$$\mathcal{E}_n^3(\epsilon) := \left\{\sup_{(\tau,h)}\left\|\widehat{\beta}_h(\tau) - \beta_h(\tau)\right\| \geq \epsilon^{1/4}\right\},$$

which is such that

$$\mathbb{P}\left(\mathcal{E}_n^3(\epsilon)\right) \leq C\exp(-Cn\epsilon)$$

by Lemma 2.2. The bounds for $\mathbb{P}(\mathcal{E}_n^1(r))$ and $\mathbb{P}(\mathcal{E}_n^2(r))$ follow from Lemma 2.3. In particular, $\lim_{n\to\infty}\mathbb{P}(\mathcal{E}_n^2(r)) = 0$ and Lemma 2.1 ensure, under Assumption X, that $b \mapsto \widehat{R}_h(b;\tau)$ is strictly convex for b in a vicinity of $\beta_h(\tau)$, for all τ in $[\underline{\tau},\bar{\tau}]$ with at least $1-(\mathbb{P}(\mathcal{E}_n^1(r))+\mathbb{P}(\mathcal{E}_n^2(r)))$. But Lemma 2.2 and Theorem 2.2 ensure that all minimizers of $\widehat{R}_h(b;\tau)$ lie in such a vicinity with a probability tending to 1. Since $1-(\mathbb{P}(\mathcal{E}_n^1(r))+\mathbb{P}(\mathcal{E}_n^2(r)))$ can be made arbitrarily close to 1 by increasing r, it follows that $\widehat{\beta}_h(\tau)$ is unique with a probability tending to 1 when n increases. It also follows that when $\mathcal{E}_n^1(r)^c$, $\mathcal{E}_n^2(r)^c$ and $\mathcal{E}_n^3(\epsilon)^c$ are all true and n is large enough, $\widehat{\beta}_h(\tau)$ satisfies the first-order condition,

$$\widehat{R}_h^{(1)}\big(\widehat{\beta}_h(\tau);\tau\big)=0.$$

It then follows that

$$\begin{aligned}-\widehat{R}_h^{(1)}(\beta_h(\tau);\tau) &= \widehat{R}_h^{(1)}\big(\widehat{\beta}_h(\tau);\tau\big)-\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\\ &=\Big[\int_0^1 \widehat{R}_h^{(2)}\Big(\beta_h(\tau)+t\big(\widehat{\beta}_h(\tau)-\beta_h(\tau)\big);\tau\Big)\,dt\Big]\Big[\widehat{\beta}_h(\tau)-\beta_h(\tau)\Big].\end{aligned}$$

Now, if ϵ in $\mathcal{E}_n^3(\epsilon)$ is small enough, the eigenvalues of the above matrix are in $[1/C, C]$ for a large C when n is large enough, uniformly in τ and h. Hence,

$$\widehat{\beta}_h(\tau)-\beta_h(\tau)=-\Big[\int_0^1 \widehat{R}_h^{(2)}\Big(\beta_h(\tau)+u\big(\widehat{\beta}_h(\tau)-\beta_h(\tau)\big);\tau\Big)\,du\Big]^{-1}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau).$$

This gives, on $\mathcal{E}_n^1(r)^c$ and $\mathcal{E}_n^2(r)^c$ and by Lemma 2.1,

$$\begin{aligned}&\Big\|\sqrt{n}\big(\widehat{\beta}_h(\tau)-\beta_h(\tau)\big)+\Big[R_h^{(2)}(\beta_h(\tau);\tau)\Big]^{-1}\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\Big\|\\ &\quad\le C\Big\|\int_0^1\Big(\widehat{R}_h^{(2)}\Big(\beta_h(\tau)+u\big(\widehat{\beta}_h(\tau)-\beta_h(\tau)\big);\tau\Big)-R_h^{(2)}\Big(\beta_h(\tau)+u\big(\widehat{\beta}_h(\tau)-\beta_h(\tau)\big);\tau\Big)\Big)\,du\Big\|\\ &\qquad\qquad\times\Big\|\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\Big\|\\ &\quad+C\Big\|\int_0^1\Big(R_h^{(2)}\Big(\beta_h(\tau)+u\big(\widehat{\beta}_h(\tau)-\beta_h(\tau)\big);\tau\Big)-R_h^{(2)}(\beta_h(\tau);\tau)\Big)\,du\Big\|\Big\|\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\Big\|\\ &\quad\le C\left\{C^2\left(\frac{\log n}{nh}\right)^{1/2}r^2+C\underbrace{\Big\|\widehat{\beta}_h(\tau)-\beta_h(\tau)\Big\|\cdot\Big\|\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\Big\|}_{\le Cn^{-1/2}\left\|\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\right\|^2}\right\}\\ &\quad\le C\left\{\left(\frac{\log n}{nh}\right)^{1/2}+n^{-1/2}\right\}r^2,\end{aligned}$$

which shows that $\mathcal{E}_n(r)^c$ holds provided C is taken large enough. □

Proof of Theorem 2.1. Uniqueness has already been established. Now observe that $\mathbb{P}(\mathcal{E}_n(r)) \le \mathbb{P}(\mathcal{E}_n(r)\cap[\mathcal{E}_n^1(r)^c\cap\mathcal{E}_n^2(r)^c])+\mathbb{P}(\mathcal{E}_n^1(r))+\mathbb{P}(\mathcal{E}_n^2(r))$ which can be made arbitrarily small for large n by fixing ϵ and increasing r. □

2.5.3 Asymptotic variance and mean squared error

Proof of Theorem 2.3. We first show that the expansion

$$\operatorname{Var}\Big(\sqrt{n}D(\tau)^{-1}\widehat{S}_h(\tau)\Big) = \Sigma(\tau) - c_k h D(\tau)^{-1} + O\Big(h^2\Big) \tag{2.20}$$

holds uniformly with respect to $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \Big[\underline{h}, \bar{h}\Big]$. Indeed, since $\mathbb{E}\Big[\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\Big] = 0$, we have

$$\begin{aligned}
&\operatorname{Var}\Big(\sqrt{n}\widehat{S}_h(\tau)\Big) = \operatorname{Var}\Big(\sqrt{n}\widehat{R}_h^{(1)}(\beta_h(\tau);\tau)\Big) \\
&= \operatorname{Var}\left(X\left\{K\left(\frac{X'\beta_h(\tau) - Y}{h}\right) - \tau\right\}\right) = \mathbb{E}\left[XX'\left(K\left(\frac{X'\beta_h(\tau) - Y}{h}\right) - \tau\right)^2\right] \\
&= \mathbb{E}\left[XX'\left(K\left(\frac{X'\beta_h(\tau) - Y}{h}\right)\right)^2\right] - 2\tau\mathbb{E}\left[XX'K\left(\frac{X'\beta_h(\tau) - Y}{h}\right)\right] + \tau^2\mathbb{E}[XX'].
\end{aligned}$$

Assumptions Q2 and K give, integrating by parts and using Theorem 2.2,

$$\begin{aligned}
&\mathbb{E}\left[K\left(\frac{X'\beta_h(\tau) - Y}{h}\right)\middle| X = x\right] \\
&= \int K\left(\frac{x'\beta_h(\tau) - y}{h}\right) f(y|x)\,dy = \int \frac{1}{h} k\left(\frac{x'\beta_h(\tau) - y}{h}\right) F(y|x)\,dy \\
&= F(x'\beta_h(\tau)|x) + \int (F(x'\beta_h(\tau) - hz|x) - F(x'\beta_h(\tau)|x)) k(z)\,dz \\
&= F(x'\beta(\tau)|x) + O\Big(h^{s+1}\Big) + O\Big(h^{s+1}\Big) = \tau + O\Big(h^{s+1}\Big),
\end{aligned} \tag{2.21}$$

since $x'\beta(\tau) = F^{-1}(\tau|x)$ and arguing as in Lemma 2.1. For the term involving $K(\cdot)^2$, define

$$\mathbf{K}(z) = 2k(z)K(z) = \frac{d}{dz}\Big[K(z)^2\Big],$$

which is such that

$$\int \mathbf{K}(z)\,dz = \lim_{z \to +\infty} K(z)^2 = 1.$$

Arguing as above now gives

$$\begin{aligned}
&\mathbb{E}\left[K\left(\frac{X'\beta_h(\tau) - Y}{h}\right)^2 \middle| X = x\right] \\
&= \int \frac{1}{h}\mathbf{K}\left(\frac{x'\beta_h(\tau) - y}{h}\right) F(y|x)\,dy \\
&= \tau + O\Big(h^{s+1}\Big) + \int (F(x'\beta_h(\tau) - hz|x) - F(x'\beta_h(\tau)|x))\mathbf{K}(z)\,dz \\
&= \tau - h(f(x'\beta_h(\tau)|x) + O(h)) \int z\mathbf{K}(z)\,dz + O\Big(h^{s+1}\Big) \\
&= \tau - h\Big(f(x'\beta(\tau)|x) + O\Big(h^{(s+1)}\Big) + O(h)\Big) \int z\mathbf{K}(z)\,dz + O\Big(h^{s+1}\Big) \\
&= \tau - hf(x'\beta(\tau)|x) \int z\mathbf{K}(z)\,dz + O\Big(h^2\Big).
\end{aligned} \tag{2.22}$$

Substituting gives the variance expansion since $K(-z) = 1 - K(z)$,

$$\begin{aligned}\int z\mathbf{K}(z)\,dz = 2\int zk(z)K(z)\,dz &= \int_{-\infty}^{0} zd\Big[K(z)^2\Big] + \int_{0}^{+\infty} zd\Big[(K(z))^2 - 1\Big] \\ &= -\int_{-\infty}^{0} K(z)^2\,dz + \int_{0}^{+\infty}\Big(1 - (K(z))^2\Big)\,dz \\ &= \int_{0}^{+\infty}\Big(-(1-K(z))^2 + 1 - (K(z))^2\Big)\,dz = 2\int_{0}^{+\infty} K(z)(1-K(z))\,dz.\end{aligned}$$

We now establish (2.10). The expansion in (2.20), Lemma 2.1, Theorem 2.2, and the Locally Lipschitz property of matrix inversion ensure that

$$\begin{aligned}&\operatorname{Var}\Big(\sqrt{n}D_h(\tau)^{-1}\widehat{S}_h(\tau)D_h(\tau)^{-1}\Big) \\ &= \Sigma(\tau) - c_k hD(\tau)^{-1} + O\Big(h^2\Big) \\ &\quad + \Big\{D_h(\tau)^{-1}\operatorname{Var}\Big(\sqrt{n}\widehat{S}_h(\tau)\Big)D_h(\tau)^{-1} - D(\tau)^{-1}\operatorname{Var}\Big(\sqrt{n}\widehat{S}_h(\tau)\Big)D(\tau)^{-1}\Big\},\end{aligned}$$

with

$$\begin{aligned}&\Big\|D_h(\tau)^{-1}\operatorname{Var}\Big(\sqrt{n}\widehat{S}_h(\tau)\Big)D_h(\tau)^{-1} - D(\tau)^{-1}\operatorname{Var}\Big(\sqrt{n}\widehat{S}_h(\tau)\Big)D(\tau)^{-1}\Big\| \\ &\leq \Big\|D_h(\tau)^{-1} - D(\tau)^{-1}\Big\| \cdot \Big\|\operatorname{Var}\Big(\sqrt{n}\widehat{S}_h(\tau)\Big)\Big\| \cdot \Big(\Big\|D_h(\tau)^{-1}\Big\| + \Big\|D(\tau)^{-1}\Big\|\Big) \\ &\leq C\Big\|D_h(\tau)^{-1} - D(\tau)^{-1}\Big\| = O(h^s),\end{aligned}$$

which yields (2.10). □

Proof of Theorem 2.4. By hypothesis, (2.9) holds. The variance expansion (2.10) and $\mathbb{E}\widehat{S}_h(\tau) = 0$ thus give

$$\text{AMSE}\Big(\lambda'\widehat{\beta}_h(\tau)\Big) = h^{2s+2}(\lambda'\mathbf{B}(\tau))^2 + \frac{1}{n}\lambda'\Big(\Sigma(\tau) - c_k hD(\tau)^{-1}\Big)\lambda + O\left(\frac{h^{s\wedge 2}}{n}\right) + o\Big(h^{2s+2}\Big).$$

Writing $g(h) = h^{2s+2}(\lambda'\mathbf{B}(\tau))^2 - n^{-1}c_k h\lambda' D(\tau)^{-1}\lambda$ and differentiating yields

$$g'(h) = (2s+2)h^{2s+1}(\lambda'\mathbf{B}(\tau))^2 - \frac{1}{n}c_k\lambda' D(\tau)^{-1}\lambda,$$

and thus solving $g'(h) = 0$ gives the desired h^* and the corresponding AMSE expansion. □

2.5.4 Asymptotic covariance estimator

Lemma 2.4. *Given Assumptions X, Q and K, the following holds uniformly with respect to* $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \Big[\underline{h}, \bar{h}\Big]$,

(i) $\Big\|\widehat{D}_h(\tau) - D(\tau)\Big\| = O_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{nh}} + \frac{1}{\sqrt{n}} + h^s\Big)$;

(ii) $\Big\|\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}\Big\| = O_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{nh}} + \frac{1}{\sqrt{n}} + h^s\Big)$;

(iii) $\left\|\widehat{V}_h(\tau) - V(\tau)\right\| = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh}} + \frac{1}{h\sqrt{n}} + h\right)$.

Proof of Lemma 2.4. For item (i), we have

$$\left\|\widehat{D}_h(\tau) - D(\tau)\right\| \le \left\|\widehat{R}_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right) - R_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right)\right\| + \left\|R_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right) - R^{(2)}(\beta(\tau);\tau)\right\|$$

Now the first term in the sum in the above sum is $O_{\mathbb{P}}\left(\sqrt{\log n/(nh)}\right)$ uniformly for $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$, by Lemma 2.3. For the second term in the sum, we have

$$\left\|R_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right) - R^{(2)}(\beta(\tau);\tau)\right\| \le$$

$$\le \left\|R_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right) - R_h^{(2)}(\beta(\tau);\tau)\right\| \tag{2.23}$$

$$+ \left\|R_h^{(2)}(\beta(\tau);\tau) - R^{(2)}(\beta(\tau);\tau)\right\|. \tag{2.24}$$

The term (2.24) is $O(h^s)$ by Lemma 2.1. Regarding (2.23), we have by Lemma 2.1 and Theorems 2.2 and 2.1

$$\begin{aligned}\left\|R_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right) - R_h^{(2)}(\beta(\tau);\tau)\right\| &\le C\left\|\widehat{\beta}_h(\tau) - \beta(\tau)\right\| \\ &\le C\left(\left\|\widehat{\beta}_h(\tau) - \beta_h(\tau)\right\| + \left\|\beta_h(\tau) - \beta(\tau)\right\|\right) \\ &= O_{\mathbb{P}}\left(n^{-1/2} + h^{s+1}\right).\end{aligned}$$

This yields the stated result.

Item (ii) is just (i) and the locally Lipschitz property of matrix inversion.

For item (iii), define

$$\begin{aligned}W(b;\tau) &:= \mathbb{E}XX'(\mathbb{I}[Y - X'b \le 0] - \tau)^2 \\ \widehat{W}_h(b;\tau) &:= \frac{1}{n}\sum_{i=1}^n X_iX_i'\left[K\left(\frac{X_i'b - Y_i}{h}\right) - \tau\right]^2 \\ W_h(b;\tau) &:= \mathbb{E}\widehat{W}_h(b;\tau),\end{aligned}$$

so $W(\beta(\tau);\tau) = V(\tau)$ and $\widehat{W}_h\left(\widehat{\beta}_h(\tau);\tau\right) = \widehat{V}_h(\tau)$. Now

$$\left\|\widehat{V}_h(\tau) - V(\tau)\right\| = \left\|\widehat{W}_h\left(\widehat{\beta}_h(\tau);\tau\right) - W\left(\beta(\tau);\tau\right)\right\|$$

$$\le \left\|\widehat{W}_h\left(\widehat{\beta}_h(\tau);\tau\right) - W_h\left(\widehat{\beta}_h(\tau);\tau\right)\right\| \tag{2.25}$$

$$+ \left\|W_h\left(\widehat{\beta}_h(\tau);\tau\right) - W_h\left(\beta_h(\tau);\tau\right)\right\| \tag{2.26}$$

$$+ \left\|W_h\left(\beta_h(\tau);\tau\right) - W\left(\beta(\tau);\tau\right)\right\|. \tag{2.27}$$

For (2.27), we have by Assumption X and equations (2.21) and (2.22),

$$\begin{aligned}\left\|W_h(\beta_h(\tau);\tau) - W(\beta(\tau);\tau)\right\| &\le C\left|\mathbb{E}\left\{\left[K\left(\frac{X'\beta_h(\tau) - Y}{h}\right) - \tau\right]^2\right\} - \tau(1-\tau)\right| \\ &= O(h).\end{aligned}$$

For (2.26), since both K and K^2 are Lipschitz, we have that $\widehat{W}_h(b;\tau)$ and $W_h(b;\tau)$ are also Lipschitz, with a Lipschitz constant which is $O(h^{-1})$. Thus Assumption X and Theorem 2.1 give

$$\left\|W_h\left(\widehat{\beta}_h(\tau);\tau\right) - W_h\left(\beta_h(\tau);\tau\right)\right\| \leq \frac{C}{h}\left\|\widehat{\beta}_h(\tau) - \beta_h(\tau)\right\| = O_{\mathbb{P}}\left(n^{-1/2}h^{-1}\right).$$

For (2.25), an argument similar to the proof of Lemma 2.3 yields the desired rate.

□

Proof of Proposition 2.1. We have

$$\begin{aligned}
&\left\|\widehat{\Sigma}_h(\tau) - \Sigma(\tau)\right\| \\
&\leq \left\|\widehat{D}_h(\tau)^{-1}\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1}\right\| \\
&\quad + \left\|D(\tau)^{-1}\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}V(\tau)D(\tau)^{-1}\right\| \\
&\leq \left\|\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}\right\| \cdot \left\|\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1}\right\| \\
&\quad + \left\|D(\tau)^{-1}\right\| \cdot \left\|\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1} - V(\tau)D(\tau)^{-1}\right\| \\
&\leq \left\|\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}\right\| \cdot \left\|\widehat{V}_h(\tau)\widehat{D}_h(\tau)^{-1}\right\| \\
&\quad + \left\|D(\tau)^{-1}\right\| \cdot \left(\left\|\widehat{D}_h(\tau)^{-1}\right\| \cdot \left\|\widehat{V}_h(\tau) - V(\tau)\right\| + \left\|V(\tau)\right\| \cdot \left\|\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}\right\|\right).
\end{aligned}$$

Now since the terms $\left\|D(\tau)^{-1}\right\|$ and $\left\|V(\tau)\right\|$ are $O(1)$ uniformly for $\tau \in [\underline{\tau}, \bar{\tau}]$, and the terms $\left\|\widehat{V}_h(\tau)\right\|$ and $\left\|\widehat{D}_h(\tau)^{-1}\right\|$ are $O_{\mathbb{P}}(1)$ uniformly for $(\tau, h) \in [\underline{\tau}, \bar{\tau}] \times \left[\underline{h}, \bar{h}\right]$, we get

$$\left\|\widehat{\Sigma}_h(\tau) - \Sigma(\tau)\right\| = O_{\mathbb{P}}\left(\left\|\widehat{D}_h(\tau)^{-1} - D(\tau)^{-1}\right\| + \left\|\widehat{V}_h(\tau) - V(\tau)\right\|\right),$$

and the result follows from Lemma 2.4. □

2.6 References

Azzalini, A. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328, 1981.

Buchinsky, M. Estimating the asymptotic covariance matrix for quantile regression models a monte carlo study. *Journal of Econometrics*, 68(2):303–338, 1995.

Buchinsky, M. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources*, 33(1):88–126, 1998.

Chaudhuri, P., et al. Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, 19(2):760–777, 1991.

Chernozhukov, V., and Hong, H. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

Cheung, K., and Lee, S. M. Bootstrap variance estimation for Nadaraya quantile estimator. *Test*, 19(1):131–145, 2010.

FALK, M. Relative deficiency of kernel type estimators of quantiles. *The Annals of Statistics*, 12(1):261–268, 1984.

FAN, Y., AND LIU, R. A direct approach to inference in nonparametric and semiparametric quantile regression models. *Preprint*, 2013.

GOH, S. C., AND KNIGHT, K. Nonstandard quantile-regression inference. *Econometric Theory*, 25(5):1415–1432, 2009.

GUERRE, E., AND SABBAH, C. Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory*, 28(1): 87–129, 2012.

HE, X., AND SHAO, Q.-M. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24 (6):2608–2630, 1996.

HOROWITZ, J. L. Bootstrap methods for median regression models. *Econometrica*, 66 (6):1327–1351, 1998.

JUREČKOVÁ, J., SEN, P. K., AND PICEK, J. *Methodology in Robust and Nonparametric Statistics*. CRC Press, 2012.

KAPLAN, D. M., AND SUN, Y. Smoothed estimating equations for instrumental variables quantile regression. 2012.

KNIGHT, K. Comparing conditional quantile estimators: first and second order considerations. Technical report, 2001.

KOENKER, R. Confidence intervals for regression quantiles. In *Asymptotic statistics*, pages 349–359. Springer, 1994.

KOENKER, R. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *Journal of Econometrics*, 95(2):347–374, 2000.

KOENKER, R. *Quantile regression*. Number 38. Cambridge university press, 2005.

KOENKER, R., AND BASSETT, G. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

KOENKER, R., AND HALLOCK, K. Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

KOENKER, R., AND PORTNOY, S. L-estimation for linear models. *Journal of the American Statistical Association*, 82(399):851–857, 1987.

KONG, E., LINTON, O., AND XIA, Y. Global Bahadur representation for nonparametric censored regression quantiles and its applications. *Econometric Theory*, 29(5):941–968, 2013.

KOZEK, A. S. How to combine M-estimators to estimate quantiles and a score function. *Sankhyā: The Indian Journal of Statistics*, 67(2):277–294, 2005.

MACHADO, J. A., AND PARENTE, P. Bootstrap estimation of covariance matrices via the percentile method. *The Econometrics Journal*, 8(1):70–78, 2005.

MACK, Y. Bahadur's representation of sample quantiles based on smoothed estimates of a distribution function. *Probabability and Mathemathical Statistics*, 8:183–189, 1987.

Mammen, E., Van Keilegom, I., and Yu, K. Expansion for moments of regression quantiles with application to nonparametric testing. *arXiv preprint arXiv:1306.6179*, 2013.

Massart, P. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

Mehra, K., Rao, M. S., and Upadrasta, S. A smooth conditional quantile estimator and related applications of conditional empirical processes. *Journal of Multivariate Analysis*, 37(2):151–179, 1991.

Nadaraya, E. A. Some new estimates for distribution functions. *Theory of Probability & its Applications*, 9(3):497–500, 1964.

Newey, W. K., and McFadden, D. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4(4):2111–2245, 1994.

Otsu, T. Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics*, 142(1):508–538, 2008.

Parzen, E. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365):105–121, 1979.

Portnoy, S. Nearly root-n approximation for regression quantile processes. *The Annals of Statistics*, 40(3):1714–1736, 2012.

Ralescu, S. S. A Bahadur–Kiefer law for the Nadaraya empiric-quantile processes. *Theory of Probability & its Applications*, 41(2):296–306, 1997.

Samanta, M. Non-parametric estimation of conditional quantiles. *Statistics & Probability Letters*, 7(5):407–412, 1989.

Sheather, S. J., and Marron, J. S. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416, 1990.

Silverman, B. W. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Stute, W. Conditional empirical processes. *The Annals of Statistics*, 14(2):638–647, 1986.

Whang, Y.-J. Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory*, 22(2):173–205, 2006.

Xiang, X. Bahadur representation of kernel quantile estimators. *Scandinavian Journal of Statistics*, 21(2):169–178, 1994.

3 CONJUGATE PROCESSES

Eduardo Horta[1] Flávio Ziegelmann[2]

November, 2015

Abstract. In this paper we provide a general approach for construction of stochastic processes driven by a second, measure-valued stochastic process. The theory allows for a rich set of examples, and includes a class of Regime Switching models. Our construction also provides a rigorous formalism for Bayesian inference.

Keywords and phrases. Random measure. Disintegration. Conditional distributions. Stochastic Processes.

JEL Classification. C1, C14, C22

3.1 Preliminaries

The concept of random measure dates back at least to Kallenberg (1973), although arguably its roots can be traced back to the theory of point processes (see Kallenberg (1974)). Random measures are an important tool in Probability Theory, being straightly linked to the notions of disintegration of measures and regular conditional probabilities – see for instance Kallenberg (1988), Chang and Pollard (1997), Pollard (2002) and Kallenberg (2006). An account on the theory of disintegration of measures can be found in Pachl (1978) and Faden (1985).

In this paper we provide a general approach for construction of stochastic processes driven by a second, measure-valued stochastic process. Our main result is an existence Theorem which states that given a measure-valued stochastic process and an appropriate compatible family of probability measures, one can construct a probability space where a conjugation property holds. The theory allows for a rich set of examples, including a class of Regime Switching models, and provides a rigorous formalism for Bayesian inference. It also yields a theoretical framework for the approach taken in Horta and Ziegelmann (2015b).

3.2 Introduction and main results

If T is a set and $(S, \mathfrak{S})$ is a measurable space, we write $(\mathscr{F}, \mathfrak{F}) \leq (S^T, \pi)$ to mean that $\mathscr{F} \subset S^T$ and that $\mathfrak{F}$ is a σ-algebra on $\mathscr{F}$ for which the coordinate projections $\pi_t : f \mapsto f(t)$ are $\mathfrak{F} \setminus \mathfrak{S}$ measurable. We will be loose on notation and use the symbol π to denote a general projection from a function space into a coordinate; the index together with the

[1]Department of Statistics – Universidade Federal do Rio Grande do Sul. eduardo.horta@ufrgs.br
[2]Department of Statistics – Universidade Federal do Rio Grande do Sul. flavioz@ufrgs.br

argument should clarify the intended interpretation. We let $M(\mathfrak{S})$ denote the set of all probability measures on the measurable space $(S, \mathfrak{S})$

Let $(S, \mathfrak{S})$ be a state-space, and T be an index set, and consider a 'space of sample paths' $(\mathscr{F}, \mathfrak{F}) \leq (S^T, \pi)$. Say that a family $(P_\lambda : \lambda \in \mathscr{L})$ of probability measures on $(\mathscr{F}, \mathfrak{F})$ is *$\mathscr{L}$-coherent* iff

(*i*) $\mathscr{L} \subset M(\mathfrak{S})^T$;

(*ii*) $P_\lambda \circ \pi_t^{-1} = \lambda_t$, for each $\lambda = (\lambda_t : t \in T) \in \mathscr{L}$ and each $t \in T$.

The family (P_λ) is to be thought of as a 'construction rule' for measures on $\mathfrak{F}$ where one 'knows what to do' provided the marginals $\lambda = (\lambda_t : t \in T) \in \mathscr{L}$ are specified.

Example 3.1 (Product Measure). Let $\mathscr{L} = M(\mathfrak{S})^T$ and $P_\lambda = \bigotimes_{t \in T} \lambda_t$. //

Example 3.2 (Copulas). Let $T = \{1, 2\}$, $S = \mathbb{R}$ and $\mathscr{L} = M(\mathfrak{S}) \times M(\mathfrak{S})$. Let C be a bivariate copula function. For $\lambda = (\lambda_1, \lambda_2) \in \mathscr{L}$, let F_i be the distribution function on $\mathbb{R}$ corresponding to λ_i. Let P_λ be the probability measure on $\mathscr{F} = \mathbb{R}^2$ corresponding to the distribution function $H^\lambda(x, y) = C(F_1(x),\ F_2(y))$. //

Example 3.3 (Copulas, again). Let $T = \mathbb{N}$, $S = \mathbb{R}$ and $\mathscr{F} = S^{\mathbb{N}}$. Consider a family $(C_n : n \in \mathbb{N})$ of n-variate copulas $(n \in \mathbb{N})$ satisfying the compatibility condition

$$C_{n+1}(u_1, \ldots, u_n, 1) = C_n(u_1, \ldots, u_n).$$

Thus the C_n are uniquely associated to a consistent family of finite dimensional distributions of a real valued, discrete time stochastic process having Uniform(0,1) marginals. Now let $\mathscr{L} = M(\mathfrak{S})^{\mathbb{N}}$ and write each $\lambda \in \mathscr{L}$ as $\lambda = (F_1, F_2, \ldots)$ via identification of a measure λ_i with its corresponding distribution function F_i. Write

$$H_n^\lambda(x_1, \ldots, x_n) := C_n(F_1(x_1), \ldots, F_n(x_n)).$$

If each C_n is continuous, then the collection $\left(H_n^\lambda : n \in \mathbb{N}\right)$ determines a consistent family of finite dimensional distributions and thus the Daniell-Kolmogorov Theorem yields a unique probability measure P_λ on the product space $\left(\mathbb{R}^{\mathbb{N}}, \mathfrak{S}^{\mathbb{N}}\right)$, having said finite dimensional distributions. In the probability space $\left(\mathbb{R}^{\mathbb{N}}, \mathfrak{S}^{\mathbb{N}}, P_\lambda\right)$ the random variable π_t has marginal distribution F_t, whereas the dependence structure of the process $(\pi_t : t \in \mathbb{N})$ is mostly determined by the collection $(C_n : n \in \mathbb{N})$. Here $\mathfrak{S}$ is the Borel σ-field on $\mathbb{R}$. Notice that this example includes the product measure (independent sequences). //

The notion of $\mathscr{L}$-coherence is thus nonempty and includes at least two important examples.

We now want to use this 'construction rule' when the marginals are selected at random. In this direction, let J be some set, and let $(T_\alpha : \alpha \in J)$ be a partition of T (so $\operatorname{card} T \geq \operatorname{card} J$). Consider yet another 'space of sample paths' $(\mathscr{M}, \mathfrak{M}) \leq \left(M(\mathfrak{S})^J, \pi\right)$, this time a space of sample paths of measures, and let Q be a probability measure on $\mathfrak{M}$. We assume throughout that $M(\mathfrak{S})$ is endowed with a σ-algebra satisfying the requirement that $\mu_0 \mapsto \mu_0(A)$ is measurable for each $A \in \mathfrak{S}$. For $\mu = (\mu_\alpha : \alpha \in J) \in \mathscr{M}$, define $\rho : \mathscr{M} \to M(\mathfrak{S})^T$ by

$$\rho(\mu)_t = \mu_\alpha, \qquad t \in T_\alpha, \quad \alpha \in J.$$

Therefore ρ is an embedding of $\mathscr{M}$ into $M(\mathfrak{S})^T$ whose range is composed by some piecewise constant maps from T to $M(\mathfrak{S})$ (the maps are constant over each T_α). In this setting say that an $\mathscr{L}$-coherent family $(P_\lambda : \lambda \in \mathscr{L})$ of probability measures on $(\mathscr{F}, \mathfrak{F})$ is *compatible with* Q iff there is a Q-null set N such that $\rho(\mathscr{M} \setminus N) \subset \mathscr{L}$. Compatibility thus means that, if one selects a family $\mu = (\mu_\alpha : \alpha \in J)$ from $\mathscr{M}$ according to Q, then one can construct a probability measure on $(\mathscr{F}, \mathfrak{F})$ using $P_{\rho(\mu)}$ and such that the random variable $\pi_t : \mathscr{F} \to S$ has, conditional on the previous selection, marginal distribution μ_α for $t \in T_\alpha$.

Theorem 3.1. *Let $(P_\lambda : \lambda \in \mathscr{L})$ be an $\mathscr{L}$-coherent family of probability measures on $(\mathscr{F}, \mathfrak{F})$, compatible with the probability measure Q on $(\mathscr{M}, \mathfrak{M})$. Assume that, for each $E \in \mathfrak{F}$, the map $\mu \mapsto P_{\rho(\mu)}(E)$ from $\mathscr{M} \setminus N$ to $\mathbb{R}$ is Q-measurable. Then there exist a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and stochastic processes $(\xi_\alpha : \alpha \in J)$ and $(X_t : t \in T)$ with sample paths lying almost surely respectively in $(\mathscr{M}, \mathfrak{M})$ and $(\mathscr{F}, \mathfrak{F})$, and such that, for $E \in \mathfrak{F}$,*

$$\mathbb{P}[X \in E \,|\, \xi] = P_{\rho \circ \xi}(E). \tag{3.1}$$

Moreover, the process $(\xi_\alpha : \alpha \in J)$ has unconditional distribution Q.

Two processes $(\xi_\alpha : \alpha \in J)$ and $(X_t : t \in T)$, defined on some probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and satisfying the conclusions in Theorem 3.1 are said to be *conjugate*. Equivalently, the pair

$$(\xi_\alpha : \alpha \in J,\, X_t : t \in T)$$

is called a *conjugate process*.

Corollary 3.1. *Assume the conditions of Theorem 3.1 hold. Let $A \in \mathfrak{S}$. If $\mathfrak{B}$ is any σ-algebra on Ω satisfying $\sigma(\xi_\alpha(A)) \subset \mathfrak{B} \subset \sigma(\xi)$, then*

$$\mathbb{P}[X_t \in A \,|\, \mathfrak{B}] = \xi_\alpha(A), \qquad t \in T_\alpha, \quad \alpha \in J. \tag{3.2}$$

In particular, $\mathbb{P}[X_t \in A \,|\, \xi_\alpha] = \mathbb{P}[X_t \in A \,|\, \xi] = \xi_\alpha(A)$.

Thus, if $t \in T_\alpha$, the conditional marginal of X_t depends on ξ through ξ_α only.

3.3 Further examples

In this section we provide a few examples of conjugate processes. In some of these examples it is implicitly assumed that the underlying probability space is the canonical one constructed in the proof of Theorem 3.1.

Example 3.4 (Copulas, yet again: continuation of Example 3.3)**.** The main motivation for introducing the partition $(T_\alpha : \alpha \in J)$ is to allow for a notion of *distributional cycle*. For instance, let $J = \mathbb{N}$, and consider a *period* $p \in \mathbb{N}$. Let $T_1 = \{1, \ldots, p\}$, $T_2 = \{p+1, \ldots, 2p\}$ and so on. Let Q be any measure on $M(\mathfrak{S})^{\mathbb{N}}$. Applying the construction of Theorem 3.1 one obtains a conjugate process $(\xi_n : n \in \mathbb{N},\, X_t : t \in \mathbb{N})$ with the following properties: conditional on ξ, (i) the dependence structure of (X_t) is captured by the family of copulas $(C_n : n \in \mathbb{N})$, and; (ii) for $t \in T_n = \{(n-1)p+1, \ldots, np\}$, the random variable X_t has marginal ξ_n: one has $\mathbb{P}[X_t \in A \,|\, \xi] = \xi_n(A)$. The interpretation is that 'at the beginning of time' a random sequence of measures (ξ_n) has been drawn from Q, and thereafter the process (X_t) evolves according to the dependence structure implied by the family of copulas from Example 3.3, with each X_t having marginal distribution ξ_n during the n-th *cycle*, T_n. This model is potentially useful in applications where a process admits a natural notion of a cyclic behavior, at least in a distributional sense. Horta and Ziegelmann (2015b) consider a similar setting. //

Example 3.5 (Bayesian inference)**.** Let $T = \{1, \ldots, n\}$, and put $(\mathscr{F}, \mathfrak{F}) = (S^n, \mathfrak{S}^n)$. Let $\mathscr{L} = M(\mathfrak{S})^n$ and, for $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathscr{L}$, define $P_\lambda = \lambda_1 \otimes \cdots \otimes \lambda_n$ on $\mathfrak{S}^n$. Now let $J = \{0\}$ and $T_0 = T$. In this scenario we want to set $\mathscr{M} := M(\mathfrak{S})^{\{0\}} \equiv M(\mathfrak{S})$, and $\rho(\mu) = (\mu, \ldots, \mu) \in \mathscr{L}$. Let Q be a *prior distribution* on $\mathfrak{M}$. Then Theorem 3.1 ensures the existence of a $M(\mathfrak{S})$ valued random element ξ having distribution Q, and S valued random elements $X_1, \ldots, X_n$ such that, for $A_i \in \mathfrak{S}$,

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n \,|\, \xi] = \xi(A_1) \cdots \xi(A_n).$$

Whenever the conditional probability $\mathbb{P}[\xi \in M \,|\, X_1, \ldots, X_n]$ is almost surely a probability measure (as a function of $M \in \mathfrak{M}$), it is called the *posterior distribution* (of the parameter ξ given the data $X_1, \ldots, X_n$). //

Example 3.6 (Bayesian inference, unconditional distribution of X)**.** Let T be the set $\mathbb{N}$ of natural numbers, and put $(\mathscr{F}, \mathfrak{F}) = \left(S^{\mathbb{N}}, \mathfrak{S}^{\mathbb{N}}\right)$. Let $\mathscr{L} = M(\mathfrak{S})^{\mathbb{N}}$ and, for each $\lambda = (\lambda_1, \lambda_2, \ldots) \in \mathscr{L}$, define P_λ to be the product measure $P_\lambda = \otimes_{n \in \mathbb{N}} \lambda_n$ on $\mathfrak{S}^{\mathbb{N}}$. Let J be a singleton as above, and again set $\mathscr{M} = M(\mathfrak{S})$. Set $\rho(\mu) = (\mu, \mu \ldots)$, and let Q be a prior distribution on $\mathfrak{M}$. Assume that the *baricenter*[1] $\int \mu \, dQ(\mu)$ of Q is well defined. Consider

[1] Recall that the baricenter of Q is defined to be the unique element $\mu^* \in M(\mathfrak{S})$ such that the equality $\int \psi(x)\, d\mu^*(x) = \int \left(\int \psi(x)\, d\mu(x) \right) dQ(\mu)$ holds for each bounded measurable $\psi : S \to \mathbb{R}$. Notation: $\mu^* = \int \mu \, dQ(\mu)$ (this is a Pettis integral).

the processes $(X_n : n \in \mathbb{N})$ and $(\xi_\alpha : \alpha \in J)$ given in Theorem 3.1. Since J is as singleton the latter process is in fact just a random element ξ in $M(\mathfrak{S})$, and here we will have $\mathbb{E}\xi = \int \mu \, dQ(\mu)$. In the language of Bayesian inference, $\mathbb{E}\xi$ is a *hyperparameter*. Now notice that, by construction, conditional on ξ the sequence (X_n) is iid $\sim \xi$. Also observe that, from Corollary 3.1, it is immediate that each X_n has unconditional marginal distribution $\mathbb{E}\xi$. It is thus tempting to conjecture that the unconditional distribution of (X_n) is $P_{\rho(\mathbb{E}\xi)}$ – i.e., that (X_n) is an independent sequence, each X_n having marginal distribution $\mathbb{E}\xi$. Unfortunately, this is not true in general. Indeed, letting $E = A_1 \times \cdots \times A_n \times S \times \cdots$ be an element of $\mathfrak{S}^{\mathbb{N}}$ we have, on the one hand,

$$\int P_{\rho(\mu)}(A_1 \times \cdots \times A_n \times S \times \cdots) \, dQ(\mu) = \int \mu(A_1) \cdots \mu(A_n) \, dQ(\mu), \tag{3.3}$$

whereas on the other hand, writing $\mu^* = \int \mu \, dQ(\mu)$, we have

$$P_{\rho(\mu^*)}(A_1 \times \cdots \times A_n \times S \times \cdots) = \mu^*(A_1) \cdots \mu^*(A_n). \tag{3.4}$$

Since $\mu^*(A_i) = \int \mu(A_i) \, dQ(\mu)$, there is no particular reason for (3.3) and (3.4) to be equal. //

Example 3.7 (Continuation of Example 3.6, Limit Theorems)**.** Let the model be as above, and let f be a bounded, real valued measurable function on S. Consider the question of whether a Law of Large Numbers holds for $(f \circ X_n : n \in \mathbb{N})$. As we have seen, the sequence (X_n) is not necessarily unconditionally iid, so there is no reason for this to happen. However, *conditionally* a LLN must hold. One possible way to interpret this fact is by saying that the 'state of the world' can be one and only one of $[\xi = \mu]$, where $\mu \in M(\mathfrak{S}) \setminus N$. For such μ, and for f as above, the set

$$E_f^\mu := \left\{ (x_1, x_2, \dots) \in S^{\mathbb{N}} : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f(x_i) = \int f \, d\mu \right\}$$

has $P_{\rho(\mu)}$ probability 1. Thus if $X(\omega) \in E_f^\mu$, the sequence $n^{-1} \sum_{i=1}^n f(X_i(\omega))$ converges to $\int f \, d\mu$. Nevertheless we cannot say that the $P_{\rho(\mu)}$ probability of such ω is 1 since the latter measure has domain $\mathfrak{F}$ rather than $\mathfrak{A}$. The usual way of resolving this is by inducing a measure $P'_{\rho(\mu)}$ on the σ-algebra $\mathfrak{A}$ (see the proof of Theorem 3.1) which concentrates on the set $[\xi = \mu]$, as follows: for $G \in \mathfrak{A}$ set $P'_{\rho(\mu)}(G) := P_{\rho(\mu)}\big\{x \in S^{\mathbb{N}} : (\mu, x) \in G\big\}$.

Alternatively, notice that $\mathbb{P}[|n^{-1} \sum_{i=1}^n f \circ X_i - \int f \, d\xi| > \epsilon \,|\, \xi] \to 0$ almost surely, and hence Lebesgue's Dominated Convergence Theorem gives $\mathbb{P}[|n^{-1} \sum_{i=1}^n f \circ X_i - \int f \, d\xi| > \epsilon] \to 0$. //

Example 3.8 (Weak Law of Large Numbers)**.** Let $J = T = \mathbb{N}$, with $\mathscr{L} = \mathscr{M} = M(\mathfrak{S})^{\mathbb{N}}$. Assume Q is weakly ergodic in a class $\mathscr{C}$ of bounded measurable functions from S to $\mathbb{R}$, in the following sense: for each $f \in \mathscr{C}$ and $\epsilon > 0$,

$$Q\left\{ (\mu_1, \mu_2, \dots) \in \mathscr{M} : \left| \frac{1}{n} \sum_{t=1}^n \mu_t(f) - \mu_1^*(f) \right| > \epsilon \right\} \xrightarrow[n\to\infty]{} 0. \tag{3.5}$$

Here $\mu_1^* = \int \mu_t \, dQ(\mu)$ is the common (by assumption) baricenter of the μ_t, and we have written $\mu_t(f)$ for $\int f \, d\mu_t$. For $\lambda = (\lambda_t : t \in \mathbb{N}) \in \mathscr{L}$, define P_λ as the product measure $\otimes_{t\in\mathbb{N}}\lambda_t$. In this setting the following Weak Law of Large Numbers holds.

Proposition 3.1. *In the conditions of Theorem 3.1, for each $f \in \mathscr{C}$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} f \circ X_t - \mathbb{E}\xi_1(f)\right| > \epsilon\right\} \underset{n\to\infty}{\longrightarrow} 0,$$

with $\mathbb{E}\xi_1 = \mu_1^*$.

Compare with Example 3.7: there the time averages $n^{-1}\sum_{t=1}^{n} f \circ X_t$ converge in probability to the *conditional* expectation $\int f \, d\xi$, whereas in Proposition 3.1 the limit is the *unconditional* $\int f \, d\mu_1^*$. //

Example 3.9 (Regime switching models)**.** The concept of conjugate process encompasses a class of regime switching models. We give an example below. Let $S = \mathbb{R}$ be endowed with the Borel σ-field $\mathfrak{S}$, and put $J = T = \mathbb{Z}$. Set $\mathscr{L} = M(\mathfrak{S})^{\mathbb{Z}}$ and define, for each $\lambda = (\lambda_t : t \in \mathbb{Z}) \in \mathscr{L}$, $P_\lambda = \otimes_{t\in\mathbb{Z}}\lambda_t$. Let $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(d)} \in M(\mathfrak{S})$ be any d distinct *regimes* ($d \in \mathbb{N}$). Let Q be the probability measure on $\mathscr{M} := \{\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(d)}\}^{\mathbb{Z}}$ corresponding to the Markov Chain with transition probability $(Q_{ij} : i, j = 1, \ldots, d)$ and stationary distribution $(p_i : i = 1, \ldots, d)$. By Theorem 3.1 and Corollary 3.1 one can obtain an independent two-sided sequence $(X_t : t \in \mathbb{Z})$ with X_t having conditional marginal distribution $\mathbb{P}[X_t \in A \,|\, \xi_t] = \xi_t(A)$, where (ξ_t) is a *hidden*, or *latent*, Markov Chain with state space $\left\{\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(d)}\right\}$, transition probability (Q_{ij}), and stationary distribution (p_i). The *innovations* (X_t) can be used to obtain more elaborate models driven by the latent regime process. For instance, assume that for any $k = 1, \ldots, d$ one has

$$\int x \, d\mu^{(k)}(x) = 0, \qquad \int x^2 \, d\mu^{(k)}(x) = \sigma_k^2 < \infty,$$

and that, if $(\epsilon_t : t \in \mathbb{Z})$ is iid $\sim \mu^{(k)}$, then $\sum_{t\in\mathbb{Z}} |\alpha\epsilon_t| < \infty$ almost surely, as long as $|\alpha| < 1$. For such α, define $Y_t := \sum_{s=0}^{\infty} \alpha X_{t-s}$. Then, conditional on ξ, we have that (Y_t) is an AR(1) process, $Y_t = \alpha Y_{t-1} + X_t$, with the distribution of the innovations X_t being driven by the Markov Chain (ξ_t). //

Example 3.10 (Non-random ξ)**.** Let J be a singleton, and set $\mathscr{M} = M(\mathfrak{S})^J \equiv M(\mathfrak{S})$. Let $\mu_0 \in M(\mathfrak{S})$, and put $Q = \delta_{\mu_0}$. If $T = J$ and $\mathscr{F} = S$, then the probability space given in Theorem 3.1 has $\Omega = M(\mathfrak{S}) \times S$, and $\mathbb{P}$ is a probability measure concentrating on the set $\{\mu_0\} \times S$. Indeed $(\Omega, \mathfrak{A}, \mathbb{P})$ is isomorphic to $(S, \mathfrak{S}, \mu_0)$. More generally, if P is any measure on a function space $(\mathscr{F}, \mathfrak{F}) \leq \left(S^T, \pi\right)$, define $\mathscr{L} = \{\mu\}$, where $\mu = \left(P \circ \pi_t^{-1} : t \in T\right)$, and set $P_\mu := P$. Putting $\mathscr{M} = \mathscr{L}$ and $Q = \delta_\mu$ yields $(\Omega, \mathfrak{A}, \mathbb{P}) = (\mathscr{F}, \mathfrak{F}, P)$. //

Example 3.11 (Conditioning on a random variable Y)**.** Let T and J be singletons. Define $(\mathscr{F}, \mathfrak{F}) = (S, \mathfrak{S})$ and $\mathscr{M} = M(\mathfrak{S})$. Let $(N, \mathfrak{N}, \nu)$ be a probability space, and let $\gamma : N \times \mathfrak{S} \to \mathbb{R}$ be a probability kernel from $(N, \mathfrak{N})$ to $(S, \mathfrak{S})$. Assume the map $y \in N \mapsto \gamma(y, \cdot) \in M(\mathfrak{S})$ is injective, and let h denote its left inverse. Induce on $M(\mathfrak{S})$ the σ-field $\mathfrak{M}$ defined by the condition

$$G \in \mathfrak{M} \iff \{y \in N : \gamma(y, \cdot) \in G\} \in \mathfrak{N}.$$

and put $Q(G) := \nu\{y \in N : \gamma(y, \cdot) \in G\}$. The measure Q is seen to concentrate on $\operatorname{ran}(y \mapsto \gamma(y, \cdot))$. In the probability space from Theorem 3.1, defining $Y := h \circ \xi$ gives $\sigma(\xi) = \sigma(Y)$ and thus, for $A \in \mathfrak{S}$,

$$\mathbb{P}[X \in A \,|\, \xi] = \mathbb{P}[X \in A \,|\, Y] = \gamma(Y, A).$$

Here the N-valued random variable Y is defined almost everywhere, and has law ν. It can be interpreted for instance as a parameter (parametric Bayesian inference), or as a (random) regressor (regression models). As a more concrete example of the latter, take $(N, \mathfrak{N}, \nu) = ([0, 1], \mathfrak{B}[0, 1], \mathrm{Leb})$, and let $\gamma(y, \cdot)$ be the normal distribution on the real line $\mathbb{R} \equiv S$ having mean $\alpha + \beta y$ and variance equal to 1, for some $\alpha, \beta \in \mathbb{R}$. This gives the regression model $\mathbb{E}[X \,|\, Y] = \alpha + \beta Y$ with Gaussian errors and Uniform$[0, 1]$ regressor. //

Follow ups. Say J is a semigroup with identity element 0. Assume Q is a stationary measure on $(\mathscr{M}, \mathfrak{M})$, that is, Q is invariant by the group action: $Q(M) = Q(\alpha M)$. For f and h in some class of bounded measurable functions on S, define

$$R_\alpha(f, h) = \int \mu_0(f)\mu_\alpha(h)\, dQ(\mu) - \int \mu_0(f)\, dQ(\mu) \int \mu_\alpha(h)\, dQ(\mu).$$

Of particular interest is the autocovariance function R_0: a crucial question is whether it can be estimated by the data $(X_t : t \in T)$. Horta and Ziegelmann (2015b) tackle this question in a different setting. //

3.4 Concluding remarks

The original insight of the concept of a conjugate process appeared when we were studying how to model the dynamics of distribution functions of high frequency asset returns in financial data; to be precise, assume asset returns share the same marginal distribution inside each day, but allow this marginal to vary from day to day (possibly in a stochastic manner). How to give a reasonable formulation of these ideas? From there to the construction presented in this paper it was a long road but at some point, after chasing the primitive conditions which would allow for a solid theory for describing the sort of process just discussed, we arrived at the concept of $\mathscr{L}$-coherence and compatibility which permit the elegant statement of Theorem 3.1. The fact that the concept of conjugate process yielded almost effortlessly a very interesting (in our opinion) set of examples may be taken as an indicative that it is a powerful tool only waiting for minds more creative that our own to find relevant applications in a wide variety of scientific fields.

3.5 Proofs

Proof of Theorem 3.1. We shall assume for simplicity that $\rho(\mathscr{M}) \subset \mathscr{L}$. Let $\Omega := \mathscr{M} \times \mathscr{F}$ and denote by $\mathfrak{A}$ the product σ-field on Ω. The collection $\big(P_{\rho(\mu)} : \mu \in \mathscr{M}\big)$ is a probability kernel from $(\mathscr{M}, \mathfrak{M})$ to $(\mathscr{F}, \mathfrak{F})$. Thus the identity

$$\mathbb{P}[M \times E] = \int_M P_{\rho(\mu)}(E)\, dQ(\mu), \qquad M \in \mathfrak{M}, \quad E \in \mathfrak{F},$$

defines a unique probability measure $\mathbb{P}$ on $\mathfrak{A}$. See Pollard (2002, Theorem 4.20). Let ξ and X be the projections from Ω respectively onto $\mathscr{M}$ and onto $\mathscr{F}$ (so $\xi_\alpha = \pi_\alpha \circ \xi$ and likewise $X_t = \pi_t \circ X$). Clearly $(\xi_\alpha : \alpha \in J)$ has its sample-paths in $\mathscr{M}$, and law Q. Likewise, $(X_t : t \in T)$ has its sample-paths in $\mathscr{F}$. Moreover, for $M \in \mathfrak{M}$ and $E \in \mathfrak{F}$,

$$\begin{aligned}\int_{\xi^{-1}(M)} P_{\rho\circ\xi(\omega)}(E)\, d\mathbb{P}(\omega) &= \int_M P_{\rho(\mu)}(E)\, dQ(\mu)\\ &= \mathbb{P}[X \in E,\, \xi \in M],\end{aligned}$$

and thus $P_{\rho(\xi)}(E)$ is a version of $\mathbb{P}[X \in E \,|\, \xi]$. This establishes (3.1). □

Proof of Corollary 3.1. Let $t \in T_\alpha$ and $E \in \mathfrak{F}$ be of the form $E = \pi_t^{-1}(A)$ for some $A \in \mathfrak{S}$. Notice that $P_{\rho(\mu)}(E) = \pi_\alpha(\mu)(A)$, and since the LHS is a measurable function of μ by assumption, so is the RHS. The rest of the proof follows easily. □

Proof of Proposition 3.1. Clearly, $\mathbb{E}\xi = \mu_1^*$. Let $Z_t = f \circ X_t - \xi_t(f)$. We have

$$\left|\frac{1}{n}\sum_{t=1}^n f \circ X_t - \mathbb{E}\xi(f)\right| \le \left|\frac{1}{n}\sum_{t=1}^n Z_t\right| + \left|\frac{1}{n}\sum_{t=1}^n \xi_t(f) - \mathbb{E}\xi(f)\right|.$$

The second term in the above sum is $o_{\mathbb{P}}(1)$ by (3.5). For the first term, we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{t=1}^n Z_t\right| > \epsilon\right\} = \mathbb{E}\left[\mathbb{P}\left\{\left|\frac{1}{n}\sum_{t=1}^n Z_t\right| > \epsilon \,\middle|\, \xi\right\}\right]. \tag{3.6}$$

But $(Z_t|\xi : t \in \mathbb{N})$ is and independent sequence with $\mathbb{E}[Z_t|\xi] = 0$, and therefore

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{t=1}^n Z_t\right| > \epsilon \,\middle|\, \xi\right\} \to 0, \qquad \text{almost surely.}$$

This quantity is bounded by 1 and hence the RHS in (3.6) goes to zero by Dominated Convergence. □

Remark. All the examples in the text should come accompanied by the sentence 'if $\mu \mapsto P_{\rho(\mu)}(E)$ is measurable for all E'. The reader should check that this is the case in each given situation.

3.6 References

Chang, J. T., and Pollard, D. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

Faden, A. M. The existence of regular conditional probabilities: necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298, 1985.

Horta, E., and Ziegelmann, F. Weakly conjugate processes – theory and application to risk forecasting. 2015.

Kallenberg, O. Characterization and convergence of random measures and point processes. *Probability Theory and Related Fields*, 27(1):9–21, 1973.

Kallenberg, O. *Lectures on random measures.* Consolidated University of North Carolina, Institute of Statistics, 1974.

Kallenberg, O. An elementary approach to the Daniell-Kolmogorov Theorem and some related results. *Mathematische Nachrichten*, 139(1):251–265, 1988.

Kallenberg, O. *Foundations of modern probability.* Springer Science & Business Media, 2006.

Pachl, J. K. Desintegration and compact measures. *Mathematica Scandinavica*, 43: 157–168, 1978.

Pollard, D. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002.

4 WEAKLY CONJUGATE PROCESSES – THEORY AND APPLICATION TO RISK FORECASTING

Eduardo Horta[1] Flávio Ziegelmann[2]

November, 2015

Abstract. Many dynamical phenomena display a cyclic behavior, in the sense that time can be partitioned into cycles in which distributional aspects of a process are homogeneous. The standard probabilistic approach to modeling the evolution of a system over time usually begins with specification of a certain probability measure on the space of sample paths, induced by a family of finite dimensional distributions. In this setting consideration of conditional probabilities commonly involves the notion of 'past information' as summarized by a filtering or the past trajectory of the process. In contrast, the class of models that we present here allows the marginal distributions of a cyclic process to evolve stochastically in time, in principle separated from the observable process itself. The connection between them is given by a compatibility condition on the conditional marginal distributions. The methodology relates to the concept of random measure and more generally to Probability in Banach spaces. From the inferential point of view our method can be seen as Functional Data Analysis. We provide a constructive example which illustrates the method. A statistical implementation of our model to risk forecasting in financial data is given. Specifically, we generate forecasts of intraday asset returns variance and Value-at-Risk. The forecasts are attainable by reducing the dimension of the conditional distribution process into a latent scalar time series.

Keywords and phrases. Random measure. Covariance operator. Dimension reduction. Functional time series. High frequency financial data. Risk forecasting.

JEL Classification. C1, C14, C22

4.1 Introduction

Many dynamical phenomena display a cyclic behavior, in the sense that time can be partitioned into cycles, over which a process 'repeats itself' except for certain characteristics specific to each cycle. This idea is the starting point of the theory developed in Bosq (2000), for instance. The standard probabilistic approach to modeling the evolution of a system over time usually begins with specification of a certain probability measure on the space of sample paths, induced by a family of finite dimensional distributions. In this setting consideration of conditional probabilities usually involves the notion of 'past information' as summarized by a filtering or the past trajectory of the process. We shall take a different approach, interpreting the cyclic character of a process in a distributional sense. We consider the following model. A sequence $F_0, F_1, \ldots, F_t, \ldots$ of random cdf's evolves stochastically in time. Associated to these distribution functions is a continuous

[1]Department of Statistics – Universidade Federal do Rio Grande do Sul. eduardo.horta@ufrgs.br

[2]Department of Statistics – Universidade Federal do Rio Grande do Sul. flavioz@ufrgs.br

time, real-valued stochastic process $(X_\tau : \tau \geq 0)$ that satisfies the following condition

$$\mathbb{P}[X_\tau \leq z | \mathscr{F}] = F_t(z), \quad \tau \in [t, t+1), \tag{4.1}$$

where $\mathscr{F}$ is the σ-algebra generated by $F_0, F_1, \ldots$. We shall call each interval $[t, t+1)$ the t-th *cycle*. Of course, equation (4.1) implies that, for $\tau \in [t, t+1)$, $\mathbb{P}[X_\tau \leq z | F_0, \ldots, F_t] = \mathbb{P}[X_\tau \leq z | F_t] = F_t(z)$. This can be interpreted as meaning that the process (X_τ) has marginal conditional distribution F_t during cycle t, and that past and future information about the $F_j's$ is in some sense irrelevant when F_t is given. Little further probabilistic structure is imposed on (X_τ). From the point of view of simulation, condition (4.1) and the latter comments say that sampling 'all' the F_t's first and then generating the process $(X_\tau : \tau \geq 0)$, or sampling the F_t's iteratively for each t and generating $(X_\tau : \tau \in [t, t+1))$ at each cycle, is an equivalent procedure. The model is potentially useful in situations where there is a natural notion of a cycle in the behavior of the process (X_τ), and where the main interest concerns statistical (i.e. distributional) aspects of the process, rather than 'sample-path' aspects, within each cycle. Possible applications include temperature measurements and intraday stock market return processes, the latter of which we illustrate below with a real data set. This model does have a Bayesian flavor, in that the distribution of the random variables X_τ are themselves random elements in a space of distribution functions. From now on we will go without saying that the index sets for t and τ are $0, 1, 2, \ldots$ and $\mathbb{R}^+$ respectively. A pair (F_t, X_τ), where (F_t) is a sequence of random distribution functions and (X_τ) is a process satisfying the compatibility condition (4.1), will be called a *weakly conjugate process*. Horta and Ziegelmann (2015a) consider a slightly different scenario where $\mathscr{F}$ determines the whole distribution of the process (X_τ) and not only the marginals. Our condition (4.1) is always satisfied in this setting (see Corollary 1 therein).

In a sense, the method herein presented is intrinsically *functional* in that we consider random elements in a space of functions. Thus our theory can be regarded both as Functional Data Analysis and as Probability in Banach spaces, which are two very important research fields in the statistics and probability literature respectively. Statistical inference on objects pertaining to function spaces has come to be known in the literature as Functional Data Analysis (hereafter FDA). In recent years, FDA has received growing attention from researchers of a wide spectrum of academic disciplines; see for instance the collection edited by Dabo-Niang and Ferraty (2008) for a discussion on recent developments and many applications. As an example, an application to implied volatility estimation can be found in Benko et al. (2009). The cornerstone monograph by Ramsay and Silverman (1998) presents a thorough treatment on the topic. A central technique in this context is that of functional principal components analysis. At short, such methodology – whose foundation lies in the Karhunen-Loève Theorem – seeks a decomposition of the observed functions as orthogonal projections onto a suitable orthonormal basis corresponding to the eigenfunctions of a covariance operator. Hall and Vial (2006) consider the case where the observed functional data are imprecise – due to roundings, experimental measurement errors, etc. – a scenario where some complications arise regarding estimation of the covariance operator. Bathia et al. (2010), tackle this issue in a functional time series framework. Our approach is inspired by their methodology. From a theoretical point of view, functional data are to be seen as realizations of function-valued random variables. The general approach is to consider random elements in a Banach space. The theory of Probability in Banach spaces first rose from the need to interpret stochastic processes as random variables with values in function spaces. The original insight is likely due to Wiener, who constructed a probability measure on the space of continuous functions (Brownian motion) yet before Kolmogorov's axiomatization of probability theory. Classic texts include Ledoux and Talagrand (1991) and Vakhania et al. (1987). It turns out that a convenient and quite general approach is to consider probability measures in metric spaces; this theory is well stablished in the classic text by Billingsley (2009), whereas a modern account would be Van Der Vaart and Wellner (1996). For stationary sequences and linear processes in Banach spaces, the monograph from Bosq (2000) is a complete account. Specialized versions of the LLN and CLT for dependent sequences can be found therein. The theory of Bochner and Pettis integrals is straightly linked to the theory of probability in Banach spaces. A very clear exposition is given in the first chapters of van Neerven (2008). For texts that blend theory and applications, see Ferraty and Vieu (2006) and Damon and Guillas (2005).

In our framework, we consider the case where the underlying distribution process (F_t) displays some degree of dependence. The precise meaning of this property is given in Assumption D below. With this consideration in mind we show that the methodology of Bathia et al. (2010) can be applied to study the dependence structure of (F_t) through observation of the process (X_τ) only. In applications this methodology allows one, at the end of each cycle, to use current information to forecast *distributional* aspects of the observable process (X_τ) in the next cycle. For each t let X_{it}, $i = 1, \dots, q_t$ denote some observations of the process (X_τ) in cycle t. We define a *sampling scheme* in terms of the collection $\{X_{it}\}$ (we are being rather loose in the definition but the meaning should be evident). Let G_t denote the empirical distribution function of the observations in cycle t,

$$G_t(x) := \frac{1}{q_t} \sum_{i=1}^{q_t} \mathbb{I}_{[X_{it} \leq x]}.$$

Writing

$$G_t(x) = F_t(x) + \varepsilon_t(x),$$

where $\varepsilon_t = G_t - F_t$ *by tautology*, we obtain the following properties.

Lemma 4.1. *Let (F_t, X_τ) be a weakly conjugate process satisfying Assumptions F, D and X. Then the following holds.*

(i) $\mathbb{E}\varepsilon_t(x) = 0$ *for all* t *and all* $x \in \mathbb{R}$*;*

(ii) $\mathrm{Cov}(F_t(x), \varepsilon_{t+k}(y)) = 0$ *for all* t*, all integers* k *and all* $x, y \in \mathbb{R}$*;*

(iii) $\mathrm{Cov}(\varepsilon_t(x), \varepsilon_{t+k}(y)) = 0$ *for all* t *and all* $x, y \in \mathbb{R}$ *provided* $k \neq 0$.

In summary, Lemma 4.1 can be interpreted as saying that, under weak assumptions, the intra-cycle empirical cdf's of weakly conjugate processes are decomposable as 'underlying, true cdf' plus 'noise'. Notice though that in general (ε_t) is not *white* noise since when $k = 0$ in item *(iii)* the covariances may depend on t.

In this setting, for a fixed, finite measure μ on $\mathbb{R}$, let R^μ be the operator acting on $L^2(\mu)$ defined by $R^\mu f(x) := \int R_\mu(x, y) f(y)\, d\mu(y)$, where

$$R_\mu(x, y) := \sum_{k=1}^{p} \int C_k(x, z) C_k(y, z) d\mu(z), \tag{4.2}$$

and C_k is the k-th lag autocovariance function of (F_t), that is

$$C_k(x, y) := \mathrm{Cov}(F_t(x), F_{t+k}(y)). \tag{4.3}$$

Under Assumptions F and D below, F_t admits the representation

$$F_t(x) = F(x) + \sum_{j=1}^{d} \eta_{tj} \psi_j(x), \tag{4.4}$$

where

$$F(x) := \mathbb{E}F_t(x), \tag{4.5}$$

and the ψ_j are the eigenfunctions of the positive, finite-rank operator R^μ. Here $\eta_{tj} := \left\langle F_t - F, \psi_j \right\rangle$ is a scalar, zero-mean random variable, and p is some fixed integer. We shall use the notation $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively for the inner-product and the norm in $L^2(\mu)$. Representation (4.4) is similar to the well-known Karhunen-Loève expansion (of each F_t seen as a process $x \mapsto F_t(x)$) but the latter is associated to the zero-lag covariance operator of F_t rather than with R^μ. Notice that although not explicitly indicated, the eigenfunctions ψ_j and the scalar random variables η_{tj} will depend on the measure μ. See the discussion on Assumption D below for further justification of introducing the operator R^μ. Observe that in our setting representation (4.4) holds uniformly, almost surely, as stated in Proposition 4.1.

The rationale for introducing the measure μ is to make it possible to see the random variables F_t as taking values in a suitable Hilbert space, and thus take advantage of the richer structure of such spaces. The measure μ is, in fact, at choice of the statistician, and should be chosen in such a way that important features of the weakly conjugate process $(F_t,\, X_\tau)$ can be captured; ideally one would take μ not too far from the measure corresponding to F, but the latter is generally unknown. A very important case occurs when the process (X_τ) is bounded. In this case there is an interval $[a, b]$ such that the F_t satisfy $F_t(x) = 0$ for $x \leq a$ and $F_t(x) = 1$ for $x \geq b$ whereas $0 < F_t(x) < 1$ if $x \in (a, b)$; the same holds for F. In this situation one can take μ = Lebesgue measure restricted to $[a, b]$ and representation (4.4) becomes easier to interpret. In fact the only reason to introduce more general measures μ is to allow for consideration of unbounded processes (X_τ).

4.2 Assumptions and main results

We consider a given probability space $(\Omega, \mathscr{A}, \mathbb{P})$ where a weakly conjugate process (F_t, X_τ) is defined and satisfies the following conditions

F (F_t) is a stationary sequence of random elements taking values in a finite dimensional subspace of the space $\mathrm{C}_b(\mathbb{R})$ of bounded, continuous functions on $\mathbb{R}$. Moreover F_t is a cdf with probability 1, for all t;

X conditional on $\mathscr{F}$, the random variables X_{τ_1} and X_{τ_2} are independent if $\lfloor \tau_1 \rfloor \neq \lfloor \tau_2 \rfloor$.

In addition to the above, and in order to exploit the dynamic structure of the process (F_t), we shall find it fruitful to impose an extra condition, one similar to the requirement in Proposition 1 in Bathia et al. (2010). In this direction, let Ω_0 be a $\mathbb{P}$-null set such that

$$\bigvee\{F_0^\omega - F;\, \omega \notin \Omega_0\} = \bigvee\{\varphi_1, \ldots, \varphi_d\}$$

for some *minimal* linearly independent set $\varphi_1, \ldots, \varphi_d$ in $\mathrm{C}_b(\mathbb{R})$. Let ξ_{tj} be the (random) coefficient of $F_t - F$ with respect to φ_j, that is $F_t - F = \xi_{t1}\varphi_1 + \cdots + \xi_{td}\varphi_d$ almost surely. Notice that $\mathbb{E}\xi_{tj} = 0$ for all t and all j. The following assumption summarizes our requirement that the sequence (F_t) displays 'enough' dependence.

D for some integer $k \geq 1$ the matrix $(\mathbb{E}\xi_{0i}\xi_{kj})_{ij}$ is of rank d.

Notice that this property is independent of the choice of basis $\varphi_1, \ldots, \varphi_d$. This condition is easier to appreciate in the case where $F_t - F$ lies in a one-dimensional subspace of $\mathrm{C}_b(\mathbb{R})$, that is the case $d = 1$. In this setting the matrix $(\mathbb{E}\xi_{0i}\xi_{kj})_{ij}$ is indeed a scalar, and the condition that it is full-rank for some k means that the univariate time series (ξ_{t1}) is correlated at some lag k. Otherwise we would find ourselves in the not very interesting scenario (for our purposes at least) of an uncorrelated scalar time series.

Assumption F above gives a precise meaning to our notion of a 'sequence of random cdf's which evolve stochastically in time'. First, the hypothesis that the process (F_t) takes its values in a finite dimensional space has a statistical motivation, relating to functional PCA and identification of finite dimensionality in functional data. See Hall and Vial (2006) and Bathia et al. (2010) for a discussion. As for the specific choice of a finite dimensional subspace of $\mathrm{C}_b(\mathbb{R})$, it is motivated mostly by adequacy. One could as well take as model, in place of $\mathrm{C}_b(\mathbb{R})$, the space of càdlàg functions with the Skorokhod metric, or the space of real probability measures with the weak* topology (with a few modifications, this is the approach taken in Horta and Ziegelmann (2015a)), but in terms of structure $C_b(\mathbb{R})$ grants us a few properties that are rather convenient for us to further develop the theory, such as *(i)* the F_t are separably valued; *(ii)* $F_t(x) \equiv \delta_x \circ F_t$ are real random variables, from which; *(iii)* the Bochner expectation $\mathbb{E}F_t$ is well defined and $(\mathbb{E}F_t)(x) = \mathbb{E}(F_t(x))$ holds; *(iv)* given any finite measure μ on $\mathbb{R}$, each F_t is a strongly measurable random element in $L^2(\mu)$, and moreover the expectation $\mathbb{E}F_t \in \mathrm{C}_b(\mathbb{R})$ 'works' as the expected value in $L^2(\mu)$. Finally, the hypothesis of weak stationarity implies that the mean function (4.5) and the autocovariance functions (4.3) are well defined and time-invariant. In particular the covariance function is continuous and induces a Mercer kernel on $L^2(\mu)$.

Assumption X appears quite restrictive at first but as seen in the proofs section it is crucial in establishing item *(iii)* in Lemma 4.1 and the LLN in Theorem 4.1. In this regard an important remark is that it does not imply that if X_{τ_1} and X_{τ_2} are in distinct cycles they will be *unconditionally* independent: conditional independence *does not* imply unconditional independence. See Horta and Ziegelmann (2015a) for a discussion in a similar setting.

Introducing the operator R^μ in turn is justified in the same fashion as in Bathia et al. (2010), and has an inferential motivation. The aim is to obtain representation (4.4) as an alternative to the Karhunen-Loève representation of F_t. To see why we want to achieve this, let

$$\widehat{C}_k(x,y) = \frac{1}{n-p}\sum_{t=1}^{n-p}\Big(G_t(x) - \widehat{F}(x)\Big)\Big(G_{t+k}(y) - \widehat{F}(y)\Big), \tag{4.6}$$

where

$$\widehat{F}(x) = \frac{1}{n}\sum_{t=1}^{n} G_t(x). \tag{4.7}$$

It is clear that $\widehat{C}_0$ is generally an illegitimate estimator for C_0, since $\mathrm{Cov}(G_t(x), G_t(y)) = C_0(x,y) + \mathrm{Cov}(\varepsilon_t(x), \varepsilon_t(y))$. For integers $k \neq 0$, however, it holds that $\mathrm{Cov}(G_t(x), G_{t+k}(y)) = C_k(x,y)$ by Lemma 4.1, and so $\widehat{C}_k$ is indeed legitimate as an estimator of C_k. From here defining R^μ via (4.2) gives that R^μ is a positive operator with $\mathrm{Ran}\Big(R^\mu\Big) = \bigvee\{\varphi_1, \ldots, \varphi_d\}$, as stated in Proposition 4.1, which finally yields (4.4). The strategy becomes thus to estimate R^μ, its associated eigenvalues and eigenfunctions, and most importantly to use the latter to recover the vector time series $\boldsymbol{\eta}_t := (\eta_{t1}, \ldots, \eta_{td})$. Regarding the integer p, it reflects the fact that in general it is not known the precise value of k for which assumption D holds. Using some of the lagged C_k in the definition of R_μ is a parsimonious way to overcome this.

Proposition 4.1. *Let* (F_t, X_τ) *be a weakly conjugate process satisfying assumptions F, X and D. If* μ *is a finite measure on* $\mathbb{R}$*, equivalent to Lebesgue measure, then provided p is large enough there are some continuous bounded functions* $\psi_1, \dots \psi_d$ *which are orthonormal in* $L^2(\mu)$ *and satisfy*

(i) $\bigvee_{j=1}^d \psi_j = \bigvee_{j=1}^d \varphi_j$*;*

(ii) $R^\mu \psi_j = \theta_j \psi_j$*, for some* $\theta_j > 0$*,* $j = 1, \dots, d$*.*

In particular, representation (4.4) *holds uniformly in* x *almost surely, for zero-mean random variables* η_{tj} *satisfying* $\mathbb{E}\eta_{ti}\eta_{tj} = 0$ *for* $i \neq j$ *and* $\mathbb{E}\eta_{ti}^2 = \theta_i$*.*

Moreover, if $\operatorname{supp}(X_0)$ *is bounded, say* $\operatorname{supp}(X_0) \subset [a, b]$*, then the above statement remains true if* μ *is equivalent to Lebesgue measure restricted to* $[a, b]$*.*

In view of Proposition 4.1, in all that follows we shall assume that μ is equivalent either to Lebesgue measure or to the restriction of Lebesgue measure to $[a,\, b] \supset \operatorname{supp}(X_0)$.

From an inferential viewpoint, especially if the goal is forecasting, a crucial aspect of the presented model is that the dynamic behavior of (F_t) is entirely determined by the vector process $\boldsymbol{\eta}_t := (\eta_{t1}, \dots, \eta_{td})$. In other words, the (in principle) infinite-dimensional process of random cdf's is driven by a d-dimensional process, from which dynamic aspects of the former can be studied. In applications this property can also be interpreted as an identification condition, which serves as further justification for modeling the F_t as taking values in a finite dimensional space. Unfortunately neither the distribution process (F_t) nor the latent process $(\boldsymbol{\eta}_t)$ are observable by the statistician. Indeed in a first stage all one observes is the process (X_τ) in each cycle, and its associated empirical distribution function G_t. As shown in Proposition 4.2 below however, under suitable conditions one can hope to recover the η_{tj}'s through observation of (X_τ) only. In this direction, define

$$\widehat{R}_\mu(x, y) = \sum_{k=1}^{p} \int \widehat{C}_k(x, z)\widehat{C}_k(y, z) d\mu(z). \tag{4.8}$$

and let $\widehat{R}^\mu$ be the integral operator with kernel $\widehat{R}_\mu$. Denote by $\widehat{\psi}_1, \dots, \widehat{\psi}_{d_0}$ its orthonormal eigenfunctions. For large sample sizes one will have $d_0 \geq d$ typically. See Section 4.2.1 below for a straightforward estimation procedure which relies on simple matrix analysis. Now put

$$\widehat{\eta}_{tj} := \left\langle G_t - \widehat{F}, \widehat{\psi}_j \right\rangle.$$

The following result shows that there is a bound on how far $\widehat{\eta}_{tj}$ and η_{tj} can be one from another.

Proposition 4.2. *Let* (F_t, X_τ) *be a weakly conjugate process satisfying Assumptions F, D and X. Then*

$$|\widehat{\eta}_{tj} - \eta_{tj}| \leq \|G_t - F_t\| + \|\widehat{F} - F\| + 2|\mu|^{1/2} \|\widehat{\psi}_j - \psi_j\|. \tag{4.9}$$

If supp(X_0) *is bounded and the* F_t *are absolutely continuous with respect to* $\mu \equiv$ *Lebesgue measure restricted to* $[a,b]$, *with* F_t' *continuous over* (a,b), *then*

$$(4.10)\quad |\widehat{\eta}_{tj} - \eta_{tj}| \le \left| \mathbb{E}[\Psi(X_{1t})|\mathscr{F}] - \frac{1}{q_t}\sum_{i=1}^{q_t} \Psi(X_{it}) \right| + \|\widehat{F} - F\| + 2(b-a)^{1/2}\|\widehat{\psi}_j - \psi_j\|,$$

where $\Psi(x) = \int_a^x \psi(v)dv$. *The bounds in* (4.9) *and* (4.10) *hold for* $j = 1,\dots,d$ *and* $t = 1,\dots,n$, *almost surely.*

Proposition 4.2 shows that the accuracy of approximating the η_{tj}'s by $\widehat{\eta}_{tj}$ will depend on further assumptions on (F_t, X_τ). Asymptotics on n alone will in general not suffice: one must also control for the term $\|G_t - F_t\|$ via a (conditional) Glivenko-Cantelli type result, or for the term in (4.10) via a (conditional) Law of Large Numbers. In any case, this will depend on q_t. Let us first consider convergence of the terms $\|\widehat{F} - F\|$ and $\|\widehat{\psi}_j - \psi_j\|$. An important property of weakly conjugate processes is that a LLN for (G_t) holds under weak assumptions on the sampling scheme.

Theorem 4.1. *Let* (F_t, X_τ) *be a weakly conjugate process satisfying Assumptions F, D and X. Suppose further that the sampling scheme satisfies the following conditions.*

(i) the sequence (F_t) *is* $\sqrt{n}$*-ergodic in probability:* $\|n^{-1}\sum_{t=1}^n F_t - F\| = O_{\mathbb{P}}\big(n^{-1/2}\big)$;

(ii) the intra-cycle sample sizes are uniformly bounded on t: $q_t \le q^*$ *for all* t;

(iii) for $k = 1,\dots,q^*$ *the limit* $\lim_{n\to\infty}(1/n)\#\{t \le n : q_t = k\}$ *exists.*

Then it holds that

$$\|\widehat{F} - F\| = O_{\mathbb{P}}\big(n^{-1/2}\big).$$

In particular, the above holds when $q_t = q^*$ *for all* t.

Regarding $\|\widehat{\psi}_j - \psi_j\|$ further assumptions on the process (F_t, X_τ) may be needed. The following result, which is a corollary to Theorem 1 in Bathia et al. (2010), gives sufficient conditions for $\sqrt{n}$-consistency of $\widehat{\psi}_j$.

Theorem 4.2 (Bathia et al. (2010), Theorem 1.)**.** *Let* (F_t, X_τ) *be a weakly conjugate process satisfying Assumptions F, D and X and the following conditions.*

(i) (G_t) *is a strongly stationary* ψ*-mixing sequence with the mixing coefficient satisfying the condition* $\sum_{k=1}^{\infty} k\,\psi^{1/2}(k) < \infty$;

(ii) the nonzero eigenvalues of R^μ *are all distinct.*

Then it holds that

$$\|\widehat{\psi}_j - \psi_j\| = O_{\mathbb{P}}\big(n^{-1/2}\big)$$

for all $j = 1,\dots,d$.

The conditions in Theorem 4.2 correspond to assumptions C1 and C3 from Theorem 1 in Bathia et al. (2010). These are quite technical and can be hard to check in each given example. It is important to notice though that since C1 imposes restrictions on the process (G_t), it will likely involve properties of both F_t and X_τ jointly.

It remains to consider the term $\|G_t - F_t\|$. First notice that whereas Theorem 4.1 and Theorem 4.2 ensure *unconditional* consistency of $\widehat{F}$ and $\widehat{\psi}_j$ respectively, one can in principle only expect that $G_t \to F_t$ *conditionally* (on $\mathscr{F}$), by making q_t large and imposing some ergodicity condition on the process $(X_\tau|\mathscr{F} : \tau \in [t, t+1))$. To illustrate how to bound $\|G_t - F_t\|$ unconditionally, assume that $(X_\tau|\mathscr{F} : \tau \in [0, 1))$ is iid F_0. If this process can be sampled at an arbitrary rate (which means we can make $q_0 \to \infty$), then the Glivenko-Cantelli Theorem gives

$$\lim_{q_0 \to \infty} \mathbb{P}[\|G_0 - F_0\| > \epsilon|\mathscr{F}] = 0, \qquad \text{almost surely.}$$

But $\mathbb{P}[\|G_0 - F_0\| > \epsilon|\mathscr{F}]$ is bounded by 1 almost surely, and thus the Lebesgue Dominated Convergence Theorem gives $\mathbb{P}[\|G_0 - F_0\| > \epsilon] \to 0$.

The above example shows that for $\|G_t - F_t\| = o_{\mathbb{P}}(1)$ to hold (unconditionally), one will typically need to rely on a not too strong dependence structure of the process $(X_\tau|\mathscr{F} : \tau \in [t, t+1))$, and on the possibility of sampling at a rate such that q_t can be taken large. In that regard let us say that a weakly conjugate process (F_t, X_τ) is δ-*conjugate* if there exist a $\delta > 0$ and a sampling scheme $\{X_{it}\}$ such that $\mathbb{P}[\|G_t - F_t\| > \delta] \leq \delta$, for all t. We can now state the following.

Corollary 4.1. *Let (F_t, X_τ) be a weakly δ-conjugate process satisfying Assumptions F, D and X. Then, for all $\epsilon > 0$, provided n is large enough,*

$$\mathbb{P}\left[\max_{1\leq t\leq n} |\widehat{\eta}_{tj} - \eta_{tj}| > \delta + \epsilon\right] \leq \delta + \epsilon,$$

for all $j = 1, \ldots, d$.

Of course every weakly conjugate process satisfying condition F is δ-conjugate with $\delta = \sqrt{|\mu|}$, but in general we will be thinking of the least such δ. What the Corollary says is that for weakly δ-conjugate processes the sample paths of $(\widehat{\eta}_{tj})$ and (η_{tj}) are eventually uniformly close, with a large probability and an approximation error of at most $\delta + \epsilon$, for arbitrary $\epsilon > 0$.

An important remark on the present methodology is that although the random variables η_{tj} can be recovered under an adequate sampling scheme and large sample sizes, recovering the cdf's F_t is not as straightforward as it would seem. First of all, there is the issue of estimating the dimension d which is unknown to the statistician. Secondly, even if d were known, the natural estimator $\widehat{F}_t := \widehat{F} + \sum_{j=1}^{d} \widehat{\eta}_{tj}\widehat{\psi}_j$ will generally not satisfy any restriction on its shape, as it should in the present case. That is, even though $\widehat{F}_t$ will be close to F_t in the $L^2(\mu)$ norm, nothing grants that it will be nondecreasing or have its values strictly between 0 and 1. If the interest were to obtain estimators for the true F_t (other than G_t), then a convolution type filter would have to be applied to $\widehat{F}_t$ to obtain a cdf.

4.2.1 Estimation procedure

This section describes how one can obtain estimates of the ψ_j and η_{tj} through straightforward matrix analysis. This approach is adopted by Bathia et al. (2010). The idea is to represent the operator $\widehat{R}^\mu$ as an infinite matrix acting on the canonical Hilbert space ℓ^2, and then obtain a $(n-p)\times(n-p)$ matrix whose spectrum coincides with that of $\widehat{R}^\mu$. The construction relies on the fact that given any operators A and B, it is always true that AB^* and B^*A share the same nonzero eigenvalues. The representation of $\widehat{R}^\mu$ is given by the $\infty\times\infty$ matrix

$$\frac{1}{(n-p)^2}\boldsymbol{G}_0\sum_{k=1}^{p}\boldsymbol{G}_k'\boldsymbol{G}_k\boldsymbol{G}_0',$$

where $\boldsymbol{G}_k = \begin{bmatrix}\boldsymbol{g}_{1+k} & \dots & \boldsymbol{g}_{n-p+k}\end{bmatrix}$ and $\boldsymbol{g}_t$ is a canonical representation of $G_t - \widehat{F}$ in ℓ^2 such that $\boldsymbol{g}_t'\boldsymbol{g}_s = \langle G_t - \widehat{F}, G_s - \widehat{F}\rangle$. Now apply the duality discussed above with $A = \boldsymbol{G}_0$ and $B = \sum_{k=1}^{p}\boldsymbol{G}_k'\boldsymbol{G}_k\boldsymbol{G}_0'$ to obtain the $(n-p)\times(n-p)$ matrix

$$\boldsymbol{M} := \frac{1}{(n-p)^2}\sum_{k=1}^{p}\boldsymbol{G}_k'\boldsymbol{G}_k\boldsymbol{G}_0'\boldsymbol{G}_0. \tag{4.11}$$

To be explicit, the entry (t,s) of $\boldsymbol{G}_k'\boldsymbol{G}_k$ is the inner product $\left\langle G_{t+k} - \widehat{F},\, G_{s+k} - \widehat{F}\right\rangle$. The preceding heuristics establishes the first claim of the following Proposition.

Proposition 4.3 (Bathia et al. (2010), Proposition 1.)**.** *The $(n-p)\times(n-p)$ matrix $\boldsymbol{M}$ shares the same nonzero eigenvalues with the operator $\widehat{R}^\mu$. Moreover, the associated eigenfunctions of $\widehat{R}^\mu$ are given by*

$$\widetilde{\psi}_j(x) = \sum_{t=1}^{n-p}\gamma_{jt}\Big(G_t(x) - \widehat{F}(x)\Big), \tag{4.12}$$

where γ_{jt} is the t-th component of the eigenvector $\boldsymbol{\gamma}_j$ associated to the j-th largest eigenvalue of $\boldsymbol{M}$.

Proof. See Bathia et al. (2010, app. B). □

We then let $\widehat{\psi}_j := \widetilde{\psi}_j/\|\widetilde{\psi}_j\|$ denote the normalized eigenfunctions of $\widehat{R}^\mu$. Notice that in order to obtain the matrix $\boldsymbol{M}$ all one needs is to calculate the inner products $\langle G_t - \widehat{F},\, G_s - \widehat{F}\rangle$ with t and s ranging from 1 to n. An important aspect in our context is that, unlike it is common in general Functional Data Analysis methodologies, the explicit formulas for this coefficients can be easily derived. Indeed,

$$\left\langle G_t - \widehat{F},\, G_s - \widehat{F}\right\rangle = \langle G_t, G_s\rangle - \langle G_t, \widehat{F}\rangle - \langle G_s, \widehat{F}\rangle + \langle \widehat{F}, \widehat{F}\rangle,$$

with

$$\langle G_t, G_s \rangle = \frac{1}{q_t q_s} \sum_{i=1}^{q_t} \sum_{j=1}^{q_s} \mu\Big[X_{it} \vee X_{js},\, +\infty\Big),$$

$$\langle \widehat{F}, \widehat{F} \rangle = \frac{1}{n^2} \sum_{t=1}^{n} \sum_{s=1}^{n} \langle G_t, G_s \rangle,$$

$$\langle G_t, \widehat{F} \rangle = \frac{1}{n} \sum_{s=1}^{n} \langle G_t, G_s \rangle.$$

The norms $\|\widetilde{\psi}_j\|$ can be calculated as well through

$$\|\widetilde{\psi}_j\|^2 = \sum_{t=1}^{n-p} \sum_{s=1}^{n-p} \gamma_{jt}\gamma_{js} \langle G_t - \widehat{F}, G_s - \widehat{F} \rangle,$$

and finally the coefficients $\widehat{\eta}_{tj}$ are given by

$$\widehat{\eta}_{tj} = \frac{1}{\|\widetilde{\psi}\|} \sum_{s=1}^{n-p} \gamma_{js} \langle G_t - \widehat{F}, G_s - \widehat{F} \rangle.$$

4.3 An example

In order to illustrate some further properties of weakly conjugate processes, let us construct a simple example. Some of the arguments of this section will motivate our application to real data below. In this construction μ is Lebesgue measure restricted to the interval $I = [-1, 1]$. Let $d = 1$ and write $\psi_1 \equiv \psi$, and likewise $\eta_{t1} \equiv \eta_t$. Assume $\eta_0, \eta_1, \ldots$ is a stationary AR(1) process, $\eta_t = \alpha\eta_{t-1} + u_t$, where u_t is some centered iid real sequence. Let F be a fixed cdf on I with $\int x\, dF(x) = 0$, and let ψ be some bounded function on $[-1, 1]$ with $\psi(-1) = \psi(1) = 0$. Write $F_t(x) = F(x) + \eta_t\psi(x)$. A straightforward calculation yields $F_t(x) = (1-\alpha)F(x) + \alpha F_{t-1}(x) + u_t\psi(x)$, that is, (F_t) is a linear process as well. It is clear that some restrictions on ψ and on the process (η_t) must be imposed to ensure that the F_t are indeed cdfs, but we relegate the details on how to achieve this to our simulation below. Assuming further that $\int \psi(x)\, dx = 0$ we obtain $\int x dF_t(x) = 0$. Thus any process $(X_\tau : \tau \geq 0)$ satisfying (4.1) will be such that $\mathbb{E}[X_\tau | \mathscr{F}] = 0$. Notice that this assumption is not restrictive. In this context an important object of interest in applications may be the variances $\sigma_t^2 := \int x^2\, dF_t(x)$. Under the linearity conditions just introduced these random variables will satisfy

$$\sigma_t^2 = \beta_0 + \beta_1\eta_t \tag{4.13}$$

$$= \beta_0 + \beta_1\alpha\eta_{t-1} + \beta_1 u_t \tag{4.14}$$

$$= (1-\alpha)\beta_0 + \alpha\sigma_{t-1}^2 + \beta_1 u_t, \tag{4.15}$$

where $\beta_0 = \int x^2\, dF(x)$ and $\beta_1 = \int x^2\, d\psi(x)$. Thus (σ_t^2) is also a linear process, and σ_t^2 and η_t are entirely determined one by another. Now, given observations X_{it}, $i = 1, \ldots, q_t$,

$t = 1, \ldots, n$ from the process (X_τ), is it possible to estimate the parameters α, β_0 and β_1? One possible way to achieve this is to set $\widehat{\beta}_0$ and $\widehat{\beta}_1$ equal to $\int x^2\, d\widehat{F}(x)$ and $\int x^2\, d\widehat{\psi}(x)$ respectively. Alternatively, let $\widehat{\sigma}_t^2$ be the sample variance of $X_{1t}, \ldots, X_{q_t,t}$ and consider the sample counterparts to equations (4.13), (4.14) and (4.15) above,

$$\widehat{\sigma}_t^2 = \beta_0 + \beta_1 \widehat{\eta}_t \quad + \quad (e_t^\sigma - \beta_1 e_t^\eta) \tag{4.16}$$

$$= \beta_0 + \beta_1 \alpha \widehat{\eta}_{t-1} + \beta_1 u_t \quad + \quad (e_t^\sigma - \beta_1 \alpha e_{t-1}^\eta) \tag{4.17}$$

$$= (1-\alpha)\beta_0 + \alpha \widehat{\sigma}_{t-1}^2 + \beta_1 u_t \quad + \quad \left(e_t^\sigma - \alpha e_{t-1}^\sigma\right), \tag{4.18}$$

and likewise

$$\widehat{\eta}_t = \alpha \widehat{\eta}_{t-1} + u_t + (e_t^\eta - \alpha e_{t-1}^\eta), \tag{4.19}$$

where $e_t^\sigma := \widehat{\sigma}_t^2 - \sigma_t^2$ and $e_t^\eta := \widehat{\eta}_t - \eta_t$. The idea is that the latter quantities can be made small if (F_t, X_τ) is δ-conjugate. Thus, if the aim is to forecast future values of σ_t^2, one could use the identities above to propose some forecasting strategies, such as

Strategy 1

(Step 1) Estimate α from data $(\widehat{\eta}_1, \ldots, \widehat{\eta}_n)$;

(Step 2) Estimate β_0 and β_1 from data $(\widehat{\sigma}_t^2, \widehat{\eta}_t : t = 1, \ldots, n)$;

(Step 3) Use the estimated $\widehat{\alpha}$ to forecast $\widehat{\eta}_{n+1}$. Use this forecast together with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to forecast $\widehat{\sigma}_{n+1}^2$ through (4.16).

Strategy 2

(Step 1) Estimate β_0 and $\beta_1\alpha$ from data $(\widehat{\sigma}_t^2 : t = 2, \ldots, n)$, and $(\widehat{\eta}_t : t = 1, \ldots, n-1)$;

(Step 2) Use $\widehat{\eta}_n$ and (4.17) to obtain the forecast for $\widehat{\sigma}_{n+1}^2$.

Strategy 3

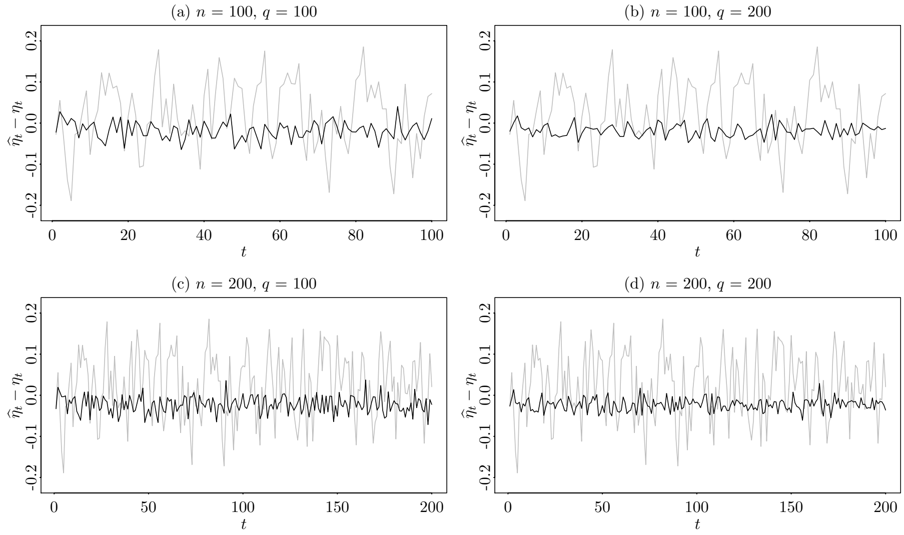
(Step 1) Estimate $(1-\alpha)\beta_0$ and α from data $(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_n^2)$;

(Step 2) Obtain the forecast for $\widehat{\sigma}_{n+1}^2$ through (4.18).

The approaches described above give a benchmark for proposing forecast procedures in more general situations, as in our application to financial data below, where the AR(1) specification may not be the more adequate one, and where $\int x\, dF_t(x) \neq 0$.

Narrowing a little further, let us consider the following special case of the above example. Let F be the cdf corresponding to the uniform distribution over $[-1, 1]$, and let $\psi(x) := \int_{-1}^{x} (1/2 - |v|)\, dv$. Let η_t be a stationary AR(1) process as above, with the innovations u_t being iid uniformly distributed over $[-1 + |\alpha|, 1 - |\alpha|]$. We may assume that the process (u_t) is indexed for $t \in \mathbb{Z}$ and set $\eta_t = \sum_{k=0}^{\infty} \alpha^k u_{t-k}$. Now put $F_t = F + \eta_t \psi$.

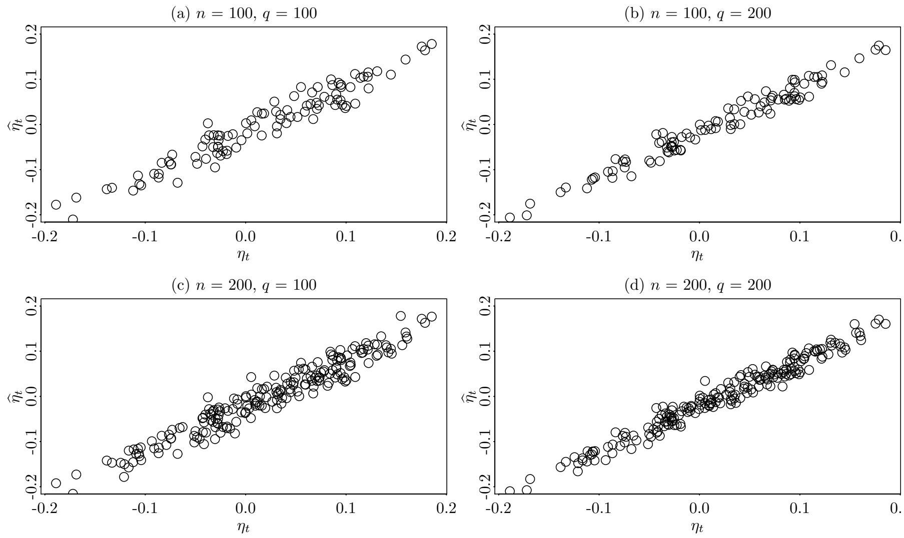
Figure 4.1: Residuals $\widehat{\eta}_t - \eta_t$ with (a) $n = 100$, $q = 100$; (b) $n = 100$, $q = 200$; (c) $n = 200$, $q = 100$; (d) $n = 200$, $q = 200$. Grey: sample path of $\widehat{\eta}_t$.



This model specification is easier to appreciate if we consider the derivatives of F and ψ over (a, b), that is, we gain better insight if we differentiate F_t and study the resulting equation, $f_t = f + \eta_t\psi'$, with $f(x) = (1/2)\mathbb{I}_{[-1,1]}(x)$ and $\psi'(x) = (1/2) - |x|$. First notice that $|\eta_t| \leq 1$ by construction. Now f_t is a probability density function obtained by adding to the Uniform$[-1, 1]$ density a random deformation where the deforming 'parameter' is the function ψ' and the random weights are given by the η_t which lie in $[-1, 1]$. The extreme cases correspond to $\eta_t = 1$, in which case f_t is the triangular distribution over $[-1, 1]$, and to $\eta_t = -1$, in which case f_t is a V-shaped distribution, $f_t(x) = |x|\mathbb{I}_{[-1,1]}(x)$. Any possible realization of f_t is thus a convex combination of the latter two densities. The interpretation is that ψ adds mass to the center of the uniform distribution when $\eta_t > 0$ and adds mass to the 'tail' of that distribution when $\eta_t < 0$. Observe that the proposed ψ is not normalized, but this does not matter since the rescaling would be passed to the η_t's.

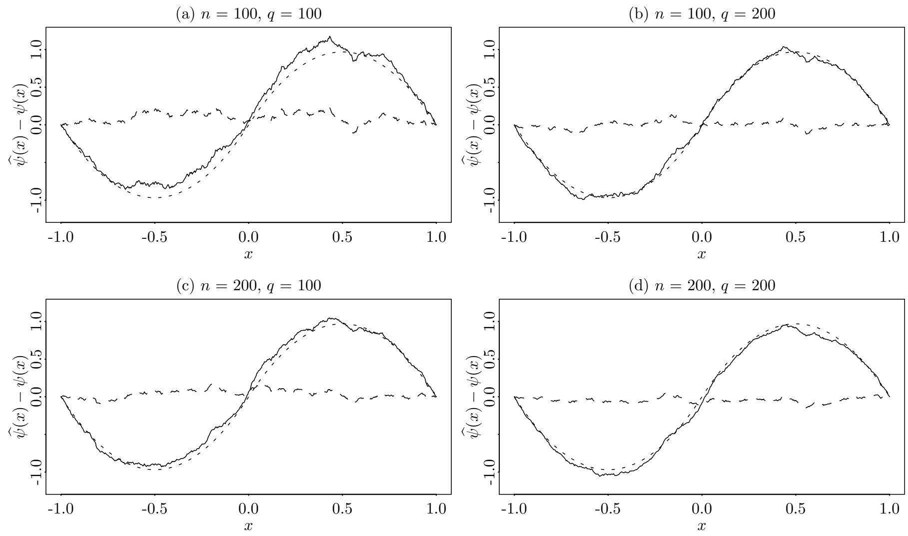
To illustrate, we set $\alpha = 0.5$ and generated a sample $F_1, \ldots, F_{200}$ from the above model and then, for each t, we sampled the X_{it}, $i = 1, \ldots, 200$, independently from F_t. Sampling independently is a simplification but not inconsistent with the present framework, as it may be the case that the process (X_τ) admits an independent sampling scheme at each cycle. Next, we estimate η_t and ψ restricting the data set to $n \leq 200$ cycles and $q \leq 200$ intracycle observations. We consider the following configurations: (i) $n = 100$, $q = 100$; (ii) $n = 100$, $q = 200$; (iii) $n = 200$, $q = 100$ and; (iv) full sample $n = 200$, $q = 200$. Figure 4.1 shows the residuals $\widehat{\eta}_t - \eta_t$, and Figure 4.2 displays the dispersion plots of $(\eta_t, \widehat{\eta}_t)$ in each of these configurations. In this figures it is apparent that increasing the intracycle sample sizes will result in more accurate estimates for the η_t, as one would expect from Proposition 4.2. Figure 4.3 displays the true eigenfunction ψ and the estimates $\widehat{\psi}$, together with the deviations $\widehat{\psi} - \psi$ for each one of the specifications (i)–(iv). These figures

Figure 4.2: Dispersion plots of $(\eta_t, \widehat{\eta}_t)$ with (a) $n = 100$, $q = 100$; (b) $n = 100$, $q = 200$; (c) $n = 200$, $q = 100$; (d) $n = 200$, $q = 200$.



point to the fact that, although Theorem 4.2 ensures that asymptotics on n will suffice for consistency of $\widehat{\psi}$, increasing the intra-cycle sample size may have a positive impact on estimation as well. In this simulation study and in the empirical application below, all computational work was carried out through the softwares `R` and `Julia`.

Figure 4.3: True eigenfunction ψ (dotted), estimated eigenfunction $\widehat{\psi}$ (solid) and deviation $\widehat{\psi} - \psi$ (dashed). (a) $n = 100$, $q = 100$; (b) $n = 100$, $q = 200$; (c) $n = 200$, $q = 100$; (d) $n = 200$, $q = 200$.



4.4 Application to financial data

We apply our methodology to forecast risk in intraday stock market trading. Our sample consists of 5–minute returns for the ITUB4 asset prices; the raw data is available at the Bovespa ftp site. ITUB4 is the main asset in the composition of the Bovespa index. Our sample ranges from July 1st 2012 to April 30 2015, encompassing 719 business days. At each day t the sample $X_{1t}, \cdots, X_{q_t,t}$ consists of $q_t = 79$ observations of the 5–minute return process, defined as the difference of logarithm prices over 5 minutes, ranging from 10:30 AM to 5:00 PM. There are 3 carnival days during the sampling period, at which the intra-day sample sizes are $q_{170} = 47$, $q_{433} = 46$ and $q_{670} = 47$ respectively. Our working assumption is that the X_{it} are sampled from a weakly conjugate process (F_t, X_τ). The empirical distribution functions of 5–minute returns for the first two days in our sample, G_1 and G_2, are plotted in Figure 4.4. In what follows μ is the Laplace(0,1) distribution on the real line.

Figure 4.4: The empirical CDF of 5–minute returns. (a) Day 1; (b) Day 2.

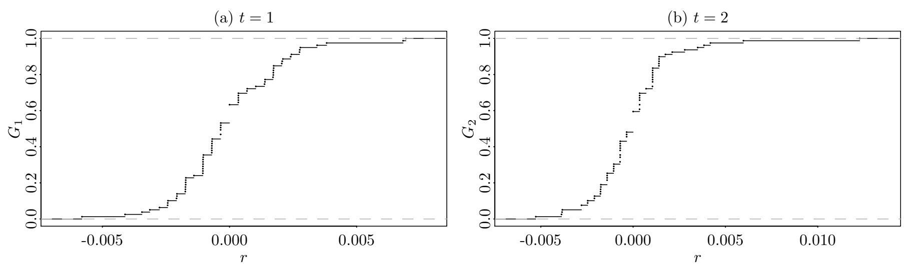


Figure 4.5: Estimated eigenvalues. (a) 1st–10th; (b) 2nd–11th

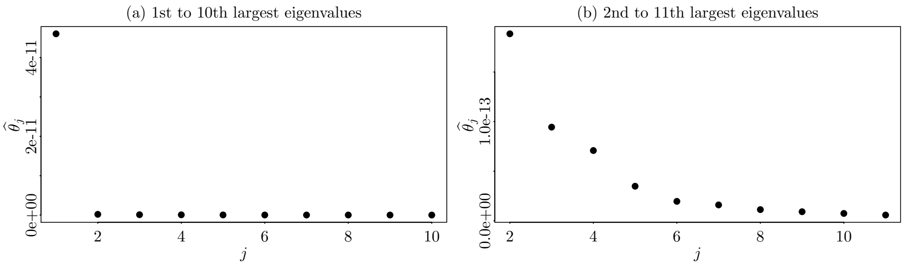
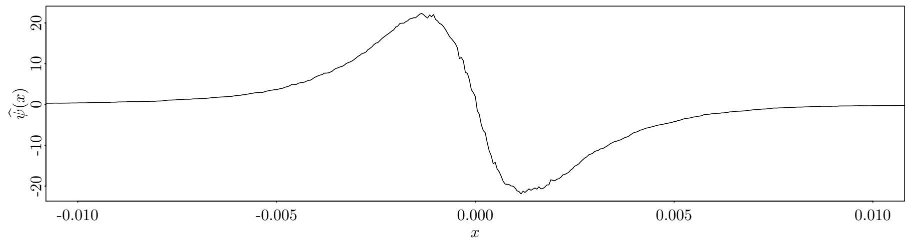


Figure 4.6: Estimated eigenfunction $\widehat{\psi}$.



4.4.1 Data analysis

Figure 4.5 displays the largest estimated eigenvalues $\widehat{\theta}_j$ of $\widehat{R}^{\mu}$. The drop in scale from the largest to the second largest eigenvalue is markedly steep, whereas from the second to the third largest and so on it decays smoothly. Moreover, the p-value from the Ljung-Box test for independence is nearly zero for the time series $(\widehat{\eta}_{t1})$, $t = 1, 2, \dots, n$, whereas for $(\widehat{\eta}_{t2})$ it is 0.8818. This indicates that indeed there is dynamic dependence in the direction of ψ_1 but not in the remaining ones. Observe though that this interpretation must be taken with caution as pointed in Bathia et al. (2010, remark 3). The sample path of the estimated $\widehat{\eta}_{1t}$ are found in Figure 4.7. The plot of the estimated eigenfunction $\widehat{\psi}_1$ is shown in Figure 4.6[1]. It displays a plausible shape whereas the eigenfunction $\widehat{\psi}_2$ is very irregular (the plot is not reported here). In any case we assume $d = 1$ and write $\widehat{\eta}_t \equiv \widehat{\eta}_{t1}$, and likewise $\widehat{\psi} \equiv \widehat{\psi}_1$. We then perform the augmented Dickey-Fuller test to the time series $\widehat{\eta}_t$, and the obtained p-values are virtually zero whatever specification is used, be it with a drift component, a drift and a trend component, or neither. Therefore we take $\widehat{\eta}_t$ to be stationary. Figure 4.8 displays the ACF and PACF plots for $\widehat{\eta}_t$.

We are first interested in forecasting the variance σ_t^2 and the 0.05-th quantile ζ_t corresponding to F_t. The forecasting strategies which we propose follow closely the arguments of Section 4.3. However in the more general scenario the identities (4.13), (4.14) and (4.15), as well as their empirical counterparts, can become more involved. In particular when $\int x\, dF(x) \neq 0$, that is when $X_\tau \big| \mathscr{F}$ is not zero-mean, the relationship between σ_t^2 and η_t is no longer linear but rather of the quadratic form

$$\sigma_t^2 = \beta_0 + \beta_1 \eta_t + \beta_2 \eta_t^2, \tag{4.20}$$

[1]This plot was obtained by considering a centered version of the returns data. If the original data is used instead (as in the rest of our analysis), the resulting eigenfunction appears slightly noisier.

Figure 4.7: Estimated coefficients $\widehat{\eta}_t$.

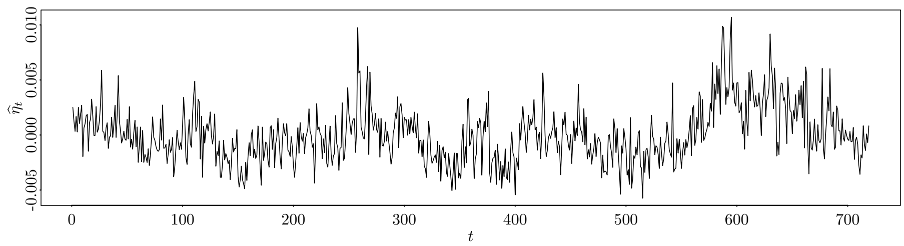
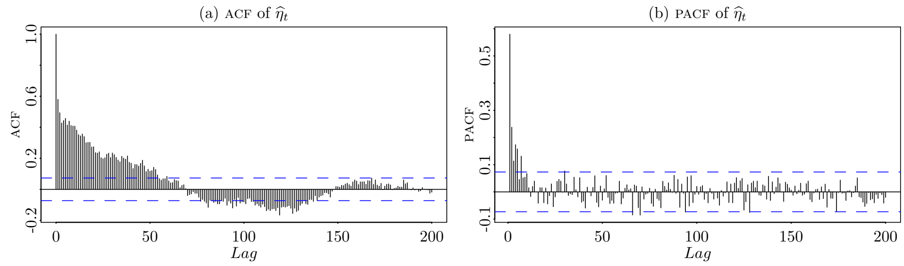


Figure 4.8: Correlation functions of $\widehat{\eta}_t$. (a) ACF; (b) PACF.



where the coefficients β_j are functions of first and second moments of F and ψ. Equations (4.14) and (4.15) in turn will depend on the dynamic behavior of η_t, and even slight deviations from an AR(1) model, say an ARMA(1,1) model, will give expressions with crossed terms in place of (4.14) and (4.15). The empirical equations (4.16), (4.17) and (4.18) will suffer from the same complications if we intend to generalize.

That said, we adopt an exploratory approach to model the time series of $\widehat{\eta}_t$, $\widehat{\sigma}_t^2$, and $\widehat{\zeta}_t$, as well as the interdependence between them. Here $\widehat{\sigma}_t^2$ and $\widehat{\zeta}_t$ are respectively the variance and the 0.05-th quantile of the distribution function G_t. Figure 4.9, panel (a) displays the dispersion plot of $(\widehat{\eta}_{t-1}, \widehat{\eta}_t)$. One sees that a linear model may give a parsimonious description of $(\widehat{\eta}_t)$. Together with the ACF and PACF plots from Figure 4.8, as well as the results from the augmented Dickey-Fuller tests discussed above, we feel authorized to assume that $(\widehat{\eta}_t)$ is an ARMA process. We choose the ARMA(1,2) specification based on the AIC criterium. The estimation results can be found in Table 4.1.

The dispersion plots of $\left(\widehat{\sigma}_{t-1}^2, \widehat{\sigma}_t^2\right)$ and $\left(\log \widehat{\sigma}_{t-1}^2, \log \widehat{\sigma}_t^2\right)$ are displayed in Figure 4.9, panels (c) and (d) respectively. It is apparent that the relationship between $\widehat{\sigma}_t^2$ and its lagged value is highly heteroskedastic; the logarithmic transformation stabilizes and linearizes this interaction, as seen in panel (d). The p-values of the augmented Dickey-Fuller test for both series are virtually zero in each specification (be it with a drift component, a drift and a trend component, or neither). The ACF and PACF plots of $(\widehat{\sigma}_t^2)$ and $(\log \widehat{\sigma}_t^2)$ are both nearly identical to those of $(\widehat{\eta}_t)$ and are not reported here. From this analyses we find it adequate to assume that $(\log \widehat{\sigma}_t^2)$ is an ARMA process, and based on the AIC criterium we select the ARMA(2,1) specification.

In Figure 4.9, panel (b), the dispersion plot of $\left(\widehat{\zeta}_{t-1}, \widehat{\zeta}_t\right)$ is shown. It is seen that a linear model may not be the more adequate description of how $\widehat{\zeta}_t$ interacts with its past

Figure 4.9: Dispersion plots of time series (a) $(\widehat{\eta}_t)$; (b) $(\widehat{\zeta}_t)$; (c) $(\widehat{\sigma}^2_t)$ and; (d) $(\log \widehat{\sigma}^2_t)$.

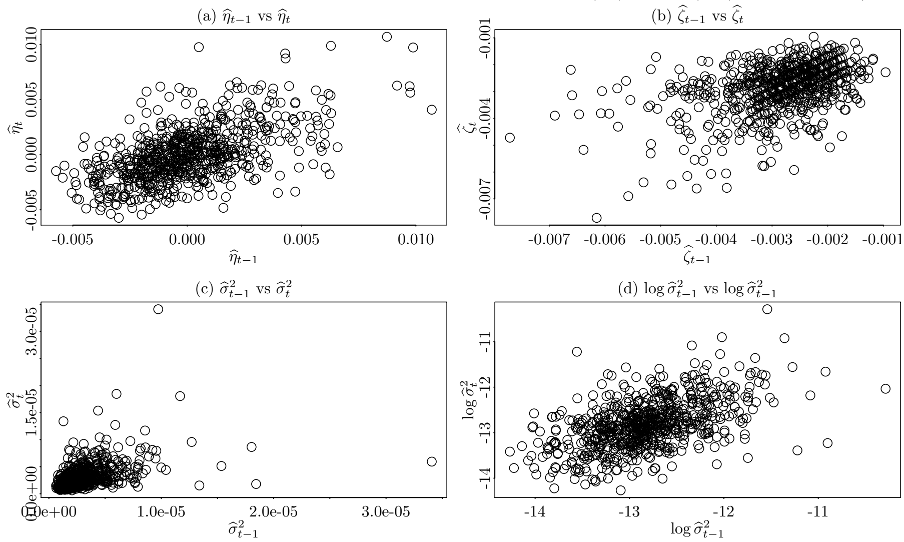
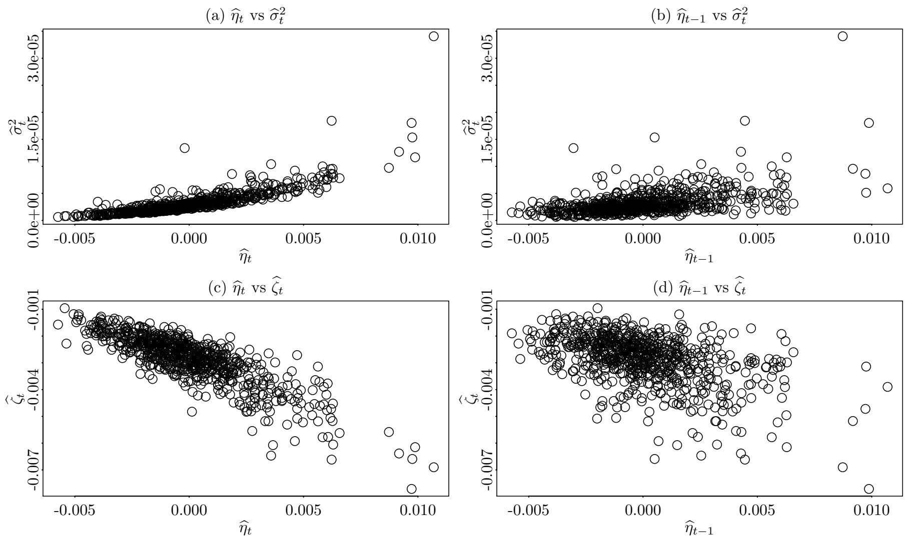


Figure 4.10: Dispersion plots: (a) $(\widehat{\eta}_t, \widehat{\sigma}^2_t)$; (b) $(\widehat{\eta}_{t-1}, \widehat{\sigma}^2_t)$; (c) $(\widehat{\eta}_t, \widehat{\zeta}_t)$ and; (d) $(\widehat{\eta}_{t-1}, \widehat{\zeta}_t)$



values, but it is also true that such a model may provide a parsimonious approximation to the actual DGP. As with the time series discussed above, the augmented Dickey-Fuller tests for the data $\left(\widehat{\zeta}_1, \ldots \widehat{\zeta}_n\right)$ reject the null of unit root in every specification. The ACF and PACF plots for $\left(\widehat{\zeta}_t\right)$ are again nearly identical to those of $\left(\widehat{\eta}_t\right)$ and are not reported here. Less confidently that in the previous cases, we assume an ARMA(3,1) for $\left(\widehat{\zeta}_t\right)$ based on the AIC criterium.

Figure 4.10 displays the dispersion plots of (a) $(\widehat{\eta}_t, \widehat{\sigma}_t^2)$; (b) $(\widehat{\eta}_{t-1}, \widehat{\sigma}_t^2)$; (c) $\left(\widehat{\eta}_t, \widehat{\zeta}_t\right)$; (d) $\left(\widehat{\eta}_{t-1}, \widehat{\zeta}_t\right)$. Panels (a) and (b) indicate that $\widehat{\sigma}_t^2$ depends on $\widehat{\eta}_t$ and on $\widehat{\eta}_{t-1}$ in a nonlinear way, as one would expect from the identity (4.20). Panels (c) and (d) indicate that the 0.05-th quantile $\widehat{\zeta}_t$ depend linearly on both $\widehat{\eta}_t$ and on $\widehat{\eta}_{t-1}$, although some heteroskedasticity appears to be at play.

In order to obtain one-step-ahead forecasts for the quantities σ_t^2 and ζ_t we adopt some forecasting strategies similar to the ones described in Section 4.3. We give the details of how we produce the forecasts for $\widehat{\sigma}_{t+1}^2$. The case of $\widehat{\zeta}_{t+1}$ is entirely analogous. Letting $n_0 := 350$, we generate forecasts $\widehat{\sigma}_{t+1|t,a}^2$, $\widehat{\sigma}_{t+1|t,b}^2$ and $\widehat{\sigma}_{t+1|t,c}^2$, with t ranging in $n_0, \ldots, n-1$, as defined according to the following strategies.

Strategy 1. $\log \widehat{\sigma}_{t+1|t,a}^2$ is the one-step-ahead forecast obtained from an ARMA(2,1) fit to the data $(\log \widehat{\sigma}_1^2, \ldots, \log \widehat{\sigma}_t^2)$;

Strategy 2. $\widehat{\sigma}_{t+1|t,b}^2 = h(\widehat{\eta}_t)$, where h is the local polinomial regression function obtained from fitting the data $((\widehat{\eta}_1, \widehat{\sigma}_2^2), \ldots, (\widehat{\eta}_{t-1}, \widehat{\sigma}_t^2))$;

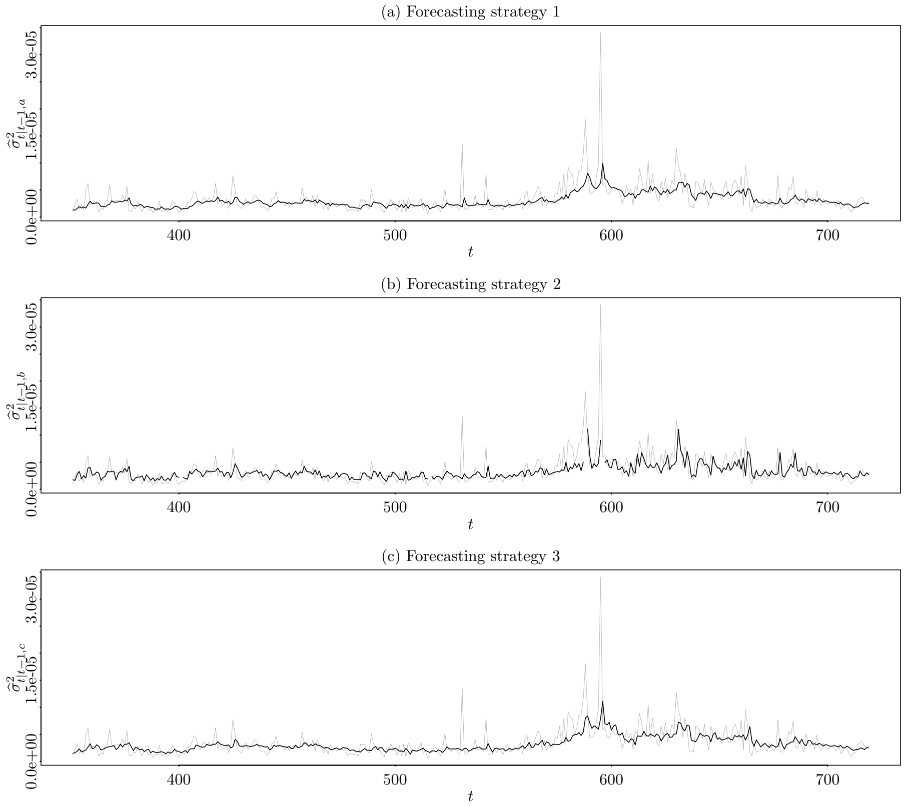
Strategy 3. $\widehat{\sigma}_{t+1|t,c}^2 = h\left(\widehat{\eta}_{t+1|t}\right)$ where h is the local polinomial regression function obtained from fitting the data $((\widehat{\eta}_1, \widehat{\sigma}_1^2), \ldots, (\widehat{\eta}_t, \widehat{\sigma}_t^2))$, and $\widehat{\eta}_{t+1|t}$ is the one-step-ahead forecast of an ARMA(1,2) fit to the data $(\widehat{\eta}_1, \ldots, \widehat{\eta}_t)$.

We are aware that the approach in Strategy 1 is not in the best statistical practice since applying the inverse transformation to a regression fit is not generally valid, but in a comparison (not reported here) the AR(1) regression curve and the median regression curve obtained for the data $(\log \widehat{\sigma}_1^2, \ldots, \log \widehat{\sigma}_n^2)$ were nearly identical, partially validating our approach. In the case of $\widehat{\zeta}_t$, the forecasts are obtained from an ARMA(1,3) fit to the untransformed data $\left(\widehat{\zeta}_1, \ldots, \widehat{\zeta}_t\right)$.

The mean squared errors and relative (to Strategy 1) mean squared errors from each forecast strategy are reported in Table 4.2. Notice that the 'true' quantity being forecasted (for example σ_{t+1}^2) is not observable, not even *ex post*. Thus our forecasts are contrasted with empirical realizations, which are taken as proxies for their population counterparts; for instance the mean squared error of forecasting Strategy 2 above is calculated as

$$\sum_{t=n_0}^{n-1} \left(\widehat{\sigma}_{t+1|t,b}^2 - \widehat{\sigma}_{t+1}^2\right)^2.$$

Figure 4.11: Volatility forecasts.



Heuristically, one would expect that the forecasting strategies which use the $\widehat{\eta}_t$ in their formulation would display better forecasting power since each of the $\widehat{\eta}_t$ is constructed using full sample information, whereas $\widehat{\sigma}_t^2$ and $\widehat{\zeta}_t$ only use information from day t. This reasoning is partially supported as seen from the results displayed in Table 4.2. In any case, applying the Diebold-Mariano test pairwise to each of the obtained forecasts, we cannot reject the null of equal forecasting accuracy. Figure 4.11 contrasts the realized $\widehat{\sigma}_t^2$ with the forecast values obtained through each of the described strategies. Figure 4.12 has a similar interpretation but for $\widehat{\zeta}_t$ instead of the empirical variance.

A last comment on how to interpret the obtained forecasts may come in handy. At the end of day t, the statistician can apply our methodology and obtain, say, a forecast $\widehat{\sigma}^2_{t+1|t}$ for the variance σ^2_{t+1}. The latter quantity is the variance of a 5-minute return at any instant during day $t+1$, as implied by the common marginal distribution of these returns. Thus someone who negotiates in the market in 5-minute intervals has 'typical' variance equal to σ^2_{t+1}; this is the quantity that we are forecasting, and thus $\widehat{\sigma}^2_{t+1|t}$ estimates the 'typical' variability someone who negotiates at each 5 minutes would expect to find next day. In this regard it is convenient to mention that a forecast obtained through a GARCH fit to the data available up to the end of day t would have a different interpretation and thus would not be comparable to our method. Indeed at the end of day t the model will give a forecast for the variance of the "opening" return rather than averaging over day $t+1$.

Figure 4.12: 0.05-th quantile forecasts.

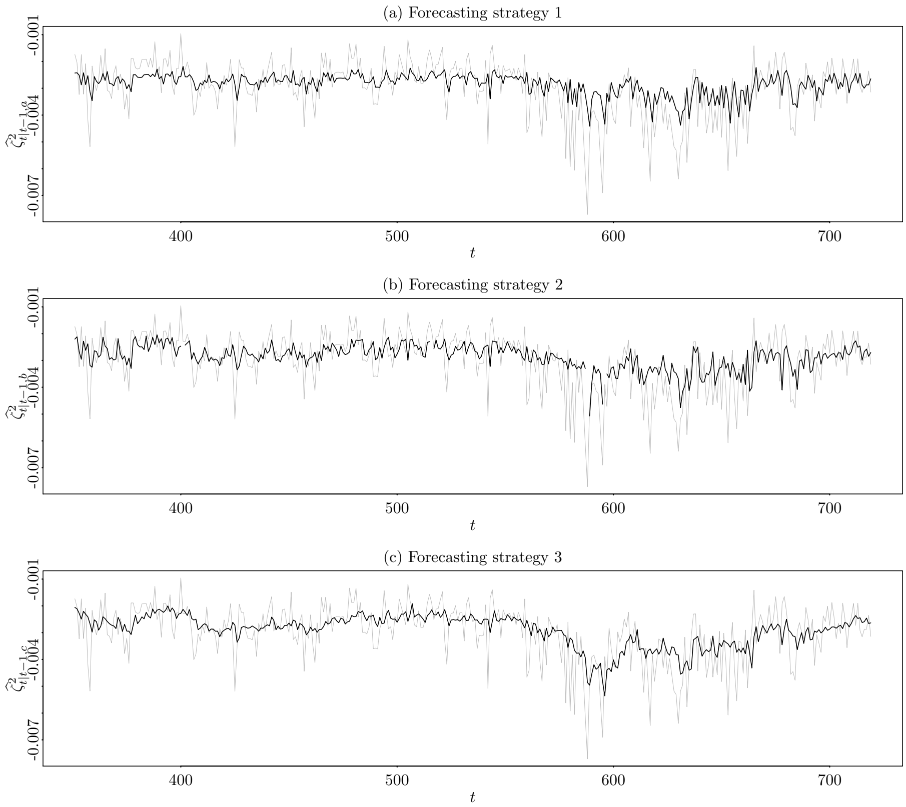


Table 4.1: Coefficient estimates and standard errors of an ARMA(1,2) fit to the data $(\widehat{\eta}_1, \ldots, \widehat{\eta}_n)$.

	AR1	MA1	MA2	intercept
Coef.	0.9709	-0.6452	-0.1302	0e+00
s.e.	0.0115	0.0378	0.0352	5e-04
AIC = -6963.55				

Table 4.2: Mean squared error and relative (to Strategy 1) mean squared error for each of the three forecasting strategies: variance (top 3 rows) and 0.05-th quantile (bottom 3 rows).

	MSE	RMSE
$\widehat{\sigma}^2_{t+1\|t,a}$ (strategy 1)	5.424022e-12 (3)	1
$\widehat{\sigma}^2_{t+1\|t,b}$ (strategy 2)	5.022838e-12 (1)	0.9260357
$\widehat{\sigma}^2_{t+1\|t,c}$ (strategy 3)	5.112571e-12 (2)	0.9425793
$\widehat{\zeta}_{t+1\|t,a}$ (strategy 1)	8.711426e-07 (3)	1
$\widehat{\zeta}_{t+1\|t,b}$ (strategy 2)	7.876192e-07 (2)	0.904122
$\widehat{\zeta}_{t+1\|t,c}$ (strategy 3)	7.256247e-07 (1)	0.8329574

Recall that, as argued, in the framework of weakly conjugate processes, a question of its own interest is identification of the dimension d and characterization of the dynamics of $(\hat{\boldsymbol{\eta}}_t)$. In this regard, we can say that there is some evidence in the data that the true dimension is indeed equal to one, and that the latent process (η_t) is linear. Testing these hypotheses is beyond the scope of the present paper.

4.5 Proofs

In the following proof we use the fact that $\mathrm{C}_b(\mathbb{R}) \subset L^2(\mu)$ for any finite measure μ on $\mathbb{R}$. Also recall that $\|\cdot\|_\infty$ is finer than $\|\cdot\|_2$ and thus a random element in $\mathrm{C}_b(\mathbb{R})$ is also a random element in $L^2(\mu)$.

Proof of Proposition 4.1 . Recall that the property in Assumption D is independent of the choice of basis, as long as the basis is taken to be minimal. Let $\varphi_1, \dots, \varphi_d$ be a minimal linearly independent set in $\mathrm{C}_b(\mathbb{R})$ such that $F_t - F = \xi_{t1}\varphi_1 + \cdots + \xi_{td}\varphi_d$ almost surely. Let C_0^μ be the covariance operator of F_t, acting on $L^2(\mu)$, i.e $C_0^\mu f(x) := \int C_0(x,y) f(y)\, d\mu(y)$, where C_0 is the covariance function $C_0(x,y) = \mathrm{Cov}(F_0(x),\, F_0(y))$. On the one hand it holds that $F_t - F \perp \mathrm{Null}\, C_0^\mu$ almost surely in $L^2(\mu)$, and so $\bigvee_{j=1}^d \varphi_j \subset \overline{\mathrm{Ran} C_0^\mu}$. On the other hand, $C_0(x,y) = \sum_{i=1}^d \sum_{j=1}^d (\mathbb{E}\xi_{0i}\xi_{0j})\varphi_i(x)\varphi_j(y)$, and thus $\bigvee_{j=1}^d \varphi_j = \mathrm{Ran} C_0^\mu$. By Mercer's Theorem (Ferreira and Menegatto (2009), Theorem 1.1), the eigenfunctions of C_0^μ are continuous on $\mathrm{supp}\, \mu$. The condition that μ is equivalent to Lebesgue measure ensures that continuity of the eigenfunctions holds on $\mathbb{R}$. Moreover any eigenfunction of C_0^μ is easily seen to be bounded. Thus we can assume to begin with that $\varphi_1, \dots, \varphi_d$ is an orthonormal set of continuous bounded eigenfunctions of C_0^μ.

For simplicity and without loss of generality, assume $p = 1$ and that Assumption D holds for $k = 1$. Thus R^μ is the integral operator with kernel

$$R_\mu(x,y) = \int C_1(x,z) C_1(y,z)\, d\mu(z).$$

It is easily seen that $R^\mu = C_1^\mu (C_1^\mu)^*$ where C_1^μ is the integral operator with kernel $C_1(x,y)$ and $*$ means adjoining. Then $\overline{\mathrm{Ran} R^\mu} = \overline{\mathrm{Ran} C_1^\mu}$. But

$$C_1(x,y) = \sum_{i=1}^d \sum_{j=1}^d (\mathbb{E}\xi_{0i}\xi_{1j})\varphi_i(x)\varphi_j(y)$$

which implies that $\mathrm{Ran} R^\mu$ is finite dimensional, and a subspace of $\bigvee_{j=1}^d \varphi_j$. Since $(\mathbb{E}\xi_{0i}\xi_{1j})_{ij}$ is full-rank, the reverse inclusion holds. The details can be found in the proof of Proposition 1 in Bathia et al. (2010). □

In the following we use the fact that $\delta_x \in \mathrm{C}_b(\mathbb{R})^*$ to justify that $\mathbb{E}[F_t(x) | \mathscr{F}] = F_t(x)$.

Proof of Lemma 4.1. For item *(i)*, we have

$$\mathbb{E}\varepsilon_t(x) = \mathbb{E}\{\mathbb{E}[G_t(x) - F_t(x)|\mathscr{F}]\} = \mathbb{E}\{\mathbb{E}[G_t(x)|\mathscr{F}] - F_t(x)\}.$$

Now

$$\mathbb{E}[G_t(x)|\mathscr{F}] = \frac{1}{q_t}\sum_{i=1}^{q_t}\mathbb{E}\Big[\mathbb{I}_{\{X_{it}\le x\}}|\mathscr{F}\Big] = F_t(x). \tag{4.21}$$

For *(ii)*, write

$$\begin{aligned}\mathbb{E}[F_t(x)\varepsilon_{t+k}(y)] &= \mathbb{E}\{\mathbb{E}[F_t(x)G_{t+k}(y) - F_t(x)F_{t+k}(y)|\mathscr{F}]\}\\ &= \mathbb{E}\{F_t(x)\mathbb{E}[G_{t+k}(y)|\mathscr{F}] - F_t(x)F_{t+k}(y)\} = 0\end{aligned}$$

by (4.21).

To establish *(iii)* write

$$\begin{aligned}\mathbb{E}[\varepsilon_t(x)\varepsilon_{t+k}(y)] &= \mathbb{E}[(G_t(x) - F_t(x))(G_{t+k}(y) - F_{t+k}(y))]\\ &= \mathbb{E}\{\mathbb{E}[G_t(x)G_{t+k}(y)|\mathscr{F}] - F_{t+k}(y)\mathbb{E}[G_t(x)|\mathscr{F}]\}\\ &\qquad + \mathbb{E}\{F_t(x)F_{t+k}(y) - F_t(x)\mathbb{E}[G_{t+k}(y)|\mathscr{F}]\}\\ &= \mathbb{E}\{\mathbb{E}[G_t(x)G_{t+k}(y)|\mathscr{F}] - F_{t+k}(y)F_t(x)\}\end{aligned}$$

via (4.21) again. Then

$$\mathbb{E}[G_t(x)G_{t+k}(y)] = \frac{1}{q_tq_{t+k}}\sum_{i=1}^{q_t}\sum_{j=1}^{q_{t+k}}\mathbb{E}\Big[\mathbb{I}_{[X_{it}\le x]}\mathbb{I}_{\left[X_{j,t+k}\le y\right]}|\mathscr{F}\Big],$$

but

$$\begin{aligned}\mathbb{E}\Big[\mathbb{I}_{[X_{it}\le x]}\mathbb{I}_{\left[X_{j,t+k}\le y\right]}|\mathscr{F}\Big] &= \mathbb{P}[X_{it}\le x,\ X_{j,t+k}\le y|\mathscr{F}]\\ &= F_t(x)F_{t+k}(y)\end{aligned}$$

by Assumption X. This yields the stated result. □

Proof of Proposition 4.2. Recall that $\|\psi_j\|$ and $\|\widehat{\psi}_j\|$ are equal to 1 by construction, and notice that both $\|\widehat{F}\|$ and $\|G_t\|$ are bounded by $|\mu|^{1/2}$ almost surely, where $|\mu| = \mu(\mathbb{R})$. Also notice that both ψ_j and $-\psi_j$ are normalized eigenfunctions of R^μ. We assume that the 'right' one has been picked.

Now we have

$$\begin{aligned}|\widehat{\eta}_{tj} - \eta_{tj}| &= \Big|\Big\langle G_t - \widehat{F}, \widehat{\psi}_j\Big\rangle - \Big\langle F_t - F, \psi_j\Big\rangle\Big|\\ &\le \Big|\Big\langle G_t - F_t, \psi_j\Big\rangle\Big| + \Big|\Big\langle G_t, \widehat{\psi}_j - \psi_j\Big\rangle\Big| + \Big|\Big\langle F, \psi_j\Big\rangle - \Big\langle \widehat{F}, \widehat{\psi}_j\Big\rangle\Big|.\end{aligned}$$

The second term in the RHS above is bounded by $\|G_t\|\cdot\|\widehat{\psi}_j - \psi_j\|$, whereas the last term is

$$\begin{aligned}\Big|\Big\langle F, \psi_j\Big\rangle - \Big\langle \widehat{F}, \widehat{\psi}_j\Big\rangle\Big| &= \Big|\Big\langle F - \widehat{F}, \psi_j\Big\rangle + \Big\langle \widehat{F}, \psi_j - \widehat{\psi}_j\Big\rangle\Big|\\ &\le \|F - \widehat{F}\|\cdot\|\psi_j\| + \|\widehat{F}\|\cdot\|\psi_j - \widehat{\psi}_j\|.\end{aligned}$$

Noticing that $\left|\left\langle G_t - F_t, \psi_j \right\rangle\right| \leq \|G_t - F_t\| \cdot \|\psi_j\|$ establishes (4.9).

Assume now that μ is Lebesgue measure restricted to $I := [a, b]$, and that the F_t are differentiable on (a, b) with a continuous derivative. We can especialize the expression for $\left|\left\langle G_t - F_t, \psi_j \right\rangle\right|$. First fix some $\omega \in \Omega$. Then

$$\left\langle G_t^\omega, \psi_j \right\rangle = \frac{1}{q_t} \sum_{i=1}^{q_t} \left\langle \mathbb{I}_{[X_{it} \leq \cdot]}(\omega), \psi_j \right\rangle,$$

with

$$\left\langle \mathbb{I}_{[X_{it} \leq \cdot]}(\omega), \psi_j \right\rangle = \int_a^b \mathbb{I}_{[X_{it} \leq v]}(\omega) \psi_j(v)\, dv = \int_{X_{it}(\omega)}^b \psi_j(v)\, dv = \Psi_j(b) - \Psi_j(X_{it}(\omega))$$

since $\mathbb{I}_{[X_{it} \leq v]}(\omega) = \mathbb{I}_{[X_{it}(\omega), +\infty)}(v)$. Regarding $\left\langle F_t, \psi_j \right\rangle$ write

$$(F_t \Psi_j)' = F_t' \Psi_j + F_t \psi_j.$$

Integrating by parts yields

$$\begin{aligned} \left\langle F_t, \psi_j \right\rangle &= \int_a^b F_t(v) \psi_j(v)\, dv \\ &= \int_a^b (F_t \Psi_j)'(v)\, dv - \int_a^b F_t'(v) \Psi_j(v)\, dv \\ &= F_t(b)\Psi_j(b) - F_t(a)\Psi_j(a) - \mathbb{E}[\Psi_j(X_{1t}) | \mathscr{F}] \\ &= \Psi_j(b) - \mathbb{E}[\Psi_j(X_{1t}) | \mathscr{F}], \end{aligned}$$

as $F(b) = 1$ and $F(a) = 0$. Thus

$$\left\langle G_t - F_t, \widehat{\psi}_j \right\rangle = \mathbb{E}[\Psi_j(X_{1t}) | \mathscr{F}] - \frac{1}{q_t} \sum_{i=1}^{q_t} \Psi_j(X_{it}),$$

from which (4.10) follows. □

Proof of Theorem 4.1. We will consider the case $q_t = q^* \equiv q$ for all t. The general case can be obtained through a similar argument by summing t over the sets $\{t \leq n : q_t = k\}$. Let

$$Z_{it}(x) := \mathbb{I}_{[X_{it} \leq x]} - F_t(x).$$

Observe that Z_{it} is a strong order 2 random element in the Hilbert space $L^2(\mu)$. Now notice that

$$\|\widehat{F} - F\| \leq \frac{1}{q} \sum_{i=1}^{q} \left\{ \left\| \frac{1}{n} \sum_{t=1}^{n} Z_{it} \right\| + \left\| \frac{1}{n} \sum_{t=1}^{n} F_t - F \right\| \right\}.$$

The second term in the above sum is $O_{\mathbb{P}}\left(n^{-1/2}\right)$ by assumption. For the first term, we will need the following result.

Lemma 4.2 (Hilbert space Hoeffding Inequality. Boucheron et al. (2013, p. 172))**.** *Let $W_1, \dots, W_n$ be independent, centered random elements in a separable Hilbert space H. If for some $c > 0$ one has $\|W_i\| \leq c/2$ for all i, then for each $\epsilon \geq c/2$ it holds that*

$$\mathbb{P}\left[\left\|\sum_{i=1}^{n} W_i\right\| > \sqrt{n}\epsilon\right] \leq \exp\left(-\frac{(\epsilon - c/2)^2}{c^2/2}\right).$$

By Assumption X, conditional on $\mathscr{F}$, $(Z_{it} : t = 1, 2, \dots)$ is an independent sequence of centered random elements in $L^2(\mu)$, with $\|Z_i\| \leq \sqrt{|\mu|}$. Thus, for $c = 2\sqrt{|\mu|}$, we have

$$\mathbb{P}\left[\left\|\sum_{t=1}^{n} Z_{it}\right\| > \sqrt{n}\epsilon \,\middle|\, \mathscr{F}\right] \leq \exp\left(-\frac{(\epsilon - c/2)^2}{c^2/2}\right), \qquad \text{almost surely.}$$

Taking expectation on both sides yields the stated result. □

Proof of Theorem 4.2. Notice that condition C2 in Bathia et al. (2010) is immediately satisfied in our setting. Their conditions C1 and C3 correspond to the assumptions in Theorem 4.2. Condition C4 there is item *(ii)* in our Lemma 4.1. It only remains to observe that their proof is valid in any separable Hilbert space and not only in $L^2([a, b])$. See Horta and Ziegelmann (2015c). □

4.6 References

Bathia, N., Yao, Q., and Zieglemann, F. A. Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, 38(6):3352–3386, 2010.

Benko, M., Härdle, W., and Kneip, A. Common functional principal components. *The Annals of Statistics*, 37(1):1–34, 2009.

Billingsley, P. *Convergence of probability measures*, volume 493. John Wiley & Sons, 2009.

Bosq, D. *Linear processes in function spaces: theory and applications.* Springer Verlag, 2000.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press, 2013.

Dabo-Niang, S., and Ferraty, F., editors. *Functional and operatorial statistics.* Springer Verlag, 2008.

Damon, J., and Guillas, S. Estimation and simulation of autoregressive Hilbertian processes with exogenous variables. *Statistical Inference for Stochastic Processes*, 8(2): 185–204, 2005.

Ferraty, F., and Vieu, P. *Nonparametric functional data analysis: theory and practice.* Springer Verlag, 2006.

Ferreira, J., and Menegatto, V. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

HALL, P., AND VIAL, C. Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):689–705, 2006.

HORTA, E., AND ZIEGELMANN, F. Conjugate processes. 2015a.

HORTA, E., AND ZIEGELMANN, F. Identifying the spectral representation of hilbertian time series. 2015b.

LEDOUX, M., AND TALAGRAND, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

RAMSAY, J., AND SILVERMAN, B. *Functional data analysis*. Springer Verlag, 1998.

VAKHANIA, N., TARIELADZE, V., AND CHOBANYAN, S. *Probability distributions on Banach spaces*, volume 14. Springer Science & Business Media, 1987.

VAN DER VAART, A. W., AND WELLNER, J. A. *Weak Convergence and Empirical Processes*. Springer, 1996.

VAN NEERVEN, J. Stochastic evolution equations. *OpenCourseWare, TU Delft*, 2008.

5 IDENTIFYING THE SPECTRAL REPRESENTATION OF HILBERTIAN TIME SERIES

EDUARDO HORTA[1] FLÁVIO ZIEGELMANN[2]

November, 2015

Abstract. We provide $\sqrt{n}$-consistency results regarding estimation of the spectral representation of covariance operators of Hilbertian time series, in a setting with imperfect measurements. This is a generalization of the method developed in Bathia et al. (2010). The generalization relies on an important property of centered random elements in a separable Hilbert space, namely, that they lie almost surely in the closed linear span of the associated covariance operator. We provide a straightforward proof to this fact. This result is, to our knowledge, overlooked in the literature. It incidentally gives a rigorous formulation of PCA in Hilbert spaces.

Keywords and phrases. Covariance operator. $\sqrt{n}$-consistency. Hilbertian Time Series. Dimension reduction.

JEL Classification. C1, C14, C22

5.1 Introduction

In this paper, we provide theoretical results regarding estimation of the spectral representation of the covariance operator of stationary Hilbertian time series. This is a generalization of the method developed in Bathia et al. (2010) to a setting of random elements in a separable Hilbert space. The approach taken in Bathia et al. (2010) relates to functional PCA and, similarly to the latter, relies strongly in the Karhunen-Loève (K-L) Theorem. The authors develop the theory in the context of curve time series, with each random curve in the sequence satisfying the conditions of the K-L Theorem which, together with a stationarity assumption, ensures that the curves can all be expanded in the same basis – namely, the basis induced by the zero-lag covariance function of the curves. The idea is to identify the dimension of the space M spanned by this basis (finite by assumption), and to estimate M, when the curves are observed with some degree of error. More specifically, it is assumed that the statistician can only observe the curve time series (Y_t), where

$$Y_t = X_t + \epsilon_t,$$

whereas the curve time series of interest is actually (X_t). Here Y_t, X_t and ϵ_t are random functions (curves) defined on $[0, 1]$. Estimation of M in this framework was previously addressed in Hall and Vial (2006) assuming the curves are iid (in t), a setting in which the problem is indeed unsolvable in the sense that one cannot separate X_t from ϵ_t. Hall and

[1]Department of Statistics – Universidade Federal do Rio Grande do Sul. eduardo.horta@ufrgs.br
[2]Department of Statistics – Universidade Federal do Rio Grande do Sul. flavioz@ufrgs.br

Vial (2006) propose a Deus ex machina solution which consists in assuming that ϵ_t goes to 0 as the sample size grows. Bathia et al. (2010) in turn resolve this issue by imposing a dependence structure in the evolution of (X_t). Their key assumption is that, at some lag k, the k-th lag autocovariance matrix of the random vector composed by the Fourier coefficients of X_t in M, is full rank. In our setting this corresponds to Assumption (A1) (see below).

In Bathia et al. (2010) it is assumed that the stochastic processes $(X_t(u) : u \in [0,1])$ satisfy the conditions of the K-L Theorem (and similarly for ϵ_t), and as a consequence the curves are in fact random elements with values in the Hilbert space $L^2[0,1]$. Therefore, since every separable Hilbert space is isomorphic to $L^2[0,1]$, the idea of a generalization to separable Hilbert spaces of the aforementioned methodology might seem at first rather dull. The issue is that *in applications transforming the data (that is, applying the isomorphism) may not be feasible nor desireable.* For instance, the isomorphism may involve calculating the Fourier coefficients in some 'rule-of-thumb' basis that might yield infinite series even when the curves are actually finite dimensional.

The approach that we take here relies instead on the key feature that a centered Hilbertian random element of strong second order, lies almost surely in the closed linear span of its corresponding covariance operator. This result allows one to dispense with considerations of 'sample path properties' of a random curve by addressing the spectral representation of a Hilbertian random element directly. In other words, the Karhunen-Loève Theorem is just a special case[1] of a more general phenomena. The result below (which motivates – and for that matter, justifies – our approach) is not a new one: it appears, for example, in a slightly different guise as an exercise in Vakhania et al. (1987). However, it is in our opinion rather overlooked in the literature. The proof that we give is straightforward and, to our knowledge, a new one. In this paper H is always assumed to be a real Hilbert space, but with minor adaptations all stated results hold for complex H.

Theorem 5.1. *Let H be a separable Hilbert space and ξ be a centered random element in H of strong second order, with covariance operator R. Then $\xi \perp \ker(R)$ almost surely.*

Corollary 5.1. *In the conditions of Theorem 5.1, let (λ_j) be the sequence of nonzero eigenvalues of R, and (φ_j) be the associated sequence of orthonormal eigenvectors. Then*

(i) $\xi(\omega) = \sum_{j=1}^{\infty} \langle \xi(\omega), \varphi_j \rangle \varphi_j$ *in H almost surely;*

(ii) $\xi = \sum_{j=1}^{\infty} \langle \xi, \varphi_j \rangle \varphi_j$ *in $L^2_{\mathbb{P}}(H)$.*

Moreover, the scalar random variables $\langle \xi, \varphi_i \rangle$ and $\langle \xi, \varphi_j \rangle$ are uncorrelated if $i \neq j$.

---

[1] This is not entirely true since the Karhunen-Loève Theorem states uniform (in $[0,1]$) $L^2(\Omega)$ convergence.

Proofs to the above and subsequent statements are presented in section 5.7.

Remark. Although it is beyond the scope of this work, we call attention to the fact that Theorem 5.1 and Corollary 5.1 provide a rigorous justification of PCA for Hilbertian random elements.

We can now adapt the methodology of Bathia et al. (2010) to a more general setting.

5.2 The model

In what follows $(\Omega, \mathscr{F}, \mathbb{P})$ is a fixed complete probability space. Consider a stationary process $(\xi_t : t \in \mathbb{T})$ of random elements with values in a separable Hilbert space H. Here $\mathbb{T}$ is either $\mathbb{N} \cup \{0\}$ or $\mathbb{Z}$. We assume throughout that ξ_0 is a centered random element in H of strong second order. Of course these conditions are true for all the ξ_t as well, by the stationarity assumption. Now let

$$R_k(h) := \mathbb{E}\langle \xi_0, h\rangle \xi_k, \qquad h \in H,$$

denote the k-th lag autocovariance operator of (ξ_t), and let (λ_j, φ_j) be the sequence of eigenvalue / eigenvector pairs of R_0 (with the φ_j being orthonormal). Corollary 5.1 and the stationarity assumption ensure that the spectral representation

$$\xi_t = \sum_{j=1}^{\infty} Z_{tj}\varphi_j$$

holds almost surely in H, for all t, where the $Z_{tj} := \langle \xi_t, \varphi_j\rangle$ are centered scalar random variables satisfying $\mathbb{E}Z_{tj}^2 = \lambda_j$ for all t, and $\mathbb{E}Z_{ti}Z_{tj} = 0$ if $i \neq j$. In applications, an important case is that in which the above sum has only finitely many terms: that is, the case in which R_0 is a finite rank operator. In this setting, the stochastic evolution of (ξ_t) is driven by a vector process $(\boldsymbol{Z}_t : t \in \mathbb{T})$, where $\boldsymbol{Z}_t = (Z_{t1}, \ldots, Z_{td})$, in $\mathbb{R}^d$ (here d is the rank of R_0). The condition that R_0 is of finite rank models the situation where the statistician's measurements lie (in principle) in an infinite dimensional space, but it is reasonable to assume that they in fact lie in a finite dimensional subspace which must be identified inferentially.

We are interested in modeling the situation where the statistician observes a process $(\zeta_t : t \in \mathbb{T})$ of H valued random elements, and we shall consider two settings; the simplest one occurs when

$$\zeta_t = \xi_t. \tag{5.1}$$

This is to be interpreted as meaning that perfect measurements of a 'quantity of interest' ξ_t are attainable. A more realistic scenario would admit that associated to every measurement there is an intrinsic error – due to rounding, imprecise instruments, etc. In that case observations would be of the form

$$\zeta_t = \xi_t + \epsilon_t. \tag{5.2}$$

In fact, the latter model nests the 'no noise' one if we allow the ϵ_t to be degenerate. Equation (5.2) is analogous to the model considered in Hall and Vial (2006) and in Bathia et al. (2010). Here $(\epsilon_t : t \in \mathbb{T})$ is assumed to be *noise*, in the following sense: (i) for all t, $\epsilon_t \in L^2_{\mathbb{P}}(H)$, with $\mathbb{E}\epsilon_t = 0$; (ii) for each $t \neq s$, ϵ_t and ϵ_s are strongly orthogonal. Moreover we also assume that ϵ_t and ξ_s are strongly orthogonal, for all t and s.

In the above setting, for $h, f \in H$ one has $\mathbb{E}\langle h, \zeta_t\rangle\langle f, \zeta_t\rangle = \langle R_0(h), f\rangle + \mathbb{E}\langle h, \epsilon_t\rangle\langle f, \epsilon_t\rangle$ and thus estimation of R_0 via (ζ_t) is spoiled (unless the ϵ_t are degenerate). This undesirable property has been addressed by Hall and Vial (2006) and Bathia et al. (2010) respectively in the iid scenario and in the time series (with dependence) setting. The clever approach by Bathia et al. (2010) relies on the fact that $\mathbb{E}\langle h, \zeta_t\rangle\langle f, \zeta_{t+1}\rangle = \langle R_1(h), f\rangle$ (lagging filters the noise) and therefore R_1 can be estimated using the data (ζ_t). Now an easy check shows that $\overline{\operatorname{ran}(R_1)} \subset \overline{\operatorname{ran}(R_0)}$. The key assumption in Bathia et al. (2010) is asking that this relation hold with equality:

$$\text{(A1)} \qquad \overline{\operatorname{ran}(R_1)} = \overline{\operatorname{ran}(R_0)}.$$

Assume (A1) holds. Consider the operator $S := R_1 R_1^*$, where $*$ denotes adjoining. It is certainly positive, and compact (indeed nuclear) since $\overline{\operatorname{ran}(R_1 R_1^*)} = \overline{\operatorname{ran}(R_1)}$. Thus, letting (ψ_j) denote the orthonormal sequence of eigenvectors of S, the representation

$$\xi_t = \sum_{j=1}^{\infty} W_{tj}\psi_j$$

is seen to hold, for all t, almost surely in H for centered scalar random variables $W_{tj} = \langle \xi_t, \psi_j\rangle$. Again, when R_0 is finite rank, say $\operatorname{rank}(R_0) = d$, then the stochastic evolution of ξ_t is driven by the vector process $(\boldsymbol{W}_t : t \in \mathbb{T})$ in $\mathbb{R}^d$, where $\boldsymbol{W}_t = (W_{t1}, \ldots, W_{td})$.

5.3 Main results

Before stating the main result, let us establish some notation. Define the estimator $\widehat{S} := \widehat{R}_1 \widehat{R}_1^*$, where $\widehat{R}_1$ is given by

$$\widehat{R}_1(h) := \frac{1}{n-1}\sum_{t=1}^{n-1}\langle \zeta_t, h\rangle \zeta_{t+1}, \qquad h \in H.$$

Let (θ_j, ψ_j) and $\left(\widehat{\theta}_j, \widehat{\psi}_j\right)$ denote the eigenvalue / eigenvector pairs respectively of S and $\widehat{S}$ (dependence on n and on the sample is implied by the 'hat' in notation). For a closed subspace $V \subset H$, let Π_V denote the orthogonal projector onto V. Let $M := \overline{\operatorname{ran}(R_0)}$, and for conformable k put $\widehat{M}_k := \vee_{j=1}^{k}\widehat{\psi}_j$.

Theorem 5.2. *Let* (A1) *and the following conditions hold.*

(A2) $(\zeta_t : t \in \mathbb{T})$ *is strictly stationary and* ψ*-mixing, with the mixing coefficient satisfying the condition* $\sum_{k=1}^{\infty} k\, \psi^{1/2}(k) < \infty$*;*

(A3) $\zeta_t \in L^4_{\mathbb{P}}(H)$, *for all* t*;*

(A4) $\ker(S - \theta_j)$ *is one-dimensional, for each nonzero eigenvalue* θ_j *of* S*;*

(A5) ϵ_t *and* ξ_s *are strongly orthogonal, for all* t *and* s*.*

Then,

(i) $\left\|\widehat{S} - S\right\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right)$*;*

(ii) $\left\|\widehat{\psi}_j - \psi_j\right\| = O_{\mathbb{P}}\left(n^{-1/2}\right)$*, for all* j *such that* $\theta_j > 0$*;*

(iii) $\left|\widehat{\theta}_j - \theta_j\right| = O_{\mathbb{P}}\left(n^{-1/2}\right)$*, for all* j *such that* $\theta_j > 0$*.*

Moreover, if S *is of rank* $d < \infty$*, then*

(iv) $\widehat{\theta}_j = O_{\mathbb{P}}(n^{-1})$*, for all* $j > d$*;*

(v) $\left\|\Pi_M\left(\widehat{\psi}_j\right)\right\| = O_{\mathbb{P}}\left(n^{-1/2}\right)$*, for all* $j > d$*.*

Remark. See Remark 2.1 in Mas and Menneteau (2003) for a comment on Assumption (A4).

Remark. In *(ii)* it is assumed that the 'correct' version of ψ_j (among ψ_j and $-\psi_j$) is being picked. See Lemma 4.3 in Bosq (2000).

Remark. It is important to notice that the operator $\widehat{S}$ is almost surely of finite rank. Hence items *(iii)* and *(iv)* imply the following. If $\mathrm{rank}(S) = d < \infty$, then for $j = 1, \ldots, d$, $\widehat{\theta}_j$ is eventually non-zero and arbitrarily close to θ_j, and the remaining nonzero $\widehat{\theta}_j$ for $j > d$ (if any) are eventually arbitrarily close to zero. Otherwise (if all the eigenvalues θ_j are nonzero) then eventually $\widehat{\theta}_j > 0$ for all j (but notice that this cannot occur uniformly in j). This property can be used to propose consistent estimators of d.

Corollary 5.2. *Let* $N_j := \ker(S - \theta_j)$ *and* $\widehat{N}_j := \ker\left(\widehat{S} - \widehat{\theta}_j\right)$*. Then,*

(i) $\left\|\Pi_{\widehat{N}_j} - \Pi_{N_j}\right\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right)$*, for all* j *such that* N_j *is one-dimensional;*

(ii) if S *is of rank* $d < \infty$*,* $\left\|\Pi_{\widehat{M}_d} - \Pi_M\right\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right)$*;*

(iii) if S *is of rank* $d < \infty$*, there exists a metric* ρ *on the collection of finite-dimensional subspaces of* H *such that* $\rho\left(\widehat{M}_d, M\right) = O_{\mathbb{P}}\left(n^{-1/2}\right)$*.*

Remark. Observe that, when the process (ξ_t) is not centered, evidently all the above results would still hold by replacing ζ_t by $\zeta_t - \mathbb{E}\xi_0$ and ξ_t by $\xi_t - \mathbb{E}\xi_0$, but this is not practical since in general $\mathbb{E}\xi_0$ is not known to the statistician. However, this does not pose a problem, since under mild conditions we have $1/n \sum_{t=1}^n \zeta_t \overset{a.s}{\to} \mathbb{E}\xi_0$, and thus all the results still hold with ζ_t and ξ_t replaced respectively by $\zeta_t - 1/n \sum_{t=1}^n \zeta_t$ and $\xi_t - 1/n \sum_{t=1}^n \zeta_t$.

Remark. The key assumption in Bathia et al. (2010) would be translated in our setting to the condition that, for some $k \geq 1$, the identity $\overline{\operatorname{ran}(R_k)} = \overline{\operatorname{ran}(R_0)}$ holds. For simplicity we have assumed that $k = 1$, but of course the stated results remain true if we take k to be any integer ≥ 1 and redefine S and $\widehat{S}$ appropriately. Indeed the stated results remain true if we define $S = (n-p)^{-1} \sum_{k=1}^{p} R_k R_k^*$, where p is an integer such that $\overline{\operatorname{ran}(R_k)} = \overline{\operatorname{ran}(R_0)}$ holds for some $k \leq p$. In statistical applications, a recommended approach would be to estimate S defined in this manner. In any case, computation of the eigenvalues and eigenvectors of $\widehat{S}$ can be carried out directly through the spectral decomposition of a convenient $n - p \times n - p$ matrix. The method is discussed in Bathia et al. (2010). Notice that if R_0 is of rank one, then asking that $\overline{\operatorname{ran}(R_k)} = \overline{\operatorname{ran}(R_0)}$ holds for some k corresponds to the requirement that the times series $(Z_{t1} : t \in \mathbb{T})$ is correlated at some lag k. Otherwise we would find ourselves in the not very interesting scenario (for our purposes) of an uncorrelated time series.

5.4 Concluding remarks

In this paper we have provided consistency results regarding estimation of the spectral representation of Hilbertian time series, in a setting with imperfect measurements. This generalizes a result from Bathia et al. (2010). The generalization relies on an important property of centered random elements in a separable Hilbert space – see Theorem 5.1. Further work should be directed at obtaining a Central Limit Theorem for the operator $\widehat{S}$, which would have the important consequence of providing Central Limit Theorems for its eigenvalues (via Theorem 2.2 in Mas and Menneteau (2003)), potentially allowing one to propose statistical tests for these parameters.

5.5 Notation and mathematical background

As in the main text we let $(\Omega, \mathscr{F}, \mathbb{P})$ denote a complete probability space, i.e. a probability space with the additional requirement that subsets $N \subset \Omega$ with outer probability zero are elements of $\mathscr{F}$. We assume that $(\Omega, \mathscr{F}, \mathbb{P})$ is large enough that it supports all the random variables considered; by Kolmogorov's Extension Theorem this assumption is legitimate. Let H be a separable Hilbert space with inner-product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. A Borel measurable[2] map $\xi : \Omega \to H$ is called a *random element with values in H* (also: Hilbertian random element). For $q \geq 1$, if $\mathbb{E}\|\xi\|^q < \infty$ we say that ξ is of *strong order* q and write $\xi \in L^q_{\mathbb{P}}(H)$. In this case, there is a unique element $h_\xi \in H$ satisfying the identity $\mathbb{E}\langle \xi, f \rangle = \langle h_\xi, f \rangle$ for all $f \in H$. The element h_ξ is called the *expectation* of ξ and is denoted be $\mathbb{E}\xi$. If $\mathbb{E}\xi = 0$ we say that ξ is *centered.* If ξ and η are centered random elements in H of strong order 2, they are said to be (mutually) strongly orthogonal if, for each $h, f \in H$, it holds that $\mathbb{E}\langle h, \xi \rangle \langle f, \eta \rangle = 0$.

[2]There are notions of strong and weak measurability but for separable spaces they coincide.

Denote by $\mathcal{L}(H)$ the Banach space of bounded linear operators acting on H. Let $A \in \mathcal{L}(H)$. If for some (and hence, all) orthonormal basis (e_j) of H one has $\|A\|_2 := \sum_{j=1}^{\infty} \|A(e_j)\|^2 < \infty$, we say that A is a *Hilbert-Schmidt operator*. The set $\mathcal{L}_2(H)$ of Hilbert-Schmidt operators is itself a separable Hilbert space with inner-product $\langle A, B\rangle_2 = \sum_{j=1}^{\infty} \langle A(e_j), B(e_j)\rangle$, with $\|\cdot\|_2$ being the induced norm. An operator $T \in \mathcal{L}(H)$ is said to be *nuclear*, or *trace-class*, if $T = AB$ for some Hilbert-Schmidt operators A and B. If $\xi \in L^2_{\mathbb{P}}(H)$, its *covariance operator* is the nuclear operator $R_\xi(h) := \mathbb{E}\langle \xi, h\rangle \xi$, $h \in H$. More generally, if $\xi, \eta \in L^2_{\mathbb{P}}(H)$, their *cross-covariance* operator is defined, for $h \in H$, by $R_{\xi,\eta}(h) := \mathbb{E}\langle \xi, h\rangle \eta$. In the main text we denote by R_k the cross-covariance operator of ξ_0 and ξ_k.

For a survey on strong mixing processes, including the definition of ψ-mixing, we refer the reader to Bradley et al. (2005).

Below is a statement of Mercer's Lemma and the Karhunen-Loève Theorem. It is to be contrasted with Theorem 5.1.

Theorem 5.3 (Mercer Lemma and Karhunen-Loève Theorem)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let $(\xi(u) : u \in [0,1])$ be a real stochastic process satisfying the following conditions.*

(KL1) the map $(\omega, u) \mapsto \xi(\omega, u)$ is measurable;

(KL2) $\mathbb{E}\xi(u)^2 < \infty$ and $\mathbb{E}\xi(u) = 0$ for all $u \in [0,1]$;

(KL3) the covariance function $r(u,v) := \mathbb{E}\xi(u)\xi(v)$ is continuous for $u, v \in [0,1]$.

Then there exists a sequence (φ_j) os continuous functions on $[0,1]$ and a non-increasing sequence (λ_j) of nonnegative numbers such that

$$\int_0^1 r(u,v)\varphi_j(v)\,dv = \lambda_j \varphi_j(u), \qquad u \in [0,1], \quad j \in \mathbb{N},$$

with, for $i \neq j$,

$$\int_0^1 \varphi_j(u)^2\,du = 1 \qquad \textit{and} \qquad \int_0^1 \varphi_i(u)\varphi_j(u)\,du = 0.$$

The covariance function r admits the representation,

$$r(u,v) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(u)\varphi_j(v), \qquad s, t \in [0,1],$$

where the series converges uniformly on $[0,1]^2$. Hence

$$\sum_{j=1}^{\infty} \lambda_j = \int_0^1 r(u,u)du < \infty.$$

Moreover,

$$\sup_{u\in[0,1]} \mathbb{E}\Big(\Big|\xi(u) - \sum_{j=1}^{n} Z_j \varphi_j(u)\Big|^2\Big) \underset{n\to\infty}{\longrightarrow} 0,$$

where (Z_j) *is a sequence of mutually uncorrelated real valued, zero mean random variables with* $\mathbb{E}Z_j^2 = \lambda_j$. *In particular,* ξ *defines a random element with values in the Hilbert space* $L^2[0,1]$.

5.6 Comments and references

Probability in Banach spaces first rose from the need to interpret stochastic processes as random variables with values in function spaces. An early insight is due to Wiener. It turns out that a convenient and quite general approach is to consider probability measures in metric spaces. This is well stablished in Billingsley (2009). A modern account can be found in Van Der Vaart and Wellner (1996). Observe that, as in the real and finite dimensional case, the approach is twofold and mostly a matter of convenience; one may either consider the random variables taking values in that spaces, or the (push-forward) probability measures induced by the random variables (and forget about random variables altogether). Usually considering the probability measures alone ends up being a more elegant treatment of the topic, but at the sacrifice of the intuitive appeal that the 'language' of random variables brings. Ledoux and Talagrand (1991) give a thorough account of the slightly more restrictive case of random variables taking values in Banach spaces (and hence of the probability measures in those spaces), and the authors are mostly interested in independent draws (product measures) of such random variables. A LLN for independent sequences of random elements in a Banach space appears there as a Corollary. A CLT in Hilbert spaces is given in Bosq (2000). For stationary sequences and linear processes in Banach spaces, the monograph from Bosq (2000) is a complete account. Specialized versions of the LLN and CLT for dependent sequences can be found therein. The theory of Bochner and Pettis integrals (expectation in Banach spaces) is straightly linked to the theory of probability in Banach spaces. A very clear exposition is given in the first chapters of van Neerven (2008).

The Mercer Lemma appeared rather early in Mercer (1909), when the connection with operator theory wasn't yet completely clear. A quite general statement can be found in Ferreira and Menegatto (2009). The connection between Mercer kernels and the theory of Reproducing Kernel Hilbert Spaces is clarified in Ferreira and Menegatto (2012). Bosq (2000) gives a proof to the Karhunen-Loève Theorem 5.3.

In the literature (Bosq (2000), Mas and Menneteau (2003), Vakhania et al. (1987)) the term *nuclear operator* is often employed; the notion of nuclear (and of p-nuclear, $0 < p \leq 1$) operator is one that applies more generally to Banach spaces and is due to Grothendieck. For Hilbert spaces, there is also the notion of p-Schatten-von Neumann operators ($0 \leq p < \infty$). The cases $p = 1$ and $p = 2$ deserve special names, respectively trace-class operator and Hilbert-Schmidt operator. The relevant fact is that, for $p = 1$, an operator acting on H is 1-nuclear if and only if it is trace-class. Notice that the definition of a trace-class operator that appears in Vakhania et al. (1987) is not the most generally used. For a review, see Hinrichs and Pietsch (2010).

The term 'spectral' in the title of this work refers, of course, to the spectral representation of the operator S and not to the spectral representation of the time series (ξ_t) in the usual sense.

5.7 Proofs

Proof of Theorem 5.1. Let (e_j) be a basis of $\ker(R)$. It suffices to show that $\mathbb{E}|\langle \xi, e_j\rangle|^2 = 0$ for each j. Indeed, this implies that there exist sets E_j, $\mathbb{P}(E_j) = 0$ and $\langle \xi(\omega), e_j\rangle = 0$ for $\omega \notin E_j$. Thus $\langle \xi(\omega), e_j\rangle = 0$ for all j as long as $\omega \notin \bigcap E_j$ with $\mathbb{P}(\bigcap E_j) = 0$. But $\mathbb{E}|\langle \xi, e_j\rangle|^2 = \mathbb{E}\langle \xi, e_j\rangle\langle \xi, e_j\rangle = \mathbb{E}\langle\langle \xi, e_j\rangle\xi, e_j\rangle = \langle \mathbb{E}\langle \xi, e_j\rangle\xi, e_j\rangle = \langle R(e_j), e_j\rangle = 0$. □

Proof of Corollary 5.1. Item (i) is just another way of stating the Lemma. For item (ii), first notice that the functions $\omega \mapsto \langle \xi(\omega), \varphi_j\rangle\varphi_j$, $j \geq 1$, form an orthogonal set in $L^2_{\mathbb{P}}(H)$ (although not *orthonormal*). We must show that $\int \|\xi(\omega) - \sum_{j=1}^n \langle \xi(\omega), \varphi_j\rangle\varphi_j\|^2 d\mathbb{P}(\omega) \to 0$. Let $g_n(\omega) := \|\xi(\omega) - \sum_{j=1}^n \langle \xi(\omega), \varphi_j\rangle\varphi_j\|$. By item (i) $g_n(\omega) \to 0$ almost surely. Also, $0 \leq g_n(\omega) \leq 2\|\xi(\omega)\|$. So $g_n^2(\omega) \to 0$ and $g_n^2(\omega) \leq 4\|\xi(\omega)\|^2$. Now apply Lebesgue's Dominated Convergence Theorem. □

Proof of Theorem 5.2. One only has to consider an isomorphism $U : H \to L^2[0, 1]$. The proof is the same as in Bathia et al. (2010). See also Theorem 2.1 and Proposition 3.1 in Mas and Menneteau (2003). □

Proof of Corollary 5.2. Items (i) and (ii) follow from Proposition 3.1 in Mas and Menneteau (2003). Item (iii) is just Theorem 2 in Bathia et al. (2010). □

Remark. The hypothesis that ξ is centered in Theorem 5.1 cannot be weakened, as the following simple example shows. Let $H = \mathbb{R}^2$ and let $\xi = (\xi_1, \xi_2)$ where ξ_1 is a (real valued) standard normal and $\xi_2 = 1$ almost surely. Then $R \equiv (R_{ij})$ is the matrix with all entries equal to zero except for R_{11} which is equal to 1, and obviously one has $\mathbb{P}(\xi \perp \ker(R)) = 0$.

5.8 References

Bathia, N., Yao, Q., and Zieglemann, F. A. Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, 38(6):3352–3386, 2010.

Billingsley, P. *Convergence of probability measures*, volume 493. John Wiley & Sons, 2009.

Bosq, D. *Linear processes in function spaces: theory and applications.* Springer Verlag, 2000.

Bradley, R. C., et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2(2):107–144, 2005.

Ferreira, J., and Menegatto, V. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

Ferreira, J., and Menegatto, V. Reproducing kernel Hilbert spaces associated with kernels on topological spaces. *Functional Analysis and its Applications*, 46(2):152–154, 2012.

Hall, P., and Vial, C. Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):689–705, 2006.

Hinrichs, A., and Pietsch, A. p-nuclear operators in the sense of Grothendieck. *Mathematische Nachrichten*, 283(2):232–261, 2010.

Ledoux, M., and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Mas, A., and Menneteau, L. *Perturbation approach applied to the asymptotic study of random operators*. Springer, 2003.

Mercer, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441–458):415–446, 1909.

Vakhania, N., Tarieladze, V., and Chobanyan, S. *Probability distributions on Banach spaces*, volume 14. Springer Science & Business Media, 1987.

Van Der Vaart, A. W., and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer, 1996.

van Neerven, J. Stochastic evolution equations. *OpenCourseWare, TU Delft*, 2008.

6 CONCLUDING REMARKS

The present Thesis is composed of four research papers in two distinct areas. The first of these papers, Horta, Guerre, and Fernandes (2015) – Chapter 2 of this Thesis – deals with the topic of quantile regression models. Quantile regression has emerged in its modern formulation through the seminal paper by Koenker and Bassett (1978), and has since become both an object of theoretical interest and an important tool in applications (see Koenker (2005)). The standard quantile regression estimator minimizes an empirical counterpart to the population objective function, of which the true parameter is a minimizer, but unfortunately smoothness properties of the population objective function are not inherited by its sample analogue, a drawback when it comes to inference. As reviewed for instance in Koenker (1994), Buchinsky (1995), Koenker (2005), Fan and Liu (2013), Goh and Knight (2009), computation of asymptotic confidence intervals for components of the standard quantile regression estimator is not straightforward, an issue which has been widely investigated in the literature (Koenker (2005), Buchinsky (1995)). In Horta, Guerre, and Fernandes (2015), we propose a convolution-type smoothing of the sample objective function, from which a smoothed estimator is attainable. We provide a uniform Bahadur-Kiefer representation of the proposed estimator and show that its asymptotic rate dominates that of the standard quantile regression estimator. Next, we prove that the bias introduced by smoothing is negligible in the sense that the bias term is first-order equivalent to the true parameter. A precise rate of convergence, which is controlled uniformly by choice of bandwidth, is provided. We then study second-order properties of the smoothed estimator, in terms of its asymptotic mean squared error, and show that it improves on the usual estimator when an optimal bandwidth is used. As corollaries to the above, one obtains that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal. Next, we provide a consistent estimator of the asymptotic covariance matrix which does not depend on ancillary estimation of nuisance parameters, and from which asymptotic confidence intervals are straightforwardly computable. The quality of the method is then illustrated through a simulation study. As for future work, important inquiries remain open. Obtaining an Edgeworth expansion for the t-statistic associated to the smoothed estimator, thus giving a finer qualitative assertion about its asymptotic normality, is a primary goal. Such expansions can be obtained, for example, through strong approximation methods as in Portnoy (2012). It is also important to study data-driven bandwidth choices, for instance bandwidths selected through cross-validation methods, as well as bootstrap techniques aimed at refined inferential results.

The research papers Horta and Ziegelmann (2015a;b;c) are all related in the sense that they stem from an initial impetus of generalizing the results in Bathia et al. (2010). In Horta and Ziegelmann (2015a), Chapter 3 of this Thesis, we address the question of existence of conjugate processes (that is, stochastic processes driven by a second, measure-valued stochastic process), given some primitive conditions. The positive answer to this question, Theorem 3.1, is the main contribution of the paper, and relies on the concept of random measure (Kallenberg (1973; 1974)) together with the machinery of disintegration of measures (Pachl (1978), Faden (1985), Chang and Pollard (1997), Pollard (2002)). A rich set of examples is provided. Subsequent work should focus in applications, both theoretical (through construction of probability models) and statistical (inference-oriented models aimed at implementations to real data sets that are potentially well described by the concept of conjugate process). In Horta and Ziegelmann (2015b), Chapter 4 of the present Thesis, we show that processes satisfying equation (4.1) – which we call *weakly conjugate processes* – fall smoothly into the methodology of Bathia et al. (2010). The main contributions of the paper are Proposition 4.2 and Theorems 4.1 and 4.2, which provide $\sqrt{n}$-consistency results for the natural estimators appearing in the theory. Additionally, we illustrate the methodology through an implementation to financial data. Specifically, our method permits us to translate the dynamic character of the distribution of an asset returns process into the dynamics of a latent scalar process, from which forecasts of quantities associated to distributional aspects of the returns process can be obtained. Further work should be directed towards empirical research. In particular, an important question is whether other financial assets display a similar representation as the one considered in our implementation, that is, can the evolution of distributional character of asset returns be described by a latent univariate time series? What about linearity? It is also relevant to provide interpretations to the shape of the associated eigenfunctions since this is related to the type of information one can extract from forecasts. Horta and Ziegelmann (2015c), Chapter 5 of this Thesis, is the byproduct of our initial effort of obtaining a generalization of the methodology of Bathia et al. (2010). Relying on the property that centered random elements of strong second order in a separable Hilbert space lie almost surely in the closed linear span of the associated covariance operator (Theorem 5.1), we provide a reformulation of the theory of Bathia et al. (2010) in a Hilbert space setting, culminating in Theorem 5.2 and Corollary 5.2, which state $\sqrt{n}$-consistency of the proposed estimators and are the main contributions of the paper. An important question that remains open is whether the conditions of Theorem 5.2 are also necessary for $\sqrt{n}$-consistency. Clearly (A1) must hold, but likely the ψ-mixing requirement can be weakened. An answer to this question can potentially settle the matter of estimation of the dynamic space associated to Hilbertian time series in the 'noisy measurements' framework.

REFERENCES

Azzalini, A. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328, 1981.

Bathia, N., Yao, Q., and Zieglemann, F. A. Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, 38(6):3352–3386, 2010.

Benko, M., Härdle, W., and Kneip, A. Common functional principal components. *The Annals of Statistics*, 37(1):1–34, 2009.

Billingsley, P. *Convergence of probability measures*, volume 493. John Wiley & Sons, 2009.

Bosq, D. *Linear processes in function spaces: theory and applications*. Springer Verlag, 2000.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Bradley, R. C., et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2(2):107–144, 2005.

Buchinsky, M. Estimating the asymptotic covariance matrix for quantile regression models a monte carlo study. *Journal of Econometrics*, 68(2):303–338, 1995.

Buchinsky, M. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources*, 33(1):88–126, 1998.

Chang, J. T., and Pollard, D. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

Chaudhuri, P., et al. Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, 19(2):760–777, 1991.

Chernozhukov, V., and Hong, H. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

Cheung, K., and Lee, S. M. Bootstrap variance estimation for Nadaraya quantile estimator. *Test*, 19(1):131–145, 2010.

Dabo-Niang, S., and Ferraty, F., editors. *Functional and operatorial statistics*. Springer Verlag, 2008.

Damon, J., and Guillas, S. Estimation and simulation of autoregressive Hilbertian processes with exogenous variables. *Statistical Inference for Stochastic Processes*, 8(2): 185–204, 2005.

Driscoll, M. F. The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Probability Theory and Related Fields*, 26(4):309–316, 1973.

Faden, A. M. The existence of regular conditional probabilities: necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298, 1985.

Falk, M. Relative deficiency of kernel type estimators of quantiles. *The Annals of Statistics*, 12(1):261–268, 1984.

Fan, Y., and Liu, R. A direct approach to inference in nonparametric and semiparametric quantile regression models. *Preprint*, 2013.

Ferraty, F., and Vieu, P. *Nonparametric functional data analysis: theory and practice*. Springer Verlag, 2006.

Ferreira, J., and Menegatto, V. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

Ferreira, J., and Menegatto, V. Reproducing kernel Hilbert spaces associated with kernels on topological spaces. *Functional Analysis and its Applications*, 46(2):152–154, 2012.

Goh, S. C., and Knight, K. Nonstandard quantile-regression inference. *Econometric Theory*, 25(5):1415–1432, 2009.

Guerre, E., and Sabbah, C. Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory*, 28(1): 87–129, 2012.

Hall, P., and Vial, C. Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):689–705, 2006.

He, X., and Shao, Q.-M. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608–2630, 1996.

Hinrichs, A., and Pietsch, A. p-nuclear operators in the sense of Grothendieck. *Mathematische Nachrichten*, 283(2):232–261, 2010.

Horowitz, J. L. Bootstrap methods for median regression models. *Econometrica*, 66 (6):1327–1351, 1998.

Horta, E., and Ziegelmann, F. Conjugate processes. 2015a.

Horta, E., and Ziegelmann, F. Weakly conjugate processes – theory and application to risk forecasting. 2015b.

Horta, E., and Ziegelmann, F. Identifying the spectral representation of hilbertian time series. 2015c.

Horta, E., Guerre, E., and Fernandes, M. Smoothing quantile regression. 2015.

Jurečková, J., Sen, P. K., and Picek, J. *Methodology in Robust and Nonparametric Statistics*. CRC Press, 2012.

Kallenberg, O. Characterization and convergence of random measures and point processes. *Probability Theory and Related Fields*, 27(1):9–21, 1973.

Kallenberg, O. *Lectures on random measures*. Consolidated University of North Carolina, Institute of Statistics, 1974.

Kallenberg, O. An elementary approach to the Daniell-Kolmogorov Theorem and some related results. *Mathematische Nachrichten*, 139(1):251–265, 1988.

Kallenberg, O. *Foundations of modern probability*. Springer Science & Business Media, 2006.

Kaplan, D. M., and Sun, Y. Smoothed estimating equations for instrumental variables quantile regression. 2012.

Knight, K. Comparing conditional quantile estimators: first and second order considerations. Technical report, 2001.

KOENKER, R. Confidence intervals for regression quantiles. In *Asymptotic statistics*, pages 349–359. Springer, 1994.

KOENKER, R. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *Journal of Econometrics*, 95(2):347–374, 2000.

KOENKER, R. *Quantile regression*. Number 38. Cambridge university press, 2005.

KOENKER, R., AND BASSETT, G. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

KOENKER, R., AND HALLOCK, K. Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

KOENKER, R., AND PORTNOY, S. L-estimation for linear models. *Journal of the American Statistical Association*, 82(399):851–857, 1987.

KONG, E., LINTON, O., AND XIA, Y. Global Bahadur representation for nonparametric censored regression quantiles and its applications. *Econometric Theory*, 29(5):941–968, 2013.

KOZEK, A. S. How to combine M-estimators to estimate quantiles and a score function. *Sankhyā: The Indian Journal of Statistics*, 67(2):277–294, 2005.

LEDOUX, M., AND TALAGRAND, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

MACHADO, J. A., AND PARENTE, P. Bootstrap estimation of covariance matrices via the percentile method. *The Econometrics Journal*, 8(1):70–78, 2005.

MACK, Y. Bahadur's representation of sample quantiles based on smoothed estimates of a distribution function. *Probabability and Mathemathical Statistics*, 8:183–189, 1987.

MAMMEN, E., VAN KEILEGOM, I., AND YU, K. Expansion for moments of regression quantiles with application to nonparametric testing. *arXiv preprint arXiv:1306.6179*, 2013.

MAS, A., AND MENNETEAU, L. *Perturbation approach applied to the asymptotic study of random operators*. Springer, 2003.

MASSART, P. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

MEHRA, K., RAO, M. S., AND UPADRASTA, S. A smooth conditional quantile estimator and related applications of conditional empirical processes. *Journal of Multivariate Analysis*, 37(2):151–179, 1991.

MERCER, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441–458):415–446, 1909.

NADARAYA, E. A. Some new estimates for distribution functions. *Theory of Probability & its Applications*, 9(3):497–500, 1964.

NEWEY, W. K., AND MCFADDEN, D. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4(4):2111–2245, 1994.

OTSU, T. Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics*, 142(1):508–538, 2008.

PACHL, J. K. Desintegration and compact measures. *Mathematica Scandinavica*, 43: 157–168, 1978.

PARZEN, E. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365):105–121, 1979.

POLLARD, D. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002.

PORTNOY, S. Nearly root-n approximation for regression quantile processes. *The Annals of Statistics*, 40(3):1714–1736, 2012.

RALESCU, S. S. A Bahadur–Kiefer law for the Nadaraya empiric-quantile processes. *Theory of Probability & its Applications*, 41(2):296–306, 1997.

RAMSAY, J., AND SILVERMAN, B. *Functional data analysis*. Springer Verlag, 1998.

SAMANTA, M. Non-parametric estimation of conditional quantiles. *Statistics & Probability Letters*, 7(5):407–412, 1989.

SHEATHER, S. J., AND MARRON, J. S. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416, 1990.

SILVERMAN, B. W. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

STUTE, W. Conditional empirical processes. *The Annals of Statistics*, 14(2):638–647, 1986.

VAKHANIA, N., TARIELADZE, V., AND CHOBANYAN, S. *Probability distributions on Banach spaces*, volume 14. Springer Science & Business Media, 1987.

VAN DER VAART, A. W., AND WELLNER, J. A. *Weak Convergence and Empirical Processes*. Springer, 1996.

VAN NEERVEN, J. Stochastic evolution equations. *OpenCourseWare, TU Delft*, 2008.

WHANG, Y.-J. Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory*, 22(2):173–205, 2006.

XIANG, X. Bahadur representation of kernel quantile estimators. *Scandinavian Journal of Statistics*, 21(2):169–178, 1994.