Multivariate wavelet-based shape-preserving estimation for dependent observations

Cosma, Antonio; Scaillet, Olivier; von Sachs, Rainer

How to cite

COSMA, Antonio, SCAILLET, Olivier, VON SACHS, Rainer. Multivariate wavelet-based shape-preserving estimation for dependent observations. In: Bernoulli, 2007, vol. 13, n° 2, p. 301–329. doi: 10.3150/07-BEJ5066

This publication URL: https://archive-ouverte.unige.ch/unige:79880
Publication DOI: 10.3150/07-BEJ5066

Multivariate wavelet-based shape preserving estimation for dependent observations

Antonio Cosma[1] Olivier Scaillet[2] Rainer von Sachs[3*]

July 13, 2006

Abstract

We introduce a new approach on shape preserving estimation of cumulative distribution functions and probability density functions using the wavelet methodology for multivariate dependent data. Our estimators preserve shape constraints such as monotonicity, positivity and integration to one, and allow for low spatial regularity of the underlying functions. We discuss conditional quantile estimation for financial time series data as an application. Our methodology can be implemented with B-splines. We show with Monte Carlo simulations that it performs well in finite samples and for a data-driven choice of the resolution level.

Keywords: Conditional quantile, time series, shape preserving wavelet estimation, B-splines, multivariate process.

Abbreviated title: Shape preserving wavelet estimation

AMS 2000 classification: 62G05, 62G07, 42C40, 41A15. *JEL classification:* C14, C15, C32.

[1]Luxembourg School of Finance, Université du Luxembourg, Luxembourg. antonio.cosma@uni.lu

[2]HEC Genève and Swiss Finance Institute, Genève, Suisse. olivier.scaillet@hec.unige.ch

[3]Institut de statistique, Université catholique de Louvain, Louvain-la-Neuve Belgique. vonsachs@stat.ucl.ac.be

*Corresponding author

1 Introduction

The construction of shape-preserving estimators of probabilistic functions, such as probability density functions (pdfs) and cumulative distribution functions (cdfs), has attracted recent interest, see for instance Cheng et al. (1999). When estimating a pdf $f(x)$, we expect an estimator which fulfills nonnegativity and integration to one. When estimating a cdf $F(x)$, we expect an estimator which fulfills monotonicity and right-continuity. Shape preserving means building functional estimators $\hat{f}(x)$ or $\hat{F}(x)$ which display such properties. In contrast to most nonparametric approaches, we aim at dealing with probabilistic functions with low spatial regularity, i.e. with occasional jumps or other discontinuities. In this set-up wavelet methods are known to be of relevance (Vidakovic (1999), Ogden (1996)), but meeting shape constraints is not clear in wavelet estimation.

In this paper we study shape-preserving wavelet estimation of pdfs and cdfs. We model the observed data as serially dependent. We do not need post-processors to implement the shape constraints. Our construction is shape-preserving but not shape-imposing. We can also deal with non-monotone or non-positive functions. We start from the construction of Dechevsky and Penev (1997, 1998), hereafter DP. Their analysis of the estimation of a *univariate* probabilistic function with low spatial regularity with non-orthogonal wavelets is a pure theoretical one and for an identically and independently distributed (*i.i.d.*) model of the data. DP do not suggest an algorithm to bring their method to the data. Our goal is to estimate a *multivariate* function under less stringent conditions than the usual ones in nonparametric approaches. Each component of the multivariate function can be either of a pdf type or a cdf type. Note that a direct transfer of the DP-construction to a multivariate framework is not possible. One of the main motivations for this extension is conditional quantile estimation, for dependent data in financial time series. There we need a general methodology which can estimate a d–dimensional function which is a cdf in one component, and a multivariate pdf in the remaining $d-1$ components.

To summarize, the three main contributions of this paper are,

- the definition of appropriate norms of convergence for an estimator of a multivariate function which is a cdf in one of its components;
- the generalization of the univariate results of DP for an *i.i.d.* model to the case of multivariate time series data subject to mixing conditions;
- the design of fast algorithms to implement our method.

Doukhan (1988) and Doukhan and Léon (1990) provide early work on wavelet density estimation with univariate independent data. Masry provides a generalization to dependent data using orthonormal bases in the univariate setting (Masry (1994)) and in the multivariate setting (Masry (1997)), but the latter analysis is limited to the case of uniform convergence on compact sets. Kerkyacharian and Picard (1992) are the first authors to derive optimality results for linear wavelet density estimation in function spaces, such as Besov spaces. Tribouley (1995) studies linear wavelet methods for multivariate density estimation. Nonlinear wavelet methodologies are applied to univariate density estimation in Donoho et al. (1996) and Kerkyacharian et al. (1996), and in Tribouley and Viennet (1998) for $\beta-$mixing data. In nonlinear methods, orthogonality or bi-orthogonality of the underlying wavelet bases has to be imposed. Other recent studies on density estimation with wavelets aim at dealing with shape preserving properties, see for instance Penev and Dechevsky (1997) and Pinheiro and Vidakovic (1997). These approaches use devices such as pre- or post-processing.

Choosing the DP shape preserving wavelets does not only overcome the need of pre- or post-processors but also yields the following advantages: 1) Since orthogonality has to be given up, we rely on a simple construction using B-splines. It allows us to have analytic expressions of our

basis functions in the time domain. This is essential for deriving the reconstruction of a cdf by integration; 2) The proof technique applies simultaneously to the pdf case and the cdf case; 3) Our approach gives general results for linear wavelet density estimation without a restriction to the Besov space framework of Kerkyacharian and Picard (1992).

Shape-preserving estimation of probabilistic functions turns out to be interesting for a variety of nonparametric estimation problems. To give only a few examples beyond multivariate density estimation, we state hazard rate estimation (Hall and Van Keilegom (2004)), and logistic regression, see for instance McFadden and Train (2000) for an application to Mixed Multinomial Logit Models (MMNL). D. McFadden in his Nobel Prize lecture (McFadden (2003)) states explicitly that an appropriate multivariate extension of the DP set-up is required in an MMNL framework. This ensures that the multivariate indirect utility functions determining the choice probabilities display the required shape restrictions. Shape preserving estimators have also been designed for functional estimation in the context of diffusion processes (Chen et al. (1998)). Another application is quantile regression. It provides robust estimators such as median estimators, and allows to characterize the heterogeneous impact of variables on different points of a distribution. It finds important applications in finance and insurance (quantiles of loss distributions), and in labor economics (measures of income inequality).

We briefly discuss quantile regression when we develop our main application, namely nonparametric estimation of conditional quantiles for time series, the conditioning information being the past observations of the time series. As pointed out by Hall et al. (1999) and Cai (2002), the shape preserving property of a cdf estimator is particularly important for quantile estimation. In our approach we strongly benefit from a direct wavelet estimate which is monotone and constrained to lie between 0 and 1. This is in contrast to other popular methods for quantile regression - see for instance the modified local linear quantile estimators of Yu and Jones (1998). Our time series framework calls for a particular care in proving consistency of our estimator. We provide such a result under less stringent assumptions on the smoothness of the conditional and marginal distribution functions of the random process.

Our paper is organized in the following way.

Section 2 gives an introduction to shape preserving wavelets and estimation of univariate probabilistic functions developed by DP. We present the relevant terminology and concepts such as moduli of smoothness, seminorms, as well as appropriate risk definitions. In Section 3 we briefly recall the main results of DP which are essential for our work. For details summarizing their work we refer to Appendix A in Cosma et al. (2005). Section 4 presents our theoretical contributions: Theorem 1 states the most general result of this article. It is an extension of the univariate results of Section 3 to higher dimensions and the case of time series data. In Section 5 we examine quantile regression and, in particular, conditional quantile estimation for financial time series data. We discuss numerical implementation via B-splines, and present a simulation study. A data-driven choice of the resolution level is investigated numerically. In a short conclusion we discuss some ideas for future research. The outline of the proofs is deferred to an appendix. The detailed proofs can be found in the technical report Cosma et al. (2005).

2 Preliminaries on shape-perserving wavelets

We start this section by introducing the concept of Multiresolution Analysis (MRA). Let $L_2(\mathbb{R})$ be the space of square integrable functions defined on the real line, that is $L_2(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} \,\middle|\, \int_{-\infty}^{+\infty} dx f(x)^2 < \infty \right\}$. An MRA is a sequence of closed subspaces $V_j \subset L_2(\mathbb{R}), j \in \mathbb{Z}$, with the following properties (Meyer (1992)): $V_j \subset V_{j+1}$, $\cap V_j = \{\mathbf{0}\}$, $\overline{\cup V_j} = L_2(\mathbb{R})$, and for all $v(x) \in L_2(\mathbb{R})$ and $j, k \in \mathbb{Z}$, $v(x) \in V_j \Leftrightarrow v(2x) \in V_{j+1}$ and $v(x) \in V_0 \Leftrightarrow v(x-k) \in V_0$. Moreover, a *scaling*

function $\varphi \in V_0$ exists such that $\{\varphi(x-l)|l \in \mathbb{Z}\}$ is a Riesz basis of V_0. It follows that in general $\varphi(2^j x) \in V_j$, and $\left\{\varphi_{jk}(x)\right\}_{k\in\mathbb{Z}} \doteq \{2^{j/2}\varphi(2^j x - k) \mid k \in \mathbb{Z}\}$ is a Riesz basis in V_j.

Orthogonal projection was the first idea exploited in wavelet analysis. It leads to a construction of orthonormal bases of scaling functions, such that $\int_{-\infty}^{+\infty} \varphi(x-l)\varphi(x-k)dx = \delta_{kl}$. However, orthogonality poses a number of constraints on the construction of the scaling functions. For instance it is not possible to construct $\{\varphi(\cdot - l)\}_{l\in\mathbb{Z}}$ that are at the same time continuous, orthogonal, nonnegative and symmetric. Moreover we want to build shape preserving operators, that is projection operators that map nonnegative functions to nonnegative functions and monotone functions to monotone functions. To achieve this, we need extra freedom in building the proper scaling functions. Hence we introduce a non-orthogonal projection operator starting from two different families of scaling functions $\{\varphi(2^j x - k)\}_{k\in\mathbb{Z}}$ and $\{\tilde{\varphi}(2^j x - k)\}_{k\in\mathbb{Z}}$. The projector on the space V_j is given by:

$$A_j\big(f\big)(x) = \sum_{k\in\mathbb{Z}} 2^j \langle f, \tilde{\varphi}(2^j \cdot - k)\rangle \; \varphi(2^j x - k). \tag{2.1}$$

The two families of functions are called the *primal* basis and the *dual* basis. The primal basis is generated from the scaling function φ, and the dual basis is generated from the scaling function $\tilde{\varphi}$. We require that the two families display the following properties. Let φ be such that

$$\varphi(x) \geqslant 0, \quad x \in \mathbb{R}; \tag{2.2}$$

$$\varphi(x) \quad \text{bounded, right continuous;} \tag{2.3}$$

$$\text{supp } \varphi \subset [-a, a), \quad a \geqslant 1/2; \tag{2.4}$$

$$\sum_{k=-\infty}^{\infty} \varphi(x-k) \equiv 1, \text{ on } \mathbb{R}; \tag{2.5}$$

$$\begin{aligned} &\text{there exists } b \in (-a, a) \text{ such that} \\ &\varphi \text{ is not decreasing for } \; x \leqslant b \; \text{ and non-increasing for } \; x \geqslant b. \end{aligned} \tag{2.6}$$

Then let the *dual scaling function* be such that:

$$\tilde{\varphi} \text{ satisfies (2.2), (2.4)}, \quad \tilde{\varphi} \in L_1, \quad \text{and } \int_{-\infty}^{+\infty} dt\, \tilde{\varphi}(t) = 1. \tag{2.7}$$

As for the primal basis, we define the scaled versions of $\tilde{\varphi}$ such that $\left\{\tilde{\varphi}_{jk}(x)\right\}_{k\in\mathbb{Z}} \doteq \{2^{j/2}\tilde{\varphi}(2^j x - k) \mid k \in \mathbb{Z}\}$.

The conditions given on $\tilde{\varphi}$ are weaker than those on φ. Condition $\sum_{-\infty}^{\infty} \varphi(x-k) = 1$ implies $\int_{-\infty}^{+\infty} dt\, \varphi(x) = 1$ (see Anastassiou and Yu (1992) for a proof). In particular, this means that both $\varphi(x)$ and $\tilde{\varphi}(x)$ are normalized in the L_1 norm.

The following notation is also used:

$$\varsigma_{\varphi,\tilde{\varphi}}(t) \doteq \int_{-\infty}^{+\infty} d\tau\, \tau\tilde{\varphi}(\tau) - \sum_{k=-\infty}^{+\infty} (t-k)\varphi(t-k). \tag{2.8}$$

Let us examine why these assumptions guarantee a shape-preserving approximation (2.1) of a pdf or a cdf. Assumption (2.2) on φ and $\tilde{\varphi}$ ensures that the reconstruction (2.1) of a pdf and a cdf is nonnegative. If assumption (2.5) is also satisfied by $\tilde{\varphi}$, then the approximation of a pdf integrates to 1. Assumptions (2.3) and (2.6) are specific to shape preserving approximation of a cdf. Assumption (2.3) guarantees that the reconstruction has the minimum regularity conditions of a cdf (boundedness and right continuity), while assumption (2.6), jointly with the nonnegativity of $\tilde{\varphi}$, guarantees the monotonicity of the reconstruction.

Assumption (2.4) is a usual compact support assumption. It boils down to the use of finite length filters in the implementation of the discrete wavelet transform and implies that $\nu_a = \#\{\varphi_{jk} \mid x \in \text{Supp}(\varphi_{jk})\}$ is independent of scale j and location k. Assumptions (2.5) and (2.8) together with the condition $\varsigma_{\varphi,\tilde{\varphi}} = 0$ *almost everywhere* are equivalent to the usual moment conditions in wavelet approximation theory. Define the moments of the dual scaling function: $\tilde{\mathscr{M}}_0 = \int_{-\infty}^{+\infty} dt\, \tilde{\varphi}(t)$, $\tilde{\mathscr{M}}_1 = \int_{-\infty}^{+\infty} dt\, t\, \tilde{\varphi}(t)$. Then (2.5) and the condition $\varsigma_{\varphi,\tilde{\varphi}} = 0$ can be rewritten in the following way:

$$\sum_{k=-\infty}^{+\infty} (t-k)^p\ \varphi(t-k) = \tilde{\mathscr{M}}_p, \quad p = 0, 1. \tag{2.9}$$

These two conditions ensure that the multiresolution analysis V_j reproduces exactly polynomials of degree less or equal to 1. We say that the multiresolution analysis fulfills a *Strang-Fix* condition of order 1.

We finish this second section by recalling the useful concept of *modulus of smoothness*. This concept is used later on to derive the approximation properties of the projection operator (2.1) in general function spaces, such as Sobolev and Besov spaces (see Nikol'skiĭ (1975)). For functions defined on a region $\Omega \in \mathbb{R}^d$, we introduce the increment of the function f in the direction $\mathbf{i}$ and the corresponding modulus of smoothness. Let $\Delta^1_{\mathbf{i}t} f(\mathbf{x}) = f(\mathbf{x} + \mathbf{i}t) - f(\mathbf{x})$, $\Delta^\mu_{\mathbf{i}t} f(\mathbf{x}) = \Delta^1(\Delta^{\mu-1}_{\mathbf{i}t} f(\mathbf{x}))$, then, for $h > 0$, $\mu \in \mathbb{N}$ and $1 \leqslant p \leqslant \infty$, the integral p-modulus of smoothness in the $\mathbf{i}$ direction is given by

$$\omega^\mu_{\mathbf{i}}(f, h)_p = \sup_{0<t\leqslant h} \left\| \Delta^\mu_{\mathbf{i}t} f(\mathbf{x}) \right\|_p, \tag{2.10}$$

with the usual convention of the $\sup_x$-norm for $p = \infty$, which is the classical modulus of continuity.

3 Shape preserving estimation of univariate probabilistic functions

First we briefly summarize the main concepts of DP in the *i.i.d.* case for the estimation of *univariate* probabilistic functions (pdfs and cdfs) by means of shape preserving wavelets. We recall these results since they have inspired our own work for multivariate time series data. Note that we need to define an estimation risk in a function norm which is appropriate for treating simultaneously the error when estimating a pdf or a cdf nonparametrically. See equation (3.5). In the two cases the inner products in equation (2.1) can be estimated from the observed data $(X_1, \ldots, X_n)$ in the following way:

$$\widehat{\langle f, \tilde{\varphi}_{jk} \rangle} = \langle f, \tilde{\varphi}_{jk}(X) \rangle = \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_{jk}(X_i), \qquad \text{if } f \text{ is a pdf,} \tag{3.1}$$

$$\widehat{\langle F, \tilde{\varphi}_{jk} \rangle} = \langle F, \tilde{\varphi}_{jk}(X) \rangle = \frac{1}{n} \sum_{i=1}^{n} 2^{-\frac{j}{2}} \left\{ 1 - \tilde{\Phi}(2^j X_i - k) \right\}, \qquad \text{if } F \text{ is a cdf,} \tag{3.2}$$

with $\tilde{\Phi}(x) = \int_{-\infty}^{x} \tilde{\varphi}(t)\, dt$. Since f is a pdf, $\langle f, \tilde{\varphi} \rangle = \mathbb{E}[\tilde{\varphi}]$. Then (3.1) is an estimator of the expected value of $\tilde{\varphi}_{jk}$. In a similar way we can obtain (3.2) by integration by parts

$$\langle F, \tilde{\varphi}_{jk} \rangle \ = \ 2^{-j/2} - 2^{-j/2} \mathbb{E}[\Phi(2^j X - k)],$$

using the boundedness of the support and the normalization properties of $\tilde{\varphi}_{jk}$. It then follows that the estimators for a univariate pdf and a univariate cdf are given by:

$$\hat{f}(x) = \hat{A}_j^{(n)}\big(f\big)(x) = \frac{1}{n}\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\tilde{\varphi}_{jk}(X_i)\varphi_{jk}(x), \qquad \text{if } f \text{ is a pdf;} \tag{3.3}$$

$$\hat{F}(x) = \hat{A}_j^{(n)}\big(F\big)(x) = \frac{1}{n}\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n} 2^{-\frac{j}{2}}\big\{1-\tilde{\Phi}(2^j X_i - k)\big\}\varphi_{jk}(x), \qquad \text{if } F \text{ is a cdf.} \tag{3.4}$$

Lemma 1. *Let f be either a pdf or a cdf. Let $\varphi, \tilde{\varphi}$ fulfill assumptions* (2.2) *to* (2.7) *and, if f is a pdf, let $\tilde{\varphi}$ fulfill also* (2.5)*. Then the estimator $\hat{A}_j^{(n)}\big(f\big)$ derived from the operator* (2.1) *using* (3.3) *or* (3.4) *is shape preserving.*

By shape preserving, we mean that if f is a pdf, then $\hat{A}_j^{(n)}\big(f\big)$ is a nonnegative function that integrates to 1, and if F is a cdf, then $\hat{A}_j^{(n)}\big(F\big)$ is a monotone, right-continuous function and $\lim_{x\to\pm\infty}\hat{A}_j^{(n)}\big(F\big)(x) = 0, 1$. For the proofs we refer to Lemma 2.2.1 in DP (1997) for the pdf case, and to Lemma 2.1.1 in DP (1997) and Lemma 3 in Anastassiou and Yu (1992) for the cdf case. The shape preserving properties of estimators (3.3) and (3.4) come from the approximation results derived in DP (1997).

To assess the behavior of the estimators we define a risk using the following quasi-norm for a function $g(x)$ defined on $\mathbb{R}$, whose random values depend on the realization of $(X_1, \ldots, X_n)$:

$$\big\|\, g\,\big|L_p(\mathscr{L}_q)\big\| = \left\{\int_{-\infty}^{+\infty} dx\,\big(\mathbb{E}\,|g(x)|^q\big)^{p/q}\right\}^{1/p},$$

with $0 < p, q \leqslant \infty$. Recall that for a quasi-norm the triangular inequality holds with $\big\|\, g+h\big\|_A \leqslant c_A(\big\|\, g\big\|_A + \big\|\, h\big\|_A)$, $c_A \geqslant 1$. In order to be able to work with the usual triangular inequality, i.e. $c_A = 1$, we move to the space $L_p(\mathscr{L}_q)^\rho$ with an appropriately chosen $\rho > 0$ (see the definition and the discussion in Appendix D).

In the $L_p(\mathscr{L}_q)$ quasi-norm the p parameter takes into account the smoothness of the function via (2.10), while the q parameter gives an additional degree of freedom to ensure that the estimation risk stays finite through a control of the tails. In the original work of DP (1998) the notation of the two parameters is inverted; in our setting we prefer to call p the "smoothness" parameter as is done for Besov spaces. Note that in contrast to usual Besov spaces, q is here connected to the stochastic dimension of the problem. Our change is notational only and does not alter the essence of the quasi-norm used in the original reference. For $p = q$, we get the usual L_p-risk, i.e. $\mathbb{E}\|\cdot\|_p$. The notation $\|\cdot\|_p$ remains associated with the usual $L_p(\mathbb{R}^d)$-norm.

From now on let $\hat{f}$ be an estimator for a pdf or a cdf. Then,

$$\begin{aligned}\big\|\,\hat{f}-f\,\big|L_p(\mathscr{L}_q)\big\|^\rho &= \big\|\,\hat{f}-\mathbb{E}(\hat{f})+\mathbb{E}(\hat{f})-f\,\big|L_p(\mathscr{L}_q)\big\|^\rho \\ &\leqslant c(p,q,\rho)\Big\{\big\|\,\hat{f}-\mathbb{E}(\hat{f})\,\big|L_p(\mathscr{L}_q)\big\|^\rho + \big\|\,\mathbb{E}(\hat{f})-f\,\big|L_p(\mathscr{L}_q)\big\|^\rho\Big\}.\end{aligned} \tag{3.5}$$

Hereafter we restrict ourselves to the ranges $1 \leqslant p \leqslant \infty$, $0 < q \leqslant 2$, and we always work with a choice of ρ such that $c(p, q, \rho) = 1$.

From equations (3.1) and (3.2), we can see that the estimators (3.1) and (3.2) are unbiased estimators of the inner products $\langle f, \tilde{\varphi}\rangle$, and that $\hat{A}_j$ is also an unbiased estimator of A_j. The triangular inequality (3.5) can thus be rewritten as:

$$\big\|\,\hat{f}-f\,\big|L_p(\mathscr{L}_q)\big\|^\rho \leqslant \Big\{\big\|\,A_j\big(f\big)(\cdot)-f(\cdot)\big\|_p^\rho + \big\|\,\hat{A}_j^{(n)}\big(f\big)(\cdot)-A_j\big(f\big)(\cdot)\,\big|L_p(\mathscr{L}_q)\big\|^\rho\Big\}. \tag{3.6}$$

The first result (DP (1997), Theorem 2.1.1) concerns the first part of equation (3.6), i.e., the bias term:

$$\left\| A_j\big(f\big)(\cdot) - f(\cdot)\right\|_p \leqslant c_1 \|\varsigma_{\varphi,\tilde{\varphi}}\|_\infty \cdot \omega^1(f, 2^{1-j}a)_p + c_2 \|\tilde{\varphi}\|_{p'} \cdot \|\varphi\|_\infty \cdot \omega^2(f, 2^{1-j}a)_p, \tag{3.7}$$

where $p' \in [1, \infty]$ is such that $\frac{1}{p} + \frac{1}{p'} = 1$, a is the length of the support of the scaling functions, and $c_1 > 0$ and $c_2 > 0$ are absolute constants that do not depend on the resolution level j.

Note that, in contrast to pdfs, cdfs are not in L_p with $1 \leqslant p < \infty$, but only in L_∞. Yet by (3.7) the L_p distance between $A_j(F)$ and F is bounded. This means that the approximation properties of $A_j(F)$ can be studied in an appropriately chosen L_p-norm.

It is possible to obtain how the bias depends explicitly on the resolution level j. To this end we would need to specify the function space to which f belongs and exploit the properties of the modulus of smoothness in this specific function space. For now, we just give a qualitative characterization of this dependence. Since $\omega^\mu(f, 2^{1-j}a)_p$ is increasing in its second argument by definition, the bias can be bounded by a *decreasing* function of j, so that equation (3.7) can be rewritten as:

$$\left\| A_j\big(f\big)(\cdot) - f(\cdot)\right\|_p \leqslant B(j), \tag{3.8}$$

where $B(j)$ is a decreasing function of j.

For the second term of (3.5), which is rewritten as

$$\left\| \hat{A}_j^{(n)}\big(f\big)(\cdot) - \mathbb{E}\big(\hat{A}_j^{(n)}\big(f\big)(\cdot)\big)\Big| L_p(\mathscr{L}_q)\right\| = \left\{ \int_{-\infty}^{+\infty} dx \left(\mathbb{E}\big|\hat{A}_j^{(n)}\big(f\big)(x) - \mathbb{E}(\hat{A}_j^{(n)}\big(f\big)(x)\big|^q\right)^{\frac{p}{q}} \right\}^{\frac{1}{p}}, \tag{3.9}$$

we have two different behaviors depending on f being a cdf or a pdf.

For a cdf. (DP (1998), Theorem 2.1.1) The variance term (3.9) has a parametric decay $O\left((\frac{1}{n})^{\rho/2}\right)$ to zero. In this case the bias rate can be adapted by appropriately choosing the increasing function $j^* = j^*(n)$ such that $B(j^*) = O(n^{-\rho/2})$. The total risk (3.6) then decays at a parametric rate:

$$\left\| \hat{A}_j^{(n)}\big(F\big)(\cdot) - F(\cdot)\,\big| L_p(\mathscr{L}_q)\right\|^\rho = O\left(\left(\frac{1}{n}\right)^{\rho/2}\right), \qquad n \to \infty\ . \tag{3.10}$$

It can be easily seen that the above convergence rate can be achieved by choosing $j \geqslant (p/2)\log_2 n$.

For a pdf. (DP (1998), Theorem 2.2.1) The variance term is an increasing function of j, that is (3.9) is bounded by a function $V(2^j/n)\big((2^j/n)^\rho\big)$. When choosing the function $j^* = j^*(n)$, we face the typical nonparametric trade-off between the competing behaviors of B and V as functions of j. In particular, $j^* = j^*(n)$ has to be an increasing function such that $V(2^{j^*}/n) = O\big(B(j^*)\big)$. The convergence rate for the estimator of a pdf is of order

$$\left\| \hat{A}_j^{(n)}\big(f\big)(\cdot) - F(\cdot)\,\big| L_p(\mathscr{L}_q)\right\|^\rho = O\ (2^{-j^*(n)\rho}), \qquad n \to \infty\ . \tag{3.11}$$

which is typically slower than the parametric one. These findings are the same as in classical wavelet estimation (Härdle et al. (1998), Kerkyacharian and Picard (1992)). For detailed results, we refer to Cosma et al. (2005), Appendix A, Corollaries 10 - 14.

4 Shape preserving estimation of multivariate probabilistic functions

This is the main section of our paper. Here we provide a multivariate extension of the results of DP for time series data. Our work is motivated by an application to quantile estimation (see Section 5).

We analyze a multivariate function $F \in \mathbb{R}^d$ which is a cdf in the last argument, and a pdf in the $d-1$ remaining arguments, i.e.,

$$F_Y(\mathbf{x}, y) = \int_{-\infty}^{y} dt\, f(\mathbf{x}, t). \tag{4.1}$$

However, our set-up allows for constructing estimators of multivariate densities $f(\mathbf{x}) \in \mathbb{R}^d$ as well. For ease of notation we present only the bivariate case. From now on the argument y will always denote the variable with respect to which $F_Y(x, y)$ is cumulated, and x the argument with respect to which $F_Y(x, y)$ is a density. The extension of the results to a bivariate pdf can be obtained with minor changes that will be given in the sequel.

Our constructions are based on *tensor product* wavelets. The primal and dual wavelet bases $\varphi, \tilde{\varphi}$ introduced in equations (2.2) - (2.7) are functions defined from $\mathbb{R}$ to $\mathbb{R}$ so that functions $f : \mathbb{R} \to \mathbb{R}$ can be approximated. It is straightforward to build scaling functions defined on $\mathbb{R}^d$, so that multivariate functions can be approximated by wavelet series, using tensor product wavelets. It is known (see, e.g., Meyer (1992) Section 3.3) that, if $\{V_j\}_{j\in\mathbb{Z}}$ is a multiresolution analysis of $L_2(\mathbb{R})$, then $L_2(\mathbb{R}^2) = \overline{\bigcup_{j_1,j_2=0}^{\infty} V_{j_1} \otimes V_{j_2}}$.

Now we define $\mathbf{j} = (j_1, j_2)$ and $V_{\mathbf{j}} = V_{j_1} \otimes V_{j_2}$, from which we can derive the 2-dimensional basis for each approximation space $V_{\mathbf{j}}$:

$$\{\varphi_{\mathbf{jk}}(x, y)\}_{\mathbf{k}\in\mathbb{R}^2} = \{\varphi_{j_1 k_1}(x)\}_{k_1\in\mathbb{Z}} \otimes \{\varphi_{j_2 k_2}(y)\}_{k_2\in\mathbb{Z}}. \tag{4.2}$$

This construction can obviously be extended to any dimension $d > 2$. The use of independent scales $j_i,\ i = 1, \ldots, d$ for each dimension allows for adaptation to possibly different regularity (over the different dimensions) of the $d-$dimensional function (see, e.g., Neumann and von Sachs (1997)). Hence the cdf part can benefit from a higher j_2 than the j_1 of the pdf part.

In the sequel we provide approximation and estimation results of bivariate probability functions treating both the *i.i.d.* case and the serially dependent case. To emphasize the latter, we suppose that we have real-valued bivariate time series observations $(Y_1, X_1), \ldots, (Y_T, X_T)$, generated from a stationary stochastic process $\{(Y_t, X_t)\}_{t\in\mathbb{Z}}$. Its dependence structure is controlled via mixing conditions. In particular we could have that $Y_t = Z_t$ and $X_t = Z_{t-1}$, where $\{Z_t\}_{t\in\mathbb{Z}}$ is a univariate stationary process. Let $\mathscr{F}_i^k$ be the sigma-field of events generated by the random variables $\{(Y_t, X_t), i \leqslant t \leqslant k\}$. The stationary process $\{(Y_t, X_t)\}_{t\in\mathbb{Z}}$ is called strongly or α- mixing if $\sup_{\substack{A\in\mathscr{F}_{-\infty}^0 \\ B\in\mathscr{F}_p^\infty}} \left|P[AB] - P[A]P[B]\right| = \alpha(p) \xrightarrow[p\to\infty]{} 0$. Below, if not differently stated, let $f(\cdot)$ be the "design" density $f_{X_t}(x)$, which is the marginal distribution of the stationary process in the univariate time series case.

Our results are then derived under the following assumptions on the stochastic process.
Assumption 1: For every integer $s > 0$ the joint distribution $F_{(X_0,Y_0),(X_s,Y_s)}$ exists and there is a positive constant M such that for every bounded zero-mean random variable $T(X_t, Y_t)$:

$$\mathbb{E}[|T(X_0, Y_0)\ T(X_s, Y_s)|] \leqslant M\ \mathbb{E}[|T(X_0, Y_0)|]\ \mathbb{E}[|T(X_s, Y_s)|]. \tag{4.3}$$

Assumption 2: The process $\{(X_t, Y_t)\}$ is α-mixing and the coefficients $\alpha(p)$ are such that:

$$\sum_{p=N}^{\infty} \left[\alpha(p)\right]^{1-2/r} = O(N^{-1}), \tag{4.4}$$

for $r > 2$.

Many processes verify the condition given on the mixing coefficients. Gaussian processes, non Gaussian autoregressive moving average processes (see Pham and Tran (1980)), many nonlinear

functionals of these processes, and various GARCH and stochastic volatility models, see Carrasco and Chen (2002).

We construct our estimators by mimicking the univariate constructions (3.3) and (3.4). We use the shape preserving scaling functions φ and $\tilde{\varphi}$ that fulfill assumptions (2.2) to (2.7). Recall that $\mathbf{j} = (j_1, j_2)$. Let $\{\varphi_{\mathbf{jk}}(x,y)\}_{\mathbf{k}\in\mathbb{R}^2}$, $\{\tilde{\varphi}_{\mathbf{jk}}(x,y)\}_{\mathbf{k}\in\mathbb{R}^2}$ be the bivariate primal and dual bases (see (4.2)), then the estimators of $F_Y(x,y)$ and $f(x,y)$ are given by:

$$\hat{A}^{(T)}_{\mathbf{j}}\big(f\big)(x,y) = \sum_{\mathbf{k}\in\mathbb{Z}^2}\Big\{\frac{1}{T}\sum_{t=1}^{T}\tilde{\varphi}_{j_1k_1}(X_t)\tilde{\varphi}_{j_2k_2}(Y_t)\Big\}\varphi_{\mathbf{jk}}(x,y),\ (x,y)\in\mathbb{R}^2, \tag{4.5}$$

$$\hat{A}^{(T)}_{\mathbf{j}}\big(F\big)(x,y) = \sum_{\mathbf{k}\in\mathbb{Z}^2}\sum_{t=1}^{T}2^{-\frac{j_2}{2}}\left(\frac{\tilde{\varphi}_{j_1k_1}(X_t)}{T} - \frac{\tilde{\varphi}_{j_1k_1}(X_t)\tilde{\Phi}(2^{j_2}Y_t - k_2)}{T}\right)\varphi_{\mathbf{jk}}(x,y),\quad (x,y)\in\mathbb{R}^2. \tag{4.6}$$

Let us further introduce the multivariate approximator of the probabilistic function f:

$$A_{\mathbf{j}}\big(f\big)(x,y) = \sum_{k\in\mathbb{Z}^d}\langle f, \tilde{\varphi}_{\mathbf{jk}}\rangle\varphi_{\mathbf{jk}}(x,y). \tag{4.7}$$

Note that the properties of the functions and the projectors $A_{\mathbf{j}}\big(f\big)(x,y)$ that allow us to approximate the cdf of one variable, cannot play the same role when we move to a multivariate analysis. In particular (3.7) does not continue to hold in the multivariate L_p-norm because it is not possible to bound the modulus of smoothness of a function which is in L_∞ in the y-argument. We could use an L_∞-norm but this would require continuity of the density part of $F_Y(x,y)$. Instead we take advantage of the different convergence rates for a pdf and a cdf as discussed in Section 3, and we work with the following risk:

$$d\big\{\hat{f}(x,y), \hat{g}(x,y)\big\}_p = \sup_{y\in\mathbb{R}}\big\|\,\hat{f}(\cdot,y) - \hat{g}(\cdot,y)\,\big|L_p(\mathscr{L}_q)\big\|. \tag{4.8}$$

In order to use this norm, we have to assume continuity of $F_Y(x,y)$ as a function of y. We will see that the results on the approximation and estimation of the bivariate $F_Y(x,y)$ look like the usual results on estimation of a univariate pdf. No additional effort is needed to interpret them. Moreover, these results can be adapted with minor changes to a bivariate pdf, where the norm (4.8) becomes the usual $L_p(\mathbb{R}^2)$-norm. Then we do not need to make a continuity assumption on the bivariate pdf. Further details on the changes needed to adapt the pdf-cdf results to the pure pdf case will be given at the end of each of the following sub-sections.

4.1 Bias in multivariate cdf-pdf approximation

Here we bound the deterministic bias made by approximating $F_Y(x,y)$. A sketch of the proof can be found in Appendix A. The proof technique is largely inspired by the one in DP (1997), but with changes requested by the use of the norm (4.8). The proof is based on approximations by Steklov means. It allows us to relate the approximation error to the modulus of smoothness. We refer to Appendix C for a definition and relevant properties of Steklov means. Recall the definitions of $\varsigma_{\varphi\tilde{\varphi}}$ in (2.8), and let $\mathbf{e}_x$ and $\mathbf{e}_y$ be the unit vectors in the x and y directions, respectively.

Lemma 2. *Let assumptions* (2.2) *to* (2.7) *hold. Let $F(x,y)$ be, for a fixed x, a continuous function in y. Then, for $1 \leqslant p \leqslant \infty$:*

$$\begin{aligned}\sup_y \|A_{\mathbf{j}}\big(F\big)(\cdot,y) - F(\cdot,y)\|_p \\ \leqslant c_1 \sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p + c_2 \sup_y \sup_{0<\delta<1}\big\|\Delta_{\mathbf{e}_y\delta(2^{1-j_2}a)}F(\cdot,y)\big\|_p \\ + c_3 \sup_y \omega^2_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p + c_4 \sup_y \sup_{0<\delta<1}\big\|\Delta^2_{\mathbf{e}_y\delta(2^{1-j_2}a)}F(\cdot,y)\big\|_p.\end{aligned} \tag{4.9}$$

The constants c_1 and c_2 include the value of the norms $\|\varsigma_{\varphi\tilde{\varphi}}(2^{j_1}x)\|_\infty$ and $\|\varsigma_{\varphi\tilde{\varphi}}(2^{j_2}y)\|_\infty$. If the latter two L_∞-norms are 0, then the constants c_1 and c_2 will be smaller but they cannot be equal to 0.

The first order terms, i.e. the terms in which appear the constants c_1 and c_2, apparently do not vanish in (4.9) even if the conditions $\|\varsigma_{\varphi\tilde{\varphi}}(2^{j_1}x)\|_\infty = 0$ and $\|\varsigma_{\varphi\tilde{\varphi}}(2^{j_2}y)\|_\infty = 0$ are fulfilled. This differs from equation (3.7) of the univariate setting. It is related to the proof technique and not to a fundamental reason.

Remark 1. For a bivariate pdf, we do not need to assume continuity of $f(x, y)$ with respect to y in Lemma 2. In (3.7) the L_p-norm is taken with respect to both arguments, and the bound is given by the usual moduli of smoothness in the two directions. This amounts to the bivariate equivalent of (3.7).

4.2 Estimation of multivariate cdf-pdf from dependent data

We come now to our main results where we study the estimation of the bivariate $F_Y(x, y)$ from time series observations $\{(X_1, Y_1), \ldots, (X_T, Y_T)\}$ generated under assumptions 1 and 2. The *i.i.d.* case is a subcase of this one.

Again we make use of the triangular inequality to split the risk into a stochastic term and a bias term:

$$\begin{aligned}&\sup_{y\in\mathbb{R}}\big\| \hat{A}_{\mathbf{j}}^{(T)}\big(F\big)(\cdot, y) - F(\cdot, y)\,\big|L_p(\mathscr{L}_q)\big\|^\rho \\ &\quad\leqslant\Big\{ \sup_y \big\|A_{\mathbf{j}}\big(F\big)(\cdot, y) - F(\cdot, y)\big\|_p^\rho + \sup_y \big\| \hat{A}_{\mathbf{j}}^{(T)}\big(F\big)(\cdot, y) - A_{\mathbf{j}}\big(F\big)(\cdot, y)\,\big|L_p(\mathscr{L}_q)\big\|^\rho\Big\}.\end{aligned}$$

The bias part is bounded in Lemma 2. We now give a bound for the stochastic part.

Lemma 3. *Let $\varphi, \tilde{\varphi}$ be as in (2.2) - (2.7). Let $\{(X_t, Y_t)\}_{t=0,\ldots,T}$ be realizations of a stationary process fulfilling assumptions 1 and 2. Let $p \geqslant 1$, $0 < q \leqslant 2$ and $\rho = \min\{2, p\}$. Assume, for fixed y, that $F_Y(x, y) \in L_{p/2} \cap L_1 \cap L_{p/r}$ for $r > 2$. Let $j_2 \geqslant (p/2)\log_2 T$. Then*

$$\sup_y \Big\| \hat{A}_{\mathbf{j}}^{(T)}\big(F\big)(\cdot, y) - \mathbb{E}\big[\hat{A}_{\mathbf{j}}^{(T)}\big(F\big)(\cdot, y)\big]\,\Big|L_p(\mathscr{L}_q)\Big\|^\rho \leqslant \bar{V}_0 + \bar{V}_1, \tag{4.10}$$

where

$$\begin{aligned}\bar{V}_0 = \left(\frac{2^{j_1}}{T}\right)^{\rho/2} \Big[& d_1 \max\Big\{\sup_y \big\|F(\cdot, y)\big\|_1, \sup_y \big\|F(\cdot, y)\big\|_{p/2}\Big\}^{\rho/2} \\ + \quad & d_2(a) \max\Big\{\sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot, y), 2^{1-j_1}a\big)_1, \sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot, y), 2^{1-j_1}a\big)_{p/2}\Big\}^{\rho/2}\Big] \\ & + O\Big(T^{-\rho/2}\Big);\end{aligned} \tag{4.11}$$

$$\frac{\bar{V}_1}{\bar{V}_0} \longrightarrow 0, \quad \textit{as} \;\; T \to \infty\ . \tag{4.12}$$

Moreover, we have

$$\max\big\{\sup_y\ \omega^1(F(\cdot, y), 2^{1-j_1}a)_1, \sup_y\ \omega^1(F(\cdot, y), 2^{1-j_1}a)_{p/2}\big\} = o(1), \quad j_1 \to \infty,$$

if $2 \leqslant p < \infty$,
or if $p = \infty$ and F is continuous also with respect to the argument x,
or if $1 < p < 2$ and $\sup_y \sup_{0\leqslant t\leqslant h} \int_{-\infty}^{+\infty} dx \Big(\int_0^1 d\alpha\, F(x + \alpha t, y)\Big)^{p/2} < \infty$.

Here d_1 and $d_2(a)$ are absolute constants that do not depend on the resolution levels $\mathbf{j} = (j_1, j_2)$, and a is the support of $\varphi, \tilde{\varphi}$. As it can be seen from equation (4.10) the stochastic component of the risk has two contributions $\bar{V}_0$ and $\bar{V}_1$. We refer to the proof for an explicit form of the latter. $\bar{V}_0$ is the only variance term if $\hat{F}_Y(x, y)$ is estimated under an *i.i.d.* model.

In Cosma et al. (2005) Appendix F, we provide a slightly more general expression for the variance term in the *i.i.d.* set-up with a parameter range $0 < q < \infty$.

A close inspection of (4.11) indicates that only the resolution parameter j_1 appears. Formally the variance term obtained in Lemma 3 is equivalent to the one we would obtain for a univariate pdf, see Appendix A in Cosma et al. (2005). This is because, as remarked in the discussion following equations (3.10) and (3.11), the cdf like component of the variance has a faster convergence, and with the choice $j_2 \geqslant (p/2)\log_2 T$ the convergence of the entire y component of the stochastic error is taken into account by the $O(T^{-\rho/2})$ term . We can finally put together the results of Lemmas 2 and 3 to obtain the convergence rates of $\hat{A}_{\mathbf{j}}^{(T)}$.

Theorem 1. *Let the assumptions of Lemmas 2 and 3 hold. Then the total risk of the estimator* $\hat{A}_{\mathbf{j}}^{(T)}$ *of* $F_Y(x, y)$ *is:*

$$\begin{aligned}
&\sup_{y\in\mathbb{R}} \big\| \hat{A}_{\mathbf{j}}^{(T)}\big(F\big)(\cdot, y) - F(\cdot, y)\,\big|L_p(\mathscr{L}_q)\big\|^{\rho} \\
&\leqslant c_1 \sup_y \omega_{\mathbf{e}_x}^1\big(F(\cdot, y), 2^{1-j_1}a\big)_p^{\rho} + c_3 \sup_y \omega_{\mathbf{e}_x}^2\big(F(\cdot, y), 2^{1-j_1}a\big)_p^{\rho} + \left(\frac{2^{j_1}}{T}\right)^{\rho/2}\Big\{C_1 + o(1)\Big\} \qquad (4.13)\\
&+ O(T^{-\rho/2}) + o(1).
\end{aligned}$$

The constants c_1, c_3 and C_1, can be made explicit by comparing (4.13) with (4.9) and (4.11). Again we see that only the j_1 parameter appears. The increments $\|\Delta_{\mathbf{e}_y}^i f\|$ of the function $F_Y(x, y)$ in the direction y that can be found in (4.9) are missing in (4.13), since they are taken into account in the $O(T^{-\rho/2})$ term once we choose j_2 to fulfill $j_2 \geqslant (p/2)\log_2 T$. The $o(1)$ term takes into account the $\bar{V}_1$ component of the variance. The risk given in Theorem 1 is formally equivalent to the risk in estimating a univariate pdf. Hence the explicit convergence rates of $\hat{A}_{\mathbf{j}}^{(T)}\big(F\big)$ can be obtained again by choosing a resolution level j_1 in the x direction that balances bias and variance. They are equivalent to the convergence rates of the estimator of a univariate pdf having the same smoothness as the pdf part of F_Y (see Corollaries 10 - 14 in Cosma et al. (2005), Appendix A).

Remark 2. Lemma 3 can be extended to a bivariate pdf. The continuity assumption on the y variable is no more required, and we assume that $f(x, y) \in L_{p/2} \cap L_1 \cap L_{p/r}$. Then in (4.11) a $(2^{j_1+j_2}/T)$ factor appears instead of $(2^{j_1}/T)$, and the $O(T^{-\rho/2})$ term disappears.

Lemma 3 and the proofs deal with the bivariate distribution function $F_Y(x, y)$, but a close look at the proof reveals that the results can be immediately extended to a dimension $d \geqslant 2$. As a multivariate density has a finite L_p-module of smoothness for $p < \infty$, we can extend Theorem 1 to functions $F_Y(\mathbf{x}, y)$, by looking at the uniform convergence in y of

$$\sup_y \left\| \hat{A}_{\mathbf{j}}^{(T)}(f)(\cdot, y) - \mathbb{E}\hat{A}_{\mathbf{j}}^{(T)}(F)(\cdot, y)\right| L_p(\mathscr{L}_q)\Big\| \,,$$

where the above L_p-norm, made explicit in equation (3.9), is now defined on $\mathbb{R}^{d-1}$.

Let us briefly comment on the results obtained for the convergence of the distribution function $F_Y(x, y)$. We have derived the upper bound and the asymptotic behavior of the stochastic term of the risk of the estimator (4.6) when the data come from a weakly dependent process. The risk is computed in the $L_p(\mathscr{L}_q)$-norm. The advantage of separating the pointwise expectation from the global norm is probably more evident here than in other contexts. This can be seen by comparing our results with the ones in Masry (1994). First of all Masry computes the risk in the Sobolev W_2^s

norm. We, however, start from the $L_p(\mathscr{L}_q)$-risk and can specify the risk in a variety of different norms, especially in the norm of the Besov spaces built from L_p. But the main difference concerns the conditions that have to be imposed on the tails of the density for the risk not to explode. While we simply need to impose that $f \in L_{p/r}, r > 2$, in Masry (1994) the decay of the density tails has to be related to the smoothness s of the density, asking for a decay of order $x^{-\beta}$, with $\beta > 0.5 + s$.

5 Statistical applications: quantile regression and estimation of conditional quantiles of financial data

We address the estimation of conditional quantiles as an important application of our methodology. Let us consider the stationary process $\{(X_t, Y_t)\}$ introduced in Section 4. We wish to estimate the conditional distribution function $F_Y(y|x) = P(Y_t \leqslant y | X_t = x)$ and, from this, the p-th conditional quantile, that is the value $Q(x,p) = \inf\{y \in \mathbb{R} | F(y|x) \geqslant p\}$, for a probability level $0 < p < 1$. The conditional median $Q(x, p = 0.5)$ has often been the main object of interest. It is used as an alternative to the conditional mean to deliver a robust estimate of the effect of the variable X on the response variable Y. In general we can build confidence intervals for the variable Y from a collection of conditional quantiles. In the case of a stationary process $\{Z_t\}_{t\in\mathbb{Z}}$, when $Y_t = Z_t$ and $X_t = Z_{t-1}$, it is possible to build prediction intervals for Z_t having observed Z_{t-1}. The estimation of conditional quantiles of financial time series motivates this work to a large extent. Conditional quantiles are used as measures of risk, and are known as conditional Value-at-Risk in the financial literature.

In this section we start by linking the problem of the estimation of conditional quantiles to the one of the estimation of the bivariate $F_Y(x,y)$ studied in the previous sections. In order to achieve this, we show that a modification of the norm (4.8) is needed. We then provide a couple of scaling functions $\varphi, \tilde{\varphi}$ fulfilling assumptions (2.2) to (2.7) and (2.9), and we explicitly build shape preserving estimators. Hence we deliver an algorithm for the implementation of the shape preserving set-up. Finally we carry out a Monte Carlo experiment to check our methodology.

5.1 Set-up and conditional quantiles

We keep the same assumptions on the stochastic process as in Section 4. In particular we consider observations $\{Y_1, \dots, Y_T\}$ of a stationary process $\{Y_t\}_{t\in\mathbb{Z}}$ such that the couples $\{(Y_t, X_t = Y_{t-1})\}_{t\in\mathbb{Z}}$ form a Markovian process of order one fulfilling assumption 1 and the mixing conditions of assumption 2. Moreover we assume that the conditional distributions $F_{Y_t}(y|x)$ and $f_{Y_t}(y|x)$ of Y_t given $X_t = x$ exist. Here we are interested in the p-th conditional quantile, which is assumed to be unique. We define the estimator $\hat{Q}(x,p)$ to be such that

$$\hat{Q}(p,x) = \inf\left\{y \in \mathbb{R} \,\middle|\, \hat{F}(y|x) \geqslant p\right\}. \tag{5.1}$$

The solution of (5.1) always exists since $\hat{F}(y|x)$ is monotone and bounded between 0 and 1.

Now, since we have assumed the existence of the conditional pdf $f(y|x)$, and since $F(y|x)$ is its integral by definition, we can write a Taylor expansion with integral remainder:

$$\begin{aligned} &F(\hat{Q}(p,x)|x) - F(Q(p,x)|x) \\ &\qquad = \big(\hat{Q}(p,x) - Q(p,x)\big) \int_0^1 d\theta f\big(Q(p,x) + \theta(\hat{Q}(p,x) - Q(p,x))\big|\, x\big). \end{aligned} \tag{5.2}$$

Denote $\tilde{f}_{\hat{Q},Q}(x) \doteq \int_0^1 d\theta f\big(Q(p,x) + \theta(\hat{Q}(p,x) - Q(p,x))\big|\, x\big)$.

In order to invert (5.2) we assume a local lower bound on the conditional densities, namely the existence of a positive c such that $f(y|x) \geq c > 0, \forall y \in |y - Q(p,x)| < \varepsilon_\delta$. Here ε_δ is chosen

as a function of the convergence of $\hat{F}$ to F (see Theorem 1). For T sufficiently large, this implies that $\hat{F}(\hat{Q}(p,x)|x)$ is in a $\delta-$neighborhood of $F(\hat{Q}(p,x)|x)$, and by a (uniform) continuity argument, $\hat{Q}(p,x)$ is in a $\varepsilon_\delta-$neighborhood of $Q(p,x)$.

Since $F(y|x) = F_Y(x,y)/f(x)$, and $\hat{F}(\hat{Q}(p,x)|x) = p = F(Q(p,x)|\,x)$ by definition, we can write

$$\hat{Q}(p,x) - Q(p,x) = \frac{1}{\tilde{f}_{\hat{Q},Q}(x)} \left\{ \frac{F_Y(x,\hat{Q}(p,x))}{f(x)} - \frac{\hat{F}_Y(x,\hat{Q}(p,x))}{\hat{f}(x)} \right\}. \tag{5.3}$$

It is not possible to evaluate the right-hand side of (5.3) with the norm (4.8), since the densities in the denominator pose a measurability problem upon integration with respect to the x variable. We limit the analysis of the convergence of the conditional quantile to a neighborhood of the conditioning value $X = x$. It remains to determine how big the neighborhood of x should be. We know that the bound on the bias of $F_Y(x,y)$ is given by $\omega^{\mu}_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p$ from Lemma 2. This is a measure of the variation of F_Y in a set of radius $2^{1-j_1}a$. Hence we compute the convergence of $\hat{Q}(x,p)$ to $Q(x,p)$ in a neighborhood of x of radius $2^{1-j_1}a$. Indeed no improvement in the estimation error could be made by restricting ourselves to a smaller set containing x. The following theorem gives the consistency result.

Theorem 2. *Let the assumptions of Lemmas 2 and 3 hold with $\{(Y_t, X_t = Y_{t-1})\}$, $\{Y_t\}_{t\in\mathbb{Z}}$ being a stationary stochastic process. Let furthermore the marginal density $f_{X_t}(x)$ of the stationary process be bounded away from 0 for $x \in |x-\xi| < 2^{1-j_1}a$, where $X_t = \xi$ is a value taken by the conditioning variable. Then*

$$\begin{aligned} \big\| \hat{Q}(p,\cdot) - Q(p,\cdot)\,\big|L_p(\mathscr{L}_q)\big\|^{\rho}_{[|x-\xi|<2^{1-j_1}a]} &\leqslant \tilde{C}(\xi)\cdot \sup_y \big\| \hat{A}^{(T)}_{\mathbf{j}}\big(F\big)(\cdot,y) - F(\cdot,y)\,\big|L_p(\mathscr{L}_q)\big\|^{\rho}_{[|x-\xi|<2^{1-j_1}a]} \\ &\leqslant \tilde{C}(\xi)\cdot \sup_y \big\| \hat{A}^{(T)}_{\mathbf{j}}\big(F\big)(\cdot,y) - F(\cdot,y)\,\big|L_p(\mathscr{L}_q)\big\|^{\rho} \end{aligned} \tag{5.4}$$

where $\tilde{C}$ depends on the value of the conditioning value ξ. The bound and convergence rate of the right hand side of (5.4) *are given in Theorem 1.*

5.2 Implementation and Monte Carlo experiments

In this section we implement our method on simulated time series, and compare it with a kernel based conditional quantile estimator.

In the following we depart from the usual dyadic definition of the resolution level. We allow the resolution level j to be chosen in a continuum instead of being constrained to be a power of 2. As shown by Hall and Penev (2001), there are marked improvements in applications by choosing the resolution level in a continuum. This is especially true for a data driven choice of the smoothing parameter. Opting for a continuum has no impact on the theoretical results since we never take advantage of the dyadic nature of the scaling factor 2^j in the proofs. This requires notational changes only.

We start by choosing a pair of primal and dual bases $\{\varphi(\cdot-k)\}_{k\in\mathbb{Z}}$, $\{\tilde{\varphi}(\cdot-k)\}_{k\in\mathbb{Z}}$ verifying properties (2.2) to (2.7). We use translates of B-Splines for the primal basis. For more details on B-Splines we refer to Chui (1992). We consider a B-Spline function of order N, from now on, ${}_N\varphi(x)$, translated so that its nodes correspond to integer values, regardless of whether N is even or odd.

To build the dual basis we use the translates of the indicator function of the support of the generator of the primal basis. That is, if ${}_N\varphi(x)$ is the B-Spline that lives on $\big[\lfloor -N/2+1\rfloor, \lfloor N/2+1\rfloor\big)$, $\lfloor t\rfloor$ being the largest integer not greater than t, then ${}_N\tilde{\varphi}(x) = \frac{1}{N}\mathbb{I}\big([\lfloor -N/2+1\rfloor, \lfloor N/2+1\rfloor]\big)$.

Consider the density estimator of equation (4.5). The coefficients of the scaling functions ${}_N\varphi_{jk}$ are then given by:

$$\frac{1}{T}\sum_{t=1}^{T} {}_N\tilde{\varphi}_{jk}(X_t) = \frac{1}{T}\cdot\{\text{number of } X_t \text{ in the support of}{}_N\varphi_{jk}\}\ .$$

Hence, the estimation procedure boils down to counting the number of points that fall within the support of the scaling functions. We could think of the operator (4.5) as a special version of a smoothed histogram, even though we are actually dealing with overlapping supports. Such a density estimator is interesting because it is fast. Indeed suppose you want to estimate the density of a random variable on a grid of M points. Since the reconstructing functions φ_{jk} are known beforehand, you know the $\nu_a \times M$ matrix of values taken on the grid by the φ_{jk}'s, where ν_a is the number of overlapping basis functions on each grid point. Then the estimation consists in counting the number of data falling in the support of every φ_{jk}, in multiplying element by element with the former $\nu_a \times M$ matrix, and finally in summing on the columns to get density estimates at the M selected points. Such an algorithm is simple and quick, and there are no constraints on the number of points, unlike in Fast Fourier Transform based algorithms. The simulations (for which codes are available on request) are carried out with the free programming software *Ox*, see Doornik (2002).

The design of the Monte Carlo experiments is as follows. We consider a centered stationary autoregressive process of order one, whose root is equal to 0.6. The innovations are chosen to be either symmetric or asymmetric. For the symmetric case, we draw from a Gaussian distribution with zero mean and unit variance. For the asymmetric case, we draw from a skewed histogram. This histogram is plotted in Figure 1. The sample size is equal to 2000, while the number of Monte Carlo replications is equal to 1000. The kernel estimator relies on a product quartic kernel. Our estimators require the choice of four smoothing parameters, two in the pdf-like direction, h_x and j_x, and two in the cdf-like direction, h_y and j_y. As discussed after equations (3.10) and (3.11), the crucial choice is the one of the smoothing parameters in the x direction. The choice of the smoothing parameters in the y direction is not as critical since there is no bias-variance trade-off for this component. The smoothing parameters in the y direction will be chosen to be smaller than all the possible choices of the parameters in the x direction.

First we opt for a "best case" framework in the sense that we select the bandwidth or resolution level which minimizes a given loss function for each Monte Carlo run. Some preliminary simulations have been made to determine suitable grids. We have chosen to select the bandwidth h_x in the grid $\{0.15, 0.30, \ldots, 2.4\}$ for the symmetric case, namely 16 values, and $\{0.5, 0.1, \ldots, 1.2\}$ for the asymmetric case, namely 24 values. The resolution level j_x is selected among 17 values $\{0, 0.25, \ldots, 4\}$ for the symmetric case and among 12 values $\{1, 1.5, \ldots, 6\}$ for the asymmetric case. We have checked both in the kernel case and the wavelet case that the smoothing parameter minimizing the given loss function lies within the selected range. The smoothing parameters in the y direction are chosen to be $h_y = 0.08$ and $j_y = 6$. We use two different loss functions, namely an Integrated Absolute Deviation and an Integrated Square Error. We integrate over the probability interval $[0, 1]$, and we condition with respect to the simulated 2000^{th} value in predicting the conditional quantile associated with the next observation. The deviation or error is computed with respect to the true conditional quantile. Loss function values are then averaged through all runs to get a Mean Integrated Absolute Deviation (MIAD) and a Mean Integrated Square Error (MISE). The results of the simulations are reported in Table 1.

As it can be seen, the wavelet estimator performs better in the two cases, and the difference is clearer in the asymmetric case. More specifically, in the asymmetric case, the wavelet estimator displays a 29% decrease in the MIAD and a 65% decrease in the MISE with respect to the kernel estimator, while, in the symmetric case, we have a 5% decrease in the MIAD and a 7% decrease in the MISE of the wavelet estimator, compared to the kernel method. In the asymmetric case the

	Symmetric case		Asymmetric case	
	Kernel	Wavelet	Kernel	Wavelet
MIAD	5.86E-02	5.60E-02	4.89E-02	3.80E-02
MISE	8.20E-03	7.65E-03	5.69E-03	3.46E-03

Table 1: MIAD and MISE for the kernel and wavelet estimators when choosing the optimal bandwidth or resolution level within each run.

better performance of the wavelet estimator could have been anticipated. It is interesting to see that also in the symmetric case the wavelet estimator behaves better, even if not as clearly as in the previous case. In particular in the symmetric case we benefit strongly from choosing the resolution level j_x in a continuum of values. If we use integer values for j_x only, the picture is reversed, and the kernel estimator wins the competition, but by just a little.

We end this section by illustrating a methodology that can be followed in practical applications to make a data-driven choice of the resolution level j_x. As discussed after Theorem 2, the convergence of the conditional quantile estimator is driven by the convergence of the estimator of the univariate "design" density $f(x)$ of the conditioning variable, so that we can try to apply methods developed for univariate densities. We use a leave-one-out cross-validation methodology. In our setup, the univariate density $f(x)$ is the marginal density of the stationary process, and so it is relatively smooth and homogeneous. Given this characteristic of $f(x)$, we choose to apply the cross-validation method on the whole support of the density, unlike Tribouley (1995) and Hall and Penev (2001) who split the support of the density into a number of subregions of relatively homogeneous smoothness. Applying this methodology to the same 1000 series used previously selects an average j_x of 2.55 for the asymmetric case and of 1.94 for the symmetric case. We compare these values with the ones that actually minimize the MISE. In order to do this we repeat the same forecast exercise as in the "best case" study, but this time we select one resolution level within a grid of values, and maintain it fixed across all 1000 runs. The grid of values for j_x is $\{0, 0.125, \ldots, 6\}$ for both the symmetric and asymmetric cases. The j_x that minimizes the MISE is 2.5 in the asymmetric case, and 1.375 in the symmetric case. On average, the cross validation procedure then delivers the exact value of j_x in the asymmetric case, and a slightly undersmoothing one in the symmetric case. Choosing $j_x = 1.94$ in the symmetric case causes an increase of 12% in the value of the MISE with respect to the one we obtain if the correct $j_x = 1.375$ is selected. The relatively small error introduced by the data driven choice of j_x speaks in favor of both the applicability of the method and the robustness of the wavelet conditional quantile estimator with respect to the choice of the resolution level. Unreported results show that in most cases, the plot of the MISE against the resolution level j_x gives a somewhat flatter curve than the one that can be obtained by a kernel estimator. We think that this robustness feature of the wavelet method can be explained by the choice of the dual basis function used to build the estimator, see equation (4.7). For the kernel estimator there is the classical variance-bias trade off. The quartic kernel function becomes narrower and hence more variable as the bandwidth decreases. The wavelet estimator takes advantage of using non-orthogonal basis functions. In fact, when increasing the resolution level j_x only the primal scaling function $\varphi_{jk}(x)$ behaves as a kernel of higher order (such as the quartic one). However, the dual scaling function $\tilde{\varphi}_{jk}(x)$, a boxcar (Haar) function used to construct the coefficients in the reconstruction (4.7), suffers less from increasing variability on finer levels. A boxcar function used to construct the local average assigns equal weights to all observations. This tends to be numerically more stable than a local average provided by a higher order scaling function or a kernel of higher order. This is likely to dampen the increase in numerical variability of the wavelet estimator compared to the kernel estimator.

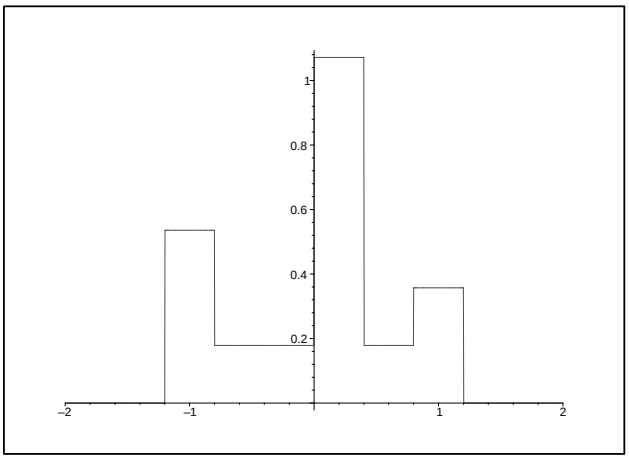

Figure 1: Distribution of the innovations for the asymmetric case.

6 Conclusion

In this paper we have further developed the DP approach of constructing shape-preserving nonparametric estimators of probabilistic functions (cdfs and pdfs). The wavelet methodology, taylored to reconstruct functions with low spatial regularity, has been extended to higher dimensions and to serially dependent data. In contrast to existing work this approach does not need to use pre- or post-processors applied to traditional wavelet estimators in order to make them positive and integrating to one for pdf estimation, and monotone for cdf estimation. We have investigated and defined appropriate norms of convergence, and we have derived rates of consistency for our estimators in these norms. We have applied our general methodology to the specific problem of conditional quantile estimation for dependent data in financial time series analysis. This has required to treat the specific situation of the intertwining of a cdf component and a pdf component in a curve estimation set-up, and to face and solve the technical difficulties of this nonparametric framework. Last but not least we have designed tractable algorithms relying on B-splines in that context.

Our method is still linear, and some words of comparison with both linear kernel estimation and nonlinear wavelet estimation seem to be necessary here. First, our linear wavelet estimators are performing uniformly not worse than kernel estimators. This means that they offer advantages for some functions with local structure without losing performance for smoother functions. Second, they provide a starting point for more flexible constructions. Indeed we have more options to adapt the construction of the estimator to the situation at hand with our non-orthogonal wavelets (scaling functions). We have seen for instance that computing the empirical inner products boils down to counting the number of observations falling in a given interval. Furthermore primal and dual bases are not as tightly related as in a biorthogonal set-up. To meet the moment conditions a change in the primal scaling function to obtain smoother reconstructions only requires a change in the support of the indicator function used as dual basis. Linear wavelet methods have already shown their practical interest in empirical analysis. Lee and Hong (2001) find that even linear wavelet methods capture irregularities in the spectral density better than kernel methods.

We have to acknowledge that the real strength of wavelet estimators shows up when it comes to nonlinear estimators, i.e., threshold estimators. Presently it is not clear how to design a neat methodology to maintain the shape-preserving property of the resulting wavelet threshold estimator. Simply deleting the non-significant empirical wavelet coefficients at "arbitrary" locations destroys this property. We believe that the "zero-tree" wavelet estimators of Shapiro (1993) could be an interesting alley for future research in that respect. This construction keeps a group of em-

pirical wavelet coefficients at a specific location and scale together with all "coarser scale parents" at the same location over all coarser scales. This yields a kind of "locally linear" complete reconstruction structure. However, how to build the wavelet functions is not clear in this non-orthogonal set-up. We think that one possibility is to follow the general device of Cohen (2003) to extend our work.

To finish we summarize again the points of methodological interest:

1. Our method can be applied to probabilistic functions belonging to a large variety of smoothness classes, and, in particular, to specific classes of non smooth functions.

2. The wavelet estimators are shape preserving (but not shape imposing). This type of wavelets are well suited for applications in many fields of statistics in which a shape restriction exists on the function to be estimated. The extension of the set-up of DP (Section 3) to a multivariate setting has therefore a theoretical interest that goes beyond the application we have investigated in this work.

3. A large flexibility is permitted in choosing the wavelet bases, which implies clear computational advantages.

Acknowledgments

O. Scaillet and A. Cosma (during his stay at the University of Lugano) acknowledge financial support by the Swiss National Science Foundation through the National Center of Competence in Research: Financial Valuation and Risk Management (NCCR FINRISK), while R. von Sachs gratefully acknowledges financial support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy). We also would like to thank the Editor and two referees for their comments which helped to improve presentation of the paper. We are grateful to A. Antoniadis, V. Delouille, seminar participants at Université de Genève and Université Catholique de Louvain, as well as participants of the WCES 2005 London conference and the EMS 2005 Oslo conference for helpful comments.

References

Anastassiou, G. and Yu, X. (1992). Monotone and probabilistic wavelet approximation. *Stochastic Analysis and Applications*, 10(3):251–264.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory*, 18:169–192.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, 18:17–39.

Chen, X., Hansen, L., and Scheinkman, J. (1998). Shape-preserving estimation of diffusions. *University of Chicago, Discussion Paper, http://citeseer.ist.psu.edu/chen98shapepreserving.html.*

Cheng, M.-Y., Gasser, T., and Hall, P. (1999). Nonparametric density estimation under unimodality and monotonicity constraints. *Journal of Computational and Graphical Statistics*, 8:1–21.

Chui, C. (1992). *An Introduction to Wavelets.* Academic Press, San Diego.

Cohen, A. (2003). *Numerical Analysis of Wavelet Methods*, volume 32 of *Studies in Mathematics and its Applications.* Elsevier, North-Holland.

Cosma, A., von Sachs, R., and Scaillet, O. (2005). Multivariate wavelet-based shape preserving estimation for dependent observations. *Discussion Paper 0516, Institut de statistique, Université catholique de Louvain, Louvain-la-Neuve; http://www.stat.ucl.ac.be/ISpub/dp/2005/dp0516.pdf.*

Dechevsky, L. and Penev, S. (1997). On shape preserving probabilistic wavelet approximators. *Stochastic Analysis and Applications*, 15(2):187–215.

Dechevsky, L. and Penev, S. (1998). On shape preserving wavelet estimators of cumulative distribution functions and densities. *Stochastic Analysis and Applications*, 16(3):423–462.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics*, 24:508–539.

Doornik, J. (2002). *Object-Oriented Matrix Programming Using Ox*. Timberlake Consultants Press and Oxford, London, 3rd edition.

Doukhan, P. and Léon, J. R. (1990). Déviation quadratique d'estimateurs de densité par projections orthogonales. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 310:425–430.

Doukhan, R. (1988). Formes de Töplitz associées à une analyse multiéchelle. *C.R. Acad. Sci. Paris Sér. A*, 306:663–666.

Hall, P. and Penev, S. (2001). Cross-validation for choosing resolution level for nonlinear wavelet curve estimators. *Bernoulli*, 7(2):317–342.

Hall, P. and Van Keilegom, I. (2004). Testing for monotone increasing hazard rate. *Annals of Statistics*, to appear.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94:154–163.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. B. (1998). *Wavelets, Approximation and Statistical Applications*, volume 129 of *Lecture Notes in Statistics*. Springer Verlag, New York.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statistics and Probability Letters*, 13:15–24.

Kerkyacharian, G., Picard, D., and Tribouley, K. (1996). l^p adaptive density estimation. *Bernoulli*, 2(3):229–247.

Lee, J. and Hong, Y. (2001). Testing for correlation of unknown form using wavelet methods. *Econometric Theory*, 17:386–423.

Masry, E. (1994). Probability density estimation from dependent observations using wavelets orthonormal bases. *Statistics and Probability Letters*, 21:181–194.

Masry, E. (1997). Multivariate probability density estimation by wavelet methods: Strong consistency and rates for stationary time series. *Stochastic Processes and their Applications*, 67:177–193.

McFadden, D. (2003). *Economic Choices*, pages 330–365. Nobel Lectures in Economic Sciences 1996-2000. World Scientific Publishing Company.

McFadden, D. and Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15:447–470.

Meyer, Y. (1992). *Ondelettes et Operateurs*, volume 1: Ondelettes. Cambridge : Cambridge University Press, Paris.

Neumann, M. H. and von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Annals of Statistics*, 25(1):38–76.

Nikol'skiĭ, S. M. (1975). *Approximations of Functions of Several Variables and Imbedding Theorems*. Springer-Verlag, Berlin Heidelberg New York.

Ogden, R. T. (1996). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser Boston Inc., Cambridge, MA, USA.

Penev, S. and Dechevsky, L. (1997). On non-negative wavelet-based density estimators. *Journal of Nonparametric Statistics*, 7:365–394.

Petrushev, P. P. and Popov, V. A. (1987). *Rational Approximation of Real Functions*. Cambridge University Press, Cambridge.

Pham, T. D. and Tran, L. T. (1980). The strong mixing property od the autoregressive moving average time series model. *Séminaire de Statistique*, pages 59–76.

Pinheiro, A. and Vidakovic, B. (1997). Estimating the square root of a density via compactly supported wavelets. *Computational Statistics and Data Analysis*, 25:399–415.

Shapiro, J. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Image Processing*, 41(12):3445–3462.

Tribouley, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statistica Neerlandica*, 49:41–62.

Tribouley, K. and Viennet, G. (1998). l_p adaptive density estimation in a beta-mixing framework. *Annales de l'Institut Henri Poincaré, Section B, Calcul des Probabilités et Statistique*, 34:179–208.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228237.

APPENDICES

A Proofs for Section 4

Proof of Lemma 2. We will bound the following quantity from above:

$$E_{\mathbf{j}}(F)(x,y) = A_{\mathbf{j}}(F)(x,y) - F(x,y)\ ,$$

through the use of the intermediate approximation $F_{\mu,h}$ by Steklov means (see Appendix C). Then from the linearity of $A(F)(x,y)$ and from Minkowsky inequality we get:

$$\begin{aligned}
&\sup_y \big\|E_{\mathbf{j}}(F)(\cdot,y)\big\|_p = \sup_y \big\|A_{\mathbf{j}}(F)(\cdot,y) - F(\cdot,y)\big\|_p \\
&= \sup_y \big\|A_{\mathbf{j}}(F-F_{\mu,h}) - (F-F_{\mu,h}) + A_{\mathbf{j}}(F_{\mu,h}) - F_{\mu,h}\big\|_p \\
&\leqslant \sup_y \big\|E_{\mathbf{j}}(F-F_{\mu,h})\big\|_p + \sup_y \big\|E_{\mathbf{j}}(F_{\mu,h})\big\|_p.
\end{aligned} \tag{A.1}$$

By taking advantage of the properties relating Steklov means to moduli of smoothness, we bound the two terms of the right hand side of inequality (A.1) separately. For the first term, following the details given in Cosma et al. (2005), application of property (C.1) of the Steklov means onto $g(x,y) \doteq F(x,y) - F_{\mu,h}(x,y)$, with $\mu = 2$ and $h = (2^{1-j_1}, 2^{1-j_2})$, gives

$$\sup_y \big\|E_{\mathbf{j}}(g(\cdot,y))\big\|_p \leqslant \nu_a^2 2^{\frac{4}{p}+2} a^{\frac{2}{p}} \big\|\widetilde{\varphi}\big\|_{p'} \big\|\varphi\big\|_\infty \sup_y \sup_{\mathbf{i}\in\mathbb{R}^2} \sup_{0<\delta<1} \big\|\Delta^2_{\mathbf{i}\delta(2^{1-j^*}a)} F(\cdot,y)\big\|_p, \tag{A.2}$$

where j^* solves the equation $(2^{1-j^*})^2 = (2^{1-j_1})^2 + (2^{1-j_2})^2$.

Studying the second term of (A.1), i.e. the approximation of the Steklov mean $F_{\mu,h}$, we assume that $F_{\mu,h}$ has almost everywhere partial and mixed derivatives up to the second order. Using a second order Taylor expansion with an integral remainder, properties (2.5) and (2.7), Minkowsky (generalized) inequality and property (C.2), we get the upper bound (see Cosma et al. (2005))

$$\begin{aligned}
&\sup_y \big\|E_{\mathbf{j}}(g(\cdot,y))\big\|_p \\
&\leqslant \frac{c_1}{2a}\Big\{\big\|\varsigma_x(2^{j_1}\cdot)\big\|_\infty \sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p + \big\|\varsigma_y(2^{j_2}\cdot)\big\|_\infty \sup_y \sup_{0<\delta<1} \big\|\Delta_{\mathbf{e}_y\delta(2^{1-j_2}a)} F(\cdot,y)\big\|_p\Big\} \\
&\quad + a^{\frac{2}{p}}\,\nu_a^2\,2^{\frac{4}{p}-1}\,\big\|\tilde{\varphi}\big\|_{L_{p'}}\big\|\varphi\big\|_\infty\, 2c_2 \cdot \Big\{\sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p + \sup_y \sup_{0<\delta<1} \big\|\Delta_{\mathbf{e}_y\delta(2^{1-j_2}a)} F(\cdot,y)\big\|_p\Big\} \\
&\quad + a^{\frac{2}{p}}\,\nu_a^2\,2^{\frac{4}{p}-1}\,\big\|\tilde{\varphi}\big\|_{L_{p'}}\big\|\varphi\big\|_\infty\, c_2 \cdot \Big\{\sup_y \omega^2_{\mathbf{e}_x}\big(g(\cdot,y)\big)_p + \sup_y \sup_{0<\delta<1} \big\|\Delta^2_{\mathbf{e}_y\,\delta\cdot(2^{1-j_2}a)} F(\cdot,y)\big\|_p\Big\},
\end{aligned} \tag{A.3}$$

where c_1 and c_2 are constants not depending on $\mathbf{j} = (j_1, j_2)$. For an explicit expression of c_1 and c_2 see the detail of the proof in Appendix B of Cosma et al. (2005). We can now put together (A.2) and (A.3) to obtain a final bound for the approximation error:

$$\begin{aligned}
&\sup_y \big\|A_{\mathbf{j}}(F)(\cdot,y) - F(\cdot,y)\big\|_p \\
&\quad \leqslant c_1'\Big\{\big\|\varsigma_x(2^{j_1}\cdot)\big\|_\infty \sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p + \big\|\varsigma_y(2^{j_2}\cdot)\big\|_\infty \sup_y \sup_{0<\delta<1} \big\|\Delta_{\mathbf{e}_y\delta(2^{1-j_2}a)} F(\cdot,y)\big\|_p\Big\} \\
&\qquad + c_2'\,\big\|\tilde{\varphi}\big\|_{L_{p'}}\big\|\varphi\big\|_\infty \cdot \Big\{\sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot,y), 2^{1-j_1}a\big)_p + \sup_y \sup_{0<\delta<1} \big\|\Delta_{\mathbf{e}_y\delta(2^{1-j_2}a)} F(\cdot,y)\big\|_p\Big\} \\
&\qquad + c_2'' \cdot \Big\{\sup_y \omega^2_{\mathbf{e}_x}\big(F(\cdot,y)\big)_p + \sup_y \sup_{0<\delta<1} \big\|\Delta^2_{\mathbf{e}_y\,\delta\cdot(2^{1-j_2}a)} F(\cdot,y)\big\|_p\Big\} \\
&\qquad + c_2'''\big\|\widetilde{\varphi}\big\|_{p'}\big\|\varphi\big\|_\infty \sup_y \sup_{\mathbf{i}\in\mathbb{R}^2} \sup_{0<\delta<1} \big\|\Delta^2_{\mathbf{i}\delta(2^{1-j^*}a)} F(\cdot,y)\big\|_p,
\end{aligned} \tag{A.4}$$

where the expressions of the constants can easily be made explicit by comparing (A.2), (A.3) and (A.4). Finally, (A.4) can be further simplified. Simple algebra (i.e., adding and subtracting $F(x+2s, y+t)$) leads to the inequality
$\sup_{0<\delta<1}\big\|\Delta^2_{\mathbf{i}\delta(2^{1-j^*a})}F(\cdot,y)\big\|_p \leqslant 2\cdot\sup_y \sup_{0<\delta<1}\big\|\Delta_{\mathbf{e}_y\delta(2^{1-j_2}a)}F(\cdot,y)\big\|_p + \sup_y \omega^2_{\mathbf{e}_x}\big(F(\cdot,y)\big)_p$. □

Proof of Lemma 3. We have to find an upper bound for the quantity:

$$\left\|\hat{A}^{(T)}_{\mathbf{j}}(F)(\cdot,y) - \mathbb{E}\hat{A}^{(T)}_{\mathbf{j}}(F)(\cdot,y)\Big|L_p(\mathscr{L}_q)\right\| = \left\|\Big(\mathbb{E}\big|\hat{A}^{(T)}_{\mathbf{j}}(F)(\cdot,y) - A_{\mathbf{j}}(F)(\cdot,y)\big|^q\Big)^{1/q}\right\|_p , \tag{A.5}$$

and start by expressing the estimator $\hat{A}^{(T)}_{\mathbf{j}}(F)(x,y)$ as:

$$\begin{aligned}
\hat{A}^{(T)}_{\mathbf{j}}(F)(x,y) &= \sum_{\mathbf{k}\in\mathbb{Z}^2} \widehat{\langle F,\tilde{\varphi}_{\mathbf{jk}}\rangle}\,\varphi_{\mathbf{jk}}(x,y) \\
&= \sum_{\mathbf{k}\in\mathbb{Z}^2}\sum_{t=1}^{T} 2^{-\frac{j_2}{2}}\left(\frac{\tilde{\varphi}_{j_1k_1}(X_t)}{T} - \frac{\tilde{\varphi}_{j_1k_1}(X_t)\tilde{\Phi}(2^{j_2}Y_t - k_2)}{T}\right)\varphi_{\mathbf{jk}}(x,y) \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{\mathbf{k}\in\mathbb{Z}^2}\widehat{\langle F,\tilde{\varphi}_{\mathbf{jk}}\rangle}_t\,\varphi_{\mathbf{jk}}(x,y) = \frac{1}{T}\sum_{t=1}^{T}\hat{A}_{\mathbf{j},t}(F)(x,y).
\end{aligned} \tag{A.6}$$

Then we determine the pointwise variance for the stochastic variable:

$$Z_t(x,y) \doteq \hat{A}_{\mathbf{j},t}(f)(x,y) - A_{\mathbf{j}}(F)(x,y)\ , \tag{A.7}$$

which fulfills $|Z| \leqslant 2\cdot 2^{j_1}$ since both $\hat{A}_{\mathbf{j}}(f)$ and $A_{\mathbf{j}}(F)$ are smaller than 2^{j_1}, and also $\mathbb{E}Z = 0$. With $q \leqslant 2$, by a stationarity argument and a convexity argument

$$\begin{aligned}
\mathbb{E}\big|\hat{A}^{(T)}_{\mathbf{j}}(x,y) - A_{\mathbf{j}}(x,y)\big|^q &= \mathbb{E}\left|\sum_t \frac{Z_t(x,y)}{T}\right|^q \leqslant \frac{1}{T^q}\left[\mathbb{E}\Big(\sum_t Z_t(x,y)\Big)^2\right]^{q/2} \\
&= \frac{1}{T^q}\left[\sum_{t=1}^{T}\mathbb{E}\big(Z_t^2(x,y)\big) + 2\sum_{p=1}^{T-1}(T-p)\mathbb{E}\big(Z_T(x,y)Z_{T-p}(x,y)\big)\right]^{q/2}.
\end{aligned}$$

Taking the L_p norm with respect to x and the supremum with respect to y, we obtain:

$$\begin{aligned}
&\sup_y \left\|\hat{A}^{(T)}_{\mathbf{j}}(F)(\cdot,y) - \mathbb{E}\hat{A}^{(T)}_{\mathbf{j}}(F)(\cdot,y)\Big|L_p(\mathscr{L}_q)\right\|^\rho \\
&\leqslant \sup_y \frac{1}{T^\rho}\left\{\Big\|\sum_{t=1}^{T}\mathbb{E}\big(Z_t^2(\cdot,y)\big)\Big\|^{\rho/2}_{p/2} + \Big\|2T\sum_{p=1}^{T-1}\Big(1-\frac{p}{T}\Big)\mathbb{E}\big(Z_T(\cdot,y)Z_{T-p}(\cdot,y)\big)\Big\|^{\rho/2}_{p/2}\right\} \\
&\doteq \bar{V}_0 + \bar{V}_1.
\end{aligned} \tag{A.8}$$

The first summand of (A.8), that we refer to as $\bar{V}_0$, is the only term in the *i.i.d.* case. $\bar{V}_1$ is the additional term obtained by estimating $\hat{A}^{(T)}_{\mathbf{j}}$ from serially dependent data. In the sequel we give a sketch of the treatment of both terms $\bar{V}_0$ and $\bar{V}_1$; for all the details we refer to Cosma et al. (2005).

A. Study of $\bar{V}_0$ term: We introduce as variance of Z, for a fixed (x,y),

$$\begin{aligned}
\sigma^2(x,y) = \mathbb{E}[\hat{A}^{(T)}_{\mathbf{j}}(F)(x,y)]^2 - A_{\mathbf{j}}(F)(x,y)^2 \\
= \sum_{\mathbf{k}\in\mathbb{Z}^2}\sum_{\mathbf{l}\in\mathbb{Z}^2}\big[\mathbb{E}\big(\widehat{\langle F,\tilde{\varphi}_{\mathbf{jk}}\rangle}\widehat{\langle F,\tilde{\varphi}_{\mathbf{jl}}\rangle}\big) - \langle F,\tilde{\varphi}_{\mathbf{jk}}\rangle\langle F,\tilde{\varphi}_{\mathbf{jl}}\rangle\big]\varphi_{\mathbf{jk}}(x,y)\varphi_{\mathbf{jl}}(x,y)\ .
\end{aligned} \tag{A.9}$$

The idea is to control each term in square brackets of the last equation, i.e.,

$$\begin{aligned}\Delta_{\mathbf{j};\mathbf{kl}} &\doteq \mathbb{E}\big(\widehat{\langle F,\tilde{\varphi}_{\mathbf{jk}}\rangle\langle F,\tilde{\varphi}_{\mathbf{jl}}\rangle}\big) - \langle F,\tilde{\varphi}_{\mathbf{jk}}\rangle\langle F,\tilde{\varphi}_{\mathbf{jl}}\rangle \\ &= 2^{-j_2}\Big\{ \operatorname{Cov}\big(\tilde{\varphi}_{j_1k_1}(S),\tilde{\varphi}_{j_1l_1}(S)\big) + \operatorname{Cov}\big(\tilde{\varphi}_{j_1k_1}(S),\tilde{\varphi}_{j_1l_1}(S)\tilde{\Phi}(2^{j_2}T-l_2)\big) \\ &+ \operatorname{Cov}\big(\tilde{\varphi}_{j_1k_1}(S)\tilde{\Phi}(2^{j_2}T-k_2),\tilde{\varphi}_{j_1l_1}(S)\big) + \operatorname{Cov}\big(\tilde{\varphi}_{j_1k_1}(S)\tilde{\Phi}(2^{j_2}T-k_2),\tilde{\varphi}_{j_1l_1}(S)\tilde{\Phi}(2^{j_2}T-l_2)\big)\Big\}\,,\end{aligned} \tag{A.10}$$

where S and T are used to denote two random variables drawn from a bivariate cdf $F_{S,T}(S,T)$, corresponding to the empirical inner product of equation (A.6) formally taken at sample size $T=1$.

Let us study each of the four terms separately. The first term is a one dimensional term which has already been studied in DP (1998). By using integration by parts we can rewrite the above in terms of the moduli of smoothness, and thus bound the variance $\sigma(x,y)$:

$$\begin{aligned}\sigma^2(x,y) &= \sum_{\mathbf{k}\in\mathbb{Z}^2}\sum_{\mathbf{l}\in\mathbb{Z}^2}\Delta_{\mathbf{j};\mathbf{kl}}\varphi_{\mathbf{jk}}(x,y)\varphi_{\mathbf{jl}}(x,y) \leqslant \sup_{\mathbf{kl}}|\Delta_{\mathbf{j};\mathbf{kl}}|\sum_{\mathbf{k}\in\mathbb{Z}^2}\sum_{\mathbf{l}\in\mathbb{Z}^2}\varphi_{\mathbf{jk}}(x,y)\varphi_{\mathbf{jl}}(x,y) \\ &=: 2^{j_1}\Delta(x,y);\end{aligned} \tag{A.11}$$

where the latter definition will now be used in the following estimate (we refer again to Cosma et al. (2005) for all the lengthy details):

$$\sup_y \frac{1}{T^\rho}\Big\| \sum_{i=1}^{T}\mathbb{E}\big(Z_i^2\big)\Big\|_{p/2}^{\rho/2} \leqslant \sup_y T^{-\frac{\rho}{2}}\Big(\int_{\mathbb{R}} dx\big(\sigma^2(x,y)\big)^{\frac{p}{2}}\Big)^{\frac{2}{p}\cdot\frac{\rho}{2}} \leqslant \sup_y\left(\frac{2^{j_1}}{T}\right)^{\rho/2}\big\|\,\Delta(\cdot,y)\big\|_{p/2}^{\rho/2}. \tag{A.12}$$

When taking the L_p-norm in (A.12), the terms contained in $\Delta(x,y)$ will be controlled by the modulus of smoothness in the x direction of $F(x,y)$. We recall that $\omega^1_{\mathbf{e}_x}\big(f,h\big)_p \leqslant \sup_y \omega^1_{\mathbf{e}_x}\big(F(\cdot,y),h\big)_p$. The terms containing $F(x,y)-F(s,t)$ can be split into two terms by adding $\pm F(x,t)$ and we obtain an increment in the x direction and one in the y direction. The increments in the x direction $F(x,t)-F(s,t)$ are controlled by the modulus $\omega^1_{\mathbf{e}_x}\big(F(\cdot,y),h\big)_p$, while the increments in the y direction $F(x,y)-F(x,t)$ can be controlled by $\|f(\cdot,y)\|_p$ through the inequality $\sup_y\sup_{|t|<h}\|\Delta_{\mathbf{e}_y,t}F(\cdot,y)\|_p \leqslant h\sup_y\|f(\cdot,y)\|_p$. Substituting the explicit expression for $\Delta(x,y)$, using Minkowsky generalized inequality and the properties of the moduli of smoothness gives:

$$\frac{1}{T^\rho}\Big\| \sum_{t=1}^{T}\mathbb{E}\big(Z_i^2\big)\Big\|_{p/2}^{\rho/2} \leqslant \left(\frac{2^{j_1}}{T}\right)^{\rho/2}\Big(d_1\big\|f(\cdot,y)\big\|_{p/2}^{\rho/2} + d_2(a)\,\omega^1_{\mathbf{e}_x}\big(f(\cdot,y),2^{1-j_1}a\big)_{p/2}^{\rho/2}\Big)\,. \tag{A.13}$$

To choose the value of ρ as a function of p and q, we refer to Appendix D. In (A.13) the parameter q is absent, and the parameter p appears through the $L_{p/2}$ norm of the quantity $\Delta(x,y)$, so that an adequate choice is such that

$$\begin{aligned}\frac{\rho^*}{2} &= \frac{1}{1+\log_2(c_p)} = \frac{1}{1+\log_2(\max\{1,2^{2/p-1}\})} \\ &= \frac{1}{1+(\max\{0,2/p-1\})} = \frac{1}{\max\{1,\frac{2}{p}\}} = \min\{p/2,1\},\end{aligned}$$

where c_p is constant in the quasi-triangular inequality, see again Appendix D. Since $f\in L_1$ for any density, we can give a more general expression for (A.13):

$$\begin{aligned}\sup_y \frac{1}{T^{\rho^*}}\Big\| \sum_{t=1}^{T}\mathbb{E}\big(Z^2(\cdot,y)\big)\Big\|_{p/2}^{\rho^*/2} &\leqslant \sup_y\left(\frac{2^{j_1}}{T}\right)^{\rho^*/2}\Big(d_1\max\Big\{\big\|f(\cdot,y)\big\|_1,\big\|f(\cdot,y)\big\|_{p/2}\Big\}^{\rho^*/2} \\ &+ d_2(a)\max\Big\{\omega^1_{\mathbf{e}_x}\big(f(\cdot,y),2^{1-j_1}a\big)_1,\omega^1_{\mathbf{e}_x}\big(f(\cdot,y),2^{1-j_1}a\big)_{p/2}\Big\}^{\rho^*/2}\Big)\,.\end{aligned} \tag{A.14}$$

The condition, for $1 < p < 2$, that $\sup_y \sup_{0\leqslant t\leqslant h} \int_{-\infty}^{+\infty} dx \Big(\int_0^1 d\alpha\, f(x+\alpha t, y)\Big)^{p/2} < +\infty$, ensures that the modulus $\omega_\lambda(f,h) \to_{h\to 0} 0$ even if $\lambda < 1$. The proof is found in DP (1998).

B. Study of $\bar{V}_1$ term: Dealing with the covariance part $\bar{V}_1$ of (A.8), we split the summation into two parts:

$$\begin{aligned}
&\sup_y T^{-\rho/2} \Big\| \sum_{p=1}^{T-1} \Big(1-\frac{p}{T}\Big)\Big(\mathbb{E}\big|Z_T(\cdot,y)Z_{T-p}(\cdot,y)\big|\Big)\Big\|_{p/2}^{\rho/2} \\
&\quad \leqslant \sup_y \frac{1}{T^{\rho/2}} \Big\| \sum_{p=1}^{n_T} \Big(1-\frac{p}{T}\Big)\Big(\mathbb{E}\big|Z_T Z_{T-p}\big|\Big)\Big\|_{p/2}^{\rho/2} \\
&\qquad + \sup_y \frac{1}{T^{\rho/2}} \Big\| \sum_{p=n_T+1}^{T-1} \Big(1-\frac{p}{T}\Big)\Big(\mathbb{E}\big|Z_T Z_{T-p}\big|\Big)\Big\|_{p/2}^{\rho/2} \\
&\quad \doteq S_1 + S_2.
\end{aligned} \tag{A.15}$$

Here the explicit dependence $Z(x,y)$ on (x,y) is omitted for readability.
To gauge S_1 we apply *assumption 1* (4.3), and get the bound

$$\big|\ \mathrm{Cov}\ \big(Z_T(x,y)Z_{T-p}(x,y)\big)\big| \ \leqslant M\ \Big(\mathbb{E}\big|Z_t(x,y)\big|\Big)^2.$$

Elementary calculations to bound $\mathbb{E}\big|Z_t(x,y)\big|$ (see Cosma et al. (2005)) lead to

$$S_1 \leqslant M\left(\frac{n_T}{T}\right)^{\rho/2} \sup_y \Big\{\omega^1_{\mathbf{e}_x}\big(f(\cdot,y), 2^{1-j_1}a\big)_p^\rho + \|f(\cdot,y)\|_p^\rho\Big\} = O\Big(\frac{n_T}{T}\Big)^{\rho/2}. \tag{A.16}$$

For the term S_2 of (A.15) we use α - mixing properties. Davydov inequality and stationarity give with $r > 2$,

$$\Big|\sum_{p=n_T+1}^{T}\Big(1-\frac{p}{T}\Big)\,\mathrm{Cov}\,\big(Z_T(x,y)Z_{T-p}(x,y)\big)\Big| \leqslant \sum_{p=n_T+1}^{T} 2\frac{r}{r-2}\big(2\alpha(p)\big)^{1-\frac{1}{r}}\Big(\mathbb{E}\big|Z_t(x,y)\big|^r\Big)^{2/r}. \tag{A.17}$$

Since $r > 2$ and since $|Z_t(x)| < 2^{j_1}$ uniformly in x, $\mathbb{E}|Z|^r \leqslant 2^{j_1\cdot(r-2)}\,\mathbb{E}Z_t(x,y)^2$, so that the right hand side of (A.17) can be bounded by

$$\begin{aligned}
C_r\cdot 2^{2\frac{r-2}{r}j_1}\big(\sigma^2(x,y)\big)^{2/r}\sum_{p=n_T+1}^{T}\alpha(p)^{1-\frac{1}{r}} &= C_r\cdot 2^{2\frac{r-2}{r}j_1}2^{\frac{2}{r}j_1}\Delta(x,y)^{\frac{2}{r}}\sum_{p=n_T+1}^{T}\alpha(p)^{1-\frac{1}{r}} \\
&\doteq C_r\cdot 2^{2\frac{r-1}{r}j_1}\Delta(x,y)^{\frac{2}{r}}S'_T(n_T).
\end{aligned}$$

We finally find the following bound for S_2:

$$S_2 \leqslant \frac{\tilde{C}_r}{T^{-\rho/2}}\Big(2^{2\frac{r-1}{r}j_1}S'_T(n_T)\Big)^{\rho/2}\sup_y \big\|\Delta(\cdot,y)\big\|_{p/r}^{\rho/r}, \tag{A.18}$$

with $\tilde{C}_r$ obtained by collecting all terms not depending on j_1 and T.

To finish the proof of Lemma 3, we compare the asymptotic behavior of $\bar{V}_0$ and S_2.

$$\frac{S_2}{\big(2^{j_1}/T\big)^{\rho/2}} = \tilde{C_r}'\frac{\Big(2^{2\frac{r-1}{r}j_1}S'_T(n_T)\Big)^{\rho/2}}{2^{j_1\cdot\rho/2}} = \tilde{C_r}'\left(\frac{2^{2\frac{r-1}{r}j_1}S'_T(n_T)}{2^{j_1}}\right)^{\rho/2} = \Big(2^{\frac{r-2}{r}j_1}S'_T(n_T)\Big)^{\rho/2}.$$

We observe that the ratio will tend to zero if $S'_T(n_T) = O\big(2^{-(\frac{r-2}{r}+\delta)j_1}\big)$ with $\delta > 0$, so, since by (4.4) $S'_T(n_T) = O(n_T^{-1})$, we need to impose $n_T = O\big(2^{(\frac{r-2}{r}+\delta)j_1}\big)$. In order to have $S_1/\bar{V}_0 \to 0$ as $T \to \infty$, n_T is constrained by (A.16) to grow such that $n_T/2^{j_1} \underset{T}{\longrightarrow} 0$, *i.e.* $n_T = O\big(2^{j_1(1-\theta)}\big)$, with $\theta > 0$. Since $r > 2$, $0 < \frac{r-2}{r} < 1$, we can choose a δ such that $\frac{r-2}{r} + \delta < 1$, that is $0 < \delta < \frac{2}{r} < 1$, and a $\theta = 1 - \frac{r-2}{r} - \delta$, that both satisfy the conditions for $S_1/\bar{V}_0, S_2/\bar{V}_0 \to 0$ simultaneously as $T \to \infty$.

□

B Proof for Section 5

Theorem 2. As seen in equation (5.3), we can write

$$\hat{Q}(p,\xi) - Q(p,\xi) = \frac{1}{\tilde{f}_{\hat{Q},Q}(\xi)}\left\{\frac{F(\xi,\hat{Q}(p,\xi))}{f(\xi)} - \frac{\hat{F}(\xi,\hat{Q}(p,\xi))}{\hat{f}(\xi)}\right\}$$
$$= \frac{1}{B(\xi)\cdot\tilde{f}_{\hat{Q},Q}(\xi)}\Big\{F(\xi,\hat{Q}(p,\xi)) - \hat{F}(\xi,\hat{Q}(p,\xi))\Big\} - \frac{A(\xi)}{B^2(\xi)\cdot\tilde{f}_{\hat{Q},Q}(\xi)}\big\{f(\xi) - \hat{f}(\xi)\big\},$$

where we have used the mean value theorem for the function $\frac{u}{v}$ in the two variables $u = \hat{F}(\xi,\hat{Q}(p,\xi)$ and $v = \hat{f}(\xi)$ with mean values $|B(\xi) - f(\xi)| \leqslant |f(\xi) - \hat{f}(\xi)|$, and $|A(\xi) - F(\xi,\hat{Q}(p,\xi))| \leqslant |F(\xi,\hat{Q}(p,\xi)) - \hat{F}(\xi,\hat{Q}(p,\xi))|$. The above equality is true for pointwise deviations $f(\xi) - \hat{f}(\xi), F(\xi,\hat{Q}(p,\xi)) - \hat{F}(\xi,\hat{Q}(p,\xi))$. Taking the $\mathscr{L}_q$ expectation on both sides we have

$$\mathbb{E}|\hat{Q}(p,x) - Q(p,x)|^q \leqslant \frac{1}{\tilde{f}_{\hat{Q},Q}(x)^q}\mathbb{E}\left|\frac{F(x,\hat{Q}) - \hat{F}(x,\hat{Q})}{B(x)} - \frac{A(x)}{B(x)^2}\big(f(x) - \hat{f}(x)\big)\right|^q.$$

Then we take the norm in a neighborhood $\mathscr{J} \doteq \{x | [\xi - x] < 2^{1-j_1}a\}$,

$$\big\|\,\hat{Q}(p,\cdot) - Q(p,x)\,\big|L_p(\mathscr{L}_q)\big\|^\rho_{\mathscr{J}} = \Big\|\,\big(\mathbb{E}|\hat{Q}(p,\cdot) - Q(p,\cdot)|^q\big)^{1/q}\Big\|^\rho_{\mathscr{J}}$$
$$\leqslant \frac{1}{\widetilde{f_{\hat{Q},Q}(\xi)}\,\widetilde{B(\xi)}}\sup_y\big\|\,F(\cdot,y) - \hat{F}(\cdot,y)\,\big|L_p(\mathscr{L}_q)\big\|^\rho_{\mathscr{J}} + \frac{\widetilde{A(\xi)}}{\widetilde{f_{\hat{Q},Q}(\xi)}\,\widetilde{B(\xi)}^2}\big\|\,f(\cdot) - \hat{f}(\cdot)\,\big|L_p(\mathscr{L}_q)\big\|^\rho_{\mathscr{J}},$$

where $A(\xi) \leqslant \widetilde{A(\xi)}$ in $\mathscr{J}$, $\tilde{f}_{\hat{Q},Q}(\xi) \geqslant \widetilde{f_{\hat{Q},Q}(\xi)}$ and $B(x) \geqslant \widetilde{B(\xi)}$ in $\mathscr{J}$.

The two norms in the last inequality are bounded by expressions identical to the right hand side of (5.4), but with $\omega^\mu(f, 2^{1-j_1}a)_{L_p(\mathscr{J})}$ instead of $\omega^\mu(f, 2^{1-j_1}a)_{L_p(\mathbb{R})}$. The result of Theorem 2 follows immediately by remarking that $\omega^\mu(f, 2^{1-j_1}a)_{L_p(\mathscr{J})} \leqslant \omega^\mu(f, 2^{1-j_1}a)_{L_p(\mathbb{R})}$.

□

C Steklov Means

For $g \in L_{1,loc}, \mu \in \mathbb{N}, 0 < h < \infty$, the Steklov functions (Steklov means) $g_{\mu,h}$ of a function in one variable is defined by:

$$g_{\mu,h}(x) = (-h)^{-\mu}\underbrace{\int_0^h\cdots\int_0^h\cdots}_{\mu}\left[\sum_{\nu=0}^{\mu-1}(-1)^{\nu-1}\binom{\mu}{\nu}g\left(x + \frac{\mu-\nu}{\mu}\sum_{\lambda=1}^{\mu}\theta_\lambda\right)\right]d\theta_1\cdots d\theta_\mu.$$

Steklov functions $g_{\mu,h}$ are related to g, and to the moduli of smoothness $\omega^\mu(g,t)_p$ by:

$$\big\|\,g - g_{\mu,h}\big\|_p \leqslant \omega^\mu(g,h)_p, \quad \big\|\,g^{(\nu)}_{\mu,h}\big\|_p \leqslant c_{\mu,\nu}h^{-\nu}\omega_\nu(g,h)_p, \quad \nu = 1,2,\cdots\mu,$$

where $c_{\mu,\nu}$ are positive constants. There exist explicit estimates from above for this constants. For more details see, for instance, Petrushev and Popov (1987).

For functions of two variables we can define an analogous function

$$g_{\mu,h}(x,y) = (-h_x h_y)^{-\mu} \underbrace{\int_0^{h_x} \cdots \int_0^{h_x}}_{\mu} \underbrace{\int_0^{h_y} \cdots \int_0^{h_y}}_{\mu} d\theta_1^x \cdots d\theta_\mu^x d\theta_1^y \cdots d\theta_\mu^y$$
$$\left[\sum_{\nu=0}^{\mu-1} (-1)^{\nu-1} \binom{\mu}{\nu} g\left(x + \frac{\mu-\nu}{\mu} \sum_{\lambda=1}^{\mu} \theta_\lambda^x, y + \frac{\mu-\nu}{\mu} \sum_{\lambda=1}^{\mu} \theta_\lambda^y \right) \right]$$

with the properties:

$$\|g - g_{\mu,h}\|_p \leqslant \sup_{\mathbf{i} \in \mathbb{R}^2} \omega_{\mathbf{i}}^{\mu}(g,h)_p, \tag{C.1}$$

$$\left\| \frac{\partial^\nu g_{\mu,h}}{\partial x_i^\nu} \right\|_p \leqslant c_{\mu,\nu} h^{-\nu} \omega_{\mathbf{e}_i}^{\nu}(g,h)_p, \quad \nu = 1,2,\cdots\mu. \tag{C.2}$$

D Properties of the space $L_p(\mathscr{L}_q)$

In Section 3 we introduced the space $L_p(\mathscr{L}_q)$. Recall that the triangular inequality holds with $\|g+h\|_A \leqslant c_A(\|g\|_A + \|h\|_A)$, $c_A \geqslant 1$. For $1 \leqslant p,q \leqslant \infty$ $L_p(\mathscr{L}_q)$ is a Banach space, while for $0 < p < 1$ and/ or $0 < q < 1$ the constant is: $c_A = c_p c_q = \max\{1, 2^{(1/p)-1}\} \cdot \max\{1, 2^{(1/q)-1}\}$. If A is a quasi normed space, then A^ρ, defined as $\{g \in A, \|g\|_{A^\rho} = \|g\|_A^\rho\}$ is a 1-quasi normed space, *i.e.* $c_A = 1$, with $\rho = 1/[1 + \log_2(c_A)]$. Now, $L_p(\mathscr{L}_q)^\rho$ is a Banach space for ρ such that $\frac{1}{\rho} = \max\{1, \frac{1}{p}, \frac{1}{q}, \frac{1}{p} + \frac{1}{q} - 1\}$. Finally we note that for $p = q$, the norm coincides with the usual L_p-risk, i.e. $\mathbb{E}\|\cdot\|_p$.