Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff *

Seok Young Hong† Oliver Linton‡

University of Cambridge

This version: 12 June 2018

Abstract

This paper studies nonparametric estimation of the infinite order regression $E(Y_t^k|\mathcal{F}_{t-1})$, $k \in \mathbb{Z}$ with stationary and weakly dependent data. We propose a Nadaraya-Watson type estimator that operates with an infinite number of conditioning variables. We propose a bandwidth sequence that shrinks the effects of long lags, so the influence of all conditioning information is modelled in a natural and flexible way, and the issues of omitted information bias and specification error are effectively handled. We establish the asymptotic properties of the estimator under a wide range of static and dynamic regressions frameworks, thereby allowing various kinds of conditioning variables to be used. We establish pointwise/uniform consistency and CLTs. We show that the convergence rates are at best logarithmic, and depend on the smoothness of the regression, the distribution of the marginal regressors and their dependence structure in a non-trivial way via the Lambert W function. We apply our methodology to examine the intertemporal risk-return relation for the aggregate stock market, and some new empirical evidence is reported. For the S&P 500 daily data from 1950-2017 using our estimator we report an overall positive risk-return relation. We also find evidence of strong time variation and counter-cyclical behaviour in risk aversion. These conclusions are possibly attributable to the allowance of further flexibility and the inclusion of otherwise neglected information in our method.

JEL Codes: C10; C58; G10.

*We gratefully acknowledge helpful comments from John Aston, Xiaohong Chen, Jeroen Dalderop, Paul Doukhan, Jiti Gao, Shuyi Ge, Hayden Lee, Yuan Liao, Mikhail Lifshits, Richard Nickl, Alexei Onatski, Hashem Pesaran, Peter Phillips, Alessio Sancetta, Philippe Vieu, the Associate Editor and two anonymous referees. We also thank Alexey Rudenko for providing an original Russian photocopy of Sytaya (1974), Hyungjin Lee for translating the paper, and the ERC for providing financial support.

†Statistical Laboratory, Faculty of Mathematics, University of Cambridge; syh30@cam.ac.uk

‡Faculty of Economics, University of Cambridge; obl20@cam.ac.uk

1 Introduction

Conditional expectations are crucially important in financial economics, with implications in many applications including asset returns predictability, market efficiency and risk management. One fundamental objective is to understand the risk/return trade-offsummarized by the relationship between the *expected excess return* relative to the *conditional variance* of returns. Due to the latency of conditional expectations however, there has been no universal agreement upon what is the best way to measure these objects. Differences in the approaches to modelling and estimating the conditional mean and variance has led to disagreement on their measurement, and also, conflicting empirical evidence on their intertemporal relation. Theoretical asset pricing models do not generally restrict the shape of the risk premium or the dynamics of the risk return trade-off. For example, Backus and Gregory (1993) show that the shape of the relation between the risk premium and the conditional variance of returns is largely unrestricted with increasing, decreasing, flat, or nonmonotonic patterns all possible. Similar conclusions are drawn by studies such as Abel (1988), Gennotte and Marsh (1993), and Veronesi (2000). Nevertheless, most empirical studies adopt a simple linear specification.

There are some issues with the usual approaches. First, there is the risk of misspecification. For instance, some studies have relied on parametric or semiparametric assumptions such as the ARCH or stochastic volatility models, where some high degree of structure is imposed on the return generating process. Other studies have typically measured the conditional mean and conditional variance as projections onto some predetermined variables. These approaches cannot be entirely justified, since they are all necessarily prone to some degree of potential specification error, see Linton and Perron (2003) and Escanciano, Pardo-Fernández and Van Keilegom (2017) for further discussions. Nonparametric modelling can be an effective solution in this context. It is a well established practical tool for analyzing time series data; see for example Härdle (1990), Bosq (1996), or Fan and Yao (2003) for a comprehensive review. A major advantage of this approach is that the relationship between the explanatory variables under study, denoted by $X = (X_1, \ldots, X_d)^\intercal$, and the response, say Y, can be modelled without assuming any restrictive parametric or linear structures. Stone (1980, 1982) showed that the best achievable convergence rate (in minimax sense) is $n^{-\beta/(2\beta+d)}$, where β is a measure of smoothness and d is the dimension of the covariates.

Secondly, there may be potential bias due to the omission of necessary information. Choosing among a few conditioning variables introduces an element of arbitrariness into the econometric modelling of expectations. In particular, if information that

investors consider important is neglected, then the corresponding estimates may be unreliable, Harvey (2001). Lettau and Ludvigson (2010) argued that contrasting conclusions on the intertemporal risk-return relation are largely due to the prevalent use of only small amount of conditional information in modelling the conditional mean and variance. Indeed, such practice greatly restricts the dynamics for the variance process and may result in poor estimates, especially when the volatility is highly persistent, Linton and Perron (2003), Giraitis et al. (2008). For example, Pagan and Hong (1990) estimated the conditional moments with nonparametric estimates of $\mathrm{E}(r_{mt} - r_{ft}|r_{m,t-1}, \ldots, r_{m,t-p})$ and $\mathrm{var}(r_{mt} - r_{ft}|r_{m,t-1}, \ldots, r_{m,t-p})$, where $r_{mt} - r_{ft}$ denotes the excess market return and $p = 1$ or 4. Having ended up with a negative risk-return relation using their estimates, they conjectured that the conclusion may have been affected by their use of only a small, finite number of conditioning variables. Noting the dependence of a GARCH process on the infinite past history of returns (with declining weights), they wrote: "*[A] nonparametric estimator of σ_t^2 appeals as a solution ..., although the fact that it operates with only a finite number of conditioning elements makes it unable to explicitly handle a GARCH type process. ... [O]ne might be able to establish consistency of the estimator [which deals with the infinite dependence]. As far as we are aware, however, there are no current theorems that would justify such a conjecture.*"

1.1 Overview of Results

This paper defines an estimation method that effectively addresses the aforementioned difficulties. We propose a Nadaraya-Watson type estimator that operates with an unrestricted number of conditioning variables. We derive large sample properties of the estimator in extensive detail, thereby providing an answer to the longstanding question in the quotation above. With a bandwidth sequence that shrinks the effects of long lags, the influence of all conditioning information is modelled in a natural and flexible way, and both issues of omitted information bias and specification error are effectively handled. It is worth noting that Harvey (2001) reported sensitivity of conditional expectations estimates on what type of conditioning variables are used in modelling the expectations. He showed with examples how several parametric/nonparametric estimates (and the estimated risk-return relationship) may vary according to the choice of different predetermined conditioning information. In this paper, we allow for various kinds of conditioning information. This is achieved by letting our model assumptions cover a wide range of static and dynamic regressions frameworks. The latter includes the autoregression framework as a special case.

Linton and Sancetta (2009) tackled this estimation problem of infinite order regression in the autoregression context. They established uniform almost sure consistency for stationary ergodic data but without rates. In the conclusion, they conjectured that the limiting distribution of nonparametric estimators could be established, and that the rate of convergence would be logarithmic. Under strict cross-sectional and temporal i.i.d. assumption, Mas (2012) derived a convergence rate that is consistent with our results in the particular case they considered.

We make several contributions. First, we establish some theorems which answer several open questions posed in the literature. Specifically, we show the pointwise consistency of our estimator under a set of mild regularity conditions. Further, we establish a central limit theorem for our estimator at a point under stronger conditions as well as for a feasibly studentized version of the estimator, thereby allowing pointwise inference to be conducted. Also, uniform consistency of the estimator is shown over a compact set of logarithmically increasing dimension. We prove that convergence rates depend on the smoothness of the regression function, the distribution of the marginal regressors and their dependence structure in a non-trivial way via the Lambert W function. We elaborate how each of those factors affects the rate of convergence, and show that the best possible rate is, nonetheless, of logarithmic order in all cases regardless of the smoothness of the regression function. This reflects the difficulty of capturing nonparametrically the effect of an infinite number of lags.

Second, using our estimation method we find some new empirical results. We reveal new evidence on the dynamics of risk-return relation and its link with the macroeconomy, and add supporting evidence for explaining some major puzzles in financial economics. To elaborate, applying our methods on the US stock market we find a positive risk-return relationship over the past 60 years overall − which is what asset pricing models generally postulate, e.g. Merton (1973). In particular, the relation turns out to be highly positive and strongly statistically significant in the recent 30 years period. Moreover, we also found that there has been a strong time variation and counter-cyclicity in risk aversion and in the conditional Sharpe ratio. The time series of estimated risk aversion tends to move in the opposite direction to the Federal Funds rate, a proxy for the business cycle, with the sample correlation being -0.5673. The quarterly Sharpe ratio is also strongly counter-cyclical, rising over most periods of recessions. By contrast, when a standard nonparametric method is employed instead, we noticed that these findings are not revealed, and different conclusions are reached. We believe our new empirical findings suggest an improvement in the econometric analysis that is attributable to allowing for extended flexibility and the inclusion of

otherwise neglected information in our method.

1.2 Technical challenges and sketch proposals for remedy

One major hurdle we face in the infinite-dimensional setting is the non-existence of the usual notion of density $p(\cdot)$ for the regressor X. Since there is no σ-finite Lebesgue measure in infinite-dimensional spaces, the Lebesgue density (with respect to the infinite product of probability measures) of the regressor cannot be defined via the Radon-Nikodym theorem. Consequently, standard asymptotic arguments for kernel estimators are no longer valid, for example: Bochner's lemma whereby under suitable regularity conditions, for $j = 1, 2$

$$\begin{aligned} \frac{1}{h^d}\mathrm{E}\left[\mathcal{K}^j\left(\frac{x-X}{h}\right)\right] &= \int \mathcal{K}^j(u)p(x-uh)\ du \\ &\to p(x)\|\mathcal{K}\|_j^j \quad \text{as} \quad h \to 0 \end{aligned} \tag{1}$$

where $\mathcal{K}$ is a multivariate kernel (see Section 2.2 below). So classical limiting theories cannot be readily extended to our setting.

We propose to adopt and apply some ideas from the functional regression literature. There is a vast statistical literature on functional data (typical examples include curves and images, which are infinite-dimensional in nature). Ferraty and Vieu (2002) first studied the case where the regressor was function-valued. Masry (2005) provided a rigorous treatment of nonparametric regression with dependent functional data in which X lies in a general semi-metric space, establishing the central limit theorem. Mas (2012) derived the minimax rate of convergence for nonparametric estimation of the regression function with strictly independent and identically distributed covariates. Ferraty and Vieu (2006) detailed a number of extensions and gave an overview of nonparametric approaches in the functional statistics literature. Geenens (2011) gave an up-to-date accessible summary of the literature on nonparametric functional regression, and introduced the term *curse of infinite dimensionality*, which reflects the evident difficulties in nonparametric estimation of infinite-dimensional objects due to extreme data sparsity. In the finite dimensional case more smoothness can mitigate completely the slower rate of convergence caused by dimensionality, but in the infinite dimensional case, additional smoothness can only mildly improve the convergence rate of estimators. We discuss in the next section the difference between the functional data framework and our discrete time framework.

There is another potential problem that may arise specifically in the infinite dimen-

sional setting. In the dynamic regression framework, the regressor vector X_t includes the infinite lags of a variable Z_t, say the response variable. Consequently, the class of mixing type assumptions, a popular notion of dependence in the econometrics literature, is generally not applicable. This is because measurable functions of X_t will depend upon the infinite time-lags of Z_t, and are not mixing in general, see e.g. Davidson (1994). Therefore, in order to establish asymptotic theories, an alternative set of dependence assumptions should be imposed on the data generating process. We defer further discussions to Section 2.1 below.

Lastly, for notations, we define $a_n \simeq b_n$ by $a_n = b_n + o(1)$, and $c_n \sim d_n$ by equivalence of order between the two sequences c_n and d_n. Also, $f \preceq g$ means there exists some constant $c > 0$ such that $\lim_{n\to\infty} f(n)/g(n) \leq c$. The term 'stationarity' is taken to mean strict stationarity. Throughout, C (or C', C'') refers to some generic constant that may take different values in different places unless defined otherwise.

2 Some Preliminaries

Consider the regression model

$$Y = m(X) + \varepsilon, \tag{2}$$

where the regressor $X = (X_1, X_2, \ldots)^\intercal$ is a random element taking values in some sequence space S, the response Y is a real-valued variable, and the stochastic error ε is such that $\mathrm{E}(\varepsilon|X) = 0$ a.s. The objective is to estimate the Borel function

$$m(\cdot) = \mathrm{E}(Y|X = \cdot) \tag{3}$$

based on n random samples observed from a strictly stationary data generating process $\{(Y_t, X_t) \in \mathbb{R} \times S\}_{t\in\mathbb{Z}}$ having some weak dependence structure. Details on the assumptions are given in Section 2.1 below.

This setting is related to the usual framework adopted for functional data, which has been widely studied by statisticians, see Ramsey and Silverman (2002), Aneiros, Bongiorno, Cao and Vieu (2017). Recently, successful attempts have been made to develop theories for nonparametric inference in the functional statistics literature; Ferraty and Romain (2010) gives a comprehensive review. A major issue in this field of research lies in extending the statistical theories applicable to $\mathbb{R}^d$ to function spaces. In this literature, attention is usually on smooth functions that are approximated and reconstructed from finely discretised grids on some compact interval. In contrast, the setup in our model (2) can be viewed as looking at a countable number of discrete

observations. Such a difference is reflected by the fact that the observed data is taken to be a discrete process $X = (X_s)$ with unbounded $s \in \mathbb{Z}^+$ so that $S = \{f | f : \mathbb{N} \to \mathbb{R}\}$, rather than $X = (X(s))$ with $s \in [0,T]^k$ so that $S = \{f | f : [0,T]^k \subset \mathbb{R}^k \to \mathbb{R}\}$, e.g. curves if $k = 1$, images if $k \geq 2$. The discrete nature of our setting has several fundamental distinctive features that allow us to look further into many specific practical issues.

An immediate consequence of our framework is that the tuning parameter can be imposed on each and every dimension, allowing one to control the marginal influence of the regressors. For instance when it is sensible to postulate that the influence of distant covariates is getting monotonically downweighted, one may set the marginal bandwidths to increase in the lag horizon so as to impose higher amount of smoothing at distant lags. Depending on the nature of the regressor, S may be taken as the space of all infinite real sequences $\mathbb{R}^\infty := \prod_{j=1}^\infty \mathbb{R}_j$ formed by taking Cartesian products of the reals, or its various linear subspaces such as ℓ_∞, ℓ_p, c. We propose to take $S = \mathbb{R}^\infty$ so as to refrain from imposing any prior restrictions with regard to the choice of the regressor; for example, taking S to be the space of bounded sequences excludes the possibility of having regressors with infinite support (e.g. Gaussian process).

2.1 Dependence structure and leading examples

A distinctive characteristic of time series data is temporal dependence between observations. In the nonparametric time series literature, Rosenblatt (1956)'s α-mixing has been the *de facto* standard choice due to it being the weakest among the class of mixing-type asymptotic independence conditions. Roussas (1990) established pointwise and uniform consistency of the local constant estimator under this condition, respectively, while Fan and Masry (1992) established asymptotic normality. The α-mixing condition has also been widely used in the context of dependent functional observations, see for instance Ferraty et al. (2010), Masry (2005), and Delsol (2009).

DEFINITION 1. *A stochastic process $\{Z_t\}_{t=1}^\infty$ defined on some probability space $(\Omega, \mathcal{F}, P)$ is called α-mixing (NB. 'jointly' α-mixing if Z_t is $\mathbb{R}^d$-valued, with $d \in (1, \infty]$), if*

$$\alpha(r) := \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{r+t}^\infty} |P(A \cap B) - P(A)P(B)|$$

is asymptotically zero as $r \to \infty$, where $\mathcal{F}_a^b$ is the σ-algebra generated by $\{Z_s; a \leq s \leq b\}$. In particular, we say the process is algebraically (respectively exponentially) α-mixing with rate k if there exists some $c, k > 0$ such that $\alpha(r) \leq cr^{-k}$ (respectively

if there exists some $\gamma, \varsigma > 0$ *such that* $\alpha(r) \leq \exp(-\varsigma r^{\gamma})$).

The popularity of the α-mixing condition (note the modifier α- will occasionally be omitted if no confusion is likely) in the literature stems from the fact that it is easy to work with, see e.g. Doukhan (1994), Rio (2000) for a comprehensive survey. However, there are several limitations that have been pointed out in the literature. First, it is a rather strong technical condition that is hard to verify in practice. Second, some basic processes are not mixing. e.g. AR(1) with Bernoulli innovations, Andrews (1984).

We turn to our setting. In the static regression case it is appropriate to assume the mixing condition, but in the dynamic case this condition is not generally applicable as we now explain. Recall that the object of estimation is the conditional mean $E(Y_t|\mathcal{F})$, cf. (2), where the information set $\mathcal{F}$ is determined by the nature of the conditioning variables. There are two leading cases: the first case is the static regression where the information set is taken to mean $\sigma(X_{jt}; j = 1, 2, \ldots)$, the σ-algebra generated by the exogenous marginal regressors. The second case is the autoregression, where $X_{tj} = Y_{t-j}$ for all j, in which case $\mathcal{F} = \mathcal{F}_{t-1}$ represents $\sigma(Y_s; s \leq t-1)$, the σ-algebra generated by the sequence of lags of the response $(Y_s)_{s \leq t-1}$. In fact, as for the latter framework we may consider a more general setup, i.e. a dynamic regression, where the information set is taken to be $\mathcal{F} = \sigma(X_{js}, Y_s;\ s \leq t-1)$ for some j. Details are formally given in Assumptions A below.

In the static regression case the usual joint α-mixing condition can be assumed on the sample data $\{Y_t, X_t\}$ as is usually done; since marginal regressors are observed at the same time t: $X_t = (X_{1t}, X_{2t}, \ldots)^{\intercal}$, assuming joint dependence does not require additional adjustments. Indeed, it can be easily shown that joint mixing implies both marginal component processes and any measurable function thereof are mixing.[1] In this paper, we do not necessarily require independence between component processes $\{X_{jt}\}$, $j = 1, 2, \ldots$; later we specify to what extent some dependence can be allowed (see Assumption C). It will turn out that the requirement is mild and allows sufficient generality in application.

Moving on to the dynamic regression setting, since the regressors are taken to be the lags of the response and/or a covariate, measurable functions of X_t depend on infinite time-lags and hence are *not* necessarily mixing.[2] Therefore an alternative set of dependence conditions is necessary to establish asymptotic theories for the second

[1]The converse is not necessarily true unless the marginal processes are independent to each other, see Bradley (2005, Section 5).

[2]Except for some very special cases; Davidson (1994, Theorem 14.9) gives a set of technical conditions under which a process with infinite (linear) temporal dependence is α-mixing.

framework. We adopt the notion of near epoch dependence due to Ibragimov (1962) for the dynamic regression setting and deal with two leading cases separately.

DEFINITION 2. *A stochastic process* $\{Z_t\}_{t=1}^{\infty}$ *defined on some probability space* $(\Omega, \mathcal{F}, P)$ *is called near-epoch dependent or stable in* L_2 *with respect to a strictly stationary* α*-mixing process* $\{\eta_t\}$ *if the stability coefficients* $v_2(r) := E|Z_t - Z_{t,(r)}|^2$ *is asymptotically zero as* $r \to \infty$*, where* $Z_{t,(r)} = \Psi_r(\eta_t, \ldots, \eta_{t-r+1})$ for some Borel function $\Psi_r : \mathbb{R}^r \to \mathbb{R}$.

A process that is *near epoch dependent* on a mixing sequence is influenced primarily by the "recent past" of the sequence and hence asymptotically resembles its dependence structure; see e.g. Billingsley (1968), Davidson (1994), or Lu (2001) for details. Andrews (1995) established uniform consistency of kernel regression estimators under near epoch dependence conditions. Following the usual convention, e.g. Bierens (1983), we shall take $\Psi_r(\eta_t, \ldots, \eta_{t-r+1}) \equiv \mathrm{E}(Z_t|\eta_t, \ldots, \eta_{t-r+1})$. In Section 2.3 it will be shown that under suitable conditions similar asymptotic theories can be derived for both static and dynamic regression frameworks.

2.2 Local Weighting

In this section we fix the notions of local weighting and the measure of closeness between the data objects. Let $K : [0, \infty) \to [0, \infty) =: \mathbb{R}_+$ be a univariate density function and for an element u of a normed sequence space, let

$$\mathcal{K}(u) := K(\|u\|). \tag{4}$$

In our setting the properties of K are crucially important. We now group the kernel functions into three subcategories depending on how they are generated. The first two, referred to as Type-I and Type-II kernels in Ferraty and Vieu (2006) generalize the usual 'window' kernels and monotonically decreasing kernels in finite dimension, respectively. Both types of kernels are continuous on a compact support $[0, \lambda]$.

DEFINITION 3. *A function* $K : [0, \infty) \to [0, \infty)$ *is called a kernel of type*$-I$ *if it integrates to 1, and if there exist real constants* C_1, C_2 *(with* $0 < C_1 < C_2$*) for which*

$$C_1 1_{[0,\lambda]}(u) \leq K(u) \leq C_2 1_{[0,\lambda]}(u), \tag{5}$$

where λ *is some fixed positive real number. A function* $K : [0, \infty) \to [0, \infty)$ *is called*

a kernel of type$-II$ if it satisfies (5) with $C_1 \equiv 0$, and is continuous on $[0, \lambda]$ and differentiable on $(0, \lambda)$ with the derivative K' that satisfies

$$C_3 \leq K'(u) \leq C_4$$

for some real constants C_3, C_4 such that $-\infty < C_3 < C_4 < 0$.

The definition above suggests that the uniform kernel on $[0, \lambda]$ is a type-I kernel, and the Epanechnikov, Biweight and Bartlett kernels belong to the class of Type-II kernels. Some of those with semi-infinite support, for example (one-sided) Gaussian, are covered by the last group, which we will call the Type-III kernels.

DEFINITION 4. *A function $K : [0, \infty) \to [0, \infty)$ is a kernel of type$-III$ if it integrates to 1, and if it is of exponential type; that is, $K(r) \propto \exp(Cr^{\beta})$ for some β and C.*

2.3 Small deviations

The *small ball (or small deviation) probability* plays a crucial role in establishing the asymptotic theory. Let S^* be a sequence space equipped with some norm $\|.\|$; then the small ball probability of an S^*-valued random element Z is a function defined as

$$\varphi_z(h) := P\left(\|z - Z\| \leq h\right), \tag{6}$$

where $h \in \mathbb{R}_+$. The probability is called *centered* if $z = 0$ (in which case we write $\varphi(h)$) and *shifted* (with respect to some fixed point $z \in S^* \backslash \{0\}$) if otherwise. The relation between the two quantities cannot be explicitly specified in general, and will be given in terms of a Radon-Nikodym derivative (See Assumption D2 below).

The name *small ball* stems from the fact that we are interested in the asymptotic behaviour of $\varphi_z(h)$ as h tends to zero. The function can be thought of as a measure for how much the observations are densely *packed* or *concentrated* around the fixed point z with respect to the associated norm and the reference distance h. From the definition it is straightforward to see that $\varphi_z(h) \to 0$ as $h \to 0$, and that $n\varphi_z(h)$ is an approximate count of the number of observations whose influence is taken into account in the smoothing procedure. When Z is a continuous random vector of fixed dimension d with density $p(\cdot) > 0$, it can be readily shown that the shifted small ball probability (with respect to the usual Euclidean norm) is given by

$$\varphi_z(h) = V_d h^d p(z) = O(h^d), \tag{7}$$

where $V_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of the d-dimensional unit sphere.

However, when Z takes values in an infinite-dimensional normed space, it is difficult to specify the exact form of the small ball probability, and its behaviour varies depending heavily on the nature of the associated space and its topological structure. Due to the non-equivalence of norms in infinite dimensional spaces, it is intuitively clear that the "speed" at which $\varphi_z(h)$ converges to zero is affected by the choice of the norm $\|.\|$. Nonetheless, a rapid decay is expected in general irrespective of the choice of the norm due to the extreme sparsity of data in infinite-dimensional spaces.

One possible example of S^* is $(\ell_r, \|.\|_r)$, the space of r-th power summable sequence equipped with the ℓ_r-norm; the centred small ball behaviour of sums of weighted i.i.d. random variables is widely studied in the literature, see for example Borovkov and Ruzankin (2008) and references therein. In this work here, we will focus our main attention on the case of $r = 2$ (and take $\|.\|$ to mean $\|.\|_2$ unless specified otherwise). Nevertheless, we note that the results derived in this paper can be extended to the case of $r > 2$ as long as the regularity conditions are adjusted appropriately.

Writing the expected value of the kernel in terms of the small ball probability

$$\mathrm{E}\mathcal{K}\left(\frac{z-Z}{h}\right) = \mathrm{E}K\left(\frac{\|z-Z\|}{h}\right) = \int K(u)\ dP_{\|z-Z\|/h}(u) = \int K(u)\ d\varphi_z(uh), \quad (8)$$

we are able to bypass the difficulties mentioned in the introduction, and to establish the convergence of the integrals without requiring the existence of the Lebesgue density.

Lemma 1 Ferraty and Vieu (2006, Lemma 4.3 & 4.4). *Suppose $\|.\|$ is some semi-norm defined on a function space. If K is type-I, then it satisfies*

$$C_1^j \leq \frac{1}{\varphi_z(h\lambda)} \int_0^\lambda K^j(v)\ d\varphi_z(vh) \leq C_2^j, \quad j = 1, 2 \quad (9)$$

where $C_1, C_2 > 0$ are as defined in Definition 3. When the kernel K is type-II, if

$$\exists\ \varepsilon_0 > 0, C_5 > 0\ \ s.t.\ \forall \varepsilon < \varepsilon_0, \quad \int_0^\varepsilon \varphi_x(u)du > C_5 \varepsilon \varphi_x(\varepsilon) \quad (10)$$

then we have

$$C_6^j \leq \frac{1}{\varphi_z(h\lambda)} \int_0^\lambda K^j(v)\ d\varphi_z(vh) \leq C_7^j, \quad j = 1, 2 \quad (11)$$

where the constants $C_6 = -C_5 C_4$ and $C_7 = \sup_{s\in[0,\lambda]} K(s)$ are strictly positive.

Under the regularity conditions of Lemma 1, (9) and (11) hold for every $h > 0$, so it follows that for any kernels of type-I and II:

Corollary 1 *If the kernel K is either type-I or type-II, then for $j = 1, 2$ we have*

$$\frac{1}{\varphi_z(h\lambda)}\mathrm{E}\left[\mathcal{K}^j\left(\frac{z-Z}{h}\right)\right] \longrightarrow \xi_j \quad as \quad h \to 0^+, \tag{12}$$

where ξ_1 and ξ_2 are some strictly positive real constants.

This result can be seen as an infinite-dimensional analogue of Bochner's lemma (1): i.e., for $Z \in \mathbb{R}^d$, $h^{-d}\mathrm{E}\mathcal{K}((z-Z)/h) \to p(z) > 0$. It is obvious that ξ_j is bounded below and above by C_1^j and C_2^j, respectively (or C_6^j and C_7^j, depending on the choice of the kernel). With specific choices of kernels and regressors we may be able to specify the exact values of the constants in some certain cases. For example, it is straightforward to see that $\xi_1 = 1/\lambda$ and $\xi_2 = 1/\lambda^2$ when K is uniform kernel supported on $[0, \lambda]$.

Remarks. (i) Lemma 1 reveals the importance of condition (10) in constructing the asymptotics when the kernel is of type-II. Whereas the condition is widely assumed in the functional statistics literature for that reason, Azais and Fort (2013) proved that it necessarily restricts the variable Z to be of finite dimension. In other words, whenever (10) is valid, the topology that governs the concentration properties of Z accounts effectively only for finite dimension. An example includes the case where Z is associated with the semi-norm $\|y\| := (y_1, \ldots, y_p, 0, 0, \ldots)$ for some fixed positive integer $p < \infty$ and $y \in \mathbb{R}^\infty$, Ferraty and Vieu (2006, Section 13.3.3). Since this severely restricts the applicability of our work, we shall not consider the case of Type-II kernels.

(ii) A natural question one may then ask is whether (12) would hold for kernels with semi-infinite support such as the Type-III kernels. In the finite $\mathbb{R}^d$-framework, it is well known that a set of assumptions including $\|u\|^d K(u) \to 0$ as $u \to \infty$ is sufficient for showing (1), see for instance Parzen (1962, Theorem 1A) and Pagan and Ullah (1999, Lemma 1). However, in the infinite-dimensional setting the answer is negative in most usual cases where the kernel is of exponential type (e.g. Gaussian kernel). Whereas the lower bound of the limit can be easily constructed via Chebyshev's inequality: with reference to Definition 4, writing $V = \|z - Z\|^\beta$, $\delta = h^\beta$ and letting c_δ be some function of δ we have

$$(0 <) \exp(-c_\delta\delta) \leq [P(V \leq \delta)]^{-1}\mathrm{E}\exp(-c_\delta V). \tag{13}$$

So the upper bound may not exist, and the rate at which the small ball probability

decays to zero may dominate the speed at which the integral (8) converges to zero. This claim cannot be formally verified for all general cases because (as aforementioned) there is no unified result for the asymptotic behaviour of small deviations available. Nevertheless, the idea can be sketched in the common case where the asymptotics of the distribution function (i.e. small deviation) is of exponential order: $P(V \leq \delta) \sim \exp(-C\delta^{-\theta})$ as $\delta \to 0$ for some constants C and $\theta > 0$. By de Bruijn's exponential Tauberian theorem (see Bingham et al. (1987), Li (2012)), a necessary and sufficient condition for such a case is the following limiting behaviour of the Laplace transform near infinity:

$$\mathrm{E}[\exp(-c_\delta V)] \sim \exp\Big(- C' \cdot c_\delta^{\theta/(1+\theta)} \Big) \quad \text{as} \quad c_\delta \to \infty$$

for some constant $C' > 0$. With $V = \|z - Z\|^2$, $\delta = h^2$, $c_\delta = 2^{-1}h^{-2}$ (which corresponds to the case of the Gaussian kernel) the difference in the order of convergence suggests that the right hand side of (13) is unbounded, and that the limit (12) diverges.

Due to the reasons above we shall confine our attention to Type-I kernels only here in this work.

2.4 Bandwidth Matrix and covariates

We aim to estimate the regression operator at a point $x \in \mathbb{R}^\infty$ with an $\mathbb{R}^\infty$-dimensional regressor $X = (X_1, X_2, \ldots)^\intercal$. Let $H := \mathrm{diag}(\underline{h}) = \mathrm{diag}(h_1, h_2, \ldots) \in \mathbb{R}^{\infty\times\infty}$ be the bandwidth matrix. We require that a norm $\|.\|$ can be admitted to the *weighted regressor* values and the *weighted point*, and for this the bandwidth sequence must be chosen appropriately. In particular, we let

$$H = hD = h \times \mathrm{diag}(\phi_1, \phi_2, \ldots), \tag{14}$$

where $D \in \mathbb{R}^{\infty\times\infty}$ and $h \in \mathbb{R}$. By Kolmogorov's three-series theorem, the sequence of weighted regressors $\{\phi_j^{-1} X_j\}$ is square summable, with probability one, provided that the marginal regressors X_j' are independent with finite variance and satisfy

$$\sum_{j=0}^{\infty} \mathrm{E} \min\left\{1, \phi_j^{-2} X_j^2\right\} < \infty, \tag{15}$$

so that $(\phi_1^{-1}X_1, \phi_2^{-1}X_2, \ldots)^\intercal =: Z$ is $(\ell_2, \|.\|_2)$-valued. In the autoregressive framework, ϕ_j can be interpreted as a weight sequence that represents the "relative influence" of the marginal regressors, which diminishes as lags get further apart.

For this purpose we assume from now on that *the bandwidth-weighted* X and x (i.e. Z and $z := (\phi_1^{-1}x_1, \phi_2^{-1}x_2, \ldots)^\intercal$, respectively) are ℓ_2-valued[3] and normed with $\|.\| = \|.\|_2$. Consequently, (with an abuse of notation) we can extend the usual definition of shifted small deviation to account for the generalized support $[0, \lambda]$ and bandwidth vector $\underline{h} = (h_1, h_2, \ldots)^\intercal$:

$$\begin{aligned}\varphi_x(\underline{h}\lambda) &:= P\left(\|H^{-1}(x - X_t)\| \leq \lambda\right) \\ &= P\left(\|D^{-1}(x - X_t)\| \leq h\lambda\right). \end{aligned} \tag{16}$$

Equivalently, $\varphi_x(\underline{h}\lambda) = P(X_t \in \mathcal{E}(x, \underline{h}\lambda))$, where $\mathcal{E}$ is the infinite-dimensional hyperellipsoid centred at $x \in \mathbb{R}^\infty$, and λ is as defined in Section 2.2. Clearly, $\varphi_x(\underline{h}\lambda) = \varphi_z(h\lambda)$. For later reference, we also define the joint small ball probability of the regressor vectors observed at different times t and s as the joint distribution

$$\psi_x(\underline{h}\lambda; t, s) := \ P\big((X_t, X_s) \in \mathcal{E}(x, \lambda\underline{h}) \times \mathcal{E}(x, \lambda\underline{h})\big). \tag{17}$$

3 The Estimator

We observe a sample $\{Y_t, X_t\}_{t=1}^n$ with $Y_t \in \mathbb{R}$ and $X_t \in \mathbb{R}^\infty$. With these data, we propose to estimate $m(x) = \mathrm{E}(Y|X = x)$, $x \in \mathbb{R}^\infty$ with the following local constant type estimator:

$$\widehat{m}(x) := \frac{\sum_{t=1}^n \mathcal{K}\Big(H^{-1}(x - X_t)\Big)Y_t}{\sum_{t=1}^n \mathcal{K}\Big(H^{-1}(x - X_t)\Big)} \equiv \frac{\sum_{t=1}^n K\Big(\|H^{-1}(x - X_t)\|\Big)Y_t}{\sum_{t=1}^n K\Big(\|H^{-1}(x - X_t)\|\Big)}. \tag{18}$$

In practice, in the autoregression case we essentially observe only $\{Y_1, Y_2, \ldots, Y_n\}$ rather than the full infinity, so further lags can be regarded as zeros. Similarly, in the static case, when X_t is in $\mathbb{R}^\tau$ for large τ we can identify this with $X_t = (X_{1t}, X_{2t}, \ldots, X_{\tau t}, 0, 0, \ldots) \in \mathbb{R}^\infty$. So for practical applications, one may for example employ a truncation argument on the regressor (as will be done in Section 4.4 - albeit with a different purpose) and let the effective dimension τ of the regressor X_t to increase in n in the theoretical analysis.

The estimator can be viewed as an infinite-dimensional generalization of the standard multivariate local linear estimator, and is a special case of the one in Ferraty and Vieu (2002), Masry (2005) and references therein for functional data. In the following

[3]This gives a mild restriction on the range of possible points at which the estimation is made; i.e. $x \in \mathbb{R}^\infty$ is such that $\sum_j j^{-2p}x_j^2 < \infty$.

section we will examine some asymptotic properties of the estimator.

4 Asymptotic Properties

In this section we introduce the main results of this paper. We derive some large sample asymptotics of the proposed estimator (18). We establish consistency in both pointwise and uniform sense, and also the asymptotic normality. All proofs are given in the appendix.

Consider two different cases: (1) the static regression and (2) the dynamic regression. Below we specify two sets of temporal dependence conditions, either of which will be assumed on the data generating process of the sample observations. Assumption A1 corresponds to the static regression case where we have exogenous regressors that are jointly observed in time in a weakly dependent manner. No restriction is needed as regards the dependence structure between the marginal regressors, although certain additional conditions can be potentially imposed at the later stage (see Assumptions C below). The second option A2 concerns with the dynamic regression framework. In this case, the notion of near epoch dependence is adopted to describe the dependence structure of the processes defined as functions of the response variables. The assumptions below suggest that there is a trade-off between the degree of mixing and the possible order of moments, we allow on the response variable, i.e. $2+\delta$.

Assumptions A

A1. *The marginal regressors $X_{1t}, X_{2t}, \ldots$ are exogenous variables, and the sample data $\{Y_t, X_t\}_{t=1}^n = \{Y_t, (X_{1t}, X_{2t}, \ldots)\}_{t=1}^n$ is stationary and jointly arithmetically α-mixing with rate $k \geq 2(\delta+2)/\delta$, where δ is as defined in Assumption B4 below.*

A2. *Each regressor is either a lag of the response variable Y_t or of a covariate V_t, i.e. $X_{jt} = Y_{t-j}$ or $X_{jt} = V_{t-j}$, $j \in \mathbb{N}$, and $\{Y_t, V_t\}_{t=1}^n$ is stationary and arithmetically α-mixing with rate $k \geq 2(\delta+2)/\delta$. Also, the process $K_t := K(\|H^{-1}(x - X_t)\|)$ is near epoch dependent on (Y_t, V_t), and there exists some $r = r_n \to \infty$ such that the rate of stability for K_t denoted $v_2(r_n) = v_2(r)$ satisfies*

$$v_2(r)^{1/2}[\varphi_x(\underline{h}\lambda)]^{-(2\delta+3)/(2\delta+2)} n^{1/(2(\delta+1))} \to 0 \qquad as \;\; n \to \infty. \tag{19}$$

Remark. Our model under Assumption A2 can be viewed as a generalization of the NAARX model in Chen and Tsay (1993). The framework nests both the fully

autoregressive framework in which $X_{jt} = Y_{t-j}$ for all j, and the case where the regressor vector consists only of the lags of a covariate V_t. Doukhan and Wintenberger (2008) studied the autoregressive model of order $d = \infty$ under a notion of weak dependence, and showed the existence of a stationary solution. This result was further studied in Wu (2011).

4.1 Pointwise consistency

Pointwise consistency of the local constant estimator was first studied by Watson (1964) and Nadaraya (1964) for i.i.d data with $d = 1$. Their result was extended to the multivariate case (finite d) by Greblicki and Krzyzak (1980) and Devroye (1981). Robinson (1983) and Bierens (1983) were amongst the earliest papers that worked on consistency of the estimator with dependent observations (both static regression and autoregression were allowed in their frameworks), followed by Roussas (1989), Fan (1990), and Phillips and Park (1998) to name a few out of numerous papers. The case of the functional regressor was first studied by Ferraty and Vieu (2002).

In this section we establish the pointwise weak consistency of the estimator (18) with dependent data satisfying either A1 or A2. A set of assumptions required for the theory is now introduced, and some introductory arguments are briefly sketched.

Assumptions B

B1. *The regression operator* $m : \mathbb{R}^\infty \to \mathbb{R}$ *is continuous in some neighbourhood of* x

B2. *The marginal bandwidths satisfy* $h_j = h_{j,n} \to 0$ *as* $n \to \infty$ *for all* $j = 1, 2, \ldots$, *where* $\operatorname{diag}(h_1, h_2, \ldots) = \operatorname{diag}(\underline{h}) = H$ *is the bandwidth matrix, and the small ball probability obeys* $n\varphi_x(\underline{h}\lambda) \to \infty$ *for every point* $x \in \mathbb{R}^\infty$, *where* $\varphi_x(\underline{h}\lambda) := P(\|H^{-1}(x - X)\| \leq \lambda) \to 0$ *as* $n \to \infty$.

B3. *The kernel* K *is type*$-I$

B4. *The response* Y_t *satisfies* $E\big(|Y_t|^{2+\delta}\big) \leq C < \infty$ for some $C, \delta > 0$.

B5. *The joint small ball probability (17) satisfies* $\psi_x(\underline{h}\lambda; i, j) \leq C\varphi_x(\lambda\underline{h})^2$, $\forall i \neq j$.

B6. *The conditional expectation* $E\big(|Y_tY_s||X_t, X_s\big) \leq C < \infty$ *for all* t, s.

Remark. The continuity assumption B1 is necessary for asymptotic unbiasedness of the estimator. It will be shown that the estimator is unbiased at every point of continuity, and that the rate of convergence for the bias term can be specified upon

imposing further smoothness condition on the regression operator, see later. Assumption B2 can be thought of as an extension of the usual bandwidth conditions that are assumed in finite-dimensional nonparametric literature, cf. (7). As discussed before, $n\varphi_x(\underline{h}\lambda)$ can be understood as an approximate number of observations that are "close enough" to x. Therefore, it is sensible to postulate that $n\varphi_x(\underline{h}\lambda) \to \infty$ as $n \to \infty$, meaning that the point x is visited many times by the sample of data as the size of the sample grows to infinity. This is in line with the usual assumption that $nh^d \to \infty$ when $X \in \mathbb{R}^d$, in which case the small ball probability is given by $\varphi_x(h) \propto h^d p_X(x)$ as noted in (7). Conditions B5 and B6 are imposed to control the asymptotics of the covariance terms. The validity of condition B5 can be easily seen in the $\mathbb{R}^d$ frameworks; for relevant discussions, see Ferraty and Vieu (2006, Remark 11.2).

To sketch the idea, we write $K_t := K(\|H^{-1}(x - X_t)\|)$ for the sake of simplicity of presentation (note its dependence upon X_t), and express the estimator (18) as

$$\widehat{m}(x) := \frac{\sum_{t=1}^n K\big(\|H^{-1}(x - X_t)\|\big)Y_t}{\sum_{t=1}^n K\big(\|H^{-1}(x - X_t)\|\big)} = \frac{\frac{1}{n}\sum_{t=1}^n \frac{K_t}{EK_1}Y_t}{\frac{1}{n}\sum_{i=1}^n \frac{K_t}{EK_1}} = \frac{\widehat{m}_2(x)}{\widehat{m}_1(x)}. \tag{20}$$

We then employ the following decomposition:

$$\begin{aligned}\widehat{m}(x) - m(x) &= \frac{\widehat{m}_2(x)}{\widehat{m}_1(x)} - m(x) = \frac{\widehat{m}_2(x) - m(x)\widehat{m}_1(x)}{\widehat{m}_1(x)} \\ &= \frac{\mathrm{E}\widehat{m}_2(x) - m(x)\mathrm{E}\widehat{m}_1(x)}{\widehat{m}_1(x)} + \frac{[\widehat{m}_2(x) - \mathrm{E}\widehat{m}_2(x)] - m(x)[\widehat{m}_1(x) - \mathrm{E}\widehat{m}_1(x)]}{\widehat{m}_1(x)}, \end{aligned} \tag{21}$$

where clearly $\mathrm{E}\widehat{m}_1(x) = 1$. Below we show consistency by proving that the 'bias part' $\mathrm{E}\widehat{m}_2(x) - m(x)$ and the 'variance part' $[\widehat{m}_2(x) - \mathrm{E}\widehat{m}_2(x)] - m(x)[\widehat{m}_1(x) - 1]$ are both negligible in large samples. As for the latter term, it suffices to show the mean squared convergence of $\widehat{m}_2(x) - \mathrm{E}\widehat{m}_2(x)$ to zero because $\widehat{m}_1(x) \to^P 1$ then readily follows.

Theorem 1 *Suppose that Assumptions B1-B5 hold. Then the estimator (18) with sample observations $\{Y_t, X_t\}_{t=1}^n$ satisfying either A1 or A2 is weakly consistent for the regression operator $m(x) = E(Y|X = x)$. That is, as $n \to \infty$*

$$\widehat{m}(x) \xrightarrow{P} m(x). \tag{22}$$

In the following section, we present the rates of convergence and asymptotic normality under additional regularity conditions.

4.2 Asymptotic Normality

Earlier studies on the limiting distribution of the standard Nadaraya-Watson estimator can be traced back to Schuster (1972) and Bierens (1987), where the case of univariate and multivariate regressors was considered, respectively. The case of dependent samples was studied in Robinson (1983), Bierens (1983), Masry and Fan (1997), and by many others under various model setups and different regularity conditions. Masry (2005, Theorem 4) and Delsol (2009) established general distribution theories for Nadaraya-Watson type estimators in a semi-metric space. Our results are different from them in two respects. First, the difference of our framework from the functional literature discussed in the beginning of Section 2.2 gives us further flexibility, without which the analysis cannot be done to meet our specific needs. Second, whereas the final results of many existing papers were given in terms of abstract functions, our results are presented with an explicit rate of convergence, allowing practical applications. In this section we outline the main theory and introduce some interesting consequences thereof.

Our objective is to construct the asymptotic distribution of our estimator. That is, to find deterministic sequences $\mathcal{V}_n(x), \mathcal{B}_n(x)$ such that

$$\mathcal{V}_n^{-1/2}(x)\Big(\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\Big) \Longrightarrow N(0,1)\,. \tag{23}$$

In fact, under certain conditions[4] we can show that the following self-normalized limiting distribution holds

$$\Delta_n^{-1}(x)\Big(\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\Big) \Longrightarrow N(0,1)\,, \tag{24}$$

where $\Delta_n^2(x) := \sum_{t=1}^n \left(\sum_{s=1}^n K_s\right)^{-2} [K_t\big(Y_t - \widehat{m}(x)\big)]^2$, and as defined previously, $K_t = K(\|H^{-1}(x - X_t)\|)$. The proof of (24) is given within the proof for Theorem 2 later. This gives pointwise confidence intervals for $\widehat{m}(x)$, which can be used as a basis for conducting standard statistical inference.

We now discuss some main assumptions needed for our distribution theory.

4.2.1 The bias and variance components

Regarding the asymptotic 'bias', we need to strengthen Assumptions B by imposing additional smoothness conditions and suitable bandwidth adjustments, just like

[4] Assumptions B7-B10 in section 4.2.1 below

the continuity assumption is extended for asymptotic normality in standard $\mathbb{R}^d$ cases. These allow the exact upper bound of the asymptotic bias to be specified. Note that alternatively, a Fréchet differentiability-type condition may be imposed. Here also we introduce two conditions (Assumptions B9, B10) that we require for studying the variance component.

Further Assumptions B

B7. *The regression operator* $m : \mathbb{R}^\infty \to \mathbb{R}$ *satisfies*

$$\left|m(x) - m(x')\right| \leq \sum_{j=1}^{\infty} c_j \left|x_j - x'_j\right|^\beta \tag{25}$$

for every $x, x' \in \mathbb{R}^\infty$*, and some constant* $\beta \in (0, 1]$*, where* $\{c_j\}$ *is some sequence of real constants that satisfies* $\sum_{j=1}^{\infty} c_j \leq 1$.

B8. *The marginal bandwidths satisfy* $h_j = \phi_j \cdot h$ *for some positive real numbers* ϕ_j*, where* $h = h_n \to 0$ *as* $n \to \infty$*. We suppose that* ϕ_j *satisfy* $\sum_{j=1}^{\infty} \phi_j^{-2} < \infty$ *and* $\sum_{j=1}^{\infty} c_j \phi_j^\beta < \infty$.

B9. *The conditional variance* $\mathrm{var}[Y_t | X_t = u] = \sigma^2(u)$ *is continuous in some neighbourhood of* x*; i.e.* $\sup_{u \in \mathcal{E}(x, \underline{h}\lambda)} [\sigma^2(u) - \sigma^2(x)] = o(1)$*. Similarly, the cross-conditional moment* $E[(Y_t - m(x))(Y_s - m(x) | X_t = u, X_s = v] = \sigma(u, v)$*,* $t \neq s$ *is continuous in some neighbourhood of* (x, x).

B10. $R_{nt} := (EK_1)^{-1}\{K_t(Y_t - m(x)) - EK_t(Y_t - m(x))\}$ *belongs to the domain of attraction of a normal distribution.*

Remark. The first two assumptions concern with the bias component $\mathcal{B}_n$ in (23). Assumption B8 extends the previous bandwidth condition B2. Obviously, it is consistent with (and stronger than) what was previously assumed in B2, since $h \to 0$ implies the coordinate-wise convergence of each marginal bandwidths. With B8 one can write the asymptotic bias and the order of the bias-variance balancing bandwidth in terms of the common factor h. It is possible to dispense with this condition at the cost of imposing minor modifications in B7; the asymptotic bias will then be written in terms of the infinite sum of a weighted marginal bandwidth h_j, whose convergence needs to be ensured. A further increment condition will be needed on ϕ_j later to elaborate the asymptotics of the variance term.

Assumption B7 replaces and strengthens B1, and can be thought of as a variant of Hölder-type continuity; the case of $c_j = 2^{-j}$ and $\beta = 1$ is implied by the Lipschitz condition. Another example of c_j includes $\exp(-j)$. Indeed, under B7 the regression operator becomes a contraction mapping, and the contribution from each marginal dimension decreases in j. This ensures summability of the bias and allows the order of convergence rate to be specified, cf. (28) below.

Note that in the context of autoregression where $X_j \equiv Y_{t-j}$ for all j, the model is given by

$$Y_t = m(Y_{t-1}, Y_{t-2}, \ldots) + \varepsilon_t. \tag{26}$$

Whether the stationary solution $\{Y_t\}$ indeed exists is an important question. In the study of a class of general nonlinear AR(d) models, Duflo (1997) and Götze and Hipp (1994) assumed what is called the Lipschitz mixing condition (or the strong contraction condition), which is essentially (25) replaced by finite d-sum on the right hand side. In our context, Assumption B7 plays a similar role; Doukhan and Wintenberger (2008) showed that (25) with $\sum_{j=1}^{\infty} c_j < 1$, is sufficient for the existence of a stationary solution: for some measurable f, $Y_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots)$, where ε_t is an i.i.d. sequence. Wu (2011) arrived at the same conclusion under the assumption of $\sum_{j=1}^{\infty} c_j = 1$; the specific restrictions on c_j are chosen to reflect their findings, despite the fact that we are not restricting the error process $\{\varepsilon_t\}$ to be an independent sequence in our model setup.

The standard conditions B9 are assumed to deal with the asymptotics of the variance and covariance terms. The last condition B10 is needed only for the self-normalized CLT (24) without assuming higher moment conditions; relevant discussions can be found for example in de la Peña et al. (2009). The condition is not affected by the temporal dependence of the DGP as the property is inherited to the approximated sum in the Bernstein's blocking procedure; see (77) later for details. Lastly, before we proceed, we briefly remark that from now on the rate condition (19) is slightly strengthened as follows (modifying Assumption A2 accordingly):

$$v_2(r)^{1/2}[\varphi_x(\underline{h}\lambda)]^{-1} n^{1/2} \to 0 \qquad as \;\; n \to \infty. \tag{27}$$

With the additional assumptions introduced in this section (B7-B9), the bias and

variance components can be specified as follows, and the CLT (23) can be constructed:

$$\mathcal{B}_n(x) := \left[\mathrm{E}\widehat{m}_2(x) - m(x)\right] \leq h^{\beta}\lambda^{\beta}\sum_{j=1}^{\infty} c_j j^{p\beta} \tag{28}$$

$$\mathcal{V}_n(x) := \mathrm{var}\left[\widehat{m}_2(x)\right] \simeq \frac{\sigma^2(x)\xi_2}{n\varphi_x(\underline{h}\lambda)\xi_1^2}, \tag{29}$$

where λ and $\widehat{m}_2(\cdot)$ are as in (5) and (20), respectively. Formal derivation is done in Section 7.2 of the appendix within the proof for Theorem 2 we introduce below.

4.2.2 Sufficient conditions for the derivation of convergence rates

Convergence rates are crucial in understanding estimators' large sample asymptotics and the evaluation of their performance. Below we introduce a set of conditions under which the rate for our estimator can be specified. We elaborate how the rate of convergence of our estimator depends crucially on (i) the distribution of marginal regressors and (ii) their "cross-sectional" dependence structure. It is important to note that the conditions are *sufficient but not necessary*; we leave other possibilities as future studies, hoping that similar theories would work in a wider range of frameworks.

(i) Dependence across marginal regressors

The way how the marginal regressors are "cross-sectionally" related to each other (given each fixed time) affects the rate of convergence. We consider and allow for the following dependence:

Assumption C. *For every fixed t, the real-valued stochastic process formed by the marginal regressors $\{X_{jt}\}_{j=1}^{\infty}$ has finite fourth moments, i.e. $EX_{jt}^4 \leq C < \infty$ $\forall j,t$, and is stationary and admits the following causal moving average representation:*

$$X_{jt} = \sum_{u=0}^{\infty} a_u \epsilon_{j-u,t}, \tag{30}$$

where a_u is square summable, and $\{\epsilon_{jt}\}_j$ is an orthogonal sequence.

Remark. The dependence structure described in the assumption above is very weak and general, and covers a large class of processes. To elaborate, a necessary and sufficient condition for a stationary sequence to have the representation (30) is

linear regularity[5], a weak notion of dependence. For example, if the stack of marginal regressors $\{X_{jt}\}_j$ (with finite second moments) are independent of each other or α-mixing between themselves, they satisfy the condition because they both imply linearly regularity. Therefore, Assumption C is consistent with both Assumptions A1 (the static regression case) where X_j's are exogenous variables and can have any dependence structure, and Assumption A2 (the dynamic regression case) where X_j's are (temporal) lags of a mixing variable. The special case when X's are Gaussian is worth noting. If $\{X_{jt}\}_j$ is Gaussian and linearly regular, then it has the representation above with *independent and normally distributed* $\epsilon_j's$. An equivalent condition for linear regularity of a Gaussian sequence is simply the existence of a spectral density.

The requirement of finite 4th moment is imposed just to ensure that the *squared* marginal regressors have finite second moments; this is related to the distributional properties of the regressors and will become clear below. Note that obviously, when a lag of the response is included as in the dynamic regression framework (A2), this makes the maximum order of moments of Y_t (i.e. $2+\delta) \geq 4$due to Assumption B4.

(ii) The distribution of regressors

We can show the rate becomes available when, in addition to Condition C above, the marginal regressors satisfy some distributional properties (and when also they are "downweighted" in a certain way via bandwidths). Below we introduce these conditions and discuss them in detail. Note that as defined above, vectors Z and z are taken to mean $(\phi_1^{-1}X_1, \phi_2^{-1}X_2, \ldots)^\intercal$ and $(\phi_1^{-1}x_1, \phi_2^{-1}x_2, \ldots)^\intercal$, respectively, where the vector $x = (x_1, x_2, \ldots)^\intercal$ is the point at which estimation is made, and ϕ_j's are the weight coefficients on bandwidths introduced in Assumption B8 above.

Assumptions D

D1. *The distribution F of X_s^2, where each X_s is the marginal regressor, is regularly varying near zero with strictly positive index $(-\rho) > 0$.*

D2. *The induced probability measure P_{z-Z} is dominated by the measure P_Z, and its Radon-Nikodym density $dP_{z-Z}/dP_Z =: p^*$ is continuous and is bounded away from zero at $0 \in \mathbb{R}^\infty$; i.e., $p^*(0) > 0$.*

D3. *Further to B8, the bandwidth satisfies $h_j = j^p h$ (i.e. $\phi_j = j^p$) with $p \in \Pi(c, \beta)$,*

[5]See Ibragimov and Linnik (1971, Chapter 17) or Davidson (1994, Part III) for details.

where

$$\Pi(c,\beta) = \left\{ p : \sum_{j=1}^{\infty} c_j j^{p\beta} < \infty, \frac{1}{2} < p \right\}.$$

REMARK. Condition D1 concerns with the marginal distributions of the regressor vector. It is equivalent to saying that

$$\lim_{x\to\infty} \frac{F(1/(\gamma x))}{F(1/x)} = \gamma^{\rho},$$

where ρ is the index of variation which is strictly negative. Under this condition, Dunker, Lifshits and Linde (1998, cf. Conditions I and L) derived the explicit behaviour of the small ball probability. We require the function $F(1/x)$ to be regularly varying in order to ensure that the small ball probability is *well-behaved* near infinity in the asymptotic sense. Since only those functions having strictly negative ρ satisfy the condition, the distribution F of the squared regressor must be such that $F(1/x)$ decreases (as $x \to \infty$) at a *reasonable speed*. By reasonable we mean that the relative weight of decrease follows a power law, and the variation should be continuous. A large class of common distributions satisfies this condition; for example: the Gamma, Beta, Pareto, Uniform, Exponential, Weibull, and also the Chi-squared distribution (in which case each marginal regressor X_s is Gaussian).

Assumption D2 is about the transition of the shifted small ball probability to the centred small deviation (whose asymptotic behaviour is more accessible), see Section 2.3 above and Mas (2012). The explicit form of the derivative (and hence of the relationship between the two probabilities) cannot be easily computed in general. Nonetheless, in the special case of the Gaussian process Z with some covariance operator Σ it is known by Sytaya (1974) and Zolotarev (1986) that

$$P\Big(\|z - Z\| \le \epsilon\Big) \simeq P\Big(\|Z\| \le \epsilon\Big) \exp\left\{ -\frac{1}{2}\|\Sigma^{-1/2} z\|^2 \right\} \quad \text{as} \quad \epsilon \to 0. \tag{31}$$

The reader is directed to Li and Shao (2001) for detailed discussion on this asymptotic equivalence relation. Note that Σ can be expressed in terms of the a_j constants (in Assumption C), which govern the dependence across the marginal regressors, and of the bandwidth weights ϕ_j:

$$\mathrm{cov}(Z) = \Sigma = (DA)(A^* D), \tag{32}$$

where $A = (a_{ij}) = (a_{i-j})$ and $D = \mathrm{diag}(\phi_1, \phi_2, \ldots)$.

Assumption D3 is concerned with how the ordered marginal regressors are down-

weighted. The specific bandwidth increment condition assumed in D3 is one framework under which the explicit behaviour of the small ball probability can be specified, (see e.g. Dunker et al. (1998)). In case the regressors are independent to each other, the probability can also be derived when the weights are of an exponential type (i.e. $h_j = e^j h$) up to an unknown function, or are non-increasing in a particular manner (see Gao et al. (2003)) similar to the polynomial decay. In this paper however, we shall confine our attention to the case of the polynomial law for expositional simplicity and consistency of presentation, since the asymptotic behaviour of the small ball is not yet known in the general case for choices other than the polynomial decay as in Assumption D3.

In practice, we would require some ordering for the marginal regressors in the static regressions case A1, since the influence of marginals is set to decrease via the bandwidth adjustments as discussed just above. One practical way of doing this is to rank them in the order of goodness of fit, or the contribution that each marginal regressor makes in the estimation. For example, one could evaluate the sample correlations between Y_t and $\widehat{\mathrm{E}}(Y_t|X_{jt})$, where X_{jt} is a marginal covariate and $\widehat{\mathrm{E}}(Y_t|X_{jt})$ is a kernel estimate of the univariate marginal regression, and order them according to the computed correlations. This way one can line up the marginal regressors in the order of their relative importance. This method is motivated by the Kernel Sure Independence Screening (KSIS) approach in Chen, Li, Linton and Lu (2017), and the reader is referred to their paper for further details.

4.2.3 The Central Limit Theorem

We now introduce the general central limit theorem. The theorem below gives the limiting distribution of the estimator (18) with respect to mixing sample data as described in either Assumption A1 or A2.

Theorem 2 *Suppose that B2-B9 and D1-D3 hold. Let the marginal regressors X_s satisfy Assumption C. Then the estimator (18) based on the sample observations $\{Y_t, X_t\}_{t=1}^n$ satisfying either Assumption A1 or A2 is asymptotically normal with the following limiting distribution:*

$$\sqrt{nh^{\frac{1+2\rho p}{2p-1}} \exp\left(-\kappa_0 h^{-\frac{2}{2p-1}}\right)}\left[\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\right] \Longrightarrow N\left(0, \kappa_1 \sigma^2(x)\right), \qquad (33)$$

where $\mathcal{B}_n(x) = O(h^\beta)$ is the 'bias part' in (28) and $\sigma^2(x) = Var(Y|X = x)$ is the conditional variance defined in Assumption B9.

Below we explain the associated constants. Recall Assumption D1; under that condition, by the characterization theorem of Karamata (1933) (see e.g. Feller (1971)), there always exists a slowly varying function $\ell(x)$ satisfying $F(1/x) = x^{\rho}\ell(x)$. Now fix some p, the order of increment constant for bandwidth in Assumption D3, and denote by $\mathcal{L}(t)$ the Laplace transform of X^2. Then we have,

$$C_\ell = \lim_{\delta \to 0} \left[\ell^{-1/2}\left(\delta^{-\frac{4p}{2p-1}}\right)\right], \quad \zeta = -\int_0^\infty \frac{u^{-1/2p}\mathcal{L}'(u)}{\mathcal{L}(u)} du$$

$$C^* = \frac{(2\pi)^{(1+2p\rho)}(2p-1)}{\Gamma^{-1}(1-\rho)\cdot(2p)^{\frac{2p(\rho+2)-1}{2p-1}}} \cdot \zeta^{\frac{2p(1+\rho)}{2p-1}}, \qquad C^{**} = (2p-1)\cdot\left(\frac{\zeta}{2p}\right)^{2p/(2p-1)}$$

$$\kappa_0(K,p,F) = C^{**}(C_{\mathcal{A}}\lambda^{-1})^{\frac{2}{2p-1}} \quad \text{and} \quad \kappa_1(K,p,F) = \frac{C^* C_\ell \xi_2}{p^*(0)\xi_1^2 (C_{\mathcal{A}}\lambda^{-1})^{\frac{1+2\rho p}{1-2p}}},$$

where $\Gamma(\cdot)$ is the Gamma function, ξ_1 and ξ_2 are the constants specified in (12) (which simplify in case of uniform kernel for example), λ is the upper bound of the support of the kernel, and $p^*(\cdot)$ is the Radon-Nikodym derivative defined in D2. Recall that for the uniform (Box) kernel $\xi_2 = \xi_1^2$, so they cancel out in κ_2. See Dunker, Lifshits and Linde (1998) and also Hong, Lifshits and Nazarov (2016) for some discussions on the underlying arguments for the formulation of these constants. To aid the exposition, we compute and present the constants for some common, regularly varying distributions in the table below.

$X_j^2 \sim F$ i.i.d.	ρ	$\lim_{x\to\infty} \ell(x) = C_\ell^{-2}$	ζ
Uniform(1,b)	-1	1	n/a
Gamma(α,β)	$-\alpha$	$\beta^\alpha \alpha^{-1}\Gamma(\alpha)^{-1}$	$\frac{\alpha\pi\beta^{-1/2p}}{\sin(\pi/2p)}$
Exponential(η)	-1	η	$\frac{\pi\eta^{-1/2p}}{\sin(\pi/2p)}$
Weibull(α,β)	$-\alpha$	β	n/a
Pareto(θ,μ)	-1	μ/θ	n/a
χ_1^2	$-1/2$	$(2/\pi)^{1/2}$	$\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)}$

Table 1: Examples of the key constants for some common distributions

For example, when X_j's are Gaussian, the constants C^* and C^{**} denoted C_G^* and C_G^{**} respectively, are given by:

$$C_G^* = \frac{(2\pi)^{(1-p)}(2p-1)}{2\cdot(2p)^{\frac{3p-1}{2p-1}}} \cdot \left[\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)}\right]^{\frac{-p}{2p-1}}, \qquad C_G^{**} = \frac{2p-1}{2}\left(\frac{\pi}{2p\sin\frac{\pi}{2p}}\right)^{\frac{2p}{2p-1}}.$$

and

$$\kappa_2(K,p,a) = \frac{C_G^* C_\ell \xi_2 \xi_1^{-2}}{e^{-\frac{1}{2}\|\Sigma^{-1/2}z\|_2^2}(C_{\mathcal{A}}\lambda^{-1})^{\frac{p-1}{2p-1}}}$$

where $z = (z_j) = (j^{-p}x_j) = D^{-1}x$. The exponential term in the denominator of the asymptotic variance arises from the asymptotic equivalence relationship between the shifted and non-shifted small deviation for ℓ_2-valued Gaussian variables, cf. (31).

The $C_{\mathcal{A}}$ constant represents the dependence across the marginal regressors and plays an important role. First of all, when the marginal regressors are identically distributed and independent to each other, then $C_{\mathcal{A}} = 1$. If not independent, the specific form of this constant is only known in the Gaussian case (by Hong, Lifshits and Nazarov (2016, Theorem 1.1)). Specifically, when $\{X_j\}_j$ is centred Gaussian and satisfies Assumption C, then given the square summable sequence a_j in (30), we have

$$C_{\mathcal{A}} = \left[\frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{j=0}^{\infty} a_j \exp(\mathrm{i}js)\right|^{1/p} ds\right]^p. \tag{34}$$

It is worth noting that $C_{\mathcal{A}}$ is a function of the spectral density of the MA(∞) process $\{X_{jt}\}_j$ denoted $S_X(\cdot)$. Specifically, some straightforward algebra gives $C_{\mathcal{A}} = (2\pi)^{p(p-2)/2} \cdot [\int_0^{2\pi} S_X(s)^{p/2} ds]^p$.

The implications of the constant $C_{\mathcal{A}}$ suggest an interesting finding that allowing for dependence does not seem to incur much penalty; we conjecture that similar conclusion would hold for regressors of different distributions than Gaussian, but leave it for future studies.

4.3 Optimal Bandwidth

We now discuss the issue of bandwidth optimality. As in the finite-dimensional framework, there is a bias-variance trade-off. As the bandwidth goes up, the variance gets smaller while the bias increases, and vice versa. Therefore we search for the optimal bandwidth h_{opt} that balances the order of those two quantities.

We first suppose that $p \in \Pi(c,\beta)$, cf. D3, is given. In the i.i.d. case with Gaussian regressor we have

$$h^\beta \sim \sqrt{\frac{\exp\left(\kappa_0 h^{-2/(2p-1)}\right)}{nh^{\frac{1-p}{2p-1}}}}, \tag{35}$$

so that

$$\left[2\beta + \frac{1-p}{2p-1}\right] \cdot \log h - \kappa_0 h^{-\frac{2}{2p-1}} \sim -\log n.$$

Taking $h \sim (\log n)^a$ for some $a < 0$ balances the leading terms on both sides:

$$\left[2\beta + \frac{1-p}{2p-1}\right] \cdot a \cdot \log\log n - \kappa_0 (\log n)^{-\frac{2}{2p-1}\cdot a} \sim -\log n. \tag{36}$$

The explicit order a that solves (36) can be expressed in terms of n, β and p. Writing $\vartheta := [2\beta + (1-p)/(2p-1)]$ and $\chi := 2/(2p-1)$ for notational simplicity, and solving for a we have

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}\left(\frac{\chi}{\vartheta} \cdot \kappa_0 \cdot n^{\chi/\vartheta}\right) - \chi \log n}{\vartheta\chi \cdot \log\log n}, \tag{37}$$

where $\mathcal{W}(y)$ is the Lambert W function (see e.g. Olver et al. (2010)), which returns the solution x of $y = x \cdot e^x$. From (37) the optimal bandwidth $h_{opt} \sim (\log n)^{a_{opt}}$ follows, in which case the asymptotic root mean squared error is of the order $(\log n)^{\beta a_{opt}}$.

Remark. We can look for the optimal bandwidth for the cases of non-Gaussian regressors by following exactly the same procedure as above; tedious details are omitted here. Lower value of ρ makes the rate better in general. Regarding the solution in (37), since the mapping $x \mapsto x \cdot e^x$ is not an injection, the solution may be multi-valued on the negative domain, i.e. $y < 0$. This does not happen in (37) provided $\beta \geq 1/4$ (however big p is), because $(1-p)/(2p-1)$ is bounded away from $-1/2$; in this case, the coefficient of the double logarithmic term in (36) is strictly smaller or equal to zero.

Since the log terms dominate the double logarithm in (36) as the sample size n increases, it can be readily expected that the optimal value of a in (37) converges to a limit in such a way that the leading orders are balanced. Below we introduce without formal justification a trivial result that gives the lower bound (infimum) of the optimal bandwidth (and hence of the optimal rate that balances the bias and variance). We remark that the result below holds for other choices of the distribution of the regressors, and also for the case of dependent (non-independent) regressors as allowed in Assumption C (i.e. when $C_{\mathcal{A}} \neq 1$), as the exponent of the leading term $-2/(2p-1)$ remains invariant as is clear from Theorem 2. Nonetheless, it is worth noting that although the order of the convergence rate remains the same, the difference in associated constants does make the speed at which a_{opt} converge to the limit in (38).

Corollary 2 *For any fixed choice of $p \in \Pi(c, \beta)$ and the distribution F of X^2 satis-*

fying Assumption D1, the order of the optimal bandwidth a_{opt} satisfies

$$a_{opt} \downarrow \left(-\frac{2p-1}{2}\right) \quad as \quad n \to \infty, \tag{38}$$

which suggests that the lower bound of the optimal bandwidth is given by

$$(\log n)^{-\frac{2p-1}{2}} \preceq h_{opt} \sim (\log n)^{a_{opt}}. \tag{39}$$

This result tells us the best possible performance we can expect from the optimal bandwidth. Higher p and β makes the rate better. But, because $n^k(\log n)^{-(2p-1)/2} \to \infty$ for any real $k > 0$, it follows that we cannot possibly estimate the regression function at a polynomial rate. This result is consistent with and complements the findings of Mas (2012, Theorem 3), which were obtained under the assumption of independence of regressors. The paper also considered the case where the bandwidth grows exponentially: $\phi_j \succeq \exp(j^q)$ for some $q > 0$. Then for $h_j = \phi_j h$, his result suggests $\exp[-(\log n)^{2q/(2q+1)}] \preceq h_{opt} \sim \exp[a_{opt} \cdot (\log n)^{b_{opt}}]$. Therefore, the performance is better in general in this case, although obviously a polynomial rate of convergence still cannot be attained. It is not clear what will happen when the regressors are allowed to be dependent in the sense of Assumption C, since the behaviour of the small ball probability for non-independent (and non-Gaussian) sequence is not known for the case of exponentially decaying weights.

Returning back to Corollary 3, we emphasize that the arguments are true for any $p \in \Pi(c, \beta)$. Let $p_{\max} = \sup_{p \in \Pi(c,\beta)} p$. Then a lower bound on the optimal bandwidth (over all p) is $(\log n)^{-\left(p_{\max} - \frac{1}{2}\right)}$. For example, when $c_j = (1/2)j^{-2}$ we have $p_{\max} = 1/\beta$. Unfortunately, it is generally the case that $p_{\max} \notin \Pi(c, \beta)$, in which case the lower bound is not quite achievable by our method.

Remark. Regarding bandwidth selection, one possibility is the Bayesian bandwidth selection methods like proposed in Zhang, King, and Hyndman (2006). We take as prior for h the density proportional to $1/(1 + \lambda h^2)$ and as prior for $p - 1/2$ the density of a $\chi^2(w)$ random variable. The hyperparameters λ, w may be chosen by experimentation. The priors are combined with a Gaussian (least squares) density to deliver a posterior for the bandwidth. The reader is referred to Section 5 for further discussions on the issue of bandwidth choice.

4.4 Uniform consistency

Uniform consistency of the Nadaraya-Watson estimator was first studied by Nadaraya (1964, 1970) and subsequently by numerous others. To mention few early papers, Devroye (1978) weakened the regularity conditions required in the previous papers, and Robinson (1983) proved uniform consistency for dependent sample data. In the functional statistics literature, uniform consistency of kernel estimators for conditional mean is established only with respect to i.i.d. sample data so far (see for example Ferraty et al. (2010), Ferraty et al. (2011), Kudraszow and Vieu (2013), and Kara-Zaïtri et al. (2017)) as far as the authors are concerned.

In this section, we show uniform consistency of our estimator under the (suitably modified) regularity conditions assumed in the previous sections. We start by introducing the notion of Kolmogorov's entropy below. For some of its earlier discussions in statistics literature, the reader is referred to Yaracos (1985) and Mammen (1991).

DEFINITION 5. *Given some $\eta > 0$, let $L(S, \eta)$ be the smallest number of open balls in E of radius η needed to cover the set $S \subset E$. Then Kolmogorov's η-entropy is defined as* $\log L(S, \eta)$.

This quantity will be used in explaining the topological restrictions we adopt to suitably accommodate infinite dimensionality. The definition implies the dependence of Kolmogorov's entropy both on the nature of the space under study and the measure of proximity. It will be shown later in this section that the entropy is closely related to the rate of convergence of the estimator, in particular, to the penalty incurred on the rate in the uniform case. It is well known that the regression function cannot be estimated uniformly over the entire space, e.g. Bosq (1996). In our infinite dimensional framework, the infinite sequence spaces, if unrestricted, cannot be covered by a finite number of balls, and that $L(S, \eta) = \infty$. We propose to consider uniform consistency over a subset of $\mathbb{R}^\infty$, whose effective dimension is truncated and is increasing in sample size n. In particular, we define the set

$$S_\tau := \left\{ u | (u_i)_{i \in \mathbb{Z}^+}, u_j = 0 \text{ for all } j > \tau, \|u\|_\infty \leq \lambda \right\} \subset \mathbb{R}^\infty, \tag{40}$$

where $\tau = \tau_n$ is some increasing sequence and λ is fixed, and consider uniform consistency over this compact set. Then Kolmogorov's entropy of the set S_τ is given as follows:

Lemma 2 *Kolmogorov's η-entropy of S_τ with $\tau = \tau_n(\to \infty)$ and $\lambda > 0$ is*

$$\log L\big(S_\tau, \eta\big) = \log \left[\left(\frac{2\lambda\sqrt{\tau}}{\eta} + 1 \right)^{\tau} \right]. \tag{41}$$

REMARK. (41) is in line with common intuition; as the effective dimension τ increases, the number of balls (with some fixed radius) required to cover the set tends to infinity. Lemma 2 can be shown by exploiting the splitting technique and then by covering the polyhedron of increasing dimension. See appendix for details. Note that for λ fixed and $\eta = \eta_n$, Kolmogorov's entropy $\log L(S_\tau, \eta)$ is of order $(\tau \log \tau - \tau \log \eta)$.

Considering the definition of the set S_τ, in the sequel (with a slight abuse of notation) we take X to denote the regressor, but with zeros after its $\tau^{th}(= \tau_n \to \infty$ as $n \to \infty)$ entry; that is, $X = (X_1, X_2, \ldots, X_\tau, 0, 0, \ldots)^\intercal$ (so that the original X is recovered as $n \to \infty$). Also, the regression operator and the estimator with respect to this truncated regressor are denoted by $m_\tau(\cdot)$ and $\widehat{m}_\tau(\cdot)$, respectively. All assumptions, including the additional one to follow below, are understood to hold under these modifications.

ASSUMPTION E

E. *For sufficiently large n, Kolmogorov's η-entropy $\log L(S_\tau, \eta)$ satisfies*

$$\frac{(\log n)^{8+2\epsilon}}{n\varphi_x(\underline{h})} \leq \log L(S_\tau, \eta) \leq \frac{\sqrt{n\varphi_x(\underline{h})}}{(\log n)^{1+\epsilon}} \quad \text{for some } \epsilon \in (0, 1/2). \tag{42}$$

Furthermore, $0 < \varphi_x(\underline{h}) \preceq h < \infty$ and $(\log n)^2/(n\varphi_x(\underline{h})) \longrightarrow 0$ as $n \to \infty$.

REMARK. The first part of Assumption E specifies the rate at which Kolmogorov's entropy should behave with sample size n (hence in dimension $\tau = \tau_n$). From the upper and lower bound it readily follows that $n\varphi(h)$ must be of order larger than $(\log n)^{6+2\epsilon}$. This assumption is sufficiently general. For example, in view of the bias-variance optimal bandwidth suppose $h \sim (\log n)^{-(2p-1)/2}$ so that $n\varphi(h) \sim (\log n)^{(2p-1)\beta}$. In this case, assumption (42) is valid as long as p is moderately large enough relative to $\beta \leq 1$ in such a way that $6 + 2\epsilon \leq (2p-1)\beta$. The second part is standard; in particular, the last condition straightforwardly follows by (42) and only slightly strengthens the bandwidth condition in Assumption B2.

For uniform consistency we shall impose a stronger condition on mixing coefficients. From hereafter, by A1′ and A2′ we mean Assumptions A1 and A2 but with the arithmetic mixing rate condition strengthened to the following exponential mixing condition (cf. Definition 1):

$$\alpha(r) \leq \exp(-\varsigma r^{\gamma_2}) \tag{43}$$

where $\varsigma > 1$ and γ_2 is a positive constant such that $\gamma := 1/(\gamma_1^{-1} + \gamma_2^{-1}) \geq 1$, with γ_1 defined as in Assumption B4′. When the response is assumed to be bounded (i.e. $|Y_t| \leq C$), γ_1 may be taken to be ∞ so that $\gamma_2 = \gamma \geq 1$. This stronger mixing condition enables us to obtain exponential bounds that decay fast enough, thereby accommodating uniformity, see appendix for details. We hope to expect the same conclusion in this section to hold under the arithmetic mixing condition we previously assumed, once some suitably sharper exponential inequality becomes available. In line with the modification on the mixing rate above, we also impose a slightly stricter condition on the response:

B4′. *The response Y_t is satisfies the following tail condition: There exists some positive constant γ_1 and C such that $P(|Y_t| > u) \leq C\exp(1 - u^{\gamma_1})$ for any $u > 0$.*

For example, a Gaussian random variable satisfies B4′ with $\gamma_1 < 2$. The condition is also satisfied by many unbounded variables and all those bounded ones. The main result of this section now follows.

Theorem 3 *Suppose that Assumptions B2, B3, B4′, B5-B9, D1-D3 and E hold. Let the marginal regressors X_s satisfy Assumption C, and take $\tau = \tau_n \sim (\log n)$. Then the estimator $\widehat{m}_\tau(\cdot)$ with respect to sample observations $\{Y_t, X_t\}_{t=1}^n$ satisfying A1′ is uniformly consistent for $m(x) = m(x_1, x_2, \ldots)$ over S_τ:*

$$\sup_{x \in S_\tau} \left| \widehat{m}_\tau(x) - m_\tau(x) \right| = O_P\left(h^\beta + \sqrt{\frac{(\log n)^2 \exp\left(\kappa_0 h^{-2/(2p-1)}\right)}{n h^{\frac{1+2p\rho}{2p-1}}}} \right). \tag{44}$$

Remark. We may choose the optimal bandwidth as before; following the same arguments in the pointwise case, choosing $h \sim (\log n)^a$ and solving for n gives

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}\left[\frac{\chi}{\vartheta} c \exp(-\frac{\chi}{\vartheta} 2\log\log n + \chi \log n)\right] + 2\chi \log\log n - \chi \log n}{\vartheta \chi \log\log n}. \tag{45}$$

And because the order of the leading terms is $(\log n)^{-(2p-1)/2}$ as in the pointwise case, it is straightforward to see that the lower bound of the optimal bandwidth in Corollary 3 still continues to hold; that is, $h_{opt} \succeq (\log n)^{-(2p-1)/2}$. This is again invariant to the choice of distribution F of the squared regressor. It is important to note that as before, potential cross-sectional dependence between marginal regressors and also their distributional properties are represented via c, the collection of constants that appear inside the exponential terms in the asymptotic variance.

The results altogether give the optimal rate of convergence of our estimator as follows.

Corollary 3 *Suppose conditions assumed in Theorem 2.4 hold. Upon choosing* $h \sim (\log n)^{a_{opt}}$*, where* a_{opt} *is as defined in (45), we have*

$$\sup_{x \in S_\tau} \left| \widehat{m}_\tau(x) - m_\tau(x) \right| = O_P\Big([\log n]^{\beta \cdot a_{opt}} \Big). \tag{46}$$

In the pointwise case the same result in Corollary 4 trivially holds, but with the different optimal a_{opt}; it is as given in (37). In that case this rate of convergence is minimax optimal in view of Theorem 3 of Mas (2012). Although both a_{opt} converge to $-(2p-1)/2$, the speed at which they converge is different as can be seen in the example in Figure 1 below.

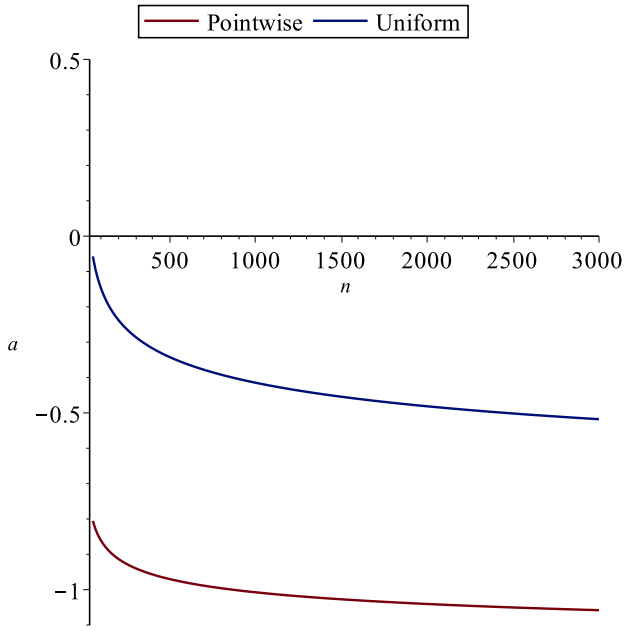

Figure 1: $a_{opt} = a_{opt}(n)$ for $\beta = 1$ and $p = 2$

5 Application to the Risk Return Relationship

The relation between the expected excess return on the aggregate stock market - the so called "equity risk premium" - and its conditional variance has long been the subject of both theoretical and empirical research in financial economics. The risk-return relation is an important ingredient in optimal portfolio choice, and is central to the development of theoretical asset-pricing models aimed at explaining a host of observed stock market patterns. Asset pricing models generally predict a positive relationship between the risk premium on the market portfolio and the variance of its return. In an influential paper, Merton (1973) obtained very simple restrictions albeit under somewhat drastic assumptions; he showed in the context of a continuous time partial equilibrium model that

$$\mu_t = E[(r_{mt} - r_{ft})|\mathcal{F}_{t-1}] = \gamma \times \text{var}[(r_{mt} - r_{ft})|\mathcal{F}_{t-1}] = \gamma\sigma_t^2, \tag{47}$$

where r_{mt}, r_{ft} are the returns on the market portfolio and risk-free asset respectively, while $\mathcal{F}_{t-1}$ is the market wide information available at time $t-1$. The positive constant γ is the Arrow–Pratt measure of relative risk aversion. The linear functional form actually only holds when σ_t^2 is constant; otherwise μ_t and σ_t^2 can be nonlinearly related, Gennotte and Marsh (1993). Further examples with a positive risk return trade-off include the external habit model of Campbell and Cochrane (1999) and the Long Run Risks model of Bansal and Yaron (2004). However, a negative risk-return relation is not inconsistent with (a general enough) equilibrium, Backus and Gregory (1993). Unfortunately, the empirical evidence on the risk-return relation is mixed and inconclusive. Ghysels, Santa-Clara, and Valkanov (2005), Lundblad (2005), Bali and Peng (2006), Pástor, Sinha, and Swaminathan (2008), and Ludvigson and Ng (2007) find a positive risk-return relation, while Campbell (1987), Glosten, Jagannathan, and Runkle (1993), Harvey (2001), and Lettau and Ludvigson (2003) find a negative relation. Still others find mixed and inconclusive evidence like French, Schwert, and Stambaugh (1987), Nelson (1991), Campbell and Hentschel (1992), Linton and Perron (2003), and Whitelaw (1994). Scruggs (1998) and Guo and Whitelaw (2006) document a positive trade-off within specifications that facilitate hedging demands. However, Scruggs and Glabadanidis (2003) find that this partial relationship is not robust across alternative volatility specifications.

As already mentioned in the beginning of this paper, the main difficulty in estimating the risk-return relation is that neither the conditional expected return nor the conditional variance of the market is directly observable. The contradictory findings

of the above studies are mostly the result of differences in the specifications and approaches to modelling the conditional mean and variance. Pagan and Ullah (1988), and Pagan and Hong (1990) initiated the use of nonparametric methods in this setting. The latter paper argued that the risk premium μ_t and the conditional variance σ_t^2 are highly nonlinear functions of the past whose form is not captured by standard parametric GARCH–M models. They estimated $E(r_{mt} - r_{ft}|r_{m,t-1}, \ldots, r_{m,t-p})$ and $\text{var}(r_{mt} - r_{ft}|r_{m,t-1}, \ldots, r_{m,t-p})$ nonparametrically, where $p \in \{1, 4\}$, finding evidence of considerable nonlinearity. They then estimated δ from the regression

$$r_{mt} - r_{ft} = \delta\sigma_t^2 + \eta_t, \tag{48}$$

by OLS and IV methods, finding a negative but insignificant δ. There are a number of drawbacks with the Pagan and Hong (1990) approach. Firstly, as aforementioned in the introduction, the conditional moments are calculated using a finite and small conditioning set. This greatly restricts the dynamics for the variance process. Secondly, they only test for linearity of the relationship between μ_t and σ_t^2; this seems to be somewhat restrictive in view of earlier findings. Linton and Perron (2003) considered the model where σ_t^2 was a parametrically specified CH process (with dependence on the infinite past) but $\mu_t = \varphi(\sigma_t^2)$ for some function φ of unknown functional form. They proposed an estimation algorithm but did not establish any statistical properties. They found some evidence of a nonlinear relationship. Conrad and Mammen (2008) develop the theory of estimation and inference for this model. Christensen, Dahl, and Iglesias (2012) developed the theoretical framework by considering volatility models that are driven by observable shocks so that a full theory can be given. Escanciano, Pardo-Fernández and Van Keilegom (2017) consider a more general class of semiparametric models. Under the semi-strong form of the efficient market hypothesis prices contain all relevant information and so the risk premium and risk themselves can be expressed in terms of only the past history of prices. We shall use this assumption to obviate the omitted variables/endogeneity issues that have limited previous applications in this area.

5.1 Empirical study on the US stock market

We apply our methods to the daily risk premium on the value weighted S&P500 index — the total return on the index minus the returns on T-bills[6], denoted Y_t — over the period 04 January 1950 to 30 August 2017, a total of 17,025 observations. The

[6]Data obtained from Yahoo Finance and Kenneth French's Data Library.

whole time period is divided into 5 subperiods: 1950:01:04-1963:02:21, 1963:02:25-1976:05:04, 1976:05:05-1989:05:24, 1989:05:25-2002:06:24, and 2002:06:25-2017:08:30, to see if there is any variation in the ex-post risk and return by decades. Except for the last subperiod where there are 3824 observations, the other five each contains 3300 observations. We suppose that both the conditional mean and variance of Y_t, denoted μ_t and σ_t^2, are unrestricted nonparametric functions of the entire information set. We estimate them for $p = 4$ and 12 at the points $X_t = (Y_{t-1}, Y_{t-2}, \ldots, Y_1, 0, 0, \ldots)$. The uniform kernel $K(\|u\|) = 1_{[0,1]}(\|u\|)$ is used, and the bandwidth sequence h of 0.00035 and 0.000125 are used for $p = 4$ and $p = 12$, respectively. These bandwidths are in accordance with the selection methods we propose below in the end of this section.

Table 2 reports some summary statistics of the nonparametric estimates $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ for $p = 4$ over the full period (1950-2017). We present the mean, standard deviation, skewness, kurtosis and the fitted AR(1) coefficients. The estimated conditional variance shows high persistence. Table 2 may be compared with Table I of Bali and Peng (2006), where they report similar descriptive statistics for their realized, GARCH, and implied volatility estimates computed using 5-minute high frequency dataset. Note that their time period is different (1982-2002), and they present excess kurtosis.

Table 2: Summary statistics of the estimates $(\widehat{\mu}_t, \widehat{\sigma}_t^2)$
Full Period (1950-2017)

	Mean	Std	Skewness	Kurtosis	AR(1)
$\widehat{\mu}_t$	3.1260×10^{-4}	2.3346×10^{-3}	1.03347	3.4194	0.0190
$\widehat{\sigma}_t^2$	6.5894×10^{-5}	4.8913×10^{-5}	6.6066	76.20624	0.7033

Figure 2 reports the (annualized) estimated values, that is, $(\sqrt{252} \cdot \widehat{\sigma}_t, 252 \cdot \widehat{\mu}_t)$, $t = 2, \ldots, n(= 17025)$ when $p = 4$. The result shows there is no noticeable disparity over different time periods, although the estimates are more spread out in the more recent periods, showing higher variability. Having a different number of observations does not seem to affect the conclusion either, seeing from the last plot. Interestingly, the number of negative expected excess returns is quite large; such estimates are not inconsistent with asset pricing theory, Boudoukh et al. (1997), Whitelaw (2000), Harvey (2001). The plot of estimates evaluated when $p = 12$ – omitted here – reports similar findings, except that the estimates are a bit more concentrated.

Note that to focus on the main "chunk" of the fitted values, where almost all observations are located, the plots in Figure 2 are magnified and truncated on the

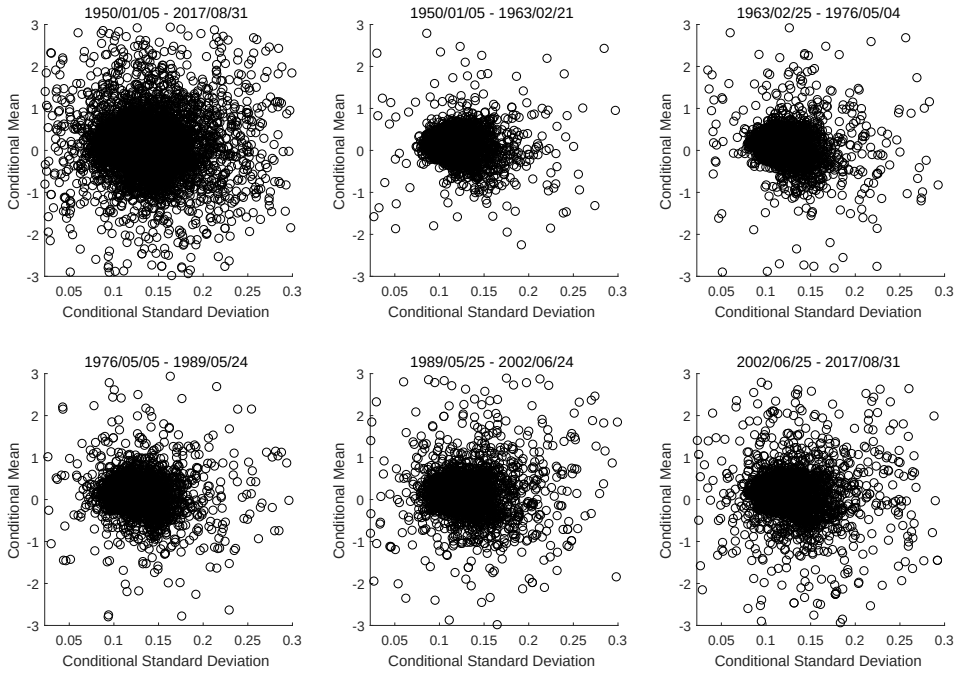

Figure 2: Annualized estimates of conditional mean and standard deviation, $p = 4$

ranges of $[-3, 3]$ of the y-axis and $[0.0225, 0.3]$ of the x-axis; around 96.1% of the entire fitted values appear in the plot. In particular, among those not appearing in the plot are those with zero estimated conditional standard deviation, which constitute around 3% of the whole estimates. This happens when, at a point of evaluation X_t, only one kernel in the sums returns a value of 1 and zero otherwise, so that the second moment equals the squared first moment. One way of reducing the number of such estimates is to increase the bandwidth. We do not proceed to this direction because it makes the bandwidth sub-optimal and the number of those observations is rather negligible.

Figure 3 and 4 show the estimated relationship obtained using local constant smoothing, with the bandwidth chosen according to Silverman's rule of thumb. The smooths are evaluated at the 100 quantiles of the marginal distribution so that the spacing of the covariate can be shown. The first and last 5 values are taken out, since they tend to be extreme values in general, and including them may make the graph look misleading. All subplots suggest that quadratic fits, i.e. including the conditional variance term, would be appropriate. From Figure 4, we note that when $p = 12$, i.e. when the influence of distant lags is "less weighted", we begin to see some negative relationships in some subplots (especially in the more recent periods), which is consistent with our findings later in this section.

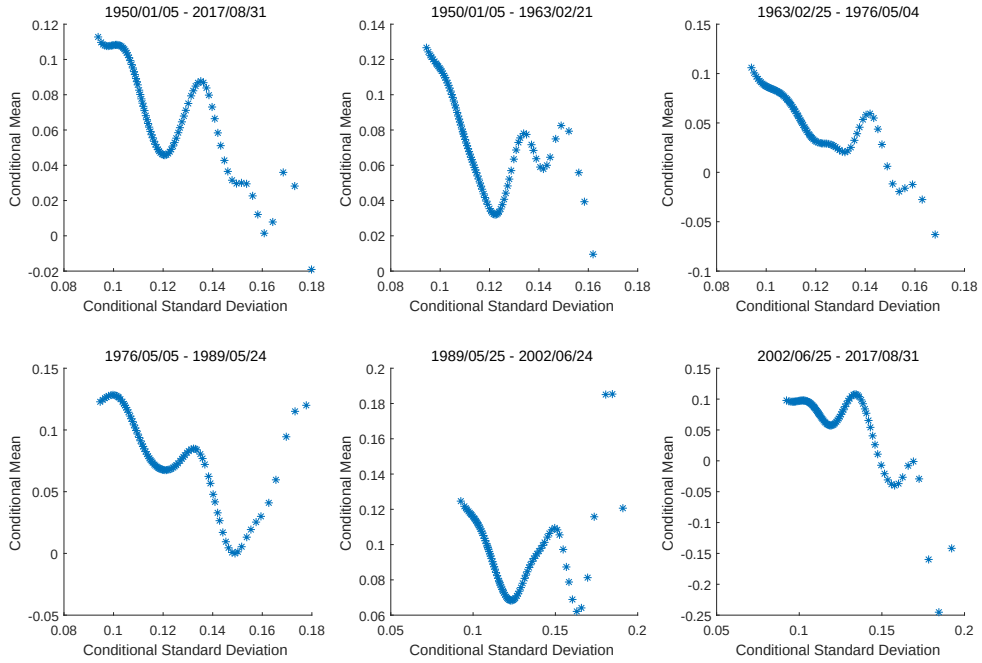

Figure 3: Estimated relationship between annualized $\widehat{\sigma}_t$ and $\widehat{\mu}_t$; $p = 4$

Now we consider some parametric analysis, and suppose the conditional mean $\mu(x) = \mathrm{E}(Y|X = x)$ and conditional variance $\sigma^2(x) = \mathrm{var}(Y|X = x)$ are related in a quadratic way, i.e.,

$$\mu(x) = \alpha + \gamma\sigma(x) + \beta\sigma^2(x), \tag{49}$$

where $\theta = (\alpha, \gamma, \beta)^\intercal$ with α, β, γ being unknown constants. Let $x_1, x_2, \ldots, x_q \in \mathbb{R}^\infty$ be some given points such that $\|D^{-1}(x_j - x_k)\| > 0$ for all j, k, and let $\widehat{\mu}(x)$ and $\widehat{\sigma}^2(x)$ be the estimated moments.

Then we take

$$\widehat{\theta} = \left(\widehat{\alpha}, \widehat{\gamma}, \widehat{\beta}\right)^\intercal = \widehat{\Sigma}_q^{-1}\widehat{U}_q$$

and

$$\widehat{\Sigma}_q = \begin{pmatrix} 1 & \sum_{i=1}^q \widehat{\sigma}(x_i) & \sum_{i=1}^q \widehat{\sigma}^2(x_i) \\ \sum_{i=1}^q \widehat{\sigma}(x_i) & \sum_{i=1}^q \widehat{\sigma}^2(x_i) & \sum_{i=1}^q \widehat{\sigma}^3(x_i) \\ \sum_{i=1}^q \widehat{\sigma}^2(x_i) & \sum_{i=1}^q \widehat{\sigma}^3(x_i) & \sum_{i=1}^q \widehat{\sigma}^4(x_i) \end{pmatrix} \quad ; \quad \widehat{U}_q = \begin{pmatrix} \sum_{i=1}^q \widehat{\mu}(x_i) \\ \sum_{i=1}^q \widehat{\sigma}(x_i)\widehat{\mu}(x_i) \\ \sum_{i=1}^q \widehat{\sigma}^2(x_i)\widehat{\mu}(x_i) \end{pmatrix},$$

where q is finite.

We next derive the limiting distribution of the vector of estimated coefficients $\widehat{\theta} := (\widehat{\alpha}, \widehat{\gamma}, \widehat{\beta})^\intercal$, which can be used for conducting statistical inference. Define:

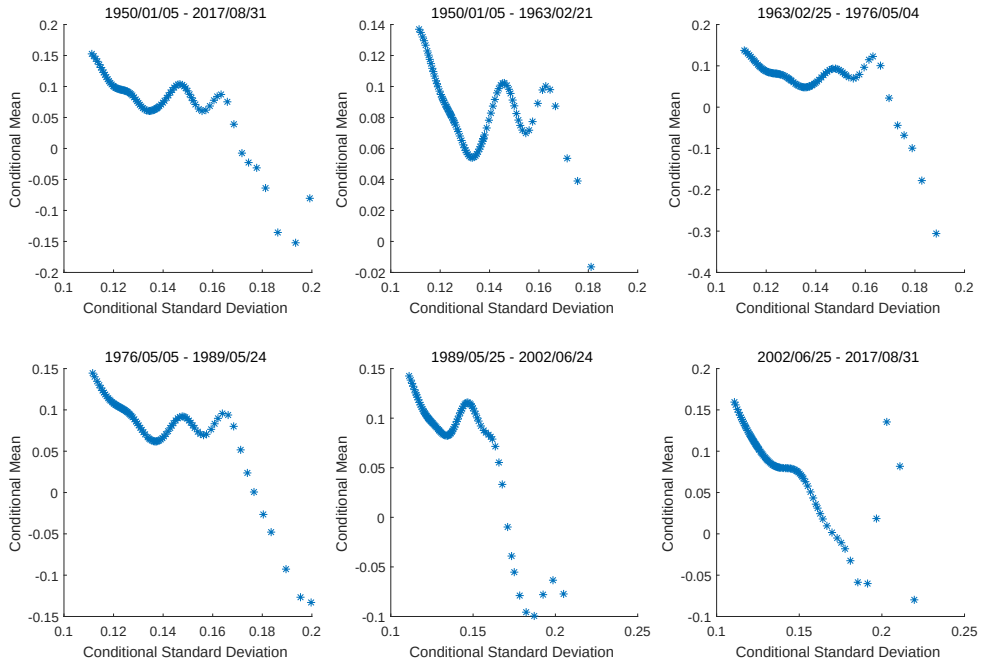

Figure 4: Estimated relationship between annualized $\widehat{\sigma}_t$ and $\widehat{\mu}_t$; $p = 12$

$$\Sigma_q = \begin{pmatrix} 1 & \sum_{i=1}^q \sigma(x_i) & \sum_{i=1}^q \sigma^2(x_i) \\ \sum_{i=1}^q \sigma(x_i) & \sum_{i=1}^q \sigma^2(x_i) & \sum_{i=1}^q \sigma^3(x_i) \\ \sum_{i=1}^q \sigma^2(x_i) & \sum_{i=1}^q \sigma^3(x_i) & \sum_{i=1}^q \sigma^4(x_i) \end{pmatrix}$$

$$\Omega(x_i) = \begin{pmatrix} \sigma^2(x_i) & \text{skew}(Y_t|X_t = x_i) \\ \text{skew}(Y_t|X_t = x_i) & \sigma^4(x_i)\left(\text{kurt}(Y_t|X_t = x_i) + 2\right) \end{pmatrix} =: \begin{pmatrix} \omega_{1,1}(x_i) & \omega_{1,2}(x_i) \\ \omega_{2,1}(x_i) & \omega_{2,2}(x_i) \end{pmatrix},$$

$$V_q = \sum_{i=1}^q J(x_i)\Omega(x_i)J(x_i)^\intercal \quad ; \quad J(x_i) = \begin{pmatrix} 1 & 0 \\ \sigma(x_i) & \frac{\mu}{2\sigma}(x_i) \\ \sigma^2(x_i) & \mu(x_i) \end{pmatrix}.$$

Here, skew and kurt denote skewness and kurtosis of Y_t (conditional on $X_t = x_i$). The result is a direct consequence of consistency of estimated moments and their asymptotic independence across i.

Theorem 4 *Let Assumptions B2, B3, B5-B9, and D1-D3 hold, and suppose B4 is strengthened to require $E(|Y_t|^{8+\delta}) \leq C < \infty$ for some $C, \delta > 0$. Suppose the operator $g(\cdot) = E(Y^2|X = \cdot)$ satisfies Assumption B7. Suppose further that $\omega_{a,b}(u)$ is continuous in some neighbourhood of x_i for all i. Then, given the sample observations*

$\{Y_t, X_t\}_{t=1}^n$ *specified in A2, we have the following limiting distribution:*

$$\sqrt{nh^{\frac{1-p}{2p-1}} \exp\left(-\kappa_0' h^{-\frac{2}{2p-1}}\right)} \left(\widehat{\theta} - \theta - B_\theta\right) \Longrightarrow N\Big(0, \kappa_2(K,p,a)\Sigma_q^{-1} V_q \Sigma_q^{-1}\Big),$$

where B_θ *is a bias terms of order* h^β*,* κ_2 *is the constant defined in Section 2.4.2.*

The parameters θ are estimated at the same rate as the functions $\mu(\cdot)$ and $\sigma^2(\cdot)$. It may be possible to achieve faster rates of convergence by allowing $q \to \infty$, as is commonly done in the semiparametric literature, but we have not yet been able to establish this rate improvement; see Chen and Christensen (2015).

With the same S&P500 data as before and the nonparametric estimates we obtained for $p = 4$ and reported in Figure 2, we fit the linear regression (49). Note that those with zero estimated variance we discussed above are removed (around 3% of whole data), since they can make the fitted estimates misleading and spurious. Also, the standard deviation term is deliberately removed to allow for a direct comparison with the results from those in the existing literature, Pagan and Ullah (1988), Pagan and Hong (1990) and Harvey (2001). We estimate the coefficients α and β, and provide the results along with the values of t-statistics that $\alpha = 0$ and $\beta = 0$ in Table 3. The first subperiod is omitted because the estimates for earlier periods may be less reliable due to being evaluated at points with many zeros. Parentheses marked with asterisks (respectively, double asterisks) mean that the corresponding estimates are statistically different from zero at 5% level (respectively, 1% level) of significance based on Newey-West standard errors.

Table 3: Estimated parameters obtained using $(252 \cdot \widehat{\mu}_t, 252 \cdot \widehat{\sigma}_t^2)$
Full (1954-2017) and Sub Periods

	Full	1954-1963	1963-1976	1976-1990	1990-2003	2003-2017
α	0.07264	0.10046	0.01672	0.14233	0.09700	0.03761
(t)	$(9.060)^{**}$	$(2.2340)^{*}$	(0.4488)	$(5.0488)^{**}$	$(6.7565)^{**}$	$(2.3038)^{*}$
β	0.40830	-1.64391	2.9360	-3.69919	0.44916	2.14394
(t)	(1.063)	(-0.5213)	(1.2656)	$(-2.9391)^{**}$	(0.6707)	$(2.3265)^{**}$

The result reports a positive effect (0.4083, with $t = 1.063$) of conditional variance on the risk premium during the period of 1954-2017 overall. For the full period we

considered a period starting from 1954 here simply because the federal rate series, which we analyse together with later, is available only from 1954. Including the period of 1950-1954 does not change the conclusion. Over the periods of 1963-1976 and 2003-2017 the risk-return relationship is strongly positive, and in the later period the estimate is statistically significant at 1% level. In fact, the estimated risk averse parameter β is 1.41294 (with $t = 2.7317$) on the last two subperiods combined (i.e. the period of around past 30 years), revealing evidence of strongly positive and statistically significant risk-return relation in the recent time after 1990. We may compare these results with the findings of Pagan and Hong (1990, page 61), where they reached a different conclusion with Monthly CRSP data over 1953-1984. They reported the estimated coefficient for conditional variance of -0.87 (with $t = -0.35$). The estimated risk aversion parameter β using our conditional expectations estimates over 1953-1984 is -0.07418 (with $t = -0.0790$), reporting a negative but much weaker risk-return relation.

To investigate how the analysis we adopt in our method may have made any difference, we repeat the same step above by computing nonparametric estimates with using only one lag as conditioning variable. This is to replicate what was done in the papers computed fitted means and variances based on nonparametric regression approaches, e.g. Pagan and Hong (1990). The local constant estimation is done with the standard Gaussian kernel and the bandwidth chosen via cross-validation. We denote by those fitted conditional mean and variance $(\widetilde{\mu}_t, \widetilde{\sigma}_t^2)$, and report the least squares estimates for the parameters α and β below in Table 4.

Table 4 reveals a very strong and persistent negative risk-return relation throughout all time periods. Over the full period, the estimated risk aversion parameter β is around -3.53, and this is statistically significant at 1% level based on Newey-West

Table 4: Estimated parameters obtained using $(252 \cdot \widetilde{\mu}_t, 252 \cdot \widetilde{\sigma}_t^2)$
Full (1954-2017) and Sub Periods

	Full	1954-1963	1963-1976	1976-1990	1990-2003	2003-2017
α	0.14977	0.15097	0.15647	0.11693	0.10808	0.19830
(t)	$(4.9068)^{**}$	$(2.0964)^{*}$	$(4.0584)^{**}$	$(2.6367)^{**}$	$(2.2444)^{**}$	$(2.3897)^{**}$
β	-3.33228	-3.14701	-4.30003	-2.38936	-0.77965	-5.13079
(t)	$(-2.2834)^{**}$	(-0.8642)	$(-2.0728)^{*}$	(-1.0904)	(-0.3297)	(-1.4274)

standard errors. This result implies that the conclusion Pagan and Hong (1990) obtained may have been influenced by the fact that they conditioned only on small, fixed

lags (when forming the nonparametric estimates). In other words, incorporating further information that are neglected in estimating conditional expectations has clearly led to some new empirical findings. This provides explanations to the conjecture Pagan and Hong (1990) raised in their paper.

5.2 Time variation and counter-cyclicity in risk aversion

Meanwhile, the results in Table 3 suggest that the risk-return relationship is strongly time-varying. In particular, over the subperiod 1976-1990 the estimate of β was significantly lower than the other periods. To take a closer look, we conducted a rolling regressions analysis. We set the rolling window to be 4000, roughly a quarter of the number of whole sample, and start estimating β from 1958:09:18. That is, we use conditional expectations estimates over 1958:09:18-1974:12:04 to estimate β for date 1974:12:05, and roll forward the window by one every time. The window size is deliberately chosen to be different from the size of 5 subperiods; this is to check if our previous results in Table 2 are driven by a particular choice of sample size. The time series of estimated parameter β shown below in Figure 5 provides an evidence that investor's average risk aversion has been varying over time.

Furthermore, we observe that interestingly, the time series of risk aversion tends to move in the opposite direction to the federal funds rate[7] f_t, which is a proxy for the business cycle fluctuations, see Figure 5. In fact, the sample correlation between $\widehat{\beta}_t$ and f_t turns out to be -0.5673, implying that the risk aversion exhibits a counter-cyclical behaviour. Also, in Figure 6 we plot the time series of quarterly Sharpe ratio and the designated recession periods by the NBER. The blue line is the ratio computed using our estimates $(\widehat{\mu}, \widehat{\sigma}^2)$, and the red line is the one computed using the standard nonparametric method $(\widetilde{\mu}, \widetilde{\sigma}^2)$. The shadings show that blue line rises over the period of recession in general, which is a finding that is consistent with Lettau and Ludvigson (2010). Note that the red line does not behave as expected in most cases and therefore does not quite capture counter-cyclicity.

[7]Data taken from Federal Reserve Bank of St. Louis http://fred.stlouisfed.org

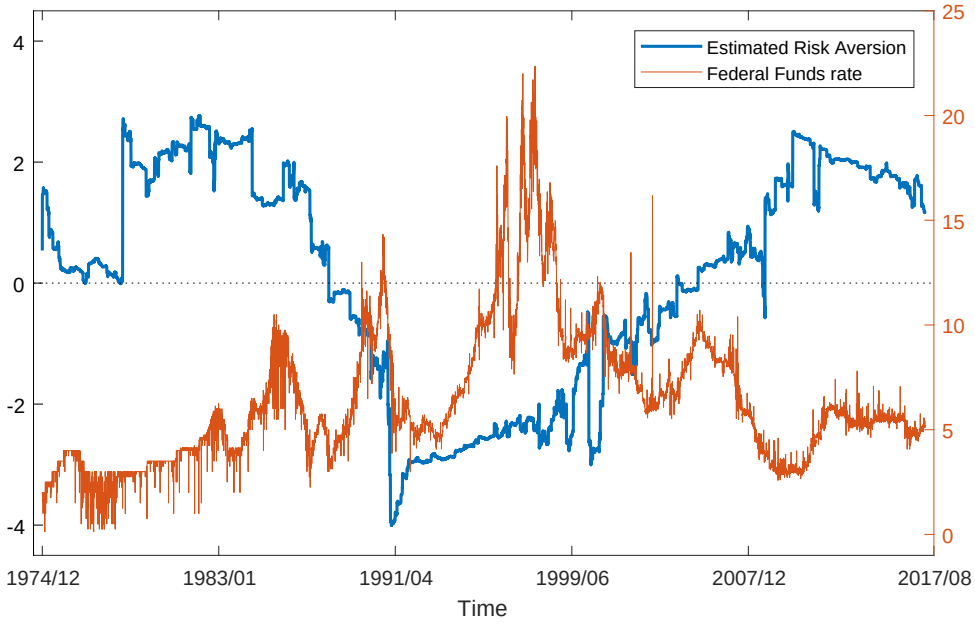

Figure 5: Estimated risk-return tradeoff and the federal funds rate

These findings are consistent with what is suggested and widely discussed in the finance literature, for example Antell and Vaihekoski (2016), Campbell and Cochrane (1999), Bliss and Panigirtzoglou (2004), Smith and Whitelaw (2009), Bollerslev, Gibson and Zhou (2011), and Guo, Wang and Yang (2013).

As noted in Mehra (2012), empirical evidence for the financial theory suggesting counter-cyclical risk return tradeoff is rather scarce and limited. Cohn et al. (2015) wrote, "*A key ingredient of many popular asset pricing models is that investors exhibit countercyclical risk aversion, which helps explain major economic puzzles such as the strong and systematic variation in risk premiums over time and the high volatility of asset prices. There is, however, surprisingly little evidence for this ...*"

Our findings on the time series dynamics of risk return tradeoff and their link with the macroeconomy add a supporting empirical evidence to this issue. We reiterate that when standard nonparametric method is employed, these evidence is not well revealed. This potential improvements in the econometric analysis are possibly attributable to the extended flexibility and the inclusion of otherwise neglected information in our method.

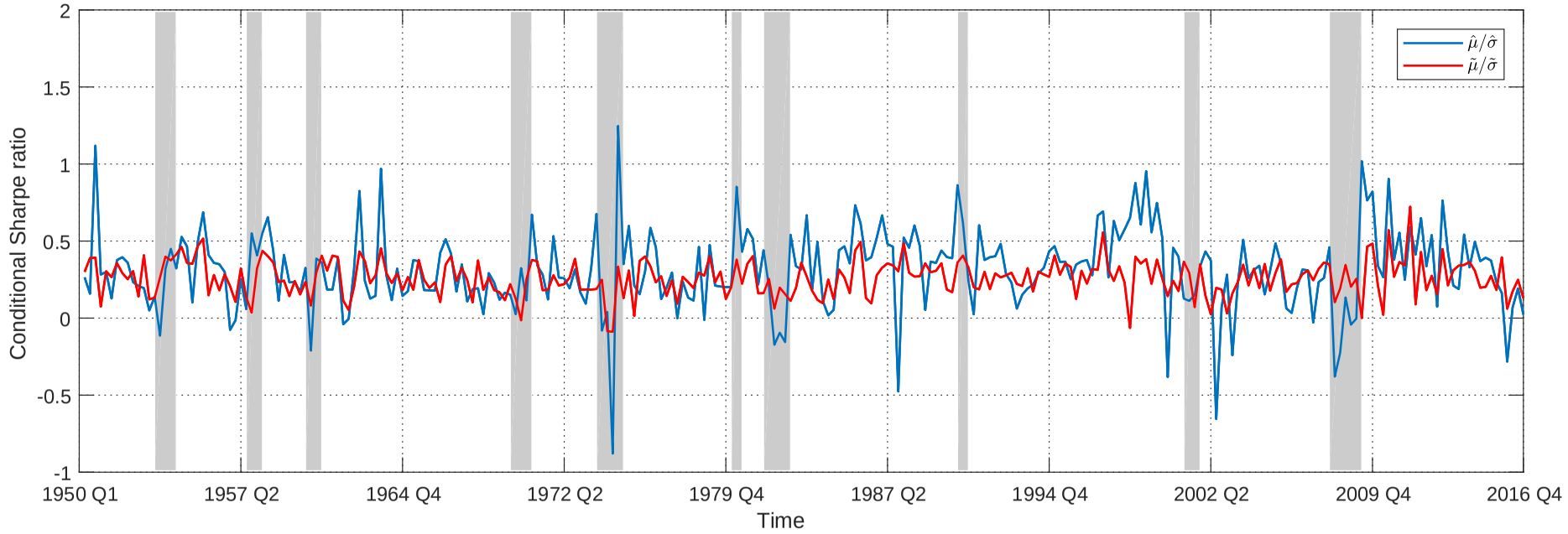

Figure 6: Quarterly sharpe ratio (1950-2016) computed based on $(\widehat{\mu}, \widehat{\sigma})$ and $(\widetilde{\mu}, \widetilde{\sigma})$, respectively.
The shades are the recession periods designated by the National Bureau of Economic Research (NBER).

5.3 Practical methods for choosing bandwidth

Lastly, we discuss some possible ways for selecting the bandwidth in practice. A natural choice would be direct extensions of two most extensively adopted approaches in the multivariate nonparametric regression, namely, a rule-of-thumb and cross-validation, Green and Silverman (1993), Fan and Gijbels (1996).

We first consider a heuristic plug-in method for obtaining a rule-of-thumb bandwidth. For simplicity, we suppose that the regressors X are i.i.d. Gaussian distributed. This way the small ball probability takes a simple form as discussed in previous sections. Further, this bypasses the need to estimate the $C_{\mathcal{A}}$ term, a function of spectral density representing the degree of dependence between regressors. With the uniform kernel supported on $[0,1]$ i.e. $\lambda = 1$, the asymptotic mean squared error of our estimator is then given by

$$\text{AMSE}(\widehat{m}) = h^{2\beta}\left(\sum_{j=1}^{\infty} c_j j^{p\beta}\right)^2 + \frac{\sigma^2(x)}{nh^{\frac{1-p}{2p-1}}\exp\big(-\kappa_0' h^{-\frac{2}{2p-1}}\big)}. \tag{50}$$

Denote by $\mathcal{C}$ the squared term in (50), and let $\beta = 1$. Now, differentiating (50) with respect to h and equating it to zero we have

$$\begin{aligned}
&\frac{\partial\{\text{AMSE}(\widehat{m})\}}{\partial h} = 2\mathcal{C}h + \frac{\sigma^2(x)}{n(2p-1)}\cdot e^{\kappa_0 h^{-\frac{2}{2p-1}}}\left[(p-1)h^{-\frac{p}{2p-1}} - 2\kappa_0 h^{-\frac{p+2}{2p-1}}\right] = 0 \\
\Leftrightarrow\ &\frac{2n(2p-1)\mathcal{C}h}{\sigma^2} = \exp\big(\kappa_0 h^{-\frac{2}{2p-1}}\big)\cdot\left[2\kappa_0 h^{-\frac{p+2}{2p-1}} - (p-1)h^{-\frac{p}{2p-1}}\right] \\
\Leftrightarrow\ &\frac{14n\cdot\mathcal{C}}{\sigma^2} = h^{-1}\exp\big(3.605h^{-2/7}\big)\left[7.21h^{-6/7} - 3h^{-4/7}\right]
\end{aligned} \tag{51}$$

where in the last line we substituted $p = 4$ and $\kappa_0 = C^{**} \approx 3.605$ (follows from a straightforward computation; see definitions in Section 4.2.2). As we can substitute the sample variance $\widehat{\sigma}^2$ into σ^2, it now suffices to replace the squared term $\mathcal{C}$ with a suitable estimate.

To proceed, we impose a further model assumption and suppose $m(x) = \sum_{j=1}^{\infty}\alpha_j|x_j|$ and $|\alpha_j| \le c_j$ ($= C\theta^j$ for some $0 < \theta < 1$ and constant C[8]). In this case, Assumption B7 in Section 2.4.2:

$$\big|m(x) - m(x')\big| \le \sum_{j=1}^{\infty} c_j\big|x_j - x_j'\big|$$

is satisfied via the reverse triangle inequality. A heuristic idea is to choose some C

[8]This is a reasonable assumption because $\text{cov}(Y_t, Y_{t-k}) = O(k^{-ck})$ for some c under Assumption A2 and by Davydov's inequality for covariance of mixing sequences.

and θ in such a way that c_j's bound statistically significant estimates of α_j's. Fitting the linear model upto lag 15, say: $Y_t = \sum_{j=1}^{15} \alpha_j |x_j| + \varepsilon_t$, the estimates of α_j at lags $1, 2, 9, 15$ are given by $0.058, 0.0443, 0.0435$ and 0.027, respectively. Therefore, we could let $C = 0.1$ and $\theta = 0.95$ for example; substituting these values back into the first order condition (51) yields

$$\frac{14n}{\widehat{\sigma}^2}\left(\frac{1}{10}\sum_{j=1}^{\infty}(0.95)^j j^4\right)^2 = u^{7/2}e^{3.605u}\Big[7.21u^3 - 3u^2\Big],$$

where $u = h^{-2/7}$, $\widehat{\sigma}^2 \approx 9.3557 \times 10^{-5}$ and $n = 16820$. Numerical approximation via Matlab yields $h = 0.000312$.

An alternative approach would be the cross-validation, where we search for the bandwidth that minimises the mean squared leave-one-out residuals:

$$g(h) := \frac{1}{n}\sum_{t=1}^{n}[Y_t - \widehat{m}_{h,-t}(X_t)]^2, \tag{52}$$

where $\widehat{m}_{h,-t}(X_t)$ is the estimate obtained by ignoring t^{th} sample. The result, as illustrated in Figure 5, suggests $h = 0.00035$ and $h = 0.000125$ for $p = 4$ and $p = 12$, respectively. These bandwidths are the ones we used earlier in the example in this section. Note that Yao and Tong (1998) proposed an different leave-one-out method for choosing optimal bandwidth for dependent data. When we applied their cross-validation for the data we considered previously however, we noticed that it suggests way lower optimal bandwidths, and the standard approach gives a more reasonable result. This might be because the returns data we consider is almost uncorrelated but nonlinearly dependent.

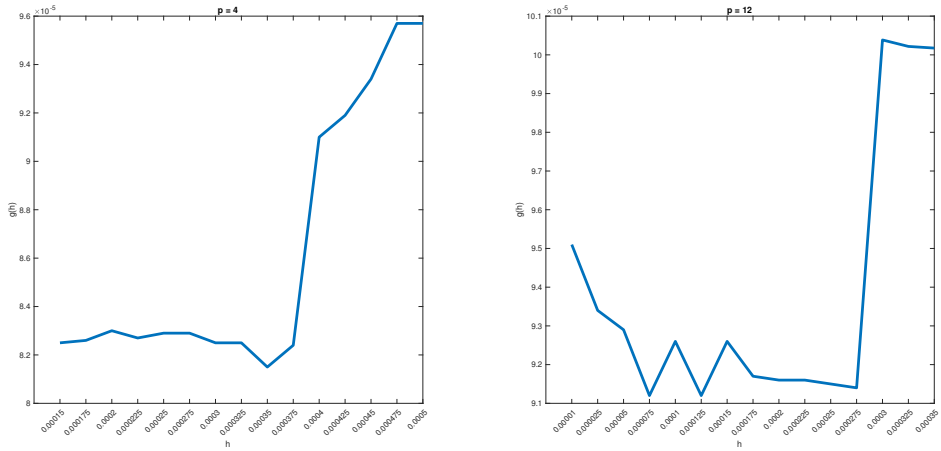

Figure 7: Cross validation choice of bandwidth given $p = 4$ and $p = 12$

6 Some concluding remarks

In this paper we studied the nonparametric estimation problem of the infinite order regression. While we answered several open questions raised in the literature, there are some remaining questions that we leave for future studies from a methodological point of view. First, it is not clear how the conclusions we obtained will be changed when the marginal bandwidth is set to decay in a way other than polynomially. It is also a non-trivial question whether the geometric mixing condition in the uniform consistency result could be relaxed to allow weaker dependence of the data. Furthermore, as Linton and Sancetta (2009) pointed out, it may be possible to achieve algebraic convergence rates for some restricted class of functions. For example, we conjecture that given the additive regression model: $\mathrm{E}(Y|X = x) = m(x) = \sum_{j=1}^{\infty} m_j(x_j)$ where $x = (x_j)_j \in \mathbb{R}^{\infty}$, the rate for estimating $m_j(\cdot)$ and $m(\cdot)$ would be the same, just as it was proven to be so in the d-dimensional case (i.e. $m_j(\cdot) = 0$, $\forall j > d \in \mathbb{Z}_+$) by Stone (1985). In this paper, we were unable to find an answer to this question, although we found the existence of the curse of infinite dimensionality under a wide range of frameworks we considered. We leave the question for future study.

Lastly, it would be interesting to come up with a practical way to choose the parameter p or more generally the rate at which the bandwidths expand in the order of the covariates. This should relate to the rate of decay of influence (mixing in the autoregression case) that prevails, and perhaps this can be addressed by using tools from the estimation of memory properties. An alternative approach is to us a penalization method combined with cross-validation, namely, the Bandwidth-LASSO method that adds the penalty $\lambda \sum_i |\phi_i^{-1}|$ to the objective function (52). The positive numbers ϕ_i are the bandwidth weights defined previously in Section 2.4 and in Assumption B8. The resulting choice of $\{\phi_i^{-1}\}$ will contain many zeros (infinite smoothing of one covariate) depending on the tuning parameter λ, which would give a much more parsimonious representation. The properties of this method will be investigated in the sequel.

Other quantities of interest in prediction such as the conditional median or mode can also be studied. This could be done via nonparametrically estimating the conditional distribution $P(Y \leq y|X = \cdot) = E(1\{-\infty, y](Y)|X = \cdot)$, but would necessarily require a slightly different set of assumptions. It is also quite easy to bring finite dimensional predictors into the theory separately. For example, one may want to allow for slow time variation whereby t/T becomes an additional covariate and the regression function is $m(x, u)$ with $u \in [0, 1]$ and $x \in \mathbb{R}^{\infty}$. In this case we modify the estimator of (18) by introducing a multiplicative kernel of the form $k_b(u - t/T)$, where b is a bandwidth and k is a symmetric probability density function.

7 Appendix: Proofs of the main results

7.1 Proof of Theorem 1

PROOF. From the decomposition in (21):

$$\widehat{m}(x) - m(x) = \frac{E\widehat{m}_2(x) - m(x)}{\widehat{m}_1(x)} + \frac{\widehat{m}_2(x) - E\widehat{m}_2(x)}{\widehat{m}_1(x)} - \frac{m(x)[\widehat{m}_1(x) - 1]}{\widehat{m}_1(x)},$$

we see that it suffices to show $E\widehat{m}_2(x) - m(x) \to 0$ and $\widehat{m}_2(x) - E\widehat{m}_2(x) \to^P 0$, since $\widehat{m}_1(x) \to^P 1$ will then follow from the latter and complete the proof.

We first consider the "bias component". It is straightforward to see

$$\begin{aligned} E\widehat{m}_2(x) - m(x) &= E\left(\frac{1}{nEK_1} \sum_{t=1}^{n} K_t Y_t - m(x) \right) \\ &= \frac{1}{EK_1} EK_1 Y_1 - \frac{EK_1}{EK_1} m(x) = \frac{1}{EK_1} E\Big[E\Big[\big(Y_1 - m(x)\big) K_1 \Big| X \Big] \Big] \\ &= \frac{1}{EK_1} E\Big[\Big[m(X) - m(x) \Big] K_1 \Big] \leq \sup_{u \in \mathcal{E}(x, \lambda \underline{h})} \big| m(u) - m(x) \big| \longrightarrow 0 \end{aligned} \tag{53}$$

as $n \to \infty$, where K_t is the shorthand notation for $K(\|H^{-1}(x - X_t)\|)$ and $\mathcal{E}(x, \lambda \underline{h})$ is the infinite dimensional hyperellipsoid centred at $x = (x_j)_j \in \mathbb{R}^\infty$ with semi-axes h_j in each direction as introduced in the main text before. The second equality is justified by stationarity that is preserved under measurable transformations, and the last inequality is due to compact support of the kernel and continuity of the regression operator at x (Assumption B1).

The next step concerns with the latter 'variance component' $\widehat{m}_2 - E\widehat{m}_2$. We show its mean-squared convergence to zero. Writing

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n} \sum_{t=1}^{n} \frac{1}{EK_1} \Big\{ K_t Y_t - E(K_t Y_t) \Big\} =: \frac{1}{n} \sum_{t=1}^{n} Q_{nt}, \tag{54}$$

we remark that the arguments to follow depend upon the temporal dependence structure of Q_{nt}. In the static regression case, Q_{nt} is a measurable function of $Y_t, X_{1t}, X_{2t}, \ldots$, and hence inherits their joint dependence structure. That is, Q_{nt} is arithmetically α-mixing with the rate specified in A1. In the dynamic regressions case (which covers the autoregression framework), the dependence of Q_{nt} is defined via K_t which is near epoch dependent on (Y_t, V_t) as specified in Assumption A2; this bypasses the issue of Q_{nt} being dependent upon infinite past of Y_t and/or V_t. We proceed with these two cases separately.

CASE 1: STATIC REGRESSION. Clearly, it is sufficient to prove $\text{var}(\widehat{m}_2 - E\widehat{m}_2) \to 0$ for the mean squared convergence. Since Q_{nt} is stationary over time we have

$$\begin{aligned}\text{var}(\widehat{m}_2 - E\widehat{m}_2) &= \frac{1}{n^2}\sum_{t=1}^{n}\text{var}(Q_{nt}) + \frac{2}{n^2}\sum\sum_{1\le i<j\le n}\text{cov}\big(Q_{ni}, Q_{nj}\big) \qquad (55)\\ &= \frac{1}{n}\text{var}(Q_{n1}) + \frac{2}{n^2}\sum\sum_{1\le j-i<n}\text{cov}\big(Q_{ni}, Q_{nj}\big)\\ &= \frac{1}{n}\text{var}(Q_{n1}) + \frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\cdot\text{cov}\big(Q_{n1}, Q_{n,s+1}\big) =: A_1 + A_2. \qquad (56)\end{aligned}$$

Now, by (9), (11) and Assumption A it follows that

$$\begin{aligned}A_1 &= \frac{1}{nE^2K_1}\text{var}\bigg(K_1Y_1 - EY_1K_1\bigg) = \frac{\text{var}\,(K_1Y_1)}{nE^2K_1}\\ &\le \frac{EK_1^2Y_1^2}{nE^2K_1} = \frac{E(E(Y_1^2|X_1)K_1^2)}{nE^2K_1} \le \frac{C}{n\varphi_x(\lambda\underline{h})} \to 0 \quad \text{as} \quad n\to\infty. \qquad (57)\end{aligned}$$

We now move on to the second term A_2 and investigate the covariance term. Since measurable transformations of mixing variables preserve the mixing property, using Davydov's inequality, see Davydov (1968, Lemma 2.1) or Bosq (1996, Corollary 1.1), and stationarity we have

$$\big|\text{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \left|\text{cov}\left(Y_1\frac{K_1}{EK_1}, Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right| \le \frac{C\{E|Y_1K_1|^{2+\delta}\}^{\frac{2}{2+\delta}}}{\varphi_x(\underline{h}\lambda)^2\cdot s^{k\delta/(2+\delta)}}. \qquad (58)$$

In the meantime,

$$\begin{aligned}\big|\text{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| &= \left|\text{cov}\left(Y_1\frac{K_1}{EK_1}, Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right|\\ &\le \left|E\left(Y_1\frac{K_1}{EK_1}Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right| + \left|E\left(Y_1\frac{K_1}{EK_1}\right)E\left(Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right|\\ &\le \frac{C}{\varphi_x(\underline{h}\lambda)^2}\left|E\left(K_1K_{s+1}\right)\right| + \frac{C'}{E^2K_1}\left|E\left(K_1\right)E\left(K_{s+1}\right)\right|\\ &\le \frac{C}{\varphi_x(\underline{h}\lambda)^2}\cdot\psi_x(\lambda\underline{h};1,s+1) + C' \le C'' \qquad (59)\end{aligned}$$

by stationarity, law of iterated expectation, boundedness of regression function, and Assumption B6, B5 (along with the upper bound $\psi(\lambda\underline{h};1,s+1)$ of EK_1K_{s+1} obtained as a direct consequence of B5 following similar arguments used for Lemma 1).

With reference to (58) and (59), we take some increasing sequence $u_n \to \infty$ such that $u_n = o(n)$, and write

$$\sum_{s=1}^{n-1} \big|\mathrm{cov}\big(Q_{n1},Q_{n,s+1}\big)\big| = \sum_{s=1}^{u_n-1} \big|\mathrm{cov}\big(Q_{n1},Q_{n,s+1}\big)\big| + \sum_{s=u_n}^{n-1} \big|\mathrm{cov}\big(Q_{n1},Q_{n,s+1}\big)\big|$$
$$\leq C''\big(u_n - 1\big) + \sum_{s=u_n}^{n-1} \frac{Cs^{-k\delta/(2+\delta)}}{\varphi_x(\underline{h}\lambda)^2} = O\bigg(u_n + \frac{u_n^{-k\delta/(2+\delta)+1}}{\varphi_x(\underline{h}\lambda)^2}\bigg), \qquad (60)$$

which is $O\big(\varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)}\big)$ upon choosing $u_n \sim \varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)}$.

Consequently, since $k \geq 2(2+\delta)/\delta$ it follows that

$$A_2 := \frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\cdot\mathrm{cov}\big(Q_{n1},Q_{n,s+1}\big) = \frac{2}{n}\sum_{s=1}^{n-1}\bigg(1-\frac{s}{n}\bigg)\cdot\mathrm{cov}\big(Q_{n1},Q_{n,s+1}\big)$$
$$= O\big(n^{-1}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)} + n^{-2}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)}\big)$$
$$= O\big(n^{-1}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)}\big) = o(1) \qquad (61)$$

by Assumption B2, and the desired result is obtained.

Case 2: Dynamic Regression.[9] We return back to (54):

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\bigg\{K_tY_t - E(K_tY_t)\bigg\} =: \frac{1}{n}\sum_{t=1}^{n}Q_{nt}. \qquad (62)$$

In this framework $K_t = K(\|H^{-1}(x-X_t)\|)$ is a (measurable) function of $(Y_{t-1}, Y_{t-2}, \ldots)$. Despite loosing the mixing property, K_t inherits stationarity of the mixing process $\{Y_t\}$. We write $K_{t,(r)} = \Psi(Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-r+1}) = E(K_t|Y_t, \ldots, Y_{t-r+1})$ with r as in Assumption A2, and the measurable map Ψ. Then, $K_{t,(r)}$ preserves the mixing dependence structure of Y_t with mixing coefficient $\alpha(\ell-(r-1))$ since $\sigma(K_{s,(r)};\ s \geq t+\ell) \subset \sigma((Y_s, \ldots, Y_{s-r+1});\ s \geq t+\ell) = \sigma(Y_s;\ s \geq t+\ell-(r-1))$.

Now write

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\bigg[K_{t,(r)}Y_t - E\big(K_{t,(r)}Y_t\big)\bigg] + \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\bigg[K_tY_t - K_{t,(r)}Y_t\bigg]$$
$$+\ \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\bigg[E\big(K_{t,(r)}Y_t\big) - E(K_tY_t)\bigg] = R_1 + R_2 + R_3, \qquad (63)$$

[9]For the sake of notational simplicity, we will write the proofs for the dynamic regression framework in terms of its autoregressive special case throughout the appendix. That is, some lags of the response variable Y_t here possibly represent lagged covariate V_t.

and first consider the last term R_3. Fix some increasing sequence $q = q_n \to \infty$, and write $Y_{t,L} := Y_t 1\{|Y_t| \le q\}$ and $Y_{t,U} = Y_t 1\{|Y_t| > q\}$. Then

$$\begin{aligned} EY_t K_{t,(r)} &= EY_t K\big(\|H^{-1}(x - X_t)\|\big) - EY_{t,U} K\big(\|H^{-1}(x - X_t)\|\big) \\ &\quad + EY_{t,L} K_{t,(r)} - EY_{t,L} K\big(\|H^{-1}(x - X_t)\|\big) \\ &\quad + EY_{t,U} K_{t,(r)} = D_1 + D_2 + D_3. \end{aligned} \tag{64}$$

The second part of D_1 is given by

$$\begin{aligned} &EY_{t,U} K\big(\|H^{-1}(x - X_t)\|\big) \le E|Y_t| 1_{\{|Y_t|>q\}} K\big(\|H^{-1}(x - X_t)\|\big) \\ &\le q^{-(\delta+1)} E|Y_t|^{2+\delta} 1_{\{|Y_t|>q\}} K_t \le C q^{-(\delta+1)} E|Y_t|^{2+\delta} 1_{\{|Y_t|>q\}} = o(q^{-(\delta+1)}) \end{aligned} \tag{65}$$

because $1_{\{|Y_t|>q\}} = o(1)$ as $n \to \infty$. Following similar arguments on D_3 we have $D_1 + D_3 = EY_t K_t + o(q^{-(\delta+1)})$. So we are now left with the middle term D_2:

$$D_2 \le E\left|Y_{t,L}\right| \left|K_t - K_{t,(r)}\right| = O\left(q\sqrt{v_2(r_n)}\right) \tag{66}$$

by Hölder's inequality. Therefore, from (64), (65) and (66) we see that

$$\begin{aligned} R_3 &= \frac{1}{nEK_1} \sum_{t=1}^{n} \Big[EK_{t,(r)} Y_t - E(K_t Y_t) \Big] \\ &= o\left(\frac{q^{-(\delta+1)}}{\varphi_x(\lambda \underline{h})} \right) + O\left(\frac{q\sqrt{v_2(r_n)}}{\varphi_x(\lambda \underline{h})} \right), \end{aligned} \tag{67}$$

and upon choosing $q = (\varphi_x(\underline{h}\lambda)/n)^{-1/(2(\delta+1))}$ we have $o(\varphi^{-1} q^{-(\delta+1)}) = o(\varphi^{-1}(\varphi/n)^{1/2}) = o(n^{-1/2}\varphi^{-1/2}) = o(1)$. Furthermore,

$$\begin{aligned} O\left(\frac{1}{\varphi_x(\underline{h}\lambda)} q\sqrt{v_2(r_n)} \right) &= O\left(\frac{1}{\varphi_x(\underline{h}\lambda)} \cdot \left(\frac{\varphi_x(\underline{h}\lambda)}{n} \right)^{-1/(2(\delta+1))} \sqrt{v_2(r_n)} \right) \\ &= O\left(\frac{\sqrt{v_2(r_n)}}{[\varphi_x(\underline{h}\lambda)]^{(2\delta+3)/(2\delta+2)} n^{-1/(2(\delta+1))}} \right) = o(1) \end{aligned} \tag{68}$$

by Assumption A2, yielding $R_3 = o(1)$ and consequently $R_2 = o_p(1)$.

As for the first term that remains,

$$
\begin{aligned}
R_1 &= \frac{1}{n}\sum_{t=1}^{n}\left[\frac{K_{t,(r)}Y_t - E(K_tY_t)}{EK_1}\right] + \frac{1}{n}\sum_{t=1}^{n}\left[\frac{E(K_tY_t) - E(K_{t,(r)}Y_t)}{EK_1}\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} E\big(Q_{nt}|Y_t, Y_{t-1}, \ldots, Y_{t-r+1}\big) - R_3 \\
&= \frac{1}{n}\sum_{t=1}^{n} Q_{nt,(r)} + o\left(\frac{q^{-(\delta+1)}}{\varphi_x(\underline{h}\lambda)}\right) + O\left(\frac{\sqrt{v_2(r_n)}}{[\varphi_x(\underline{h}\lambda)]^{(2\delta+3)/(2\delta+2)}n^{-1/(2(\delta+1))}}\right). \qquad (69)
\end{aligned}
$$

Since $Q_{nt,(r)}$ is α-mixing, we can work with the first term by following similar arguments in the regression case. Specifically, due to boundedness of the kernel and the mixing properties, the bound in (58) can be constructed. As for the constant bound constructed in (59), we rewrite

$$
\begin{aligned}
\frac{\operatorname{cov}\big(Y_1K_{1,(r)}, Y_{s+1}K_{s+1,(r)}\big)}{\varphi_x(\lambda\underline{h})^2} &= \frac{\operatorname{cov}\big(Y_1[K_{1,(r)} - K_1],\ Y_{s+1}[K_{s+1,(r)} - K_{s+1}]\big)}{\varphi_x(\lambda\underline{h})^2} \\
&+ \frac{\operatorname{cov}\big(Y_1[K_{1,(r)} - K_1], Y_{s+1}K_{s+1,(r)}\big)}{\varphi_x(\lambda\underline{h})^2} \\
&+ \frac{\operatorname{cov}\big(Y_1, Y_{s+1}[K_{s+1,(r)} - K_{s+1}]\big)}{\varphi_x(\lambda\underline{h})^2} + \frac{\operatorname{cov}\big(Y_1K_1, Y_{s+1}K_{s+1}\big)}{\varphi_x(\lambda\underline{h})^2} \\
&= \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 + \mathcal{G}_4.
\end{aligned}
$$

The fourth term $\mathcal{G}_4 \leq C$ by (59). Further,

$$
\begin{aligned}
\mathcal{G}_1 \leq &\left|\frac{E(Y_1Y_{s+1}[K_{1,(r)} - K_1][K_{s+1,(r)} - K_{s+1}])}{\varphi_x(\lambda\underline{h})^2}\right| \\
&+ \left|\frac{E(Y_1[K_{1,(r)} - K_1]) \cdot E(Y_{s+1}[K_{s+1,(r)} - K_{s+1}])}{\varphi_x(\lambda\underline{h})^2}\right| \leq C'\frac{v_2(r)}{\varphi_x(\lambda\underline{h})^2} \to 0
\end{aligned}
$$

by Assumption B6 and by the fact that

$$
\left(\frac{\sqrt{v_2(r_n)}}{\varphi_x(\underline{h}\lambda)}\right) \leq \left(\frac{\sqrt{v_2(r_n)}}{\varphi_x(\underline{h}\lambda)}\right) \cdot (n/\varphi)^{1/(2\delta+2)} \longrightarrow 0
$$

by (19) in Assumption A2. Following similar steps it can be shown that $\mathcal{G}_2$ and $\mathcal{G}_3$ converge to zero.

Now choosing an increasing sequence $u_n \sim [\varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)} + r_n] \to \infty$ such that

$r_n/u_n = o(1)$, we see that (ignoring the array notation in $Q_{nt,(r)}$ for simplicity)

$$\sum_{s=1}^{n-1}\big|\mathrm{cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\big| = \sum_{s=1}^{u_n-1}\big|\mathrm{cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\big| + \sum_{s=u_n}^{n-1}\big|\mathrm{cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\big|$$
$$\leq C\big(\varphi_x(\underline{h}\lambda)^{-\frac{2(2+\delta)}{(k\delta)}} + r_n\big) + \sum_{s=u_n}^{n-1}\frac{C(s-r_n+1)^{-k\delta/(2+\delta)}}{\varphi_x(\underline{h}\lambda)^2} = O\Big(\varphi_x(\underline{h}\lambda)^{-\frac{2(2+\delta)}{(k\delta)}}\Big),$$

since the mixing coefficient for $Q_{nt,(r)}$ denoted $\alpha'(n)$ is given by $\alpha(n-(r-1))$ for $n \geq r$. It now follows by the same arguments in (61) that the first term in (69) converges to zero, yielding $R_1 = o_p(1)$, which is the result we desired. ■

7.2 Proof of Theorem 2 and 3

Proof of Theorem 2 and 3. We start by recalling the bias component (53). Additional assumptions B7, B8 and D3 allow us to proceed further as follows:

$$\begin{aligned}
\mathcal{B}_n(x) &= E\widehat{m}_2(x) - m(x) = E\bigg(\frac{1}{nEK_1}\sum_{t=1}^{n} K_tY_t - m(x)\bigg)\\
&= \frac{1}{EK_1}EK_1Y_1 - \frac{EK_1}{EK_1}m(x) = \frac{1}{EK_1}E\bigg[E\Big[\big(Y_1 - m(x)\big)K_1\Big|X\Big]\bigg]\\
&= \frac{1}{EK_1}E\Big[\Big[m(X) - m(x)\Big]K_1\Big] \leq \sup_{u\in\mathcal{E}(x,\lambda\underline{h})}\big|m(u) - m(x)\big|\\
&\leq \sup_{u\in\mathcal{E}(x,\lambda\underline{h})}\sum_{j=1}^{\infty} c_j\big|u_j - x_j\big|^{\beta} = \sum_{j=1}^{\infty} c_j(\lambda h\phi_j)^{\beta} = h^{\beta}\bigg(\lambda^{\beta}\sum_{j=1}^{\infty} c_j j^{p\beta}\bigg) < \infty. \qquad (70)
\end{aligned}$$

Now rewriting the decomposition (21) as

$$\begin{aligned}
&\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\\
&\qquad = \frac{\mathcal{B}_n(x)\cdot\big[1 - \widehat{m}_1(x)\big]}{\widehat{m}_1(x)} + \frac{\widehat{m}_2(x) - E\widehat{m}_2(x) - m(x)\big[\widehat{m}_1(x) - 1\big]}{\widehat{m}_1(x)},
\end{aligned}$$

and noting that $\widehat{m}_1(x) \to^p 1$ (an immediate consequence of Theorem 1), we see that it suffices to derive the limiting distribution of

$$\begin{aligned}
&\widehat{m}_2(x) - E\widehat{m}_2(x) - m(x)[\widehat{m}_1(x) - 1]\\
&\qquad = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[K_tY_t - m(x)K_t - E(K_tY_t) + m(x)EK_t\Big] =: \frac{1}{n}\sum_{t=1}^{n} R_{nt}. \qquad (71)
\end{aligned}$$

In the rest of the proof, the way how we construct the general CLT *under Assumption*

A1 is quite similar to the proofs of theorems in Masry (2005), where asymptotic normality is established in a functional context for mixing data sample. For completeness of the proof however, we will go over some of the main arguments; some relatively less important details will only be briefly sketched to prevent being repetitive.

By Assumption B6, B8, D3, and the law of iterated expectations, the asymptotic variance of the triangular array R_{nt} is given by

$$\begin{aligned}
\mathrm{var}(R_{nt}) &= \frac{\mathrm{var}[K_t(Y_t - m(x))]}{E^2 K_1} \\
&= \frac{1}{E^2 K_1}\left\{ E\Big[K_t(Y_t - m(x))\Big]^2 - E^2\big[K_t(Y_t - m(x))\big] \right\} \\
&\simeq \frac{1}{E^2 K_1}\left\{ E\big[\sigma^2(X)K_1^2\big] + E\bigg(\Big[m(X) - m(x)\Big]^2 K_1^2\bigg) \right\} \\
&= \frac{1}{E^2 K_1}\left\{ \sigma^2(x) E K_1^2 + E\bigg(\Big[\sigma^2(X) - \sigma^2(x)\Big] K_1^2\bigg) + o(1) E K_1^2 \right\} \\
&= \frac{E K_1^2}{E^2 K_1}(\sigma^2(x) + o(1)) \simeq \frac{\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2}. \qquad (72)
\end{aligned}$$

Using the latter assumption of B9 and Assumptions B, and following similar arguments as in the above and those in the proof of Theorem 1, it can be readily shown that the covariance term is of negligible order, which together with (72) shows (29).

Meanwhile, under Assumption D2 the small ball probability can be written in terms of the centered small deviation and $p^*(\cdot)$, the Radon-Nikodym derivative of the induced probability measure P_{z-Z} with respect to P_Z:

$$\begin{aligned}
\varphi_x(\lambda\underline{h}) &= P\big(X \in \mathcal{E}(x, \lambda\underline{h})\big) \\
&= P\bigg(\sum_{j=1}^{\infty} j^{-2p}\big(x_j - X_j\big)^2 \le h^2\lambda^2 \bigg) = P\Big(\|z - Z\| \le h\lambda \Big) \\
&= \int_{B(0,h\lambda)} dP_{z-Z}(u) = \int_{B(0,h\lambda)} p^*(u)\, dP_Z(u) \\
&\simeq p^*(0) \cdot P(\|Z\| \le h\lambda) = p^*(0) \times P\left(\sum_{j=1}^{n} j^{-2p} X_j^2 \le h^2\lambda^2 \right). \qquad (73)
\end{aligned}$$

Given that the fourth moment of X_j is finite by Assumptions C, the latter probability in (73) can be explicitly specified by substituting $r = h^2\lambda^2$, $A = 2p$, and $a = 2p/(2p-1)$ in Proposition 4.1 of Dunker et al. (1998) for the i.i.d. case. When the marginal regressors are dependent as in Assumption C, the small ball probability can be specified (by letting $r = h^2\lambda^2 C_{\mathcal{A}}^{-2}$ and leaving the others the same) in view of Theorem 1.1 of

Hong, Lifshits and Nazarov (2016). In the general i.i.d. case (under Assumptions A1 and independence across marginal covariates) we have

$$\frac{\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2} = \frac{1}{\phi(h)} \cdot \frac{\sigma^2(x)\xi_2}{p^*(0)\xi_1^2} \cdot \frac{C^* C_\ell}{\lambda^{\frac{1+2\rho p}{2p-1}}},$$

where $\phi(h) = h^{(1+2\rho p)/(2p-1)} \exp\{-C^{**}(\lambda h)^{-2/(2p-1)}\}$ and

$$C_\ell = \lim_{h\to 0}\left[\ell^{-1/2}\left(h^{-\frac{4p}{2p-1}}\right)\right] \qquad C^* = \frac{(2\pi)^{(1+2p\rho)}(2p-1)}{\Gamma^{-1}(1-\rho)\cdot(2p)^{\frac{2p(\rho+2)-1}{2p-1}}} \cdot \zeta^{\frac{2p(1+\rho)}{2p-1}}.$$

$\Gamma(\cdot)$ is the Gamma function, ξ_1 and ξ_2 are the constants specified in (12), and λ is the upper bound of the support of the kernel. The constants for the dependent case can be specified similarly.

In constructing the central limit theorem we consider the normalized statistic $R^*_{nt} := \sqrt{\phi(h)} \cdot R_{nt}$ and derive the limiting distribution of $(1/\sqrt{n}) \cdot R^*_{nt}$. We shall prove under Assumption A2 as it involves some further arguments, without which the proof just serves as the proof under Assumption A1. We make use of the standard Bernstein's blocking method and partition $\{1, \ldots, n\}$ by $2k(= 2k_n \to \infty)$ number of blocks of two different sizes that alternate (hereafter referred to as the "big" and "small" blocks) and lastly a single block (the "last block") that covers the remainder. The size of the alternating blocks is given by a_n and b_n respectively, where the one for the "big-blocks" a_n is set to dominate that for the "small-blocks" b_n in large sample, i.e. $b_n = o(a_n)$. Specifically, take $k_n = \lfloor n/(a_n + b_n) \rfloor$ and $a_n = \lfloor \sqrt{n\phi(h)}/q_n \rfloor$, where $q_n \to \infty$ is a sequence of integer; it then clearly follows that $a_n/n \to 0$ and $a_n/\sqrt{n\phi(h)} \to 0$. We also assume $(n/a_n) \cdot \alpha^*(b_n) = (n/a_n) \cdot \alpha(b_n - r + 1) \to 0$, where α^* is the mixing coefficient of $R^*_{nt,(r)} = E(R^*_{nt} | \mathcal{F}^{t-1}_{t-r+1})$.

By construction above we can write $\sqrt{n}^{-1} \sum_{t=1}^n R^*_{nt}$ as the sum of the groups of big-blocks $\mathcal{B}$, small-blocks $\mathcal{S}$ and the remainder block $\mathcal{R}$ defined as

$$\begin{aligned}
\mathcal{B} &:= \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \Xi_{1,j} = \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R^*_{nt} \right) \\
\mathcal{S} &:= \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \Xi_{2,j} = \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R^*_{nt} \right) \\
\mathcal{R} &:= \frac{1}{\sqrt{n}} \Xi_{3,j} = \frac{1}{\sqrt{n}} \left(\sum_{t=k(a+b)+1}^{n} R^*_{nt} \right).
\end{aligned}$$

The aim is to show that the contributions from the small and the last remaining block are negligible, and that the big-blocks are asymptotically independent. Consider the big blocks $\mathcal{B}$. Given r as in Assumption A2, and $R^*_{nt,(r)} = E(R^*_{nt}|Y_t, \ldots, Y_{t-r+1})$,

$$\mathcal{B} = \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R^*_{nt,(r)} \right) + \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+1}^{j(a+b)+a} \left[R^*_{nt,(r)} - R^*_{nt} \right] \right) = \mathcal{Q}_1 + \mathcal{Q}_2.$$

As for the second term, consider

$$\begin{aligned}
\frac{1}{\sqrt{n}} E\mathcal{Q}_2 &\leq \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} E\left| R^*_{nt,(r)} - R^*_{nt} \right| \\
&= \frac{1}{EK_1} \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} E\left| K_t Y_t - Y_t E(K_t | Y_t, Y_{t-1}, \ldots, Y_{t-r+1}) \right| \\
&\leq \frac{1}{\sqrt{n}} \frac{1}{\varphi_x(\underline{h}\lambda)} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} E|Y_t||K_t - K_{t,(r)}| \\
&\leq \frac{1}{\sqrt{n}} \frac{1}{\varphi_x(\underline{h}\lambda)} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} \left(E|Y_t|^2 \right)^{1/2} \left(E|K_t - K_{t,(r)}|^2 \right)^{1/2} \\
&\leq C \cdot \frac{1}{\sqrt{n}} k_n a_n \frac{\sqrt{v_2(r_n)}}{\varphi_x(\lambda \underline{h})} = O\left(\frac{\sqrt{n \cdot v_2(r_n)}}{\varphi_x(\lambda \underline{h})} \right) = o(1),
\end{aligned}$$

which implies that $\sqrt{n}^{-1} \mathcal{Q}_2 = o_p(1)$.

We now show asymptotic independence of terms in $\mathcal{Q}_1$, on noting that $\Xi'_{1,j}s$ are independent if for all real t_j

$$\left| E\left[\sum_{j=0}^{k-1} \exp\left(\mathrm{i} t_j \Xi_{1,j} \right) \right] - \prod_{j=0}^{k-1} E\left[\exp(\mathrm{i} t_j \Xi_{1,j}) \right] \right| \tag{74}$$

is zero. Applying the Volkonskii-Rozanov inequality (see Fan and Yao (2003, page 72) for example), it can be shown that (74) is bounded above by $C(n/a_n) \cdot \alpha(b_n - r + 1) \to 0$, implying asymptotic independence.

Moving on to the small blocks, due to stationarity we have

$$\begin{aligned}
\text{var}\,(\mathcal{S}) &= \frac{1}{n}\text{var}\left(\sum_{j=0}^{k-1}\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^*\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\text{var}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^*\right) + \frac{1}{n}\sum\sum_{j\neq l}^{k-1}\text{cov}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^*, \sum_{s=l(a+b)+a+1}^{(l+1)(a+b)} R_{ns}^*\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\left(b_n\text{var}(R_{nt}^*) + \sum_{t\neq l}^{b_n}\text{cov}(R_{nt}^*, R_{nl}^*)\right) + \frac{1}{n}\sum\sum_{j\neq l}^{k-1}\sum_{i,j=1}^{b_n}\text{cov}\big(R_{n,i+w_j}^*, R_{n,r+w_l}^*\big) \\
&= Q_1 + Q_2 + Q_3.
\end{aligned}$$

where $w_j = j(a+b)+a$.

Regarding the first term, similar arguments used in deriving (72) yield

$$Q_1 = \frac{1}{n}k_n b_n \frac{\left[\varphi_x(\underline{h}\lambda)^{1/2}\right]^2\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2} = \frac{k_n b_n \sigma^2(x)\xi_2}{n\xi_1^2} \longrightarrow 0 \tag{75}$$

because $k_n b_n/n \sim b_n/(a_n+b_n) \to 0$. Now moving on to Q_2 and Q_3, the sum of covariances can be dealt with in the same manner as we did for the variance using (72), so $Q_2 \to 0$. Similarly for Q_3, implying $\text{var}(\mathcal{S}) \to 0$ as desired. Convergence result for the remainder $\mathcal{R}$ can be established similarly, and is bounded by $C(a_n+b_n)/n \to 0$.

The results above suggest that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} R_{nt}^* = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^*\right) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j + o_p(1), \tag{76}$$

and the desired result holds in view of (62) and the CLT for triangular array upon checking the Lindeberg condition (which is omitted here due to its similarity with Masry (2005, page 174-175)). Corollary 2 now follows because

$$\begin{aligned}
\sqrt{n\phi(h)}\left(\frac{\widehat{m}-m-\mathcal{B}_n}{\sqrt{n\phi(h)}\Delta_n}\right) &= \frac{\sqrt{n}\frac{1}{n}\sum_{t=1}^{n} R_{nt}^*}{\sqrt{\frac{1}{n}\sum_t \widehat{R}_{nt}^{*,2}}} = \frac{\frac{1}{\sqrt{n}}\sum_{t=1}^{n} R_{nt}^*}{\sqrt{\frac{1}{n}\sum_t R_{nt}^{*,2}} + o_p(1)} \\
&= \frac{\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^* + o_p(1)}{\sqrt{\frac{1}{n}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^*\right)^2} + o_p(1)} = \frac{\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j + o_p(1)}{\sqrt{\frac{1}{n}\sum_{j=0}^{k-1}\eta_j^2} + o_p(1)} \Longrightarrow N(0,1)
\end{aligned} \tag{77}$$

by Theorem 4.1 of de la Peña et al. (2009), since the denominator converges in probability to a strictly positive quantity ($\sigma^2(x)\xi_2/\xi_1^{-2}$), and that η_j belongs to the

domain of attraction of a normal distribution by definition and (76). ■

7.3 Proof of Lemmas 1 and 2

PROOF. Lemma 1 is a straightforward extension of Lemma 4.3 and 4.4 of Ferraty and Vieu (2006), and hence is omitted. Lemma 2 can be shown by noting that for each n the τ_n-dimensional polyhedron $D := \{w = (w_i)_{i\leq\tau} \in \mathbb{R}^\tau, |w_i| \leq \lambda\}$ can be covered by $([2\lambda\sqrt{\tau}/\varepsilon + 1])^\tau$ number of balls of radius ε, see Chaté and Courbage (1997), and then following the arguments of the proof of Theorem 2 in Jia et al. (2003). ■

7.4 Proof of Theorem 4 and 5

PROOF OF THEOREM 4. In the sequel, we omit the subscript τ in the notations for truncated regressor and its estimator, i.e. $m_\tau(\cdot)$ and $\widehat{m}_\tau(\cdot)$ for notational simplicity. As before, we start from the decomposition (21):

$$\widehat{m}(x) - m(x) = \frac{1}{\widehat{m}_1(x)}\left(\Big[\widehat{m}_2(x) - E\widehat{m}_2(x)\Big] + \Big[E\widehat{m}_2(x) - m(x)\Big] - m(x)\Big[\widehat{m}_1(x) - 1\Big]\right).$$

We recall from (73) that $\varphi_x(\lambda\underline{h}) \sim \varphi(\lambda\underline{h})$. Further, notice that the small deviation for the truncated regressor $X = (X_1, \ldots, X_\tau, 0, 0, \ldots)$ denoted $\varphi^{\mathcal{T}}(\lambda\underline{h})$ satisfies

$$\varphi(\lambda\underline{h}) = P\left(\sum_{j=1}^{\infty} j^{-2p} X_j^2 \leq h^2\right) \leq P\left(\sum_{j=1}^{\tau} j^{-2p} X_j^2 \leq h^2\right) = \varphi^{\mathcal{T}}(\lambda\underline{h}). \tag{78}$$

Note that as implicitly mentioned in the main text, (42) is meant to hold for $\varphi^{\mathcal{T}}(\lambda\underline{h})$.

In the first step of the proof we show

$$\sup_{x\in\mathcal{S}_\tau} \left|\widehat{m}_2(x) - E\widehat{m}_2(x)\right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \tag{79}$$

We cover the set $\mathcal{S}_\tau$ defined in (40) with $L = L(\mathcal{S}_\tau, \eta)$ number of balls of radius η denoted by I_k, each of which is centred at x_k, $k = 1, \ldots, L$. i.e. $S_\tau \subset \bigcup_{k=1}^{L_n} B(x_k, \eta)$.

Then it follows that

$$\begin{aligned}
\sup_{x\in\mathcal{S}_\tau}\left|\widehat{m}_2(x)-E\widehat{m}_2(x)\right| &= \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\left|\widehat{m}_2(x)-E\widehat{m}_2(x)\right| \\
&= \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\left|\widehat{m}_2(x)-\widehat{m}_2(x_k)+\widehat{m}_2(x_k)-E\widehat{m}_2(x_k)+E\widehat{m}_2(x_k)-E\widehat{m}_2(x)\right| \\
&\le \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\left|\widehat{m}_2(x)-\widehat{m}_2(x_k)\right|+\max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\left|E\widehat{m}_2(x_k)-E\widehat{m}_2(x)\right| \\
&\quad+\max_{1\le k\le L_n}\left|\widehat{m}_2(x_k)-E\widehat{m}_2(x_k)\right| \;=:\; R_1+R_2+R_3, \qquad (80)
\end{aligned}$$

where $\widehat{m}_2(x_k)=(nEK_1)^{-1}\sum_{t=1}^n Y_tK_{t,k}$ and $K_{t,k}=K(\|H^{-1}(x_k-X_t)\|)$.

We first consider R_1. By Lemma 1,

$$\begin{aligned}
R_1 &= \max_{1\le k\le L_n}\sup_{x\in I_k\cap S_\tau}\left|\widehat{m}_2(x)-\widehat{m}_2(x_k)\right| \\
&= \max_{1\le k\le L_n}\sup_{x\in I_k\cap S_\tau}\left|\frac{1}{nEK_1}\sum_{t=1}^n Y_tK\Big(\|H^{-1}(x-X_t)\|\Big)-Y_tK\Big(\|H^{-1}(x_k-X_t)\|\Big)\right| \\
&\le \max_{1\le k\le L_n}\sup_{x\in I_k\cap S_\tau}\frac{C}{n\varphi^{\mathcal{T}}(\lambda\underline{h})}\sum_{t=1}^n\left|Y_tK_t-Y_tK_{t,k}\right|\cdot 1_{\mathcal{E}(x,\lambda\underline{h})\cup\mathcal{E}(x_k,\lambda\underline{h})}(X_t).
\end{aligned}$$

Now because type-I kernels are Lipschitz continuous on $[0,\lambda]$, by the triangle inequality we have

$$R_1\le\frac{1}{n}\sum_{t=1}^n\frac{C'|Y_t|}{\varphi^{\mathcal{T}}(\underline{h}\lambda)}\eta h^{-1}\cdot 1_{\mathcal{E}(x,\lambda\underline{h})\cup\mathcal{E}(x_k,\lambda\underline{h})}(X_t)=:\frac{1}{n}\sum_{t=1}^n J_t,$$

where J_t is α-mixing under both assumptions A1′ and A2′ (with a different rate under A2′: $\alpha^*(n)=\alpha(n-\tau+1)$, where $\alpha(\cdot)$ is the mixing rate under A1′). Let $\eta=\log n/n^2$. Using Assumption B6 and the law of iterated expectations it is straightforward to see that

$$E|J_t|\le\frac{C\eta}{h}. \qquad (81)$$

Using Lemma 2 we can specify the Kolmogorov's entropy of S_τ for $\eta=\log n/n^2$:

$$\log L\left(S,\frac{\log n}{n^2}\right)=C\log\left[\left(\frac{2\lambda n^2}{\sqrt{\log n}}+1\right)^{\log n}\right]\sim\log n\times\log\left[\frac{2\lambda n^2}{\sqrt{\log n}}\right],$$

implying that the order of Kolmogorov's $\frac{\log n}{n^2}$ entropy is of order $(\log n)^2$.[10]

We now apply the Fuk-Nagaev inequality (see for example, Fuk and Nagaev (1971),

[10]Notice that in this case (42) is indeed satisfied with $\beta=1,p=4,\epsilon=1/4$, for example.

or Rio (2000)) for exponentially mixing variables in Merlevède, Peligrad and Rio (2011, 1.7) with $\varepsilon = \varepsilon_0[\log L\big(S, \frac{\log n}{n^2}\big)/(n\varphi(\lambda\underline{h}))]^{1/2}$ and $r = (\log L)/\varphi(\lambda\underline{h})$, where ε_0 is some positive constant. Since

$$s_n^2 := \sum_{t=1}^{n}\sum_{s=1}^{n} \operatorname{cov}(J_t, J_s) = O\big(n\varphi^{\mathcal{T}}(\lambda\underline{h})^{-1}\log n\big)$$

and the required tail condition holds, under Assumption A1′ we obtain

$$\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{t=1}^{n} J_t - EJ_t\right| > \varepsilon\right) \\
&\quad = P\left(\left|\sum_{t=1}^{n}(J_t - EJ_t)\right| > n\varepsilon_0\sqrt{\frac{\log L\big(S, \frac{\log n}{n^2}\big)}{n\varphi(\lambda\underline{h})}}\right) \\
&\quad \le\ 4\left(1 + \frac{n^2\varepsilon_0^2 \log L\big(S, \frac{\log n}{n^2}\big)}{16 r s_n^2 n\varphi(\lambda\underline{h})}\right)^{-\frac{r}{2}} + \frac{16C\sqrt{n\varphi(\lambda\underline{h})}}{\varepsilon_0\sqrt{\log L}}\exp\left(-\varsigma\left[\frac{\frac{1}{4}n\varepsilon_0\sqrt{\frac{\log L}{n\varphi(\lambda\underline{h})}}}{\log L/\varphi(\lambda\underline{h})}\right]^{\gamma}\right) \\
&\quad \le\ 4\left(1 + \frac{C\varepsilon_0^2\log n}{16r}\right)^{-\frac{r}{2}} + \frac{16C\sqrt{n\varphi(\lambda\underline{h})}}{\varepsilon_0\log n}\exp\left(-\varsigma\left[\frac{\varepsilon_0\sqrt{n\varphi(\lambda\underline{h})}}{4\log n}\right]^{\gamma}\right) \\
&\quad \le\ 4\left(1 + \frac{C\varepsilon_0^2\varphi(\lambda\underline{h})}{16\log n}\right)^{-\frac{\log L}{2\varphi(\lambda\underline{h})}} + \frac{16C}{\varepsilon_0}\left(\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right)\exp\left(-\varsigma\varepsilon_0^{\gamma}4^{-\gamma}\cdot\left[\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right]^{\gamma}\right) \\
&\quad \le\ 4\exp\left(-\frac{\varepsilon_0^2 C\log n}{32}\right) + \frac{16C}{\varepsilon_0}\left(\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right)\cdot e^{-C'(\sqrt{n\varphi}/\log n)} \longrightarrow 0, \qquad (82)
\end{aligned}$$

where $\varsigma > 1$ and $\gamma \ge 1$ are as defined in Section 2.4.4, by choosing ε_0 sufficiently large. In the last inequality we exploited the fact that $\log(1+\epsilon) = \epsilon + o(\epsilon^2)$ as $\epsilon \to 0$.

Under Assumption A2′, a penalty of $(-\log n)$ is incurred in the squared brackets in the inequalities above. This does not affect the conclusion (82) because $\tau = \log n \le (\log n)^2 \le \sqrt{n\varphi}/(\log n)^{1+\epsilon} \le \sqrt{n\varphi}/\log n$ by (42) in Assumption E.[11]
Therefore, in view of (81) it now follows that

$$\begin{aligned}
R_1 = \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\left|\widehat{m}_2(x) - \widehat{m}_2(x_k)\right| &\le O\left(\frac{\eta}{h}\right) + O_P\left(\sqrt{\frac{\log L\big(S, \frac{\log n}{n^2}\big)}{n\varphi(\lambda\underline{h})}}\right) \\
&= O\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) + O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\ \right). \qquad (83)
\end{aligned}$$

[11] To elaborate, this is due to the fact that $y\exp(-(y-g(y))) \to 0$ as $y\to\infty$, as long as $(y-g(y))$ tends to $+\infty$ as $y\to\infty$ at the speed strictly faster than $\log y$.

As for the second term R_2, we have

$$R_2 \leq \max_{1\leq k\leq L_n} \sup_{x\in I_k\cap\mathcal{S}_\tau} E\big|\widehat{m}_2(x) - \widehat{m}_2(x_k)\big| = O\left(\frac{\eta}{h}\right) = O\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \tag{84}$$

Next we move on to the last component:

$$R_3 = \max_{1\leq k\leq L_n} \big|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\big| =: \max_{1\leq k\leq L_n} \big|W_n(x_k)\big| \tag{85}$$

where

$$\begin{aligned} W_n(x) = \widehat{m}_2(x) - E\widehat{m}_2(x) &= \frac{1}{nEK_1}\sum_{t=1}^{n}\Big[Y_tK_t - EY_tK_t\Big] \\ &\leq \frac{C}{n\varphi^{\mathcal{T}}(\underline{h}\lambda)}\sum_{t=1}^{n}\Big[Y_tK_t - EY_tK_t\Big] = \frac{1}{n}\sum_{t=1}^{n}U_{nt} \end{aligned}$$

where $U_{nt} = (\varphi^{\mathcal{T}}(\underline{h}\lambda))^{-1}C(Y_tK_t - EY_tK_t)$.

By following similar arguments in the proof of Theorem 1, it can be readily seen that

$$s_n^2 = \sum_{t=1}^{n}\sum_{s=1}^{n}\operatorname{cov}(U_{nt}, U_{ns}) = O\left(n\varphi^{\mathcal{T}}(\underline{h}\lambda)^{-1}\right).$$

With the exponential tail condition in B4, we apply the same Fuk-Nagaev inequality for exponentially mixing sequences we referred to in the above. Writing $L_n := L\left(S, \frac{\log n}{n^2}\right)$ and taking $\varepsilon = \varepsilon_0[\log L\big(S, \frac{\log n}{n^2}\big)/(n\varphi(\lambda\underline{h}))]^{1/2}$ and $r = (\log n)^{2+\epsilon}/\varphi^{\mathcal{T}}(\lambda\underline{h})$, $\epsilon \in (0, 1/2)$ for some $\varepsilon_0 > 0$, under Assumption A1′ we have

$$\begin{aligned} P\left(\max_{1\leq k\leq L_n}|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)| > \varepsilon\right) &\leq L_n\cdot\sup_{x\in\mathcal{S}}P\left(|W_n(x)| > \varepsilon_0\sqrt{\frac{\log L_n}{n\varphi(\lambda\underline{h})}}\right) \\ &\leq L_n\cdot\sup_{x\in\mathcal{S}}P\left(\left|\sum_{t=1}^{n}U_{nt}\right| > n\varepsilon_0\sqrt{\frac{\log L_n}{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}\right) \\ &\leq L_n\cdot 4\left(1 + \frac{n^2\varepsilon_0^2\log L_n}{16rs_n^2n\varphi^{\mathcal{T}}(\lambda\underline{h})}\right)^{-\frac{r}{2}} \\ &\quad + \frac{16L_nCn\sqrt{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}{n\varepsilon_0\log n}\exp\left(-\varsigma\left\{\frac{\varepsilon_0\sqrt{n}\log n/\sqrt{\varphi^{\mathcal{T}}(\lambda\underline{h})}}{4(\log n)^{2+\epsilon}/\varphi^{\mathcal{T}}(\lambda\underline{h})}\right\}^{\gamma}\right) \\ &\leq L_n\cdot 4\left(1 + \frac{\varepsilon_0^2C\log L_n}{16(\log n)^{2+\epsilon}/\varphi^{\mathcal{T}}(\lambda\underline{h})}\right)^{-\frac{(\log n)^{2+\epsilon}}{2\varphi^{\mathcal{T}}(\lambda\underline{h})}} \end{aligned}$$

$$
\begin{aligned}
&+ \frac{16 L_n C}{\varepsilon_0} \frac{\sqrt{n\varphi^{\mathcal{T}}(\lambda \underline{h})}}{\log n} \exp\left(-\varsigma \frac{\varepsilon_0^{\gamma}}{4^{\gamma}} \left\{ \frac{\sqrt{n\varphi^{\mathcal{T}}(\lambda \underline{h})}}{(\log n)^{1+\epsilon}} \right\}^{\gamma} \right) \\
\leq\ & L_n \cdot 4 \exp\left(-\frac{\varepsilon_0^2 C \log L_n}{32} \right) + C L_n^2 \exp(-\varsigma \varepsilon_0^{\gamma} / 4^{\gamma} \log L) \\
\leq\ & 4 L_n^{-C\varepsilon_0^2/32} + C L_n^{-\frac{\varsigma \varepsilon_0}{4}+2}. \qquad (86)
\end{aligned}
$$

Here we used the fact that $\gamma \geq 1$ and (42) in Assumption E. Note that in the special case when the response Y_t is assumed to be bounded, the same result continues to hold with $\gamma_1 = \infty$ (so that $\gamma_2 = \gamma (\geq 1)$). Now noting that $\varsigma > 1$, by choosing ε_0 large enough it follows that

$$
R_3 = \max_{1 \leq k \leq L_n} \left| \widehat{m}_2(x_k) - E\widehat{m}_2(x_k) \right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda \underline{h})}} \right). \qquad (87)
$$

Same conclusion holds in the dynamic regression case (i.e. under Assumption A2′) because of the following reason. The penalty term (due to the penalised mixing rate) that incurs inside the curly bracket results in an additional multiplicative term of $\exp(-c(-\tau)) = \exp(c \log n) = n^c$ in the second term of the final bound in (86), where $c := \varsigma(\varepsilon_0/4)$ is fixed, and this diverges to infinity at the slower rate than $L_n^{(c-2)} = (n^2/\sqrt{\log n})^{\log n \cdot (c-2)}$.

Returning back to where we started, viewing $\widehat{m}_1(x)$ as a special case of $\widehat{m}_2(x)$ with $Y_t = 1\ \forall t$, we can repeat the above procedure, yielding (since $E\widehat{m}_1(x) = 1$)

$$
\sup_{x \in \mathcal{S}_\tau} \left| \widehat{m}_1(x) - 1 \right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda \underline{h})}} \right). \qquad (88)
$$

The proof is now complete in view of (79), (80), (83), (84), (87), (88), contributions from the bias component, and either Proposition 4.1 of Dunker, Lifshits and Linde (1998) under independence across marginal covariates, or Theorem 1.1 of Hong, Lifshits and Nazarov (2016) under general Assumption C. ■

Proof of Theorem 5. Given the extended moment condition upto $8 + \delta$, it is straightforward to see (from Theorem 1 and 2 & 3) the consistency of $\widehat{\sigma}^j(x_i)$ for $\sigma^j(x_i)$ for $j = 1, 2, 3, 4$ at every point of continuity x_i, and the asymptotic normality of $(\widehat{\mu}, \widehat{\sigma}^2)$ with limiting variance $\Omega(x_i)$.

Hence it suffices to show asymptotic independence of $\widehat{m}(x_i)$and $\widehat{m}(x_i')$ across i, where x_i and x_i' are continuity points of m such that $\|D^{-1}(x_i - x_i')\| > 0$. Following the notations of the proof of Theorem 2 and 3, the asymptotic covariance matrix is

given by $\text{Var}\big[(\sqrt{\phi(h)}/\sqrt{n})\sum_{t=1}^{n} R_{nt}\big]$, and

$$\text{Var}(R_{nt}) = \text{Var}\begin{pmatrix} \frac{1}{EK_{1,x}} \cdot K_{t,x}[Y_t - m(x)] \\ \frac{1}{EK_{1,x'}} \cdot K_{t,x'}[Y_t - m(x')] \end{pmatrix} = E\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \tag{89}$$

We know from Theorem 2 and 3 that as for $A_{11} \simeq \sigma^2(x)$ and $A_{22} \simeq \sigma^2(x')$. So we just consider the off-diagonal terms. Due to stationarity we see that

$$\begin{aligned}
&E\Big[K_{t,x}K_{t,x'}(Y_t - m(x))(Y_t - m(x'))\Big] \\
&= E\Big[K_{1,x}K_{1,x'}\Big\{Y_1 - m(X_1) + m(X_1) - m(x)\Big\}\Big\{Y_1 - m(X_1) + m(X_1) - m(x')\Big\}\Big] \\
&= E\Big[K_{1,x}K_{1,x'}(Y_t - m(X_1))(Y_t - m(X_1))\Big] + o(1) = E\Big[K_{1,x}K_{1,x'}\sigma^2(X_1)\Big] + o(1) \\
&\leq \sup_{u \in B(x,h) \cap B(x',h)} \sigma^2(u) E[K_{1,x'}K_{1,x}] \to 0
\end{aligned}$$

as $h \to 0$ since the kernels return 0 outside its compact support and $\|D^{-1}(x_i - x_i')\| > 0$. The desired result now directly follows via the delta method. ■

References

[1] Abel, A. B. (1988): "Stock prices under time-varying dividend risk : An exact solution in an infinite-horizon general equilibrium model," Journal of Monetary Economics, 22(3), 375–393.

[2] Andrews, D. W. K. (1984). *Non-Strong Mixing Autoregressive Processes*. Journal of Applied Probability, 21(4), 930-934.

[3] Andrews, D. W. K. (1995). *Nonparametric kernel estimation for semiparametric models*. Econometric Theory, 11(3), 560-586.

[4] Aneiros, G., Bongiorno, E. G., Cao, R. and Vieu, P. (eds.). (2017). *Functional Statistics and Related Fields*. New York: Springer.

[5] Antell, J. and Vaihekoski, M. (2016). *Countercyclical and Time-Varying Risk Aversion and the Equity Premium*. SSRN Working paper available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2753537.

[6] Azais, J. M. and Fort, J. C. (2013). *Remark on the finite-dimensional character of certain results of functional statistics*. Comptes Rendus Mathematique, 351(3), 139-141.

[7] Backus, K. and Gregory, A, W. (1993). *Theoretical Relations Between Risk Premiums and Conditional variances*. Journal of Business and Economic Statistics, 11(2), 177-185.

[8] Bali, T. G. and Peng, L. (2006). *Is there a Risk-Return Trade-Off? Evidence from High-frequency data*. Journal of Applied Econometrics, 21, 1169-1198.

[9] Bansal, R. and Yaron, A. (2004): *Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles*. Journal of Finance, 59(4), 1481-1509.

[10] Bierens, H. J. (1983). *Uniform consistency of kernel estimators of a regression function under generalized conditions*. Journal of the American Statistical Association, 78(383), 699-707.

[11] Bierens, H. J. (1987). *Kernel estimators of regression functions*. In Bewley, T. F. (ed.) *Advances in econometrics: Fifth world congress, 99-144, Vol. 1*, Cambridge: Cambridge University Press.

[12] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987). *Regular variation*. Cambridge: Cambridge University Press.

[13] Billingsley, P. (1968). *Convergence of Probability Measures*. New York: John Wiley.

[14] Bliss, R. R. and Panigirtzoglou, N. (2004). *Option-implied risk aversion estimates*. Journal of Finance, 59(1), 407-446.

[15] Bollerslev, T. (1996). *Modeling and pricing long memory in stock market volatility*. Journal of Econometrics, 73(1), 151-184.

[16] Bollerslev, T., Gibson, M. and Zhou, H. *Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities*. Journal of Econometrics, 160(1), 235-245.

[17] Borovkov, A. A. and Ruzankin, P. S. (2008). *On small deviations of series of weighted random variables*. Journal of Theoretical Probability, 21(3), 628-649.

[18] Bosq, B. (1996). *Nonparametric statistics for stochastic processes: estimation and prediction*. New York: Springer-Verlag.

[19] Boudoukh, J., Richardson, M. and Whitelaw, R. F. (1997). *Nonlinearities in the Relation between the Equity Risk Premium and the Term Structure*. Management Science, 43(3), 371-385.

[20] Bradley, R. C. (2005). *Basic properties of strong mixing conditions. A survey and some open questions*. Probability Surveys, 2(2), 107-144.

[21] Campbell, J. Y. (1987). *Stock Returns and the Term Structure*. Journal of Financial Economics, 18(2), 373-399.

[22] Campbell, J. Y. and Cochrane, J. H. (1999). *By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior*. Journal of Political Economy, 107(2), 205-251.

[23] Campbell, J. Y. and Hentschel, L. (1992). *No News is Good News: An Asymmetric Model of Changing Volatility in Stock Returns*. Journal of Financial Economics, 31(3), 281-318.

[24] Chaté, H. and Courbage, M. (1997). *Special issue on lattice dynamics*. Physica D, 103, 1-611.

[25] Chen, J., Li, D., Linton, O. B. and Lu, Z. (2017). *Semiparametric Ultra-High Dimensional Model Averaging of Nonlinear Dynamic Time Series*. Journal of the American Statistical Association, Forthcoming.

[26] Chen, R. and Tsay, R. S. (1993). *Nonlinear additive ARX models*. Journal of the American Statistical Association, 88(423), 955-967.

[27] Chen, X. and Christensen, T. M. (2015). *Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions*. Journal of Econometrics, 188(2), 447-465.

[28] Chou, R., Engle, R. F. and Kane, A. (1992). *Measuring risk aversion from excess returns on a stock index*. Journal of Econometrics, 52(1-2), 201-224.

[29] Christensen, B. J., Dahl, C. M. and Iglesias, E. M. (2012). *Semiparametric inference in a GARCH-in-mean model*. Journal of Econometrics, 167(2), 458-472.

[30] Cohn, A., Engelmann, J., Fehr, E. and Maréchal, M. (2015). *Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals*. American Economic Review, 105(2), 860-885.

[31] Conrad, C. and Mammen, E. (2008), *Nonparametric regression on latent covariates with an application to semiparametric GARCH-in-Mean models*. Discussion Paper No. 473, University of Heidelberg, Department of Economics.

[32] Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

[33] Davydov, Y. A. (1968). *Convergence of distributions generated by stationary stochastic processes*. Theory of Probability and Its Applications, 13(4), 691-696.

[34] Delsol, L. (2009). *Advances on asymptotic normality in non-parametric functional time series analysis*. Statistics, 43(1), 13-33.

[35] Devroye, L. P. (1978). *The uniform convergence of the Nadaraya-Watson regression function estimate*. Canadian Journal of Statistics, 6(2), 179-191.

[36] Devroye, L. (1981). *On the almost everywhere convergence of nonparametric regression function estimates*. The Annals of Statistics, 9(6), 1310-1319.

[37] Doukhan, P. (1994). *Mixing*. New York: Springer.

[38] Doukhan, P. and Wintenberger, O. (2008). *Weakly dependent chains with infinite memory*. Stochastic Processes and their Applications, 118(11), 1997-2013.

[39] Duflo, M. (1997). *Random iterative models*. Berlin: Springer-Verlag.

[40] Dunker, T., Lifshits, M. A. and Linde, W. (1998). *Small deviation probabilities of sums of independent random variables*. In: Eberlein, E. (ed.), High dimensional probability, volume 43 of Progress in Probability., Birkhauser, Basel, 59-74.

[41] Escanciano, J.C., Pardo-Fernández, J.C. and Van Keilegom, I. (2017). *Semiparametric estimation of risk return relationships*. Journal of Business and Economic Statistics, 35(1), 40-52.

[42] Fan, J. and Masry, E. (1992). *Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes*. Journal of Multivariate Analysis, 43(2), 237-271.

[43] Fan, J. (1990). *A remedy to regression estimators and nonparametric minimax efficiency*. Technical Report 161, Department of Statistics, University of North Carolina at Chapel Hill.

[44] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. New York: Chapman and Hall.

[45] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

[46] Feller, W. (1971). *Introduction to Probability Theory and Its Applications*, Vol. 2. New York: Wiley.

[47] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2010). *Rate of uniform consistency for nonparametric estimates with functional variables*. Journal of Statistical Planning and Inference, 140(2), 335-352.

[48] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011). *Kernel regression with functional response*. Electronic Journal of Statistics, 5, 159-171.

[49] Ferraty, F. and Romain, Y. (2010). *The Oxford Handbook of Functional Data Analysis*. New York: Oxford University Press.

[50] Ferraty, F. and Vieu, P. (2002). *The functional nonparametric model and application to spectrometric data*. Computational Statistics, 17(4), 545-564.

[51] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: Theory and Practice*. New York: Springer.

[52] French, K. R., Schwert, G. W. and Stambaugh, R. F. (1987) *Expected Stock Returns and Volatility*. Journal of Financial Economics, 19(1), 3-29.

[53] Fuk, D. K. and Nagaev, S. V. (1971). *Probability inequalities for sums of independent random variables*. Theory of Probability and its applications, 16(4), 643-660.

[54] Gao, F., Hannig, J. and Torcaso, F. (2003). *Comparison Theorems for Small Deviations of Random Series*. Electronic Journal of Probability, 8(21), 1-17.

[55] Gennotte, G. and Marsh, T. A. (1993). *Variations in economic uncertainty and risk premiums on capital assets*. European Economic Review, 37(5), 1021-1041.

[56] Geenens, G. (2011). *Curse of dimensionality and related issues in nonparametric functional regression*. Statistics Surveys, 5, 30-43.

[57] Ghysels, E., Santa-Clara, P. and Valkanov, R. (2005). *There is a Risk-Return Trade-Off After All*. Journal of Financial Economics, 76(3), 509-548.

[58] Giraitis, L., Leipus, R. and Surgailis, D. (2008). *ARCH*(∞) *and long memory properties*. In: Andersen, T. G., Davis, R. A., Kreiß, J., Mikosch, T. (Eds.), Handbook of Financial Time Series. Berlin: Springer-Verlag.

[59] Glosten, L., Jagannathan, R. and Runkle, D. E. (1993). *On The Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks*. Journal of Finance, 48(5), 1779-1801.

[60] Götze, F. and Hipp, C. (1994). *Asymptotic distribution of statistics in time series*. The Annals of Statistics, 22(4), 2062-2088.

[61] Greblicki, W. and Krzyzak, A. (1980). *Asymptotic properties of kernel estimates of a regression function*. Journal of Statistical Planning and Inference, 4(1), 81-90.

[62] Green, P. J. and Silverman, B. (1993). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. New York: Chapman and Hall.

[63] Guo, H., Wang, Z. and Yang, J. (2013). *Time-Varying Risk-Return Trade-off in the Stock Market*. Journal of Money, Credit and Banking, 45(4), 623-650.

[64] Guo, H. and Whitelaw, R. (2006). *Uncovering the Risk-Return Relation in the Stock Market*. Journal of Finance, 61(3), 1433-1463.

[65] Härdle, W. K. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

[66] Harvey, C. (2001). *The Specification of Conditional Expectations*. Journal of Empirical Finance, 8(5), 573-638.

[67] Hong, S. Y., Lifshits, M. and Nazarov. A. (2016). *Small deviations in L_2-norm for Gaussian dependent sequences*. Electronic Communications in Probability, 21(41), 1-9.

[68] Ibragimov, I. A. (1962). *Some limit theorems for stationary processes*. Theory of Probability & Its Applications, 23, 291-300.

[69] Ibragimov, I. A. and Linnik, Y. B. (1971). *Independent and Stationary Sequences of Random variables*. Groningen: Wolters-Noordhoff.

[70] Jia, Q., Zhou, S. and Yin. F. (2003). *Kolmogorov entropy of global attractor for dissipative lattice dynamical systems*. Journal of Mathematical Physics, 44, 5804-5801.

[71] Kandel, S. and Stambaugh, R. F. (1990). *Expectations and Volatility of Consumption and Asset Returns*. Review of Financial Studies, 3(2), 207-232.

[72] Karamata, J. (1933). *Sur un mode de croissance regulire. Theoremes fondamentaux* (in French). Bulletin de la Societe Mathematique de France, 61, 55-62.

[73] Kara-Zaïtri, L., Laksaci, A., Rachdi, M. and Vieu, P. (2017). *Uniform in the smoothing parameter consistency results in functional regression.* Functional Statistics and Related Fields, 161-167, Springer.

[74] Kudraszow, N. L. and Vieu, P. (2013). *Uniform consistency of kNN regressors for functional variables.* Statistics and Probability Letters, 83(8), 1863-1870.

[75] Lettau, M. and Ludvigson, S. (2003). *Measuring and modeling variation in the risk-return tradeoff.* In: Ait-Sahalia, Y., Hansen, L. (Eds.), Handbook of Financial Econometrics. Amsterdam: North-Holland.

[76] Li, W. V. (2012). *Small value probabilities in analysis and mathematical physics.* Presented at the Arizona School of Analysis and Mathematical Physics, Tucson, Arizona, United States in March 15, 2012. Retrieved from the link `http://math.arizona.edu/~mathphys/school_2012/WenboLi.pdf`.

[77] Li, W. V. and Shao, Q. M. (2001). *Gaussian processes: inequalities, small ball probabilities and applications.* Handbook of Statistics, 19, 533-598.

[78] Linton, O. B. (2009). *Semiparametric and Nonparametric ARCH Modeling.* In: Andersen, T. G., Davis, R. A., Kreiß, J., Mikosch, T. (Eds.), Handbook of Financial Time Series. Berlin: Springer-Verlag.

[79] Linton, O. B. and Perron, B. (2003). *The Shape of the Risk Premium: Evidence From a Semiparametric Generalized Autoregressive Conditional Heteroscedasticity Model.* Journal of Business and Economic Statistics, 21(3). 354-367.

[80] Linton, O. B. and Sancetta, A. (2009). *Consistent estimation of a general nonparametric regression function in time series.* Journal of Econometrics, 152(1), 70-78.

[81] Lu, Z. (2001). *Asymptotic normality of kernel density estimators under dependence.* Annals of the Institute of Statistical Mathematics, 53(3), 447-468.

[82] Ludvigson, S. and Ng, S. (2007) *The Empirical Risk-Return Relation: A Factor Analysis Approach.* Journal of Financial Economics, 83(1), 171-222.

[83] Lundblad, C. (2005). *The Risk-Return Tradeoff in the Long Run: 1836-2003.* Journal of Financial Economics, 85(1), 123-150.

[84] Mammen E. (1991) *Nonparametric Curve Estimation and Simple Curve Characteristics*. In: Roussas G. (eds) Nonparametric Functional Estimation and Related Topics. NATO ASI Series (Series C: Mathematical and Physical Sciences), vol 335. Springer, Dordrecht.

[85] Mas, A. (2012). *Lower bound in regression for functional data by small ball probability representation in Hilbert space*, Electronic Journal of Statistics, 6, 1745-1778.

[86] Masry, E. (2005). *Nonparametric regression estimation for dependent functional data: asymptotic normality*. Stochastic Processes and their Applications, 115(1), 155-177.

[87] Masry, E. and Fan, J. (1997). *Local polynomial estimation of regression function for mixing processes*. Scandinavian Journal of Statistics, 24(2), 1965-1979.

[88] Mehra, R. (2012). *Consumption-Based Asset PRicing Models'*. Annual Review of Financial Economics, 4(1), 385-409.

[89] Merlevède, F., Peligrad, M., Rio, E. (2011). *A Bernstein type inequality and moderate deviations for weakly dependent sequences*. Probability Theory and Related Fields, 151(3-4), 435-474.

[90] Merton, R. C. (1973). *An Intertemporal Capital Asset Pricing Model*. Econometrica, 41(5), 867-887.

[91] Nadaraya, E. A. (1964). *On estimating regression*. Theory of Probability & Its Applications, 9(1), 141-142.

[92] Nadaraya, E. A. (1970). *Remarks on non-parametric estimates for density functions and regression curves*. Theory of Probability & Its Applications, 15(1), 134-137.

[93] Nelson, D. (1991). *Conditional Heteroscedasticity in Asset Returns: a New Approach*. Econometrica, 59(2), 347-370.

[94] Olver, F. W. J., Lozier, D. W., Boisvert, R. F. and Clark, C. W. (2010). *NIST Handbook of Mathematical Functions*. New York: Cambridge University Press.

[95] Pagan, A. R. and Hong, Y. S. (1990). *Nonparametric Estimation and the Risk Premium*. In W. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics*. Cambridge University Press.

[96] Pagan, A. R. and Ullah, A. (1988). *The econometric analysis of models with risk terms*. Journal of Applied Econometrics, 3(2), 87-105.

[97] Pagan, A. R. and Ullah, A. (1999). *Nonparametric econometrics*. Cambridge: Cambridge University Press.

[98] Parzen, E. (1962). *On estimation of a probability density function and mode*. Annals of Mathematical Statistics, 33(3), 1065-1076.

[99] Pástor, L., Sinha, M. and Swaminathan, B. (2008). *Estimating the Intertemporal Risk-Return Tradeoff Using the Implied Cost of Capital*. Journal of Finance, 63(6), 2859-2897.

[100] de la Peña, V. H., Lai, T. L. and Shao, Q. M. (2009). *Self-normalized processes: Limit theory and Statistical Applications*. New York: Springer.

[101] Phillips, P. C. and Park, J. Y. (1998). *Nonstationary density estimation and kernel autoregression*. Cowles Foundation discussion paper.

[102] Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. New York: Springer.

[103] Rio, E. (2000). *Theorie asymptotique des processus alé atoires faiblement dependants* (in French). Berlin: Springer Verlag.

[104] Robinson, P. M. (1983). *Nonparametric estimators for time series*. Journal of Time Series Analysis, 4(3), 185-207.

[105] Rosenblatt, M. (1956). *A central limit theorem and a strong mixing condition*. Proceedings of the National Academy of Sciences, 42(1), 43-47.

[106] Roussas, G. G. (1989). *Consistent regression estimation with fixed design points under dependence conditions*. Statistics & Probability Letters, 8(1), 41-50.

[107] Roussas, G. G. (1990). *Nonparametric regression estimation under mixing conditions*. Stochastic Processes and Their Applications, 36(1), 107-116.

[108] Schuster, E. F. (1972). *Joint asymptotic distribution of the estimated regression function at a finite number of distinct points*. Annals of Mathematical Statistics, 43(1), 84-88.

[109] Scruggs, J. (1998). *Resolving the Puzzling Intertemporal Relation Between the Market Risk Premium and the Conditional Market variance: A Two-Factor Approach*. Journal of Finance, 53(2), 575-603.

[110] Scruggs, J. and Glabadanidis, P. (2003). *Risk Premia and the Dynamic covariance Between Stock and Bond Returns*. Journal of Financial and Quantitative Analysis, 38(2), 295-316.

[111] Smith, D. R. and Whitelaw, R. (2009). *Time-varying risk aversion and the risk-return relation*. SSRN Working paper available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1663542

[112] Stone, C. (1980). *Optimal rates of convergence for nonparametric estimators*. The Annals of Statistics, 8(6), 1348-1360.

[113] Stone, C. (1982). *Optimal global rates of convergence for nonparametric regression*. The Annals of Statistics, 10(4), 1040-1053.

[114] Stone, C. J. (1985). *Additive Regression and Other Nonparametric Models*. The Annals of Statistics, 13(2), 689-705.

[115] Sytaya, G. N. (1974). *On certain asymptotic representations for a Gaussian measure in Hilbert space* (in Russian). Theory of Random Process, 2, 93-104.

[116] Veronesi, P. (2000): "How Does Information Quality Affect Stock Returns?," The Journal of Finance, 55(2), 807–837.

[117] Watson, G. S. (1964). *Smooth regression analysis*. Sankhya Series A, 26(4), 359-372.

[118] Whitelaw, R. F. (1994). *Time variations and covariations in the Expectation and Volatility of Stock Market Returns*, Journal of Finance, 49(2), 515-541.

[119] Whitelaw, R. F. (2000). *tock Market Risk and Return: An Equilibrium Approach*, Review of Financial Studies, 13(3), 521-547.

[120] Wu, W. B. (2011). *Asymptotic theory for stationary processes*. Statistics and Its Interface, 4, 207-226.

[121] Yao, Q and Tong, H. (1998). *Cross-validatory bandwidth selections for regression estimation based on dependent data*. Journal of Statistical Planning and Inference, 68(2), 387-415.

[122] Yaracos, Y. G. (1985). *Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy.* The Annals of Statistics, 13(2), 768-774.

[123] Yu, J. and Yuan, Y. (2011). *Investor Sentiment and the Mean-Variance Relation.* Journal of Financial Economics, 100(2), 367-381.

[124] Zhang, X., King, M. L. and Hyndman, R. J. (2006). *A Bayesian approach to bandwidth selection for multivariate kernel density estimation.* Computational Statistics and Data Analysis, 50(11), 3009-3031.

[125] Zolotarev, V. M. (1986). *Asymptotic behavior of the Gaussian measure in* ℓ_2. Journal of Soviet Mathematics, 35(2), 2330-2334.