

Density estimation with locally identically distributed data and with locally stationary data

Jussi Klemelä

University of Mannheim, Department of Economics

L7 3-5, 68131 Mannheim, Germany

Email: klemela@rumms.uni-mannheim.de

Fax +49 621 1811931

September 29, 2006

Abstract

We consider multivariate density estimation when the assumptions of identically distributed data or stationary data are relaxed to the assumptions of locally identically distributed data or locally stationary data. We assume that the distribution of the data is changing continuously as function of time. To estimate densities non-parametrically with these local regularity conditions we need time localization in addition to the usual space localization. We define a time localized kernel estimator which estimates the density non-parametrically at any given point of time. The consistency of the time localized kernel estimator is proved and the rates of convergence of the estimator are derived under conditions on the β - and α -mixing coefficients. Both the time series setting and spatial setting are covered.

Key Words: Kernel estimator, mixing coefficients, non-parametric density estimation, spatial data, time series data.

Short title: Locally identically distributed data

1 Introduction

We assume to observe a realization of sequence $X^1, \dots, X^T \in \mathbf{R}^p$ of random vectors. We do not assume that the random vectors are identically distributed or that they are independent. We will assume that the random vectors are *locally identically distributed*. When there exists a distribution P such that a sufficient number of random vectors in the sequence have a distribution close to P , then we say that the random vectors are locally identically distributed. Then we may find an estimator which accurately estimates P . We call distribution P the true distribution. We assume that P and the distributions of X^t have densities with respect to Lebesgue measure.

Let \mathbf{T} be a metric space with metric $\delta : \mathbf{T} \times \mathbf{T} \rightarrow [0, \infty)$. We assume that there exists a mapping $\tau : \{1, \dots, T\} \rightarrow \mathbf{T}$, which gives the locations of the observations in time or space. In the time series setting we may take $\mathbf{T} = [0, 1]$ and $\tau(t) = t/T$, assuming that X^t is observed earlier than X^u when $t < u$, and that the time steps are equal. In the spatial setting we may have for example $\mathbf{T} = [0, 1]^2$, $\mathbf{T} = [0, 1]^3$, or $\mathbf{T} = S_3$, where S_3 is the unit sphere in \mathbf{R}^3 , and $\tau(t)$ gives the spatial location of the observation. We use the time series terminology, and call $\tau(t)$ the time point of the observation.

We may imagine that there exists a collection of density functions $\{g_u : u \in T\}$, and we want to estimate density function g_{τ_0} , for a fixed $\tau_0 \in T$. If the density of X^t is close to g_{τ_0} , when $\tau(t)$ is close to τ_0 , and if there is enough such indexes t that $\tau(t)$ is close to τ_0 , then sequence X^1, \dots, X^T may be used to estimate g_{τ_0} .

For example, in the time series setting we may want to estimate the density of X^{t_0} , where t_0 is a fixed point of time. When we want to estimate the current density, then $t_0 = T$. Denote the density of X^{t_0} with f_{t_0} . The estimation of f_{t_0} is possible when the distribution of random vector X^t is changing continuously and sufficiently slowly as a function of t . In this case the time series is called locally identically distributed.

When the sequence of random vectors is locally identically distributed we may estimate the true distribution with a time localized kernel estimator. A time localized kernel estimate is a moving average of localized kernels. We prove the consistency of the time localized kernel estimator and give the rates of convergence of the estimator under smoothness assumptions for the densities.

We may also want estimate the joint distribution of consecutive observations $X^{t_0}, X^{t_0-1}, \dots, X^{t_0-k}$ for a fixed $k \geq 1$. When the joint distribution of the vector of consecutive observations is locally identically distributed, we call the times series locally stationary. We may again apply the time

localized kernel estimator.

A common univariate nonstationary time series model is the regression model $Y^t = m(w^t) + Z^t$, where m is a real valued smooth regression function, w^t are fixed design points, and $Z^t \in \mathbf{R}$ are i.i.d random noise. Since regression function m is assumed to be smooth, the time series is in a certain sense locally identically distributed and function m may be estimated with a moving average, see Remark 5. A different concept of a nonstationary time series was introduced by Priestly (1965), in terms of the spectral density. Dahlhaus (1997) studied local stationarity in the context of covariance evolution and developed an asymptotic theory with the help of time rescaling. Thus, locally stationary time series models have been considered both in terms of the mean and covariance. We consider more general local stationarity in terms of density functions.

Density estimation is often studied under the assumption of independent observations but this assumption is usually not satisfied in the time series setting or in the spatial setting. The assumption of independence may be relaxed to an assumption on the largeness of the β -mixing coefficients or the α -mixing coefficients, see Viennet (1997), Bosq (1998). We make an assumption concerning the largeness of these coefficients in the time localized setting. The assumptions state that the joint distribution of nearby observations is not too far away from the product distribution, in a given neighborhood of the space \mathbf{T} . β -mixing coefficients are more convenient for the statistical analysis but α -mixing coefficients provide a natural relaxation of the β -mixing coefficients and thus we study also the α -mixing coefficients.

In Section 2 we study the problem of the estimation of the joint distribution of the components of a vector time series. Section 2.1 contains the definition of locally identically distributed data and the definition of the time localized kernel estimator. Section 2.2 contains the definitions of the dependency concepts. Section 2.3 contains the results on consistency and on the rate of convergence. In Section 3 we state the definition of locally stationary data and we apply the previous results to the estimation of the joint distribution of consecutive observations.

We denote $a_T \preceq b_T$ when $\limsup_{T \rightarrow \infty} a_T/b_T < \infty$, that is, when $a_T = O(b_T)$. We denote $a_T = o(b_T)$ when $\lim_{T \rightarrow \infty} a_T/b_T = 0$. We denote $a_T \sim b_T$ when $\lim_{T \rightarrow \infty} a_T/b_T = 1$. We denote $a_T \lesssim b_T$ when $\limsup_{T \rightarrow \infty} a_T/b_T \leq 1$. The L_q norm of a function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ will be denoted with $\|f\|_q$. The i th partial derivative of a function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ is denoted by $D_i f$ and $D^\alpha f = D_1^{\alpha_1} \cdots D_d^{\alpha_d} f$, where $\alpha = (\alpha_1, \dots, \alpha_d)$ is a vector of nonnegative integers.

2 Estimation of the joint distribution of the components of a vector time series

2.1 Locally identically distributed observations and the time localized kernel estimator

Let $X^1, \dots, X^T \in \mathbf{R}^p$ be a sequence of random vectors. We write $X^t = (X_1^t, \dots, X_p^t)$, $t = 1, \dots, T$. We want to estimate a density $g : \mathbf{R}^p \rightarrow \mathbf{R}$. This is possible if there is enough observations in the sequence whose density is close to g . We define the concept of *locally identically distributed* random vectors, which formalizes this idea. The concept of locally identically distributed random vectors is defined with the help of a fixed rate function r which quantifies the similarity of distributions.

Definition 1 *Random vectors $X^1, \dots, X^T \in \mathbf{R}^p$ are locally identically distributed, with rate function $r : [0, \infty) \rightarrow [0, \infty)$, with rate $o(1)$, at time point $\tau_0 \in \mathbf{T}$, when there exists density $g : \mathbf{R}^p \rightarrow \mathbf{R}$ such that*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] \|f_t - g\|_2 = 0 \quad (1)$$

and with rate $O(T^{-1/2})$, when

$$\limsup_{T \rightarrow \infty} T^{-1/2} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] \|f_t - g\|_2 < \infty, \quad (2)$$

where $f_t : \mathbf{R}^p \rightarrow \mathbf{R}$, $t = 1, \dots, T$, is the density function of X^t .

In order estimation of g to be possible we need to assume that around the time point τ_0 of interest there is asymptotically a non-negligible amount of indexes $\tau(t)$.

Assumption 1 *Let $\tau_0 \in \mathbf{T}$ be the point of interest, and let $r : [0, \infty) \rightarrow [0, \infty)$ be the rate function in Definition 1. We assume that*

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] > 0.$$

For example, when $\mathbf{T} = [0, 1]$, $\delta(t, s) = |t - s|$, $\tau(t) = t/T$, $\tau_0 = 1$, and $0 < \int_0^1 r < \infty$, then

$$\frac{1}{T} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] = \frac{1}{T} \sum_{t=1}^T r(|t/T - 1|) \sim \int_0^1 r(|s - 1|) ds = \int_0^1 r, \quad (3)$$

as $T \rightarrow \infty$, and Assumption 1 holds.

Remark 1 To construct a consistent estimate for g we need to assume (1) and to construct an estimate with fast rates of convergence we need to assume (2).

Remark 2 It is natural to assume that rate function $r : [0, \infty) \rightarrow [0, \infty)$ is monotonically decreasing, for example $r(x) = (1 - x^q)_+$ or $r(x) = \exp(-x^q)$, for $q \geq 1$. Then Definition 1 says that when $\tau(t)$ are close to τ_0 , then f_t are close to g , up to a negligible amount of indexes t . Rate function r quantifies the closeness of f_t to g . If Definition 1 holds for rate function r , then it holds for some other rate functions as well. If r satisfies (1) or (2), then r^* satisfies (1) or (2) when

$$\limsup_{T \rightarrow \infty} \max_{t=1, \dots, T} \frac{r^*[\delta(\tau(t), \tau_0)]}{r[\delta(\tau(t), \tau_0)]} < \infty.$$

Remark 3 We have defined locally identically distributed random vectors as an asymptotic concept, as $T \rightarrow \infty$. The type of asymptotics we used is called time rescaling (Dahlhaus (1997)) or in-fill asymptotics. The in-fill asymptotics is often used to analyze estimators in the setting where one assumes that there exists a continuous time stochastic process $Z(t)$, $t \in [0, 1]$, and the observations are sampled from this process: $X^t = Z(t/T)$, $t = 1, \dots, T$.

Remark 4 We have measured the distance between distributions with the L_2 distance between the density functions of the distributions. In particular, we assume that the densities exist. It is natural to consider different definitions of locally identically distributed random vectors, based on various distances between distributions.

Remark 5 Two nonstationary models which have been used in statistics include the mean-nonstationary model

$$X^t = \mu^t + Z^t,$$

where $\mu^t \in \mathbf{R}^p$ and Z^t are i.i.d with $EZ^t = 0$, and the volatility-nonstationary model

$$X^t = \sigma_t Z^t,$$

where $\sigma_t > 0$ and Z^t are i.i.d. This volatility-nonstationary model has been used as an alternative for stationary GARCH-modeling, see Stărică and Granger (2005). For the mean-nonstationary model one has $f_t(x) = f_Z(x - \mu_t)$ where f_Z is the density of Z^t , and thus one can write, under smoothness assumptions on f_Z ,

$$\|f_t - f_u\|_2 \leq \|\mu^t - \mu^u\| \cdot \|Df_Z\|_2.$$

Thus X^t are locally identically distributed in the sense of Definition 1, if μ^t is sufficiently smooth as function of t . For the volatility-nonstationary model one has $f_t(x) = f_Z(x/\sigma_t)/\sigma_t^p$ and thus one can write, under smoothness assumptions on f_Z ,

$$\begin{aligned} \|f_t - f_u\|_2 &\leq \|f_Z(x/\sigma_t)/\sigma_t^p - f_Z(x/\sigma_u)/\sigma_u^p\|_2 + \|f_Z(x/\sigma_u)/\sigma_u^p - f_Z(x/\sigma_u)/\sigma_t^p\|_2 \\ &= \sigma_t^{-p/2} \|f_Z(x) - f_Z(x\sigma_t/\sigma_u)\|_2 + |\sigma_t^{-p} - \sigma_u^{-p}| \sigma_u^{p/2} \|f_Z\|_2 \\ &= |1 - \sigma_t/\sigma_u| \sigma_t^{-p/2} \|x^T Df_Z(\xi)\|_2 + |1 - (\sigma_u/\sigma_t)^p| \sigma_u^{-p/2} \|f_Z\|_2, \end{aligned}$$

where $\xi = \xi_{x, \sigma_t, \sigma_u}$ is between x and $x\sigma_t/\sigma_u$ and we use the abuse of notation $\|f\|_2 = \|f(x)\|_2$ when x is the integration variable. Thus X^t are locally identically distributed in the sense of Definition 1, if the ratio σ_u/σ_t is close to one and σ_t^{-1} are bounded.

If the random vectors are locally identically distributed, then a *time localized kernel estimator* may be used to estimate density g . We define the time localized kernel estimator by associating weights p_t , $t = 1, \dots, T$, to the observations, which replace the usual weights T^{-1} .

Definition 2 *The time localized kernel estimator is defined with*

$$\hat{f}_T(x) = \sum_{t=1}^T p_t K_h(x - X^t), \quad x \in \mathbf{R}^p,$$

where $h > 0$ is the smoothing parameter, $K_h(x) = h^{-p} K(x/h)$, $K : \mathbf{R}^p \rightarrow \mathbf{R}$ is the kernel function, and $p_t \geq 0$, $\sum_{t=1}^T p_t = 1$.

The time localized kernel estimator is used to estimate density g . The optimal choice of the smoothing parameter h and the weights p_t depend on the density g and on the the rate function r . See (13) for a sufficient condition on the weights to reach consistency and fast rates of convergence. An application of a time localized kernel estimator is given in Klemelä (2006).

2.2 Dependency concepts

We will not assume that the observations are independent but consider two concepts of dependence: conditions on the largeness of the β -mixing coefficients and on the α -mixing coefficients.

2.2.1 Regular mixing coefficients

We make an assumption on the largeness of β -mixing coefficients, which are also called regular mixing coefficients. These coefficients were introduced by Kolmogorov and Rozanov (1961). Let Z, \tilde{Z} be random vectors and let $P_Z, P_{\tilde{Z}}$ be the corresponding probability measures. Let $P_{Z, \tilde{Z}}$ be the probability measure of the joint distribution of Z, \tilde{Z} and let $P_Z \otimes P_{\tilde{Z}}$ be the product measure of P_Z and $P_{\tilde{Z}}$. The β -mixing coefficient is defined by

$$\beta(Z, \tilde{Z}) = \|P_{Z, \tilde{Z}} - P_Z \otimes P_{\tilde{Z}}\|_{tot} = \sup_W \int W d(P_{Z, \tilde{Z}} - P_Z \otimes P_{\tilde{Z}}), \quad (4)$$

where the supremum is over all random variables $0 \leq W \leq 1$, measurable with respect to the product sigma-algebra of the sigma-algebras generated by Z and \tilde{Z} . We set $\beta(Z, Z) = 1$.

Assumption 2 Let $r : [0, \infty) \rightarrow [0, \infty)$ be the rate function of Definition 1 and let $\tau_0 \in \mathbf{T}$. Assume that

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t, u=1}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] \beta(X^t, X^u) < \infty.$$

Remark 6 Assumption 2 defines a localized dependency property, since it depends on $\tau_0 \in \mathbf{T}$. We will take rate function r same as in Definition 1. We have multiplied with $r[\delta(\tau(t), \tau_0)]$ and $r[\delta(\tau(u), \tau_0)]$ which are small when $\tau(t)$ or $\tau(u)$ are far away from τ_0 , when r is similar as in Remark 2. Thus the assumption of the smallness of $\beta(X^t, X^u)$ is not so restrictive as in the case we would take $r \equiv 1$.

Remark 7 Let us consider the case when X^1, \dots, X^T is strictly stationary. Define $b_k = \beta(X^t, X^{t+k})$, for $t = 1, \dots, T, k = 1 - t, \dots, T - t$. Numbers b_k do not depend on t due to the strict stationarity. We have that

$$\frac{1}{T} \sum_{t, u=1}^T \beta(X^t, X^u) = \frac{1}{T} \sum_{t=1}^T \sum_{k=1-t}^{T-t} \beta(X^t, X^{t+k}) \leq 2 \sum_{k=0}^{\infty} b_k. \quad (5)$$

Thus the assumption $\sum_{k=0}^{\infty} b_k < \infty$ implies Assumption 2 with rate function r as identically 1: $r \equiv 1$. Viennet (1997) makes the assumption $\sum_{k=0}^{\infty} (k+1)^{q-2} b_k < \infty$ when studying density estimation with the L_q loss.

2.2.2 Strong mixing coefficients

Strong mixing coefficients are also called α -mixing coefficients. These coefficients were introduced by Rosenblatt (1956). The α -mixing coefficient is defined by

$$\alpha(Z, \tilde{Z}) = \sup_{A, B} \left| P(Z \in A, \tilde{Z} \in B) - P(Z \in A)P(\tilde{Z} \in B) \right|,$$

where Z, \tilde{Z} are random vectors taking values in \mathbf{R}^p and the supremum is taken over Borel measurable subsets of \mathbf{R}^p . We may define $\alpha(Z, \tilde{Z})$ analogously to (4), by writing

$$\alpha(Z, \tilde{Z}) = \sup_{A, B} \int I_{A \times B} d(P_{Z, \tilde{Z}} - P_Z \otimes P_{\tilde{Z}}). \quad (6)$$

α -mixing coefficients provide a natural relaxation for the β -mixing coefficients because

$$\alpha(Z, \tilde{Z}) \leq \beta(Z, \tilde{Z}),$$

which can be seen by comparing (4) and (6). For the statistical theory β -mixing coefficients are more convenient because we have to make stronger assumptions on the convergence of the sums of α -mixing coefficients than we have to make on the convergence of the sums of β -mixing coefficients.

Assumption 3 Let $r : [0, \infty) \rightarrow [0, \infty)$ be the rate function of Definition 1 and let $\tau_0 \in \mathbf{T}$. Assume that

$$T^{-1} \sum_{t=1}^T \sum_{u=1, |u-t|>U}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] \alpha(X^t, X^u) \leq C \cdot U^{1-a}, \quad (7)$$

and

$$T^{-1} \sum_{t=1}^T \sum_{u=1, |u-t| \leq U}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] \leq C \cdot U, \quad (8)$$

for integers $1 \leq U < T$, where C and a are positive constants.

Remark 8 Assumption (8) holds when r is bounded.

Remark 9 Let us consider the case when X^1, \dots, X^T is strictly stationary. Define $a_k = \alpha(X^t, X^{t+k})$, for $t = 1, \dots, T, k = 1-t, \dots, T-t$. We have similarly as in (5)

$$\frac{1}{T} \sum_{t=1}^T \sum_{u=1, |u-t|>U}^T \alpha(X^t, X^u) = \frac{1}{T} \sum_{t=1}^T \sum_{k=1-t, |k|>U}^{T-t} \alpha(X^t, X^{t+k}) \leq 2 \sum_{k=U+1}^{\infty} a_k.$$

Bosq (1998), Theorem 2.1, derives the rates of convergence of a kernel estimator under assumption $a_k \leq C'k^{-a}$ where $a > 1$ and C' is a positive constant. This assumption implies

$$\sum_{k=U+1}^{\infty} a_k \leq C' \frac{(U+1)^{1-a}}{a-1}.$$

Thus the assumption $a_k \leq C'k^{-a}$ implies assumption (7) with rate function r as identically 1: $r \equiv 1$.

In the case of α -mixing coefficients we have to make the following technical assumptions, which are similar to the assumptions in Bosq (1998), Theorem 2.1. We assume that the densities f_t have compact supports which are contained in a fixed compact set:

$$\text{supp}(f_t) \subset A, \quad t = 1, 2, \dots, \quad \text{for some compact set } A. \quad (9)$$

For $t \neq u$ and for some $r > 2$,

$$\|d(P_{t,u} - P_t \otimes P_u)\|_r \leq C_r, \quad (10)$$

where P_t is the distribution of X^t , P_u is the distribution of X^u , $P_{t,u}$ is the distribution of (X^t, X^u) , and C_r is a positive constant. Constant a in Assumption 3 satisfies

$$a > 2 \frac{r-1}{r-2}. \quad (11)$$

The dual use of r as a symbol for the rate function and in (10) will not cause confusion.

2.3 Results

2.3.1 Consistency

We prove the consistency of the time localized kernel estimator, for the L_2 error.

Theorem 1 *Let sequence $X^1, \dots, X^T \in \mathbf{R}^p$ locally identically distributed in the sense of Definition 1, Eq. (1), with rate function r which satisfies Assumption 1. Let Assumption 2 on the β -mixing coefficients or Assumption 3 with (9)-(11) on the α -mixing coefficients be satisfied. Assume that for all $t = 1, \dots, T$,*

$$\|f_t\|_2 \leq C, \quad (12)$$

for a positive constant C . We apply the time localized kernel estimator \hat{f}_T defined in Definition 2 with the following parameters:

1. kernel function K satisfies $\|K\|_2 < \infty$, $\int_{\mathbf{R}^p} K = 1$, and the support of K is included in $[-1/2, 1/2]^p$,
2. smoothing parameter $h = h_T$ is such that $\lim_{T \rightarrow \infty} h_T = 0$ and $\lim_{T \rightarrow \infty} Th_T^p = \infty$,
3. weights $p_t = p_{t,T}$ satisfy

$$\limsup_{T \rightarrow \infty} \max_{t=1, \dots, T} \frac{p_t}{\pi_t} < \infty, \quad (13)$$

where

$$\pi_t = \frac{r[\delta(\tau(t), \tau_0)]}{\sum_{u=1}^T r[\delta(\tau(u), \tau_0)]}. \quad (14)$$

Then,

$$\lim_{T \rightarrow \infty} E \|\hat{f}_T - g\|_2 = 0.$$

Proof. We consider first the bias and then the variance.

Bias. We have for $x \in \mathbf{R}^p$

$$\begin{aligned} E \hat{f}_T(x) - g(x) &= \sum_{t=1}^T p_t \int_{\mathbf{R}^p} K_h(x-y) f_t(y) dy - g(x) \\ &= \sum_{t=1}^T p_t \left(\int_{\mathbf{R}^p} K_h(x-y) f_t(y) dy - f_t(x) \right) \\ &\quad + \sum_{t=1}^T (p_t - \pi_t) (f_t(x) - g(x)) \\ &\quad + \sum_{t=1}^T \pi_t (f_t(x) - g(x)) \\ &\stackrel{def}{=} A_1(x) + A_2(x) + A_3(x). \end{aligned} \quad (15)$$

The assumption in (12) and Lemma 1 below imply that

$$\max_{t=1, \dots, T} \int_{\mathbf{R}^p} \left(\int_{\mathbf{R}^p} K_h(x-y) f_t(y) dy - f_t(x) \right)^2 dx \rightarrow 0$$

as $h \rightarrow 0$. This and the assumption $\lim_{T \rightarrow \infty} h = 0$ imply that $\lim_{T \rightarrow \infty} \|A_1\|_2 = 0$. By Assumption 1,

$$\limsup_{T \rightarrow \infty} \max_{t=1, \dots, T} \pi_t \cdot \frac{T}{r[\delta(\tau(t), \tau_0)]} < \infty. \quad (16)$$

Applying assumption (13) on the weights and Eq. (16),

$$\limsup_{T \rightarrow \infty} \max_{t=1, \dots, T} \frac{T}{r[\delta(\tau(t), \tau_0)]} |p_t - \pi_t| < \infty. \quad (17)$$

Thus, applying (17),

$$\|A_2\|_2 \leq \sum_{t=1}^T |p_t - \pi_t| \|f_t - g\|_2 \preceq T^{-1} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] \|f_t - g\|_2, \quad (18)$$

where we denote $a_T \preceq b_T$ when $\limsup_{T \rightarrow \infty} a_T/b_T < \infty$. Thus, by assumption (1) of locally identically distributed random vectors, $\lim_{T \rightarrow \infty} \|A_2\|_2 = 0$. We have by (16) that

$$\begin{aligned} \|A_3\|_2 &= \left\| \sum_{t=1}^T \pi_t (f_t - g) \right\|_2 \\ &\preceq \left\| T^{-1} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] (f_t - g) \right\|_2 \\ &\leq T^{-1} \sum_{t=1}^T r[\delta(\tau(t), \tau_0)] \|f_t - g\|_2. \end{aligned} \quad (19)$$

It follows, by assumption (1) of locally identically distributed random vectors, that $\lim_{T \rightarrow \infty} \|A_3\|_2 = 0$. We have proved

$$\lim_{T \rightarrow \infty} \|E \hat{f}_T - g\|_2 = 0. \quad (20)$$

Variance. Denote

$$\text{Cov}_{tu} = \int_{\mathbf{R}^p} \text{Cov}(K_h(x - X^t), K_h(x - X^u)) dx.$$

We have that

$$\begin{aligned} \int_{\mathbf{R}^p} \text{Var}(\hat{f}_T) &= \sum_{t,u=1}^T p_t p_u \text{Cov}_{tu} \\ &\preceq \sum_{t,u=1}^T \pi_t \pi_u \text{Cov}_{tu} \end{aligned} \quad (21)$$

$$\preceq T^{-2} \sum_{t,u=1}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] \text{Cov}_{tu}, \quad (22)$$

where in (21) we applied assumption (13) on the weights and in (22) we applied (16). Applying Lemma 1 below,

$$\begin{aligned} \text{Cov}_{tt} &= \int_{\mathbf{R}^p} \text{Var}(K_h(x - X^t)) dx \\ &\leq \int_{\mathbf{R}^p} dx \int_{\mathbf{R}^p} K_h^2(x - y) f_t(y) dy \end{aligned} \quad (23)$$

$$\sim h^{-p} \int_{\mathbf{R}^p} K^2, \quad (24)$$

as $T \rightarrow \infty$. It is left to calculate Cov_{tu} when $t \neq u$. The calculation is different for the β -mixing coefficients and for the α -mixing coefficients.

Covariance with β -mixing. Now we assume that Assumption 2 holds. We generalize Theorem 2.1 of Viennet (1997) to the nonstationary case. We have for $x \in \mathbf{R}^p$, $t, u = 1, \dots, T$, $t \neq u$, applying the Cauchy-Schwartz inequality,

$$\begin{aligned} &\text{Cov}(K_h(x - X^t), K_h(x - X^u)) \\ &= \int_{\mathbf{R}^p \times \mathbf{R}^p} K_h(x - y) K_h(x - z) d(P_{t,u} - P_t \otimes P_u)(y, z) \quad (25) \\ &\leq \left(\int_{\mathbf{R}^p \times \mathbf{R}^p} K_h^2(x - y) d|P_{t,u} - P_t \otimes P_u|(y, z) \right)^{1/2} \\ &\quad \times \left(\int_{\mathbf{R}^p \times \mathbf{R}^p} K_h^2(x - z) d|P_{t,u} - P_t \otimes P_u|(y, z) \right)^{1/2}, \end{aligned}$$

where P_t is the distribution of X^t , P_u is the distribution of X^u , and $P_{t,u}$ is the distribution of (X^t, X^u) . Denote with $Q_{t,u}^{(t)}$ the 1:st marginal distribution of $|P_{t,u} - P_t \otimes P_u|$: $Q_{t,u}^{(t)}(A) = \int_{A \times \mathbf{R}^p} d|P_{t,u} - P_t \otimes P_u|(y, z)$, and denote with $Q_{t,u}^{(u)}$ the second marginal distribution. Denote with $g_{t,u}^{(l)} : \mathbf{R}^p \rightarrow \mathbf{R}$ the Radon-Nikodym density of $Q_{t,u}^{(l)}$ with respect to P_l , $l = t, u$:

$$g_{t,u}^{(l)} = \frac{dQ_{t,u}^{(l)}}{dP_l}, \quad l = t, u.$$

Then we have

$$\int_{\mathbf{R}^p \times \mathbf{R}^p} K_h^2(x - y) d|P_{t,u} - P_t \otimes P_u|(y, z) = \int_{\mathbf{R}^p} K_h^2(x - y) g_{t,u}^{(t)}(y) f_t(y) dy,$$

and

$$\int_{\mathbf{R}^p \times \mathbf{R}^p} K_h^2(x-z) d|P_{t,u} - P_t \otimes P_u|(y, z) = \int_{\mathbf{R}^p} K_h^2(x-z) g_{t,u}^{(u)}(z) f_u(z) dz.$$

Thus,

$$\begin{aligned} & \text{Cov}(K_h(x - X^t), K_h(x - X^u)) \\ & \leq \int_{\mathbf{R}^p} K_h^2(x - y) (g_{t,u}^{(t)}(y) f_t(y) + g_{t,u}^{(u)}(y) f_u(y)) dy. \end{aligned}$$

We have that

$$\int_{\mathbf{R}^p} g_{t,u}^{(t)} f_t = \int_{\mathbf{R}^p \times \mathbf{R}^p} d|P_{t,u} - P_t \otimes P_u| = 2 \|P_{t,u} - P_t \otimes P_u\|_{tot} = 2\beta(X^t, X^u)$$

and also $\int_{\mathbf{R}^p} g_{t,u}^{(u)} f_u = 2\beta(X^t, X^u)$, where $\beta(X^t, X^u)$ is defined in (4). Thus, applying Lemma 1 below,

$$\text{Cov}_{tu} = \int_{\mathbf{R}^p} \text{Cov}(K_h(x - X^t), K_h(x - X^u)) dx \lesssim h^{-p} 4\beta(X^t, X^u) \int_{\mathbf{R}^p} K^2, \quad (26)$$

where $a_T \lesssim b_T$ means that $\limsup_{T \rightarrow \infty} a_T/b_T \leq 1$. Eq. (26) holds also for $t = u$, as noted in (24). Applying (22) and (26) we get

$$\begin{aligned} & \int_{\mathbf{R}^p} \text{Var}(\hat{f}_T) \\ & \preccurlyeq (Th^p)^{-1} T^{-1} \sum_{t,u=1}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] \beta(X^t, X^u) \\ & \preccurlyeq O((Th^p)^{-1}) \\ & = o(1), \end{aligned} \quad (27)$$

as $T \rightarrow \infty$, where we applied also Assumption 2 on the β -mixing coefficients and the assumption $\lim_{T \rightarrow \infty} Th^p = \infty$.

Covariance with α -mixing. We modify the proof of Theorem 2.1 in Bosq (1998) for the nonstationary case. We apply Billingsley's inequality

$$\text{Cov}(X, Y) \leq 4\|X\|_\infty \|Y\|_\infty \alpha(X, Y), \quad (28)$$

which is proved in Doukhan (1994), Lemma 3, page 10 and Bosq (1998), Corollary 1.1. Eq. (28) implies that

$$\text{Cov}(K_h(x - X^t), K_h(x - X^u)) \leq 4h^{-2p} \|K\|_\infty^2 \alpha(X^t, X^u).$$

On the other hand, using Eq. (25) and Hölder's inequality for $1/r + 1/q = 1$ we get

$$\text{Cov}\left(K_h(x - X^t), K_h(x - X^u)\right) \leq \|d(P_{t,u} - P_t \otimes P_u)\|_r \cdot h^{-2p/r} \|K\|_q^2,$$

since $\|K_h(x - \cdot)\|_q = h^{-p} h^{p/q} \|K\|_q$. Since f_t have compact support (assumption (9)) the volume of the support of the kernel estimator can be bounded by a constant C_s (we will assume from now on that $h \leq 1$) and we have

$$\text{Cov}_{tu} \leq C_s \cdot \max\left\{C_r h^{-2p/r} \|K\|_q^2, 4h^{-2p} \|K\|_\infty^2 \alpha(X^t, X^u)\right\},$$

where we used also assumption (10). Let

$$U = \lceil h^{-2p/(qa)} \rceil.$$

The integer U solves asymptotically $Uh^{-2p/r} \approx U^{1-a} h^{-2p}$. We have, continuing from (22), applying Assumption 3,

$$\begin{aligned} & T^{-2} \sum_{t,u=1, t \neq u}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] \text{Cov}_{tu} & (29) \\ & = \\ & \leq C_s T^{-2} \left[\sum_{t=1}^T \sum_{u=1, 1 \leq |u-t| \leq U}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] C_r h^{-2p/r} \|K\|_q^2 \right. \\ & \quad \left. + \sum_{t=1}^T \sum_{u=1, |u-t| > U}^T r[\delta(\tau(t), \tau_0)] \cdot r[\delta(\tau(u), \tau_0)] 4h^{-2p} \|K\|_\infty^2 \alpha(X^t, X^u) \right] \\ & \leq CC_s T^{-1} \left[C_r h^{-2p/r} \|K\|_q^2 \cdot U + 4h^{-2p} \|K\|_\infty^2 \cdot U^{1-a} \right]. \\ & = O\left((Th^p)^{-1}\right), & (30) \end{aligned}$$

since

$$Uh^{-2p/r} < h^{-2p[1/(qa)+1/r]} < h^{-p},$$

since $1/(qa) < (r-2)/(2r)$ (remember $q = r/(r-1)$ and $a > 2(r-1)/(r-2)$) by assumption (11) and

$$U^{1-a} h^{-2p} < h^{2p(a-1)/(qa)} < h^{-p},$$

since $(a-1)/a > q/2$ (note $1 - 1/a > r/[2(r-1)] = q/2$). Combining (22), (24), and (30) gives

$$\int_{\mathbf{R}^p} \text{Var}(\hat{f}_T) = O\left((Th^p)^{-1}\right) = o(1). \quad (31)$$

Collecting the results. The theorem follows from (20) and from (27) or from (31). \square

We have applied the following lemma, whose proof may be found for example in Stein (1970), or in the Appendix.

Lemma 1 *Let $f : \mathbf{R}^p \rightarrow \mathbf{R}$ and $\|f\|_q < \infty$ for some $q \in [1, \infty)$. Let $K : \mathbf{R}^p \rightarrow \mathbf{R}$ have compact support and $\|K\|_1 < \infty$. Then*

$$\lim_{h \rightarrow 0} \int_{\mathbf{R}^p} \left| \int_{\mathbf{R}^p} K_h(x-y) f(y) dy - f(x) \int_{\mathbf{R}^p} |K| \right|^q dx = 0,$$

where $K_h(x) = h^{-p} K(x/h)$.

2.3.2 Rates of convergence

We derive the rates of convergence of the time localized kernel estimator.

Theorem 2 *Let sequence $X^1, \dots, X^T \in \mathbf{R}^p$ be locally identically distributed in the sense of Definition 1, Eq. (2), with rate function r which satisfies Assumption 1. Let Assumption 2 on the β -mixing coefficients or Assumption 3 with (9)-(11) on the α -mixing coefficients be satisfied. Assume that for all $t = 1, \dots, T$, for integer $s \geq 1$, density f_t is s times continuously differentiable and for $|\alpha| = s$,*

$$\|D^\alpha f_t\|_2 \leq C, \quad (32)$$

for a positive constant C , where $\alpha = (\alpha_1, \dots, \alpha_p)$ is a multi-index and we denote $|\alpha| = \alpha_1 + \dots + \alpha_p$. We apply the time localized density estimator \hat{f}_T defined in Definition 2 with the following parameters:

1. kernel function $K : \mathbf{R}^p \rightarrow \mathbf{R}$ satisfies $\|K\|_2 < \infty$, $\int_{\mathbf{R}^p} K = 1$, the support of K is included in $[-1/2, 1/2]^p$, $\int_{\mathbf{R}^p} x^\alpha K(x) dx = 0$, for $|\alpha| = 1, \dots, s-1$, $\int_{\mathbf{R}^p} |x^\alpha| |K(x)| dx < \infty$, for $|\alpha| = s$.
2. smoothing parameter is $h = h_T = C_h T^{-1/(2s+p)}$, for a positive constant C_h ,
3. weights p_t , $t = 1, \dots, T$, satisfy (13).

Then,

$$\lim_{T \rightarrow \infty} T^{2s/(2s+p)} E \|\hat{f}_T - g\|_2^2 < \infty.$$

Proof. We have for the bias that

$$\|E\hat{f}_T - g\|_2^2 = O(h^{2s} + T^{-1}),$$

as $T \rightarrow \infty$. Indeed, by the Taylor expansion we get

$$\int_{\mathbf{R}^p} K(z)(f_t(x + hz) - f_t(x)) dz = h^s \sum_{|\alpha|=s} \frac{1}{\alpha!} \int_{\mathbf{R}^p} z^\alpha K(z) D^\alpha f_t(x + \xi_\alpha hz) dz$$

where $0 < \xi_\alpha < 1$, and thus assumption (32) implies $\|A_1\|_2^2 = O(h^{2s})$, where A_1 is defined in (15). Then we combine (18) and (19) with assumption (2), to get $\|A_2\|_2^2 + \|A_3\|_2^2 = O(T^{-1})$. Similarly as in the proof of Theorem 1 we have that

$$\int_{\mathbf{R}^p} \text{Var}(\hat{f}_T) = O(T^{-1}h^{-p}),$$

as $T \rightarrow \infty$. The theorem follows by the choice of $h = C_h T^{-1/(2s+p)}$. \square

3 Estimation of the joint distribution of consecutive observations

When X^1, X^2, \dots is a time series, then one may be interested in the prediction of, for example, one step ahead with one explanatory variable, and consider new time series $Y^1 = (X^1, X^2), Y^2 = (X^2, X^3), \dots$. If the original time series is stationary, then the new time series consists of identically distributed random vectors. We say that that the original time series is locally stationary if the new time series is locally identically distributed. Local stationarity seems more natural in the time series setting than in the spatial setting, and thus we discuss first the time series setting. The general setting is addressed in Remark 10.

Let $l \geq 1$ and let $k_i \geq 1$ be such that $1 + k_1 + \dots + k_l \leq T$. Denote $|k| = k_1 + \dots + k_l$. We create data matrix

$$A = \begin{bmatrix} X^1 & X^{1+k_1} & \dots & X^{1+|k|} \\ X^2 & X^{2+k_1} & \dots & X^{2+|k|} \\ \vdots & \vdots & & \vdots \\ X^{T-|k|} & X^{T-|k|+k_1} & \dots & X^T \end{bmatrix}.$$

Data matrix A is of dimension $n \times d$, where

$$n = T - |k|, \quad d = (l + 1) \cdot p,$$

when X^t are interpreted as row vectors of length p . Denote with

$$Y^i = (X^i, X^{i+k_1}, \dots, X^{i+|k_i|}) \in \mathbf{R}^d, \quad i = 1, \dots, n, \quad (33)$$

the i :th row of matrix A . We call sequence $X^1, \dots, X^T \in \mathbf{R}^p$ (strongly, strictly) *stationary* when for all choices of $l \geq 1$ and $k_i \geq 1$ such that $1+k_1+\dots+k_l \leq T$,

$$Y^1, \dots, Y^n \text{ are identically distributed.}$$

The local stationarity will be defined to mean that time series Y^1, \dots, Y^n is locally identically distributed. In addition, the definition of local stationarity will be made to depend on a given choice of k_1, \dots, k_l . This is natural since time series $Y^1 = (X^1, X^2), Y^2 = (X^2, X^3), \dots$ might be locally identically distributed but the time series $Y^1 = (X^1, X^{101}), Y^2 = (X^2, X^{102}), \dots$ might not.

Definition 3 *Time series $X^1, \dots, X^T \in \mathbf{R}^p$ is locally stationary, with shifting steps k_1, \dots, k_l , with rate function $r : [0, \infty) \rightarrow [0, \infty)$, with rate $o(1)$ or $O(T^{-1/2})$, at time point $\tau_0 \in \mathbf{T}$, when sequence (Y^1, \dots, Y^n) , defined with (33), is locally identically distributed with rate function r , rate $o(1)$ or $O(T^{-1/2})$, and time point τ_0 , as defined in Definition 1.*

When the sequence is stationary, then we may apply the multivariate $((l+1) \cdot p)$ -dimensional kernel density estimator to estimate the distribution of

$$Y^1 = (X^1, X^{1+k_1}, \dots, X^{1+k_1+\dots+k_l}).$$

When the sequence is locally stationary we may apply the time localized kernel estimator to estimate the density g . We state the consistency and the rates of convergence of the time localized kernel estimator.

Theorem 3 *Let time series $X^1, \dots, X^T \in \mathbf{R}^p$ be locally stationary in the sense of Definition 3 and assume that sequence Y^1, \dots, Y^n defined with (33) satisfies also the other assumptions of Theorem 1 and Theorem 2. Then the corresponding time localized kernel estimator satisfies the statements of Theorem 1 and Theorem 2.*

Theorem 3 is basically a restatement of Theorem 1 and Theorem 2, and thus it needs not to be proved.

Remark 10 Stationarity and local stationarity may be also defined in the general setting, which includes the time series setting and the spatial setting as special cases. For $t \in \{1, \dots, T\}$ and $\mu \in \mathbf{T}$, let $t_\mu \in \{1, \dots, T\}$ be the index which corresponds to translation with μ :

$$t_\mu = \tau^{-1}(\tau(t) + \mu),$$

where $\tau : \{1, \dots, T\} \rightarrow \mathbf{T}$ is assumed to be injective. Here we have to assume that $\tau(t) + \mu \in \text{range}(\tau)$, where $\text{range}(\tau) \subset \mathbf{T}$ is the range of τ : $\text{range}(\tau) = \{\tau(t) : t = 1, \dots, T\}$. Let μ_1, \dots, μ_l be translation elements, and let $I \subset \{1, \dots, T\}$ be the set of indices which may be translated by μ_1, \dots, μ_l :

$$I = I_{\mu_1, \dots, \mu_l} = \{t = 1, \dots, T : \tau(t) + \mu \in \text{range}(\tau) \text{ for all } \mu = \mu_1, \dots, \mu_l\}.$$

Denote

$$Y^t = (X^{t_{\mu_1}}, \dots, X^{t_{\mu_l}}), \quad t \in I_{\mu_1, \dots, \mu_l}.$$

Now sequence X^1, \dots, X^T is called stationary if for all choices of μ_1, \dots, μ_l ,

$$\{Y^t : t \in I_{\mu_1, \dots, \mu_l}\} \text{ are identically distributed.}$$

Sequence X^1, \dots, X^T is called locally stationary, with shifting steps μ_1, \dots, μ_l , if

$$\{Y^t : t \in I_{\mu_1, \dots, \mu_l}\} \text{ is locally identically distributed.}$$

For example, in the time series setting, $\mathbf{T} = [0, 1]$, $\tau(t) = t/T$, $t_{\mu} = t + T\mu$, $\text{range}(\tau) = \{1/T, \dots, 1 - 1/T, 1\}$. If $\mu \in \{1/T, \dots, 1 - 1/T\}$, then $I_{\mu} = \{1, \dots, T - T\mu\}$.

Acknowledgments

Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/8-1. I wish to thank the referees for helpful comments.

References

- Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, Vol. 110 of *Lecture Notes in Statistics*, Springer.
- Dahlhaus, R. (1997), 'Fitting time series models to nonstationary processes', *Ann. Statist.* **25**, 1–37.
- Doukhan, P. (1994), *Mixing: Properties and Examples*, Vol. 85 of *Lecture Notes in Statistics*, Springer.
- Klemelä, J. (2006), 'Visualization of multivariate density estimates with shape trees', *J. Comput. Graph. Statist.* **15**(2), 372–397.

- Kolmogorov, A. N. and Rozanov, Y. A. (1961), 'On the strong mixing conditions for stationary gaussian sequences', *Theory Probab. Appl.* **5**, 204–207.
- Priestly, M. B. (1965), 'Evolutionary spectra and non-stationary processes', *J. Roy. Statist. Soc. Ser. B* **27**, 204–237.
- Rosenblatt, M. (1956), 'A central limit theorem and a strong mixing condition', *Proc. National Academic Science* **42**, 43–47.
- Stein, E. M. (1970), *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, New Jersey.
- Stărică, C. and Granger, C. (2005), 'Nonstationarities in stock returns', *Rev. Economics Statist.* **87**(3), 503–522.
- Viennet, G. (1997), 'Inequalities for absolutely regular sequences: application to density estimation', *Probab. Theory Relat. Fields* **107**, 467–492.

A Proof of Lemma 1

Denote $f * K_h(x) = \int_{\mathbf{R}^p} K_h(x - y)f(y) dy$. We have for $x \in \mathbf{R}^p$ that

$$f * K_h(x) - f(x) \int_{\mathbf{R}^p} K = \int_{\mathbf{R}^p} K(z) (f(x + hz) - f(x)) dz.$$

Thus

$$\left\| f * K_h - f \int_{\mathbf{R}^p} K \right\|_q \leq \sup_{z \in \text{supp}(K)} \|f(\cdot + hz) - f\|_q \int_{\mathbf{R}^p} |K|,$$

where $\text{supp}(K)$ is the support of K . When f is continuous then the claim holds by the dominated convergence theorem. The claim holds for all $f \in L_q(\mathbf{R}^d)$ because continuous functions are dense in the set of such functions with respect to the L_q -norm. We have proved the lemma.