DOI 10.1007/s10986-020-09480-y
Lithuanian Mathematical Journal, Vol. 60, No. 3, July, 2020, pp. 291–314

On nonparametric ridge estimation for multivariate long-memory processes

Jan Beran [a] **and Klaus Telkmann** [a,b]

[a] Department of Mathematics and Statistics, University of Konstanz, Universitaetsstrasse 10, 78457 Konstanz, Germany
[b] Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, 6210 Donald Bren Hall Irvine, CA 92697-3425, USA
(e-mail: jan.beran@uni-konstanz.de)

Received November 1, 2018; revised November 11, 2019

Abstract. We consider nonparametric estimation of the ridge of a probability density function for multivariate linear processes with long-range dependence. We derive functional limit theorems for estimated eigenvectors and eigenvalues of the Hessian matrix. We use these results to obtain the weak convergence for the estimated ridge and asymptotic simultaneous confidence regions.

MSC: 62M10, 62G05, 62G15, 62G07, 60G18, 60F17

Keywords: kernel density estimation, linear process, long-range dependence, multivariate time series, ridge

1 Introduction

Let $\mu_1, \ldots, \mu_n$ be an i.i.d. sample generated by a probability distribution with density p_μ that has a compact support $M \subset \mathbb{R}^m$. Furthermore, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be generated by an m-dimensional linear process

$$\mathbf{X}_t = \sum_{j=0}^{\infty} A_j \boldsymbol{\varepsilon}_{t-j} \quad (t \in \mathbb{Z}) \tag{1.1}$$

with probability density function p_X, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,m})^{\mathrm{T}} \in \mathbb{R}^m$ denote i.i.d. zero mean random vectors, and A_j are suitable $m \times m$-matrices. Define the process

$$\mathbf{Y}_t = (Y_{t,1}, \ldots, Y_{t,m})^{\mathrm{T}} = \mu_t + \mathbf{X}_t \quad (t \in \mathbb{Z}) \tag{1.2}$$

with corresponding probability density function $p_Y = p_\mu \star p_X$, where $\star$ denotes convolution. For $k < m$, the k-dimensional ridge of p_Y is the set of points that are local maxima of p_Y in at least $m-k$ directions. In this paper, we consider kernel estimation of the ridge under long-memory assumptions. We use the i.i.d. assumption on μ_t for simplicity of presentation. Analogous results can be derived under more general conditions, including correlated or deterministic locations $\mu_t \in M$. In contrast to standard smoothing methods in time series

analysis, the method developed in this paper is very general in the sense that the order in which μ_t traverses M does not have to be known.

Processes defined by (1.2) occur, for instance, in spatio-temporal remote sensing where temporal correlations can be observed even at the level of individual pixels (see, e.g., [35] and references therein). A much discussed issue is, for example, the statical analysis of time series of the so-called Synthetic Aperture Radar (SAR) satellite data (see, e.g., [32]). Other applications, possibly with modified conditions on μ_t, include, for instance, dynamic systems with random perturbations X_t. For example, in [30] parameter estimation in m-dimensional ordinary differential equation (ODE) models is studied, where μ_t follows an ODE with unknown parameters, and observations are of the form $\mathbf{Y}_t = \mu_t + \mathbf{X}_t$ with i.i.d. errors $\mathbf{X}_t$. Uncertainty about the dynamic system, together with random perturbations of observations, often leads to questions that go beyond parameter estimation, including in particular the topological structure of orbits. For further references, see, for example, [30].

Kernel density estimation for long-range dependent linear processes has been studied in [14, 22, 34] and [44], among others (also see [19]; for the i.i.d. case, see, for example, [42] and references therein). Nonparametric estimation of multivariate densities and their derivatives is considered in [8] under i.i.d. assumptions (also see, e.g., [7, 21, 26, 39]). The asymptotic distribution of multivariate kernel density estimators and their derivatives under long-memory assumptions is derived in [6]. A general introduction to long-memory processes can be found, for instance, in [1,16,20] and [4]. For multivariate linear long-memory processes, see, for example, [27] and reference therein. Ridge detection has a long history in image analysis and was introduced in particular by [25]. Further developments can be found in [15,17,24,31,43] and [37], among others. So far statistical inference for density ridges has been considered only under i.i.d. assumptions. For instance, [10,11,12] consider kernel estimation of density ridges. [18] establish bounds on the Hausdorff distance between the true and estimated ridge. [10] calculate confidence regions via bootstrapping. The special case of one-dimensional curves in a two-dimensional Euclidean space is addressed in [38]. In this paper, we consider error processes that exhibit long-range dependence. As it turns out, the assumption of long memory simplifies the construction of simultaneous confidence regions for density ridges.

The paper is organized as follows. General definitions and results on kernel density estimation for multivariate linear long-memory processes are summarized in Section 2. The asymptotic distributions of estimated eigenvectors and eigenvalues of the Hessian matrix are derived in Section 3. Confidence regions for the ridge are obtained in Section 4. Simulation results are discussed in Section 5. Proofs, tables, and figures are given in the appendix.

2 Basic definitions and notation

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a twice continuously differentiable function. The k-dimensional ridge of f is the set of points that are local maxima in at least $m-k$ dimensions. Thus let $\nabla f(\mathbf{y})$ and $\nabla^2 f(\mathbf{y})$ denote the gradient and the Hessian matrix of f, respectively. For a vector $\mathbf{u} \in \mathbb{R}^m$, the directional derivative of f (at $\mathbf{y}$) in direction $\mathbf{u}$ is given by $\partial_u f(\mathbf{y}) = \langle \nabla f(\mathbf{y}), \mathbf{u} \rangle$. A necessary condition for a local maximum in $m-k$ dimensions is that the directional derivatives $\partial_u f(\mathbf{y})$ vanish for at least $m-k$ orthonormal vectors. Since $\nabla^2 f(\mathbf{y})$ is real and symmetric for each $\mathbf{y}$, the spectral theorem guarantees the existence of a set of orthonormal eigenvectors $\mathbf{u}_1(\mathbf{y}), \ldots, \mathbf{u}_m(\mathbf{y})$ of $\nabla^2 f(\mathbf{y})$. Therefore the ridge points are defined in [15] as follows.

DEFINITION 1. Let $f : \mathbb{R}^m \to \mathbb{R}$ be twice continuously differentiable. Denote by $\rho_i(\mathbf{y})$ and $\mathbf{u}_i(\mathbf{y})$ $(i = 1, \ldots, m)$ the eigenvalues and orthonormal eigenvectors of $\nabla^2 f(\mathbf{y})$, and assume that $\rho_1(\mathbf{y}) \geqslant \cdots \geqslant \rho_m(\mathbf{y})$. Then a point $\mathbf{y} \in \mathbb{R}^m$ lies on the k-dimensional ridge of f if

$$\partial_{\mathbf{u}_i} f(\mathbf{y}) = 0 \quad (i = k+1, \ldots, m) \tag{2.1}$$

and

$$\rho_{k+1}(\mathbf{y}) < 0. \tag{2.2}$$

Remark 1. Note that condition (2.2) implies $\rho_{k+2}(\mathbf{y}), \ldots, \rho_m(\mathbf{y}) < 0$ because the eigenvalues are ordered. Thus $f(\mathbf{y})$ is locally maximal in these directions.

The process $\mathbf{Y}_t$ defined in (1.2) has an m-dimensional marginal distribution function

$$F_Y(\mathbf{y}) = \mathbf{P}(\mathbf{Y}_t \leqslant \mathbf{y}) = \mathbf{P}(Y_{t,1} \leqslant y_1,\ \ldots,\ Y_{t,m} \leqslant y_m) \quad \big(\mathbf{y} \in \mathbb{R}^m\big)$$

with density p_Y. We will use the notation ∇p_Y and $\nabla^2 p_Y$ for the gradient and Hessian matrix of p_Y, respectively. Also, $\bar{\mathbf{Y}}_n = n^{-1}\sum_{t=1}^n \mathbf{Y}_t$ is the sample mean,

$$\dot{F}_Y(\mathbf{y}) = \left(\frac{\partial}{\partial y_1}F(\mathbf{y}), \ldots, \frac{\partial}{\partial y_m}F(\mathbf{y})\right)^{\mathrm{T}}$$

is the gradient of F_Y, and the eigenvalues and eigenvectors of $\nabla^2 p_Y(\mathbf{y})$ are denoted by $\lambda_1(\mathbf{y}) \geqslant \cdots \geqslant \lambda_m(\mathbf{y})$ and $v_1(\mathbf{y}), \ldots, v_m(\mathbf{y})$ respectively. Finally, $\mathcal{M}(m, \mathbb{R})$ denotes the set of $m \times m$ matrices with real-valued coefficients, $\mathrm{GL}(m, \mathbb{R})$ is the general linear group, and $I = I_m$ is the $m \times m$ identity matrix. If P is a symmetric positive semidefinite matrix, then we denote its unique symmetric positive semidefinite square root by $P^{1/2}$. For $\eta > 0$ and $M, M_j \in \mathcal{M}(m, \mathbb{R})$ $(j \in \mathbb{N})$, we write

$$M_j \underset{j\to\infty}{\sim} j^{-\eta} M \quad \text{if} \quad \lim_{j\to\infty} j^{\eta} M_j = M.$$

Let $G_k : \mathbb{R}^m \to \mathbb{R}^{m-k}$ be defined by

$$G_k(\mathbf{y}) = \begin{pmatrix} g_{k+1}(\mathbf{y}) \\ \vdots \\ g_m(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} \langle \nabla p_Y(\mathbf{y}), v_{k+1}(\mathbf{y}) \rangle \\ \vdots \\ \langle \nabla p_Y(\mathbf{y}), v_m(\mathbf{y}) \rangle \end{pmatrix}.$$

Condition (2.1) can be written as $G_k(\mathbf{y}) = 0$, and the k-dimensional ridge of p_Y is the set

$$R_k = \big\{\mathbf{y} \in \mathbb{R}^m \;\big|\; G_k(\mathbf{y}) = 0,\ \lambda_{k+1}(\mathbf{y}) < 0\big\}. \tag{2.3}$$

We will use the following assumptions:

(A1) $\mathbf{X}_t$ is a linear process defined by (1.1), μ_t are i.i.d. random vectors with compact support $M \subset \mathbb{R}^m$ and independent of $\mathbf{X}_t$ $(t \in \mathbb{Z})$, and

$$\mathbf{Y}_t = \mu_t + \mathbf{X}_t = \mu_t + \sum_{j=0}^{\infty} A_j \boldsymbol{\varepsilon}_{t-j},$$

where

$$A_0 = I, \qquad A_j = \big[a_{kl}^{(j)}\big]_{k,l=1,\ldots,m} \quad (j \in \mathbb{N}).$$

and $\boldsymbol{\varepsilon}_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,m})^{\mathrm{T}} \in \mathbb{R}^m$ are i.i.d. zero mean random vectors with

$$\Sigma_\varepsilon = \mathrm{var}(\boldsymbol{\varepsilon}_t) \in \mathrm{GL}(m, \mathbb{R}).$$

Moreover,

$$\mathbf{E}[\mu_t] = 0.$$

(A2) There exists a matrix $A_\infty \in \mathrm{GL}(m, \mathbb{R})$ such that

$$A_j \underset{j\to\infty}{\sim} j^{d-1} A_\infty$$

for some $d \in (0, 1/2)$.

(A3) $$\mathbf{E}\big[|\varepsilon_{t,j}|^{4+\kappa}\big] < \infty \quad (j = 1, \ldots, m)$$

for some $\kappa > 0$.

(A4) $$F_\varepsilon(\mathbf{u}) = \mathbf{P}(\boldsymbol{\varepsilon}_t \leqslant \mathbf{u}) = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_m} p_\varepsilon(\mathbf{y})\,\mathrm{d}y_1 \cdots \mathrm{d}y_m \quad \big(\mathbf{u} \in \mathbb{R}^m\big).$$

(A5) The density function p_ε is infinitely differentiable with all bounded and square-integrable partial derivatives.

(A6) F_Y is infinitely differentiable with all bounded and square-integrable partial derivatives.

To derive asymptotic expressions for the kernel density estimator and its derivatives, an extension of the reduction principle for empirical processes is required. Denote by

$$F_{Y,n}(\mathbf{y}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\mathbf{Y}_t \leqslant \mathbf{y}\}$$

the empirical distribution function of $\mathbf{Y}_t$ $(t \in \mathbb{Z})$.

Theorem 1. *Under* (A1)–(A6)*, we have*

$$n^{1/2-d} \sup_{\mathbf{y}\in\mathbb{R}^m} \big|F_{Y,n}(\mathbf{y}) - F_Y(\mathbf{y}) + \dot{F}_Y^{\mathrm{T}}(\mathbf{y})\bar{\mathbf{Y}}_n\big| \xrightarrow[\mathrm{P}]{} 0.$$

Remark 2. Assume without loss of generality that $\mathbf{E}[\mu_t] = 0$ and note that $n^{1/2-d}\bar{\mathbf{Y}}_n$ converges in distribution to a zero-mean Gaussian random variable with covariance matrix V (see [13, Thm. 1]). Theorem 1 implies

$$n^{1/2-d}\big[F_{Y,n}(\mathbf{y}) - F_Y(\mathbf{y})\big] \Rightarrow -\dot{F}_Y^{\mathrm{T}}(\mathbf{y})V^{1/2}\boldsymbol{\xi},$$

where "$\Rightarrow$" denotes weak convergence in Skorokhod space $D(-\infty,\infty)$ equipped with the supremum norm, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ is a vector of i.i.d. standard normal random variables. The covariance matrix V is of the form

$$V = c(d) \cdot A_\infty \Sigma_\varepsilon A_\infty^{\mathrm{T}},$$

where [13]

$$c(d) = \frac{1}{d^2}\left\{\frac{1}{1+2d} + \int_{-\infty}^{0} \big[(1-t)^d - (-t)^d\big]^2\,\mathrm{d}t\right\}.$$

In the following let $K : \mathbb{R}^m \to \mathbb{R}_+$ be a kernel function, H a symmetric and positive definite $m \times m$ bandwidth matrix and $|H|$ its determinant. A kernel density estimator of p_Y is defined by

$$\hat{p}_Y(\mathbf{y}) = \frac{1}{n|H|^{1/2}}\sum_{t=1}^{n} K\big(H^{-1/2}(\mathbf{y} - \mathbf{Y}_t)\big) \quad \big(\mathbf{y} \in \mathbb{R}^m\big).$$

More generally, using Kronecker powers, all rth partial derivatives of p_Y may be stacked into an m^r-dimensional vector $p_Y^{(r)}$ $(r \geqslant 0)$. More specifically,

$$p_Y^{(r)}(\mathbf{y}) = D^{\otimes r} p_Y(\mathbf{y}) = \frac{\partial p_Y}{(\partial \mathbf{y})^{\otimes r}} \in \mathbb{R}^{m^r}.$$

The corresponding kernel estimator is of the form

$$\begin{aligned}\hat{p}_Y^{(r)}(\mathbf{y}) &= D^{\otimes r}\hat{p}_Y(\mathbf{y}) \\ &= \frac{1}{n|H|^{1/2}}\big(H^{-1/2}\big)^{\otimes r}\sum_{t=1}^{n} D^{\otimes r}K\big(H^{-1/2}(\mathbf{y}-\mathbf{Y}_t)\big) \in \mathbb{R}^{m^r}.\end{aligned}$$

For details on the Kronecker product, we refer to [33] and [28], and in the context of multivariate kernel density derivative estimation, we refer to [7]. In particular, the gradient of a real-valued function f is given by $\nabla f = D^{\otimes 1}f$, whereas the relationship between the Hessian matrix $\nabla^2 f$ and $D^{\otimes 2}f$ can be expressed by

$$\operatorname{vec}\big(\nabla^2 f\big) = D^{\otimes 2}f,$$

where the vectorization operator vec stacks the elements $\nabla^2 f$ columnwise into a vector. We will use the following assumptions on K and H:

(K1) $-\infty < K(\mathbf{u}) < \infty$.
(K2) K is a symmetric kernel of order $\nu \geqslant 2$, that is, $K(-\mathbf{u}) = K(\mathbf{u})$, and

$$\int_{\mathbb{R}^m} K(\mathbf{u})\,\mathrm{d}\mathbf{u} = 1, \qquad \int_{\mathbb{R}^m} K(\mathbf{u})\mathbf{u}^{\otimes\nu}\,\mathrm{d}\mathbf{u} \neq 0,$$

$$\int_{\mathbb{R}^m} K(\mathbf{u})\mathbf{u}^{\otimes j}\,\mathrm{d}\mathbf{u} = 0 \quad (j = 1,\dots,\nu-1).$$

(K3) K is infinitely differentiable, and all partial derivatives are square integrable.
(K4) There is a compact set $\Omega_K \subset \mathbb{R}^m$ such that $K(\mathbf{u}) = 0$ ($\mathbf{u} \notin \Omega_K$).
(K5) $H = H(n) = [h_{jl}]_{j,l=1,\dots,m}$ is a symmetric positive definite matrix such that

$$\lim_{n\to\infty} \max_{j,l=1,\dots,m} |h_{jl}| = 0.$$

Some additional notation will be needed in the following. Define the process

$$\hat{\mathbf{Y}}_t = \mathbf{Y}_t - \boldsymbol{\varepsilon}_t = \mu_t + \sum_{j=1}^{\infty} A_j \boldsymbol{\varepsilon}_{t-j} \quad (t \in \mathbb{Z})$$

with sample mean $\bar{\hat{\mathbf{Y}}}_n$. Since $A_0 = I$, it is easy to verify that the asymptotic behavior of $n^{1/2-d}\bar{\mathbf{Y}}_n$ and $n^{1/2-d}\bar{\hat{\mathbf{Y}}}_n$ is the same. In particular, denote by $F_{\hat{Y},n}$, $F_{\hat{Y}}$, and $\dot{F}_{\hat{Y}}$ the empirical distribution function, the marginal distribution function, and the gradient of the marginal distribution function of $\hat{\mathbf{Y}}_t$, respectively. Then Theorem 1 implies

$$n^{1/2-d}\big[F_{\hat{Y},n}(\mathbf{y}) - F_{\hat{Y}}(\mathbf{y})\big] \Rightarrow -\dot{F}_{\hat{Y}}^{\mathrm{T}}(\mathbf{y})V^{1/2}\boldsymbol{\xi},$$

where V and $\boldsymbol{\xi}$ are as in Remark 2. For $\mathbf{y} \in \mathbb{R}^m$, we define $\mathbf{w}(\mathbf{z}) = (w_1(\mathbf{z}),\dots,w_m(\mathbf{z}))^{\mathrm{T}} \in \mathbb{R}^m$, $\mathbf{b}_j^{(r)}(\mathbf{y}) \in \mathbb{R}^{m^r}$ ($j = 1,\dots,m$) and $M(\mathbf{y}) \in \mathcal{M}(m^r,\mathbb{R})$ by

$$\mathbf{w}(\mathbf{z}) = V^{1/2}\dot{F}_{\hat{Y}}(\mathbf{z}) \in \mathbb{R}^m,$$

$$\mathbf{b}_j^{(r)}(\mathbf{y}) = \int w_j(\mathbf{z}) \frac{\partial^m}{\partial z_1 \dots \partial z_m} p_\varepsilon^{(r)}(\mathbf{y}-\mathbf{z})\, \mathrm{d}z_1 \cdots \mathrm{d}z_m \in \mathbb{R}^{m^r},$$

and

$$M(\mathbf{y}) = p_Y(\mathbf{y}) \int \big[D^{\otimes r} K(\mathbf{u})\big] \big[D^{\otimes r} K(\mathbf{u})\big]^{\mathrm{T}} \mathrm{d}u_1 \cdots \mathrm{d}u_m.$$

Note that $M(\mathbf{y})$ is symmetric and positive semidefinite for every $\mathbf{y}$. Thus we denote by $M^{1/2}(\mathbf{y})$ its positive semidefinite matrix square root. Furthermore, we will write "$\Rightarrow_{\mathrm{C_b}}$" for weak convergence in the space of bounded continuous functions and "$\to_{\mathrm{d}}$" for (pointwise) convergence in distribution.

The bias of $\hat{p}_Y^{(r)}$ does not depend on the autocorrelation structure. The asymptotic distribution of $\hat{p}_Y^{(r)} - \mathbf{E}[\hat{p}_Y^{(r)}]$ follows from Theorem 1 and the results in [6].

Theorem 2. *For $r \geqslant 0$, let Z_i $(i = 1, \dots, m^r)$ and $\xi_1, \dots, \xi_m$ be i.i.d. standard normal random variables and set $\boldsymbol{Z} = (Z_1, \dots, Z_{m^r})^{\mathrm{T}}$. Also, let $\lambda_{H,1}(n), \dots, \lambda_{H,m}(n) > 0$ denote the eigenvalues of $H = H(n)$. Then, under* (A1)–(A6) *and* (K1)–(K5)*, we have*:

(i) *If*

$$\lim_{n\to\infty} n^{2d/(m+2r)} \max_{j=1,\dots,m} \lambda_{H,j}^{1/2}(n) = 0, \tag{2.4}$$

then, for any fixed $\mathbf{y} \in \mathbb{R}^m$,

$$\sqrt{n|H|^{1/2}} \big(H^{1/2}\big)^{\otimes r} \big[\hat{p}_Y^{(r)}(\mathbf{y}) - \mathbf{E}\big[\hat{p}_Y^{(r)}(\mathbf{y})\big]\big] \underset{\mathrm{d}}{\longrightarrow} M^{1/2}(\mathbf{y}) \boldsymbol{Z}.$$

(ii) *If*

$$\lim_{n\to\infty} n^{2d/(m+2r)} \min_{j=1,\dots,m} \lambda_{H,j}^{1/2}(n) = \infty, \tag{2.5}$$

then

$$n^{1/2-d} \big[\hat{p}_Y^{(r)}(\mathbf{y}) - \mathbf{E}\big[\hat{p}_Y^{(r)}(\mathbf{y})\big]\big] \underset{\mathrm{C_b}(\mathbb{R}^m, \mathbb{R}^{m^r})}{\Longrightarrow} \boldsymbol{\xi}^{(r)}(\mathbf{y}), \tag{2.6}$$

where

$$\boldsymbol{\xi}^{(r)}(\mathbf{y}) = (-1)^{m+1} \sum_{j=1}^{m} \mathbf{b}_j^{(r)}(\mathbf{y}) \xi_j \quad \big(\mathbf{y} \in \mathbb{R}^m\big).$$

Remark 3. We call a bandwidth H small if conditions (2.4) and (K5) are met and large if (2.5) and (K5) hold. Since $2d/(m+2r) < 1$ for all $d \in (0, 1/2)$, conditions (2.5) and (K5) imply that the elements of $H^{1/2}$ converge to zero rather slowly. Note that for large bandwidths, we have a functional limit theorem. This is in sharp contrast to small bandwidths, where only pointwise convergence can be achieved. Moreover, the limiting process obtained for large bandwidths is degenerate. Given the standard normal random variables $\xi_1, \dots, \xi_m$, the deterministic functions $\mathbf{b}_j^{(r)}(\cdot)$ fully determine how the sample path of $\boldsymbol{\xi}^{(r)}(\mathbf{y})$ changes as a function of $\mathbf{y}$. Note also that $\xi_1, \dots, \xi_m$ are the same random variables as in Remark 2.

Remark 4. Conditions (2.4) and (2.5) follow from an additive decomposition of $\hat{p}_Y^{(r)}(\mathbf{y}) - \mathbf{E}[\hat{p}_Y^{(r)}(\mathbf{y})]$ into a martingale and a long memory part. Consider, for instance, $H = \mathrm{diag}(h, \dots, h)$. Then the martingale part is of order $O_p(n^{-1/2} h^{-m/4} h^{-r/2})$, whereas the long memory part is of order $O_p(n^{d-1/2})$. If (2.4) holds, then the order of the martingale part is larger, whereas the opposite is true under condition (2.5). Note also that there is a typo in [6] since the cases $r \geqslant 1$ and $m \geqslant 2$ are not taken into account in the conditions for small and large bandwidths.

Remark 5. For simplicity of presentation, the asymptotic expressions in Theorems 1 and 2 are derived under the assumption that the random vectors μ_t are i.i.d. The same results can be derived for short-range dependent processes $\mu_t \in M$ under mild regularity conditions.

Remark 6. Assumption (K4) is needed to apply multivariate integration by parts in the proof of Theorem 2. This assumption can be weakened by assuming that the kernel decays fast enough so that the boundary terms in the divergence theorem are asymptotically negligible. For instance, we can use Gaussian kernels.

To determine the bias $C_n^{(r)}(\mathbf{y}) = \mathbf{E}[\hat{p}_Y^{(r)}(\mathbf{y})] - p_Y^{(r)}(\mathbf{y})$, we will use the following notation: For $f, g\colon \mathbb{R}^p \to \mathbb{R}^q$, we write $f(x) = o(g(x))$ if $\|f(x)\| = o(\|g(x)\|)$. Then $C_n^{(r)}$ can be written asymptotically as follows (see, e.g., [7, 8]):

Lemma 1. *Under the assumptions of Theorem* 2*,*

$$C_n^{(r)}(\mathbf{y}) = \frac{1}{\nu!}\biggl[I_{m^r} \otimes \biggl(\int\limits_{\mathbb{R}^m} K(\mathbf{w})\mathbf{w}^{\otimes\nu}\,\mathrm{d}\mathbf{w}\biggr)^{\mathrm{T}}\biggr]\bigl[I_{m^r} \otimes \bigl(-H^{1/2}\bigr)^{\otimes\nu}\bigr]p_Y^{(r+\nu)}(\mathbf{y}) + r_n,$$

where

$$r_n = o\biggl(\frac{1}{\nu!}\bigl[I_{m^r} \otimes \bigl(-\mathbf{w}^{\mathrm{T}}H^{1/2}\bigr)^{\otimes\nu}\bigr]p_Y^{(r+\nu)}(\mathbf{y})\biggr).$$

In the following, we will use the notation $\xi_n \doteq \eta_n$ if $\zeta_n = \xi_n\eta_n^{-1}$ converges in distribution to a stochastically bounded random variable or random vector ζ with $\mathbf{P}(\zeta \neq 0) > 0$. To understand the practical implications of Theorem 2 and Lemma 1, consider for simplicity a diagonal bandwidth matrix $H = \operatorname{diag}(h_1, \dots, h_m)$ with $h_1 \doteq h_2 \doteq \cdots \doteq h_m = h$. For a discussion in the univariate case, see [5]. From Theorem 2 (and its proof) and Lemma 1 we obtain

$$\hat{p}_Y^{(r)}(\mathbf{y}) - p_Y^{(r)}(\mathbf{y}) = A_n^{(r)}(\mathbf{y}) + B_n^{(r)}(\mathbf{y}) + C_n^{(r)}(\mathbf{y}),$$

where $A_n^{(r)}$ and $B_n^{(r)}$ denote the martingale and long-memory components of $\hat{p}_Y^{(r)} - \mathbf{E}[\hat{p}_Y^{(r)}]$, respectively. More specifically,

$$A_n^{(r)}(\mathbf{y}) = \frac{1}{n|H|^{1/2}}\bigl(H^{-1/2}\bigr)^{\otimes r}\sum_{t=1}^{n}\bigl\{\tau_r(\mathbf{Y}_t, \mathbf{y}) - E\bigl[\tau_r(\mathbf{Y}_t, \mathbf{y}) \bigm| \boldsymbol{\varepsilon}_s, s \leqslant t-1\bigr]\bigr\},$$

where

$$\tau_r(\mathbf{Y}_t, \mathbf{y}) = D^{\otimes r}K\bigl(H^{-1/2}(\mathbf{y} - \mathbf{Y}_t)\bigr)$$

and

$$B_n^{(r)}(\mathbf{y}) = \hat{p}_Y^{(r)}(\mathbf{y}) - \mathbf{E}\bigl[\hat{p}_Y^{(r)}(\mathbf{y})\bigr] - A_n^{(r)}(\mathbf{y}).$$

The asymptotic orders of the three terms are

$$A_n^{(r)}(\mathbf{y}) \doteq n^{-1/2}h^{-m/4-r/2}, \qquad B_n^{(r)}(\mathbf{y}) \doteq n^{d-1/2}, \qquad C_n^{(r)}(\mathbf{y}) \doteq h^{1/2\nu}.$$

The order of $B_n^{(r)}$ does not depend on h. Therefore h (or more generally H) is called asymptotically optimal if

$$\max\bigl\{A_n^{(r)}(\mathbf{y}), C_n^{(r)}(\mathbf{y})\bigr\} = o_p\bigl(n^{d-1/2}\bigr). \tag{2.7}$$

Due to the trade-off between bias and variance, the smallest order of $A_n^{(r)} + C_n^{(r)}$ is obtained for $h \doteq n^{-2/(2\nu+m+2r)}$, which leads to

$$\max\{A_n^{(r)}(\mathbf{y}), C_n^{(r)}(\mathbf{y})\} \doteq n^{-\nu/(2\nu+m+2r)}.$$

Thus (2.7) can be achieved if and only if

$$d > g(\nu, m, r) = \frac{1}{2} - \frac{\nu}{2\nu + m + 2r}. \tag{2.8}$$

Since $g(\nu, m, r)$ decreases to zero monotonically as a function of ν, and the range of d is bounded from above by $1/2$, (2.8) can always be satisfied by choosing ν large enough. For a general bandwidth matrix H with eigenvalues $\lambda_{H,1}, \ldots, \lambda_{H,n} \doteq h$, analogous arguments apply. In the following, we will therefore use the additional assumption:

(K6) The order ν of the kernel and the bandwidth matrix are such that (2.7) holds.

Remark 7. Note that the bound for the bias $C_n^{(r)}(\mathbf{y})$ is uniform in $\mathbf{y}$, provided that $p_Y^{(r)}$ is smooth enough. The order of the bias can therefore be reduced uniformly by choosing ν large enough. Moreover, (2.6) implies that, for bandwidths satisfying (2.5), $A_n^{(r)}(\mathbf{y})$ is uniformly negligible. Thus, for kernels with a sufficiently large order ν, (2.7) can be achieved uniformly.

3 Asymptotic results for eigenvectors and eigenvalues

The detection of ridge points involves checking conditions on eigenvectors and eigenvalues of the Hessian matrix. We therefore need asymptotic results for estimators of these quantities. As before, we write the derivatives of p_Y as vectors. In particular, $D^{\otimes 2} f = \mathrm{vec}(\nabla^2 f) \in \mathbb{R}^{m^2}$. Denoting by vec^{-1} the operator reversing this vectorization, the Hessian matrix of p_Y at $\mathbf{y} \in \mathbb{R}^m$ can be written as

$$\nabla^2 p_Y(\mathbf{y}) = \mathrm{vec}^{-1}\left(p_Y^{(2)}(\mathbf{y})\right).$$

Similarly, we will write $\nabla^2 \hat{p}_Y(\mathbf{y}) = \mathrm{vec}^{-1}(\hat{p}_Y^{(2)}(\mathbf{y}))$. Denote by $\lambda_1(\mathbf{y}) \geqslant \cdots \geqslant \lambda_m(\mathbf{y})$ and $v_1(\mathbf{y}), \ldots, v_m(\mathbf{y})$ the eigenvalues and eigenvectors of $\nabla^2 p_Y(\mathbf{y})$ and by $\hat{\lambda}_1(\mathbf{y}) \geqslant \cdots \geqslant \hat{\lambda}_m(\mathbf{y})$ and $\hat{v}_1(\mathbf{y}), \ldots, \hat{v}_m(\mathbf{y})$ the corresponding quantities for $\nabla^2 \hat{p}_Y(\mathbf{y})$. The following specific assumptions will be used:

(V1) For all $i, j = 1, \ldots, m$ and $\mathbf{y} \in \mathbb{R}^m$, $\langle v_i(\mathbf{y}), v_j(\mathbf{y})\rangle = \delta_{ij}$, $\langle \hat{v}_i(\mathbf{y}), \hat{v}_j(\mathbf{y})\rangle = \delta_{ij}$.
(V2) For $l = 1, \ldots, m$, $\mathrm{sign}\big(\langle v_l(\mathbf{y}), \hat{v}_l(\mathbf{y})\rangle\big) = 1$.

Since $\nabla^2 p_Y(\mathbf{y})$ and $\nabla^2 \hat{p}_Y(\mathbf{y})$ are symmetric, the spectral theorem guarantees the existence of an orthonormal system of eigenvectors as in (V1). Note that the eigenvectors are determined only up to a sign. Condition (V2) ensures without loss of generality that the estimated eigenvectors are adjusted in the correct direction.

The perturbation behavior of eigenvectors corresponding to nonsimple eigenvalues can be rather complicated (see, e.g., [40]). The reason is that $v_l(\cdot)$ may not be continuous at points $\mathbf{y}$ where $\lambda_l(\mathbf{y})$ is nonsimple. We therefore exclude this case. Let $\mathrm{supp}(p_Y)$ be the support of p_Y, define

$$\Lambda_k := \left\{\mathbf{y} \in \mathrm{supp}(p_Y) \;\middle|\; \lambda_k(\mathbf{y}) > \cdots > \lambda_m(\mathbf{y})\right\} \subseteq \mathbb{R}^m,$$

and denote by Λ_k^0 the interior of Λ_k. Then, restricted to Λ_k^0, the functions $v_l(\cdot)$ $(l = k+1, \ldots, m)$ are continuous. For $\mathbf{y} \in \Lambda_k^0$, $l = k+1, \ldots, m$ and $j = 1, \ldots, m$, we then define

$$\mathbf{B}_j(\mathbf{y}) = \mathrm{vec}^{-1}\left(\mathbf{b}_j^{(2)}(\mathbf{y})\right) \in \mathcal{M}(m, \mathbb{R}),$$

$$c_{\lambda;l,j}(\mathbf{y}) = v_l^{\mathrm{T}}(\mathbf{y})\mathbf{B}_j(\mathbf{y})v_l(\mathbf{y}) \in \mathbb{R},$$

$$c_{v;l,j}(\mathbf{y}) = \sum_{\substack{1\leqslant i\leqslant m\\ i\neq l}} \frac{1}{\lambda_l(\mathbf{y}) - \lambda_i(\mathbf{y})} v_i(\mathbf{y})v_i^{\mathrm{T}}(\mathbf{y})\mathbf{B}_j(\mathbf{y})v_l(\mathbf{y}) \in \mathbb{R}^m,$$

and the processes

$$\xi_{\nabla p_Y}(\mathbf{y}) = (-1)^{m+1}\sum_{j=1}^m \mathbf{b}_j^{(1)}(\mathbf{y})\xi_j, \xi_{\nabla^2 p_Y}(\mathbf{y}) = (-1)^{m+1}\sum_{j=1}^m \mathbf{B}_j(\mathbf{y})\xi_j \quad (\mathbf{y}\in\mathbb{R}^m),$$

$$\xi_{\lambda_l}(\mathbf{y}) = (-1)^{m+1}\sum_{j=1}^m c_{\lambda;l,j}(\mathbf{y})\xi_j \in \mathbb{R} \quad (\mathbf{y}\in \Lambda_k^0),$$

and

$$\xi_{v_l}(\mathbf{y}) = (-1)^{m+1}\sum_{j=1}^m c_{v;l,j}(\mathbf{y})\xi_j \in \mathbb{R}^m \quad (\mathbf{y}\in \Lambda_k^0),$$

where ξ_j are the same i.i.d. $N(0,1)$ random variables as in Theorem 2. Under assumption (K6), Theorem 2 implies

$$n^{1/2-d}\big[\nabla^2\hat{p}_Y(\mathbf{y}) - \nabla^2 p_Y(\mathbf{y})\big] \underset{C_b}{\Longrightarrow} \xi_{\nabla^2 p_Y}(\mathbf{y}),$$

where "$\Rightarrow_{C_b}$" denotes weak convergence in the space $C_b(\mathbb{R}^m, \mathcal{M}(m,\mathbb{R}))$ of bounded continuous functions $f : \mathbb{R}^m \to \mathcal{M}(m,\mathbb{R})$. The asymptotic distribution of estimated eigenvalues and eigenvectors is given in the following theorem.

Theorem 3. *Suppose that* (K6), (V1), (V2), *and the assumptions of Theorem* 2 *hold, with* $r = 2$ *in* (2.5) *equal to* 2. *Then, for* $l = k+1,\dots,m$,

$$n^{1/2-d}\big[\hat{\lambda}_l(\mathbf{y}) - \lambda_l(\mathbf{y})\big] \underset{C_b(\Lambda_k^0,\mathbb{R})}{\Longrightarrow} \xi_{\lambda_l}(\mathbf{y}) \tag{3.1}$$

and

$$n^{1/2-d}\big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y})\big] \underset{C_b(\Lambda_k^0,\mathbb{R}^m)}{\Longrightarrow} \xi_{v_l}(\mathbf{y}).$$

Remark 8. Theorem 3 means that, under long-memory assumptions, weak convergence of kernel estimators of the Hessian matrix carries over to estimated eigenvalues and eigenvectors. In contrast, for independent or short-range dependent observations, we have no functional limit theorem for kernel estimators and therefore also no weak convergence result for eigenvalues and eigenvectors. Note also that the limiting processes in Theorem 3 are degenerate in the sense that the coefficients $c_{\lambda;l,j}(\cdot)$ and $c_{v;l,j}(\cdot)$, respectively, determine how the sample paths change as functions of $\mathbf{y}$.

Remark 9.

$$\lim_{n\to\infty} n^{1-2d}\operatorname{cov}\big(\hat{\lambda}_l(\mathbf{y}), \hat{\lambda}_l(\tilde{\mathbf{y}})\big) = \sigma_{\lambda_l}^2(\mathbf{y},\tilde{\mathbf{y}}) = \sum_{j=1}^m c_{\lambda;l,j}(\mathbf{y})c_{\lambda;l,j}(\tilde{\mathbf{y}}),$$

$$\lim_{n\to\infty} n^{1-2d}\operatorname{var}\big(\hat{v}_l(\mathbf{y})\big) = \Sigma_{v_l}(\mathbf{y}) = \sum_{j=1}^m c_{v;l,j}(\mathbf{y})c_{v;l,j}^{\mathrm{T}}(\mathbf{y}).$$

4 Ridge estimation

Following the definition of R_k in (2.3), the estimated ridge is given by

$$\hat{R}_k = \big\{\mathbf{y} \in \mathbb{R}^m \;\big|\; \hat{G}_k(\mathbf{y}) = 0,\ \hat{\lambda}_{k+1}(\mathbf{y}) < 0\big\},$$

where

$$\hat{G}_k(\mathbf{y}) = \begin{pmatrix} \hat{g}_{k+1}(\mathbf{y}) \\ \vdots \\ \hat{g}_m(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} \langle \nabla \hat{p}_Y(\mathbf{y}), \hat{v}_{k+1}(\mathbf{y})\rangle \\ \vdots \\ \langle \nabla \hat{p}_Y(\mathbf{y}), \hat{v}_m(\mathbf{y})\rangle \end{pmatrix}.$$

To obtain confidence regions for R_k, we derive a functional limit theorem for $\hat{G}_k(\mathbf{y})$. We will use the notation

$$\zeta_l(\mathbf{y}) = \xi^{\mathrm{T}}_{\nabla p_Y}(\mathbf{y}) v_l(\mathbf{y}) + \xi^{\mathrm{T}}_{v_l}(\mathbf{y}) \nabla p_Y(\mathbf{y}) \quad (l = k+1, \ldots, m,\ \mathbf{y} \in \Lambda^0_k)$$

and

$$\zeta(\mathbf{y}) = \big[\zeta_{k+1}(\mathbf{y}), \ldots, \zeta_m(\mathbf{y})\big]^{\mathrm{T}} \in \mathbb{R}^{m-k}.$$

Theorem 4. *Suppose that* (K6), (V1), and (V2) *hold. Moreover, let* H_j $(j = 1, 2)$ *be the bandwidth matrices used for calculating* $\nabla \hat{p}_Y(\mathbf{y})$ *and* $\nabla^2 \hat{p}_Y(\mathbf{y})$*, respectively. Assume that the assumptions of Theorem* 2 *hold, with* r *in* (2.5) *equal to* 1 *for* H_1 *and equal to* 2 *for* H_2*. Then*

$$n^{1/2-d}\big[\hat{G}_k(\mathbf{y}) - G_k(\mathbf{y})\big] \underset{C_b}{\Longrightarrow} \zeta(\mathbf{y}),$$

where "$\Rightarrow_{C_b}$*" denotes weak convergence in* $C_b(\Lambda^0_k, \mathbb{R}^{m-k})$.

Remark 10. Theorem 4 means that long memory leads to a functional limit theorem for $\hat{G}_k$ with a degenerate limiting process. In contrast, under short-range dependence, a functional limit theorem cannot be obtained. Note in particular that setting

$$c_{l,j}(\mathbf{y}) = v_l^{\mathrm{T}}(\mathbf{y}) b_j^{(1)}(\mathbf{y}) + c^{\mathrm{T}}_{v;l,j}(\mathbf{y}) \nabla p_Y(\mathbf{y}),$$

$\zeta_l(\mathbf{y})$ can be written as

$$\zeta_l(\mathbf{y}) = (-1)^{m+1} \sum_{j=1}^{m} c_{l,j}(\mathbf{y}) \xi_j,$$

where ξ_j are the same i.i.d. $N(0,1)$ random variables as in Theorem 2.

Remark 11.

$$\Sigma_\zeta(\mathbf{y}, \tilde{\mathbf{y}}) = \big[\mathrm{cov}\big(\zeta_{l_1}(\mathbf{y}), \zeta_{l_2}(\tilde{\mathbf{y}})\big)\big]_{l_1,l_2=k+1,\ldots,m} = \big[\sigma_{l_1 l_2}(\mathbf{y}, \tilde{\mathbf{y}})\big]_{l_1,l_2=k+1,\ldots,m}$$

with

$$\sigma_{l_1 l_2}(\mathbf{y}, \tilde{\mathbf{y}}) = \sum_{j=1}^{m} c_{l_1,j}(\mathbf{y}) c_{l_2,j}(\tilde{\mathbf{y}}).$$

Consider the decomposition

$$\Lambda^0_k = \mathcal{G}_k \cup \mathcal{G}_k^C,$$

where

$$\mathcal{G}_k = \big\{\mathbf{y} \in \Lambda_k^0 \;\big|\; G_k(\mathbf{y}) = 0\big\}$$

and

$$\mathcal{G}_k^C = \big\{\mathbf{y} \in \Lambda_k^0 \;\big|\; G_k(\mathbf{y}) \neq 0\big\}.$$

The following corollary is useful for obtaining simultaneous confidence regions for R_k.

Corollary 1. *Suppose the assumptions of Theorem* 4 *hold. Then*

$$\sup_{\mathbf{y}\in\mathcal{G}_k} \big\| n^{1/2-d}\big[\hat{G}_k(\mathbf{y}) - G_k(\mathbf{y})\big] - \zeta(\mathbf{y})\big\| \xrightarrow[\mathrm{P}]{} 0$$

and

$$\sup_{\mathbf{y}\in\mathcal{G}_k^C} \big\| n^{1/2-d}\big[\hat{G}_k(\mathbf{y}) - G_k(\mathbf{y})\big] - \zeta(\mathbf{y})\big\| \xrightarrow[\mathrm{P}]{} 0.$$

Corollary 1 is an immediate consequence of Theorem 4. It implies that $n^{1/2-d}\hat{G}_k(\mathbf{y})$ converges to a zero-mean random variable if $G_k(\mathbf{y}) = 0$. On the other hand, if $G_k(\mathbf{y}) \neq 0$, then

$$\big\| n^{1/2-d}\hat{G}_k(\mathbf{y})\big\| \xrightarrow[\mathrm{P}]{} \infty.$$

Assuming that $R_k \subset \Lambda_k^0$, we may write

$$R_k = \mathcal{G}_k \cap \big\{\mathbf{y} \in \Lambda_k^0 \;\big|\; \lambda_{k+1}(\mathbf{y}) < 0\big\}.$$

We therefore adopt an approach based on testing whether directional derivatives are zero and $\lambda_{k+1}(\mathbf{y}) \leqslant 0$. If $\Sigma_\zeta(\mathbf{y},\mathbf{y})$ is invertible, then a simultaneous $(1-\alpha)$-confidence region for points $\mathbf{y} \in \mathcal{G}_k$ can be defined by

$$A_\alpha = \big\{\mathbf{y} \;\big|\; \hat{G}_k^{\mathrm{T}}(\mathbf{y})\big[\Sigma_\zeta(\mathbf{y},\mathbf{y})\big]^{-1}\hat{G}_k(\mathbf{y}) \leqslant n^{2d-1}\chi^2_{m-k}(1-\alpha)\big\},$$

where $\chi^2_{m-k}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with $m-k$ degrees of freedom. Similarly, based on (3.1), we define a simultaneous $(1-\alpha)$-confidence set for points with $\lambda_{k+1}(\mathbf{y}) \leqslant 0$ by

$$B_\alpha = \big\{\mathbf{y} \;\big|\; \hat{\lambda}_{k+1}(\mathbf{y}) \leqslant n^{d-1/2}\sigma_{\lambda_{k+1}}(\mathbf{y},\mathbf{y})\Phi^{-1}(1-\alpha)\big\},$$

where Φ denotes the cumulative standard normal distribution. A ridge point has to satisfy both conditions. Using a Bonferroni correction, we therefore define a simultaneous $(1-\alpha)$-confidence region for R_k by

$$C_\alpha = A_{\alpha/2} \cap B_{\alpha/2}.$$

Remark 12. The simple construction of simultaneous confidence regions for density ridges relies on the functional limit theorems obtained (Theorems 3 and 4). These results are only valid under long-memory assumptions. For weakly dependent error processes $\mathbf{X}_t$, only pointwise convergence can be obtained.

Remark 13. The Bonferroni correction is used here for multiple testing with two simultaneous tests only, thus replacing α by $\alpha/2$ in the individual tests. For practical purposes, this is reasonable and justifies the use of such a simple correction. The development of more refined methods that would be simple to implement would however be worth investigating in future research. For instance, in principle, it would be possible to design a test based on the joint distribution of $\hat{G}_k(\mathbf{y})$ and $\hat{\lambda}_{k+1}(\mathbf{y})$.

5 Simulations

The asymptotic results are illustrated by a small simulation study. We consider the following models:

Model 1. $\mathbf{Y}_t = \mu_t + \mathbf{X}_t$ where μ_t ($t = 1, \ldots, n$) are points on a spiral parameterized by arc length, that is,

$$\mu_t = \mu(u_t) = \begin{pmatrix} \cos\sqrt{2u_t} + \sqrt{2u_t}\sin\sqrt{2u_t} - e_1 \\ \sin\sqrt{2u_t} - \sqrt{2u_t}\cos\sqrt{2u_t} - e_2 \end{pmatrix},$$

where u_t are i.i.d. uniformly distributed on $[0, (2\pi)^2/2]$, and e_1, e_2 are constants such that $E[\mu_t] = 0$. The error process $\mathbf{X}_t$ is of the form $\mathbf{X}_t = (X_{t,1}, X_{t,2})^{\mathrm{T}} \in \mathbb{R}^2$, where $X_{t,1}$ and $X_{t,2}$ are independent univariate Gaussian FARIMA$(0, d, 0)$ processes with variance $\sigma_X^2 = 4$.

Model 2. $\mathbf{Y}_t = \mu_t + \mathbf{X}_t$ with

$$\mu_t = \mu(u_t) = \frac{3}{2}\begin{pmatrix} \sin u_t \\ \cos u_t \end{pmatrix} \quad (0 \leqslant u_t \leqslant 2\pi),$$

$$\mu_t = \mu(u_t) = \frac{5}{2}\begin{pmatrix} \sin u_t \\ \cos u_t \end{pmatrix} \quad (2\pi < u_t \leqslant 4\pi),$$

where u_t are i.i.d. uniformly distributed on $[0, 4\pi]$. The error process $\mathbf{X}_t$ is of the form $\mathbf{X}_t = (X_{t,1}, X_{t,2})^{\mathrm{T}} \in \mathbb{R}^2$, where $X_{t,1}$ and $X_{t,2}$ are independent univariate Gaussian FARIMA$(0, d, 0)$ processes with variance $\sigma_X^2 = 0.1$.

In both cases, we use a bivariate Gaussian kernel. Following a simple adaptation of Silverman's rule of thumb for multivariate kernel density estimators (see, e.g., [8]), we choose the scalar bandwidth matrices $H_0 = h_0 I$ (for the density), $H_1 = h_1 I$ (for the gradient), and $H_2 = h_2 I$ (for the Hessian matrix) such that $h_0, h_1, h_2 > 0$ are proportional to $\sigma_X n^{-2/6}$, $\sigma_X n^{-2/8}$, and $\sigma_X n^{-2/10}$, respectively. Since we use a second-order kernel, condition (2.8) implies that the asymptotic results above are applicable for $d > 0.3$. For $d = 0.35$, 0.4, and 0.45 and sample sizes $n = 500$, 1000, 2000, 3000, 5000, and 10000, we carried out four hundred simulations. We summarize numerical results in Tables 1 and 2.

For Model 1, the one-dimensional ridge R of p_Y has a strong decay at the end points (see Fig. 1(a). We therefore only consider the part of R where the density is larger than $0.6 \max p_Y(\mathbf{y})$ to exclude a possible eigenvalue crossing. Figures 1(a) and 1(b) show the true and estimated densities for $d = 0.4$ and $n = 10000$. Simulated point clouds and the corresponding confidence regions are displayed in Figs. 2, 3, and 4 for $n = 1000$ and 10000 and $d = 0.35$, 0.4, and 0.45 respectively. We calculated the confidence regions using the equally spaced grid consisting of 250×250 points in $[-11, 11]^2$. We summarize numerical results in Table 1. For all

Table 1. Model 1 – proportions of simulated data where the true ridge of p_Y was inside the 95%-confidence region

	$n = 500$	$n = 1000$	$n = 2000$	$n = 3000$	$n = 5000$	$n = 10000$
$d = 0.35$	0.9775	0.9575	0.9750	0.9775	0.9725	0.9700
$d = 0.4$	0.9825	0.9900	0.9850	0.9600	0.9875	0.9675
$d = 0.45$	1	0.9975	0.9850	0.9800	0.9675	0.9500

Table 2. Model 2 – proportions of simulated data where the true ridge of p_Y was inside the 95%-confidence region

	$n = 500$	$n = 1000$	$n = 2000$	$n = 3000$	$n = 5000$	$n = 10000$
$d = 0.35$	0.9725	0.9800	0.9900	0.9750	0.9850	0.9675
$d = 0.4$	0.9725	0.9875	0.9950	0.9925	0.9950	0.9975
$d = 0.45$	0.9825	0.9675	0.9475	0.9325	0.9100	0.9200

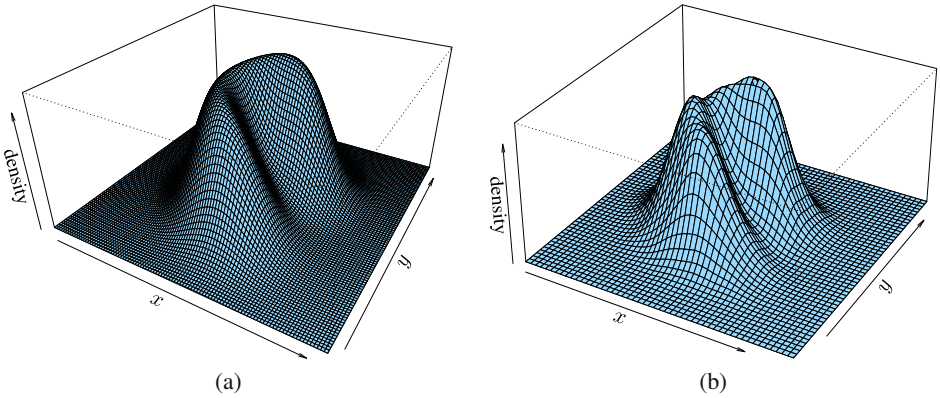

Figure 1. True and estimated density for Model 1 ($d = 0.4$ and $n = 10000$).

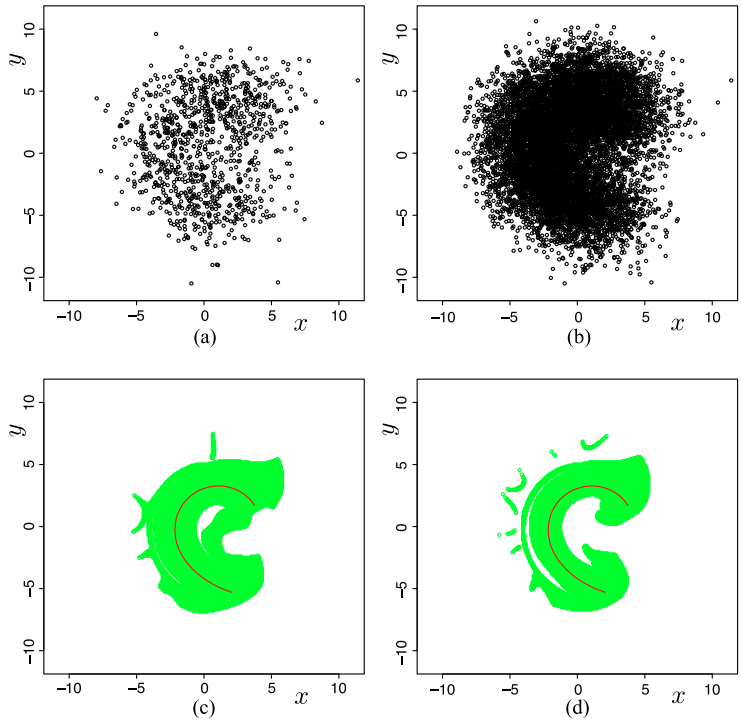

Figure 2. Points generated by Model 1 with $d = 0.35$ and (a) $n = 1000$ and (b) $n = 10000$. Corresponding 95%-confidence regions (green) are displayed in (c) and (d), respectively, together with the true ridge (red). (Online version in color.)

three values of d, the coverage probability turns out to be very close to the desired value of 0.95. On the other hand, since $n^{d-1/2}$ converges to zero rather slowly, the size of the confidence regions shrinks slowly as well.

For Model 2, the one-dimensional ridge consists of two concentric circles (see Fig. 5(a)). Figure 5(b) shows the estimated density for $d = 0.4$ and $n = 10000$. Table 2 summarizes the results for the 95%-confidence regions. Simulated point clouds and the corresponding confidence regions for $n = 1000$ and 10000 and $d = 0.35$, $d = 0.4$, and $d = 0.45$ are displayed in Figs. 6, 7, and 8. An equally spaced grid consisting of 250×250 points in $[-4, 4]^2$ was used. The same comments as for Model 1 apply.

An alternative method for constructing confidence sets for a ridge has been proposed, for instance, in [10]. Their approach is based on bootstrapped Hausdorff distances. In principle, similar ideas could be developed for the model considered in the present paper. However, some nontrivial adjustments should be made. The

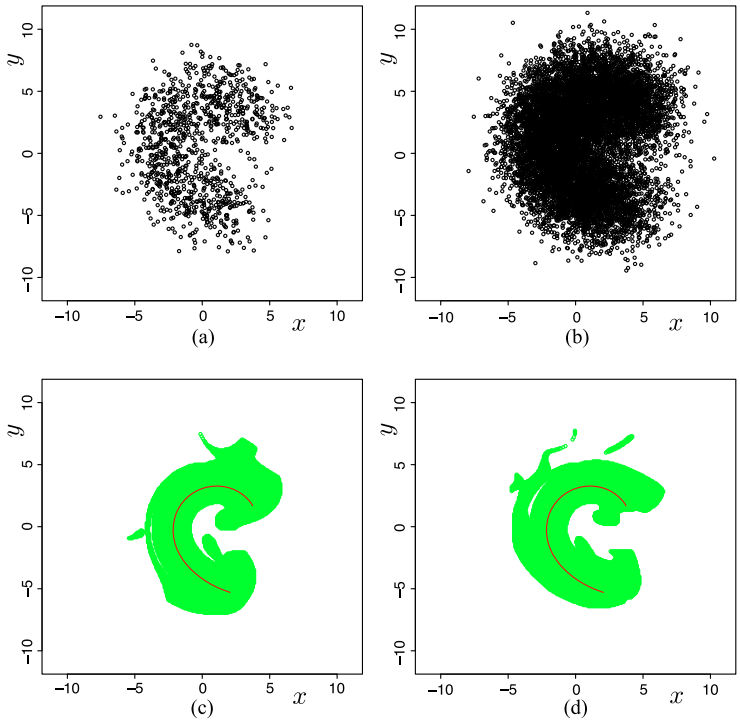

Figure 3. Points generated by Model 1 with $d = 0.4$ and (a) $n = 1000$ and (b) $n = 10000$. Corresponding 95%-confidence regions (green) are displayed in (c) and (d), respectively, together with the true ridge (red). (Online version in color.)

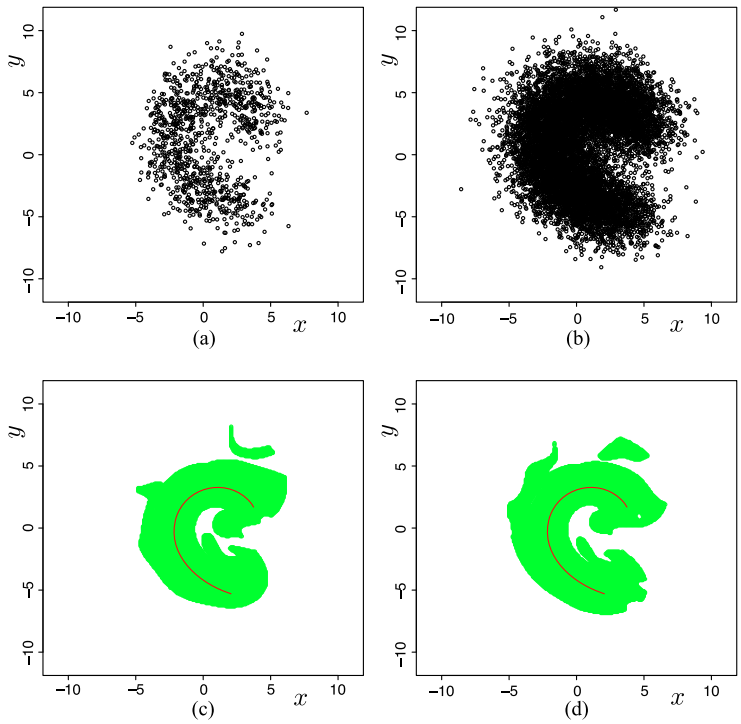

Figure 4. Points generated by Model 1 with $d = 0.45$ and (a) $n = 1000$; (b) $n = 10000$. Corresponding 95%-confidence regions (green) are displayed in (c) and (d), respectively, together with the true ridge (red). (Online version in color.)

main problem is that [10] use the simplest version of the bootstrap that destroys all temporal dependence. We may therefore expect that confidence sets based on their method would tend to have coverage probabilities far below the nominal one. To illustrate this, 95%-confidence regions based on [10] were computed for simulated series generated by Model 1 with $n = 500$ and $d = 0.35$, 0.4, and 0.45. As expected, the simulated coverage probabilities based on 100 simulations turned out to be very low, namely 0.81, 0.75, and 0.77, respectively. Moreover, the coverage probability appears to decrease with increasing sample size. For instance, for $d = 0.4$ and sample size $n = 2000$, the simulated coverage probability was 0.65. An interesting question for future research is defining a suitable modification of [10], which would be applicable to the case of long-range dependence. This could be done by designing suitable bootstrap algorithms. In simple situations such as location estimation, this has been done, for instance, in [29] and [23].

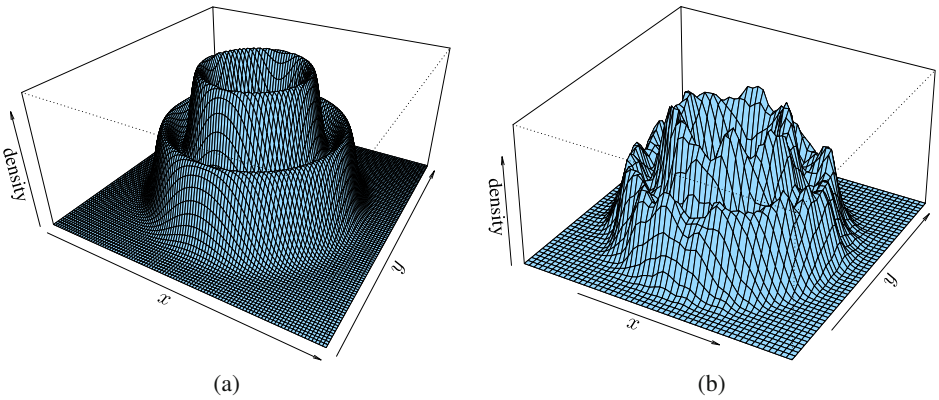

Figure 5. True and estimated density for Model 2 ($d = 0.4$ and $n = 10000$).

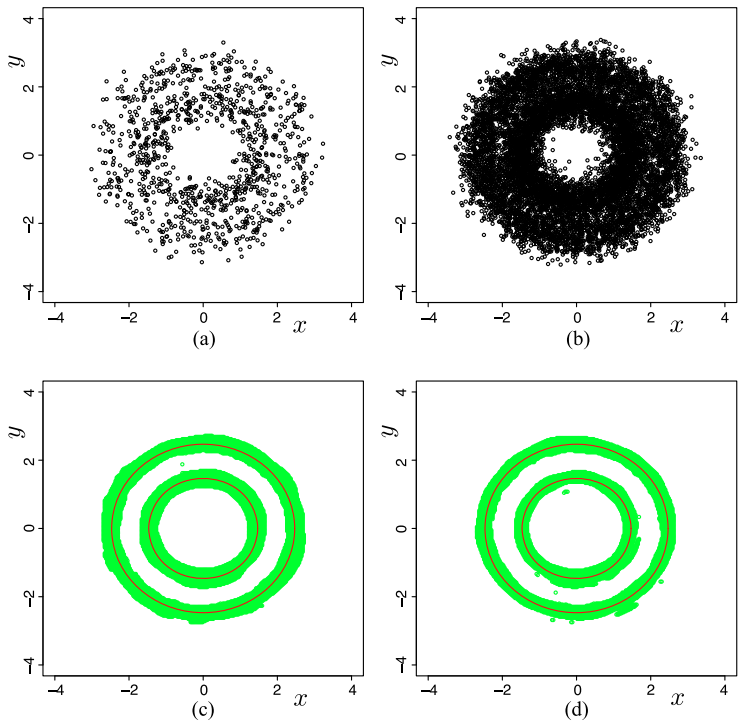

Figure 6. Points generated by Model 2 with $d = 0.35$ and (a) $n = 1000$; (b) $n = 10000$. Corresponding 95%-confidence regions (green) are displayed in (c) and (d) respectively, together with the true ridge (red). (Online version in color.)

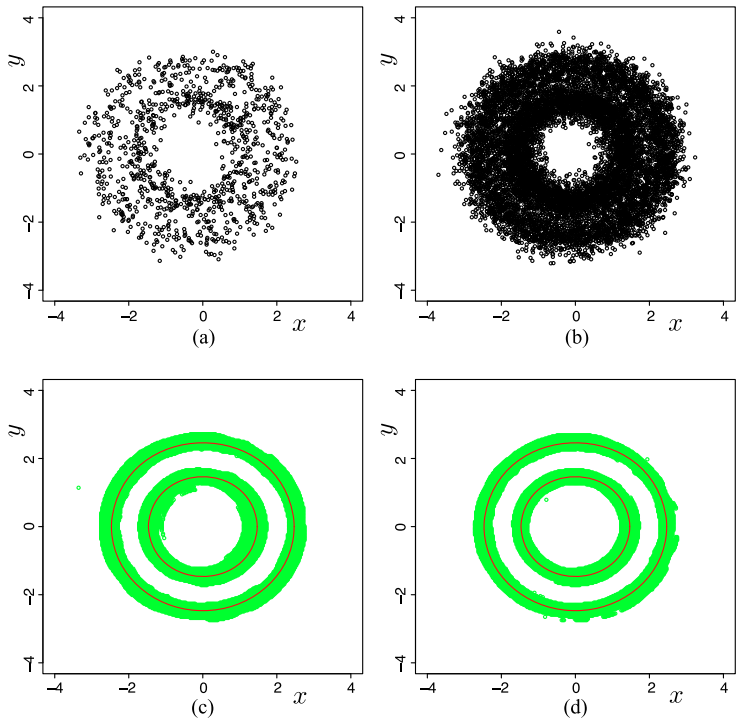

Figure 7. Points generated by Model 2 with $d = 0.4$ and (a) $n = 1000$ and (b) $n = 10000$. Corresponding 95%-confidence regions (green) are displayed in (c) and (d), respectively, together with the true ridge (red). (Online version in color.)

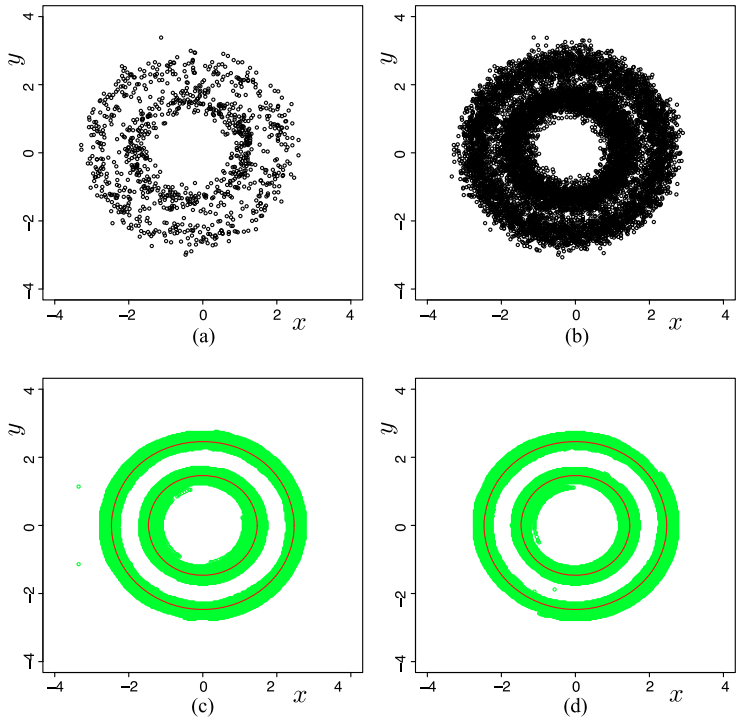

Figure 8. Points generated by Model 2 with $d = 0.45$ and (a) $n = 1000$ and (b) $n = 10000$. Corresponding 95%-confidence regions (green) are displayed in (c) and (d), respectively, together with the true ridge (red). (Online version in color.)

6 Final remarks

In this paper, we considered nonparametric inference for the ridge of a probability density function. This approach can be used in particular for manifold estimation and topological inference. For instance, suppose that i.i.d. observations on a compact differentiable k-manifold M embedded in $\mathbb{R}^m$ are perturbed by a linear process X_t with long memory. Then, under mild assumptions on the curvature of M and the variance of X_t, the k-dimensional ridge of the associated probability density function is homotopy equivalent to the manifold M (see, e.g., [9]).

The method presented here relies on using a kernel of sufficiently high order such that inequality (2.8) holds. Although (2.8) depends on the unknown parameter d, very precise knowledge of d is not required for choosing a sufficiently high order. For instance, given a confidence interval $[d_{\mathrm{low}}, d_{\mathrm{up}}]$ for d, we may evaluate (2.8) using the lower value d_{low} to identify a minimal required order ν. In the limiting case, that is, with a kernel of infinite order, there is no restriction on $d > 0$. Similarly, no precise knowledge of d is required for choosing a bandwidth for which the functional limit theorem (Theorem 2 with large bandwidths) holds.

The present paper is a first step toward statistical inference for ridge functions under long-memory errors. We expect that, given consistent estimators of the unknown quantities, asymptotic theory should not change. A more difficult question is how to design data driven algorithms to allow for optimal estimation of all parameters. In nonparametric regression with strongly dependent errors, [2,3] defined iterative algorithms that allow for simultaneous estimation of nuisance parameters and optimal trend estimation with a data-driven asymptotically optimal bandwidth. A similar approach may be adapted in the more complex context considered here. A detailed development of such data driven methods will have to be considered in future research.

Another open problem that will be worth pursuing in future research is ridge estimation at points with multiple eigenvalues. A possible approach may be adapting results in [36] and [41]

Acknowledgment. We would like to thank the referees for their insightful and constructive comments.

Appendix: Proofs

Proof of Theorem 1. Let

$$\mathcal{F}_i = \sigma(\boldsymbol{\varepsilon}_s,\, s \leqslant i)$$

denote the σ-algebra generated by $\boldsymbol{\varepsilon}_s$ $(s \leqslant i)$, whereas

$$\tilde{\mathcal{F}}_i = \sigma(\boldsymbol{\varepsilon}_s, \mu_s,\, s \leqslant i)$$

is generated by $\boldsymbol{\varepsilon}_s$ and μ_s $(s \leqslant i)$. We define the m-dimensional vectors

$$\mathbf{u}_j = (u_{j,1}, \dots, u_{j,m})^{\mathrm{T}} = \mathbf{y} - \sum_{s=j}^{\infty} A_s \boldsymbol{\varepsilon}_{t-s}$$

and the functions

$$F_j(\mathbf{u}_j) = \mathbf{P}\big(\mathbf{Y}_t \leqslant \mathbf{y} \,\big|\, \tilde{\mathcal{F}}_{t-j}\big) = \mathbf{P}(\mu_t + \mathbf{X}_t \leqslant \mathbf{y} \mid \tilde{\mathcal{F}}_{t-j}).$$

Note that the random variables μ_t are i.i.d. Hence, for $j \geqslant 1$, we have

$$\mathbf{P}(\mathbf{Y}_t \leqslant \mathbf{y} \mid \tilde{\mathcal{F}}_{t-j}) = \mathbf{P}(\mathbf{Y}_t \leqslant \mathbf{y} \mid \mathcal{F}_{t-j}),$$

whereas for $j = 0$,

$$F_0(\mathbf{u}_0) = \mathbf{P}(\mathbf{Y}_t \leqslant \mathbf{y} \mid \tilde{\mathcal{F}}_t) = \mathbf{1}\{\mathbf{Y}_t \leqslant \mathbf{y}\}.$$

Setting

$$\zeta_t(j) = F_j(\mathbf{u}_j) - F_{j+1}(\mathbf{u}_{j+1}),$$

we obtain the decomposition

$$\mathbf{1}\{\mathbf{Y}_t \leqslant \mathbf{y}\} - F_Y(\mathbf{y}) = \sum_{j=0}^{\infty} \zeta_t(j),$$

where equality is in the L^2-space of random variables. The rest of the proof follows by the same arguments as in the proof of Theorem 1 in [6]. □

Proof of Theorem 2. The proof follows from Theorem 1 by the same arguments as in the proof of Theorem 2 in [6]. □

Proof of Lemma 1.

$$\begin{aligned}
&\mathbf{E}\big[|H|^{-1/2}\big(H^{-1/2}\big)^{\otimes r} D^{\otimes r} K\big(H^{-1/2}(\mathbf{y} - \mathbf{Y}_t)\big)\big] \\
&\quad = \int_{\mathbb{R}^m} |H|^{-1/2}\big(H^{-1/2}\big)^{\otimes r} D^{\otimes r} K\big(H^{-1/2}(\mathbf{y} - \mathbf{u})\big) p_Y(\mathbf{u})\,\mathrm{d}\mathbf{u} \\
&\quad = \int_{\mathbb{R}^m} \big(H^{-1/2}\big)^{\otimes r} D^{\otimes r} K(\mathbf{w}) p_Y\big(\mathbf{y} - H^{1/2}\mathbf{w}\big)\,\mathrm{d}\mathbf{w},
\end{aligned}$$

where the last line is obtained by substituting $\mathbf{w} = H^{-1/2}(\mathbf{y} - \mathbf{u})$. Applying the divergence theorem leads to

$$\begin{aligned}
&\mathbf{E}\big[|H|^{-1/2}\big(H^{-1/2}\big)^{\otimes r} D^{\otimes r} K\big(H^{-1/2}(\mathbf{y} - \mathbf{Y}_t)\big)\big] \\
&\quad = \int_{\mathbb{R}^m} (-1)^r \big(-H^{1/2}\big)^{\otimes r}\big(H^{-1/2}\big)^{\otimes r} K(\mathbf{w}) p_Y^{(r)}\big(\mathbf{y} - H^{1/2}\mathbf{w}\big)\,\mathrm{d}\mathbf{w} \\
&\quad = \int_{\mathbb{R}^m} K(\mathbf{w}) p_Y^{(r)}\big(\mathbf{y} - H^{1/2}\mathbf{w}\big)\,\mathrm{d}\mathbf{w}.
\end{aligned}$$

Now consider the following Taylor expansion of the density derivative (see, e.g., [7])

$$p_Y^{(r)}\big(\mathbf{y} - H^{1/2}\mathbf{w}\big) = p_Y^{(r)}(\mathbf{y}) + \sum_{k=1}^{\nu} \frac{1}{k!}\big[I_{m^r} \otimes \big(-\mathbf{w}^{\mathrm{T}} H^{1/2}\big)^{\otimes k}\big] p_Y^{(r+k)}(\mathbf{y}) + r_n,$$

where

$$r_n = o\bigg(\frac{1}{\nu!}\big[I_{m^r} \otimes \big(-\mathbf{w}^{\mathrm{T}} H^{1/2}\big)^{\otimes \nu}\big] p_Y^{(r+\nu)}(\mathbf{y})\bigg).$$

Then, under (K2),

$$\begin{aligned}
&\int_{\mathbb{R}^m} K(\mathbf{w}) p_Y^{(r)}\big(\mathbf{y} - H^{1/2}\mathbf{w}\big)\,\mathrm{d}\mathbf{w} \\
&\quad = \int_{\mathbb{R}^m} K(\mathbf{w})\bigg[p_Y^{(r)}(\mathbf{y}) + \sum_{k=1}^{\nu} \frac{1}{k!}\big[I_{m^r} \otimes \big(-\mathbf{w}^{\mathrm{T}} H^{1/2}\big)^{\otimes k}\big] p_Y^{(r+k)}(\mathbf{y})\bigg]\,\mathrm{d}\mathbf{w} + r_n
\end{aligned}$$

$$= p_Y^{(r)}(\mathbf{y}) + \sum_{k=1}^{\nu} \frac{1}{k!} \left[I_{m^r} \otimes \left(\int\limits_{\mathbb{R}^m} K(\mathbf{w}) \mathbf{w}^{\otimes k} \, \mathrm{d}\mathbf{w} \right)^{\mathrm{T}} \right] \left[I_{m^r} \otimes \left(-H^{1/2} \right)^{\otimes k} \right] p_Y^{(r+k)}(\mathbf{y}) + r_n$$

$$= p_Y^{(r)}(\mathbf{y}) + \frac{1}{\nu!} \left[I_{m^r} \otimes \left(\int\limits_{\mathbb{R}^m} K(\mathbf{w}) \mathbf{w}^{\otimes \nu} \, \mathrm{d}\mathbf{w} \right)^{\mathrm{T}} \right] \left[I_{m^r} \otimes \left(-H^{1/2} \right)^{\otimes \nu} \right] p_Y^{(r+\nu)}(\mathbf{y}) + r_n. \qquad \square$$

Proof of Theorem 3. For the matrices $\nabla^2 \hat{p}_Y(\mathbf{y})$ and $\nabla^2 p_Y(\mathbf{y})$ with eigenvectors $\hat{v}_l(\mathbf{y})$ and $v_l(\mathbf{y})$ and corresponding eigenvalues $\hat{\lambda}_l(\mathbf{y})$ and $\lambda_l(\mathbf{y})$ for some fixed $l \in \{k+1, \dots, m\}$, we have

$$\nabla^2 \hat{p}_Y(\mathbf{y}) \hat{v}_l(\mathbf{y}) = \hat{\lambda}_l(\mathbf{y}) \hat{v}_l(\mathbf{y}), \tag{A.1}$$

$$\nabla^2 p_Y(\mathbf{y}) v_l(\mathbf{y}) = \lambda_l(\mathbf{y}) v_l(\mathbf{y}). \tag{A.2}$$

Then (A.1) can be written as

$$\begin{aligned} &\big\{ \nabla^2 p_Y(\mathbf{y}) + \big[\nabla^2 \hat{p}_Y(\mathbf{y}) - \nabla^2 p_Y(\mathbf{y}) \big] \big\} \big\{ v_l(\mathbf{y}) + \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big] \big\} \\ &\quad = \big\{ \lambda_l(\mathbf{y}) + \big[\hat{\lambda}_l(\mathbf{y}) - \lambda_l(\mathbf{y}) \big] \big\} \big\{ v_l(\mathbf{y}) + \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big] \big\}, \end{aligned}$$

which by (A.2) leads to

$$\begin{aligned} &\big[\nabla^2 \hat{p}_Y(\mathbf{y}) - \nabla^2 p_Y(\mathbf{y}) \big] v_l(\mathbf{y}) + \nabla^2 p_Y(\mathbf{y}) \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big] \\ &\qquad \times \big[\nabla^2 \hat{p}_Y(\mathbf{y}) - \nabla^2 p_Y(\mathbf{y}) \big] \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big] \\ &\quad = \big[\hat{\lambda}_l(\mathbf{y}) - \lambda_l(\mathbf{y}) \big] v_l(\mathbf{y}) + \lambda_l(\mathbf{y}) \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big] \\ &\qquad + \big[\hat{\lambda}_l(\mathbf{y}) - \lambda_l(\mathbf{y}) \big] \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big]. \end{aligned}$$

Neglecting higher-order terms, we obtain

$$\begin{aligned} &\big[\nabla^2 \hat{p}_Y(\mathbf{y}) - \nabla^2 p_Y(\mathbf{y}) \big] v_l(\mathbf{y}) + \nabla^2 p_Y(\mathbf{y}) \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big] \\ &\quad \approx \big[\hat{\lambda}_l(\mathbf{y}) - \lambda_l(\mathbf{y}) \big] v_l(\mathbf{y}) + \lambda_l(\mathbf{y}) \big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) \big]. \end{aligned}$$

By (V1) the eigenvectors $v_1(\mathbf{y}), \dots, v_m(\mathbf{y})$ form an orthonormal basis of $\mathbb{R}^m$. Therefore

$$\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y}) = \sum_{j=1}^{m} e_{lj} v_j(\mathbf{y}) \tag{A.3}$$

for some constants $e_{lj} \in \mathbb{R}$. Representation (A.3) leads to

$$\begin{aligned} &\big[\nabla^2 \hat{p}_Y(\mathbf{y}) - \nabla^2 p_Y(\mathbf{y}) \big] v_l(\mathbf{y}) + \nabla^2 p_Y(\mathbf{y}) \sum_{j=1}^{m} e_{lj} v_j(\mathbf{y}) \\ &\quad \approx \big[\hat{\lambda}_l(\mathbf{y}) - \lambda_l(\mathbf{y}) \big] v_l(\mathbf{y}) + \lambda_l(\mathbf{y}) \sum_{j=1}^{m} e_{lj} v_j(\mathbf{y}), \end{aligned}$$

which by (A.2) simplifies to

$$\begin{aligned}&\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]v_l(\mathbf{y})+\sum_{j=1}^{m}\lambda_j(\mathbf{y})e_{lj}v_j(\mathbf{y})\\&\quad\approx\big[\hat{\lambda}_l(\mathbf{y})-\lambda_l(\mathbf{y})\big]v_l(\mathbf{y})+\lambda_l(\mathbf{y})\sum_{j=1}^{m}e_{lj}v_j(\mathbf{y}).\end{aligned}$$

Hence, multiplying by $v_l^{\mathrm{T}}(\mathbf{y})$, we get

$$\begin{aligned}&v_l^{\mathrm{T}}(\mathbf{y})\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]v_l(\mathbf{y})+v_l^{\mathrm{T}}(\mathbf{y})e_{ll}\lambda_l(\mathbf{y})v_l(\mathbf{y})\\&\quad\approx v_l^{\mathrm{T}}(\mathbf{y})\big[\hat{\lambda}_l(\mathbf{y})-\lambda_l(\mathbf{y})\big]v_l(\mathbf{y})+v_l^{\mathrm{T}}(\mathbf{y})e_{ll}\lambda_l(\mathbf{y})v_l(\mathbf{y}),\end{aligned}$$

so that

$$\hat{\lambda}_l(\mathbf{y})-\lambda_l(\mathbf{y})\approx v_l^{\mathrm{T}}(\mathbf{y})\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]v_l(\mathbf{y}).$$

From Theorem 2 we have

$$n^{1/2-d}\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]\Rightarrow(-1)^{m+1}\sum_{j=1}^{m}\mathbf{B}_j(\mathbf{y})\xi_j. \tag{A.4}$$

Since the eigenvalues $\lambda_l(\mathbf{y})$ are simple for all $\mathbf{y}\in\Lambda_k^0$ and $l\in\{k+1,\dots,m\}$, the functions v_l are continuous functions of $\mathbf{y}$. The continuous mapping theorem then implies

$$n^{1/2-d}\big[\hat{\lambda}_l(\mathbf{y})-\lambda_l(\mathbf{y})\big]\Rightarrow(-1)^{m+1}\sum_{j=1}^{m}\big[v_l^{\mathrm{T}}(\mathbf{y})\mathbf{B}_j(\mathbf{y})v_l(\mathbf{y})\big]\xi_j,$$

completing the first part of the proof.

To derive the asymptotic distribution of $\hat{v}_l$, we need to determine the coefficients e_{lj} in the approximation

$$\begin{aligned}&\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]v_l(\mathbf{y})+\sum_{j=1}^{m}\lambda_j(\mathbf{y})e_{lj}v_j(\mathbf{y})\\&\quad\approx\big[\hat{\lambda}_l(\mathbf{y})-\lambda_l(\mathbf{y})\big]v_l(\mathbf{y})+\lambda_l(\mathbf{y})\sum_{j=1}^{m}e_{lj}v_j(\mathbf{y}).\end{aligned}$$

Multiplying by $v_i^{\mathrm{T}}(\mathbf{y})$ $(i\neq l)$ from the left leads to

$$v_i^{\mathrm{T}}(\mathbf{y})\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]v_l(\mathbf{y})+\lambda_i(\mathbf{y})e_{li}\approx\lambda_l(\mathbf{y})e_{li}.$$

Thus, for $i\neq l$ and $l\in\{k+1,\dots\}$,

$$e_{li}\approx\frac{1}{\lambda_l(\mathbf{y})-\lambda_i(\mathbf{y})}v_i^{\mathrm{T}}(\mathbf{y})\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]v_l(\mathbf{y}).$$

In particular,

$$n^{1/2-d}\big[\nabla^2\hat{p}_Y(\mathbf{y})-\nabla^2 p_Y(\mathbf{y})\big]\underset{C_b}{\Longrightarrow}\xi_{\nabla^2 p_Y}(\mathbf{y}),$$

together with $d < 1/2$, implies

$$n^{1/2-d} \sum_{j \neq l} e_{lj}^2 = n^{1/2-d} \sum_{j \neq l} e_{lj}^2(\mathbf{y}) \underset{C_b}{\Longrightarrow} 0. \tag{A.5}$$

For e_{ll}, note that

$$\begin{aligned}
1 &= \langle \hat{v}_l, \hat{v}_l \rangle = \big\langle v_l + (\hat{v}_l - v_l),\, v_l + (\hat{v}_l - v_l) \big\rangle \\
&= \left\langle v_l + \sum_{j=1}^{m} e_{lj} v_j,\; v_l + \sum_{j=1}^{m} e_{lj} v_j \right\rangle \\
&= \langle v_l, v_l \rangle + 2 \sum_{j=1}^{m} e_{lj} \langle v_l, v_j \rangle + \left\langle \sum_{j=1}^{m} e_{lj} v_j,\; \sum_{j=1}^{m} e_{lj} v_j \right\rangle \\
&= 1 + 2e_{ll} + {e_{ll}}^2 + \sum_{j \neq l} {e_{lj}}^2 = \big(1 + e_{ll}(\mathbf{y})\big)^2 + \sum_{j \neq l} {e_{lj}}^2,
\end{aligned}$$

so that

$$1 - \big(1 + e_{ll}(\mathbf{y})\big)^2 = \sum_{j \neq l} {e_{lj}}^2.$$

Due to (A.5), we have

$$-n^{1/2-d}\big[2e_{ll}(\mathbf{y}) + {e_{ll}}^2(\mathbf{y})\big] = n^{1/2-d}\big[1 - \big(1 + e_{ll}(\mathbf{y})\big)^2\big] \underset{C_b}{\Longrightarrow} 0.$$

Since, asymptotically, $\hat{v}_l$ is assumed to have the same orientation as v_l by (V2), this implies

$$n^{1/2-d} e_{ll}(\mathbf{y}) \underset{C_b}{\Longrightarrow} 0.$$

Thus, recalling the asymptotic distribution of the Hessian matrix in (A.4), we obtain

$$n^{1/2-d}\big[\hat{v}_l(\mathbf{y}) - v_l(\mathbf{y})\big] \underset{C_b(\Lambda_k^0, \mathbb{R}^m)}{\Longrightarrow} \xi_{v_l}(\mathbf{y}).$$

Proof of Theorem 4. For $\mathbf{y} \in \Lambda_k^0$ and $i = k+1, \ldots, m$, we have

$$\begin{aligned}
\hat{g}_i(\mathbf{y}) &= \big\langle \nabla \hat{p}_Y(\mathbf{y}),\, \hat{v}_i(\mathbf{y}) \big\rangle \\
&= \big\langle \nabla \hat{p}_Y(\mathbf{y}) - \nabla p_Y(\mathbf{y}) + \nabla p_Y(\mathbf{y}),\, \hat{v}_i(\mathbf{y}) - v_i(\mathbf{y}) + v_i(\mathbf{y}) \big\rangle \\
&= g_i(\mathbf{y}) + \big\langle \nabla p_Y(\mathbf{y}),\, \hat{v}_i(\mathbf{y}) - v_i(\mathbf{y}) \big\rangle + \big\langle \nabla \hat{p}_Y(\mathbf{y}) - \nabla p_Y(\mathbf{y}),\, v_i(\mathbf{y}) \big\rangle \\
&\quad + \big\langle \nabla \hat{p}_Y(\mathbf{y}) - \nabla p_Y(\mathbf{y}),\, \hat{v}_i(\mathbf{y}) - v_i(\mathbf{y}) \big\rangle.
\end{aligned}$$

Recall that

$$n^{1/2-d}\big[\nabla \hat{p}_Y(\mathbf{y}) - \nabla p_Y(\mathbf{y})\big] \underset{C_b(\mathbb{R}^m, \mathbb{R}^m)}{\Longrightarrow} \xi_{\nabla p_Y}(\mathbf{y})$$

and

$$n^{1/2-d}\big[\hat{v}_i(\mathbf{y}) - v_i(\mathbf{y})\big] \underset{C_b(\Lambda_k^0, \mathbb{R}^m)}{\Longrightarrow} \xi_{v_i}(\mathbf{y}).$$

Note also that both limit theorems, for $\nabla\hat{p}_Y$ and for $\hat{v}_i$, followed from the functional limit theorem for $\nabla^2\hat{p}_Y$. The proof of Theorem 3 can be extended along the same lines to obtain the joint weak convergence of $\psi_n(\mathbf{y}) = n^{1/2-d}[\nabla\hat{p}_Y(\mathbf{y}) - \nabla p_Y(\mathbf{y}), \hat{v}_{k+1}(\mathbf{y}) - v_{k+1}(\mathbf{y}), \ldots, \hat{v}_m(\mathbf{y}) - v_m(\mathbf{y})]$ to $\psi(\mathbf{y}) = [\xi_{\nabla p_Y}(\mathbf{y}), \xi_{v_{k+1}}(\mathbf{y}), \ldots, \xi_{v_m}(\mathbf{y})]$. To save space, we omit the details. Next, set $\zeta_i(\mathbf{y}) = \langle\nabla p_Y(\mathbf{y}), \xi_{v_i}(\mathbf{y})\rangle + \langle\xi_{\nabla p_Y}(\mathbf{y}), v_i(\mathbf{y})\rangle$. Since the scalar product is continuous, the continuous mapping theorem, together with weak convergence of $\psi_n(\mathbf{y})$ to $\psi(\mathbf{y})$, leads to

$$n^{1/2-d}\big[\hat{g}_{k+1}(\mathbf{y}) - g_{k+1}(\mathbf{y}), \ldots, \hat{g}_m(\mathbf{y}) - g_m(\mathbf{y})\big]^{\mathrm{T}} \underset{C_b(\Lambda_k^0,\mathbb{R})}{\Longrightarrow} \zeta(\mathbf{y}),$$

and hence

$$n^{1/2-d}\big[\hat{G}_k(\mathbf{y}) - G_k(\mathbf{y})\big] \underset{C_b(\Lambda_k^0,\mathbb{R}^{m-k})}{\Longrightarrow} \zeta(\mathbf{y}). \qquad \square$$

References

1. J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, CRC Press, New York, 1994.
2. J. Beran and Y. Feng, SEMIFAR models – A semiparametric framework for modelling trends, long-range dependence and nonstationarity, *Comput. Stat. Data Anal.*, **40**(2):393–419, 2002.
3. J. Beran and Y. Feng, Data driven bandwidth choice for SEMIFAR models, *J. Comput. Graph. Stat.*, **11**(2):690–713, 2002b.
4. J. Beran, Y. Feng, S. Ghosh, and R. Kulik, *Long-Memory Processes*, Springer, New York, 2013.
5. J. Beran and N. Schumm, On non parametric statistical inference for densities under long-range dependence, *Commun. Stat., Theory Methods*, **46**(22):11296–11314, 2017.
6. J. Beran and K. Telkmann, On nonparametric density estimation for multivariate linear long-memory processes., *Commun. Stat., Theory Methods*, **47**(22):5460–5473, 2018.
7. J.E. Chacón and T. Duong, *Multivariate Kernel Smoothing and Its Applications*, Chapman & Hall, CRC Press, New York, 2018.
8. J.E. Chacón, T. Duong, and M.P. Wand, Asymptotics for general multivariate kernel density derivative estimators, *Stat. Sin.*, **21**(2):807–840, 2011.
9. F. Chazal, D. Cohen-Steiner, and Q. Mérigot, Geometric inference for probability measures, *Found. Comput. Math.*, **11**(6):733–751, 2011.
10. Y.-C. Chen, C.R. Genovese, and L. Wasserman, Asymptotic theory for density ridges, *Ann. Stat.*, **43**(5):1896–1928, 10 2015.
11. Y.C. Chen, C.R. Genovese, S. Ho, and L. Wasserman, Optimal ridge detection using coverage risk, in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, Neural Information Processing Systems Foundation, 2015, pp. 316–324.
12. Y.C. Chen, C.R. Genovese, and L.A. Wasserman, Generalized mode and ridge estimation, 2014, arXiv:abs/1406.1803.
13. C.-F. Chung, Sample means, sample autocovariances, and linear regression of stationary multivariate long memory processes, *Econom. Theory*, **18**(1):51–78, 2002.
14. S. Csorgo and J. Mielniczuk, Density estimation under long-range dependence, *Ann. Stat.*, **23**(3):990–999, 1995.
15. P. Doukhan, G. Oppenheim, and M.S. Taqqu, *Ridges in Image and Data Analysis*, Springer, Dordrecht, 1996.

16. P. Doukhan, G. Oppenheim, and M.S. Taqqu (Eds.), *Theory and Application of Long-Range Dependence*, Birkhäuser, Basel, 2003.

17. D. Eberly, R. Gardner, B. Morse, S. Pizer, and C. Scharlach, Ridges for image analysis, *J. Math. Imaging Vis.*, **4**(4): 353–373, December 1994, ISSN 0924-9907.

18. C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman, Nonparametric ridge estimation, *Ann. Stat.*, **42**(4):1511–1545, 2014.

19. S. Ghosh, *Kernel Smoothing*, Wiley, New York, 2018.

20. L. Giraitis, H.L. Koul, and D. Surgailis, *Large Sample Inference for Long Memory Processes*, Imperial College Press, London, 2012.

21. A. Gramacki, *Nonparametric Kernel Density Estimation and Its Computational Aspects*, Springer, New York, 2018.

22. P. Hall and J.D. Hart, Convergence rates in density estimation for data from infinite-order moving average processes, *Probab. Theory Relat. Fields*, **87**(2):253–274, 1990.

23. P. Hall, B.-Y. Jing, and S.N. Lahiri, On the sampling window method for long-range dependent data, *Stat. Sin.*, **8**(4): 1189–1204, 1998.

24. P. Hall, W. Qian, and D.M. Titterington, Ridge finding from noisy data, *J. Comput. Graph. Stat.*, **1**(3):197–211, 1992.

25. R.M. Haralick, Ridges and valleys on digital images, *Comput. Vis. Graph. Image Process.*, **22**(1):28 – 38, 1983.

26. I. Horová, J. Kolácek, and J. Zelinka, *Kernel Smoothing In Matlab*, World Scientific, River Edge, NJ, 2012.

27. S. Kechagias and V. Pipiras, Definitions and representations of multivariate long-range dependent time series, *J. Time Ser. Anal.*, **36**(1):1–25, 2015.

28. T. Kollo and D. von Rosen, *Advanced Multivariate Statistics with Matrices*, Springer, Dordrecht, 2005.

29. S.N. Lahiri, On the moving block bootstrap under long range dependence, *Stat. Probab. Lett.*, **18**(5):405–413, 1993.

30. H. Liang and H. Wu, Parameter estimation for differential equation models using a framework of measurement error in regression models, *J. Am. Stat. Assoc.*, **103**(484):1570–1583, 2008.

31. T. Lindeberg, Edge detection and ridge detection with automatic scale selection, *Int. J. Comput. Vis.*, **30**(2):117–156, November 1998.

32. M. Lu, E. Pebesma, A. Sánchez, and J. Verbesselt, Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from MODIS time series, *ISPRS J. Photogramm. Remote Sens.*, **117**:227–236, 2016.

33. X. Magnus and H. Neudecker, *Matrix Differential Calculus*, Wiley, New Jersey, 1998.

34. D. Marinucci and P.M. Robinson, Weak convergence of multivariate fractional processes, *Stochastic Processes Appl.*, **86**(1):103 – 120, 2000.

35. A.F. Militino, M.D. Ugarte, and U. Pérez-Goya, An introduction to the spatio-temporal analysis of satellite remote sensing data for geostatisticians, in B.S. Daya Sagar, Q. Cheng, and F. Agterberg (Eds.), *Handbook of Mathematical Geosciences: Fifty Years of IAMG*, Springer, Cham, 2018, pp. 239–253.

36. Y. Nakatsukasa, Perturbation behavior of a multiple eigenvalue in generalized Hermitian eigenvalue problems, *BIT*, **50**(1):109–121, 2010.

37. G. Norgard and P.-T. Bremer, Ridge-valley graphs: Combinatorial ridge detection using Jacobi sets, *Comput. Aided Geom. Des.*, **30**:597–608, 2013.

38. W. Qiao and W. Polonik, Theoretical analysis of nonparametric filament estimation, *Ann. Stat.*, **44**:1269–1297, 2016.

39. D.W. Scott, *Multivariate Density Estimation*, Wiley, New Jersey, 2015.

40. G.W. Stewart and J.G. Sun, *Matrix Perturbation Theory*, Academic Press, Cambridge, MA, 1990.

41. J.-G. Sun, Multiple eigenvalue sensitivity analysis, *Linear Algebra Appl.*, **137–138**:183–211, 1990.

42. M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman & Hall, CRC Press, New York, 1995.

43. E.J. Wegman and Q. Luo, On methods of computer graphics for visualizing densities, *J. Comput. Graph. Stat.*, **11**(1): 137–162, 2002.

44. W.B. Wu and J. Mielniczuk, Kernel density estimation for linear processes, *Ann. Stat.*, **30**(5):1441–1459, 2002.