On Bandwidth Choice for Spatial Data Density Estimation

Zhenyu Jiang[a]*, Nengxiang Ling[b]†, Zudi Lu[a,c]‡, Dag Tjøstheim[d]§, Qiang Zhang[e]¶

[a]Statistical Sciences Research Institute, University of Southampton, SO17 1BJ, UK

[b]School of Mathematics, Hefei University of Technology, Hefei 230009, China

[c]School of Mathematical Sciences, University of Southampton, SO17 1BJ, UK

[d] Department of Mathematics, University of Bergen, Norway

[e] School of Management, Beijing University of Chemical Technology, China

Abstract: Bandwidth choice is crucial in spatial kernel estimation in exploring non-Gaussian complex spatial data. This paper investigates the choice of adaptive and non-adaptive bandwidths for density estimation given data on a spatial lattice. An adaptive bandwidth depends on local data and hence adaptively conforms with local features of the spatial data. We propose a spatial cross validation (SCV) choice of a global bandwidth. This is done first with a pilot density involved in the expression for the adaptive bandwidth. The optimality of the procedure is established, and it is shown that a non-adaptive bandwidth choice comes out as a special case. Although the CV idea has been popular for choosing a non-adaptive bandwidth in data-driven smoothing of independent and time series data, its theory and application have not been much investigated for spatial data. For the adaptive case, there is little theory even for independent data. Conditions that ensure asymptotic optimality of the SCV selected bandwidth are derived, actually, also extending time series and independent data optimality results. Further, for the adaptive bandwidth with an estimated pilot density, oracle properties of the resultant density estimator are obtained asymptotically as if the true pilot were known. Numerical simulations show that finite-sample performance of the SCV adaptive bandwidth choice works rather well. It outperforms the existing R-routines such as the 'rule of thumb' and the so-called 'second-generation' Sheather-Jones bandwidths for moderate and big data. An empirical application to a set of spatial soil data is further implemented with non-Gaussian features significantly identified.

Keywords: Optimal bandwidth; spatially adaptive bandwidth choice; kernel density estimation; cross-validation; spatial lattice data.

*E-mail address: zhenyujiang1@gmail.com (Z. Jiang)

†E-mail address: hfut.lnx@163.com (N. X. Ling)

‡Corresponding author. E-mail address: Z.Lu@soton.ac.uk (Z. Lu)

§E-mail address: Dag.Tjostheim@math.uib.no (D. Tjøstheim)

¶E-mail address: jqx_zhq@buaa.edu.cn (Q. Zhang)

1 Introduction

Nonparametric methods have become increasingly popular in exploring spatially dependent complex data. In particular, voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information (Guo and Mennis 2009). Therefore various data collected over the earth's surface arise in different disciplines, including environmental science, econometrics, epidemiology, image analysis, oceanography, to list a few. Important applications and developments in the general area of spatial statistics, under linear and/or Gaussian assumptions, can be found widely; see, e.g., Cressie (1993), Basawa (1996a; 1996b), Guyon (1995) and Gelfand *et al.* (2010) for comprehensive reviews. On the other hand, it has been recognised in the literature that linear and/or Gaussian structures may quite often be violated, or only serve as a crude approximation (c.f., Section 5.2). Nonparametric smoothing methods have provided a powerful methodology for circumventing these shortcomings and gaining insights into spatially dependent data; see, for example, Tran (1990), Hallin *et al.* (2001, 2004a, 2004b), Gao *et al.* (2006), Hallin *et al.* (2009), Robinson (2008, 2011) and Jenish (2012) and Lu and Tjøstheim (2014), among others.

Effective use of nonparametric smoothing methods requires choice of a smoothing parameter (bandwidth) (c.f., Jones et al. 1996). Arguably, it is the most important aspect of nonparametric density estimation. In this paper, our objective is to develop an adaptive (i.e., local data dependent) as well as a fixed non-adaptive bandwidth choice for spatial data density estimation. The advantage of an adaptive bandwidth is that it is attempting to enhance local, or observation-wise, smoothing, rather than a one-bandwidth-fits-all type of smoothing. Here our attention is on spatial data observed on a lattice. Many agricultural experimental data, such as the soil data to be analysed in Section 5.2 below, and in particular, with modern data acquisition techniques, remotely sensed data are on a regular grid (c.f., Zhu *et al.* 2010). For generality, let us consider the data to be the observations from $\{X_{\mathbf{i}} = (X_{\mathbf{i}}^{(1)}, X_{\mathbf{i}}^{(2)}, ..., X_{\mathbf{i}}^{(d)})\}$, a d-dimensional stationary random field with index $\mathbf{i} = (i_1, i_2, ..., i_N) \in \mathbb{Z}^N$ $(N \geq 1)$, which is defined on a common probability space $(\Omega, \mathscr{F}, \mathscr{P})$. We denote by f the common density function of $X_{\mathbf{i}}$ to be estimated, where a point $\mathbf{i} = (i_1, i_2, ..., i_N)$ will be referred to as a site in $\mathbb{Z}^N$ with $\mathbb{Z}$ the set of all integers. When $N = 1$, one may think of the time series data case or data on a line, including independent data, while $N = 2$ or 3 correspond to the data observed on a plane or three-dimensional space.

We will propose generalising the popular cross validation (CV) idea (c.f., Stone 1974) to the choice of adaptive bandwidth, which includes the fixed non-adaptive one as a special case, for spatial lattice data. There do exist bandwidth selection methods that are theoretically justified based on resampling methods for spatial data. Such methods have targeted spectral density or variance estimation (c.f., Nordman and Lahiri (2004)).

But these developments are concerned with second order moments or linear dependence properties, whereas we are seeking to give a theoretical justification with accompanying finite-sample numerical experiments for the adaptive and non-adaptive bandwidth choice in general density based nonparametric analysis of spatial data.

In the special case of density estimation for independent or time series data, bandwidth selection is well known to be a key question when applying any nonparametric smoother, and this question has been addressed in many papers. According to Jones et al. (1996), these methods can be categorised into two generations.

Cross-validation methods (c.f., Silverman (1986), Fan and Gijbels (1996), Fan and Yao (2003)) have been classified as 'first generation' methods that also include Silverman's (1986) rule of thumb bandwidth choice. The non-adaptive CV bandwidth selection has been extensively studied with independent and time series data; see, e.g., Hall (1983), Stone (1984), Marron and Härdle (1986), Marron (1987), Hart and Vieu (1990), Kim and Cox (1997), and the references therein. In the present paper, we propose extending this idea to adaptive density estimation for spatial lattice data.

From the perspective of optimal selection of a non-adaptive bandwidth, many so-called 'second generation' bandwidth selection methods (e.g., Sheather and Jones (1991) and Marron (1992)) have been proposed. These methods are basically plug-in or bootstrap based, which rely on selection of pilot bandwidths (often by rule of thumb) and have been shown to perform better than the 'first generation' methods for independent data; see Sheather and Jones (1991), Cao et al. (1994) and Jones et al. (1996) for excellent reviews. However, we have not seen any 'second generation' methods for adaptive density estimation. This may be due to the fact that the asymptotic mean integrated squared error (MISE) for the adaptive density estimator is more involved than that for the standard density estimator. It therefore may appear that cross-validation is more natural than the alternative second-generation procedures for choosing the adaptive bandwidth. In view of our use of a pilot density, the spatial CV based adaptive bandwidth choice in this paper may be seen as a 'second generation' method.

The main contributions of this paper are summarised: First, we will propose a spatial bandwidth choice by generalising the cross validation to the dependent lattice data density estimation. Our proposed method works for an adaptive bandwidth choice, but also functions for a non-adaptive bandwidth choice as a special case. Second, conditions that ensure asymptotic optimality of the spatial CV selected bandwidth in terms of various error measures are derived, actually, also extending time series optimality results. Further, taking a non-adaptive kernel density estimate as a pilot for an adaptive estimate of spatial data, oracle properties for the resultant density estimate are obtained as if the pilot density were known. Third, numerical simulations carried out will demonstrate that finite-sample performance of the proposed spatial adaptive CV bandwidth choice works rather well. It outperforms the existing non-adaptive R-routines such as the 'rule of thumb' and the so-called 'second-generation' Sheather-Jones (1991) bandwidths both for moderate sizes of spatial samples and in particular for big spatial data sets. Our empirical application

to a set of spatial soil data will further illustrate that non-Gaussian features of the data are more significantly identified by spatial adaptive density estimation.

These results require an essentially different theoretical approach with a modified, albeit quite mild, set of assumptions. Fundamentally, unlike time series data in which time is uni-directional from the past to the future, space is multi-directional. With spatial data, the fact that the spatial sites do not have a natural ordering makes the problem of CV bandwidth selection more challenging. A number of conditions are needed for an asymptotic theory to be established. These are described in detail in Appendix A in the Online Supplementary Material, but for the convenience of the reader we summarize them here. First of all a spatial mixing condition is required. Because of the multi-directionality the spatial mixing conditions are more elaborate than the time series strong mixing condition, but by now there are some rather standard spatial mixing conditions as stated in Appendix A. Further, there is a set of conditions on the spatial kernel function to be introduced in equation (2.1). This pertains to the symmetry, continuity, boundedness, differentiability and support (a compact support is assumed) of the kernel function, and an integrability condition on the convolution kernel. There is also a condition on the characteristic function of the kernel linking it to the the characteristic function of the d-dimensional standard normal. The order r of the kernel is related to the differentiability of the density function. The density function to be estimated is bounded, with a deviation from the density in the independent case, which is also bounded. Moreover, there is a boundedness condition on conditional densities. In addition there is a weight function to be introduced in equation (3.1), this function being bounded and integrable, and having a compact support with the density function being bounded away from zero on this support. Finally, the bandwidth is restricted to an interval $[a\tilde{\mathbf{n}}^{-\frac{1}{2r+d}}, b\tilde{\mathbf{n}}^{-\frac{1}{2r+d}}]$ for some constants a and b, $0 < a < b < \infty$, and with $\tilde{\mathbf{n}}$ being the total number of observations. All of these conditions are quite standard. More details and a discussion of the feasibility of the conditions are given in Appendix A.

The organization of the paper is as follows. The basic idea on spatial adaptive density estimation for lattice data, using an adaptive bandwidth that involves a global bandwidth h_0 and an estimated pilot density $\hat{f}_{\mathbf{n}}$, will be introduced in Section 2. In Section 3, a spatial CV (SCV) selection of the global bandwidth h_0 will be suggested, where it is noted that a CV selection of the non-adaptive bandwidth $h = h_0$ can be seen as a special case. The conditions to ensure optimality results of the bandwidth selected by the SCV method, in terms of the integrated squared error (ISE), mean integrated squared error (MISE) and average squared error (ASE), will be established in Subsection 3.1 with a known pilot f. The SCV selection for the global bandwidth h_0 with an estimated pilot $\hat{f}_{\mathbf{n}}$ and the oracle properties associated with spatial adaptive kernel density estimation will be presented in Section 3.2. In Section 4, some further discussions on data involving spatial trends and a potential extension to bandwidth choice for spatial regression will be provided. The numerical finite-sample performances of these estimates by using both simulated and real spatial data sets are examined in Section 5. The technical conditions on spatial mixing

and additional regularity assumptions for the optimality theorems will be collected in Appendix A. The technical proofs of the main results together with other supplementary materials are relegated to Appendix B. Here some useful lemmas are given in Appendix B1, where a moment inequality with lattice random fields established by extending that in Gao et al (2008) will play an important role in the technical proof of the main theorems in Appendix B2. Some supplementary figures and table to Subsection 5.1 are collected in Appendix B3. The code used in the paper is copied in Appendix C in a separate text form and available on a URL https://sites.google.com/site/zudiluwebsite/ while the used data set of soil250 is available from the R package geoR (c.f., Ribeiro Jr and Diggle, 2016).

2 Spatial adaptive density estimation: Basic principles

Throughout the paper, let the random field $\{X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N\}$ be observed over a rectangular region defined by $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, i_2, ..., i_N) \in \mathbb{Z}^N \mid 1 \leq i_k \leq n_k, k = 1, 2, ..., N\}$. Thus, the total sample size in $\mathcal{I}_{\mathbf{n}}$ is denoted as $\tilde{\mathbf{n}} = \prod_{k=1}^N n_k$ for $\mathbf{n} = (n_1, n_2, ..., n_N) \in \mathbb{Z}^N$ with strictly positive integer coordinates $n_1, n_2, ..., n_N$. As in Hallin *et al.* (2004), we write that $\mathbf{n} \to \infty$ if $\min_{1\leq k\leq N}\{n_k\} \to \infty$, without requiring $\max_{1\leq j,k\leq N}\{n_j/n_k\} \leq C$ for some $0 < C < \infty$ given in Tran (1990), allowing for multi-directional convergence in the sample size.

The idea of adaptive density estimation is popular in application (c.f., Davies and Hazelton (2010)), where the bandwidth (see Abramson (1982a,b) in the independent data case) is defined adaptively depending on the sample $X_{\mathbf{i}} = X_{(i_1,i_2,\cdots,i_N)}$ spatially in estimating $f(x)$:

$$\check{f}_{\mathbf{n}}(x) = \frac{1}{\tilde{\mathbf{n}}} \sum_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} \frac{1}{h^d_{(i_1,i_2,\cdots,i_N)}} K\left(\frac{x - X_{(i_1,i_2,\cdots,i_N)}}{h_{(i_1,i_2,\cdots,i_N)}}\right), x \in \mathbb{R}^d, \tag{2.1}$$

where $\sum_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k}$ stands for the N-fold summations $\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_N=1}^{n_N}$, $K : \mathbb{R}^d \to \mathbb{R}^+$ is a kernel function. Here the bandwidth, being location-dependent and inversely related to the population density $f(\cdot)$, is taken as (c.f., Abramson, 1982)

$$h_{\mathbf{i}} \equiv h_{(i_1,i_2,\cdots,i_N)} \equiv h_{\mathbf{i}}(h_0; f, \delta) = \frac{h_0}{f(X_{(i_1,i_2,\cdots,i_N)})^{\delta}\gamma_f}, \tag{2.2}$$

where $h_0 \equiv h_{0\mathbf{n}}$ is a smoothing multiplier referred to as the global bandwidth, and

$$\gamma_f = \left(\prod_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} f(X_{(i_1,i_2,\cdots,i_N)})^{-\delta}\right)^{1/\tilde{\mathbf{n}}}. \tag{2.3}$$

In adaptive estimation for independent data, it has been suggested by Abramson (1982) that $\delta = 1/2$ is optimal when f is second-order differentiable. For an introductory exposition on the adaptive methods, the reader is referred to, e.g., Silverman (1986), Bowman and Azzalini (1997), Pagan and Ullah (1999) and Davies and Hazelton (2010). Inclusion of the geometric mean term γ_f in (2.2), as noted in Silverman (1986), is to free the bandwidth factors from dependence on the scale of the data, allowing the global bandwidth h_0 to be considered on the same scale as in the corresponding fixed non-adaptive bandwidth estimate of f defined below.

In application of the adaptive density estimator (2.1), we need to have a pilot density estimator of f required in (2.2) and resulting in a kernel density estimator $\hat{f}_{\mathbf{n}}$ used in (3.13) below. This pilot estimator will be chosen as the traditional kernel density estimator $\hat{f}_{\mathbf{n}}(x)$ of f, with a fixed bandwidth, which can be seen as a special case of (2.1) with $\delta = 0$ in (2.2) and (2.3), defined (c.f., Tran, 1990) as follows:

$$\hat{f}_{\mathbf{n}}(x) = \frac{1}{\tilde{\mathbf{n}}h^d} \sum_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} K\left(\frac{x - X_{(i_1,i_2,\cdots,i_N)}}{h}\right), x \in \mathbb{R}^d, \tag{2.4}$$

where $h \equiv h_{\mathbf{n}}$ (depending on $\mathbf{n}$) is used to distinguish it from h_0 above, for a sequence of fixed bandwidths tending to zero as $\mathbf{n} \to \infty$. The study of the asymptotic properties for the spatial nonparametric kernel density estimator $\hat{f}_{\mathbf{n}}(x)$ is an interesting problem in statistical inference. With a given bandwidth series, Tran (1990) may be the first paper to establish the asymptotic normality of multivariate kernel density estimator and Carbon et al (1996) for the kernel-type estimator with the convergence in L_1, under stationary spatial random fields satisfying mixing conditions. See also the similar issues investigated by Hallin et al (2001, 2004a) under alternative conditions on the spatial processes, and Lu and Tjøstheim (2014) and Harel et al. (2016) for some more recent developments. In this paper, we will focus on the spatial mixing processes, which will be introduced with assumptions in Appendix A.

In order to make the spatial adaptive bandwidth $h_{\mathbf{i}}$ defined in (2.2) applicable for adaptive density estimation in (2.1), we need to select the global bandwidth h_0 and have $\hat{f}_{\mathbf{n}}$ in (2.4) with a selected bandwidth h as a pilot estimate for f in (2.2) and (2.3). Therefore the selection of h_0 as well as h is what we are concerned with in the next section, and selection of h in $\hat{f}_{\mathbf{n}}$ (c.f., (2.4)) can be seen as the selection of $h = h_0$ corresponding to $\delta = 0$ in (2.2). As commented in Section 1, cross-validation appears to be more natural than the plug-in procedures for selection of h_0, which will be developed in the next section. For the pilot bandwidth h, it seems that any 'second generation' method such as the Sheather-Jones (1991) can also be alternatively applied, which itself, however, need a pilot bandwidth again and will be examined in the numerical section (Section 5).

3 Spatial cross-validation bandwidth choice

In what follows, we are investigating selection of the bandwidths needed in the adaptive kernel density estimation $\check{f}_{\mathbf{n}}$ in (2.1) by proposing a spatial cross validation (SCV) method with choice of the adaptive bandwidths given in (2.2) and (2.3), for spatial lattice processes ($N > 1$). Although some CV methods for non-adaptive bandwidth that leave more observations out were suggested for the time series case of $N = 1$ in the literature (c.f., Hart and Vieu 1990), leave-one-out CV is often preferred for its simplicity in applications. Moreover, by our experience with the real spatial data numerical example in Lu *et al.*(2014) with the non-adaptive case, the performance of leave-five-out CV seems very similar to that of the leave-one-out CV. Further, the leave-one-out CV is more popular with spatial data (c.f., Le Rest *et al.* 2014), and we are mainly extending the leave-one-out CV for bandwidth selection below.

3.1 Cross-validation choice for adaptive bandwidth with a given pilot density

We first consider the choice of the adaptive bandwidth $h_{\mathbf{i}}$ by cross-validation with a given pilot density f in (2.2), where the choice reduces to the selection of the global bandwidth h_0. We can propose a CV bandwidth selection for the global bandwidth h_0 as defined in Section 2, by extending the leave-one-out CV criterion from time series to spatial lattice as follows:

$$SCV_\delta(h_0) \equiv CV_\delta(h_0) = \int_{\mathbb{R}^d} \check{f}^2_{\mathbf{n}}(x)w(x)dx - \frac{2}{\tilde{\mathbf{n}}} \sum_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} \check{f}^{(\mathbf{i})}_{\mathbf{n}}(X_{\mathbf{i}})w(X_{\mathbf{i}}), \tag{3.1}$$

where SCV_δ stands for spatial cross validation for the global bandwidth h_0 with a given δ in (2.2), and $\check{f}^{(\mathbf{i})}_{\mathbf{n}}(x)$ is the adaptive kernel estimator of f based on $X'_{\mathbf{j}}s, \mathbf{j} = (j_1, \cdots, j_N) \neq \mathbf{i} = (i_1, \cdots, i_N)$, i.e.,

$$\check{f}^{(\mathbf{i})}_{\mathbf{n}}(x) = \frac{1}{\tilde{\mathbf{n}} - 1} \sum_{\substack{j_k=1 \\ \forall k=1,2,\cdots,N \\ \exists k: j_k \neq i_k}}^{n_k} \frac{1}{h^d_{(j_1,j_2,\cdots,j_N)}} K\Big(\frac{x - X_{(j_1,j_2,\cdots,j_N)}}{h_{(j_1,j_2,\cdots,j_N)}}\Big),$$

with $h_{\mathbf{j}} = h_{(j_1,j_2,\cdots,j_N)} \equiv h_{\mathbf{j}}(h_0; f, \delta)$, as defined in (2.2) and (2.3) with $\mathbf{j}$ replacing $\mathbf{i}$, depending on h_0 (and where f is the assumed pilot density and δ is given, e.g., $\delta = 1/2$ as explained following (2.3) for an optimal adaptive bandwidth), and $w(.)$ is some nonnegative weight function. Then, the extended CV optimal smoothing parameter is defined by

$$\check{h}_0(f, \delta) \equiv \check{h}_0 = \arg \min_{h_0 \in \mathcal{H}_{\mathbf{n}}} SCV_\delta(h_0), \tag{3.2}$$

where $\mathcal{H}_{\mathbf{n}}$ is an interval defined in assumption (H) in Appendix A.

As noted in Section 2, the adaptive bandwidth defined above includes the fixed bandwidth as a special case. When $\delta = 0$, the adaptive KDE (2.1) reduces to the non-adaptive (fixed bandwidth) KDE (2.4) with $h = h_0$, which is independent of the pilot f. Therefore, the bandwidth h of $\hat{f}_{\mathbf{n}}(x)$ in (2.4) can be selected following from (3.1) and (3.2) with $\delta = 0$, i.e., to minimize an estimated Integrated Squared Error defined by

$$SCV_0(h) \equiv CV_0(h) = \int_{\mathbb{R}^d} \hat{f}^2_{\mathbf{n}}(x)w(x)dx - \frac{2}{\tilde{\mathbf{n}}} \sum_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} \hat{f}^{(i_1,\cdots,i_N)}_{\mathbf{n}}(X_{(i_1,\cdots,i_N)})w(X_{(i_1,\cdots,i_N)}), \tag{3.3}$$

where $\hat{f}^{(i_1,\cdots,i_N)}_{\mathbf{n}}(x)$ is the kernel estimator of f based on $X'_{\mathbf{j}}s, \mathbf{j} = (j_1, \cdots, j_N) \neq \mathbf{i} = (i_1, \cdots, i_N)$, i.e.,

$$\hat{f}^{(i_1,\cdots,i_N)}_{\mathbf{n}}(x) = \frac{1}{\tilde{\mathbf{n}} - 1} \sum_{\substack{j_k=1 \\ \forall k=1,2,\cdots,N \\ \exists k: j_k \neq i_k}}^{n_k} \frac{1}{h^d} K\Big(\frac{x - X_{(j_1,j_2,\cdots,j_N)}}{h}\Big).$$

Then, the SCV optimal smoothing parameter for (2.4) is defined by

$$\hat{h} = \arg\min_{h \in \mathcal{H}_{\mathbf{n}}} SCV_0(h), \tag{3.4}$$

where $\mathcal{H}_{\mathbf{n}}$ is as defined in assumption (H) in Appendix A. Note that $\hat{h} = \check{h}_0$ (c.f., (3.2)) with $\delta = 0$.

In the non-adaptive case, it has been argued (c.f., Xia and Li (2002)) that despite the fact that in the sense of mean integrated squared error (that is data-independent), the relative error of a CV-bandwidth may be higher than that of, for example, the plug-in selector and the method of Ruppert et al. (1995), there is a growing body of opinion of using performance criteria, which are not just mean-integrated squared error (c.f., Mammen 1990, Jones 1991, Härdle and Vieu 1992, and Loader 1999), targeting at estimating the unknown probability density function (in the context of this paper). From this point of view, the CV-bandwidth performs reasonably well (Hall and Johnstone 1992, page 479). In the time series context, the argument for why cross-validation is an appropriate bandwidth selection method can be found in Hart and Vieu (1990), Kim and Cox (1997), Quintela-del-Rio (1996) and Xia and Li (2002), among others. However, for adaptive kernel density estimation, there has been no plug-in method seen even for independent data in the literature. The CV-bandwidth looks a more natural and implementable option for the adaptive KDE. There has been little investigation into the issue even with non-adaptive KDE for spatial data, except for some spatial spectral density or variance estimation concerning second order moments properties (c.f., Nordman and Lahiri, 2004).

We shall concretely measure the optimality of the selected bandwidth by considering the widely used integrated squared error (ISE), mean integrated squared error (MISE) and average squared error (ASE), defined, respectively, by

$$d_I(\check{f}_{\mathbf{n}}, f)(h) = ISE(h) = \int_{\mathbb{R}^d} \big(\check{f}_{\mathbf{n}}(x) - f(x)\big)^2 w(x)dx, \tag{3.5}$$

$$d_M(\check{f}_{\mathbf{n}}, f)(h) = MISE(h) = E\int_{\mathbb{R}^d} (\check{f}_{\mathbf{n}}(x) - f(x))^2 w(x)dx \tag{3.6}$$

and

$$\begin{aligned} &d_A(\check{f}_{\mathbf{n}}, f)(h) = ASE(h) \\ &\quad = \frac{1}{\tilde{n}} \sum_{\substack{j_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} [\check{f}_{\mathbf{n}}(X_{(j_1,\cdots,j_N)}) - f(X_{(j_1,\cdots,j_N)})]^2 w(X_{(j_1,\cdots,j_N)}), \end{aligned} \tag{3.7}$$

where $\check{f}_{\mathbf{n}}(x)$ is the adaptive KDE defined in (2.1), with $h_0 = h$ for ease of notation below, under a given pilot density f and δ.

We will show that the SCV selected bandwidth $\check{h}_0 = \check{h}_0(f, \delta)$ enjoys optimality for the adaptive kernel density estimator (2.1). Before presenting the main results, we denote

$$R = \frac{1}{\tilde{\mathbf{n}}} \sum_{\substack{j_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} f(X_{(j_1,\cdots,j_N)})w(X_{(j_1,\cdots,j_N)}) - Ef(X_{(j_1,\cdots,j_N)})w(X_{(j_1,\cdots,j_N)})$$

and

$$T = -\int_{\mathbb{R}^d} (f(x))^2 w(x)dx - 2R.$$

All the technical assumptions are listed in Appendix A below.

First, Theorem 3.1 establishes the convergence rate of the so-called vertical error for the nonparametric density estimation of lattice data. It demonstrates that the proposed SCV criterion is asymptotically equivalent to the criterion of integrated squared error in selection of bandwidth with spatial kernel density estimation, when compared with the mean integrated squared error. Based on this result, we shall establish the appropriate conditions under which $\check{h}_0$ and $\hat{h}$ are asymptotically optimal not only in terms of the ISE as in Hart and Vieu (1990), but also in terms of MISE and ASE, respectively, defined in (3.5)–(3.7), thus also extending the time series results of these authors.

Theorem 3.1 (i) When $\delta = 0$, under assumptions (K1)-(K2), (D1)-(D2), (M), (H) and (W) listed in Appendix A, we have

$$\frac{|SCV_\delta(h) - ISE(h) - T|}{MISE(h)} = \mathcal{O}_P(\tilde{\mathbf{n}}^{-\frac{d}{2(2r+d)}}), \tag{3.8}$$

where r is the order of the continuous differentiation of the probability density function f, defined in assumption (D1) in Appendix A. Further, if assumption (K3) in Appendix A is satisfied, we have

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|SCV_{\delta}(h) - ISE(h) - T|}{MISE(h)} = o_P(1) \tag{3.9}$$

as $\mathbf{n} \to \infty$.

(ii) When $\delta > 0$, in addition to the conditions in (i), if assumptions (K4) and (D3) in Appendix A are satisfied, then the conclusions (3.8) and (3.9) hold true.

Remark: In this theorem, conditions have been derived with spatial lattice data for the vertical error convergence needed below. (i) Note that even in the special case for time series with $N = 1$ and $\delta = 0$, our derived convergence rate is much faster than that in the literature. For example, the convergence rate in Kim and Cox (1997, Theorem 1) corresponding to our (3.8) is $O_P(n^{d(-1/2+\tilde{r}/\mu)/(d+2r)})$ (in the notation of this paper), where $\tilde{r}$ is a positive integer $< \mu/2$ with μ given in the mixing assumption (M) in Appendix A. This rate of $O(n^{d(-1/2+\tilde{r}/\mu)/(d+2r)})$ is much slower than our $O_P(\tilde{\mathbf{n}}^{-\frac{d}{2(2r+d)}})$, the latter being the same rate as obtained by Marron (1987) for the i.i.d. data. (ii) Moreover, our derived conditions in the case of $\delta = 0$ are much weaker than those in the literature. For instance, Kim and Cox (1997, page 195) commented that one may note that the i.i.d. convergence rates found in Marron (1987) and Marron and Hardle (1987) correspond to the case of $\mu = \infty$ in their Theorem 1. Interestingly, Theorem 3.1 above only requires $\mu > 2Nr(2 - 2/q)/(1 - 2/q)$ with $q > 2$, roughly corresponding to $\mu > 8$ under $N = 1$, $r = 2$ and $q = \infty$.

Second, the following theorem establishes that both the integrated squared error and the averaged squared error are asymptotically equivalent to the criterion of mean integrated squared error in selection of bandwidth with spatial kernel density estimation.

Theorem 3.2 (i) When $\delta = 0$, under assumptions (K1)-(K3), (D1)-(D2), (M), (H) and (W) listed in Appendix A, we have

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}} \mid \frac{ISE(h) - MISE(h)}{MISE(h)} \mid = o_p(1) \tag{3.10}$$

and

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}} \mid \frac{ASE(h) - MISE(h)}{MISE(h)} \mid = o_p(1), \tag{3.11}$$

as $\mathbf{n} \to \infty$.

(ii) When $\delta > 0$, in addition to the conditions in (i), if assumptions (K4) and (D3) in Appendix A are satisfied, then the conclusions (3.10) and (3.11) hold true.

Finally, Theorem 3.3 establishes that the bandwidth $\hat{h}$ selected by the suggested CV (see (3.4)) is asymptotically optimal in terms of the criteria involving the integrated squared error and the averaged squared error as well as the mean integrated squared error for spatial kernel density estimation.

Theorem 3.3 (i) When $\delta = 0$, under the conditions for Part (i) of Theorem 3.2, we have

$$\frac{d(\check{f}_{\mathbf{n}}, f)(\check{h}_0)}{\inf_{h\in\mathcal{H}_{\mathbf{n}}} d(\check{f}_{\mathbf{n}}, f)(h)} \to 1, in \;\; probability \tag{3.12}$$

as $\mathbf{n} \to \infty$, where d is any of d_I, d_A and d_M, and $\check{f}_{\mathbf{n}}(x) = \hat{f}_{\mathbf{n}}(x)$ and $\check{h}_0 = \hat{h}$ as defined in (3.2) and (3.4), respectively.

(ii) When $\delta > 0$, in addition to the conditions in (i), if assumptions (K4) and (D3) in Appendix A are satisfied, then the conclusion (3.12) holds true.

Remark: In the case of $N = 1$ for time series data with $\delta = 0$, Hart and Vieu (1990) showed that the CV selected bandwidth is asymptotically optimal in term of the ISE. This theorem above thus extends the time series asymptotic optimality to the ASE and the MISE.

3.2 Cross-validation choice for adaptive bandwidth with an estimated pilot

In practice, for the adaptive KDE with $\delta > 0$, we cannot apply (2.1) directly because the true unknown density f is involved in the adaptive bandwidth via (2.2) in the above procedure. It needs to be replaced by a pilot estimator, say $\hat{f}_{\mathbf{n}}$ given in (2.4), where a pilot fixed-bandwidth can be selected by any reasonable method, e.g., by CV in (3.4) or others such as Sheather-Jones (1991) plug-in method (Sheather-Jones only considered the non-adaptive case). Thus the practical adaptive KDE can be defined as follows:

$$\check{\hat{f}}_{\mathbf{n}}(x) = \frac{1}{\tilde{\mathbf{n}}} \sum_{\substack{i_k=1 \\ \forall k=1,2,\cdots,N}}^{n_k} \frac{1}{\check{h}^d_{(i_1,i_2,\cdots,i_N)}} K\left(\frac{x - X_{(i_1,i_2,\cdots,i_N)}}{\check{h}_{(i_1,i_2,\cdots,i_N)}}\right), x \in \mathbb{R}^d, \tag{3.13}$$

where $\check{h}_{(i_1,i_2,\cdots,i_N)} = h_{(i_1,i_2,\cdots,i_N)}(h_0; \hat{f}_{\mathbf{n}}, \delta)$ with $\hat{f}_{\mathbf{n}}$ replacing f in (2.2). Then the SCV optimal smoothing parameter is defined by

$$\check{\hat{h}}_0(\hat{f}_{\mathbf{n}}, \delta) \equiv \check{\hat{h}}_0 = \arg\min_{h_0\in\mathcal{H}_{\mathbf{n}}} S\check{C}V_\delta(h_0), \tag{3.14}$$

where $S\check{C}V_\delta(h_0) \equiv C\check{V}_\delta(h_0)$ is as defined in (3.1) with $h_{(i_1,i_2,\cdots,i_N)}$ replaced by $\check{h}_{(i_1,i_2,\cdots,i_N)}$ in (2.2), and $\mathcal{H}_{\mathbf{n}}$ is the same interval as defined in assumption (H) in Appendix A. On the basis of Theorem 3.2, we will show that Theorem 3.4 below is true with $\hat{f}_{\mathbf{n}}$ replacing f in (2.2).

We now state the theorem on the selected $\check{\hat{h}}_0$ for h_0 in terms of ISE, ASE and MISE.

Theorem 3.4 Under the conditions for Part (ii) of Theorem 3.3 with $\sup_{x\in S_w} |\hat{f}_{\mathbf{n}}(x) - f(x)| \to 0$ in probability, we have

$$\frac{d(\check{\hat{f}}_{\mathbf{n}}, f)(\check{\hat{h}}_0)}{\inf_{h_0\in\mathcal{H}_{\mathbf{n}}} d(\check{\hat{f}}_{\mathbf{n}}, f)(h_0)} \to 1, in \;\; probability \tag{3.15}$$

as $\mathbf{n} \to \infty$, where d is any of d_I, d_A and d_M defined in (3.5)-(3.7).

This theorem confirms that the global bandwidth $\check{h}_0$ selected by the suggested CV (see (3.14)) is also asymptotically optimal in terms of the criteria involving the integrated squared error and the averaged squared error as well as the mean integrated squared error for the adaptive spatial kernel density estimation. Notice from Theorem 3.4 that the CV-bandwidth for the adaptive kernel density estimate by using $\hat{f}_{\mathbf{n}}$ instead of f in (2.2) is asymptotically optimal as that in Theorem 3.3. We call this property an oracle property in the sense that the asymptotic optimality is achieved as if f were known. In the simulation section (Section 5.1), we will call the CV selected $\check{h}_0$ in (3.2) with the true f used in (2.2) the oracle CV bandwidth, while the one with f replaced by $\hat{f}_{\mathbf{n}}$ the estimated adaptive CV bandwidth for h_0, and the corresponding spatial adaptive kernel density estimates (AKDEs) are called the oracle and the (estimated) AKDEs, respectively.

4 Some further discussion

We here provide some more discussions regarding the application of the proposed bandwidth selection.

(1) For real spatial data, there may be spatial trends, which make the data non-stationary over space (c.f., Harel et al. 2016). For instance, suppose on the plane of $N = 2$ that the observed $\tilde{X}_{\mathbf{i}}$ has a spatial trend, i.e., $\tilde{X}_{\mathbf{i}} = g(\mathbf{s}_{\mathbf{i}}) + X_{\mathbf{i}}$, where $\mathbf{i} = (i_1, i_2)$ with $1 \le i_1 \le n_1$ and $1 \le i_2 \le n_2$, $\mathbf{s}_{\mathbf{i}} = (i_1/n_1, i_2/n_2)$, $g(\cdot)$ is a smooth trend function and $X_{\mathbf{i}}$ is an unobserved stationary field. In this case, we need to estimate and remove the spatial trends before applying the methodology given in Sections 2–3. For example, as done in Hallin et al (2009) and Lu et al. (2014), the R package 'sm' can be used to remove the spatial trends, and it can be proved that the theoretical results for the (estimated) de-trended data, say $\hat{X}_{\mathbf{i}}$, replacing the unobservable stationary $X_{\mathbf{i}}$ in the above sections can still hold under some appropriate conditions (c.f., Section 3 of Hallin et al. (2009)). This is the situation for the real soil data set in Section 5.2 below.

(2) In this paper, we have developed the CV-based adaptive bandwidth selection for density estimation of lattice data. Similarly, with a more involved regression setting, the ideas and techniques developed from this paper will be useful in adaptive bandwidth selection for conditional regression of lattice data (c.f., Hallin et al. 2004). The CV-based adaptive bandwidth selection can be extended more naturally than other procedures for spatial regression, which is beyond the scope of this paper and will be left for future research.

5 Numerical finite-sample performances

In the previous sections, asymptotic optimality for the CV-bandwidth based kernel and CV-pilot adaptive kernel density estimates was presented for spatial lattice data. In this

section, we turn to the numerical finite-sample performances of these estimates by using both simulated and real data sets. To simplify our discussion, we are considering the lattice data on the plane with $N = 2$ and $d = 1$, and denote X_{i_1,i_2} by $X_{i,j}$ for notational simplicity in this section. We take $\delta = 0$ for the usual (non-adaptive) KDE and $\delta = 1/2$ for the adaptive KDE as in Abramson (1982) in the following numerical examples.

5.1 Monte Carlo simulation

To evaluate the performance of our CV bandwidth selection procedure with non-Gaussian spatial data, we need to use a spatial lattice process $\{X_{i,j}\}$ on the plane, where the theoretical probability density function of $X_{i,j}$ can be computed analytically. Therefore we are considering a special spatial lattice process, $X_{i,j}$, that is generated through a mixture of Gaussian spatial moving average processes, in a way similar to Step 2 of Section 8.1 of Lu and Tjøstheim (2014), as follows:

Step 1 Generate three intermediate processes $Y_{ij,1}$, $Y_{ij,2}$ and $Y_{ij,3}$ from three independent Gaussian spatial moving averages, with $1 \le i \le m_1,\ 1 \le j \le m_2,\ k = 1, 2, 3,$

$$Y_{ij,k} = \mu_k + a_{1,k}Z_{i-1,j;k} + a_{2,k}Z_{i,j-1;k} + a_{3,k}Z_{i,j;k} + a_{4,k}Z_{i+1,j;k} + a_{5,k}Z_{i,j+1;k}, \quad (5.1)$$

where $Z_{i,j;k}$'s are three independent $i.i.d.$ samples from $\mathcal{N}(0, \sigma^2_{Z,k})$ for $k = 1, 2, 3$, respectively. We take $\mu_1 = -1$, $\mu_2 = 0.4$ and $\mu_3 = 1.5$, let $a_{m,1}$, $a_{m,2}$ and $a_{m,3}$ be the mth elements of $a_1 = (1/5, 2/5, 3/5, 4/5, -4/5)$, $a_2 = (-1, -4/5 - 3/5, -2/5, -1/5)$ and $a_3 = (1/5, 2/5, 3/5, 4/5, 1)$, respectively, and $\sigma_{Z,1} = 0.3$, $\sigma_{Z,2} = 0.2$ and $\sigma_{Z,3} = 0.4$. Here the marginal distribution of $Y_{ij,1}$ is Gaussian $\mathcal{N}(\mu_1 = -1, \sigma_1^2 = 0.1656)$, $Y_{ij,2}$ is $\mathcal{N}(\mu_2 = 0.4, \sigma_2^2 = 0.088)$, and $Y_{ij,3}$ is $\mathcal{N}(\mu_3 = 1.5, \sigma_3^2 = 0.352)$, where $\mathcal{N}(\mu, \sigma^2)$ stands for the univariate Gaussian distribution of mean μ and variance σ^2.

Step 2 We then generate spatial process by first generating independent $R_{ij} = (R_{ij,1}, R_{ij,2}, R_{ij,3}) \sim Multinomial(1, (p_1, p_2, p_3) = (0.4, 0.3, 0.3))$, $1 \le i \le m_1,$, $1 \le j \le m_2$, and then defining

$$X_{ij} = Y_{ij,1} \times R_{ij,1} + Y_{ij,2} \times R_{ij,2} + Y_{ij,3} \times R_{ij,3},\ 1 \le i \le m_1,\ 1 \le j \le m_2, \quad (5.2)$$

where $R_{ij,k}$'s are independent of $Z_{i,j;k}$'s and hence of $Y_{ij,k}$'s, with $k = 1, 2, 3$.

Note that the distribution of X_{ij} is a mixture of normal distributions, in the form

$$f(x) = 0.4 \times \phi_{(\mu_1=-1,\sigma_1^2=0.1656)}(x) + 0.3 \times \phi_{(\mu_2=0.4,\sigma_2^2=0.088)}(x) + 0.3 \times \phi_{(\mu_3=1.5,\sigma_3^2=0.352)}(x), \quad (5.3)$$

where $\phi_{(\mu,\sigma^2)}(x)$ stands for the probability density function of normal distribution $\mathcal{N}(\mu, \sigma^2)$.

We generate the simulated spatial data by using the values of the parameters in the above models. Note, as commented by a referee, that the simulated spatial process is m-dependent and therefore satisfying the α-mixing assumption in Appendix A. This follows

Table 1: Summary of selected bandwidths for 100 simulations with $m_1 = 25, m_2 = 10$

Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
H.CV	0.0669	0.1471	0.1779	0.1732	0.1980	0.2680
H.SJ	0.1470	0.1824	0.1958	0.1989	0.2126	0.3008
H.R	0.3105	0.3302	0.3383	0.3381	0.3434	0.3747
H0.CV	0.0679	0.2104	0.2429	0.2338	0.2685	0.3242
H0.SJ	0.1055	0.2145	0.2438	0.2399	0.2689	0.3264
H0.orac	0.1024	0.2246	0.2704	0.2638	0.3052	0.4679

because the Y_{ijk}'s (and R_{ijk}'s too) are m-dependent within k and independent across k (with m appropriately defined), following from the simulating models (5.1) and (5.2) and using the fact that $Z_{i,j;1}$, $Z_{i,j;2}$, $Z_{i,j;3}$ and R_{ij} are independent *i.i.d.* processes. It is also easy to see similarly to Section 8.1 of Lu and Tjøstheim (2014) that the resultant marginal and the joint density functions are mixture of Gaussian distributions, which satisfy the assumptions in (D1), (D2) and (M) of Appendix A below. We repeat the simulation 100 times, with different sizes of samples of $(m_1, m_2) = (25, 10)$, $(m_1, m_2) = (20, 20)$, $(m_1, m_2) = (50, 50)$ and $(m_1, m_2) = (100, 100)$, respectively, from smaller to larger sample sizes.

We are examining the performance of the density estimation with the spatial CV selection for the global bandwidth h_0 in (3.2) with pilot densities of different bandwidths, including the bandwidth h given in (3.4). We are first considering the case of $(m_1, m_2) = (25, 10)$ with total sample size $n = m_1 m_2 = 250$. For the CV based bandwidth selection, after some experiments on the bandwidth interval so that the selected bandwidths are within the interval, we set the bandwidth interval $\mathcal{H}_{\mathbf{n}} = [0.02, 0.5]$ in the computations below, which roughly corresponds to taking $a = 0.06$, $b = 1.51$ with $r = 2$, $d = 1$ and $\tilde{\mathbf{n}} = n = 250$ in Assumption (H) for the case of $(m_1, m_2) = (25, 10)$. (We also tried $\mathcal{H} = [0.001, 1]$, having the same outcomes.) The spatial CV based global bandwidths for adaptive density estimates of $X_{i,j}$ with different pilot densities, as well as the selected bandwidths with non-adaptive KDE, for the kernel K taken as a standard normal density function, are summarised in Table 1 with the no-adaptive case in the upper and adaptive case in the lower part of the table. Note that the R function 'density' with the default (rule of thumb) and the Sheather-Jones (1991) bandwidths from Package stats (R Core Team 2015) are also used for comparison. It is interesting to notice from Table 1 that for the non-adaptive KDE method, the selected bandwidths by CV (H.CV) tend to be smaller than those by the Sheather-Jones (SJ) method (H.SJ) while the rule of thumb (rot) bandwidth (H.R) is much larger than both H.CV and H.SJ, which may explain that the rule of thumb (rot) bandwidths tend to lead to a largely biased KDE (as displayed in Figure 1 below).

For the adaptive KDEs, the spatial CV based global bandwidths for h_0 either with CV or SJ pilot KDEs (simply denoted by H0.CV and H0.SJ, which may be more clear by H0.CVCV and H0.CVSJ, respectively) look quite similar overall in Table 1, which

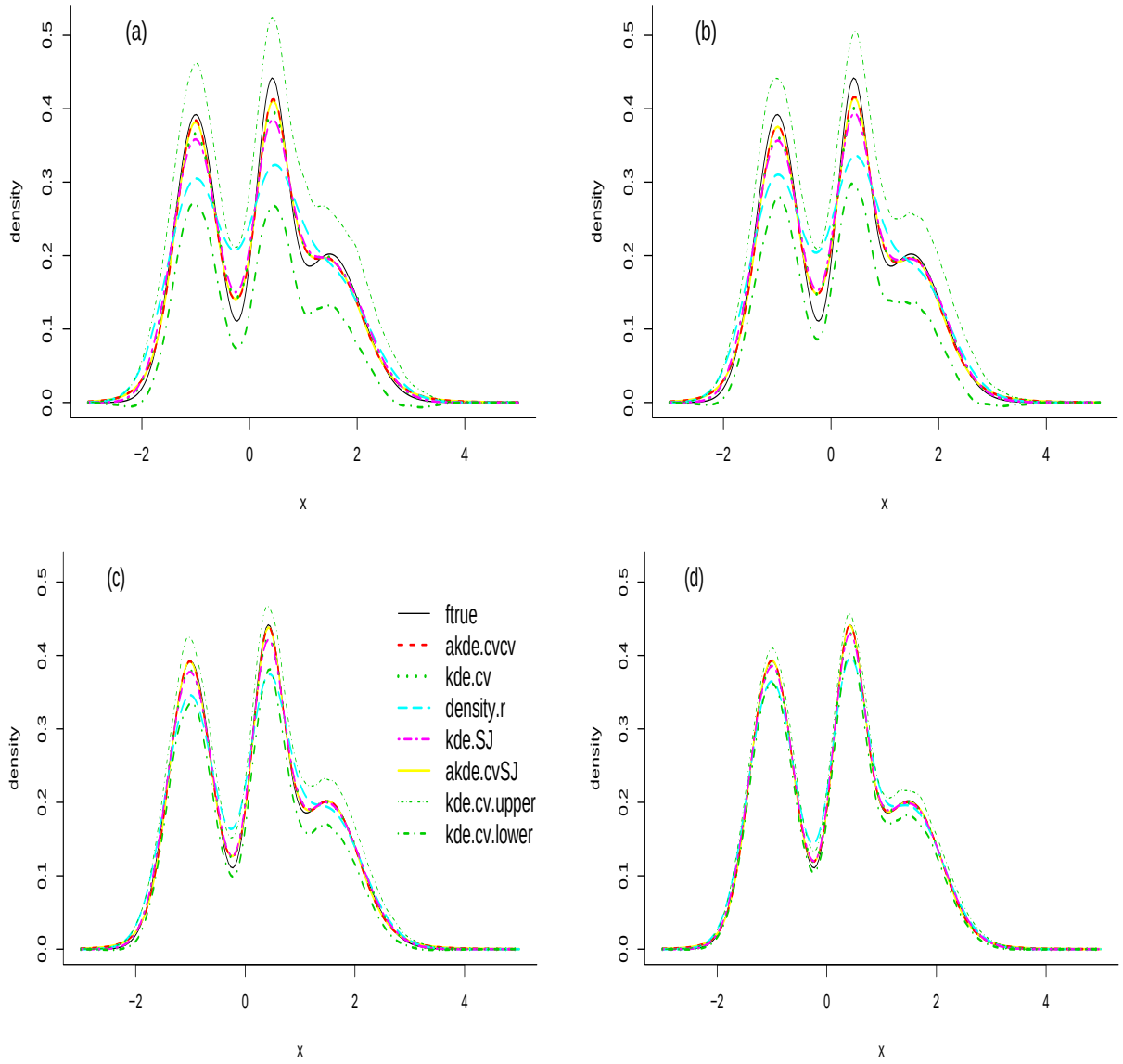


Figure 1: *Mean of 100 simulated estimates of density for different selected bandwidths of 100 simulations of sample sizes of* $n = m_1 * m_2$*: (a)* $(m_1, m_2) = (25, 10)$*, (b)* $(m_1, m_2) = (20, 20)$*, (c)* $(m_1, m_2) = (50, 50)$ *and (d)* $(m_1, m_2) = (100, 100)$.

shows that the adaptive KDEs corresponding to those bandwidths either with CV or SJ pilot KDEs perform similarly (as again indicated in Figure 1 below). In Table 1, we also provided the unpractical oracle bandwidths H0.orac (denoted by H0.or in Figure 2 below),

i.e., the spatial CV based global bandwidths with true pilot density, for the adaptive KDE. It follows from Table 1 that both H0.CV and H0.SJ are relatively smaller than the H0.orac.

As a benchmark comparison, we can also calculate the optimal bandwidths that minimise the MISEs for the non-adaptive and the adaptive KDEs, respectively, where the optimality is in terms of the MISE as defined in (3.6). For example, given the sample size of $(m_1, m_2) = (25, 10)$, the obtained optimal bandwidth, H.op=0.1788, is the one that minimises the MISE of the non-adaptive KDE (2.4) w.r.t. h. Similarly, the optimal bandwidth, H0.op=0.2659, is the one that minimises the MISE of the adaptive KDE (2.1) with oracle (i.e. true) pilot density f in (2.2) and (2.3), w.r.t. the global bandwidth h_0. This leads to the smallest MISE value among all the adaptive KDEs (including those with other pilot densities, say, the CV-based or SJ-based KDE as pilot; c.f., Figure B.1). Clearly the MISE-optimal bandwidths are theoretical quantities independent of data (given the sample size of (m_1, m_2)). These theoretically MISE-optimal bandwidths, H.op (for the non-adaptive KDE) and H0.op (for the adaptive KDE), used for benchmark only, together with boxplots of other bandwidths summarised in Table 1, can be found in Figure 2 with dotted horizontal lines indicated for ease of comparison of H.op and H0.op with others.

The means of the adaptive KDEs and non-adaptive KDEs corresponding to those bandwidths in Table 1 for 100 simulations are depicted in Figure 1; see also boxplots of those bandwidths in Figure 2. Here the true density function (5.3) together with the upper and lower limits, say, for the mean of the CV based density estimates as an example (by adding and subtracting respectively the double standard deviation of the CV based density estimates of 100 simulations) is also provided for comparison. To save space and make the figure legible, the simulated point-wise confidence interval for the mean density estimate over 100 simulations is provided for one KDE only. The sampling variability and simulation error are basically similar for other estimates. Further it can be observed from Figure 1(a) that the adaptive KDEs using spatial CV based global bandwidths either with the CV or the SJ pilot KDE (denoted by akde.cvcv and akde.cvSJ) are very similar in the means of the 100 simulations, which are much closer to the true density (f-true) than other (non-adaptive) KDEs such as the KDEs with the CV, the SJ and the rule of thumb bandwidths (denoted by kde.cv, kde.SJ and density.r, respectively). Also, we can clearly see from Panels (a)-(d) of Figure 1 that these facts remain true for the sample sizes sufficiently large, with the adaptive KDEs outperforming the non-adaptive KDEs. In particular, for the case of $(m_1, m_2) = (100, 100)$, we cannot distinguish the means of the akde.cvcv and the akde.cvSJ from the f-true in Panel (d). Note in Figure 2 that boxplots of different bandwidths are provided, where the dotted (red) horizontal lines, standing for the corresponding values of the benchmark MISE-optimal bandwidths, are plotted for ease of comparison with other bandwidths. Interestingly, the mean and median of the CV and the SJ selected bandwidths given in Table 1 and Figure 2 are much closer to the optimal bandwidth (say H.op=0.1788 in Panel (a) of Figure 2) than the rule of thumb bandwidths (the CV ones are a bit closer than the SJ's) for the non-adaptive KDE. As expected, the optimal bandwidth (say H0.op=0.2659 in Panel (a) of Figure 2) for the adaptive KDE

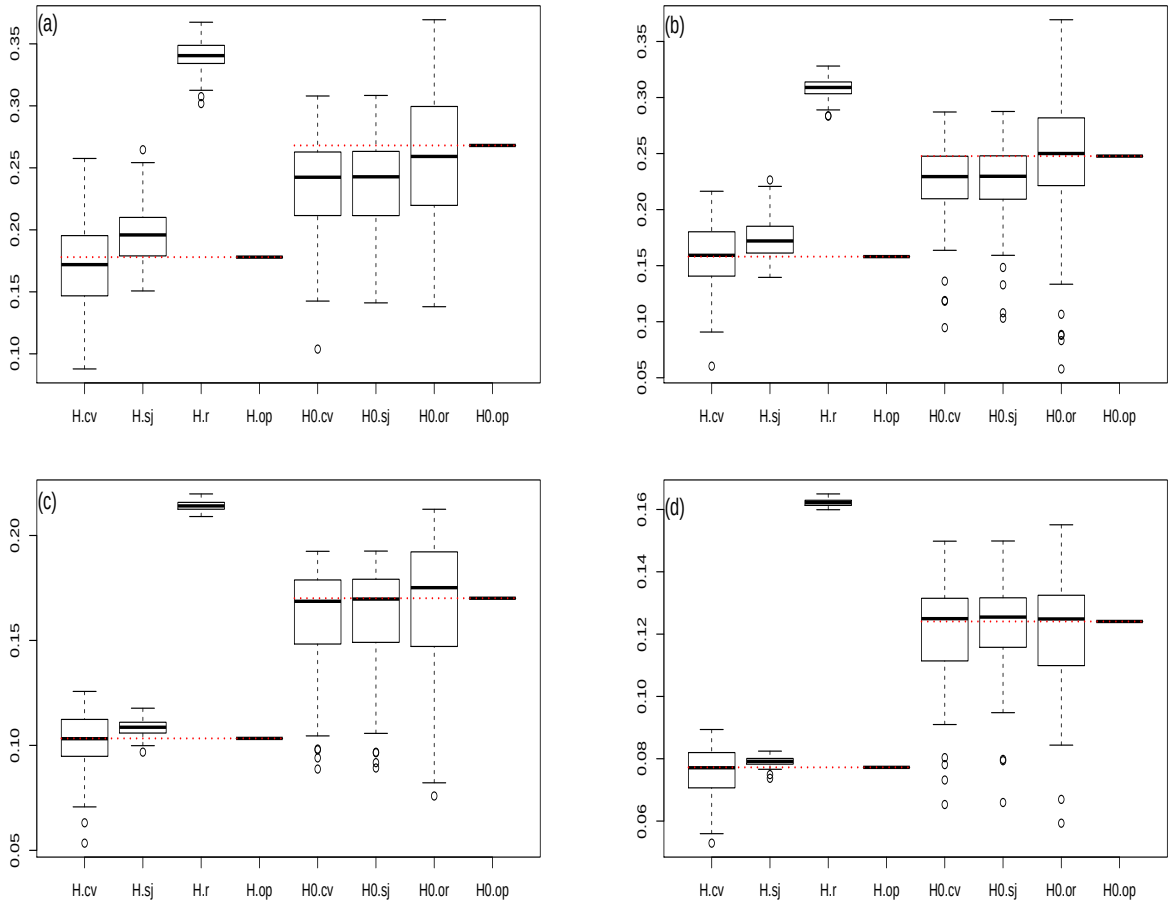


Figure 2: *Boxplot of different selected bandwidths of 100 simulations of sample sizes of* $n = m_1 * m_2$*: (a)* $(m_1, m_2) = (25, 10)$*, (b)* $(m_1, m_2) = (20, 20)$*, (c)* $(m_1, m_2) = (50, 50)$ *and (d)* $(m_1, m_2) = (100, 100)$*. Here H.CV, H.SJ and H.R are the CV, SJ and R-default (rule of thumb) bandwidths for* h *respectively, and H.op is the optimal bandwidth of* h *minimising MISE, in KDE; H0.CV, H0.SJ and H0.or are the CV bandwidths for* h_0 *in AKDE with pilot estimates by the CV, the SJ and the oracle true density, respectively. The (red) dotted horizontal lines are for the MISE-optimal H.op and H0.op of the KDE and the AKDE with the pilot of oracle (true) density, respectively, provided for ease of comparison with others.*

appears to be quite close to the mean and median of the oracle bandwidths (H0.or), which are similar to (only slightly larger than) those for H0.CV and H0.SJ given in Table 1. We can also see from Figure 2 that with the sample sizes increasing, the medians for H0.CV

and H0.SJ as well as the oracle bandwidths (H0.or) are becoming similar and close to H0.op for the adaptive KDEs (c.f., Panel (d)).

We are now further examining the finite-sample optimality of the spatial CV based adaptive bandwidths compared with other bandwidths in estimation of density in terms of ISE, ASE and MISE, which are also examined in Theorem 3.3. Here, for a given density estimator $\hat{f}_{\mathbf{n}}$ with the true density function f specified in (5.3), $d_I(\hat{f}_{\mathbf{n}}, f)(h) = ISE(h)$ and $d_A(\hat{f}_{\mathbf{n}}, f)(h) = ASE(h)$ are easily calculated as defined in (3.5) and (3.7) with $d = 1$ and $w(x) \equiv 1$ taken in this section. However, for $d_M(\hat{f}_{\mathbf{n}}, f)(h) = MISE(h)$ defined in (3.6), because of spatial dependence with the simulated random field (5.2), it becomes very complex and cannot be calculated as simply as with independent or time series observations. But fortunately, as a referee suggested, if ISE, as a function of h, is easily found, then MISE can be approximated through simulation. We do this by approximation through the average of the resultant ISE values based on 1000 simulations and we can therefore deal with MISE for the KDE and the AKDE defined in (2.4) and (2.1), respectively. In order to compare the performance of the estimates, AKDE (2.1) and KDE (2.4) with different bandwidths, defined in Sections 2–3, we will consider the AKDEs defined in (2.1) with the adaptive bandwidths given in (3.2) by using different pilot estimates of f, including the KDE (2.4) with the CV (see (3.4)) and the 'second generation' Sheather-Jones (1991, SJ hereafter) bandwidth and the oracle (true f) taken as a benchmark, respectively. As a comparison, we have also looked at the KDE in (2.4) with bandwidths of the CV (see (3.4)) and the Sheather-Jones (1991), respectively (those bandwidths outperforming the R default (rule of thumb) bandwidth as shown in Figure 1). We did not consider the 'second generation' bandwidth by bootstrap as bootstrapping the dependent non-Gaussian spatial data is still quite problematic, not as simple as for independent data (see, e.g., Faraway and Jhun (1990)).

Boxplots of the ISE's, ASE's and MISE's (scaled up by 100, 300 and 150 respectively for easy presentation) with different bandwidth based density estimates are displayed in Figure 3 for the 100 simulations of different sample sizes of (m_1, m_2) as specified above (to save space, the rule of thumb is not reported here, which performs worse than those reported in this figure, as indicated in Figure 1 above). We have also provided boxplots for the minimal values of the ISE's, ASE's and MISE's in Figure B.1 and their corresponding optimal bandwidths (i.e., the bandwidths minimising the ISE, ASE and MISE, respectively) in Figure B.2 for the AKDEs and the KDEs of the simulated data. Here in Figures 3, k.cv (k.SJ) is for the KDE with the CV-based (SJ-based) bandwidth, while a.cv (a.SJ or a.o) is a simple notation for the AKDE using the spatial CV global bandwidth with the pilot density being the CV-based KDE (the SJ-based KDE or the oracle true density) in Section 3. The conclusion of all of these experiments is that in the KDE case, CV and SJ perform similarly (with the SJ a bit better as is known). This similarity holds true in the AKDE case as well, when CV and SJ are used in the pilot estimation stage with the global bandwidth by the spatial CV, but the AKDE pair does significantly better than the KDE pair.

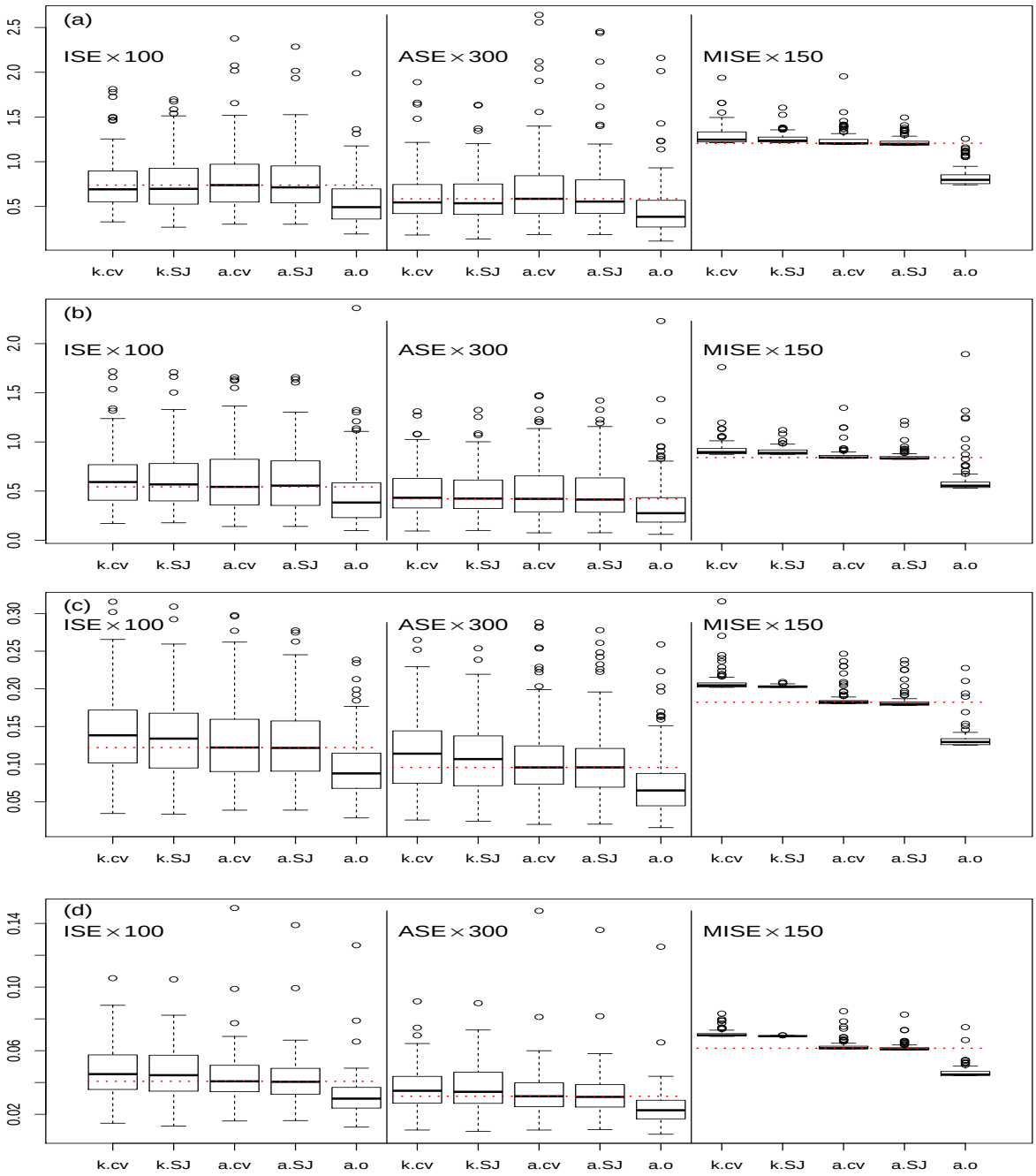


Figure 3: *Boxplots of ISE(×100), ASE(×300) and MISE(×150) for density estimates with 100 simulations of sample sizes of* (m_1, m_2)*: (a)* $(m_1, m_2) = (25, 10)$*, (b)* $(m_1, m_2) = (20, 20)$*, (c)* $(m_1, m_2) = (50, 50)$ *and (d)* $(m_1, m_2) = (100, 100)$*. Here k.cv and k.SJ are for the KDE with CV and SJ-based bandwidths, and a.cv, a.SJ and a.o are for the AKDE using the SCV global bandwidths with the CV, the SJ based KDE, and the oracle true, pilot densities. The (red) dotted horizontal lines are for the medians of the 100 simulated ISEs, ASEs and MISEs of the AKDE with a CV-KDE pilot respectively for ease of comparison, and the two vertical lines for separation of ISE, ASE and MISE.*

Also, as a referee suggested, we have examined the minimal ISE (ASE or MISE) values of the KDE and the AKDE together with their corresponding optimal bandwidths, which are provided, to save space, in Figures B.1 and B.2 in Section B3 of Appendix B. Note that in Figure B.1, v.k stands for the minimal ISE (ASE or MISE) value of the KDE, while the corresponding optimal bandwidth, denoted by hik (hak or hmk), minimises with respect to h the ISE (ASE or MISE) of the KDE (in acronyms), and v.acv for the minimal ISE (ASE or MISE) value of the AKDE using the corresponding optimal global bandwidth, denoted by h0icv (h0acv or h0mcv), minimising with respect to h_0 the ISE (ASE or MISE) of the AKDE with the CV-based KDE as the pilot density (in acronyms) in Figure B.2. Here v.asj (v.a0) can be defined similarly to v.acv, and h0isj, h0asj or h0msj (h0io, h0ao or h0mo) to h0icv, h0acv or h0mcv, with the SJ-based KDE (the oracle true density) replacing the CV-based KDE as the pilot.

It is again observed from Figure 3 and Figure B.1 that: (i) Overall the AKDEs by using the proposed spatial CV adaptive bandwidths either with the CV or the SJ piloted KDE outperform the KDEs with the CV-based and the SJ-based bandwidths. This phenomenon is particularly observed in terms of MISE even with the sample sizes as small as $(m_1, m_2) = (25, 10)$, which is the sample size in the real data below, in Panel (a) of these figures. (ii) With the sample sizes increasing, the AKDEs are clearly preferred to the KDEs, even in terms of ISE and ASE, as obviously seen in Panel (d) of these figures. (iii) For the AKDEs in Panels (a)–(d) of these figures, their performances by using the proposed spatial CV adaptive bandwidths either with the CV or the SJ KDE for pilot appear quite similar although using the SJ-KDE for pilot is sometimes slightly preferred. But for the non-adaptive KDEs, as indicated in the literature mentioned above, the SJ-based bandwidth is usually preferred. (iv) Note that, as expected, the impractical oracle AKDE performs best, but with the sample size becoming large as in Panel (d) of these figures, all the estimates have very small ISE, ASE and MISE (much smaller than those in Panels (a)–(c)), so the performance of all the practical estimates including the AKDE and the KDE approaches in these figures looks acceptable for the large sample sizes. Also, it is interesting to observe from Figure B.2 that as the sample sizes increase, the optimal global bandwidths for AKDE either with the CV-KDE or the SJ-KDE for a pilot are approaching to that of the oracle AKDE with true density as a pilot, all of which are larger than the optimal KDE bandwidths. This phenomenon is seen particularly clearly under MISE in Figure B.2 (note that the asymptotic MISE is the base of the SJ bandwidth method, c.f., Jones et al. (1996), in the non-adaptive case).

Now we can define the ratios of the left-hand side of (3.12) for d_I, d_A and d_M, respectively, by

$$R_{dI} = \frac{d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h})}{\inf_{h\in\mathcal{H}_{\mathbf{n}}} d_I(\hat{f}_{\mathbf{n}}, f)(h)},\ R_{dA} = \frac{d_A(\hat{f}_{\mathbf{n}}, f)(\hat{h})}{\inf_{h\in\mathcal{H}_{\mathbf{n}}} d_A(\hat{f}_{\mathbf{n}}, f)(h)},\ R_{dM} = \frac{d_M(\hat{f}_{\mathbf{n}}, f)(\hat{h})}{\inf_{h\in\mathcal{H}_{\mathbf{n}}} d_M(\hat{f}_{\mathbf{n}}, f)(h)}$$

The histograms of these ratios for 100 simulations with different cases of sample sizes for the non-adaptive KDE are depicted in Figure 4, while the histograms for the AKDE, to

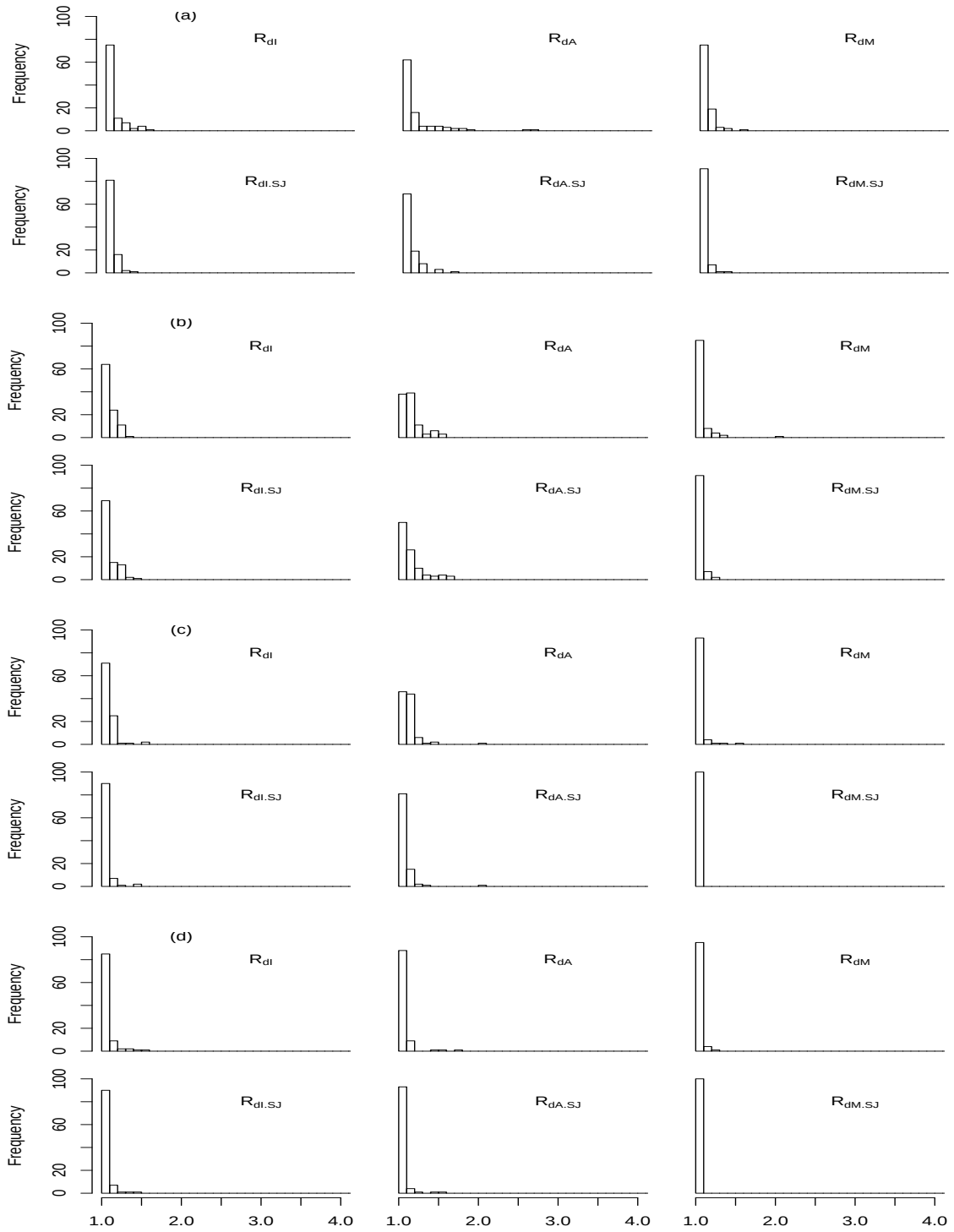


Figure 4: *Histograms of R_{dI}'s, R_{dA}'s and R_{dM}'s, and $R_{dI.SJ}$'s, $R_{dA.SJ}$'s and $R_{dM.SJ}$'s. respectively, for the CV and the SJ selected bandwidths of 100 simulations with different sample sizes of $n = m_1 * m_2$: (a) $(m_1, m_2) = (25, 10)$, (b) $(m_1, m_2) = (20, 20)$, (c) $(m_1, m_2) = (50, 50)$ and (d) $(m_1, m_2) = (100, 100)$.*

save space, in Figure B.3 in Section B3 of Appendix B, both of which do look very similar. It follows from these figures that these ratios tend to be closer to 1 as the sample size increases. In particular R_{dM} tends to converge faster than R_{dI}, and R_{dI} faster than R_{dA}, which seem understandable as R_{dA} is more sample dependent than R_{dI}, and R_{dI} than R_{dM}. Clearly, the proposed spatial CV based adaptive bandwidth selection (the non-adaptive bandwidth is its special case) works reasonably well in terms of all d_I, d_A and d_M, and it improves with the sample size becoming larger as the asymptotic optimality is indicated in Theorems 3.3 and 3.4 with the oracle property . In particular, in terms of the MISE distance d_M in (3.6), the spatial CV selected bandwidth (either with the CV or SJ as a pilot) can approximate the optimal bandwidth well even for the sample size as small as $m_1 = 25$ and $m_2 = 10$, the sample size of the real soil data set examined in Section 5.2.

Before ending this subsection, we here have a simple look at the computational cost of the approach. As can be seen from (3.1), given a pilot density, the leading order of the cost of calculating the SCV_δ in (3.1) for AKDE is $O(n^2)$, where $n = m_1 m_2$ is the total sample size of an $m_1 \times m_2$ lattice data set here. Likewise, so is the leading order for calculating the SCV_0 in (3.3) with a non-adaptive KDE. Therefore, in terms of the leading order in computational costs, the adaptive bandwidth selection by (3.14) in Section 3.2 either with a CV- or an SJ-pilot density is of the same order $O(n^2)$. In general, the procedure used above appears to run well in practice. The real elapsed time for a bandwidth selection run in R by a Dell Precision 7520 laptop of COREi7 for the simulated data sets of different sample sizes is reported in Table B.1 in Appendix B. As is seen from this table, for the sample size of $m_1 \times m_2 = 20 \times 20 = 400$, the real elapsed time needed for our proposed spatial CV procedure is about 2.5 seconds, while for the sample size as large as $m_1 \times m_2 = 100 \times 100 = 10,000$, the real elapsed time is about 110 seconds, which appears quite acceptable. This indicates that our spatial CV approach can work fast for the analysis of the real soil data of sample size $(m_1, m_2) = (25, 10)$ given in the subsection below.

5.2 An application to soil data analysis

We are analysing a spatial soil data set, soil250, in the R package geoR (c.f., Ribeiro Jr and Diggle 2016), which consists in uniformity trials with 250 undisturbed soil samples collected at 25cm soil depth of spacing of 5 meters, resulting in a regular grid of 25×10 points. The data consist of 250 observations on 22 variables concerning soil chemistry properties measured on the grid. In this analysis, we only consider 8 columns of the data table, involving the columns named Linha (the column for x-coordinate), Coluna (the column for y-coordinate), pHKCl (soil pH by potassium chloride (KCl) solution), Ca (calcium content), K (potassio content), H (hydrogen content), C (carbon content), and CTC (catium exchange capability). Here the pH by KCl measures the acidity in the soil solution, plus the reserve acidity in the colloids, and is therefore more acid than

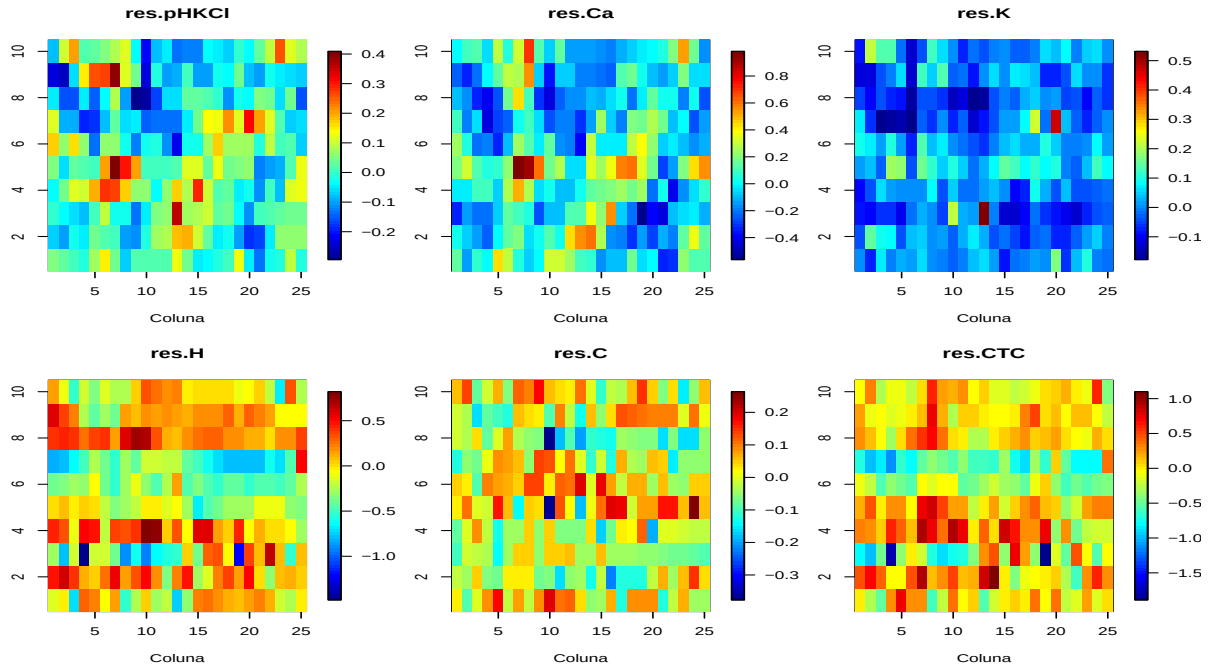


Figure 5: Soil data: The images of 6 soil properties variables after spatial trend removal by sm.regression, plotted over space, (Linha, Coluna)

pH (water), although both are neutral at pH 7.0 (c.f., DAIS 2007). Zheng *et al.* (2010) recently studied the spatial spectral density for the CTC variable, and Lu *et al.* (2014) analysed the impacts of soil chemistry properties of pHKCl, Ca, K and C as well as other variables on the CTC, an important soil property for soil conservation, of concern in agriculture science.

In the original data, as shown in Lu *et al.* (2014), there seem to be some spatial trends for all variables, so we apply sm.regression in the R package sm (c.f., Bowman and Azzalini 2014) to remove the spatial trends. The resulting spatial data of these soil chemical variables, denoted by prefix "res." standing for residual, are plotted in Figure 5, and appear to be stationary. We hence analyse the distribution of these variables based on the residual data; the different density estimates are plotted in Figure 6: (a) res.pHKCl, (b) res.Ca, (c) res.K, (d) res.H, (e) res.C, (f) res.CTC, where 'cv.akde' and 'cv.akde.SJ' are for adaptive kernel density estimate (2.1) with adaptive bandwidths combined by the cv piloted and the Sheather-Jones (SJ) piloted densities, respectively, and 'cv.kde' for the cv-based kernel density estimate (2.4) in Section 3. Here the CV selected bandwidths in Sections 3.1 and 3.2 are given in Table 2. In addition, as benchmark comparisons, the R function 'density' estimates, denoted by 'r.density' and 'r.SJ', with the default (rule of thumb) and the SJ bandwidths, respectively, and the normal density, denoted by 'normal', with the same mean and variance of the sample, are also depicted in Figure 6. It is clear that the CV-based estimated densities of all these variables indicate that distributions are basically non-Gaussian, where the densities of res.Ca, res.K, res.H, res.C and res.CTC

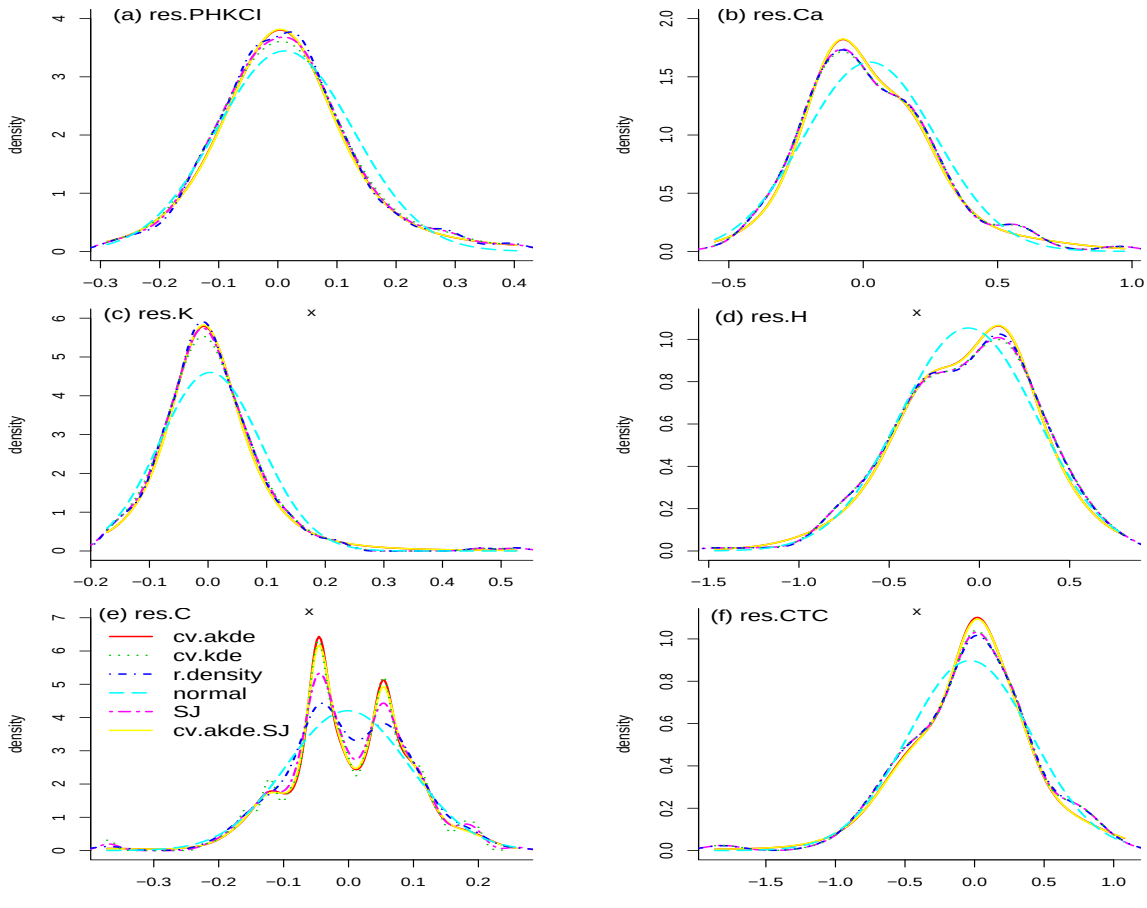


Figure 6: *Density estimation by different methods for 6 soil chemistry properties of (a) pHKCl (b) Ca, (c) K, (d) H, (e) C and (f) CTC after spatial trends removed.*

are skewed and one of them is even multi-modal (e.g., res.C). Also, from Figure 6 it is interesting to note that the estimated densities by the proposed AKDE given in (2.1) either with a CV-pilot or an SJ-pilot adaptive bandwidth are similar, which are more significantly illustrating the non-Gaussianity of these residual data than the KDEs in (2.4), including the rule of thumb, the CV and the SJ bandwidths, for the six soil variables. Furthermore, in view of the simulation above with the sample size of $m_1 = 25$ and $m_2 = 10$, the (estimated) adaptive CV-piloted AKDE strengthens this impression, demonstrating that nonlinear quantile analysis of res.CTC in relation to other covariates is needed, as indicated in Lu *et al.* (2014), for a better understanding of the data. Furthermore, from the estimated probability densities plotted in Figure 6, we can learn more on the soil properties. With the possible exception of part (a) of Figure 6, the densities in this figure are all non-Gaussian. From (a), the soil pH by potassium chloride (KCl) solution (pHKCl)

Table 2: The CV selected bandwidths for KDE and AKDE with 6 soil variables

	(a) res.pHKCl	(b) res.Ca	(c) res.K	(d) res.H	(e) res.C	(f) res.CTC
h in KDE	0.04256665	0.07467	0.02776	0.12733	0.00969	0.11465
h_0 in AKDE	0.06730755	0.09817	0.04451	0.15311	0.01712	0.16290

is basically symmetric around its spatial trend; from (b), relative to the spatial trend, the calcium content (Ca) density is positively skewed with a thicker right tail; from (c), the potassio content (K) may be seen to be basically symmetric around its spatial trend or only slightly positively skewed; from (d), relative to the spatial trend, the hydrogen content (H) density is obviously negatively skewed with a thicker left tail; from (e), the property of the carbon content (C) is more complex having at least two peaks with the negatively located peak higher than the positive one; and finally from (f), the catium exchange capability (CTC) is basically negatively skewed around its spatial trend. These properties would be helpful for soil management and conservation.

Acknowledgments: The authors are grateful to the Editor, an Associate Editor and two referees for their insightful comments and suggestions, which have led to great improvement of this present version. This research was partially supported by an European Research Agency's Marie Curie Career Integration Grant PCIG14-GA-2013-631692, and Ling's research by the National Social Science Fund of China (14ATJ005) and the National Natural Science Foundation of China (11171001), which are acknowledged.

References

[1] Abramson, I. (1982a). On bandwidth variation in kernel estimates – a square root law. Annals of Statistics 10(4), 1217-1223.

[2] Abramson, I. (1982b). Arbitrariness of the Pilot Estimator in Adaptive Kernel Methods. Journal of Multivariate Analysis, 12, 562-567.

[3] Basawa, I.V. (1996a). Special issue on spatial statistics. Part 1. J. Statist. Plann. Inference, 50, 311-411.

[4] Basawa, I.V. (1996b). Special issue on spatial statistics. Part 2. J. Statist. Plann. Inference, 51, 1-97.

[5] Bowman, A. W. and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, vol. 18. Oxford, UK: Oxford University Press.

[6] Bowman, A. W. and Azzalini, A. (2014). R package 'sm': nonparametric smoothing methods (version 2.2-5.4). URL http://www.stats.gla.ac.uk/ adrian/sm, http://azzalini.stat.unipd.it/Book_sm

[7] Cao, R., Cuevas, A,, and GonzLlez-Manteiga, W. (1994). A Comparative Study of Several Smoothing Methods in Density Estimation. Computational Statistics and Data Analysis, 17, 153-176.

[8] Carbon, M., Hallin, M., Tran, L. T. (1996). Kernel density estimation for random fields: L_1 theory, J. Nonparametric Statistics, 6, 157-170.

[9] Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. Statistica Sinica, 6, 129-145.

[10] Cox, D. D., Kim, T. Y. (1995). Moment bounds for mixing random variables useful in nonparametric function estimation, Stochastic Processes and their applications, 56, 151-158.

[11] Cressie, N. (1993). Statistics for Spatial Data, revised edition. Wiley, New York.

[12] DAIS (Directorate Agricultural Information Services) (2007). Acid soil and lime. Department of Agriculture, South Africa. http://www.nda.agric.za/docs/Infopaks/lime.pdf.

[13] Davies, T.M. and Hazelton, M.L. (2010). Adaptive kernel estimation of spatial relative risk. Statistics in Medicine, 29, 2423-2437.

[14] Doukhan, P. (1994). Mixing: Properties and Examples. New York: Springer.

[15] Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and its Applications . Chapman and Hall, London.

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

[17] Faraway, J.I. and Jhun, M. (1990). Bootstrap Choice of Bandwidth for Densitv Estimation. Journal of the American Statistical Association, 85, 1119-1122.

[18] Gao, J. (2007). Nonlinear time series: semiparametric and nonparametric methods. Chapman & Hall.

[19] Gao, J., Lu, Z. and Tjøstheim, D. (2008). Moment inequalities for spatial processes, Statistics and Probability Letters, 78, 687-697.

[20] Gao, J., Lu, Z. and Tjøstheim, D. (2006). Estimation in Semi-parametric Spatial Regression. Annals of Statistics, 34, 1395-1435.

[21] Gelfand, A.E., Diggle, P.J., Fuentes, M. and Guttorp, P. (2010). Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton, FL: CRC Press.

[22] Guo, D. and Mennis, J. (2009) Spatial data mining and geographic knowledge discovery – An introduction. *Computers, Environment and Urban Systems*, **33**, 403–408.

[23] Guyon, X. (1995). Random Fields on a Network: Modeling, Statistics, and Application. Springer-Verlag, New York.

[24] Guyon, X. (1987). Estimation dun champ par pseudo-vraisemblance conditionnelle: Etude asymptotique et application au cas Markovien. In Spatial Processes and Spatial Time Series Analysis. Proc. 6th FrancoBelgian Meeting of Statisticians (J.-J. Droesbeke et al., eds.) 1562. FUSL, Brussels.

[25] Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation, Ann. Statist. 11, 1156-1174.

[26] Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). J. R. Statist. Soc. B, 54, 475-530.

[27] Hallin, M., Lu, Z., Tran, L.T. (2001). Density estimation for spatial linear processes, Bernoulli, 7, 657-668.

[28] Hallin, M., Lu, Z., Tran, L.T. (2004a). Kernel density estimation for spatial processes: the L_1 theory, Journal of Multivariate Analysis, 88,61-75.

[29] Hallin, M., Lu, Z. and Tran, L.T. (2004b). Local Linear Spatial Regression. Annals of Statistics, 32, 2469-2500.

[30] Hallin, M., Lu, Z. and Yu, K. (2009). Local linear spatial quantile regression. Bernoulli, 15, 659-686.

[31] Hardle, W. and Vieu, P. (1992). Kernel regression smoothing of time series, Journal of Time Series Analysis, 13: 209-232.

[32] Harel, M., Lenain, J.-F., & Ngatchou-Wandji, J. (2016). Asymptotic behaviour of binned kernel density estimators for locally non-stationary random fields. Journal of Nonparametric Statistics, 28, 296–321.

[33] Hart, J. D., Vieu, P. (1990), Data-driven bandwidth choice for density estimation based on dependent data, Ann. Statist. 18, 873-890.

[34] Jenish, N. (2012). Nonparametric spatial regression under near-epoch dependence. Journal of Econometrics, 167, 224-239.

[35] Jones, M.C. (1991). The roles of ISE and MISE in density estimation. Statist. Probab. Lett., 12, 51-56.

[36] Jones, M.C, Marron, J.S. and Sheather, S.J. (1996). A Brief Survey of Bandwidth Selection for Density Estimation. Journal of the American Statistical Association, 91, 401–407.

[37] Kim, T.Y., Cox, D.D. (1995). Asymptotic behaviors of some measures of accuracy in nonparametric curve estimation with dependent observations, Journal of Multivariate Analysis, 53, 67-73.

[38] Kim, T. Y., Cox, D. D. (1997). A study on Bandwidth selection in density estimation under dependence, Journal of Multivariate Analysis, 62, 190-203.

[39] Lemke, D., Mattauch, V., Heidinger, O., Pebesma, E. and Hense, H.-W. (2015). Comparing adaptive and fixed bandwidth-based kernel density estimates in spatial cancer epidemiology. International Journal of Health Geographics, 14:15, DOI 10.1186/s12942-015-0005-9 .

[40] Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J. and Bretagnolle, V. (2014), Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecology and Biogeography, 23: 811-820.

[41] Loader, C R. (1999). Bandwidth selection: classical or plug-in? Ann. Statist., 27, 415-438.

[42] Lu, Z. (1998). On geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term. Statist. Sinica, 8, 1205-1217.

[43] Lu, Z., Tang, Q. and Cheng, L. (2014). Estimating Spatial Quantile Regression with Functional Coefficients: A robust semiparametric framework. Bernoulli, 20, 164-189.

[44] Lu, Z., Tjøstheim, D.(2014). Nonparametric estimation of probability density functions for irregularly observed spatial data, Journal of the American Statistical Association, 109, 1546-1564.

[45] Mammen, E.(1990). A short note on optimal bandwidth selection for kernel estimators, Statistics and Probability Letters, 9(1), 23-25.

[46] Marron, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. Ann. Statistics, 13, 101l-1023.

[47] Marron, J. S. (1987). A comparison of cross-validation techniques in density estimation, Ann. Statist., 15, 152-162.

[48] Marron, J. S. and Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation, Journal of Multivariate Analysis, 20, 91-113.

[49] Masry, E. (1986). Recursive probability density estimation for weakly dependent processes. IEEE Trans. Inform. Theory, 32, 254-267.

[50] Nakhapetyan, B.S. (1980). The central limit theorem for random fields with mixing conditions. In R. L. Dobrushin and Ya. G. Sinai, Eds., Advances in Probability 6, Multicomponent Systems, 531-548. New York: Marcel Dekker.

[51] Neaderhouser, C C (1980). Convergence of blocks spins defined on random fields. Journal of Statistical Physics 22, 673-684.

[52] Nordman, D.J. and Lahiri, S.N. (2004). On optimal spatial subsample size for variance estimation. The Annals of Statistics, 32, 1981–2027.

[53] Pagan, A. and Ullah, A. (1999). Nonparametric Econometrics. New York: Cambridge University Press.

[54] Pham, T.D. (1986). The mixing properties of bilinear and generalized random coefficient autoregressive models. Stochastic Process. Appl., 23, 291-300.

[55] Pham, T.D. and Tran, L. T. (1985). Some mixing properties of time series models. Stochastic Process. Appl., 19, 297-303.

[56] Quintela-del-Rio, A. (1996), Comparison of bandwidth selectors in nonparametric regression under dependence, Computational Statistics and Data Analysis, 21, 563-580.

[57] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[58] Ribeiro Jr, P.J. and Diggle, P.J. (2016). geoR: Analysis of Geostatistical Data. R package version 1.7-5.2. http://CRAN.R-project.org/package=geoR

[59] Robinson, P. (2008) Developments in the Analysis of Spatial Data. Journal of the Japan Statistical Society (issue in honour of H. Akaike), 38, 87-96.

[60] Robinson, P. (2011) Asymptotic Theory for Nonparametric Regression with Spatial Data. Journal of Econometrics, 165, 5-19

[61] Rosenblatt, M. (1985). Stationary Sequences and Random Fields. Boston: Birkhauser.

[62] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995), An Effective Bandwidth Selector for Local Least Squares Regression, Journal of the American Statistical Association , 90, 1257-1270.

[63] Scott, D. W. (1992). Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley.

[64] Sheather, S. J. and Jones M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. B, 53, 683–690.

[65] Silverman B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall: New York.

[66] Stone, C J. (1984). An asymptotically optimal window selection rule for kernel density estimation, Ann. Statist. 12, 1285-1297.

[67] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society B, 36, 111-147.

[68] Takahata, H. (1983). On the rates in the central limit theorem for weakly dependent random fields. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 64 445-456.

[69] Tjøstheim, D. (1990). Nonlinear time series and Markov chains. Adv. in Appl. Probab., 22, 587-611.

[70] Tran, L. T. (1990). Kernel density estimation on random fields, Journal of Multivariate Analysis, 34, 37-53.

[71] Vieu, P.(1991). Quadratic error for nonparametric estimates under dependence , Journal of Multivariate analysis, 39, 324-347.

[72] Withers, C S. (1981). Conditions for linear processes to be strong mixing. Z. Wahrsch. Verw. Gebiete, 57, 477-480.

[73] Xia, Y. and Li, W. K. (2002) Asymptotic behavior of bandwidth selected by cross-validation method under dependence. Journal Multivariate Analysis 83, 265-287.

[74] Zheng, Y., Zhu, J. and Roy, A. (2010). Nonparametric Bayesian inference for the spectral density function of a random field. Biometrika, 97, 238–245.

[75] Zhu, J., Huang, H.-C., and Reyes, P.E. (2010). On selection of spatial linear models for lattice data. Journal of the Royal Statistical Society Series B, 72, 389-402.

[76] Zhu, J. and Morgan, G.D. (2004). A nonparametric procedure for analyzing repeated-measures of spatially correlated data. Environmental and Ecological Statistics, 11, 431-443.

Web-based Supplementary Materials for

"On Bandwidth Choice for Spatial Data Density Estimation"

Zhenyu Jiang, Nengxiang Ling, Zudi Lu, Dag Tjøstheim, Qiang Zhang

Appendix A: Spatial dependence and assumptions

Differently from the time series data, spatial data become much more complex and involved. Fundamentally, time is only uni-directional from the past to the future, while space is multi-directional. With spatial data, the fact that the spatial sites do not have a natural ordering makes the problem of CV bandwidth selection more challenging.

We first introduce the necessary assumptions on the spatial processes. Given the stationary spatial lattice process $\{X_{\mathbf{i}}\}$ that is d-dimensional with index $\mathbf{i} = (i_1, i_2, ..., i_N) \in \mathbb{Z}^N$ $(N \geq 1)$, for any set of sites $\mathcal{S} \subset \mathbb{Z}^N$, denote by $\mathscr{B}(\mathcal{S})$ the Borel σ-field generated by $\{X_{\mathbf{i}}, \mathbf{i} \in \mathcal{S}\}$. For each couple $\mathcal{S}, \mathcal{S}'$, let $\rho(\mathcal{S}, \mathcal{S}')$ be the Euclidean distance between $\mathcal{S}$ and $\mathcal{S}'$. We assume that $\{X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N\}$ satisfies the following spatial mixing condition: There exists a function $\varphi(t) \downarrow 0$ as $t \to \infty$, such that whenever $\mathcal{S}$, $\mathcal{S}' \subset \mathbb{Z}^N$,

$$\begin{aligned} \alpha_{\mathcal{S},\mathcal{S}'}(\mathscr{B}(\mathcal{S}), \mathscr{B}(S')) &= \sup\{|\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \mathscr{B}(S), B \in \mathscr{B}(\mathcal{S}')\} \\ &\leq \psi(Card(\mathcal{S}), Card(S'))\varphi(\rho(\mathcal{S}, \mathcal{S}')), \end{aligned} \tag{A.1}$$

where $Card(\mathcal{S})$ denotes the cardinality of $\mathcal{S}$, and ψ is a symmetric positive function from $\mathbb{N} \times \mathbb{N} \to \mathbb{R}^+$ nondecreasing in each variable. For the function ψ, we assume that it satisfies either

$$\psi(n, m) \leq \min\{n, m\} \tag{A.2}$$

or

$$\psi(n, m) \leq C(n + m + 1)^k \tag{A.3}$$

for some $k > 1$ and $C > 0$; the same letter C is used for various positive constants, which may take different values in different place. Obviously, if $\psi \equiv 1$, then random field $\{X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N\}$ is called strongly mixing. See Doukhan (1994) for more discussion.

The concept of spatial mixing above concerns the conditions on the spatial data generating process, which is standard in the context of the problem under study. The α-mixing condition (A.1) is similar to (A4) of Hallin *et al.*(2004b). This is a technical assumption widely used in both nonlinear time series and spatial literature to characterize the data dependence. In the serial case, many stochastic processes and time series are shown to be strongly mixing. Withers (1981) has obtained various conditions for linear processes to be strongly mixing. Under certain weak assumptions, autoregressive and more general nonlinear time-series models are strongly mixing with exponential mixing rates; see Pham and Tran (1985), Pham (1986), Tjøstheim (1990) and Lu (1998). Guyon (1987) has shown that the results of Withers (1981) under certain conditions can be extended to linear random fields, of the form $X_{\mathbf{n}} = \sum_{\mathbf{j} \in \mathbb{Z}^2} g_{\mathbf{j}} Z_{\mathbf{n}-\mathbf{j}}$, with $\mathbf{n} \in \mathbb{Z}^2$, over gridded space, where the $Z_{\mathbf{j}}$'s are independent random variables. In addition, either (A.2) or (A.3) is

OK for our proof; not both are needed (c.f., (B.15) in Appendix B). Here conditions (A.2) and (A.3) are only some examples on the ψ part in (A.1), which are the same as the mixing conditions used by Neaderhouser (1980) and Takahata (1983), respectively, and are weaker than the uniform strong mixing condition considered by Nakhapetyan (1980). They are satisfied by many spatial models, as shown by Neaderhouser (1980), Rosenblatt (1985) and Guyon (1987). See also Lu and Tjøstheim (2014) for more discussion and some examples.

In what follows, we derive the main results of this paper under some mild assumptions.

(K1) The kernel function K is a bounded function symmetric with respect to zero, as well as Hölder continuous and compactly supported, satisfying $\int_{\mathbb{R}^d} K(t)dt = 1$.

(K2) For any non-negative components of $(i_1, i_2, ..., i_d)$ with $i_1+i_2+...+i_d \leq r$, denote

$$S(K, i_1, i_2, ..., i_d) = \int_{\mathbb{R}^d} t_1^{i_1}...t_d^{i_d} K(t)dt,$$

which satisfies the properties of rth-order kernels that

$$S(K, i_1, \cdots, i_d) = 0, \quad when \;\; for \;\; any \;\; j, \;\; i_j < r, \; with \; i_1 + i_2 + ... + i_d > 0,$$

and

$$0 < |S(K, i_1, \cdots, i_d)| < \infty, \quad if \;\; there \;\; is \;\; some \;\; j \;\; such \;\; that \;\; i_j = r,$$

where r is a positive integer given in (D1) below.

(K3) The convolution of kernel function K with itself, $\tilde{K}$, is absolutely integrable.

(K4) The kernel function K is differentiable, and its characteristic function, $\psi_K(t) = \int_{\mathbb{R}^d} e^{\iota t'u} K(u)du$, with $\iota^2 = -1$, satisfies $|\psi_K(t)| \leq c_K|\psi_{\mathcal{N}}(t)|$, where $\psi_{\mathcal{N}}(t) = e^{-t't/2}$ is the characteristic function of d-dimensional standard normal distribution, $c_K > 0$ is a constant and t' is the transpose of $t \in \mathbb{R}^d$.

(D1) The bounded density function f is Hölder continuous with rth order continuous differentiations.

(D2)(i) The joint probability density function $f_{\mathbf{i},\mathbf{j}}(x, y)$ of $X_{\mathbf{i}}$ and $X_{\mathbf{j}}$ exists and satisfies $|f_{\mathbf{i},\mathbf{j}}(x, y) - f(x)f(y)| \leq C$ for all x, y and all $\mathbf{i} \neq \mathbf{j}$. (ii) The conditional probability density function $\hat{f}_{\mathbf{i}_1,\cdots,\mathbf{i}_s|\mathbf{j}_1,\cdots,\mathbf{j}_s}(x_1, \cdots, x_s|y_1, \cdots, y_s)$ of $(X_{\mathbf{i}_1}, \cdots, X_{\mathbf{i}_s})$ given $(X_{\mathbf{j}_1} = y_1, \cdots, X_{\mathbf{j}_s} = y_s)$ is bounded for all $x_k, y_k \in \mathbb{R}^d$ and all $\mathbf{i}_k \neq \mathbf{j}_k$, $k = 1, \cdots, s$, with $1 \leq s \leq 2r$.

(W) $w(.)$ is bounded and integrable with a compact support $S_w \subset \mathbb{R}^d$.

(D3) The density function $f(\cdot)$ is bounded away from zero on S_w, that is $\inf_{x \in S_w} f(x) \geq c_w > 0$.

(H) $h \in \mathcal{H}_{\mathbf{n}} = [a\tilde{\mathbf{n}}^{-\frac{1}{2r+d}}, b\tilde{\mathbf{n}}^{-\frac{1}{2r+d}}]$ for some constants a and b with $0 < a < b < \infty$.

(M) The mixing coefficient $\varphi(t)$ satisfies

$$\varphi(t) = \mathcal{O}(t^{-\mu})$$

with $\mu > 2Nr(2-2/q)/(1-2/q)$ for some $q > 2$.

Remarks: The assumptions above are quite mild:

(i) Assumptions (K1), (K2), (K3) and (D1) that are imposed on the kernel function K and the density function f were used to obtain the bias of the estimator and asymptotical equivalence of different squared errors, respectively; see, for instance, Vieu (1991), Kim and Cox (1995, 1997), among others for details, in the non-spatial case. Assumption (K4) is a technical one needed for the case of adaptive bandwidth choice with $\delta > 0$, which is easily satisfied, say K taken as the normal probability density function itself, or having a bounded support with the constant $c_K > 0$ being sufficiently large.

(ii) Assumption (D2) is a technical condition with uniform boundedness imposed on the joint probability density functions to ensure a uniform consistency with respect to the different spatial sites. It has been similarly adopted, for instance, by Masry (1986) in the time series case, and by Tran (1990), Hallin et al. (2004b) and Lu et al. (2014) in the gridded spatial context. Obviously, (D2)(i)(ii) are valid in independent case under (D1).

(iii) Assumption (W) is general and was adopted by Marron (1987), Kim and Cox (1995), etc. Assumption (D3) is imposed on the lower bound of the density f on S_w, which is a technical condition needed for the proof in the case of $\delta > 0$ with an adaptive KDE.

(iv) Assumption (M) requires the mixing coefficients of the spatial dependent data tending to zero at a suitable rate, which is mild. For example, if $N = 1$ and $r = 2$ with a relatively large q taken, it follows from Assumption (M) that $\varphi(t) = \mathcal{O}(t^{-\mu})$ with $\mu > 2Nr(2-2/q)/(1-2/q) \approx 8$ is sufficient, which is much weaker than $\mu > 922$ required in Remark 2.1 of Hart and Vieu (1990, page 877) for the time series data.

Assumption (H) is chosen so that the vertical error (c.f. (3.8) in Theorem 3.1) contracts according to the convergence order of the optimal bandwidth. Assumption (H) is a technical assumption with $\mathcal{H}_{\mathbf{n}} = [a\tilde{\mathbf{n}}^{-1/(2r+d)}, b\tilde{\mathbf{n}}^{-1/(2r+d)}]$ for $0 < a < b < \infty$, similarly to those used in Marron and Härdle (1986), Kim and Cox (1997), Vieu (1991) and others in theory (c.f., Lemma B.4 in Appendix B). In practice, according to our empirical experience in the simulation and real data examples, as a can be chosen small while b may be large, it looked that the range $\mathcal{H}_{\mathbf{n}}$ could well include the optimal bandwidth even for the adaptive KDE.

Appendix B: Technical Proofs and Additional Figures and Table

We provide the proof of the theorems in Section 3. We will be mainly focusing on the case of $\delta = 0$ for ease of exposition below as the proofs for $\delta = 0$ and $\delta > 0$ are basically similar, where for the latter case $\delta > 0$, assumptions (K4) and (D3) in Appendix A are additionally needed (see the Proof of Theorem 3.1 in Section B2 below). Some additional figures and table are given in Section B3.

B1 Some useful lemmas and their proofs

In this section, we collect some necessary lemmas and their proofs, which are needed for the proof of the main results in the next section. Here Lemmas B.2 – B.3 given below are only stated for the case of $\delta = 0$ for saving of space, but they are easily shown to be true with slight modifications by taking notice of $\check{f}_{\mathbf{n}}$ in (2.1) (instead of $\hat{f}_{\mathbf{n}}$) and, without loss of generality, putting $h_{\mathbf{i}} = h/[f^{\delta}(X_{\mathbf{i}})]$ with $h_0 = h$ following from (2.2) for the case of $\delta > 0$ (under additional assumptions (K4) and (D3)).

Lemma B.1 (i) Suppose (3.1) holds. Let $\mathscr{L}_{\gamma}(\mathscr{F})$ denote the class of $\mathscr{F}$-measurable r.v.'s X satisfying $\|X\|_{\gamma} = (E|X|^{\gamma})^{1/\gamma} < \infty$. Let $X \in \mathscr{L}_{\gamma}(\mathscr{B}(S))$ and $Y \in \mathscr{L}_{\zeta}(\mathscr{B}(S'))$. Suppose $1 \leq \gamma, \zeta, \eta < \infty$ and $\gamma^{-1} + \zeta^{-1} + \eta^{-1} = 1$, then

$$|EXY - EXEY| \leq C\|X\|_{\gamma}\|Y\|_{\zeta} \times \{\psi(Card(\mathcal{S}), Card(\mathcal{S}'))\varphi(\rho(\mathcal{S}, \mathcal{S}'))\}^{1/\eta}. \tag{B.1}$$

(ii) For r.v.'s bounded with probability 1, the right-hand side of (B.1) can be replaced by $C\psi(Card(\mathcal{S}), Card(\mathcal{S}'))\varphi(\rho(\mathcal{S}, \mathcal{S}'))$.

Proof. See Tran (1990) and its reference.

Lemma B.2 Suppose the assumptions (K1), (D1) and (D2) hold.
(i) For any $\mathbf{i}, \mathbf{j} \in \mathcal{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{j}$, we have

$$EK^{qr}(\frac{X_{\mathbf{i}} - X_{\mathbf{j}}}{h}) = \mathcal{O}(h^d). \tag{B.2}$$

(ii) For any $\mathbf{i}_k, \mathbf{j}_k \in \mathcal{I}_{\mathbf{n}}, \mathbf{i}_k \neq \mathbf{j}_k$ and any positive integer ν_k's, with $k = 1, 2, \cdots, s$, we have

$$EK^{\nu_1}(\frac{X_{\mathbf{i}_1} - X_{\mathbf{j}_1}}{h})K^{\nu_2}(\frac{X_{\mathbf{i}_2} - X_{\mathbf{j}_2}}{h})...K^{\nu_s}(\frac{X_{\mathbf{i}_s} - X_{\mathbf{j}_s}}{h}) = \mathcal{O}(h^{ds}). \tag{B.3}$$

Proof. (i) By (K1), (D1) and (D2), we can check that

$$\begin{aligned} EK^{qr}(\frac{X_{\mathbf{i}} - X_j}{h}) &= \int\int_{\mathbb{R}^d \times \mathbb{R}^d} K^{qr}(\frac{x-y}{h}) f_{\mathbf{i},\mathbf{j}}(x,y)dxdy \\ &= \int\int_{\mathbb{R}^d \times \mathbb{R}^d} K^{qr}(u)\hat{f}_{\mathbf{i},\mathbf{j}}(y+uh, y)h^d dudy \end{aligned}$$

$$= \int\int_{\mathbb{R}^d\times\mathbb{R}^d} K^{qr}(u)\hat{f}_{\mathbf{i}|\mathbf{j}}(y+uh|y)f(y)h^d dudy = \mathcal{O}(h^d).$$

(ii) The proof for this part of (B.3) is similar to that of (i) with $f_{\mathbf{i},\mathbf{j}}(x,y)$ there replaced by the joint probability density function $\hat{f}_{\mathbf{i}_1,\mathbf{j}_1,\cdots,\mathbf{i}_s,\mathbf{j}_s}(x_1,y_1,\cdots,x_s,y_s)$ of $(X_{\mathbf{i}_1},X_{\mathbf{j}_1},\cdots,X_{\mathbf{i}_s},X_{\mathbf{j}_s})$ together with application of assumption (D2). The detail is omitted.

The following lemma plays a key role in the proof of the main results, and is of independent interest.

Lemma B.3 Suppose that Assumptions (D1), (D2), (H) and (M) hold. Let $\xi_{\mathbf{i},\mathbf{j}} = \xi(X_{\mathbf{i}},X_{\mathbf{j}},h_{\mathbf{n}})$ be a measurable function of $(X_{\mathbf{i}},X_{\mathbf{j}},h_{\mathbf{n}})$, satisfying $E\xi_{\mathbf{i},\mathbf{j}} = 0$ and $E\prod_{k=\ell}^{s}|\xi_{\mathbf{i}_k,\mathbf{j}_k}|^{\nu_k} = O(h_{\mathbf{n}}^{d(s-\ell+1)})$, uniformly with respect to $\mathbf{i}_k \neq \mathbf{j}_k \in \mathcal{I}_{\mathbf{n}}$, as $\mathbf{n}\to\infty$, where $\nu_k \leq 2r$, for $k = \ell,\cdots,s$, are positive integers, with non-negative integers $\ell \leq s$. Then we have

$$E\left(\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\right)^{2r} \leq C_1\tilde{\mathbf{n}}^2h^d + C_2(\tilde{\mathbf{n}}^2h^d)^r$$
$$+ C_3(\tilde{\mathbf{n}}^2h^d)^r\big(P^{2Nr}h^{rd} + h^{(\frac{2}{q}-1)rd}\sum_{t=P+1}^{\infty}t^{2Nr-1}\varphi(t)^{1-\frac{2}{q}}\big), \tag{B.4}$$

with $1 \leq P \leq \max\{n_1,n_2,...n_N\}$.

Proof. The result together with its proof is an extension of that of Gao et al (2008) in which the summation is only over $\mathbf{i}$. With the summation over both $\mathbf{i}$ and $\mathbf{j}$ in this lemma, the proof details become much lengthier, so we only sketch the proof here. Firstly, we have the decomposition as follows:

$$E\big(\sum_{\mathbf{i},\mathbf{j}\in I_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\big)^{2r} = \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}E[\xi_{\mathbf{i},j}^{2r}] + \sum_{s=1}^{2r-1}\sum_{\nu_0+\nu_1+\ldots+\nu_s=2r}V_s\big(\nu_0,\nu_1,...,\nu_s\big) := A+B, \tag{B.5}$$

where $\sum_{\nu_0+\nu_1+\ldots+\nu_s=2r}$ is the summation over $(\nu_0,\nu_1,...,\nu_s)$ with positive integer components satisfying $\nu_0+\nu_1+...+\nu_s = 2r$, and

$$V_s\big(\nu_0,\nu_1,...,\nu_s\big) = \sum_{(\mathbf{i}_0,\mathbf{j}_0)\neq(\mathbf{i}_1,\mathbf{j}_1)\neq\ldots\neq(\mathbf{i}_s,\mathbf{j}_s)}E\big(\xi_{\mathbf{i}_0,\mathbf{j}_0}^{\nu_0}.\xi_{\mathbf{i}_1,\mathbf{j}_1}^{\nu_1}...\xi_{\mathbf{i}_s,\mathbf{j}_s}^{\nu_s}\big), \tag{B.6}$$

where the summation $\sum_{(\mathbf{i}_0,\mathbf{j}_0)\neq(\mathbf{i}_1,\mathbf{j}_1)\neq\ldots\neq(\mathbf{i}_s,\mathbf{j}_s)}$ is over sites indexes $(\mathbf{i}_0,\mathbf{i}_1,...,\mathbf{i}_s)$ and $(\mathbf{j}_0,\mathbf{j}_1,...,\mathbf{j}_s)$, respectively with each index $\mathbf{i}_x$ and $\mathbf{j}_y$ taking value in $\mathbb{Z}^N$ from $\mathbf{1}$ to $\mathbf{n}$; and satisfying $\mathbf{i}_c \neq \mathbf{i}_d$ for any $c\neq d$, and $\mathbf{j}_k \neq \mathbf{j}_m$, for any $k\neq m$, $0\leq c,d,k,m\leq s$.

Secondly, we treat the first term A in (B.5). By the moment condition specified in the lemma, it follows that

$$A = \mathcal{O}(\tilde{\mathbf{n}}^2h^d). \tag{B.7}$$

Next, we deal with the second term B in (B.5) for $1 \le s \le r-1$. In fact, by (B.6),

$$V_s(\nu_0,\nu_1,...,\nu_s) = \sum_{(\mathbf{i}_0,\ j_0)\neq(\mathbf{i}_1,\mathbf{j}_1)\neq...\neq(\ i_s,\mathbf{j}_s)} [E(\prod_{m=0}^{s} \xi_{\mathbf{i}_m,\ j_m}^{\nu_m}) - \prod_{m=0}^{s} E\xi_{\mathbf{i}_m,\ j_m}^{\nu_m}]$$
$$+ \sum_{(\mathbf{i}_0,\mathbf{j}_0)\neq(\mathbf{i}_1,\ j_1)\neq...\neq(\mathbf{i}_s,\ j_s)} \prod_{m=0}^{s} E\xi_{\mathbf{i}_m,\ j_m}^{\nu_m} := V_{s1} + V_{s2}. \tag{B.8}$$

Obviously, by the moment condition specified in the lemma again, it follows that

$$|V_{s2}| \le C(\tilde{\mathbf{n}}^2 h^d)^{s+1}. \tag{B.9}$$

As for the first term V_{s1} in (B.8), by extending the derivation of (3.9) of Gao et al (2008), it follows that

$$|V_{s1}| \le \sum_{l=0}^{s-1} (\tilde{\mathbf{n}}^2 h^d)^l . \sum_{(\mathbf{i}_l,\mathbf{j}_l)\neq...\neq(\mathbf{i}_s,\mathbf{j}_s)} |E[\prod_{m=l}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}] - E\xi_{\mathbf{i}_l,\mathbf{j}_l}^{\nu_l} E[\prod_{m=l+1}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}]|$$
$$:= \sum_{l=0}^{s-1} (\tilde{\mathbf{n}}^2 h^d)^l V_{ls1}. \tag{B.10}$$

Thus, we need to consider that, for some $P > 0$,

$$V_{ls1} = \sum_{0<\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\})\le P} |E[\prod_{m=l}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}] - E\xi_{\mathbf{i}_l,\mathbf{j}_l}^{\nu_l} E[\prod_{m=l+1}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}]|$$
$$+ \sum_{0<\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\})>P} |E[\prod_{m=l}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}] - E\xi_{\mathbf{i}_l,\mathbf{j}_l}^{\nu_l} E[\prod_{m=l+1}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}]|$$
$$:= V_{ls11} + V_{ls12}. \tag{B.11}$$

By (B.3), we have

$$|E[\prod_{m=l}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}] - E\xi_{\mathbf{i}_l,\mathbf{j}_l}^{\nu_l} E[\prod_{m=l+1}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}]| \le |E[\prod_{m=l}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}]| + |E\xi_{\mathbf{i}_l,\mathbf{j}_l}^{\nu_l} E[\prod_{m=l+1}^{s} \xi_{\mathbf{i}_m,\mathbf{j}_m}^{\nu_m}]|$$
$$= \mathcal{O}(h^{(s-l+1)d}) + O\big(h^d\big) O\big(h^{(s-l)d}\big) = \mathcal{O}\big(h^{d(s-l+1)}\big). \tag{B.12}$$

Hence, if $\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\}) \le P$, then it follows that

$$V_{ls11} \le C h^{d(s-l+1)} \sum_{k=1}^{P} \sum_{k\le\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\})=t<k+1} 1. \tag{B.13}$$

Note that if $\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\}) = t$, then there exists some location, say $(\mathbf{i}_{l+1},\mathbf{j}_{l+1}) \in \{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\}$ such that $\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1})\}) =$

t, which leads to

$$V_{ls11} \le Ch^{d(s-l+1)}\ \tilde{n}^{2(s-(l+2)+1)} \sum_{k=1}^{P} \sum_{k\le\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1})\})=t<k+1} 1$$
$$\le Ch^{d(s-l+1)}\ \tilde{n}^{2(s-l-1)} \sum_{k=1}^{P} \tilde{\mathbf{n}}^2 k^{2N-1} \le Ch^{d(s-l+1)}\tilde{\mathbf{n}}^{2(s-l)}P^{2N}. \tag{B.14}$$

If $\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\}) = t > P$, by $q > 2$, Lemma B.1 and (B.3), we obtain

$$|E[\prod_{m=l}^{s}\xi_{\mathbf{i}_m,\,j_m}^{\nu_m}] - E\xi_{\mathbf{i}_l,\,j_l}^{\nu_l}E[\prod_{m=l+1}^{s}\xi_{\mathbf{i}_m,\,j_m}^{\nu_m}]| \le C\{E|\xi_{\mathbf{i}_l,\,j_l}|^{qr}\}^{\frac{\nu_l}{qr}}\big(E\prod_{m=l+1}^{s}|\xi_{i_m,\,j_m}|^{\frac{\nu_m qr}{2r-\nu_l}}\big)^{\frac{2r-\nu_l}{qr}}\alpha_{2,2(s-l)}^{1-\frac{2}{q}}(t)$$
$$\le Ch^{\frac{d\nu_l}{qr}}h^{d(s-l)\frac{2r-\nu_l}{qr}}\alpha_{2,2(s-l)}^{1-\frac{2}{q}}(t) \le Ch^{\frac{d\tilde{\nu}_{ls}}{qr}}\alpha_{2,2(s-l)}^{1-\frac{2}{q}}(t),$$

where $\tilde{\nu}_{ls} = \nu_l + (s-l)(2r-\nu_l)$. Thus, we have

$$V_{ls12} \le C \sum_{\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\,i_{l+1},\mathbf{j}_{l+1}),(\mathbf{i}_{l+2},\,j_{l+2}),...,(\mathbf{i}_s,\mathbf{j}_s)\})=t>P} h^{\frac{d\tilde{\nu}_{ls}}{qr}}\alpha_{2,2(s-l)}^{1-\frac{2}{q}}(t)$$
$$\le Ch^{\frac{d\tilde{\nu}_{ls}}{qr}}\sum_{k=P+1}^{\infty}\sum_{k\le\rho(\{(\mathbf{i}_l,\mathbf{j}_l)\},\{(\mathbf{i}_{l+1},\,j_{l+1}),(\mathbf{i}_{l+2},\mathbf{j}_{l+2}),...,(\mathbf{i}_s,\,j_s)\})=t<k+1}\alpha_{2,2(s-l)}^{1-\frac{2}{q}}(t)$$
$$\le Ch^{\frac{d\tilde{\nu}_{ls}}{qr}}\sum_{k=P+1}^{\infty}\tilde{n}^{2(s-(l+2)-1)}\ \tilde{n}^2\sum_{k\le\|(\mathbf{i},\mathbf{j})\|=t<k+1}\alpha_{2,2(s-l)}^{1-\frac{2}{q}}(t)$$
$$\le Ch^{\frac{d\tilde{\nu}_{ls}}{qr}}\tilde{\mathbf{n}}^{2(s-l)}\sum_{t=P+1}^{\infty}t^{2N-1}\varphi^{1-\frac{2}{q}}(t). \tag{B.15}$$

Note that in the last inequality of (B.15), the assumption of either (A.2) or (A.3) is needed but they do not really make a difference in the proof. Here we need the $\psi(n,m)$ part in the spatial mixing of (A.1) with $n = 2$ and $m = 2(s-l) \le 4r$ in (A.2) or (A.3) because $0 \le s \le 2r-1$ and $0 \le l \le s$ with r being the positive integer given in assumption (D1). Here $\psi(n,m) \le \min(n,m) \le C$ (generic positive constant) under (A.2) and $\psi(n,m) \le C(n+m+1)^k \le C$ (generic positive constant) under (A.3). In any case,

$\psi(n,m)$ is controlled by a positive generic constant C. Then, it follows that

$$
\begin{aligned}
|V_{s1}| &\le C\sum_{l=0}^{s-1}(\tilde{\mathbf{n}}^2h^d)^l V_{ls1} \le C\sum_{l=0}^{s-1}(\tilde{\mathbf{n}}^2h^d)^l[h^{d(s-l+1)}\tilde{\mathbf{n}}^{2(s-l)}P^{2N}] \\
&+ C\sum_{l=0}^{s-1}(\tilde{\mathbf{n}}^2h^d)^l[h^{\frac{d\tilde{v}_{ls}}{qr}}\tilde{\mathbf{n}}^{2(s-l)}\sum_{t=P+1}^{\infty}t^{2N-1}\varphi^{1-\frac{2}{q}}(t)] \\
&\le (\tilde{\mathbf{n}}^2h^d)^{s+1}\sum_{l=0}^{s-1}[C_1(\tilde{\mathbf{n}}^2h^d)^{l-s-1}h^{d(s-l+1)}\tilde{\mathbf{n}}^{2(s-l)}P^{2N} \\
&+ C_2(\tilde{\mathbf{n}}^2h^d)^{l-s-1}h^{\frac{d\tilde{v}_{ls}}{qr}}\tilde{\mathbf{n}}^{2(s-l)}\sum_{t=P+1}^{\infty}t^{2N-1}\varphi^{1-\frac{2}{q}}(t)] \\
&\le C(\tilde{\mathbf{n}}^2h^d)^{s+1}\sum_{l=0}^{s-1}[\tilde{\mathbf{n}}^{-2}P^{2N}+\tilde{\mathbf{n}}^{-2}h^{-Q_{ls}d}\sum_{t=P+1}^{\infty}t^{2N-1}\varphi^{1-\frac{2}{q}}(t)] \\
&= \mathcal{O}((\tilde{\mathbf{n}}^2h^d)^{s+1}) = O((\tilde{\mathbf{n}}^2h^d)^r), \qquad \text{(B.16)}
\end{aligned}
$$

where $Q_{ls} = -[(l-s-1)+\tilde{v}_{ls}/(qr)] = [(s-l)r(q-2)+qr+\nu_l(s-l-1)]/(qr) > 1$ for $0 \le l \le (s-1)$ and $q > 2$.

Finally, we treat the second term B in (B.5) for $r \le s \le 2r-1$, we only show the proof for $s = 2r-1$ in the case of $N = 2$, where we denote $\mathbf{i}_k = (i_k^1, i_k^2)$ and $\mathbf{j}_k = (j_k^1, j_k^2)$. The proof is done by extending the argument in Gao et al (2008), but some details are different. Concretely, let us denote all different (number $2r$) pairs of locations $(\mathbf{i}_0,\mathbf{j}_0),...,(\mathbf{i}_{2r-1},\mathbf{j}_{2r-1})$ as $\mathbf{s}_0,\mathbf{s}_1,...\mathbf{s}_{2r-1}$, where we denote $\mathbf{s}_k = (i_k^1, i_k^2, i_k^3, i_k^4)$ with $i_k^3 := j_k^1$ and $i_k^4 := j_k^2$ for $k = 0,1,2,...2r-1$, and $\mathbf{s}_k \ne \mathbf{s}_l$ for $k \ne l$. Arrange each of the $\mathbf{s}$ component index sets, $i_0^\ell, i_1^\ell, ..., i_{2r-1}^\ell$, for $\ell = 1,2,3,4$, in ascending orders, say, as $i_{0,\ell} \le i_{1,\ell} \le ... \le i_{2r-1,\ell}$. Then write the index increments

$$\Delta i_{k,\ell} = i_{k,\ell} - i_{k-1,\ell}, \quad \ell = 1,2,3,4,$$

for $1 \le k \le 2r-1$. Also arrange $\Delta i_{1,\ell}, \Delta i_{2,\ell}, ..., \Delta i_{2r-1,\ell}$ in decreasing orders, for $\ell = 1,2,3,4$, respectively as (for notational simplicity, we use the same notation)

$$\Delta i_{l_1^\ell,\ell} \ge \Delta i_{l_2^\ell,\ell} \ge ... \ge \Delta i_{l_{2r-1}^\ell,\ell}, \quad \ell = 1,2,3,4,$$

where $(l_1^\ell, \cdots, l_{2r-1}^\ell)$ is a permutation of $(1,\cdots,2r-1)$ for the ℓ-th component of index $\mathbf{s}$. Take $t_\ell = \Delta i_{l_r^\ell,\ell}$, for $\ell = 1,2,3,4$, and $t = max\{t_1,t_2,t_3,t_4\}$. Without loss of generality, assume $t = t_1 \ge max\{t_2,t_3,t_4\}$, then it follows that

$$
\begin{aligned}
&0 \le i_{l_k^1,1} - i_{l_k^1-1,1} \le t \le n_1, for \;\; k \in S_1 = \{r+1,...,2r-1\}; \\
&0 \le i_{l_k^2,2} - i_{l_k^2-1,2} \le t \le n_2, for \;\; k \in S_2 = \{r,...,2r-1\}; \\
&0 \le i_{l_k^3,3} - i_{l_k^3-1,3} \le t \le n_1, for \;\; k \in S_3 = \{r,...,2r-1\}; \\
&0 \le i_{l_k^4,4} - i_{l_k^4-1,4} \le t \le n_2, for \;\; k \in S_4 = \{r,...,2r-1\};
\end{aligned}
$$

that is

$$i_{l_k^\ell-1,\ell} \le i_{l_k^\ell,\ell} \le t + i_{l_k^\ell-1,\ell}, \quad \ell = 1,2,3,4, \quad for \;\; k \in S_1 = \{r+1, ..., 2r-1\}.$$

Thus, for the term of (B.6) with $s = 2r-1$, we may arrange $\mathbf{s}_0 \neq \mathbf{s}_1 \neq ... \neq \mathbf{s}_{2r-1}$ according to the order of $i_{0,1} \le i_{1,1} \le ... \le i_{2r-1,1}$. Note that $(i_{0,1}, i_{1,1}, ..., i_{2r-1,1})$ is only a permutation of $(i_0^1, i_1^1, ..., i_{2r-1}^1)$. For notational convenience, we still denote the re-ordered $\mathbf{s}_k$'s by $\mathbf{s}_k$'s, where if i_j^1 corresponds to $i_{k,1}$, then the original $\mathbf{s}_j$ corresponds to the re-ordered $\mathbf{s}_k$. Set $\mathscr{I}^\ell = \{i_{l_1^\ell,\ell}, i_{l_2^\ell,\ell}, ...i_{l_{2r-1}^\ell,\ell}\}$, $\mathscr{I}_\ell = \{i_{l_1^\ell,\ell}, i_{l_2^\ell,\ell}, ...i_{l_r^\ell,\ell}\}$, $\mathscr{I}_\ell^c = \mathscr{I}^\ell - \mathscr{I}_\ell = \{i_{l_{r+1}^\ell,\ell}, i_{l_{r+2}^\ell,\ell}, ...i_{l_{2r-1}^\ell,\ell}\}$, and $i_{l_0^\ell,\ell} = i_{0,\ell}$, for $\ell = 1,2,3,4$. We then obtain that

$$\begin{aligned}
W &= \sum_{(\mathbf{i}_0,\mathbf{j}_0)\neq(\mathbf{i}_1,\, j_1)\neq...\neq(\mathbf{i}_{2r-1},\, j_{2r-1})} |E[\xi_{\mathbf{i}_0,\mathbf{j}_0}.\xi_{i_1,\mathbf{j}_1}...\xi_{\mathbf{i}_{2r-1},\, j_{2r-1}}]| \\
&\le C \sum_{t=1}^{max\{n_1,n_2\}} \sum_{i_{l_0^1,1}=1}^{n_1} \sum_{\substack{i=1 \\ i\in\mathscr{I}_1-\{i_{l_r^1,1}\}}}^{n_1} \sum_{\substack{i_{l_k^1,1}=i_{l_k^1-1,1} \\ k\in S_1}}^{i_{l_k^1-1,1}+t} \sum_{i_{l_0^2,2}=1}^{n_2} \sum_{\substack{i=1 \\ i\in\mathscr{I}_2-\{i_{l_r^2,2}\}}}^{n_2} \sum_{\substack{i_{l_k^2,2}=i_{l_k^2-1,2} \\ k\in S_2}}^{i_{l_k^2-1,2}+t} \\
&\quad \sum_{i_{l_0^3,3}=1}^{n_1} \sum_{\substack{i=1 \\ i\in\mathscr{I}_3-\{i_{l_r^3,3}\}}}^{n_1} \sum_{\substack{i_{l_k^3,3}=i_{l_k^3-1,3} \\ k\in S_3}}^{i_{l_k^3-1,3}+t} \sum_{i_{l_0^4,4}=1}^{n_2} \sum_{\substack{i=1 \\ i\in\mathscr{I}_4-\{i_{l_r^4,4}\}}}^{n_2} \sum_{\substack{i_{l_k^4,4}=i_{l_k^4-1,4} \\ k\in S_4}}^{i_{l_k^4-1,4}+t} |E[\xi_{s_0}\xi_{\mathbf{s}_1}...\xi_{\mathbf{s}_{2r-1}}]|. \qquad \text{(B.17)}
\end{aligned}$$

Taking positive constant P such that $0 < P < \max\{n_1, n_2\}$, then the right-hand-side of (B.17) can be divided into two parts depending on $1 \le t \le P$ and $t > P$, denoted by W_1 and W_2, respectively.

For $1 \le t \le P$, by (B.3), we have

$$\begin{aligned}
W_1 &= C \sum_{t=1}^{P} \sum_{i_{l_0^1,1}=1}^{n_1} \sum_{\substack{i=1 \\ i\in\mathscr{I}_1-\{i_{l_r^1,1}\}}}^{n_1} \sum_{\substack{i_{l_k^1,1}=i_{l_k^1-1,1} \\ k\in S_1}}^{i_{l_k^1-1,1}+t} \sum_{i_{l_0^2,2}=1}^{n_2} \sum_{\substack{i=1 \\ i\in\mathscr{I}_2-\{i_{l_r^2,2}\}}}^{n_2} \sum_{\substack{i_{l_k^2,2}=i_{l_k^2-1,2} \\ k\in S_2}}^{i_{l_k^2-1,2}+t} \\
&\quad \sum_{i_{l_0^3,3}=1}^{n_1} \sum_{\substack{i=1 \\ i\in\mathscr{I}_3-\{i_{l_r^3,3}\}}}^{n_1} \sum_{\substack{i_{l_k^3,3}=i_{l_k^3-1,3} \\ k\in S_3}}^{i_{l_k^3-1,3}+t} \sum_{i_{l_0^4,4}=1}^{n_2} \sum_{\substack{i=1 \\ i\in\mathscr{I}_4-\{i_{l_r^4,4}\}}}^{n_2} \sum_{\substack{i_{l_k^4,4}=i_{l_k^4-1,4} \\ k\in S_4}}^{i_{l_k^4-1,4}+t} |E[\xi_{s_0}\xi_{\mathbf{s}_1}...\xi_{\mathbf{s}_{2r-1}}]| \\
&\le C(n_1 n_1^{r-1} n_2 n_2^{r-1})^2 \sum_{t=1}^{P} t^{4r-1} h^{2rd} \le C(n_1 n_2)^{2r} P^{4r} h^{2rd}. \qquad \text{(B.18)}
\end{aligned}$$

[For general N, the bound is $C(n_1 \cdots n_N)^{2r} P^{2Nr} h^{2rd}$.]

For $t > P$, extending the derivation for (3.19) in Gao et al (2008, page 695), clearly, if $i_{1,1}$ and $i_{2r-1,1}$ are not in the set of indices $\mathscr{I}_1$, there exist two successive indices, say $i_{k^*,1}$ and $i_{k^*+1,1}$ in $\mathscr{I}_1$, or else one of $i_{1,1}$ and $i_{2r-1,1}$ belongs to $\mathscr{I}_1$. Let $\rho(\cdot,\cdot)$ be

the generic Euclidean distance. Therefore, either $\rho(\{\mathbf{s}_0,\cdots,\mathbf{s}_{k^*-1}\},\{\mathbf{s}_{k^*}\}) \geq \Delta i_{k^*,1} \geq t$, $\rho(\{\mathbf{s}_{k^*}\},\{\mathbf{s}_{k^*+1},\cdots,\mathbf{s}_{2r-1}\}) \geq \Delta i_{k^*+1,1} \geq t$ and $\rho(\{\mathbf{s}_0,\cdots,\mathbf{s}_{k^*-1}\},\{\mathbf{s}_{k^*},\cdots,\mathbf{s}_{2r-1}\}) \geq \Delta i_{k^*+1,1} \geq t$, or $\rho(\{\mathbf{s}_0\},\{\mathbf{s}_1,\cdots,\mathbf{s}_{2r-1}\}) \geq \Delta i_{1,1} \geq t$, or $\rho(\{\mathbf{s}_0,\cdots,\mathbf{s}_{2r-2}\},\{\mathbf{s}_{2r-1}\}) \geq \Delta i_{2r-1,1} \geq t$. Set $A_{\mathbf{s}_{k^*-1}} := \xi_{\mathbf{s}_0}\xi_{\mathbf{s}_1}\cdots\xi_{\mathbf{s}_{k^*-1}}$, $B_{\mathbf{s}_{k^*+1}} := \xi_{\mathbf{s}_{k^*+1}}\cdots\xi_{\mathbf{s}_{2r-1}}$. Then for the case of i_{k^*} and i_{k^*+1} in $\mathscr{I}_1$, by (B.3) and noting that $E\xi_{\mathbf{s}_{k^*}} = 0$,

$$
\begin{aligned}
&E\xi_{\mathbf{s}_0}\xi_{\mathbf{s}_1}\cdots\xi_{\mathbf{s}_{2r-1}} = EA_{\mathbf{s}_{k^*-1}}\xi_{\mathbf{s}_{k^*}}B_{\mathbf{s}_{k^*+1}} = cov(A_{\mathbf{s}_{k^*-1}},\xi_{\mathbf{s}_{k^*}}B_{\mathbf{s}_{k^*+1}}) + EA_{\mathbf{s}_{k^*-1}}E\xi_{\mathbf{s}_{k^*}}B_{\mathbf{s}_{k^*+1}} \\
&\leq (E(A_{\mathbf{s}_{k^*}})^q)^{1/q}(E(\xi_{\mathbf{s}_{k^*}}B_{\mathbf{s}_{k^*+1}})^q)^{1/q}\alpha_{k^*,2r-k^*}^{1-2/q}(t) \\
&\qquad + EA_{\mathbf{s}_{k^*-1}}(E\xi_{\mathbf{s}_{k^*}}^q)^{1/q}(EB_{\mathbf{s}_{k^*+1}}^q)^{1/q}\alpha_{1,2r-k^*-1}^{1-2/q}(t) \\
&\leq C(h^{d\{k^*/q+(2r-k^*)/q\}}\varphi^{1-2/q}(t) + C(h^{d\{k^*+1/q+(2r-k^*-1)/q\}}\varphi^{1-2/q}(t) \leq Ch^{2rd/q}\varphi^{1-2/q}(t);
\end{aligned}
$$

and for the case of i_1 or i_{2r-1} in $\mathscr{I}_1$, the same can be got more easily. Note that

$$
W_2 = C\sum_{t=P+1}^{\max\{n_1,n_2\}}\sum_{i_{l_0^1,1}=1}^{n_1}\sum_{\substack{i=1\\ i\in\mathscr{I}_1-\{i_{l_r^1,1}\}}}^{n_1}\sum_{\substack{i_{l_k^1,1}=i_{l_k^1-1,1}\\ k\in S_1}}^{i_{l_k^1-1,1}+t}\sum_{i_{l_0^2,2}=1}^{n_2}\sum_{\substack{i=1\\ i\in\mathscr{I}_2-\{i_{l_r^2,2}\}}}^{n_2}\sum_{\substack{i_{l_k^2,2}=i_{l_k^2-1,2}\\ k\in S_2}}^{i_{l_k^2-1,2}+t}
$$

$$\sum_{i_{l_0^3,3}=1}^{n_1}\sum_{\substack{i=1\\ i\in\mathscr{I}_3-\{i_{l_r^3,3}\}}}^{n_1}\sum_{\substack{i_{l_k^3,3}=i_{l_k^3-1,3}\\ k\in S_3}}^{i_{l_k^3-1,3}+t}\sum_{i_{l_0^4,4}=1}^{n_2}\sum_{\substack{i=1\\ i\in\mathscr{I}_4-\{i_{l_r^4,4}\}}}^{n_2}\sum_{\substack{i_{l_k^4,4}=i_{l_k^4-1,4}\\ k\in S_4}}^{i_{l_k^4-1,4}+t}|E[\xi_{s_0}\xi_{\mathbf{s}_1}...\xi_{\mathbf{s}_{2r-1}}]|.$$

Therefore, we have

$$W_2\le C(n_1n_2)^{2r}\sum_{t=P+1}^{\infty}t^{4r-1}h^{2rd/q}\varphi^{1-2/q}(t)=C(n_1n_2)^{2r}h^{2rd/q}\sum_{t=P+1}^{\infty}t^{4r-1}\varphi^{1-2/q}(t). \quad \text{(B.19)}$$

[For general N, the bound is $C(n_1\cdots n_N)^{2r}h^{2rd/q}\sum_{t=P+1}^{\infty}t^{2Nr-1}\varphi^{1-2/q}(t)$.] Thus, by (B.17), (B.18) and (B.19), for $N=2$, it follows that

$$W\le C(n_1^2n_2^2h^d)^r\big(P^{4r}h^{rd}+h^{(\frac{2}{q}-1)rd}\sum_{t=P+1}^{\infty}t^{4r-1}\varphi(t)^{1-\frac{2}{q}}\big).$$

Similarly, for general N, we can have

$$W\le C(\tilde{\mathbf{n}}^2h^d)^r(P^{2Nr}h^{rd}+h^{(\frac{2}{q}-1)rd}\sum_{t=P+1}^{\infty}t^{2Nr-1}\varphi(t)^{1-\frac{2}{q}}). \quad \text{(B.20)}$$

Hence, by (B.7), (B.8), (B.9),(B.16) and (B.20), (B.4) is valid.

B2 Proofs of the main results

In this section, we present the proofs of the main results. We need the following lemma for the main proofs.

Lemma B.4 (i) In the case of $\delta=0$, if the assumptions (K1)-(K3), (W) and (M) hold, then we have

$$\begin{aligned}MISE(h)&=E\int(\hat{f}_{\mathbf{n}}(x)-f(x))^2w(x)dx=h^{2r}\int_{\mathbb{R}^d}B_f^2(x)w(x)dx+o(h^{2r})\\&\quad+\frac{1}{\tilde{\mathbf{n}}h^d}\int_{\mathbb{R}^d}f(x)w(x)dx.\int_{\mathbb{R}^d}K^2(u)du+o(\frac{1}{\tilde{\mathbf{n}}h^d}),\end{aligned} \quad \text{(B.21)}$$

where $B_f(x)=\sum_{i=1}^d C_1^{(i)}(K)T_{f,r}^{(i)}(x)$, with $C_1^{(i)}(K)=\frac{1}{r!}\int_{\mathbb{R}^d}t_i^rK(t)dt$ and $T_{f,r}^{(i)}(x)=\frac{\partial^r f(x)}{\partial x_i^r}$, for $x=(x_1,x_2,...,x_d)$, $t=(t_1,t_2,...,t_d)$ and $i=1,2,...,d$.

(ii) In the case of $\delta>0$, for $h_{\mathbf{i}}=h/[f^{\delta}(X_{\mathbf{i}})]$ with $h_0=h$, in addition to the assumptions in (i), if Assumptions (K4)) and (D3) hold, then we have

$$\begin{aligned}MISE(h)&=E\int(\check{f}_{\mathbf{n}}(x)-f(x))^2w(x)dx=h^{2r}\int_{\mathbb{R}^d}B_f^2(x)w(x)dx+o(h^{2r})\\&\quad+\frac{1}{\tilde{\mathbf{n}}h^d}\int_{\mathbb{R}^d}(f(x))^{1+d\delta}w(x)dx.\int_{\mathbb{R}^d}K^2(u)du+o(\frac{1}{\tilde{\mathbf{n}}h^d}),\end{aligned} \quad \text{(B.22)}$$

where $B_f(x) = (1-r\delta)f(x)^{-r\delta}\sum_{i=1}^{d}\frac{\partial^r f(x)}{\partial x_i^r}C_1^{(i)}(K)$. Note that under the optimal $\delta = 1/r$, $B_f(x) \equiv 0$.

Proof. (i) First, we note that $MISE(h)$ admits the variance-bias square decomposition as follows:

$$MISE(h) = \int_{\mathbb{R}^d} Var(\hat{f}_{\mathbf{n}}(x))w(x)dx + \int_{\mathbb{R}^d}(E\hat{f}_{\mathbf{n}}(x) - f(x))^2 w(x)dx. \tag{B.23}$$

Second, by stationarity, we have

$$E\hat{f}_{\mathbf{n}}(x) = \frac{1}{h^d}EK\big(\frac{x-X_{\mathbf{1}}}{h}\big) = \int_{\mathbb{R}^d}\frac{1}{h^d}K\big(\frac{x-y}{h}\big)f(y)dy = \int_{\mathbb{R}^d}K\big(y\big)f(x-yh)dy.$$

Then, by (K1), (K2) and (K3), applying Lemma 9.1 of Vieu (1991) to $\mu = f$ leads to

$$B_{\mathbf{n}}(x) := E\hat{f}_{\mathbf{n}}(x) - f(x) = h^r B_f(x) + o(h^r). \tag{B.24}$$

Hence, together with assumption (W), we obtain

$$\int_{\mathbb{R}^d}(E\hat{f}_{\mathbf{n}}(x) - f(x))^2 w(x)dx = h^{2r}\int_{\mathbb{R}^d}B_f^2(x)w(x)dx + o(h^{2r}) \tag{B.25}$$

On the other hand, by (M) with $\varphi(t) = t^{-\mu}$, it follows that $\sum_{t=1}^{\infty}t^{N-1}(\varphi(t))^{1-\frac{2}{q}} < \infty$ for $\mu > N/(1-2/q)$ and $q > 2$. Applying Lemma 2.2 of Tran (1990), we have

$$\lim_{\mathbf{n}\to\infty}\tilde{\mathbf{n}}h^d Var(\hat{f}_{\mathbf{n}}(x)) = f(x)\int_{\mathbb{R}^d}K^2(u)du,$$

which implies

$$\int_{\mathbb{R}^d}Var(\hat{f}_{\mathbf{n}}(x))w(x)dx = \frac{1}{\tilde{\mathbf{n}}h^d}\int_{\mathbb{R}^d}f(x)w(x)dx.\int_{\mathbb{R}^d}K^2(u)du + o(\frac{1}{\tilde{\mathbf{n}}h^d}). \tag{B.26}$$

Finally, by (B.23), (B.25) and (B.26), (B.21) is valid.

(ii) For $\delta > 0$, notice that the only difference between this proof and that of (i) is to replace $h^{-d}K\left(\frac{x-X_{\mathbf{i}}}{h}\right)$ in KDE with (i) by $h_{\mathbf{i}}^{-d}K\left(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}}\right)$ in ADKE for this part with $h_{\mathbf{i}} = h(f(X_{\mathbf{i}}))^{-\delta}$ (without loss of generality). With this in mind, for example, corresponding to (B.24), by assumptions (K1), (K2), (K3), (W) and (D1) together with assumptions (K4) and (D3),

$$\begin{aligned}Eh_{\mathbf{i}}^{-d}K\left(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}}\right) &= Eh_{\mathbf{i}}^{-d}K\left(\frac{X_{\mathbf{i}}-x}{h_{\mathbf{i}}}\right) = h^{-d}\int(f(u))^{d\delta}K\left(\frac{(u-x)(f(u))^{\delta}}{h}\right)f(u)du \\ &= \int(f(x+hu))^{1+d\delta}K\left(uf^{\delta}(x+hu)\right)du,\end{aligned}$$

and then we can show that

$$\begin{aligned} B_{\mathbf{n}}(x) &:= Eh_{\mathbf{i}}^{-d} K\left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{i}}}\right) - f(x) \\ &= \int \left[(f(x+hu))^{1+d\delta} K\left(uf^{\delta}(x+hu)\right) - (f(x))^{1+d\delta} K\left(uf^{\delta}(x)\right)\right] du \\ &= \int \left[J_u(f(x+hu)) - J_u(f(x))\right] du, \end{aligned} \tag{B.27}$$

where $J_u(y) = K(y^{\delta}u)y^{1+d\delta}$. Then, note that $J_u^{(1)}(y) = \sum_{i=1}^{d} K_i(y^{\delta}u)\delta y^{\delta-1}u_i y^{1+d\delta} + K(y^{\delta}u)(1+d\delta)y^{1+d\delta-1} = y^{d\delta}[\delta \sum_{i=1}^{d} K_i(y^{\delta}u)y^{\delta}u_i + (1+d\delta)K(y^{\delta}u)]$ and by assumption (D1)

$$\begin{aligned} E_{\mathbf{n}}(u) \equiv f(x+hu) - f(x) &= \sum_{j=1}^{r-1} \frac{1}{j!} \sum_{i_1=1}^{d} \cdots \sum_{i_j=1}^{d} f_{i_1 \cdots i_j}(x) h^j u_{i_1} \cdots u_{i_j} \\ &+ \frac{1}{r!} \sum_{i_1=1}^{d} \cdots \sum_{i_r=1}^{d} f_{i_1 \cdots i_r}(x+thu) h^r u_{i_1} \cdots u_{i_r}, \end{aligned} \tag{B.28}$$

where $|t| \leq 1$, and $K_i(x)$ and $f_{i_1 \cdots i_j}(x)$ stand for the first order partial derivative of $K(x)$ with respect to $x_{\mathbf{i}}$ and the j-order partial derivative of $f(x)$ with respect to $x_{i_1}, \cdots, x_{i_j}$, respectively. We then have with application of Taylor's expansion that

$$J_u(f(x+hu)) - J_u(f(x)) = J_u^{(1)}(f(x) + t_1 E_{\mathbf{n}}(u))(E_{\mathbf{n}}(u)), \tag{B.29}$$

where $|t_1| \leq 1$. Thus as done in Abramson (1982a, page 1220), by assumption (K2), it easily follows from (B.27)–(B.29) that

$$\begin{aligned} B_{\mathbf{n}}(x) :&= \int \left[J_u(f(x+hu)) - J_u(f(x))\right] du \\ &= (1+o(1))f(x)^{(1+d)\delta}\delta \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} \sum_{j=1}^{d} \int K_j(uf^{\delta}(x)) u_j u_i^r du \ \frac{h^r}{r!} \\ &\quad + (1+o(1))(1+\delta d)f(x)^{d\delta} \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} \int K(uf^{\delta}(x)) u_i^r du \ \frac{h^r}{r!} \\ &= (1+o(1))f(x)^{(1+d)\delta}\delta \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} \sum_{j=1}^{d} \int K_j(v) \frac{v_j}{f^{\delta}(x)} \frac{v_i^r}{f^{r\delta}(x)} \frac{1}{f^{d\delta}(x)} dv \ \frac{h^r}{r!} \\ &\quad + (1+o(1))(1+\delta d)f(x)^{d\delta} \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} \int K(v) \frac{v_i^r}{f^{r\delta}(x)} \frac{1}{f^{d\delta}(x)} dv \ \frac{h^r}{r!} \\ &= (1+o(1))f(x)^{-r\delta}\delta \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} (-(r+1) - (d-1)) \int K(v) v_i^r dv \ \frac{h^r}{r!} \end{aligned}$$

$$+ (1+o(1))(1+\delta d) f(x)^{-r\delta} \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} \int K(v) v_i^r dv \; \frac{h^r}{r!}$$

$$= (1+o(1))(1-r\delta) f(x)^{-r\delta} \sum_{i=1}^{d} \frac{\partial^r f(x)}{\partial x_i^r} \int K(v) v_i^r dv \; \frac{h^r}{r!} =: h^r B^*(x), \qquad \text{(B.30)}$$

and hence the bias term (B.24) is still true (and actually of order $o(h^r)$ when taking the optimal $\delta = 1/r$ in the context of Section 3.1, which is an extension of Abramson (1982a) who considered $r = 2$). For the asymptotic variance, similarly to Lemma 2.2 of Tran (1990), we have

$$\lim_{\mathbf{n}\to\infty} \tilde{\mathbf{n}} h^d Var(\check{f}_n(x)) = \lim_{\mathbf{n}\to\infty} h^d E h_{\mathbf{i}}^{-2d} K^2 \left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{i}}} \right)$$

$$= \lim_{\mathbf{n}\to\infty} h^d \int_{\mathbb{R}^d} \frac{1}{h^{2d}} f^{2d\delta}(y) K^2\big(\frac{x-y}{h} f^{\delta}(y)\big) f(y) dy = f^{1+d\delta}(x) \int_{\mathbb{R}^d} K^2(u) du.$$

Now (B.22) easily follows as done for Part (i) above.

Proof of Theorem 3.1: Note that without loss of generality we put $h_{\mathbf{i}} = h/[f^{\delta}(X_{\mathbf{i}})]$ with $h_0 = h$. Define

$$U_{\mathbf{i},\mathbf{j}} := \frac{1}{h_{\mathbf{i}}^d} K(\frac{X_{\mathbf{i}} - X_{\mathbf{j}}}{h_{\mathbf{i}}}) w(X_{\mathbf{j}}) - \int_{\mathbb{R}^d} \frac{1}{h_{\mathbf{i}}^d} K(\frac{X_{\mathbf{i}} - x}{h_{\mathbf{i}}}) f(x) w(x) dx - f(X_{\mathbf{j}}) w(X_{\mathbf{j}})$$
$$+ \int_{\mathbb{R}^d} f(x)^2 w(x) dx, \qquad \text{(B.31)}$$

$$W_{\mathbf{j}} := \frac{1}{h^d} \int_{\mathbb{R}^d} f^{d\delta}(u) K(\frac{u - X_{\mathbf{j}}}{h} f^{\delta}(u)) w(X_{\mathbf{j}}) f(u) du$$
$$- \frac{1}{h^d} \int \int_{\mathbb{R}^d \times \mathbb{R}^d} f^{d\delta}(u) K(\frac{u-x}{h} f^{\delta}(u)) f(x) f(u) w(x) dx du$$
$$- f(X_{\mathbf{j}}) w(X_{\mathbf{j}}) + \int_{\mathbb{R}^d} f(x)^2 w(x) dx \qquad \text{(B.32)}$$

and

$$V_{\mathbf{i},\mathbf{j}} := U_{\mathbf{i},\mathbf{j}} - W_{\mathbf{j}} \qquad \text{(B.33)}$$

for all $\mathbf{i}, \mathbf{j} \in \mathcal{I}_{\mathbf{n}}$ and $\mathbf{i} \neq \mathbf{j}$.

Then, by simple arithmetical calculations, we have

$$|CV_{\delta}(h) - ISE(h) - T| = \frac{2}{\tilde{\mathbf{n}}(\tilde{\mathbf{n}} - 1)} |\sum_{\mathbf{i},\mathbf{j} \in \mathcal{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{j}} U_{\mathbf{i},\mathbf{j}}|. \qquad \text{(B.34)}$$

Hence, by (B.33) and (B.34), it is sufficient to show that the following equalities hold:

$$\frac{1}{\tilde{\mathbf{n}}(\tilde{\mathbf{n}} - 1)} MISE^{-1}(h) |\sum_{\mathbf{i},\mathbf{j} \in \mathcal{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{j}} V_{\mathbf{i},\mathbf{j}}| = \mathcal{O}_p(b_{\mathbf{n}}) \qquad \text{(B.35)}$$

and

$$\frac{1}{\tilde{\mathbf{n}}} MISE^{-1}(h) |\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n}} W_\mathbf{j}| = \mathcal{O}_p(b_\mathbf{n}), \tag{B.36}$$

respectively, where $b_\mathbf{n} = \tilde{\mathbf{n}}^{-\frac{d}{2(2r+d)}}$.

Verification of (B.35): In the non-spatial case ($N = 1$), the argument given by Hart and Vieu (1990, page 885) for a similar result to (B.35) heavily depends on the natural ordering of i and j of one dimension (c.f., (4.6) of Hart and Vieu (1990)), so their argument cannot simply apply to our spatial case ($N > 1$) where $\mathbf{i}$ and $\mathbf{j}$ lack a natural ordering between them. Also note that the argument by Marron (1987, page 159) does not work as $E(V_{\mathbf{i},\mathbf{j}}) \neq 0$ and hence $E(V_{\mathbf{i},\mathbf{j}}|X_\mathbf{i}) \neq 0$ owing to spatial dependence between $X_\mathbf{i}$ and $X_\mathbf{j}$ in our case.

First, we deal with $EV_{\mathbf{i},\mathbf{j}}$. By Lemma B.2, $EV_{\mathbf{i},\mathbf{j}} = O(1)$, and also note that

$$\begin{aligned}
EV_{\mathbf{i},\mathbf{j}} &= \frac{1}{h^d} \int\int_{\mathbb{R}^d\times\mathbb{R}^d} f^{d\delta}(u) K(\frac{u-x}{h} f^\delta(u)) [f_{\mathbf{i},\mathbf{j}}(u,x) - f(u)f(x)] w(x) dx du \\
&= \frac{1}{h^d} \int\int_{\mathbb{R}^d\times\mathbb{R}^d} f^{d\delta}(u) \frac{1}{(2\pi)^d} \int e^{-\iota t' \frac{u-x}{h} f^\delta(u)} \psi_K(t) [f_{\mathbf{i},\mathbf{j}}(u,x) - f(u)f(x)] w(x) dt dx du \\
&= \int\int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{(2\pi)^d} \int e^{-\iota t'(u-x)} \psi_K(ht/f^\delta(u)) [f_{\mathbf{i},\mathbf{j}}(u,x) - f(u)f(x)] w(x) dt dx du \\
&= \frac{1}{(2\pi)^d} \int \text{ cov}\left(e^{-\iota t' X_\mathbf{i}} \psi_K(h_\mathbf{i} t), e^{\iota t' X_\mathbf{j}} w(X_\mathbf{j}) \right) dt,
\end{aligned}$$

where $\psi_K(t) = \int_{\mathbb{R}^d} e^{\iota t' u} K(u) du$, with $\iota^2 = -1$ and t' is the transpose of $t \in \mathbb{R}^d$.

We first look at the case of $\delta = 0$, i.e., $h_\mathbf{i} \equiv h$ (fixed). Then by Lemma B.1(ii),

$$\begin{aligned}
|\text{cov}\left(e^{-\iota t' X_\mathbf{i}} \psi_K(h_\mathbf{i} t), e^{\iota t' X_\mathbf{j}} w(X_\mathbf{j}) \right)| &= |\psi_K(ht)\ \text{cov}\left(e^{-\iota t' X_\mathbf{i}}, e^{\iota t' X_\mathbf{j}} w(X_\mathbf{j}) \right)| \\
&\leq C |\psi_K(ht)| \varphi(\|\mathbf{i}-\mathbf{j}\|),
\end{aligned}$$

and therefore

$$\begin{aligned}
\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n}, \mathbf{i}\neq\mathbf{j}} |EV_{\mathbf{i},\mathbf{j}}| &\leq \sum_{k=1}^{P} \sum_{\substack{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n} \\ k\leq\|\mathbf{i}-\mathbf{j}\|=t<k+1}} O(1) + \sum_{k=P+1}^{\infty} \sum_{\substack{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n} \\ k\leq\|\mathbf{i}-\mathbf{j}\|=t<k+1}} C\varphi(\|\mathbf{i}-\mathbf{j}\|) \int |\psi_K(ht)| dt \\
&\leq C \left[\tilde{\mathbf{n}} P^N + \tilde{\mathbf{n}} h^{-d} \sum_{k=P}^{\infty} k^{N-1} \varphi(k) \right].
\end{aligned} \tag{B.37}$$

For the case $\delta > 0$, by assumptions (D1) and (K4), it follows that $h_\mathbf{i} = h/f^\delta(X_\mathbf{i}) \geq h/c_f^\delta$ with $c_f > 0$ standing for the upper bound of f due to (D1), and $|\psi_K(h_\mathbf{i} t)| \leq c_K |\psi_\mathcal{N}(h_\mathbf{i} t)| = c_K e^{-h_\mathbf{i}^2 t' t/2} \leq c_K e^{-h^2 t' t/(2c_f^2)} = \psi_K^*(ht)$, with $\psi_K^*(t) = c_K e^{-t' t/(2c_f^2)}$, by which

together with Lemma B.1(i) with $\gamma \to \infty$, $\zeta \to \infty$ and $\eta \to 1$, we easily have

$$|\mathrm{cov}\left(e^{-\iota t' X_{\mathbf{i}}}\psi_K(h_{\mathbf{i}}t), e^{\iota t' X_{\mathbf{j}}}w(X_{\mathbf{j}})\right)| \le C\|e^{-\iota t' X_{\mathbf{i}}}\psi_K(h_{\mathbf{i}}t)\|_\infty \|e^{\iota t' X_{\mathbf{j}}}w(X_{\mathbf{j}})\|_\infty \varphi(\|\mathbf{i}-\mathbf{j}\|)$$
$$\le C|\psi_K^*(ht)|\varphi(\|\mathbf{i}-\mathbf{j}\|).$$

Thus replacing $\psi_K(ht)$ in (B.37) by $\psi_K^*(ht)$, we easily see that (B.37) still holds true.

Therefore $\frac{1}{\tilde{\mathbf{n}}(\tilde{\mathbf{n}}-1)}MISE^{-1}(h)|\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} EV_{\mathbf{i},\mathbf{j}}| \le C[h^d P^N + \sum_{k=P}^\infty k^{N-1}\varphi(k)] = \mathcal{O}(b_{\mathbf{n}})$, under $P^N\sum_{k=P}^\infty k^{N-1}\varphi(k) = O(1)$ as $P \to \infty$, where we take $P = \lfloor b_{\mathbf{n}}^{-1/N} \rfloor$ and $b_{\mathbf{n}} = O(h^{d/2})$, and $\lfloor b \rfloor$ stands for the integer part of b.

Now, to verify (B.35), we only need to show

$$\tilde{\mathbf{n}}^{-2}MISE^{-1}(h)\left|\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}(V_{\mathbf{i},\mathbf{j}} - EV_{\mathbf{i},\mathbf{j}})\right| = O_P(b_{\mathbf{n}}). \tag{B.38}$$

Set $\eta_{\mathbf{i},\mathbf{j}} = K(\frac{X_{\mathbf{i}}-X_{\mathbf{j}}}{h_{\mathbf{i}}})w(X_{\mathbf{j}}) - EK(\frac{X_{\mathbf{i}}-X_{\mathbf{j}}}{h_{\mathbf{i}}})w(X_{\mathbf{j}})$, $\eta_{\mathbf{i}} = \int\left(K(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}}) - EK(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}})\right)w(x)f(x)dx$, $\eta_{\mathbf{j}} = \int\left(K(\frac{u-X_{\mathbf{j}}}{h/f^\delta(u)})w(X_{\mathbf{j}}) - EK(\frac{u-X_{\mathbf{j}}}{h/f^\delta(u)})w(X_{\mathbf{j}})\right)f(u)du$. Then note that $\xi_{\mathbf{i},\mathbf{j}} := h_{\mathbf{i}}^d(V_{\mathbf{i},\mathbf{j}} - EV_{\mathbf{i},\mathbf{j}}) = \eta_{\mathbf{i},\mathbf{j}} - \eta_{\mathbf{i}} - \eta_{\mathbf{j}}$.

We first consider the case $\delta = 0$, i.e. $h_{\mathbf{i}} = h$. Then by Assumptions (K1) and (D1), it easily follows that $|\eta_{\mathbf{i}}| \le Ch_{\mathbf{n}}^d$ and $|\eta_{\mathbf{j}}| \le Ch_{\mathbf{n}}^d$. We now check the moment conditions for $\xi_{\mathbf{i},\mathbf{j}}$ specified in Lemma B.3. Obviously $E\xi_{\mathbf{i},\mathbf{j}} = 0$. Further, together with Lemma B.2, for any positive integers $\nu_k \le 2r$, $|\xi_{\mathbf{i}_k,\mathbf{j}_k}|^{\nu_k} \le C(|\eta_{\mathbf{i}_k,\mathbf{j}_k}|^{\nu_k} + |\eta_{\mathbf{i}_k}|^{\nu_k} + |\eta_{\mathbf{j}_k}|^{\nu_k}) \le C(|K_{\mathbf{i}_k,\mathbf{j}_k}|^{\nu_k} + Ch_{\mathbf{n}}^d)$, for $k = \ell, \cdots, s$, where $K_{\mathbf{i},\mathbf{j}} = K(\frac{X_{\mathbf{i}}-X_{\mathbf{j}}}{h})$. Then by Lemma B.2, it follows that, as $\mathbf{n} \to \infty$,

$$E\prod_{k=\ell}^s |\xi_{\mathbf{i}_k,\mathbf{j}_k}|^{\nu_k} \le E\prod_{k=\ell}^s C(|K_{\mathbf{i}_k,\mathbf{j}_k}|^{\nu_k} + Ch_{\mathbf{n}}^d) = O(h_{\mathbf{n}}^{d(s-\ell+1)}). \tag{B.39}$$

Similarly, for $\delta > 0$ with $h_{\mathbf{i}} = h/f^\delta(X_{\mathbf{i}})$, by Assumptions (K1), (D1), (W) and (D3), it easily follows that $\int\left(K(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}})\right)w(x)f(x)dx = \int(K(y))\,w(X_{\mathbf{i}}+h_{\mathbf{i}}y)f(X_{\mathbf{i}}+h_{\mathbf{i}}y)h_{\mathbf{i}}^d dy = O(h^d)$ owing to $h_{\mathbf{i}} \le h/c_w$ for $X_{\mathbf{i}} \in S_w$ by assumption (D3). Thus

$$|\eta_{\mathbf{i}}| \le \int\left(K(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}}) + EK(\frac{x-X_{\mathbf{i}}}{h_{\mathbf{i}}})\right)w(x)f(x)dx \le Ch_{\mathbf{n}}^d,$$

and $|\eta_{\mathbf{j}}| \le Ch_{\mathbf{n}}^d$. With $K_{\mathbf{i},\mathbf{j}}^w = K(\frac{X_{\mathbf{i}}-X_{\mathbf{j}}}{h_{\mathbf{i}}})w(X_{\mathbf{j}})$ replacing $K_{\mathbf{i},\mathbf{j}}$ above, it similarly follows that (B.39) holds true.

Thus by Lemma B.3 together with (B.39) we have

$$E\left(\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\, i\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\right)^{2r} \le C_1\tilde{\mathbf{n}}^2h^d + C_2(\,\tilde{n}^2h^d)^r$$
$$+ C_3(\,\tilde{n}^2h^d)^r\left(P^{2Nr}h^{rd} + h^{(\frac{2}{q}-1)rd}\sum_{t=P+1}^\infty t^{2Nr-1}\varphi(t)^{1-\frac{2}{q}}\right) = O(\,\tilde{n}^2h^d)^r, \tag{B.40}$$

where $P = \lfloor h^{(\frac{2}{q}-2)rd/[\mu(1-2/q)]} \rfloor$ for $\mu > 2Nr(2-2/q)/(1-2/q)$ with $q > 2$, under which $P^{2Nr}h^{rd} = O(1)$ and $h^{(\frac{2}{q}-1)rd}\sum_{t=P+1}^{\infty} t^{2Nr-1}\varphi(t)^{1-\frac{2}{q}} = O(1)$ with $\varphi(t) = O(t^{-\mu})$ in assumption (M).

To show (B.38), by Chebyschev inequality, Lemma B.4 and (B.40), we have

$$\begin{aligned}
&P\left(\tilde{\mathbf{n}}^{-2}MISE^{-1}(h)\left|\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}(h^{-d}\xi_{\mathbf{i},\mathbf{j}})\right| > \varepsilon b_{\mathbf{n}}\right) \\
&\leq (\varepsilon b_{\mathbf{n}})^{-2r}\tilde{\mathbf{n}}^{-4r}h^{-2dr}(MISE(h))^{-2r}E\left(\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\right)^{2r} \leq C(\varepsilon b_{\mathbf{n}})^{-2r}\tilde{\mathbf{n}}^{-2r}E\left|\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\right|^{2r} \\
&\leq C(\varepsilon b_{\mathbf{n}})^{-2r}\tilde{\mathbf{n}}^{-2r}\left[(\tilde{\mathbf{n}}^2h^d)^r\right] \leq C(\varepsilon b_{\mathbf{n}})^{-2r}h^{dr} \to 0 \qquad \text{(B.41)}
\end{aligned}$$

as $\varepsilon \to \infty$, with $b_{\mathbf{n}} = O(h^{d/2})$, i.e., $b_{\mathbf{n}} = \tilde{\mathbf{n}}^{-\frac{d}{2(2r+d)}}$ under assumption (H). Then, (B.35) follows from (B.41).

Verification of (B.36): When $\delta = 0$, for $x \in R^d$, write

$$\begin{aligned}
B_{\mathbf{n}}(x) &:= E\hat{f}_{\mathbf{n}}(x) - f(x) = \int_{\mathbb{R}^d} K(y)[f(x+yh) - f(x)]dy \\
&= \frac{1}{r!}h^r\sum_{i=1}^{d}\int_{\mathbb{R}^d} K(y)[f_{i^r}^{(r)}(x+\zeta hy)y_i^r]dy := h^r B_{\mathbf{n}}^*(x), \qquad \text{(B.42)}
\end{aligned}$$

by application of Taylor's expansion together with assumptions (K1), (K2) and (D1), where $f_{i^r}^{(r)}(x) = f_{i_1=i,\cdots,i_r=i}^{(r)}(x)$ and $f_{i_1,\cdots,i_j}^{(j)}(x)$ is the j-th order partial differentiation of $f(x)$ w.r.t. $(x_{i_1},\cdots,x_{i_j})$, with x_i the i-th component of $x \in \mathbb{R}^d$ and $|\zeta| < 1$.

Similarly, when $\delta > 0$, it follows from (B.30) that

$$\begin{aligned}
B_{\mathbf{n}}(x) &:= E\check{f}_{\mathbf{n}}(x) - f(x) \\
&= \frac{1}{h^d}\int_{\mathbb{R}^d} f^{d\delta}(y)K(\frac{y-x}{h}f^{\delta}(y))f(y)dy - f(x) := h^r B_{\mathbf{n}}^*(x),
\end{aligned}$$

which is still of a similar form to (B.42).

Then by (B.42) note

$$\begin{aligned}
W_{\mathbf{j}} &= B_{\mathbf{n}}(X_{\mathbf{j}})w(X_{\mathbf{j}}) - \int_{\mathbb{R}^d} B_{\mathbf{n}}(x)f(x)w(x)dx \\
&= h^r[B_{\mathbf{n}}^*(X_{\mathbf{j}})w(X_{\mathbf{j}}) - \int_{\mathbb{R}^d} B_{\mathbf{n}}^*(x)f(x)w(x)dx] \qquad \text{(B.43)}
\end{aligned}$$

with $EW_{\mathbf{j}} = 0$ and $|W_{\mathbf{j}}| \leq C$ by the assumptions (K1), (D1), (W). Similar to the proof of Lemma B.3, it is easy to show that $E(\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{j}})^2 = O(\tilde{\mathbf{n}}h^{2r})$ as $\mathbf{n} \to \infty$. Then

$$\begin{aligned}
&P(\mathbf{n}^{-1}MISE^{-1}(h)|\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{j}}| > \varepsilon b_{\mathbf{n}}) \leq (\varepsilon b_{\mathbf{n}})^{-2}(\tilde{\mathbf{n}}MISE(h))^{-2}E(\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{j}})^2 \\
&\leq (\varepsilon b_{\mathbf{n}})^{-2}h^{2d}O(\tilde{\mathbf{n}}h^{2r}) \to 0 \qquad \text{(B.44)}
\end{aligned}$$

as $\varepsilon \to \infty$, with $b_{\mathbf{n}} = O(h^{d/2}) = \tilde{\mathbf{n}}^{-\frac{d}{2(2r+d)}}$ under assumption (H). Thus, (B.36) follows from (B.44).

Finally, we can, by (3.8), have (3.9) converging to zero in probability uniformly with respect to $h \in \mathcal{H}_{\mathbf{n}}$ by a similar argument to the proof of Theorem 3.2 (as shown for (B.46)–(B.48)) below, the detail of which is omitted here.

Proof of Theorem 3.2. As seen in the proof of Theorem 3.1 above, the proof in the case of $\delta > 0$ is similar to that for $\delta = 0$ under the additional assumptions (K4) and (D3). To save space, only the proof for $\delta = 0$ in this theorem is provided below. Theorem 3.2 corresponds to Lemma B of Marron (1985, page 1018) under independent data. We only sketch the proof of this theorem with the difference of spatial dependence highlighted.

Let $\delta_h^*(x, X_{\mathbf{j}}) = K(\frac{x-X_{\mathbf{j}}}{h}) - h^d f(x)$ (we shall define $\delta_h^*(x, X_{\mathbf{j}}) = K(\frac{x-X_{\mathbf{j}}}{h} f^{\delta}(X_{\mathbf{j}})) f^{d\delta}(X_{\mathbf{j}}) - h^d f(x)$ in the case of $\delta > 0$). Then

$$\hat{f}_{\mathbf{n}}(x) - f(x) = \frac{1}{\tilde{\mathbf{n}} h^d} \sum_{\mathbf{j} \in \mathcal{I}_{\mathbf{n}}} \delta_h^*(x, X_{\mathbf{j}}) := \hat{f}_{\mathbf{n}}^*, \tag{B.45}$$

and $MISE(h) = d_M(\hat{f}_{\mathbf{n}}^*, 0)(h)$, $ISE(h) = d_I(\hat{f}_{\mathbf{n}}^*, 0)(h)$. First, notice the decomposition

$$\begin{aligned} ISE(h) - MISE(h) &= \int_{\mathbb{R}^d} \{\hat{f}_{\mathbf{n}}^*\}^2 w(x)dx - E \int_{\mathbb{R}^d} \{\hat{f}_{\mathbf{n}}^*\}^2 w(x)dx \\ &= \int_{\mathbb{R}^d} (\hat{f}_{\mathbf{n}}^* - E\hat{f}_{\mathbf{n}}^*)^2 w(x)dx + 2 \int_{\mathbb{R}^d} (\hat{f}_{\mathbf{n}}^* - E\hat{f}_{\mathbf{n}}^*) E\hat{f}_{\mathbf{n}}^* w(x)dx \\ &\quad - \int_{\mathbb{R}^d} E(\hat{f}_{\mathbf{n}}^* - E\hat{f}_{\mathbf{n}}^*)^2 w(x)dx = T_1 + T_2 + 2T_3 - T_4 - T_5, \end{aligned}$$

where

$$\begin{aligned} T_1 &= (\tilde{\mathbf{n}} h^d)^{-2} \int_{\mathbb{R}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} [\delta_h^*(x, X_{\mathbf{i}}) - E\delta_h^*(x, X_{\mathbf{i}})]^2 w(x)dx, \\ T_2 &= (\tilde{\mathbf{n}} h^d)^{-2} \int_{\mathbb{R}^d} \sum_{\mathbf{i},\mathbf{j} \in \mathcal{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{j}} [\delta_h^*(x, X_{\mathbf{i}}) - E\delta_h^*(x, X_{\mathbf{i}})][\delta_h^*(x, X_{\mathbf{j}}) - E\delta_h^*(x, X_{\mathbf{j}})] w(x)dx, \\ T_3 &= \int_{\mathbb{R}^d} (\hat{f}_{\mathbf{n}}^* - E\hat{f}_{\mathbf{n}}^*) E\hat{f}_{\mathbf{n}}^* w(x)dx = \frac{1}{\tilde{\mathbf{n}} h^d} \sum_{j \in \mathcal{I}_{\mathbf{n}}} \int_{\mathbb{R}^d} R_{\mathbf{j}}^*(x)(E\hat{f}_{\mathbf{n}}(x) - f(x)) w(x)dx, \\ T_4 &= (\ \tilde{n} h^d)^{-2} \int_{\mathbb{R}_d} E \sum_{\mathbf{i} \in I_{\mathbf{n}}} [\delta_h^*(x, X_{\mathbf{i}}) - E\delta_h^*(x, X_{\mathbf{i}})]^2 w(x)dx, \\ T_5 &= (\tilde{\mathbf{n}} h^d)^{-2} \int_{\mathbb{R}^d} E \sum_{\mathbf{i},\mathbf{j} \in \mathcal{I}_{\mathbf{n}}, \mathbf{i} \neq \mathbf{j}} [\delta_h^*(x, X_{\mathbf{i}}) - E\delta_h^*(x, X_{\mathbf{j}})] w(x)dx. \end{aligned}$$

We will show that: as $\mathbf{n} \to \infty$,

$$\sup_{h \in \mathcal{H}_{\mathbf{n}}} \left| \frac{T_1 - T_4}{MISE(h)} \right| \to 0 \quad in \quad probability, \tag{B.46}$$

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}}|\frac{T_3}{MISE(h)}|\to 0 \quad in \quad probability, \tag{B.47}$$

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}}|\frac{T_2-T_5}{MISE(h)}|\to 0 \quad in \quad probability. \tag{B.48}$$

Similarly to Marron (1985, section 8, page 1019) and Marron and Härdle (1986), by the Hölder continuity of K and f under assumptions (K1) and (D1), note that it is straightforward to extend to the supremum over $\mathcal{H}_{\mathbf{n}}$ in (B.46) - (B.48) if they hold for the supremum taken over any finite set ($\subset \mathcal{H}_{\mathbf{n}}$) of h whose cardinality increases algebraically fast with the sample size, the proof of which is outlined. For this, we can take the finite set $\mathfrak{B}_{\mathbf{n}} \subset \mathcal{H}_{\mathbf{n}}$ consisting of the end points of the sub-intervals by partitioning $\mathcal{H}_{\mathbf{n}}$ into $\lfloor \tilde{\mathbf{n}}^{\Xi} \rfloor$ sub-intervals for some $\Xi > 0$, with $\mathrm{Card}(\mathfrak{B}_{\mathbf{n}}) = O(\tilde{\mathbf{n}}^{\Xi})$.

Now, for (B.46), it suffices to show that

$$\tilde{\mathbf{n}}^{\Xi}\sup_{h\in\mathfrak{B}_{\mathbf{n}}} MISE(h)^{-2r}E|T_1-T_4|^{2r}\to 0. \tag{B.49}$$

Note that

$$T_1 - T_4 = (\tilde{\mathbf{n}}h^d)^{-2}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Z^*_{\mathbf{i}}, \tag{B.50}$$

where $Z^*_{\mathbf{i}} = \int_{\mathbb{R}^d} Y_{\mathbf{i}}(x)w(x)dx$, with $Y_{\mathbf{i}}(x) = (\delta^*_h(x, X_{\mathbf{i}}) - E\delta^*_h(x, X_{\mathbf{i}}))^2 - E(\delta^*_h(x, X_{\mathbf{i}}) - E\delta^*_h(x, X_{\mathbf{i}}))^2$. Obviously, $EZ^*_{\mathbf{i}} = 0$, and by the assumptions (K1), (D1) and (W), $|Y_{\mathbf{i}}(x)| \le K^2((X_{\mathbf{i}} - x)/h) + Ch^d_{\mathbf{n}}$ and hence $|Z^*_{\mathbf{i}}| \le Ch^d_{\mathbf{n}}$. We can deal with the spatial dependence of $Z^*_{\mathbf{i}}$ by using the similar techniques in the proof of Lemma B.3 above to show that

$$\begin{aligned} E(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Z^*_{\mathbf{i}})^{2r} &\le C_1\tilde{\mathbf{n}}h^d + C_2(\tilde{\mathbf{n}}h^d)^r \\ &\quad + C_3(\tilde{\mathbf{n}}h^d)^r\left(P^{Nr}h^{rd} + h^{(\frac{2}{q}-1)rd}\sum_{t=P+1}^{\infty} t^{Nr-1}\varphi(t)^{1-\frac{2}{q}}\right) = O((\tilde{\mathbf{n}}h^d)^r), \end{aligned}$$

where the last equality holds as shown in (B.40). Thus, by (B.50), it follows that

$$\begin{aligned} \tilde{\mathbf{n}}^{\Xi}|MISE(h)|^{-2r}E|T_1-T_4|^{2r} &\le C\tilde{\mathbf{n}}^{\Xi}(\tilde{\mathbf{n}}h^d)^{-2r}E(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Z^*_i)^{2r} \\ &\le \tilde{\mathbf{n}}^{\Xi}(\tilde{\mathbf{n}}h^d)^{-2r}(\tilde{\mathbf{n}}h^d)^r = \tilde{\mathbf{n}}^{\Xi}(\tilde{\mathbf{n}}h^d)^{-r} = \tilde{\mathbf{n}}^{\Xi}(\tilde{\mathbf{n}})^{-2r^2/(2r+d)} \to 0 \end{aligned} \tag{B.51}$$

as $\mathbf{n}\to\infty$ when $0 < \Xi < \frac{2r^2}{2r+d}$ under assumption (H), and hence (B.49) holds.

Second, for (B.47), it suffices to show that

$$\tilde{\mathbf{n}}^{\Xi}\sup_{h\in\mathfrak{B}_{\mathbf{n}}} MISE(h)^{-2}E|T_3|^{2}\to 0. \tag{B.52}$$

In fact, writing $T^*_{\mathbf{j}} := \int_{\mathbb{R}^d} R^*_{\mathbf{j}}(x)B_{\mathbf{n}}(x)w(x)dx$, with $R^*_{\mathbf{j}}(x) = \delta^*_h(x, X_{\mathbf{j}}) - E\delta^*_h(x, X_{\mathbf{j}})$ and $B_{\mathbf{n}}(x) = E\hat{f}_{\mathbf{n}}(x) - f(x) = h^r B^*_{\mathbf{n}}(x)$ (c.f. (B.42)), we have $T_3 = \frac{1}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} T^*_{\mathbf{j}} =$

$\frac{1}{\tilde{\mathbf{n}}h^d}h^r\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n}}T_\mathbf{j}^{**}$, where $T_\mathbf{j}^{**}:=\int_{\mathbb{R}^d}R_\mathbf{j}^*(x)B_\mathbf{n}^*(x)w(x)dx$. Note that $ER_\mathbf{j}^*(x)=0$. Thus, $ET_\mathbf{j}^{**}=0$ and $|T_\mathbf{j}^{**}|\le C\int_{\mathbb{R}^d}|R_\mathbf{j}^*(x)|w(x)dx\le Ch^d$. Further, $E\left(\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n}}T_\mathbf{j}^{**}\right)^2=T_{3,1}+T_{3,2}$, where $T_{3,1}=\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n}}E(T_\mathbf{j}^{**})^2$ and $T_{3,2}=\sum_{\mathbf{i}\neq\mathbf{j}\in\mathcal{I}_\mathbf{n}}\mathrm{cov}(T_\mathbf{i}^{**},T_\mathbf{j}^{**})$. Note that $E(T_\mathbf{j}^{**})^2=\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}ER_\mathbf{j}^*(x)R_\mathbf{j}^*(y)B_\mathbf{n}^*(x)B_\mathbf{n}^*(y)w(x)w(y)dxdy=O(h^{2d})$, and hence $T_{3,1}=O(\tilde{\mathbf{n}}h^{2d})$. Extending the derivation for the cross terms in the proof of Lemma B.3 and Gao et al. (2008), we can have $T_{3,2}=O(\tilde{\mathbf{n}}h^{2d})$. It then follows that

$$\begin{aligned}\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}MISE(h)^{-2}E|T_3|^2&\le\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}h^{2r}E\left(\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n}}T_\mathbf{j}^{**}\right)^2\\&\le C\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}h^{2r}(\tilde{\mathbf{n}}h^{2d})\le C\tilde{\mathbf{n}}^\Xi(\tilde{\mathbf{n}})^{-d/(2r+d)}\to0\end{aligned}\tag{B.53}$$

as $\mathbf{n}\to\infty$ when $0<\Xi<\frac{d}{2r+d}$ under assumption (H), and hence (B.52) holds.

Finally, for (B.48), it suffices to show that

$$\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}MISE(h)^{-2}E|T_2-T_5|^2\to0.\tag{B.54}$$

Denote

$$\begin{aligned}S_{\mathbf{i},\mathbf{j}}&:=\int_{\mathbb{R}^d}(\delta_h^*(x,X_\mathbf{i})-E\delta_h^*(x,X_\mathbf{i}))(\delta_h^*(x,X_\mathbf{j})-E\delta_h^*(x,X_\mathbf{j}))w(x)dx\\&=\int_{\mathbb{R}^d}K_\mathbf{i}(x)K_j(x)w(x)dx,\end{aligned}$$

where $K_\mathbf{i}(x):=K(\frac{x-X_\mathbf{i}}{h})-EK(\frac{x-X_\mathbf{i}}{h})$. Then we have

$$T_2-T_5=(\tilde{\mathbf{n}}h^d)^{-2}\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}}(S_{\mathbf{i},\mathbf{j}}-ES_{\mathbf{i},\mathbf{j}}).\tag{B.55}$$

Set $\xi_{\mathbf{i},\mathbf{j}}=S_{\mathbf{i},\mathbf{j}}-ES_{\mathbf{i},\mathbf{j}}$. Then, $E\left(\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\right)^2=S_1^*+S_2^*$, where $S_1^*=\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}}E\xi_{\mathbf{i},\mathbf{j}}^2$ and $S_2^*=\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}}\sum_{\mathbf{i}',\mathbf{j}'\in\mathcal{I}_\mathbf{n},\mathbf{i}'\neq\mathbf{j}',(\mathbf{i},\mathbf{j})\neq(\mathbf{i}',\mathbf{j}')}E\xi_{\mathbf{i},\mathbf{j}}\xi_{\mathbf{i}',\mathbf{j}'}$. Note that $E\xi_{\mathbf{i},\mathbf{j}}^2\le C(ES_{\mathbf{i},\mathbf{j}}^2+(ES_{\mathbf{i},\mathbf{j}})^2)$, $ES_{\mathbf{i},\mathbf{j}}=O(h^{2d})$ and $ES_{\mathbf{i},\mathbf{j}}^2\le C(\int\int\int\int K((x-u)/h)K((x-v)/h)K((y-u)/h)K((y-v)/h)f_{\mathbf{i},\mathbf{j}}(u,v)w(x)w(y)dudvdxdy+O(h^{4d}))=O(h^{4d})$. Hence $E\xi_{\mathbf{i},\mathbf{j}}^2=O(h^{4d})$ and $S_1^*=O(\tilde{\mathbf{n}}^2h^{4d})$. Similarly to the proof of Lemma B.3, it can be shown that $S_2^*=O(\tilde{\mathbf{n}}^2h^{4d})$. Thus $E\left(\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}}\right)^2=O(\tilde{\mathbf{n}}^2h^{4d})$. Thus it follows together with (B.55) that

$$\begin{aligned}\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}MISE(h)^{-2}E|T_2-T_5|^2&\le\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}(\tilde{\mathbf{n}}h^d)^{-2}O(\tilde{\mathbf{n}}^2h^{4d})\\&\le C\tilde{\mathbf{n}}^\Xi\sup_{h\in\mathfrak{B}_\mathbf{n}}h^{2d}\le C\tilde{\mathbf{n}}^\Xi\tilde{\mathbf{n}}^{-2d/(2r+d)}\to0\end{aligned}$$

as $\mathbf{n}\to\infty$, when $0<\Xi<2d/(2r+d)$, under assumption (H). Therefore (B.54) holds.

Thus, (B.49), (B.52) and (B.54) hold provided $0 < \Xi < \min\{\frac{d}{2r+d}, \frac{2r^2}{2r+d}\}$. This complete the proof of the theorem.

Proof of Theorem 3.3 As seen in the proof of Theorem 3.1 above, the proof in the case of $\delta > 0$ is similar to that for $\delta = 0$ under the additional assumptions (K4) and (D3). To save space, only the proof for $\delta = 0$ in this theorem is provided below, where we note that $\check{f}_{\mathbf{n}}(x) = \hat{f}_{\mathbf{n}}(x)$.

The proof is easily done by applying Theorems 3.1 and 3.2. We just prove the result for $d(\hat{f}_{\mathbf{n}}, f)(h) = d_I(\hat{f}_{\mathbf{n}}, f)(h)$ (we shall replace $\hat{f}_{\mathbf{n}}$ by $\check{f}_{\mathbf{n}}$ with $h_0 = h$ as above in the case of $\delta > 0$). Let $h^* = \arg\inf_{h\in\mathcal{H}_{\mathbf{n}}} d_I(\hat{f}_{\mathbf{n}}, f)(h)$, then it is sufficient to show that

$$\frac{|d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h}) - d_I(\hat{f}_{\mathbf{n}}, f)(h^*)|}{d_I(\hat{f}_{\mathbf{n}}, f)(h^*)} \to 0 \quad in \quad probability. \tag{B.56}$$

Since $d_I(\hat{f}_{\mathbf{n}}, f)(h) \geq d_I(\hat{f}_{\mathbf{n}}, f)(h^*)$ and $CV_\delta(h^*) \geq CV_\delta(\hat{h})$, then (B.56) will follow from

$$\frac{|d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h}) - d_I(\hat{f}_{\mathbf{n}}, f)(h^*) + CV_\delta(h^*) - CV_\delta(\hat{h})|}{d_I(\hat{f}_{\mathbf{n}}, f)(h^*)} \to 0 \quad in \quad probability, \tag{B.57}$$

which, in turn, is implied by

$$\frac{|d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h}) - CV_\delta(\hat{h}) - T - (d_I(\hat{f}_{\mathbf{n}}, f)(h^*) - CV_\delta(h^*) - T)|}{d_I(\hat{f}_{\mathbf{n}}, f)(h^*)} \to 0 \quad in \quad probability. \tag{B.58}$$

Note that

$$\frac{CV_\delta(\hat{h}) + T - d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h})}{d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h})} \leq \frac{CV_\delta(\hat{h}) + T - d_I(\hat{f}_{\mathbf{n}}, f)(\hat{h})}{d_I(\hat{f}_{\mathbf{n}}, f)(h^*)} \leq \frac{CV_\delta(h^*) + T - d_I(\hat{f}_{\mathbf{n}}, f)(h^*)}{d_I(\hat{f}_{\mathbf{n}}, f)(h^*)}.$$

Thus the relation (B.58) follows directly from (3.9) and (3.10). Hence, the proof is completed.

Proof of Theorem 3.4 The proof of this theorem can be done similarly to that of Theorem 3.3 with details checked. We only give the proof for the case of distance d_I, which can be verified by showing that $\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|d_I(\check{\hat{f}}_{\mathbf{n}}, f)(h) - d_I(\check{f}_{\mathbf{n}}, f)(h)|}{MISE(h)} = o_P(1)$ and $\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|\check{CV}_\delta(h) - CV_\delta(h)|}{MISE(h)} = o_P(1)$ as $\mathbf{n} \to \infty$, together with Theorems 3.1 and 3.2. Here $\check{\hat{f}}_{\mathbf{n}}$ and $\check{CV}_\delta(h)$ with $h_0 = h$ are defined in Section 3.2 and $MISE(h)$ is in (3.6) for $\check{f}_{\mathbf{n}}$.

Here notice that the only difference between this proof and that of Theorem 3.3 is to replace $h_{\mathbf{i}}^{-d} K\left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{i}}}\right)$ in AKDE with $h_{\mathbf{i}} = h(f(X_{\mathbf{i}}))^{-\delta}$ in Theorem 3.3 by $\hat{h}_{\mathbf{i}}^{-d} K\left(\frac{x - X_{\mathbf{i}}}{\hat{h}_{\mathbf{i}}}\right)$ in ADKE for this theorem with $\hat{h}_{\mathbf{i}} = h(\hat{f}(X_{\mathbf{i}}))^{-\delta}$ and $h_0 = h$ (without loss of generality). With this in mind, note from (2.1) and (3.13) that, by Taylor's expansion together with

assumption (K4),

$$\begin{aligned}
\breve{\hat{f}}_{\mathbf{n}}(x) - \check{f}_{\mathbf{n}}(x) &= \frac{1}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}[\hat{f}^{d\delta}(X_{\mathbf{i}})K(\frac{X_{\mathbf{i}}-x}{h}\hat{f}^{\delta}(X_{\mathbf{i}})) - f^{d\delta}(X_{\mathbf{i}})K(\frac{X_{\mathbf{i}}-x}{h}f^{\delta}(X_{\mathbf{i}}))] \\
&= \frac{1}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}[J^*_{\frac{X_{\mathbf{i}}-x}{h}}(\hat{f}(X_{\mathbf{i}})) - J^*_{\frac{X_{\mathbf{i}}-x}{h}}(f(X_{\mathbf{i}}))] \\
&= \frac{1}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}[J^*_{\frac{X_{\mathbf{i}}-x}{h}}{}^{(1)}(f(X_{\mathbf{i}}) + t(\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}})))\ (\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))], \qquad \text{(B.59)}
\end{aligned}$$

where $|t| < 1$, $J^*_u(y) = y^{d\delta}K(uy^{\delta})$, and

$$J^{*(1)}_u(y) = d\delta y^{d\delta-1}K(uy^{\delta}) + y^{d\delta}\sum_{j=1}^{d}K_j(uy^{\delta})u_j\delta y^{\delta-1} = \delta y^{d\delta-1}K^*(uy^{\delta}), \qquad \text{(B.60)}$$

with $K_j(x)$ being the partial derivative of $K(x)$ with respect to the j-th component x_j of $x \in \mathbb{R}^d$, and $K^*(x) = d\ K(x) + \sum_{j=1}^{d}K_j(x)x_j$. Let $D_{\mathbf{i}}(u;g) = J^{*(1)}_u(f(X_{\mathbf{i}}) + t(g(X_{\mathbf{i}}) - f(X_{\mathbf{i}})))\ (g(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))$. Thus, by (B.59) together with (K4) and the assumption that $\sup_{x\in S_w}|\hat{f}(x) - f(x)| \to 0$ as $\mathbf{n}\to\infty$,

$$\begin{aligned}
&\int(\breve{\hat{f}}_{\mathbf{n}}(x) - \check{f}_{\mathbf{n}}(x))^2 w(x)dx = \int\left\{\frac{1}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}D_{\mathbf{i}}(\frac{X_{\mathbf{i}}-x}{h};\hat{f})\right\}^2 w(x)dx \\
&= \int(\frac{1}{\tilde{\mathbf{n}}h^d})^2\left\{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}D^2_{\mathbf{i}}(\frac{X_{\mathbf{i}}-x}{h};\hat{f}) + \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}D_{\mathbf{i}}(\frac{X_{\mathbf{i}}-x}{h};\hat{f})D_{\mathbf{j}}(\frac{X_{\mathbf{j}}-x}{h};\hat{f})\right\}w(x)dx \\
&= \int(\frac{1}{\tilde{\mathbf{n}}h^d})^2\left\{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}D^2_{\mathbf{i}}(u;\hat{f})w(X_{\mathbf{i}}-uh) + \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}D_{\mathbf{i}}(u;\hat{f})D_{\mathbf{j}}(\frac{X_{\mathbf{j}}-X_{\mathbf{i}}}{h}+u;\hat{f})w(X_{\mathbf{i}}-uh)\right\}h^d du \\
&= (1+o_P(1))\int(\frac{1}{\tilde{\mathbf{n}}h^d})^2\left\{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}[J^{*(1)}_u(f(X_{\mathbf{i}}))]^2(\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))^2 w(X_{\mathbf{i}})\right. \\
&\quad \left. + \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}J^{*(1)}_u(f(X_{\mathbf{i}}))J^*_{u+X_{\mathbf{j}\mathbf{i}h}}{}^{(1)}(f(X_{\mathbf{j}}))(\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))(\hat{f}(X_{\mathbf{j}}) - f(X_{\mathbf{j}}))w(X_{\mathbf{i}})\right\}h^d du \\
&= (1+o_P(1))\left\{\frac{1}{\tilde{\mathbf{n}}^2h^d}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\int[J^{*(1)}_u(f(X_{\mathbf{i}}))]^2 du\ (\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))^2 w(X_{\mathbf{i}})\right. \\
&\quad \left. + \frac{1}{\tilde{\mathbf{n}}^2h^d}\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}}\int J^{*(1)}_u(f(X_{\mathbf{i}}))J^*_{u+X_{\mathbf{j}\mathbf{i}h}}{}^{(1)}(f(X_{\mathbf{j}}))du\ (\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))(\hat{f}(X_{\mathbf{j}}) - f(X_{\mathbf{j}}))w(X_{\mathbf{i}})\right\} \\
&:= (1+o_P(1))\left\{\mathcal{D}_{\mathbf{n}1} + \mathcal{D}_{\mathbf{n}2}\right\}, \qquad \text{(B.61)}
\end{aligned}$$

where $X_{\mathbf{j}\mathbf{i}h} = \frac{X_{\mathbf{j}}-X_{\mathbf{i}}}{h}$.

By the assumption that $\Upsilon_{\mathbf{n}} := \sup_{x\in S_w} |\hat{f}(x) - f(x)| = o_P(1)$ as $\mathbf{n} \to \infty$, it is obvious by Lemma B.4 that

$$\frac{\mathcal{D}_{\mathbf{n}1}}{MISE(h)} \leq \Upsilon_{\mathbf{n}}^2 \frac{1/(\tilde{\mathbf{n}}h^d)}{MISE(h)} \frac{1}{\tilde{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \int [J_u^{*(1)}(f(X_{\mathbf{i}}))]^2 du \; w(X_{\mathbf{i}}) = o_P(1), \tag{B.62}$$

where $MISE(h)$ is given in Lemma B.4 for $\check{f}_{\mathbf{n}}(x)$, and o_P holds uniformly with respect to $h \in \mathcal{H}_{\mathbf{n}}$ by assumption (H).

Also, by the assumption that $\Upsilon_{\mathbf{n}} := \sup_{x\in S_w} |\hat{f}(x) - f(x)| = o_P(1)$ together with the bounded supports for $K(\cdot)$ and $w(\cdot)$ in assumptions (K1) and (W),

$$\begin{aligned}
\mathcal{D}_{\mathbf{n}2} &= \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \int J_u^{*(1)}(f(X_{\mathbf{i}})) {J^*_{u+X_{\mathbf{ji}h}}}^{(1)}(f(X_{\mathbf{i}} + hX_{\mathbf{ji}h}))du \\
&\qquad \times (\hat{f}(X_{\mathbf{i}}) - f(X_{\mathbf{i}}))(\hat{f}(X_{\mathbf{i}} + hX_{\mathbf{ji}h}) - f(X_{\mathbf{i}} + hX_{\mathbf{ji}h}))w(X_{\mathbf{i}}) \\
&= o_P(1) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \int J_u^{*(1)}(f(X_{\mathbf{i}})) {J^*_{u+X_{\mathbf{ji}h}}}^{(1)}(f(X_{\mathbf{i}} + hX_{\mathbf{ji}h}))du \; w(X_{\mathbf{i}}) \\
&= o_P(1) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \int J_u^{*(1)}(f(X_{\mathbf{i}})) {J^*_{u+X_{\mathbf{ji}h}}}^{(1)}(f(X_{\mathbf{j}}))du \; w(X_{\mathbf{i}}) \\
&= o_P(1) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \int {J^*_{u+X_{\mathbf{i}}/h}}^{(1)}(f(X_{\mathbf{i}})) {J^*_{u+X_{\mathbf{j}}/h}}^{(1)}(f(X_{\mathbf{j}}))du \; w(X_{\mathbf{i}}) \\
&= o_P(1) \left\{ \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} [D_{2\mathbf{ji}} - ED_{2\mathbf{ji}}] + \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} ED_{2\mathbf{ji}} \right\} \\
&:= o_P(1) \left\{ \mathcal{D}_{\mathbf{n}21} + \mathcal{D}_{\mathbf{n}22} \right\},
\end{aligned} \tag{B.63}$$

where $J_u^{*(1)}(y) = \delta y^{d\delta-1} K^*(uy^\delta)$, defined in (B.60), and

$$D_{2\mathbf{ji}} = \int {J^*_{u+X_{\mathbf{i}}/h}}^{(1)}(f(X_{\mathbf{i}})) {J^*_{u+X_{\mathbf{j}}/h}}^{(1)}(f(X_{\mathbf{j}}))du \; w(X_{\mathbf{i}}).$$

Then for $\mathcal{D}_{\mathbf{n}21}$, we can easily apply Lemma B.3 with $\xi_{\mathbf{ji}} = D_{2\mathbf{ji}} - ED_{2\mathbf{ji}}$ to obtain

$$E\mathcal{D}_{\mathbf{n}21}^2 = E\left\{ \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \xi_{\mathbf{ji}} \right\}^2 = \left\{ \frac{1}{\tilde{\mathbf{n}}^2 h^d} \right\}^2 O(\tilde{\mathbf{n}}^2 h^d) = O\left(\frac{1}{\tilde{\mathbf{n}}^2 h^d} \right),$$

and hence

$$\mathcal{D}_{\mathbf{n}21} = O_P\left(\frac{h^{d/2}}{\tilde{\mathbf{n}}h^d} \right). \tag{B.64}$$

Now for $\mathcal{D}_{\mathbf{n}22}$, letting $J_{1\mathbf{i}}(u) = {J^*_{u+X_{\mathbf{i}}/h}}^{(1)}(f(X_{\mathbf{i}}))w(X_{\mathbf{i}})$ and $J_{2\mathbf{j}}(u) = {J^*_{u+X_{\mathbf{j}}/h}}^{(1)}(f(X_{\mathbf{j}}))$, we note that

$$ED_{2\mathbf{ji}} = \int [\mathrm{cov}(J_{1\mathbf{i}}(u), J_{2\mathbf{j}}(u)) + EJ_{1\mathbf{i}}(u) EJ_{2\mathbf{j}}(u)] du.$$

Hence

$$\mathcal{D}_{\mathbf{n}22} = \int \left[\frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}} \mathrm{cov}(J_{1\mathbf{i}}(u), J_{2\mathbf{j}}(u)) + \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{i}\neq\mathbf{j}} EJ_{1\mathbf{i}}(u) EJ_{2\mathbf{j}}(u) \right] du$$
$$:= \int \left[\mathcal{D}_{\mathbf{n}221}(u) + \mathcal{D}_{\mathbf{n}222}(u) \right] du \tag{B.65}$$

It is easy to show that

$$E|J_{1\mathbf{i}}(u)|^q \leq Ch^d (f(-u))^{(d\delta-1)(q-1)} w(-u), \quad E|J_{2\mathbf{j}}(u)|^q \leq Ch^d (f(-u))^{(d\delta-1)(q-1)},$$

and by Lemma B.1, $|\mathrm{cov}(J_{1\mathbf{i}}(u), J_{2\mathbf{j}}(u))| \leq (E|J_{1\mathbf{i}}(u)|^q)^{1/q} (E|J_{2\mathbf{j}}(u)|^q)^{1/q} \varphi^{1-2/q}(\|\mathbf{j}-\mathbf{i}\|)$ for $q > 2$, and $|\mathrm{cov}(J_{1\mathbf{i}}(u), J_{2\mathbf{j}}(u))| \leq (E|J_{1\mathbf{i}}(u)|^2)^{1/2} (E|J_{2\mathbf{j}}(u)|^2)^{1/2}$. Thus

$$\begin{aligned}
|\mathcal{D}_{\mathbf{n}221}(u)| &= (f(-u))^{(d\delta-1)} w^{1/2}(-u) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n}, 0<\|\mathbf{i}-\mathbf{j}\|\leq P} O(h^d) \\
&\quad + (f(-u))^{2(d\delta-1)(q-1)/q} w^{1/q}(-u) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n}, \|\mathbf{i}-\mathbf{j}\|\leq P} O(h^{2d/q}) \varphi^{1-2/q}(\|\mathbf{j}-\mathbf{i}\|) \\
&= (f(-u))^{(d\delta-1)} w^{1/2}(-u) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \tilde{\mathbf{n}} P^N O(h^d) \\
&\quad + (f(-u))^{2(d\delta-1)(q-1)/q} w^{1/q}(-u) \frac{1}{\tilde{\mathbf{n}}^2 h^d} \tilde{\mathbf{n}} \sum_{t=P+1}^{\infty} t^{N-1} \varphi^{1-2/q}(t) O(h^{2d/q}) \\
&= O(\frac{h^{d/2}}{\tilde{\mathbf{n}} h^d}) [w^{1/2}(-u) + w^{1/q}(-u)],
\end{aligned} \tag{B.66}$$

which easily follows by taking $P = h^{-d/(2N)}$ together with assumptions (D1) and (D3). Moreover, by noticing that $\int K^*(x)dx = 0$ and assumptions (W), (D1) and (D3), we have

$$\begin{aligned}
&EJ_{1\mathbf{i}}(u) = EJ^*_{u+X_\mathbf{i}/h}{}^{(1)}(f(X_\mathbf{i}))w(X_\mathbf{i}) \\
&= E\delta f^{d\delta-1}(X_\mathbf{i}) K^*((u+X_\mathbf{i}/h) f^\delta(X_\mathbf{i})) w(X_\mathbf{i}) \\
&= \delta \int f^{d\delta-1}(x) K^*((u+x/h) f^\delta(x)) w(x) f(x) dx \\
&= \delta h^d \int f^{d\delta}(xh-u) K^*(x f^\delta(xh-u)) w(xh-u) dx \\
&= (1+o(1)) w(-u) \delta h^d \int [f^{d\delta}(xh-u) K^*(x f^\delta(xh-u)) - f^{d\delta}(-u) K^*(x f^\delta(-u))] dx \\
&= (1+o(1)) w(-u) \delta h^d \int [J^{**}_x(f(xh-u)) - J^{**}_x(f(-u))] dx = w(-u) O(h^{d+r}),
\end{aligned} \tag{B.67}$$

the last equality of which follows by a similar argument to (B.30), where $J^{**}_x(y) = y^{d\delta} K^*(xy^\delta)$. Similarly to (B.67), $EJ_{2\mathbf{j}}(u) = O(h^{d+r})$. Thus together with (B.64), (B.65)

and (B.66), it follows from Lemma B.4(ii) that $\mathcal{D}_{\mathbf{n}21} + \mathcal{D}_{\mathbf{n}22} = O(\frac{h^{d/2}}{\tilde{\mathbf{n}}h^d}) + O(h^{d+2r}) = O(h^{d/2})MISE(h)$, where $MISE(h)$ is given in Lemma B.4 for $\check{f}_{\mathbf{n}}(x)$. Hence, by (B.63) with a similar argument to that in the proof of Theorem 3.2, we have

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{\mathcal{D}_{\mathbf{n}2}}{MISE(h)} = o_P(1). \tag{B.68}$$

Finally it follows from (B.61), (B.62) and (B.68) that

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{\int(\check{\check{f}}_{\mathbf{n}}(x) - \check{f}_{\mathbf{n}}(x))^2 w(x)dx}{MISE(h)} = o_P(1). \tag{B.69}$$

By (B.69) together with Theorem 3.2, it easily follows that

$$\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|d_I(\check{\check{f}}_{\mathbf{n}}, f)(h) - d_I(\check{f}_{\mathbf{n}}, f)(h)|}{MISE(h)} = o_P(1), \quad \sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|d_I(\check{\check{f}}_{\mathbf{n}}, f)(h) - d_I(\check{f}_{\mathbf{n}}, f)(h)|}{d_I(\check{f}_{\mathbf{n}}, f)(h)} = o_P(1). \tag{B.70}$$

Now we are showing $\sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|\check{CV}_\delta(h) - CV_\delta(h)|}{MISE(h)} = o_P(1)$ as $\mathbf{n}\to\infty$, which, by (B.70), is equivalent to showing

$$\begin{aligned}\mathcal{C}_{\mathbf{n}} &:= \sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|\check{CV}_\delta(h) - CV_\delta(h) - (\check{ISE}(h) - ISE(h))|}{MISE(h)} \\ &= \sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|\check{CV}_\delta(h) - \check{ISE}(h) - T - (CV_\delta(h) - ISE(h) - T)|}{MISE(h)} = o_P(1),\end{aligned} \tag{B.71}$$

where $\check{ISE}(h) = d_I(\check{\check{f}}_{\mathbf{n}}, f)(h)$ and $ISE(h) = d_I(\check{f}_{\mathbf{n}}, f)(h)$. By (B.34), in order to have (B.71), it suffices to show that

$$\mathcal{C}_{\mathbf{n}} = \sup_{h\in\mathcal{H}_{\mathbf{n}}} \frac{|\frac{2}{\tilde{\mathbf{n}}(\tilde{\mathbf{n}}-1)} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} (\check{U}_{\mathbf{i},\mathbf{j}} - U_{\mathbf{i},\mathbf{j}})|}{MISE(h)} = o_P(1), \tag{B.72}$$

where recall we put $\hat{h}_{\mathbf{i}} = h/[\hat{f}^\delta(X_{\mathbf{i}})]$ with $h_0 = h$, and

$$\begin{aligned}\check{U}_{\mathbf{i},\mathbf{j}} := \frac{1}{\hat{h}_{\mathbf{i}}^d} K(\frac{X_{\mathbf{i}} - X_{\mathbf{j}}}{\hat{h}_{\mathbf{i}}}) w(X_{\mathbf{j}}) - \int_{\mathbb{R}^d} \frac{1}{\hat{h}_{\mathbf{i}}^d} K(\frac{X_{\mathbf{i}} - x}{\hat{h}_{\mathbf{i}}}) f(x) w(x) dx - f(X_{\mathbf{j}}) w(X_{\mathbf{j}}) \\ + \int_{\mathbb{R}^d} f(x)^2 w(x) dx.\end{aligned}$$

Recalling the definition of $U_{\mathbf{i},\mathbf{j}}$ in (B.31), it follows that

$$\begin{aligned}&\frac{2}{\tilde{\mathbf{n}}(\,\tilde{n}-1)} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} (\check{U}_{\mathbf{i},\mathbf{j}} - U_{\mathbf{i},\mathbf{j}}) \\ &= \frac{2}{\tilde{\mathbf{n}}(\tilde{\mathbf{n}}-1)h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \left[J^*_{\frac{X_{\mathbf{i}} - X_{\mathbf{j}}}{h}}(\hat{f}^\delta(X_{\mathbf{i}})) - J^*_{\frac{X_{\mathbf{i}} - X_{\mathbf{j}}}{h}}(f^\delta(X_{\mathbf{i}})) \right] w(X_{\mathbf{j}}) \\ &\quad - \frac{2}{\tilde{\mathbf{n}}(\tilde{\mathbf{n}}-1)h^d} \sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_{\mathbf{n}},\mathbf{i}\neq\mathbf{j}} \int \left[J^*_{\frac{X_{\mathbf{i}} - x}{h}}(\hat{f}^\delta(X_{\mathbf{i}})) - J^*_{\frac{X_{\mathbf{i}} - x}{h}}(f^\delta(X_{\mathbf{i}})) \right] w(x) f(x) dx\end{aligned}$$

$$:= \mathcal{C}_{\mathbf{n}1}(h) - \mathcal{C}_{\mathbf{n}2}(h). \tag{B.73}$$

It follows from (B.59) together with (3.1) and the assumption $\sup_{x\in S_w}|\hat{f}(x)-f(x)| = o_P(1)$ that, for some $|\tilde{t}|<1$,

$$\begin{aligned}
\mathcal{C}_{\mathbf{n}1}(h) &= \frac{2}{\tilde{\mathbf{n}}}\frac{1}{(\tilde{\mathbf{n}}-1)h^d}\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}[J^*_{\frac{X_\mathbf{i}-X_\mathbf{j}}{h}}{}^{(1)}(f(X_\mathbf{i})+\tilde{t}(\hat{f}(X_\mathbf{i})-f(X_\mathbf{i})))\ (\hat{f}(X_\mathbf{i})-f(X_\mathbf{i}))]w(X_\mathbf{j}) \\
&= \frac{2}{\tilde{\mathbf{n}}}\frac{1}{(\tilde{\mathbf{n}}-1)h^d}\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}[J^*_{X_{\mathbf{ij}h}}{}^{(1)}(f(X_\mathbf{i})+\tilde{t}(\hat{f}(X_\mathbf{j}+hX_{\mathbf{ij}h})-f(X_\mathbf{j}+hX_{\mathbf{ij}h})) \\
&\qquad\qquad \times(\hat{f}(X_\mathbf{j}+hX_{\mathbf{ij}h})-f(X_\mathbf{j}+hX_{\mathbf{ij}h}))]w(X_\mathbf{j}) \\
&= o_P(1)\frac{2}{\tilde{\mathbf{n}}}\frac{1}{(\tilde{\mathbf{n}}-1)h^d}\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}[J^*_{X_{\mathbf{ij}h}}{}^{(1)}(f(X_\mathbf{i}))]w(X_\mathbf{j}),
\end{aligned} \tag{B.74}$$

and similarly,

$$\begin{aligned}
\mathcal{C}_{\mathbf{n}2}(h) &= \frac{2}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\int[J^*_{\frac{X_\mathbf{i}-x}{h}}{}^{(1)}(f(X_\mathbf{i})+\tilde{t}(\hat{f}(X_\mathbf{i})-f(X_\mathbf{i})))\ (\hat{f}(X_\mathbf{i})-f(X_\mathbf{i}))]w(x)f(x)dx \\
&= o_P(1)\frac{2}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\int[J^*_{\frac{X_\mathbf{i}-x}{h}}{}^{(1)}(f(X_\mathbf{i}))]w(x)f(x)dx.
\end{aligned} \tag{B.75}$$

Then, by (B.74) and (B.75),

$$\begin{aligned}
\mathcal{C}_{\mathbf{n}1}(h)-\mathcal{C}_{\mathbf{n}2}(h) = o_P(1)\frac{2}{\tilde{\mathbf{n}}h^d}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\Bigg\{&\frac{1}{\tilde{\mathbf{n}}-1}\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}[J^*_{X_{\mathbf{ij}h}}{}^{(1)}(f(X_\mathbf{i}))]w(X_\mathbf{j}) \\
&-\int[J^*_{\frac{X_\mathbf{i}-x}{h}}{}^{(1)}(f(X_\mathbf{i}))]w(x)f(x)dx\Bigg\}.
\end{aligned} \tag{B.76}$$

Let $J^*_{\mathbf{ij}} = [J^*_{X_{\mathbf{ij}h}}{}^{(1)}(f(X_\mathbf{i}))]w(X_\mathbf{j})$ and $J^*_\mathbf{i} = \int[J^*_{\frac{X_\mathbf{i}-x}{h}}{}^{(1)}(f(X_\mathbf{i}))]w(x)f(x)dx$. Set $\psi_{K^*}(t) = \int_{\mathbb{R}^d} e^{\iota t'u}K^*(u)du$. Then note that

$$\begin{aligned}
J^*_{\mathbf{ij}} &= \delta f^{d\delta-1}(X_\mathbf{i})K^*\left(\frac{X_\mathbf{i}-X_\mathbf{j}}{h}f^\delta(X_\mathbf{i})\right)w(X_\mathbf{j}) \\
&= \delta f^{d\delta-1}(X_\mathbf{i})\left(\frac{1}{2\pi}\right)^d\int_{\mathbb{R}^d}e^{-\iota t'\frac{X_\mathbf{i}-X_\mathbf{j}}{h}f^\delta(X_\mathbf{i})}\psi_{K^*}(t)dt\ w(X_\mathbf{j}) \\
&= \delta\left(\frac{1}{2\pi}\right)^d\int_{\mathbb{R}^d}f^{-1}(X_\mathbf{i})e^{-\iota t'X_\mathbf{i}/h}\psi_{K^*}(t/f^\delta(X_\mathbf{i}))\ e^{\iota t'X_\mathbf{j}/h}w(X_\mathbf{j})dt \\
&= \delta\left(\frac{1}{2\pi}\right)^d h^d\int_{\mathbb{R}^d}J^*_{1\mathbf{i}}(t)\ J^*_{2\mathbf{j}}(t)dt,
\end{aligned}$$

where $J^*_{1\mathbf{i}}(t) = f^{-1}(X_\mathbf{i})e^{-\iota t' X_\mathbf{i}}\psi_{K^*}(ht/f^\delta(X_\mathbf{i}))$ and $J^*_{2\mathbf{i}}(t) = e^{\iota t' X_\mathbf{j}} w(X_\mathbf{j})$, and similarly

$$J^*_\mathbf{i} = \delta\left(\frac{1}{2\pi}\right)^d h^d \int_{\mathbb{R}^d} J^*_{1\mathbf{i}}(t)\ EJ^*_{2\mathbf{j}}(t)dt.$$

Thus it follows from (B.76) that

$$\mathcal{C}_{\mathbf{n}1}(h) - \mathcal{C}_{\mathbf{n}2}(h) = o_P(1)\frac{2}{\tilde{\mathbf{n}}h^d}\frac{1}{\tilde{\mathbf{n}}-1}\delta\left(\frac{1}{2\pi}\right)^d h^d \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}\left[\int_{\mathbb{R}^d} J^*_{1\mathbf{i}}(t)\ (J^*_{2\mathbf{j}}(t) - EJ^*_{2\mathbf{j}}(t))dt\right]$$

$$= o_P(1)2\delta\left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d}\left[\frac{1}{\tilde{\mathbf{n}}h^d}\frac{1}{\tilde{\mathbf{n}}-1}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}(\xi_{\mathbf{ij}}(t) - E\xi_{\mathbf{ij}}(t) + E\xi_{\mathbf{ij}}(t))\right]dt, \tag{B.77}$$

where $\xi_{\mathbf{ij}}(t) = h^d J^*_{1\mathbf{i}}(t)\ (J^*_{2\mathbf{j}}(t) - EJ^*_{2\mathbf{j}}(t))$. Then, similarly to (B.64), by Lemma B.3, we can show that

$$\mathcal{C}_{\mathbf{n}11}(h) = \int_{\mathbb{R}^d}\left[\frac{1}{\tilde{\mathbf{n}}h^d}\frac{1}{\tilde{\mathbf{n}}-1}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}(\xi_{\mathbf{ij}}(t) - E\xi_{\mathbf{ij}}(t))\right]dt = O_P\left(\frac{h^{d/2}}{\tilde{\mathbf{n}}h^d}\right), \tag{B.78}$$

and, similarly to (B.37), we can show

$$\begin{aligned}
\mathcal{C}_{\mathbf{n}12}(h) &= \int_{\mathbb{R}^d}\left[\frac{1}{\tilde{\mathbf{n}}h^d}\frac{1}{\tilde{\mathbf{n}}-1}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\sum_{\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}E\xi_{\mathbf{ij}}(t)\right]dt \\
&= \frac{1}{\tilde{\mathbf{n}}h^d}\frac{1}{\tilde{\mathbf{n}}-1}h^d\sum_{\mathbf{i},\mathbf{j}\in\mathcal{I}_\mathbf{n},\mathbf{j}\neq\mathbf{i}}\int_{\mathbb{R}^d}\mathrm{cov}(J^*_{1\mathbf{i}}(t), J^*_{2\mathbf{j}}(t))dt \\
&\leq C\frac{1}{\tilde{\mathbf{n}}h^d}\frac{1}{\tilde{\mathbf{n}}-1}h^d\left[\tilde{\mathbf{n}}P^N + \tilde{\mathbf{n}}h^{-d}\sum_{k=P}^{\infty}k^{N-1}\varphi(k)\right] = O_P\left(\frac{h^{d/2}}{\tilde{\mathbf{n}}h^d}\right),
\end{aligned} \tag{B.79}$$

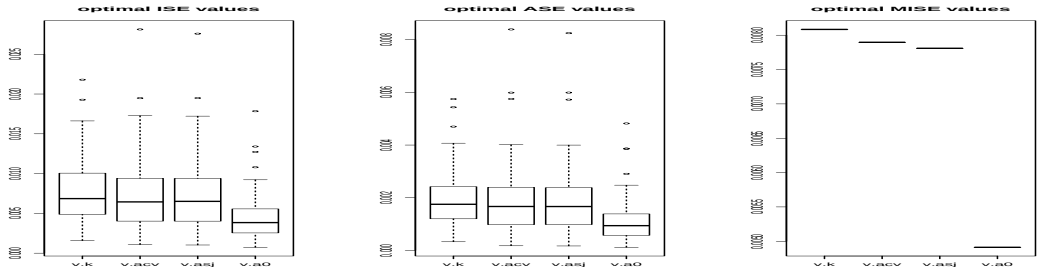
by taking $P^N = O(h^{-d/2})$. It follows from (B.77), (B.74) and (B.75) that $\mathcal{C}_{\mathbf{n}1}(h) - \mathcal{C}_{\mathbf{n}2}(h) = O(h^{d/2})MISE(h)$, where $MISE(h)$ is given in Lemma B.4 for $\check{f}_\mathbf{n}(x)$. Hence, by a similar argument to that in the proof of Theorem 3.2, we have

$$\sup_{h\in\mathcal{H}_\mathbf{n}}\frac{|\mathcal{C}_{\mathbf{n}1}(h) - \mathcal{C}_{\mathbf{n}2}(h)|}{MISE(h)} = o_P(1),$$

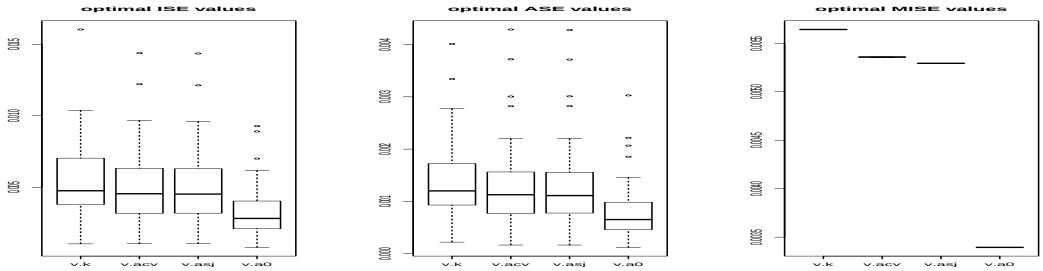
by which, together with (B.73), we can easily see that (B.72) and hence (B.71) hold true. The proof is done.

B3 Supplementary figures and table

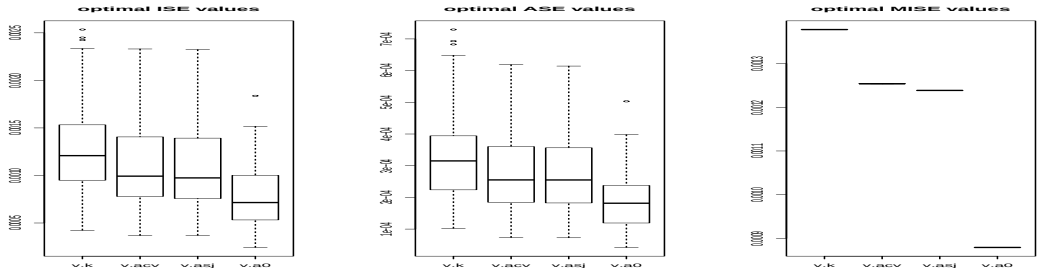
(a) $m_1 = 25,\ m_2 = 10$



(b) $m_1 = 20,\ m_2 = 20$



(c) $m_1 = m_2 = 50$



(d) $m_1 = m_2 = 100$

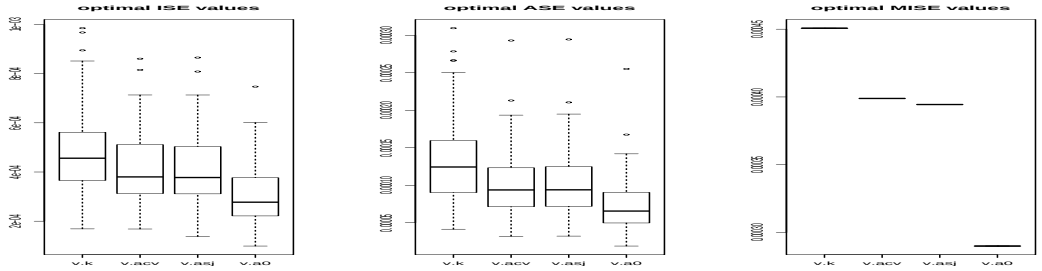
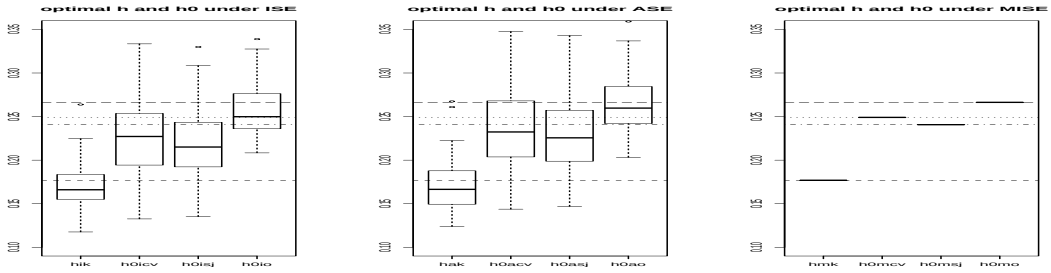
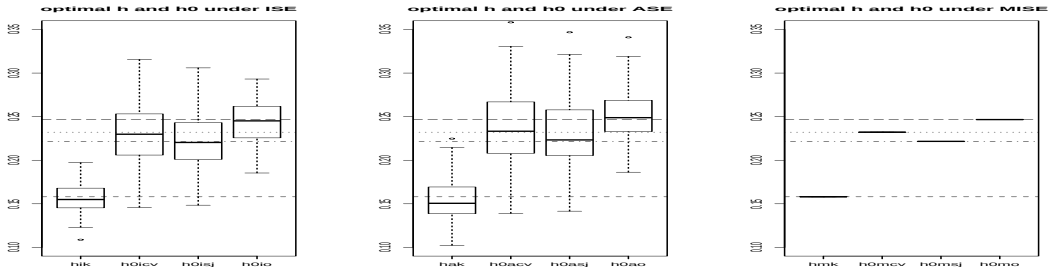


Figure B.1: *Boxplots of optimal ISE, ASE and MISE values with kernel and adaptive kernel density estimates for 100 simulations of different sample sizes of* (m_1, m_2)*: (a)* $(m_1, m_2) = (25, 10)$*, (b)* $(m_1, m_2) = (20, 20)$*, (c)* $(m_1, m_2) = (50, 50)$ *and (d)* $(m_1, m_2) = (100, 100)$.

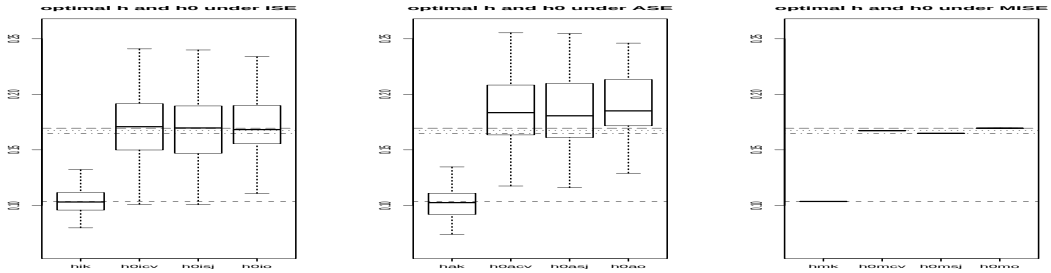
(a) $m_1 = 25,\ m_2 = 10$



(b) $m_1 = 20,\ m_2 = 20$



(c) $m_1 = m_2 = 50$



(d) $m_1 = m_2 = 100$

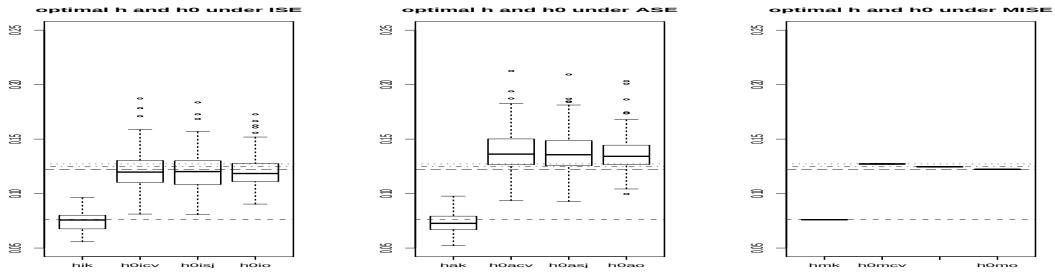
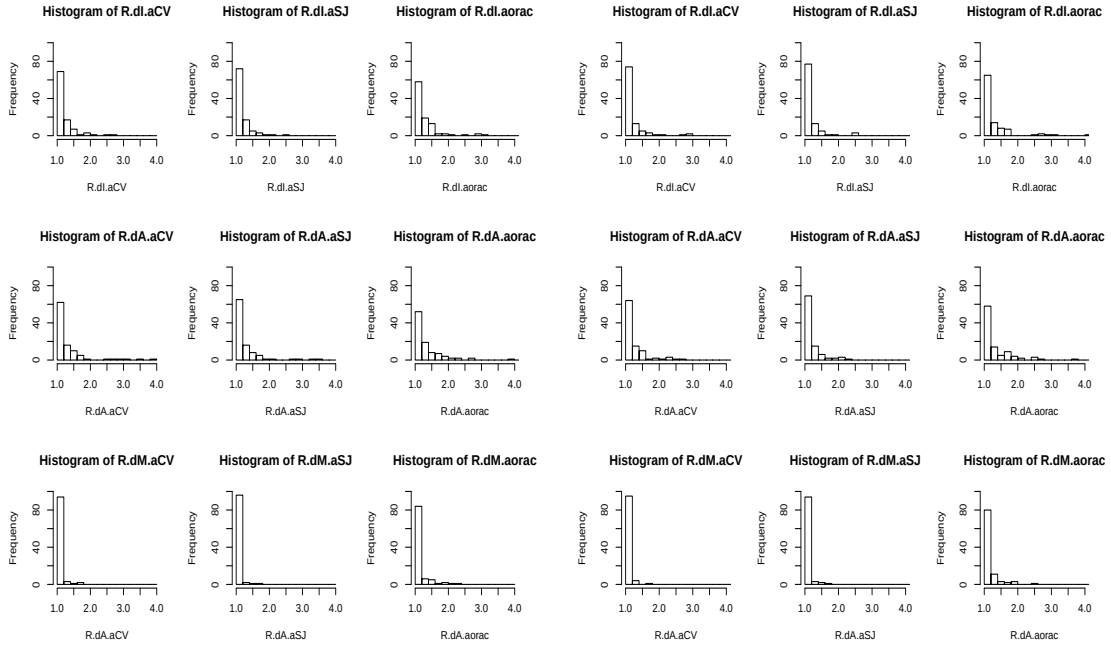


Figure B.2: *Boxplots of optimal bandwidths h or h_0 under ISE, ASE and MISE, respectively, with kernel and adaptive kernel density estimates for 100 simulations of different sample sizes of (m_1, m_2): (a) $(m_1, m_2) = (25, 10)$, (b) $(m_1, m_2) = (20, 20)$, (c) $(m_1, m_2) = (50, 50)$ and (d) $(m_1, m_2) = (100, 100)$. The horizontal lines are for the MISE-optimal bandwidths of h and h_0 for the non-adaptive KDE and the AKDEs with the pilots of CV- and SJ-based KDEs and oracle (true) density, respectively.*

(a) $m_1 = 25, m_2 = 10$ (b) $m_1 = m_2 = 20$



(c) $m_1 = m_2 = 50$ (d) $m_1 = m_2 = 100$

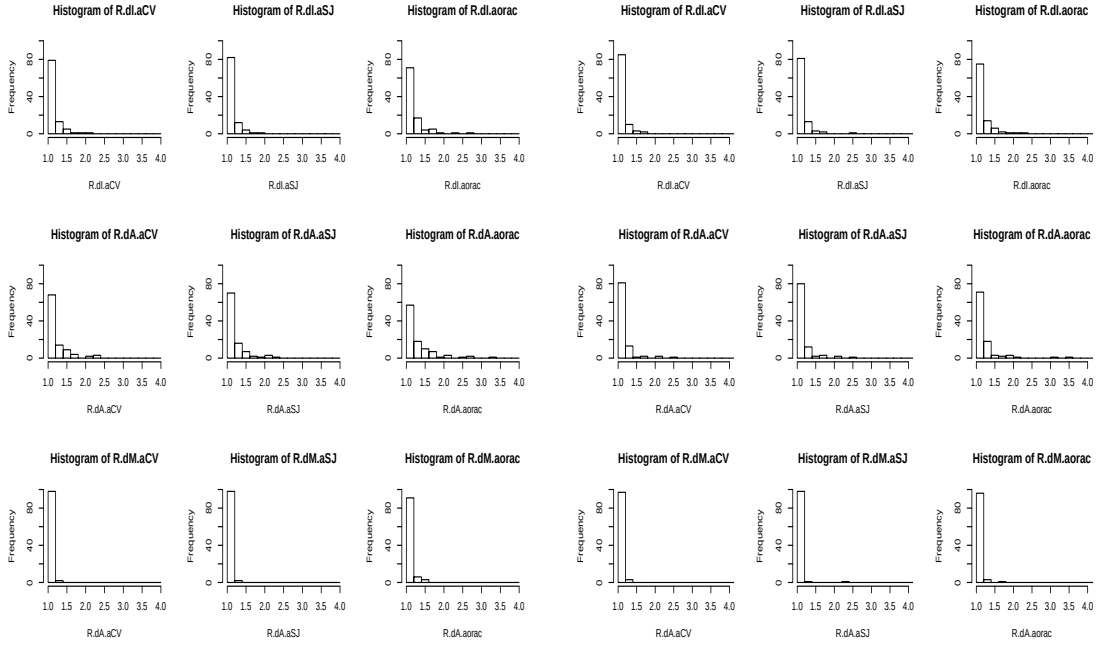


Figure B.3: *Histograms of R_{dI}'s and R_{dA}'s, respectively, for the CV selected h_0 in the adaptive bandwidth of the CV, SJ and oracle pilot estimates with 100 simulations of different sample sizes of (m_1, m_2): (a) $(m_1, m_2) = (25, 10)$, (b) $(m_1, m_2) = (20, 20)$, (c) $(m_1, m_2) = (50, 50)$ and (d) $(m_1, m_2) = (100, 100)$.*

Table B.1: The real elapsed time for bandwidth selection run in R by a Dell Precision 7520 laptop of COREi7. Note: H0.SJ is a shorthand notation for H0.CVSJ, denoting spatial CV for h_0 in the AKDE with a pilot density estimate by a SJ (plug-in) h; H0.CV is a shorthand notation for H0.CVCV, denoting spatial CV for h_0 in the AKDE with a pilot density estimate by a CV h; H.CV denotes a CV for h in the (non-adaptive) kernel density estimate.

Method	sample size	elapsed time (in seconds)
H0.SJ	m1*m2=20*20=400	2.12
	m1*m2=50*50=2500	22.75
	m1*m2=100*100=10000	106.39
H0.CV	m1*m2=20*20=400	2.17
	m1*m2=50*50=2500	15.69
	m1*m2=100*100=10000	108.01
H.CV	m1*m2=20*20=400	2.50
	m1*m2=50*50=2500	10.96
	m1*m2=100*100=10000	50.59