A Weighted k-Nearest Neighbor Density Estimate for Geometric Inference

Gérard Biau[1]
Université Pierre et Marie Curie[2] *& Ecole Normale Supérieure*[3]*, France*
gerard.biau@upmc.fr

Frédéric Chazal
INRIA Saclay – Ile-de-France, France
frederic.chazal@inria.fr

David Cohen-Steiner
INRIA Sophia-Antipolis, France
David.Cohen-Steiner@sophia.inria.fr

Luc Devroye
McGill University[4]*, Canada*
lucdevroye@gmail.com

Carlos Rodríguez
State University of New York at Albany, USA
carlos@math.albany.edu

Abstract

Motivated by a broad range of potential applications in topological and geometric inference, we introduce a weighted version of the k-nearest neighbor density estimate. Various pointwise consistency results of this estimate are established. We present a general central limit theorem under the lightest possible conditions. In addition, a strong approximation result is obtained and the choice of the optimal set of weights is discussed. In particular, the classical k-nearest neighbor estimate is not optimal in a sense described in the manuscript. The proposed method has been implemented to recover level sets in both simulated and real-life data.

Index Terms — Geometric inference, level sets, density estimation, k-nearest neighbor estimate, weighted estimate, consistency, rates of convergence, central limit theorem, strong approximation.

2010 Mathematics Subject Classification: 62G07, 62G05, 62G20.

[1]Corresponding author.

[2]Research partially supported by the French "Agence Nationale pour la Recherche" under grant ANR-09-BLAN-0051-02 "CLARA".

[3]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

[4]Research sponsored by NSERC Grant A3456 and FQRNT Grant 90-ER-0291.

1 Introduction and motivations

The problem of recovering topological and geometric information from multivariate data has attracted increasing interest in recent years.

Taking a statistical point of view, data points are usually considered as independent observations drawn according to a common distribution μ on the space $\mathbb{R}^d$. In this stochastic framework, the problem of estimating the support of μ and its geometric properties (e.g., dimension, number of connected components, volume) has been widely studied during the last two decades (for a review of the literature, see for instance Cuevas and Rodríguez-Casal [15], and Biau, Cadre, and Pelletier [5]). There are set-ups in which sets or boundaries are to be estimated from samples drawn from within and outside the set itself. Various models exist in this respect—it is the point of view taken, e.g., by Cuevas, Fraiman, and Rodríguez-Casal [14]. Korostelev and Tsybakov [30] provide detailed analysis of the rate of convergence of various set or boundary estimation errors under several scenarios. Many approaches are rooted in kernel methods, placing a small weight, often in a carefully selected ball of small radius, around data points inside the support set (Devroye and Wise [19]). Object estimation can also be attacked by methods that are based on level sets of densities. Cuevas, Fraiman, and Rodríguez-Casal [14] provide a consistent estimate of the Minkowski content that turns out to also provide an estimate of the boundary of the studied object. However, this boundary estimate does not come with topological guarantees. Approaches like principal curves and surfaces (Hastie and Stuetzle [28]), multiscale geometric analysis (Arias-Castro, Donoho, and Huo [1]) and density-based methods (Genovese, Perone-Pacifico, Verdinelli, and Wasserman [26]) have been successfully used to detect "simple" geometric structures such as one-dimensional curves or filaments in data corrupted by noise.

On the other hand, taking a slightly different and nonstochastic point of view, purely geometric methods have also been developed to infer the geometry of general compact subsets of $\mathbb{R}^d$ from point cloud data. In this context Chazal, Cohen-Steiner, and Lieutier [9, 10] and Chazal, Cohen-Steiner, and Mérigot [11, 13] argue that the study of distance functions to the data provides precise and robust information about the geometry of the sampled objects.

While statistical methods provide efficient tools to deal with noisy data, they do not however come with strong guarantees on the inferred geometric properties or are restricted to the inference of geometrically simple objects such as pieces of smooth curves or topologically trivial manifolds. On the other hand, purely geometric methods offer strong guarantees but, since they do

not integrate any statistical model, they usually rely on sampling assumptions that cannot be met by data corrupted by noise.

In the so-called distance function approach, the unknown object is estimated by the union of balls centered on the data points or, equivalently, by an appropriate sublevel set of the distance function to the data. Thanks to classical properties of distance functions, this procedure has revealed fruitful both from the statistical (Devroye and Wise [19], Biau, Cadre, Mason, and Pelletier [4]) and geometric (Chazal, Cohen-Steiner, and Lieutier [10]) points of view. Unfortunately, the distance function approach obviously fails when the observations are corrupted by "background noise" (as shown for example in Figure 1 and Figure 2), or when the observed data is not exactly drawn from a unique distribution μ but from the convolution of μ with a noncompactly supported noise measure. Different solutions have been proposed to get rid of this problem. These solutions generally rely on statistical models assuming a strong knowledge on the nature of the noise. For example, Niyogi, Smale, and Weinberger [39] show that it is possible to infer the homology of a low-dimensional submanifold $M \subset \mathbb{R}^d$ from data uniformly sampled on M and corrupted by a Gaussian noise in the normal direction to M. In lower dimensions, motivated by applications ranging from the inference of networks of blood vessels to the characterization of filaments in distributions of galaxies, the detection of filamentary structures has been carefully considered. For example, Genovese, Perone-Pacifico, Verdinelli, and Wasserman [26] address this problem using the gradient of a density estimate to exhibit filamentary structures in data; lately, these authors also proposed in [27] an asymptotically consistent geometric approach for the same problem but in dimension 2. Unfortunately, when the data is corrupted by outliers, the latter method requires a—usually tricky—preprocessing step consisting in identifying and eliminating these outliers.

Recently, Chazal, Cohen-Steiner, and Mérigot [12] proposed a framework to bridge the gap between the statistical and geometric points of view. The approach of the authors avoids the cleaning step by replacing the usual distance function by another distance-like function which is robust to the addition of a certain amount of outliers. This function extends the notion of distance functions from compact sets to probability distributions, allowing to robustly infer geometric properties of the distribution μ using independent observations drawn according to a distribution μ' "close" to μ. In this framework, the closeness between probability distributions is assessed by a Wasserstein

distance W_p defined by

$$W_p(\mu, \mu') = \inf_{\pi \in \Pi(\mu,\mu')} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^p \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \right)^{\frac{1}{p}},$$

where $\Pi(\mu, \mu')$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ that have marginals μ and μ', $\|.\|$ is a norm and $p \geq 1$ is a real number (see Villani [50]).

In the approach of Chazal, Cohen-Steiner, and Mérigot [12], given the probability distribution μ in $\mathbb{R}^d$ and a parameter $0 \leq m \leq 1$, the notion of distance to the support of μ is generalized by the function $\delta_{\mu,m} : \mathbf{x} \in \mathbb{R}^d \mapsto \inf\{r > 0 : \mu(\mathcal{B}(\mathbf{x}, r)) > m\}$, where $\mathcal{B}(\mathbf{x}, r)$ denotes the closed ball of center $\mathbf{x}$ and radius r. To avoid trouble due to discontinuities of the map $\mu \mapsto \delta_{\mu,m}$, the distance function to μ with parameter m_0 is defined by

$$d^2_{\mu,m_0} : \mathbb{R}^d \to \mathbb{R}^+, \ \mathbf{x} \mapsto \frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(\mathbf{x})^2 \mathrm{d}m,$$

where $0 < m_0 \leq 1$ is a real number. The function d_{μ,m_0} shares many properties with classical distance functions that make it well-suited for geometric inference purposes. In particular, if the space $\mathcal{P}(\mathbb{R}^d)$ of probability measures in $\mathbb{R}^d$ is equipped with the Wasserstein distance W_2 and the space of real-valued functions is equipped with the supremum norm, then the map $\mu \mapsto d_{\mu,m_0}$ is $1/\sqrt{m_0}$-Lipschitz, i.e.,

$$\sup_{\mathbf{x} \in \mathbb{R}^d} |d_{\mu,m_0}(\mathbf{x}) - d_{\mu',m_0}(\mathbf{x})| = \|d_{\mu,m_0} - d_{\mu',m_0}\|_\infty \leq \frac{1}{\sqrt{m_0}} W_2(\mu, \mu').$$

This property ensures that W_2-close measures have close sublevel sets in $\mathbb{R}^d$. The function d^2_{μ,m_0} is also seen to be semiconcave (that is $\mathbf{x} \mapsto \|\mathbf{x}\|^2 - d^2_{\mu,m_0}(\mathbf{x})$ is convex, see Petrunin [40] for more information on geometric properties of semiconcave functions). This regularity property implies that d^2_{μ,m_0} is of class $\mathcal{C}^2$ almost everywhere, thus ensuring strong regularity properties on the geometry of the level sets of d_{μ,m_0}.

Using these properties, Chazal, Cohen-Steiner, and Mérigot prove, under some general assumptions, that if μ' is a probability distribution approximating μ, then the sublevel sets of d_{μ',m_0} provide a topologically correct approximation of the support of μ (see [12][Corollary 4.11]). Figure 1 and Figure 2 below show some examples of level sets of distance functions to a measure illustrating this result.

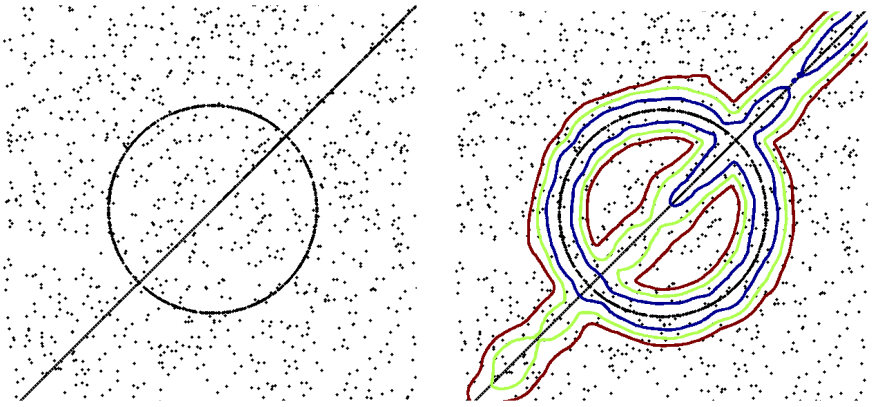

Figure 1: **Left**: A two-dimensional data set where 50% of the points have been uniformly randomly sampled on the union K of a circle and a segment, and 50% have been uniformly randomly sampled in a square containing K. **Right**: Three different level sets of the distance function d_{μ,m_0}, where μ stands for the empirical measure based on the observations and $m_0 = 0.02$, showing that the topology of the union of the circle and the segment can be correctly inferred.

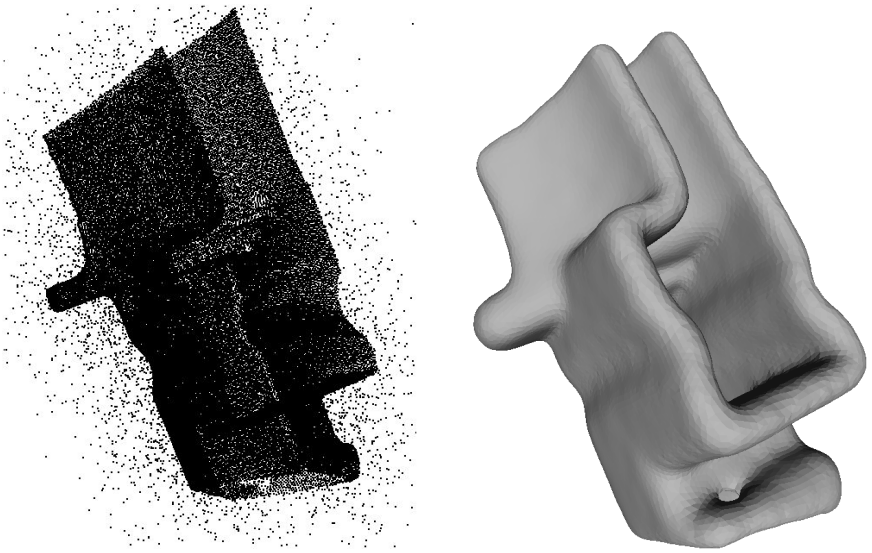

Figure 2: **Left**: A three-dimensional set of points uniformly sampled on the surface of a mechanical part to which 10% of points sampled uniformly at random in a box enclosing the mechanical part have been added. **Right**: An isosurface of the distance function d_{μ,m_0} to the empirical measure based on the observations. This isosurface successfully recovers the topology of the mechanical part. In this example, $m_0 = 0.003$.

2 Connection with density estimation

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent identically distributed observations with common distribution μ on $\mathbb{R}^d$, equipped with the standard Euclidean norm $\|.\|$. The empirical measure μ_n based on $\mathbf{X}_1, \cdots, \mathbf{X}_n$ is defined, for any Borel set $A \subset \mathbb{R}^d$, by

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A]}.$$

This empirical distribution is known to provide a suitable approximation of μ with respect to the Wasserstein distance (Bolley, Guillin, and Villani [7]). Moreover, given a sequence of positive integers $\{k_n\}$ such that $1 \leq k_n \leq n$, for $m_n = k_n/n$ the function d_{μ_n,m_n} takes the simple form

$$d^2_{\mu_n,m_n}(\mathbf{x}) = \frac{1}{k_n} \sum_{j=1}^{k_n} \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|^2$$

where $\mathbf{X}_{(j)}(\mathbf{x})$ is the j-th nearest neighbor to $\mathbf{x}$ among $\mathbf{X}_1, \cdots, \mathbf{X}_n$ and ties are broken arbitrarily. Thus, $\|\mathbf{X}_{(1)}(\mathbf{x}) - \mathbf{x}\| \leq \cdots \leq \|\mathbf{X}_{(n)}(\mathbf{x}) - \mathbf{x}\|$. In other words, the value of $d^2_{\mu_n,m_n}$ at $\mathbf{x}$ is just a weighted sum of the squares of the distances from $\mathbf{x}$ to its first k_n nearest neighbors.

Assume now that the common probability measure μ of the sequence is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with a probability density f. In this context, it turns out that the function d_{μ_n,m_n} is intimately connected to both the geometric properties of μ and to the density f. To see this, observe that in the regions where f is high, the function d_{μ_n,m_n} takes small values while in the regions where f is low, d_{μ_n,m_n} takes larger values. Observe also that the function δ_{μ_n,m_n} is just the distance function to the k_n-th nearest neighbor, i.e., $\delta_{\mu_n,m_n}(\mathbf{x}) = \|\mathbf{X}_{(k_n)} - \mathbf{x}\|$. These remarks motivate the introduction of the density estimate $\hat{f}_n$ of f, defined by

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nV_d} \left(\frac{\sum_{j=1}^{k_n} j^{2/d}}{k_n d^2_{\mu_n,m_n}(\mathbf{x})} \right)^{d/2}, \quad \mathbf{x} \in \mathbb{R}^d,$$

where V_d is the volume of the unit ball in $\mathbb{R}^d$. Put differently, for $\mathbf{x} \in \mathbb{R}^d$,

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nV_d} \left(\frac{\sum_{j=1}^{k_n} j^{2/d}}{\sum_{j=1}^{k_n} \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|^2} \right)^{d/2}.$$

From the geometric inference point of view, the estimate $\hat{f}_n$ allows to infer both the geometric properties of the support of μ and the geometry of the upperlevel sets of f, i.e., sets of the form $\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \geq t\}$. A more general density estimate is given, for $p > 0$, by

$$\tilde{f}_n(\mathbf{x}) = \frac{1}{nV_d} \left(\frac{\sum_{j=1}^{k_n} p_{nj} j^{p/d}}{\sum_{j=1}^{k_n} p_{nj} \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|^p} \right)^{d/p}, \quad \mathbf{x} \in \mathbb{R}^d,$$

where, for $j = 1, \cdots, k_n$,

$$p_{nj} = \int_{\left]\frac{j-1}{k_n}, \frac{j}{k_n}\right]} \nu(\mathrm{d}t)$$

and ν is a given probability measure on $[0, 1]$ with no atom at 0. To avoid trivial complications in the proofs, we assume throughout the document that $p = d$, leaving the reader the opportunity to adapt the results to the case $p \neq d$. Therefore, we will consider the following generalized version of the k-nearest neighbor density estimate of Fix and Hodges [23] and Loftsgaarden and Quesenberry [31], defined by

$$f_n(\mathbf{x}) = \frac{\sum_{j=1}^{k_n} p_{nj} j}{nV_d \sum_{j=1}^{k_n} p_{nj} \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|^d}, \quad \mathbf{x} \in \mathbb{R}^d.$$

This estimate is but a special case of a larger class of estimates proposed by Rodríguez [43] and Rodríguez and Van Ryzin [41, 42] that combine kernel smoothing with nearest neighbor smoothing.

When ν is the Dirac measure at 1, we obtain

$$\sum_{j=1}^{k_n} p_{nj} j = k_n$$

and, consequently,

$$f_n(\mathbf{x}) = \frac{k_n}{n V_d \|\mathbf{X}_{(k_n)}(\mathbf{x}) - \mathbf{x}\|^d}, \quad \mathbf{x} \in \mathbb{R}^d.$$

This is the original Loftsgaarden and Quesenberry k-nearest neighbor density estimate. Its properties are well-understood (Fukunaga and Hostetler [25], Devroye and Wagner [20, 21], Moore and Yackel [36, 37], Mack and Rosenblatt [33], Mack [32], Bhattacharya and Mack [3], Devroye, Györfi, and Lugosi [18], Rodríguez [43, 44, 45]). For example, at Lebesgue-almost all $\mathbf{x}$, we have $f_n(\mathbf{x}) \to f(\mathbf{x})$ in probability as $n \to \infty$, if $k_n \to \infty$ and $k_n/n \to 0$. On the other hand, taking for ν the uniform measure on $[0, 1]$, we obtain

$$\sum_{j=1}^{k_n} p_{nj} j = \frac{1 + k_n}{2},$$

and

$$f_n(\mathbf{x}) = \frac{k_n(1 + k_n)}{2n V_d \sum_{j=1}^{k_n} \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|^d}, \quad \mathbf{x} \in \mathbb{R}^d.$$

The remainder of the paper establishes various properties of f_n (Section 3). In particular, we look at pointwise consistency, and derive a general central limit theorem under the lightest possible conditions. In addition, a strong approximation is obtained as well. The asymptotic mean square error, when optimized with respect to k_n, reduces to a product of three factors, $n^{-4/(d+4)}$ (the rate of convergence), a factor depending upon the local shape of f (which involves the trace of the Hessian), and a factor depending upon ν only. The third factor is invariant for all $\mathbf{x}$, and should thus be optimized once and for all—at least if performance is measured by local mean square error. Attempts at such an optimization are rare—we will optimize ν within a large parametric class of weight functions that also play a role in the optimal shapes of kernels in kernel density estimates as established in the classical papers of Bartlett [2] and Epanechnikov [22]. Using simulations, we finally show in Section 4 the suitability of the class of estimates in a number of important applications. For the sake of clarity, proofs are postponed to Section 5.

Our approach is close in spirit to the one of Samworth [46], who derived asymptotic expansions for the excess risk of a weighted nearest neighbor classifier and found the asymptotically optimal vector of weights. In contrast, we are considering density estimation and our optimization is quite different.

3 Some asymptotic results

Our goal in this section is to establish some pointwise asymptotic properties of the estimate f_n. To this aim, we note once and for all that for any $\rho > 0$, all quantities of the form

$$\int_{[0,1]} t^\rho \nu(\mathrm{d}t)$$

are finite and positive. Moreover, for $\rho \geq 1$, as $k_n \to \infty$,

$$\frac{1}{k_n^\rho} \sum_{j=1}^{k_n} p_{nj} j^\rho = \int_{[0,1]} t^\rho \nu(\mathrm{d}t) \left(1 + \mathcal{O}\left(\frac{1}{k_n}\right)\right).$$

The symbol λ stands for the Lebesgue measure on $\mathbb{R}^d$. We start by establishing the weak pointwise consistency of f_n.

Theorem 3.1 *If $k_n \to \infty$ and $k_n/n \to 0$, then the generalized k-nearest neighbor estimate f_n is weakly consistent at λ-almost all $\mathbf{x}$, that is $f_n(\mathbf{x}) \to f(\mathbf{x})$ in probability at λ-almost all $\mathbf{x}$ as $n \to \infty$.*

Our next result states the mean square consistency of the generalized k-nearest neighbor estimate.

Theorem 3.2 *We have, at λ-almost all $\mathbf{x}$,*

$$\mathbb{E}\left[f_n^2(\mathbf{x})\right] < \infty$$

whenever $k_n \geq 5$. Furthermore, if $k_n \to \infty$ and $k_n/n \to 0$, then, for such $\mathbf{x}$,

$$\mathbb{E}\left[f_n(\mathbf{x}) - f(\mathbf{x})\right]^2 \to 0 \quad \text{as } n \to \infty.$$

The asymptotic normality of the original Loftsgaarden and Quesenberry k-nearest neighbor estimate has been established by Moore and Yackel in [37]. These authors proved that for f sufficiently smooth in a neighborhood of $\mathbf{x}$, $f(\mathbf{x}) > 0$, $k_n \to \infty$ and $k_n/n^{2/(d+2)} \to 0$ as $n \to \infty$, then

$$\sqrt{k_n} \frac{f_n(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} \xrightarrow{\mathcal{D}} N,$$

where N is a standard normal random variable. This result was obtained for the generalized k-nearest neighbor estimate by Rodríguez [43, 45]. The novelty in Theorem 3.3 is that it is a strong approximation result, which is interesting by itself and implies the classical central limit theorem. We let $\Gamma(.)$ be the gamma function and denote by $[\partial^2 f(\mathbf{x})/\partial \mathbf{x}^2]$ the Hessian matrix of f at $\mathbf{x}$, which is given by

$$\left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\right]_{i,j} \stackrel{\text{def}}{=} \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}.$$

Notation $\mathrm{tr}(A)$ stands for the trace of the square matrix A. For a sequence of random variables $\{\zeta_n\}$ and a deterministic sequence $\{u_n\}$, notation $\zeta_n = \mathrm{o}_{\mathbb{P}}(u_n)$ means that ζ_n/u_n goes to 0 in probability as n tends to infinity, and notation $\zeta_n = \mathcal{O}_{\mathbb{P}}(u_n)$ means that ζ_n/u_n is bounded in probability as n tends to infinity.

Theorem 3.3 *Let $\mathbf{x} \in \mathbb{R}^d$ and assume that f has derivatives of second order at $\mathbf{x}$, with $f(\mathbf{x}) > 0$. Let*

$$v^2 = \frac{\displaystyle\int_0^1 (1-\Phi(t))^2 \,\mathrm{d}t}{\displaystyle\left[\int_{[0,1]} t\nu(\mathrm{d}t)\right]^2} \quad \textit{and} \quad b = \frac{\displaystyle\int_{[0,1]} t^{1+2/d}\nu(\mathrm{d}t)}{\displaystyle\int_{[0,1]} t\nu(\mathrm{d}t)},$$

with

$$\Phi(t) = \int_{[0,t]} \nu(\mathrm{d}u), \quad t \in [0,1].$$

Let also

$$c(\mathbf{x}) = \frac{1}{2(d+2)\pi}\Gamma^{2/d}\left(\frac{d+2}{2}\right)\mathrm{tr}\left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\right].$$

Then, if N denotes a standard normal random variable, and if $k_n \to \infty$ and $k_n/n \to 0$,

$$f_n(\mathbf{x}) - f(\mathbf{x}) \stackrel{\mathcal{D}}{=} \frac{f(\mathbf{x})v}{\sqrt{k_n}}N + \frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})}\left(\frac{k_n}{n}\right)^{2/d} + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k_n}} + \left(\frac{k_n}{n}\right)^{2/d}\right).$$

Theorem 3.3 can be used when k_n is at its optimal value (about $n^{4/(d+4)}$). It can also be used when k_n is below this optimal value, that is

$$\sqrt{k_n}\frac{f_n(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} \stackrel{\mathcal{D}}{\to} vN,$$

and above it, when

$$\left(\frac{n}{k_n}\right)^{2/d}(f_n(\mathbf{x})-f(\mathbf{x}))\to\frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})}\quad\text{in probability.}$$

The usual k-nearest neighbor estimate has $v^2=1$. Consequently, for this estimate,

$$\sqrt{k_n}\frac{f_n(\mathbf{x})-f(\mathbf{x})}{f(\mathbf{x})}\xrightarrow{\mathcal{D}}N, \tag{3.1}$$

provided $k_n\to\infty$ and $k_n/n^{4/(d+4)}\to 0$ as $n\to\infty$. This is precisely the asymptotic normality result of Moore and Yackel [37]. Note however that our condition $k_n/n^{4/(d+4)}\to 0$ is less severe than the condition $k_n/n^{2/(d+2)}\to 0$, which is imposed by these authors at the price of a less stringent smoothness condition on f however. In any case, consistency (3.1) deals with the uninteresting case of a k_n which is suboptimal (that is, the bias in $f_n(\mathbf{x})$ is negligible with respect to the variance term). Note, in addition, that analogues of Theorem 3.1 (yet with different rates) may be obtained in the somewhat degenerated situations where $f(\mathbf{x})$ and/or $c(\mathbf{x})=0$ by pushing the asymptotic expansions.

Theorem 3.3 has also interesting consequences for the analysis of the mean square error development of the estimate f_n. Let $\lceil\cdot\rceil$ be the nearest larger integer (or ceiling) function.

Theorem 3.4 *With the notation and conditions of Theorem 3.3, if $k_n\to\infty$ and $k_n/n\to 0$, then*

$$\mathbb{E}\left[f_n(\mathbf{x})-f(\mathbf{x})\right]^2=\frac{f^2(\mathbf{x})v^2}{k_n}+\frac{c^2(\mathbf{x})b^2}{f^{4/d}(\mathbf{x})}\left(\frac{k_n}{n}\right)^{4/d}+\mathrm{o}\left(\frac{1}{k_n}+\left(\frac{k_n}{n}\right)^{4/d}\right)$$

whenever $f(\mathbf{x})>0$. Thus, for such $\mathbf{x}$, assuming that $c(\mathbf{x})\neq 0$ and for the choice

$$k_n=\left\lceil\left(\frac{df^{2+4/d}(\mathbf{x})v^2}{4c^2(\mathbf{x})b^2}\right)^{\frac{d}{d+4}}n^{\frac{4}{d+4}}\right\rceil,$$

we have

$$\mathbb{E}\left[f_n(\mathbf{x})-f(\mathbf{x})\right]^2=\Delta(\mathbf{x})n^{-\frac{4}{d+4}}+\mathrm{o}\left(n^{-\frac{4}{d+4}}\right),$$

where

$$\Delta(\mathbf{x})=\left(1+\frac{d}{4}\right)\left(\frac{4f^{4/d}(\mathbf{x})c^2(\mathbf{x})v^{8/d}b^2}{d}\right)^{\frac{d}{d+4}}.$$

For the standard k-nearest neighbor estimate,

$$\Delta(\mathbf{x}) = \left(1 + \frac{d}{4}\right) \left(\frac{4f^{4/d}(\mathbf{x})c^2(\mathbf{x})}{d}\right)^{\frac{d}{d+4}}$$

and thus, as $n \to \infty$,

$$\mathbb{E}\left[f_n(\mathbf{x}) - f(\mathbf{x})\right]^2 = \Delta(\mathbf{x}) n^{-\frac{4}{d+4}} + \mathrm{o}\left(n^{-\frac{4}{d+4}}\right),$$

for the optimal choice

$$k_n = \left\lceil \left(\frac{df^{2+4/d}(\mathbf{x})}{4c^2(\mathbf{x})}\right) n^{\frac{4}{d+4}} \right\rceil.$$

For the original Loftsgaarden and Quesenberry's density estimate, the optimization problem of k_n with respect to the mean square error criterion is thoroughly discussed in Fukunaga and Hostetler [25]. The best possible asymptotic quadratic error for the generalized k-nearest neighbor estimate, as given in Theorem 3.4, consists of a product of three factors: The first factor depends upon n and d only, and is the general rate of convergence. The second factor depends upon $f(\mathbf{x})$ and $c(\mathbf{x})$, and we have no control over that. The third factor is

$$\left(v^{8/d} b^2\right)^{\frac{d}{d+4}},$$

which depends directly on our measure ν. It is clear that we would like to minimize this factor. It is more convenient to work with a power of it, namely

$$A \stackrel{\text{def}}{=} b v^{\frac{4}{d}}.$$

For the Dirac measure at 1 (the classical k-nearest neighbor estimate), we note that $v = b = 1$, so $A = 1$.

The first important consequence of the factorization is that the optimal ν is the same at all points $\mathbf{x}$ with $f(\mathbf{x}) > 0$ and $c(\mathbf{x}) \neq 0$. A similar property has been noted a long time ago for the form of the best positive kernel in the Parzen-Rosenblatt density estimate (see Bartlett [2] and Epanechnikov [22] for $d = 1$ and Deheuvels [16] for $d > 1$).

The functional optimization of A seems daunting, but one can make a good guess in the following manner. Assume that we let ν be the measure of U^α, where U is uniform on $[0, 1]$, and $\alpha \geq 0$ is a parameter. The case $\alpha = 0$ again yields the atomic measure at 1. Repeatedly using the fact that $\mathbb{E}[U^s] = 1/(s+1)$, simple calculations show that

$$A = A(\alpha) = \frac{d 2^{\frac{2}{d}} (1+\alpha)^{1+\frac{2}{d}}}{(2+\alpha)^{\frac{2}{d}} (2\alpha + d\alpha + d)}.$$

The behavior of A as a function of α (see Figure 3 below) is best captured by studying $\log A$ and taking derivatives. This reveals that $A(0) = 1$, that A decreases initially to reach a minimum at $\alpha = d/2$, that the minimal value is

$$\frac{1}{2}\left(\frac{2+d}{2+\frac{d}{2}}\right)^{1+\frac{2}{d}},$$

and that A increases again to a limiting value given by

$$\frac{2^{\frac{2}{d}}}{1+\frac{2}{d}}.$$

The latter limit is ≤ 1 for $d \geq 2$. The value of A at the minimum is a strictly increasing function in d with limit tending to one (see Figure 4).

In other words, except for $d = 1$, any value of $\alpha > 0$ is better than $\alpha = 0$: The classical k-nearest neighbor estimate is actually the worst possible in this entire class of natural weights! Furthermore, for any $d \geq 1$, by taking $\alpha = d/2$, we obtain an improvement over the classical k-nearest neighbor estimate that is most outspoken for $d = 1$. It is interesting that for $d = 1$, ν is the law of $\sqrt{U}$, which has a triangular (increasing) density on $[0, 1]$. Rodríguez [43] obtained a similar result for the best weights in a weighted k-nearest neighbor rule for density estimation. For $d = 2$, ν is the uniform law on $[0, 1]^2$: So it is best to weigh all of the k-nearest neighbors equally. We do not know whether $U^{d/2}$ is in fact the optimal value.

Note also that in this paper, we are fixing the distance metric which determines the ranking among neighbors. There is ample evidence, especially from practicing nonparametric statisticians, that in moderate and high dimensions, a lot can be gained by considering variable metrics, such as Euclidean metrics applied after performing a locally affine (or matrix multiplication) transformation, and letting the data select to some extent the metric. This strategy was already present in the work of Short and Fukunaga [47] and Fukunaga and Flick [24]. Kernel estimates are better adapted to take advantage of local second order or Hessian structure. Combinations of nearest neighbor and kernel estimates that incorporate these ideas are being considered by a subset of the authors in [45].

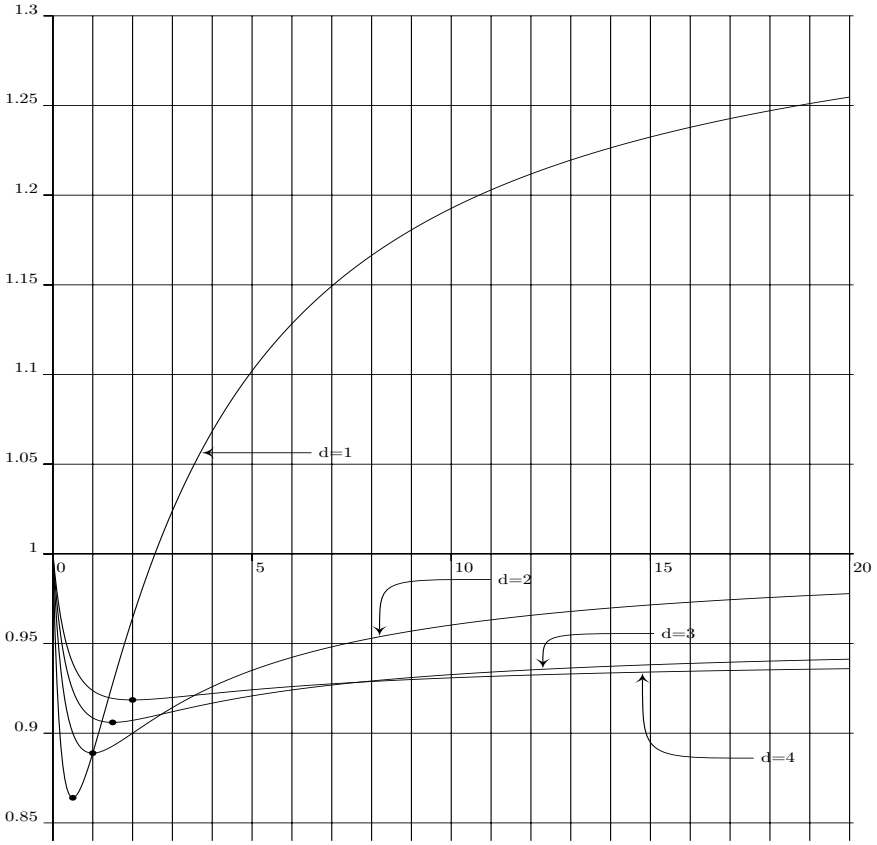

Figure 3: This figure shows A versus α for $1 \leq d \leq 4$. Note that A exceeds 1 only for $d = 1$ and α large enough.

4 Numerical illustrations

A series of experiments were conducted in order to compare the performance of our weighted estimate with that of the standard k-nearest neighbor estimate of Fix and Hodges [23] and Loftsgaarden and Quesenberry [31]. We provide numerical illustrations regarding both the geometric and convergence properties of the estimates.

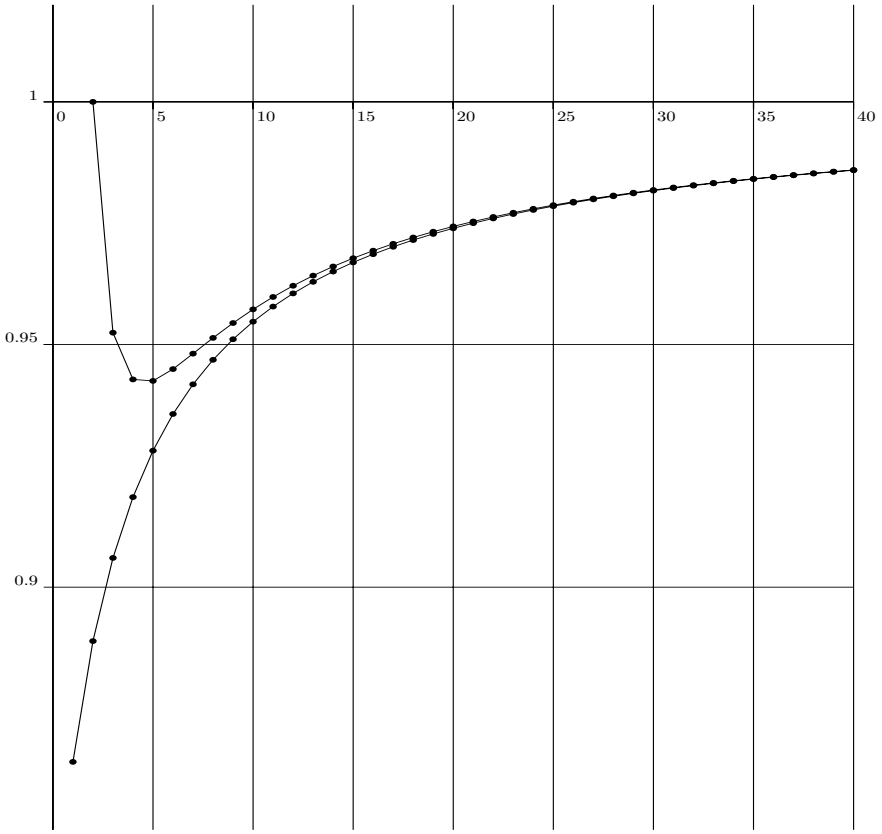

Figure 4: This figure shows the minimal value of A and the limiting value of A versus d. Note that both are nearly indistinguishable for $d \geq 10$.

On the geometrical side, particular attention was paid on the comparison of the geometry of the level sets of the various estimates. To this aim, we investigated three synthetic data sets, sampled according to known probability density models, and one real-life data set. These four data sets are denoted **D1**, **D2**, **D3** and **D4** hereafter and are described below.

D1: A two-dimensional data set of 5,000 points, randomly sampled according to a bivariate standard normal distribution (see Figure 5, **left**).

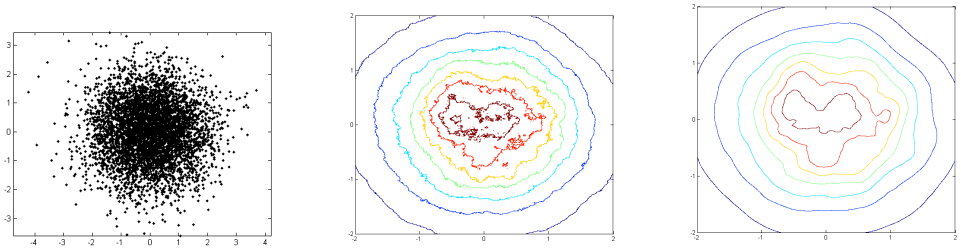

Figure 5: **Left**: Data set **D1**. **Middle and right**: Level sets of the standard (middle) and weighted (right) k-nearest neighbor estimates corresponding to level values $0.02, 0.04, 0.06, 0.08, 0.10, 0.12$ and 0.14.

D2: A two-dimensional data set of 8,000 points, randomly sampled according to an equal mixture of two bivariate normal distributions, with respective means $(-0.7, -0.7)$, $(0.7, 0.7)$ and covariance matrices

$$\begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$$

and

$$\begin{pmatrix} 0.8 & -0.4 \\ -0.4 & 0.8 \end{pmatrix}$$

(see Figure 6, **left**).

D3: A two-dimensional real data set representing the epicenters of 12,790 earthquakes registered during the period 1970-2010 on the longitude-latitude rectangle $[-170, 10] \times [-70, 70]$ (see Figure 7, **left**). This data has been extracted from the US Geological Survey database [48].

D4: A three-dimensional data set of 50,000 points, randomly sampled according to a standard normal distribution in $\mathbb{R}^3$.

The computing program codes were implemented in C++ using the Approximate Nearest Neighbor library developed by Mount and Arya [38]. Due to the efficiency of this library, all computations took a few seconds to a few minutes on a standard laptop. The programs are available upon request from the authors.

For the two-dimensional data sets **D1**, **D2** and **D3**, the density estimates were first evaluated on the vertices of a regular 2000×2000 grid, and the level sets were extracted using the *contour* function in Matlab. Figures 5, 6 and 7 depict, for each of these data sets, some level sets of the standard k-nearest

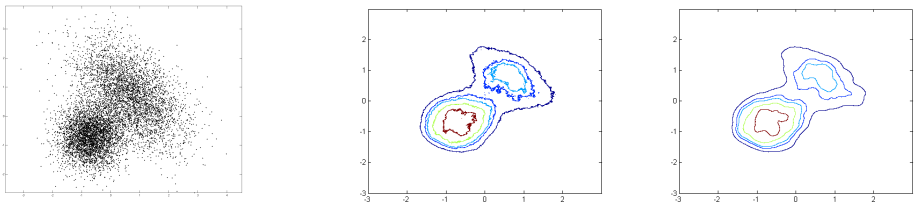

Figure 6: **Left**: Data set **D2**. **Middle and right**: Level sets of the standard (middle) and weighted (right) k-nearest neighbor estimates corresponding to level values $0.06, 0.085, 0.10, 0.14$ and 0.21.

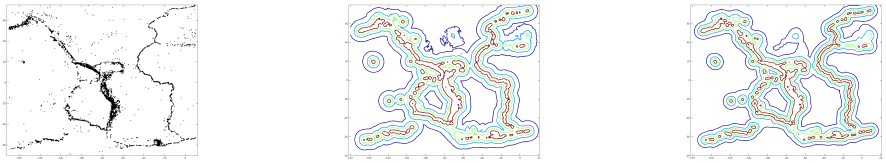

Figure 7: **Left**: Earthquakes data set **D3**. **Middle and right**: Level sets of the standard (middle) and weighted (right) k-nearest neighbor estimates corresponding to level values $27 \cdot 10^{-7}, 82 \cdot 10^{-7}, 20 \cdot 10^{-6}$ and $110 \cdot 10^{-6}$.

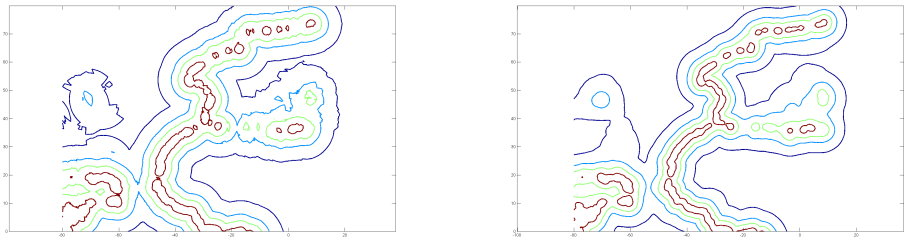

Figure 8: A zoom on the level sets of Figure 7.

neighbor estimate (**middle** column in the figures) and some level sets of the weighted estimate with uniform weights (**right** column in the figures). For the data sets **D1** and **D2**, Figures 9 and 10 (**left**) also show the details of one selected level set of the true density and their corresponding density-based estimates. Regarding the three-dimensional data set **D4**, we also used the uniform weights for the generalized estimate and meshed some level sets of the estimates using an implicit surface mesher from the C++ CGAL library [8] (Figure 11).

An important issue regarding k-nearest neighbor-based density estimates is how to select the number k_n of neighbors. In our experiments, this pa-

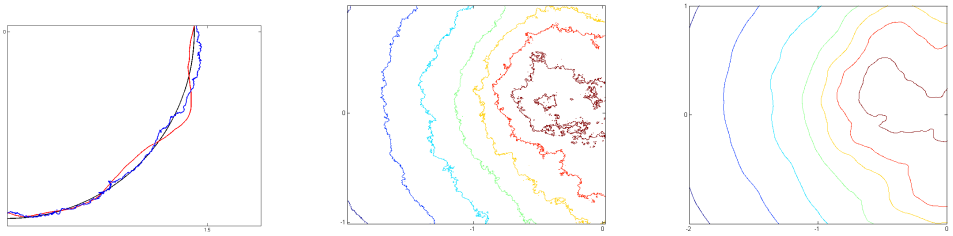

Figure 9: **Left**: Plot of the 0.06-level sets of the true density (green), the standard (blue) and weighted (red) k-nearest neighbor estimates for the data set **D1**. **Middle and right**: A zoom on the level sets of Figure 5, showing that the unweighted estimate does not allow to correctly infer the connectedness of the level sets of the true density.

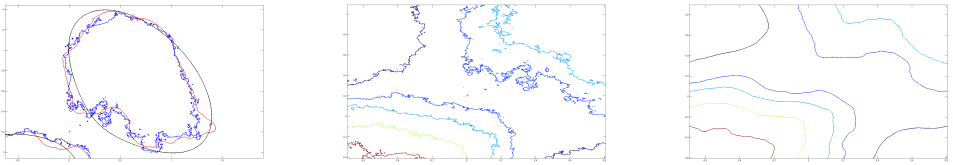

Figure 10: **Left**: Plot of the 0.085-level sets of the true density (black), the standard (blue) and weighted (red) k-nearest neighbor estimates for the data set **D1**. **Middle and right**: A zoom on the level sets of Figure 6. Here again, the unweighted estimate does not allow to correctly infer the connectedness of the level sets of the true density.

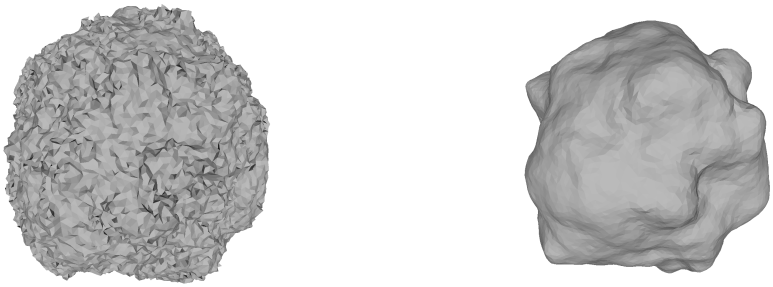

Figure 11: Level sets of the standard (left) and weighted (right) k-nearest neighbor estimates for the data set **D4**. As in the two-dimensional case, the weighted estimate provides much smoother level sets.

rameter was selected using a standard leave-one-out cross validation method

performing on the (global) L_2 criterion

$$\int_{\mathbb{R}^d} [f_n(\mathbf{x}) - f(\mathbf{x})]^2 \, \mathrm{d}\mathbf{x}.$$

As this procedure does not come with any theoretical guarantee (as far as we know), we also evaluated the errors between the cross-validated estimates and the true density when it was known (i.e., for data sets **D1**, **D2** and **D4**). In all cases, the selected value of k_n appears to be very close to the optimal oracle $k_n^\star$, which minimizes the L_2 norm between the targeted density and the estimate. The selected values of k_n are shown in Table 1, together with the L_2 norm between the estimates (respectively, the oracles) and the true density (models **D1**, **D2** and **D4** only).

Data set	**D1**	**D2**	**D3**	**D4**
Selected k_n (cross-validation)	107	127	7	410
L_2 error (standard estimate)	0.0296	0.0326	-	0.091
Selected k_n (cross-validation)	210	184	11	500
L_2 error (weighted estimate)	0.0261	0.0309	-	0.010
Oracle k_n^* (standard estimate)	183	180	-	3050
Oracle L_2 error (standard estimate)	0.0276	0.0312	-	0.062
Oracle k_n^* (weighted estimate)	250	222	-	550
Oracle L_2 error (weighted estimate)	0.0258	0.0295	-	0.009

Table 1: Cross-validated selected k_n and associated L_2 errors.

A general observation is that, in all cases, the classical k-nearest neighbor estimate provides a pretty poor geometric approximation of the level sets of the true density. In the 2D case, these sets are very jagged and contain spurious small connected components (Figures 7, 8, 9 and 10), thereby preventing any direct inference on the geometry of the level sets of the true density (such as, for instance, their connectedness). On the other hand, the level sets of the generalized estimate are much smoother and, for values that are not too close to the critical values of the true density, they appear to be homeomorphic to the ones of the target.

In the 3D situation, it is noteworthy that the level sets of the weighted estimate are smoother than the ones of the standard k-nearest neighbor (Figure 11). For technical reasons, the surface mesher was only able to mesh the component of the level sets containing the origin of $\mathbb{R}^3$. As a consequence,

the spurious small components of the standard k-nearest neighbor estimate (similar to the ones depicted in the 2D figures) are not represented on Figure 11, **left**.

Finally, in order to illustrate the convergence properties of the generalized k-nearest neighbor estimate, we generated, for each $n \in \{1 \cdot 10^4, 2 \cdot 10^4, \cdots, 15 \cdot 10^4\}$, 100 data sets of n points randomly sampled according to a standard normal distribution in $\mathbb{R}^2$. These observations were used to estimate $\mathbb{E}[f_n(\mathbf{x}) - f(\mathbf{x})]^2$ at 900 points $\mathbf{x}$ distributed on a 30×30 regular grid G on $[-3, 3] \times [-3, 3]$, where f_n was either the standard k-nearest neighbor density estimate or the generalized estimate with uniform weights. We took $k_n = n^{2/3}$.

For each $\mathbf{x}$ and each n, we first computed the average value of $[f_n(\mathbf{x}) - f(\mathbf{x})]^2$ over the 100 data sets of size n and then averaged the outcomes over the 900 points of the grid G. Figure 12 shows the results as a function of n: The red curve corresponds to the weighted estimate while the blue one refers to the unweighted one. In both cases, we see that the estimates converge to the true density with a smaller error for the generalized k-nearest neighbor estimate.

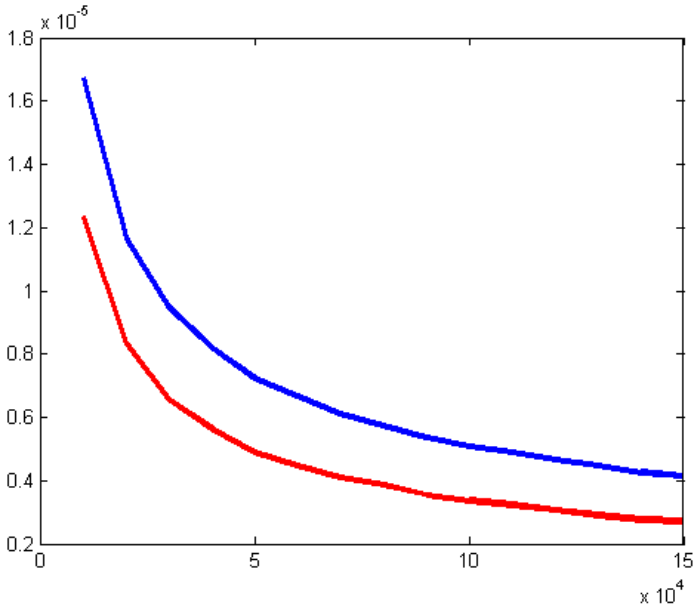

Figure 12: Estimation of $\mathbb{E}[f_n(\mathbf{x}) - f(\mathbf{x})]^2$ averaged over 900 points of the regular grid G, as a function of n. The blue curve corresponds to the standard k-nearest neighbor estimate ($k_n = n^{2/3}$) and the red one to the weighted estimate (uniform weights and $k_n = n^{2/3}$).

5 Proofs

Throughout this section, we let $\mathcal{B}(\mathbf{x}, r)$ be the closed ball in $\mathbb{R}^d$ of radius r centered at $\mathbf{x}$ and denote by μ the probability measure associated with the density f. The collection of all $\mathbf{x}$ with $\mu(\mathcal{B}(\mathbf{x}, \varepsilon)) > 0$ for all $\varepsilon > 0$ is called the support of μ. We denote it by supp μ and note that it may alternatively be defined as the smallest closed subset of $\mathbb{R}^d$ of μ-measure 1.

5.1 Two basic lemmas

We will make repeated use of the following two lemmas.

Lemma 5.1 *Let $U_1, \cdots, U_n$ be i.i.d. uniform $[0, 1]$ random variables with order statistics $U_{(1)} \leq \cdots \leq U_{(n)}$. Then*

$$\left(U_{(1)}, \cdots, U_{(n)}\right) \stackrel{\mathcal{D}}{=} \left(\frac{\sum_{j=1}^1 E_j}{n+1}, \cdots, \frac{\sum_{j=1}^n E_j}{n+1}\right)(1 + \zeta_n),$$

where $E_1, \cdots, E_n$ is a sequence of i.i.d. standard exponential random variables and $\zeta_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ as $n \to \infty$. Furthermore, for all positive integers r,

$$\sup_{n \geq 2r} \left[n^{r/2} \mathbb{E}|\zeta_n|^r\right] < \infty.$$

Proof of Lemma 5.1 It is well known that if $E_1, \cdots, E_{n+1}$ is a sequence of i.i.d. standard exponential random variables, then

$$\left(U_{(1)}, \cdots, U_{(n)}\right) \stackrel{\mathcal{D}}{=} \left(\frac{\sum_{j=1}^1 E_j}{\sum_{j=1}^{n+1} E_j}, \cdots, \frac{\sum_{j=1}^n E_j}{\sum_{j=1}^{n+1} E_j}\right)$$

(see, e.g., Devroye [17, Chapter 5]). Let G_{n+1} be the gamma $(n+1)$ random variable $\sum_{j=1}^{n+1} E_j$. Then, by the central limit theorem,

$$\sqrt{n}\left(\frac{G_{n+1}}{n+1} - 1\right) \stackrel{\mathcal{D}}{\to} N,$$

where N is a standard normal random variable. Thus, by an application of the delta method, we obtain

$$\sqrt{n}\left(\frac{n+1}{G_{n+1}} - 1\right) \stackrel{\mathcal{D}}{\to} N,$$

and the first part of the lemma follows by setting

$$\zeta_n = \frac{n+1}{G_{n+1}} - 1.$$

To prove the second statement, observe that by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left|\frac{n+1}{G_{n+1}}-1\right|^r \leq \sqrt{\mathbb{E}\left|G_{n+1}-(n+1)\right|^{2r}} \times \sqrt{\mathbb{E}G_{n+1}^{-2r}}.$$

The first term in the above product is $\mathcal{O}(n^{r/2})$ (see, e.g., Willink [52]) whereas the second one is infinite for $n+1 \leq 2r$ and $\mathcal{O}(1/n^r)$ otherwise. It follows that

$$\sup_{n\geq 2r}\left[n^{r/2}\mathbb{E}\left|\frac{n+1}{G_{n+1}}-1\right|^r\right] < \infty.$$

■

Lemma 5.2 *Let $E_1, E_2, \cdots$ be a sequence of i.i.d. standard exponential random variables and let $\{k_n\}$ be a sequence of positive integers. For $j = 1, \cdots, k_n$, let*

$$p_{nj} = \int_{\left]\frac{j-1}{k_n}, \frac{j}{k_n}\right]} \nu(\mathrm{d}t),$$

where ν is a given probability measure on $[0,1]$ with no atom at 0. Fix $\rho \geq 1$. Then, if $k_n \to \infty$,

$$\frac{\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)^\rho}{\sum_{j=1}^{k_n} p_{nj} j^\rho} = 1+\zeta_n,$$

where $\zeta_n = \mathcal{O}_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r,

$$\sup_{n\geq 1}\left[k_n^{r/2}\mathbb{E}|\zeta_n|^r\right] < \infty.$$

In addition, letting

$$\Phi(t) = \int_{[0,t]} \nu(\mathrm{d}u), \quad t \in [0,1]$$

and

$$\sigma^2 = \int_0^1 (1-\Phi(t))^2 \mathrm{d}t,$$

then, on an appropriate probability space, there exists a standard normal random variable N such that

$$\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j) = \int_{[0,1]} t\nu(\mathrm{d}t) + \frac{\sigma}{\sqrt{k_n}}N + \zeta_n',$$

where $\zeta'_n = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2})$ *and, for all positive integers* r*,*

$$\sup_{n\geq 1}\left[k_n^{r/2}\mathbb{E}|\zeta'_n|^r\right] < \infty.$$

Proof of Lemma 5.2 Denote by $\lceil . \rceil$ the ceiling function and observe that, since ν has no atom at 0,

$$\begin{aligned}\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)^\rho &= \int_{[0,1]} \left(E_1+\cdots+E_{\lceil tk_n\rceil}\right)^\rho \nu(\mathrm{d}t)\\ &= \int_{[0,1]} \left(\lceil tk_n\rceil + S_{\lceil tk_n\rceil}\right)^\rho \nu(\mathrm{d}t),\end{aligned}$$

where we set

$$S_{\lceil tk_n\rceil} = \sum_{j=1}^{\lceil tk_n\rceil}(E_j-1).$$

Note that $S_{\lceil tk_n\rceil}$ is a sum of i.i.d. zero mean random variables. Therefore,

$$\begin{aligned}\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)^\rho = &\int_{[0,1]} \lceil tk_n\rceil^\rho \nu(\mathrm{d}t)\\ &+ \int_{[0,1]} \left[\left(1+\frac{S_{\lceil tk_n\rceil}}{\lceil tk_n\rceil}\right)^\rho - 1\right]\lceil tk_n\rceil^\rho\nu(\mathrm{d}t).\end{aligned}$$

By an application of Donsker's and continuous mapping theorems (see, e.g., van der Vaart and Wellner [49]), as $k_n\to\infty$,

$$\int_{[0,1]} \left[\left(1+\frac{S_{\lceil tk_n\rceil}}{\lceil tk_n\rceil}\right)^\rho - 1\right]\lceil tk_n\rceil^\rho\nu(\mathrm{d}t) = \int_{[0,1]} \rho\frac{S_{\lceil tk_n\rceil}}{\lceil tk_n\rceil}\lceil tk_n\rceil^\rho\nu(\mathrm{d}t) + k_n^\rho\zeta_{n1},$$

where $\zeta_{n1} = \mathcal{O}_{\mathbb{P}}(k_n^{-1})$ and, for all positive integers r,

$$\sup_{n\geq 1}\left[k_n^r\mathbb{E}|\zeta_{n1}|^r\right] < \infty.$$

Similarly,

$$\int_{[0,1]} \rho\frac{S_{\lceil tk_n\rceil}}{\lceil tk_n\rceil}\lceil tk_n\rceil^\rho\nu(\mathrm{d}t) = k_n^\rho\zeta_{n2},$$

where $\zeta_{n2} = \mathcal{O}_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r,

$$\sup_{n\geq 1}\left[k_n^{r/2}\mathbb{E}|\zeta_{n2}|^r\right] < \infty.$$

Consequently,

$$\frac{1}{k_n^\rho}\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)^\rho = \int_{[0,1]} t^\rho \nu(\mathrm{d}t) + \zeta_n,$$

where $\zeta_n = \mathcal{O}_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r,

$$\sup_{n\geq 1}\left[k_n^{r/2}\mathbb{E}|\zeta_n|^r\right] < \infty.$$

The conclusion of the first assertion follows by observing that, for $\rho \geq 1$,

$$\frac{1}{k_n^\rho}\sum_{j=1}^{k_n} p_{nj} j^\rho = \int_{[0,1]} t^\rho \nu(\mathrm{d}t)\left(1+\mathcal{O}\left(\frac{1}{k_n}\right)\right).$$

The proof of the second assertion requires a bit more care. We already know that

$$\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j) = \int_{[0,1]} t\nu(\mathrm{d}t) + \frac{1}{k_n}\int_{[0,1]} S_{\lceil tk_n\rceil}\nu(\mathrm{d}t) + \zeta_{n3}, \quad (5.1)$$

where $\zeta_{n3} = \mathcal{O}(k_n^{-1})$. With respect to the second term on the right-hand side of (5.1), we have

$$\frac{1}{k_n}\int_{[0,1]} S_{\lceil tk_n\rceil}\nu(\mathrm{d}t) = \frac{1}{k_n}\sum_{j=1}^{k_n}\left[(E_j-1)\int_{\left]\frac{j-1}{k_n},1\right]}\nu(\mathrm{d}t)\right].$$

Clearly, letting

$$\sigma_{nj} = \int_{\left]\frac{j-1}{k_n},1\right]}\nu(\mathrm{d}t), \quad j=1,\cdots,k_n$$

and

$$\Phi(t) = \int_{[0,t]}\nu(\mathrm{d}u), \quad t\in[0,1],$$

we may write

$$\sum_{j=1}^{k_n}\sigma_{nj}^2 = \sum_{j=1}^{k_n}\left(1-\Phi\left(\frac{j-1}{k_n}\right)\right)^2.$$

As a consequence, setting

$$\sigma^2 = \int_0^1 (1-\Phi(t))^2\mathrm{d}t$$

and using the fact that $0 \leq (1-\Phi(t))^2 \leq 1$ is a monotone nonincreasing function, a Riemannian argument shows that

$$\frac{1}{k_n}\sum_{j=1}^{k_n}\sigma_{nj}^2 \in \left[\sigma^2, \sigma^2 + \frac{1}{k_n}\right]. \tag{5.2}$$

Therefore, we obtain via the Komlós, Major, and Tusnády strong approximation result (see Komlós, Major, and Tusnády [29] and Mason [34]) that, on the same probability space, there exists a sequence $E_1, E_2, \cdots$ of i.i.d. standard exponential random variables and a sequence $N_1, N_2, \cdots$ of standard normal random variables such that, for positive constants C_1 and λ_1 and for all $x \geq 0$,

$$\mathbb{P}\left(\sqrt{k_n}\left|\frac{1}{\sqrt{\sum_{j=1}^{k_n}\sigma_{nj}^2}}\sum_{j=1}^{k_n}\sigma_{nj}(E_j-1) - N_{k_n}\right| > x\right) \leq C_1 e^{-\lambda_1 x}.$$

Using (5.2), we deduce that, for positive constants λ_2, λ_3 and all n large enough,

$$\begin{aligned}\mathbb{P}&\left(\sqrt{k_n}\left|\frac{1}{\sqrt{k_n}}\sum_{j=1}^{k_n}\sigma_{nj}(E_j-1) - \sigma N_{k_n}\right| > x\right)\\ &\leq C_1 e^{-\lambda_2 x} + \mathbb{P}\left(|N_{k_n}| > \lambda_3\sqrt{k_n}x\right).\end{aligned}$$

Thus, writing

$$\zeta_{n4} = \frac{1}{k_n}\sum_{j=1}^{k_n}\sigma_{nj}(E_j-1) - \frac{\sigma}{\sqrt{k_n}}N_{k_n},$$

we see that

$$\frac{1}{k_n}\int_{[0,1]} S_{\lceil tk_n\rceil}\nu(\mathrm{d}t) = \frac{\sigma}{\sqrt{k_n}}N_{k_n} + \zeta_{n4},$$

where $\zeta_{n4} = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2})$ and

$$\sup_{n\geq 1}\left[k_n^{r/2}\mathbb{E}|\zeta_{n4}|^r\right] < \infty$$

for all positive integers r. Plugging this identity into (5.1) leads to the desired result. ■

5.2 Proof of Theorem 3.1

Let $\mathbf{x}$ be a Lebesgue point of f, that is, an $\mathbf{x}$ for which

$$\lim_{r\to 0}\frac{\mu\left(\mathcal{B}(\mathbf{x},r)\right)}{\lambda\left(\mathcal{B}(\mathbf{x},r)\right)}=\lim_{r\to 0}\frac{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)}f(\mathbf{y})\mathrm{d}\mathbf{y}}{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)}\mathrm{d}\mathbf{y}}=f(\mathbf{x}).$$

As f is a density, we know that λ-almost all $\mathbf{x}$ satisfy the property given above (see for example Wheeden and Zygmund [51]).

Assume first that $f(\mathbf{x})>0$. Fix $\varepsilon\in(0,1)$ and find $\delta>0$ such that

$$\sup_{0<r\leq\delta}\left|\frac{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)}f(\mathbf{y})\mathrm{d}\mathbf{y}}{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)}\mathrm{d}\mathbf{y}}-f(\mathbf{x})\right|\leq\varepsilon f(\mathbf{x}). \tag{5.3}$$

Let F be the (continuous) univariate distribution function of $W\stackrel{\text{def}}{=}\|\mathbf{X}-\mathbf{x}\|^d$. Note that if $w\leq\delta^d$, then

$$\begin{aligned}F(w)&=\mathbb{P}\left(\|\mathbf{X}-\mathbf{x}\|^d\leq w\right)=\mathbb{P}\left(\mathbf{X}\in\mathcal{B}\left(\mathbf{x},w^{1/d}\right)\right)\\&=\int_{\mathcal{B}(\mathbf{x},w^{1/d})}f(\mathbf{y})\mathrm{d}\mathbf{y}\in\left[(1-\varepsilon)V_df(\mathbf{x})w,(1+\varepsilon)V_df(\mathbf{x})w\right].\end{aligned}$$

Define $W_j=\|\mathbf{X}_j-\mathbf{x}\|^d$, $j=1,\cdots,n$, and let $W_{(1)}\leq\cdots\leq W_{(n)}$ be the order statistics for $W_1,\cdots,W_n$. If $U_{(1)}\leq\cdots\leq U_{(n)}$ are uniform $[0,1]$ order statistics, we have in fact the representation

$$W_{(j)}\stackrel{\mathcal{D}}{=}F^{\text{inv}}\left(U_{(j)}\right)$$

jointly for all j. Thus, provided $U_{(j)}\leq F(\delta^d)$,

$$\frac{U_{(j)}}{(1+\varepsilon)V_df(\mathbf{x})}\leq F^{\text{inv}}\left(U_{(j)}\right)\leq\frac{U_{(j)}}{(1-\varepsilon)V_df(\mathbf{x})}. \tag{5.4}$$

Therefore, on the event $[U_{(k_n)}\leq F(\delta^d)]$, the generalized k-nearest neighbor estimate may be written as follows:

$$f_n(\mathbf{x})\stackrel{\mathcal{D}}{=}\frac{\theta f(\mathbf{x})}{n}\frac{\displaystyle\sum_{j=1}^{k_n}p_{nj}j}{\displaystyle\sum_{j=1}^{k_n}p_{nj}U_{(j)}},$$

where θ denotes some arbitrary random variable with values in $[1-\varepsilon, 1+\varepsilon]$. Observe that $F(\delta^d) > 0$ and, as $k_n/n \to 0$, $\mathbb{P}(U_{(k_n)} \leq F(\delta^d)) \to 1$ as $n \to \infty$ (see, e.g., Devroye et al. [18, Chapter 5]). Thus, to prove that $f_n(\mathbf{x}) \to f(\mathbf{x})$ in probability, it suffices to show that

$$\frac{\displaystyle\sum_{j=1}^{k_n} p_{nj} j}{n \displaystyle\sum_{j=1}^{k_n} p_{nj} U_{(j)}} \to 1 \quad \text{in probability.}$$

But, by Lemma 5.1, we know that

$$\left(U_{(1)}, \cdots, U_{(n)}\right) \stackrel{\mathcal{D}}{=} \left(\frac{\sum_{j=1}^{1} E_j}{n+1}, \cdots, \frac{\sum_{j=1}^{n} E_j}{n+1}\right)(1+\zeta_n),$$

where $E_1, \cdots, E_n$ are i.i.d. standard exponential random variables and $\zeta_n \to 0$ in probability. Consequently,

$$\frac{\displaystyle\sum_{j=1}^{k_n} p_{nj} j}{n \displaystyle\sum_{j=1}^{k_n} p_{nj} U_{(j)}} \stackrel{\mathcal{D}}{=} \frac{n+1}{n} \times \frac{\displaystyle\sum_{j=1}^{k_n} p_{nj} j}{\displaystyle\sum_{j=1}^{k_n} p_{nj}(E_1 + \cdots + E_j)} \times \frac{1}{1+\zeta_n},$$

which goes to 1 in probability as $k_n \to \infty$ according to the first statement of Lemma 5.2.

If $f(\mathbf{x}) = 0$, two cases are possible. Suppose first that $\mathbf{x}$ belongs to the complement of supp μ. Then, clearly, for some positive constant C and all $n \geq 1$, almost surely,

$$f_n(\mathbf{x}) \leq \frac{C k_n}{n}.$$

But $f(\mathbf{x}) = 0$ and, using the condition $k_n/n \to 0$, we deduce that $f_n(\mathbf{x}) \to f(\mathbf{x})$ in probability as $n \to \infty$.

If $\mathbf{x}$ belongs to supp μ, the proof is similar to the case $f(\mathbf{x}) > 0$. Just fix $\varepsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta} \left| \frac{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)} f(\mathbf{y}) \mathrm{d}\mathbf{y}}{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)} \mathrm{d}\mathbf{y}} \right| \leq \varepsilon.$$

5.3 Proof of Theorem 3.2

Choose $\mathbf{x}$ a Lebesgue point of f. Assume first that $f(\mathbf{x}) > 0$ and fix ε and δ as in (5.3). Note that

$$f_n^2(\mathbf{x}) = \frac{1}{n^2 V_d^2} \left(\frac{\sum_{j=1}^{k_n} p_{nj} j}{\sum_{j=1}^{k_n} p_{nj} \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|^d} \right)^2 .$$

Using

$$\frac{1}{k_n} \sum_{j=1}^{k_n} p_{nj} j \to \int_{[0,1]} t\nu(\mathrm{d}t)$$

and

$$\liminf_{n\to\infty} \sum_{\lceil k_n/2 \rceil}^{k_n} p_{nj} \geq \int_{[\frac{1}{2},1]} \nu(\mathrm{d}t),$$

we have, for some positive constant C_1 and all $n \geq 1$,

$$\mathbb{E}\left[f_n^2(\mathbf{x})\right] \leq \frac{C_1 k_n^2}{n^2} \mathbb{E}\left[\frac{1}{\|\mathbf{X}_{(\lceil k_n/2 \rceil)}(\mathbf{x}) - \mathbf{x}\|^{2d}} \right].$$

If $U_{(1)} \leq \cdots \leq U_{(n)}$ are uniform $[0,1]$ order statistics, we may write, using inequality (5.4),

$$\mathbb{E}\left[\frac{1}{\|\mathbf{X}_{(\lceil k_n/2 \rceil)}(\mathbf{x}) - \mathbf{x}\|^{2d}} \right] \leq C_2 \left(\mathbb{E}\left[\frac{1}{U^2_{(\lceil k_n/2 \rceil)}} \right] + \frac{1}{\delta^{2d}} \right)$$

for some positive constant C_2. It is known (see, e.g., Devroye [17, Chapter 1]) that $U_{(\lceil k_n/2 \rceil)}$ is beta distributed, with parameters $\lceil k_n/2 \rceil$ and $n+1-\lceil k_n/2 \rceil$. Consequently, for $\lceil k_n/2 \rceil > 2$,

$$\mathbb{E}\left[\frac{1}{\|\mathbf{X}_{(\lceil k_n/2 \rceil)}(\mathbf{x}) - \mathbf{x}\|^{2d}} \right] \leq C_3 \left(\frac{n^2}{k_n^2} + \frac{1}{\delta^{2d}} \right),$$

whence, for $k_n \geq 5$,

$$\mathbb{E}\left[f_n^2(\mathbf{x})\right] \leq C_4$$

for some positive constant C_4.

Next, if $f(\mathbf{x}) = 0$, two cases are possible. If $\mathbf{x}$ belongs to the complement of supp μ, then, clearly, for some positive constant C_5 and all $n \geq 1$,

$$\mathbb{E}\left[f_n^2(\mathbf{x})\right] \leq \frac{C_5 k_n^2}{n^2} \leq C_5.$$

If $\mathbf{x}$ belongs to supp μ, the proof is similar to the case $f(\mathbf{x}) > 0$. Just fix $\varepsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta} \left| \frac{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)} f(\mathbf{y})\mathrm{d}\mathbf{y}}{\displaystyle\int_{\mathcal{B}(\mathbf{x},r)} \mathrm{d}\mathbf{y}} \right| \leq \varepsilon.$$

This shows the first part of the theorem. One proves, with similar arguments, that there exists a positive constant C_6 such that, for all n large enough,

$$\mathbb{E}\left[f_n^3(\mathbf{x})\right] \leq C_6.$$

Consequently, for all n large enough, the sequence $\{f_n^2(\mathbf{x})\}$ is uniformly integrable and, since $f_n(\mathbf{x}) - f(\mathbf{x}) \to 0$ in probability (by Theorem 3.1), this implies $\mathbb{E}\left[f_n(\mathbf{x}) - f(\mathbf{x})\right]^2 \to 0$ as $n \to \infty$ (see, e.g., Billingsley [6, Chapter 5]).

5.4 Proof of Theorem 3.3

Fix $\mathbf{x} \in \mathbb{R}^d$ and assume that f has derivatives of second order at $\mathbf{x}$, with $f(\mathbf{x}) > 0$. Let

$$G(u) = \mathbb{P}\left(\|\mathbf{X} - \mathbf{x}\| \leq u\right) = \int_{\mathcal{B}(\mathbf{x},u)} f(\mathbf{y})\mathrm{d}\mathbf{y}$$

be the univariate distribution function of $\|\mathbf{X} - \mathbf{x}\|$. We may write, by a Taylor-Young expansion of f around $\mathbf{x}$,

$$\begin{aligned} G(u) = V_d f(\mathbf{x}) u^d &+ \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right]^T \int_{\mathcal{B}(\mathbf{x},u)} (\mathbf{y} - \mathbf{x})\mathrm{d}\mathbf{y} \\ &+ \frac{1}{2} \int_{\mathcal{B}(\mathbf{x},u)} (\mathbf{y} - \mathbf{x})^T \left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\right] (\mathbf{y} - \mathbf{x})\mathrm{d}\mathbf{y} + \mathrm{o}\left(u^{d+2}\right) \quad \text{as } u \to 0, \end{aligned} \tag{5.5}$$

where the symbol T denotes transposition and $[\partial f(\mathbf{x})/\partial \mathbf{x}]$ and $[\partial^2 f(\mathbf{x})/\partial \mathbf{x}^2]$ are a vector and a matrix given by

$$\left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right] = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \cdots, \frac{\partial f(\mathbf{x})}{\partial x_d}\right)^T$$

and

$$\left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\right]_{i,j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}.$$

In view of the symmetry of the ball $\mathcal{B}(\mathbf{x}, u)$, the first term in (5.5) is seen to be zero. Using the linearity of trace and relations $\mathrm{tr}(A\mathbf{Z}\mathbf{Z}^T) = \mathbf{Z}^T A\mathbf{Z}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for matrices A, B and vector $\mathbf{Z}$, (5.5) becomes

$$G(u) = V_d f(\mathbf{x}) u^d + \frac{1}{2}\mathrm{tr}\left\{\left[\int_{\mathcal{B}(\mathbf{x},u)} (\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^T \mathrm{d}\mathbf{y}\right]\left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\right]\right\} + \mathrm{o}\left(u^{d+2}\right).$$

Letting $\mathbf{z} = (\mathbf{y}-\mathbf{x})/u$, that maps $\mathcal{B}(\mathbf{x}, u)$ to $\mathcal{B}(\mathbf{0}, 1)$, and using a hyperspherical coordinate change of variables (see, e.g., Miller [35, Chapter 1]), the integral inside the trace term simplifies to

$$\int_{\mathcal{B}(\mathbf{0},1)} u^2 \mathbf{z}\mathbf{z}^T u^d \mathrm{d}\mathbf{z} = \left[\frac{V_d}{d+2} u^{d+2}\right] \mathrm{Id},$$

where Id is the $d \times d$ identity matrix. Thus, denoting by $\Gamma(.)$ the gamma function and recalling that, for the Euclidean norm,

$$V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)},$$

we obtain

$$G(u) = V_d f(\mathbf{x}) u^d + c(\mathbf{x}) V_d^{1+2/d} u^{d+2} + \mathrm{o}\left(u^{d+2}\right) \quad \text{as } u \to 0,$$

where

$$c(\mathbf{x}) = \frac{1}{2(d+2)\pi} \Gamma^{2/d}\left(\frac{d+2}{2}\right) \mathrm{tr}\left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\right].$$

Consequently,

$$G^{\mathrm{inv}}(u) = \frac{1}{V_d^{1/d} f^{1/d}(\mathbf{x})} u^{1/d} - \frac{c(\mathbf{x})}{d V_d^{1/d} f^{1+3/d}(\mathbf{x})} u^{3/d} + \mathrm{o}\left(u^{3/d}\right) \quad \text{as } u \to 0$$

and

$$\left[G^{\mathrm{inv}}(u)\right]^d = \frac{1}{V_d f(\mathbf{x})} u - \frac{c(\mathbf{x})}{V_d f^{2+2/d}(\mathbf{x})} u^{1+2/d} + \mathrm{o}\left(u^{1+2/d}\right) \quad \text{as } u \to 0.$$

Let F be the univariate distribution function of $W \stackrel{\mathrm{def}}{=} \|\mathbf{X}-\mathbf{x}\|^d$. Clearly,

$$F^{\mathrm{inv}}(u) = \left[G^{\mathrm{inv}}(u)\right]^d.$$

Define $W_j = \|\mathbf{X}_j - \mathbf{x}\|^d$, $j = 1, \cdots, n$, and let $W_{(1)} \leq \cdots \leq W_{(n)}$ be the order statistics for $W_1, \cdots, W_n$. If $U_{(1)} \leq \cdots \leq U_{(n)}$ are uniform $[0,1]$ order statistics, using the representation

$$W_{(j)} \stackrel{\mathcal{D}}{=} F^{\text{inv}}\left(U_{(j)}\right)$$

jointly for all j, we may write

$$f_n(\mathbf{x}) \stackrel{\mathcal{D}}{=} \frac{1}{n} \frac{\sum_{j=1}^{k_n} p_{nj} j}{f^{-1}(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} U_{(j)} + c'(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} U_{(j)}^{1+2/d} + \sum_{j=1}^{k_n} p_{nj} \mathrm{o}\left(U_{(j)}^{1+2/d}\right)},$$

where

$$c'(\mathbf{x}) = -\frac{c(\mathbf{x})}{f^{2+2/d}(\mathbf{x})}.$$

Thus,

$$f_n^{-1}(\mathbf{x}) \stackrel{\mathcal{D}}{=} n \left(\frac{f^{-1}(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} U_{(j)}}{\sum_{j=1}^{k_n} p_{nj} j} + \frac{c'(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} U_{(j)}^{1+2/d}}{\sum_{j=1}^{k_n} p_{nj} j} + \frac{\sum_{j=1}^{k_n} p_{nj} \mathrm{o}\left(U_{(j)}^{1+2/d}\right)}{\sum_{j=1}^{k_n} p_{nj} j} \right).$$

Consequently, by Lemma 5.1, letting $E_1, \cdots, E_{n+1}$ be i.i.d. standard exponential random variables and

$$V_{(j)} = \frac{\sum_{i=1}^{j} E_i}{\sum_{i=1}^{n+1} E_i},$$

we obtain

$$f_n^{-1}(\mathbf{x}) \stackrel{\mathcal{D}}{=} \frac{f^{-1}(\mathbf{x})\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)}{\sum_{j=1}^{k_n} p_{nj}j}(1+\zeta_{n1})$$
$$+\frac{c'(\mathbf{x})\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)^{1+2/d}}{n^{2/d}\sum_{j=1}^{k_n} p_{nj}j}(1+\zeta_{n2})$$
$$+\frac{n\sum_{j=1}^{k_n} p_{nj}\mathrm{o}\left(V_{(j)}^{1+2/d}\right)}{\sum_{j=1}^{k_n} p_{nj}j}.$$

Besides, for $j=1,2$, $\zeta_{nj}=\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ and, for all positive integers r,

$$\limsup_{n\to\infty}\left[n^{r/2}\mathbb{E}|\zeta_{nj}|^r\right]<\infty.$$

On the one hand, using the second statement of Lemma 5.2 and the identity

$$\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj}j=\int_{[0,1]} t\nu(\mathrm{d}t)\left(1+\mathcal{O}\left(\frac{1}{k_n}\right)\right)\quad\text{as } k_n\to\infty,$$

we may write, on an appropriate probability space,

$$\frac{f^{-1}(\mathbf{x})\sum_{j=1}^{k_n} p_{nj}(E_1+\cdots+E_j)}{\sum_{j=1}^{k_n} p_{nj}j}=f^{-1}(\mathbf{x})+\frac{f^{-1}(\mathbf{x})v}{\sqrt{k_n}}N+\zeta_{n3},$$

where N is a standard normal random variable,

$$v^2=\frac{\int_0^1\left(1-\Phi(t)\right)^2\mathrm{d}t}{\left[\int_{[0,1]} t\nu(\mathrm{d}t)\right]^2},$$

and $\zeta_{n3} = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2})$ with, for all positive integers r,

$$\sup_{n\geq 1}\left[k_n^{r/2}\mathbb{E}|\zeta_{n3}|^r\right] < \infty.$$

Next, recalling that, for $\rho \geq 1$,

$$\frac{1}{k_n^\rho}\sum_{j=1}^{k_n} p_{nj} j^\rho = \int_{[0,1]} t^\rho \nu(\mathrm{d}t)\left(1 + \mathcal{O}\left(\frac{1}{k_n}\right)\right),$$

and applying the first statement of Lemma 5.2, we obtain

$$\frac{c'(\mathbf{x})\sum_{j=1}^{k_n} p_{nj}(E_1 + \cdots + E_j)^{1+2/d}}{n^{2/d}\sum_{j=1}^{k_n} p_{nj} j} = c'(\mathbf{x}) b \left(\frac{k_n}{n}\right)^{2/d} + \left(\frac{k_n}{n}\right)^{2/d}\zeta_{n4},$$

where

$$b = \frac{\displaystyle\int_{[0,1]} t^{1+2/d}\nu(\mathrm{d}t)}{\displaystyle\int_{[0,1]} t\nu(\mathrm{d}t)}$$

and $\zeta_{n4} = \mathrm{o}_{\mathbb{P}}(1)$ with, for all positive integers r,

$$\sup_{n\geq 1}\mathbb{E}|\zeta_{n4}|^r < \infty.$$

Similarly,

$$\left|\frac{n\displaystyle\sum_{j=1}^{k_n} p_{nj}\mathrm{o}\left(V_{(j)}^{1+2/d}\right)}{\displaystyle\sum_{j=1}^{k_n} p_{nj} j}\right| \leq \frac{(E_1 + \cdots + E_{k_n})^{1+2/d}}{n^{2/d}\displaystyle\sum_{j=1}^{k_n} p_{nj} j} \times \frac{\mathrm{o}\left(V_{(k_n)}^{1+2/d}\right)}{V_{(k_n)}^{1+2/d}} \times (1 + \zeta_{n5}),$$

where $\zeta_{n5} = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ and, for all positive integers r,

$$\limsup_{n\to\infty}\left[n^{r/2}\mathbb{E}|\zeta_{n5}|^r\right] < \infty.$$

Thus,

$$\left|\frac{n\displaystyle\sum_{j=1}^{k_n} p_{nj}\mathrm{o}\left(V_{(j)}^{1+2/d}\right)}{\displaystyle\sum_{j=1}^{k_n} p_{nj} j}\right| \leq \left(\frac{k_n}{n}\right)^{2/d}\zeta_{n6},$$

where $\zeta_{n6} = \mathrm{o}_{\mathbb{P}}(1)$. Moreover, we clearly have, for some $\varepsilon_0 \in (0,1)$ and all $r > 0$,

$$\limsup_{n\to\infty} \mathbb{E}\left[|\zeta_{n6}|^r \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}\right] < \infty.$$

Thus, putting all the pieces together, we obtain

$$f_n^{-1}(\mathbf{x}) \stackrel{\mathcal{D}}{=} f^{-1}(\mathbf{x}) + \frac{f^{-1}(\mathbf{x})v}{\sqrt{k_n}} N + c'(\mathbf{x})b\left(\frac{k_n}{n}\right)^{2/d} + \zeta_{n7} + \left(\frac{k_n}{n}\right)^{2/d} \zeta_{n8},$$

where $\zeta_{n7} = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2})$ and $\zeta_{n8} = \mathrm{o}_{\mathbb{P}}(1)$. Besides, for all positive integers r,

$$\limsup_{n\to\infty} \left[k_n^{r/2}\mathbb{E}|\zeta_{n7}|^r\right] < \infty \quad \text{and} \quad \limsup_{n\to\infty} \mathbb{E}\left[|\zeta_{n8}|^r \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}\right] < \infty.$$

We see in particular that, for all positive integers r and all n large enough, the sequence $\{k_n^{r/2}\zeta_{n7}^r\}$ is uniformly integrable and, consequently, that $\mathbb{E}|\zeta_{n7}|^r = \mathrm{o}(k_n^{-r/2})$ (see, e.g., Billingsley [6, Chapter 5]). Likewise, $\mathbb{E}[|\zeta_{n8}|^r \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}] = \mathrm{o}(1)$. It follows that

$$f_n^{-1}(\mathbf{x}) \stackrel{\mathcal{D}}{=} f^{-1}(\mathbf{x}) + \frac{f^{-1}(\mathbf{x})v}{\sqrt{k_n}} N + c'(\mathbf{x})b\left(\frac{k_n}{n}\right)^{2/d} + \zeta_{n9},$$

where $\zeta_{n9} = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2} + (k_n/n)^{2/d})$ and

$$\mathbb{E}\left[|\zeta_{n9}|^r \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}\right] = \mathrm{o}\left(\frac{1}{k_n^{r/2}} + \left(\frac{k_n}{n}\right)^{2r/d}\right) \tag{5.6}$$

as $k_n \to \infty$ and $k_n/n \to 0$. Note that, by definition, $f_n^{-1}(\mathbf{x})$ is almost surely finite and positive. Therefore, setting

$$T_n(\mathbf{x}) = \frac{v}{\sqrt{k_n}} N + f(\mathbf{x})c'(\mathbf{x})b\left(\frac{k_n}{n}\right)^{2/d} + f(\mathbf{x})\zeta_{n9}$$

and using the identity

$$\frac{1}{1+t} = 1 - t + \frac{t^2}{1+t}$$

valid for $t \neq -1$, we finally get

$$f_n(\mathbf{x}) \stackrel{\mathcal{D}}{=} f(\mathbf{x}) - \frac{f(\mathbf{x})v}{\sqrt{k_n}} N + \frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})}\left(\frac{k_n}{n}\right)^{2/d} + \zeta_{n10} + \frac{f(\mathbf{x})T_n^2(\mathbf{x})}{1 + T_n(\mathbf{x})},$$

where $\zeta_{n10} = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2} + (k_n/n)^{2/d})$ and

$$\mathbb{E}\left[\zeta_{n10}^2 \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}\right] = \mathrm{o}\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{4/d}\right).$$

Clearly,

$$\frac{T_n^2(\mathbf{x})}{1 + T_n(\mathbf{x})} = \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k_n}} + \left(\frac{k_n}{n}\right)^{2/d}\right).$$

Next, observing that

$$\mathbb{E}\left[\frac{1}{1 + T_n(\mathbf{x})}\right]^4 = f^{-4}(\mathbf{x})\mathbb{E}[f_n^4(\mathbf{x})],$$

it follows from an immediate adaptation of the proof of Theorem 3.2 that

$$\limsup_{n \to \infty} \mathbb{E}\left[\frac{1}{1 + T_n(\mathbf{x})}\right]^4 < \infty.$$

Thus, using the Cauchy-Schwarz inequality and (5.6), we see that

$$\mathbb{E}\left[\left(\frac{T_n^2(\mathbf{x})}{1 + T_n(\mathbf{x})}\right)^2 \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}\right] = \mathrm{o}\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{4/d}\right).$$

In conclusion,

$$f_n(\mathbf{x}) \stackrel{\mathcal{D}}{=} f(\mathbf{x}) - \frac{f(\mathbf{x})v}{\sqrt{k_n}} N + \frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})}\left(\frac{k_n}{n}\right)^{2/d} + \zeta_n, \tag{5.7}$$

where $\zeta_n = \mathrm{o}_{\mathbb{P}}(k_n^{-1/2} + (k_n/n)^{2/d})$, as desired. In addition,

$$\mathbb{E}\left[\zeta_n^2 \mathbf{1}_{[V_{(k_n)} \leq \varepsilon_0]}\right] = \mathrm{o}\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{4/d}\right). \tag{5.8}$$

5.5 Proof of Theorem 3.4

An immediate adaptation of the proof of Theorem 3.2 shows that for some positive constant C_1 and all n large enough,

$$\mathbb{E}\left[f_n(\mathbf{x}) - f(\mathbf{x})\right]^4 \leq C_1.$$

It follows, coming back to identity (5.7), that for some positive constant C_2 and all n large enough,

$$\mathbb{E}\zeta_n^4 \leq C_2.$$

Therefore, using identity (5.8) and the Cauchy-Schwarz inequality, for all n large enough,

$$\mathbb{E}\zeta_n^2 \leq \sqrt{C_2}\sqrt{\mathbb{P}\left(V_{(k_n)} > \varepsilon_0\right)} + \mathrm{o}\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{4/d}\right).$$

We know that $V_{(k_n)}$ is beta distributed, with parameters k_n and $n+1-k_n$ (see, e.g., Devroye [17, Chapter 1]). Thus, by Markov's inequality,

$$\mathbb{P}\left(V_{(k_n)} > \varepsilon_0\right) = \mathbb{P}\left(V_{(k_n)}^{9/d} > \varepsilon_0^{9/d}\right) = \mathcal{O}\left(\left(\frac{k_n}{n}\right)^{9/d}\right)$$

and, consequently,

$$\sqrt{\mathbb{P}\left(V_{(k_n)} > \varepsilon_0\right)} = \mathrm{o}\left(\left(\frac{k_n}{n}\right)^{4/d}\right)$$

as $k_n/n \to 0$. In conclusion, as $k_n \to \infty$ and $k_n/n \to 0$,

$$\mathbb{E}\zeta_n^2 = \mathrm{o}\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{4/d}\right),$$

and squaring and taking the expectation on both sides of identity (5.7) leads to the desired statement. The last assertion of Theorem 3.4 is clear.

Acknowledgments. We thank an Associate Editor and a referee for valuable comments and insightful suggestions. We also thank David Mason for stimulating discussions on the KMT approximations.

References

[1] E. Arias-Castro, D.L. Donoho, and X. Huo. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *The Annals of Statistics*, 34:326–349, 2006.

[2] M.S. Bartlett. Statistical estimation of density functions. *Sankhyā. Series A*, 25:245–254, 1963.

[3] P.K. Bhattacharya and Y.P. Mack. Weak convergence of k-NN density and regression estimators with varying k and applications. *The Annals of Statistics*, 15:976–994, 1987.

[4] G. Biau, B. Cadre, D.M. Mason, and B. Pelletier. Asymptotic normality in density support estimation. *Electronic Journal of Probability*, 14:2617–2635, 2009.

[5] G. Biau, B. Cadre, and B. Pelletier. Exact rates in density support estimation. *Journal of Multivariate Analysis*, 99:2185–2207, 2008.

[6] P. Billingsley. *Probability and Measure. Third Edition.* John Wiley and Sons, New York, 1995.

[7] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.

[8] CGAL. Computational Geometry Algorithms Library. `http://www.cgal.org`.

[9] F. Chazal, D. Cohen-Steiner, and A. Lieutier. Normal cone approximation and offset shape isotopy. *Computational Geometry: Theory and Applications*, 42:566–581, 2009.

[10] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in Euclidean space. *Discrete and Computational Geometry*, 41:461–479, 2009.

[11] F. Chazal, D. Cohen-Steiner, A. Lieutier, and B. Thibert. Stability of curvature measures. *Computer Graphics Forum*, 28:1485–1496, 2009.

[12] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for measures based on distance functions. Research Report 6930, INRIA, 2009. `http://hal.inria.fr/inria-00383685`.

[13] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Boundary measures for geometric inference. *Journal on Foundations of Computational Mathematics*, 10:221–240, 2010.

[14] A. Cuevas, R. Fraiman, and A. Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *The Annals of Statistics*, 35:1031–1051, 2007.

[15] A. Cuevas and A. Rodríguez-Casal. Set estimation: An overview and some recent developments. In M.G. Akritas and D.N. Politis, editors, *Recent Advances and Trends in Nonparametric Statistics*, pages 251–264, Amsterdam, 2003. North-Holland.

[16] P. Deheuvels. Estimation non paramétrique de la densité par histogrammes généralisés. *Publications de l'Institut de Statistique de l'Université de Paris*, 22:1–24, 1977.

[17] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[18] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[19] L. Devroye and G. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38:480–488, 1980.

[20] L.P. Devroye and T.J. Wagner. *Nonparametric discrimination and density estimation*. Technical Report 183, Information Systems Research Laboratory, Electronics Research Center, The University of Texas at Austin, Austin, 1976.

[21] L.P. Devroye and T.J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, 5:536–540, 1977.

[22] V.A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14:153–158, 1969.

[23] E. Fix and J.L. Hodges. *Discriminatory analysis. Nonparametric discrimination: Consistency properties*. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[24] K. Fukunaga and T.E. Flick. An optimal global nearest neighbor metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:314–318, 1984.

[25] K. Fukunaga and L.D. Hostetler. Optimization of k-nearest neighbor density estimates. *IEEE Transactions on Information Theory*, 19:320–326, 1973.

[26] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. On the path density of a gradient field. *The Annals of Statistics*, 37:3236–3271, 2009.

[27] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. *Nonparametric filament Estimation*, 2010. `http://arxiv.org/abs/1003.5536`.

[28] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[29] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV'-s, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32:111–131, 1975.

[30] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

[31] D.O. Loftsgaarden and C.P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36:1049–1051, 1965.

[32] Y.P. Mack. Asymptotic normality of multivariate k-NN density estimates. *Sankhyā. Series A*, 42:53–63, 1980.

[33] Y.P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9:1–15, 1979.

[34] D.M. Mason. *Coupling via the KMT quantile inequality*. Technical Report, University of Delaware, Newark, 2011.

[35] K.S. Miller. *Multidimensional Gaussian Distributions*. Wiley, New York, 1964.

[36] D.S. Moore and J.W. Yackel. Consistency properties of nearest neighbor density function estimators. *The Annals of Statistics*, 5:143–154, 1977.

[37] D.S. Moore and J.W. Yackel. Large sample properties of nearest neighbor density function estimators. In S.S. Gupta and D.S. Moore, editors, *Statistical Decision Theory and Related Topics. II*, pages 269–279, New York, 1977. Academic Press.

[38] D.M. Mount and S. Arya. ANN: A Library for Approximate Nearest Neighbor Searching. `http://www.cs.umd.edu/~mount/ANN`.

[39] P. Niyogi, S. Smale, and S. Weinberger. *A topological view of unsupervised learning from noisy data*. Technical Report, University of Chicago, Chicago, 2008. `http://www.math.uchicago.edu/~shmuel/noise.pdf`.

[40] A. Petrunin. Semiconcave functions in Alexandrov's geometry. In *Surveys in Differential Geometry. 11*, pages 137–201. International Press, Somerville, 2007.

[41] C. Rodríguez and J. Van Ryzin. Maximum entropy histograms. *Statistics and Probability Letters*, 3:117–120, 1985.

[42] C. Rodríguez and J. Van Ryzin. Large sample properties of maximum entropy histograms. *IEEE Transactions on Information Theory*, 32:751–759, 1986.

[43] C.C. Rodríguez. *On a new class of density estimators*. Technical Report, Department of Mathematics and Statistics, State University of New York at Albany, New York, 1986. `http://omega.albany.edu:8008/npde.ps`.

[44] C.C. Rodríguez. *Weak convergence of multivariate k-nn*. Technical Report, Department of Mathematics and Statistics, State University of New York at Albany, New York, 1992. `http://omega.albany.edu:8008`.

[45] C.C. Rodríguez. Optimal recovery of local truth. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 19th International Workshop*, volume 567, pages 80–115. AIP Conference Proceedings, 2001.

[46] R.J. Samworth. *Optimal weighted nearest neighbour classifiers*. Technical Report, University of Cambridge, Cambridge, 2011. `http://arxiv.org/abs/1101.5783`.

[47] R.D. Short and K. Fukunaga. The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory*, 27:622–627, 1981.

[48] USGS. US Geological Survey. `http://earthquake.usgs.gov/earthquakes/eqarchives/epic`.

[49] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer-Verlag, New York, 1996.

[50] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, 2003.

[51] R.L. Wheeden and A. Zygmund. *Measure and Integral. An Introduction to Real Analysis*. Marcel Dekker, New York, 1977.

[52] R. Willink. Relationships between central moments and cumulants, with formulae for the central moments of gamma distributions. *Communications in Statistics - Theory and Methods*, 32:701–704, 2003.