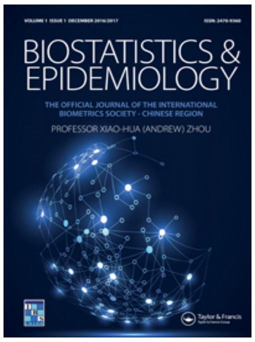

Taylor & Francis
Taylor & Francis Group

Biostatistics & Epidemiology

ISSN: 2470-9360 (Print) 2470-9379 (Online) Journal homepage: https://www.tandfonline.com/loi/tbep20

A tutorial on kernel density estimation and recent advances

Yen-Chi Chen

To cite this article: Yen-Chi Chen (2017) A tutorial on kernel density estimation and recent advances, Biostatistics & Epidemiology, 1:1, 161-187, DOI: 10.1080/24709360.2017.1396742

To link to this article: https://doi.org/10.1080/24709360.2017.1396742

BIOSTATISTICS & EPIDEMIOLOGY, 2017
VOL. 1, NO. 1, 161–187
https://doi.org/10.1080/24709360.2017.1396742

Taylor & Francis
Taylor & Francis Group

A tutorial on kernel density estimation and recent advances

Yen-Chi Chen

Department of Statistics, University of Washington, Seattle, WA, USA

ABSTRACT

This tutorial provides a gentle introduction to kernel density estimation (KDE) and recent advances regarding confidence bands and geometric/topological features. We begin with a discussion of basic properties of KDE: the convergence rate under various metrics, density derivative estimation, and bandwidth selection. Then, we introduce common approaches to the construction of confidence intervals/bands, and we discuss how to handle bias. Next, we talk about recent advances in the inference of geometric and topological features of a density function using KDE. Finally, we illustrate how one can use KDE to estimate a cumulative distribution function and a receiver operating characteristic curve. We provide R implementations related to this tutorial at the end.

ARTICLE HISTORY
Received 12 April 2017
Accepted 17 August 2017

KEYWORDS
Kernel density estimation; nonparametric statistics; confidence bands; bootstrap

1. Introduction

Kernel density estimation (KDE), also known as the Parzen's window [1], is one of the most well-known approaches to estimate the underlying probability density function of a data-set. KDE is a nonparametric density estimator requiring no assumption that the underlying density function is from a parametric family. KDE will learn the shape of the density from the data automatically. This flexibility arising from its nonparametric nature makes KDE a very popular approach for data drawn from a complicated distribution.

Figure 1 illustrates KDE using a part of the NACC (National Alzheimer's Coordinating Center) Uniform Data-Set [2], version 3.0 (March 2015). Because the purpose of using this data-set is to illustrate the effectiveness of KDE, we will draw no scientific conclusion but will just use KDE as a tool to explore the pattern of the data. We focus on two variables, 'CRAFTDTI' (Craft Story 21 Recall – delay time), and 'CRAFTDVR' (Craft Story 21 Recall – total story units recalled, verbatim scoring). Although these two variables take integer values, we treat them as continuous and use KDE to determine the density function. We consider only the unique subject with scores on both variables, resulting in a sample of size 4,044. In the left panel of Figure 1, we display the estimated density function of 'CRAFTDTI' using KDE. We see that there are two modes in the distribution. In the right panel of Figure 1, we show the scatter plot of the data and overlay it with the result of bivariate KDE (blue contours). Because many subjects have identical values for the two variables, the scatter plot (gray dots) provides no useful information regarding

CONTACT Yen-Chi Chen yenchic@uw.edu

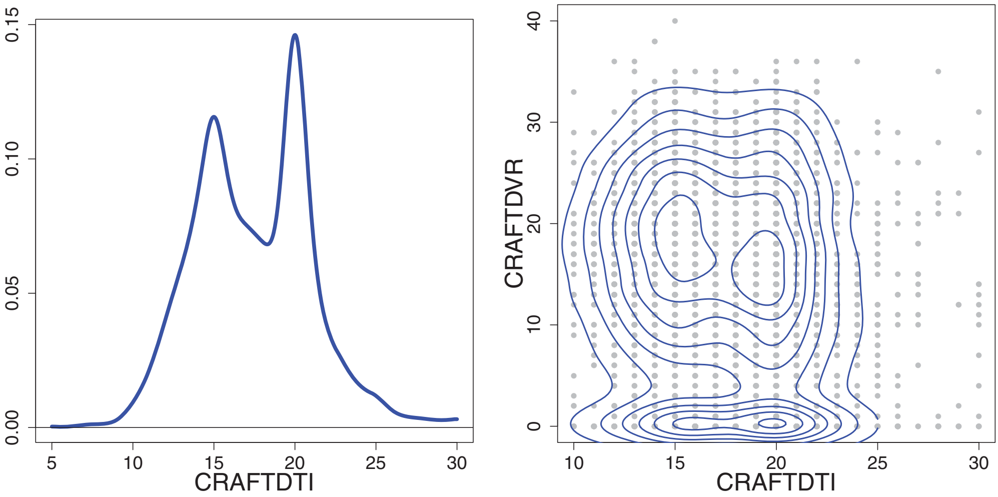

Figure 1. Examples of KDE using the NACC Uniform Data-Set. We focus on variables 'CRAFTDTI' and 'CRAFTDVR' and use those subjects who have non-missing value of either variables. Left: We show the marginal density function of variable 'CRAFTDTI' from one-dimensional (1D) KDE. There are two bumps in for this density function. Right: We show the scatter plot along with the bivariate density function of both variables from the two-dimensional (2D) KDE, showing the multi-modality feature of this bivariate density function.

the underlying distribution. However, KDE shows the multi-modality of this bivariate distribution, which contains multiple bumps that cannot be captured easily by any parametric distribution.

The remainder of the tutorial is organized as follows. In Section 2, we present the definition of KDE, followed by a discussion of its basic properties: convergence rates, density derivative estimations, and bandwidth selection. Then, in Section 3, we introduce common approaches to the construction of confidence regions, and we discuss the problem of bias in statistical inference. Section 4 provides an introduction to the use of KDE to estimate geometric and topological features of a density function. In Section 5, we study how one can use KDE to estimate the cumulative distribution function (CDF) and the receiver operating characteristic (ROC) curve. Finally, in Section 6, we discuss open problems. At the end of this tutorial, we provide R codes for implementing the presented analysis of KDE.

2. Statistical properties

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be an independent, identically distributed random sample from an unknown distribution P with density function p. Formally, KDE can be expressed as

$$\widehat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $K : \mathbb{R}^d \mapsto \mathbb{R}$ is a smooth function called the kernel function and $h > 0$ is the smoothing bandwidth that controls the amount of smoothing. Two common examples of

$K(x)$ are

$$(\text{Gaussian kernel}) \quad K(x) = \frac{\exp(-\| x \|^2/2)}{v_{1,d}}, \quad v_{1,d} = \int \exp(-\| x \|^2/2)dx,$$

$$(\text{Spherical kernel}) \quad K(x) = \frac{I(\| x \| \le 1)}{v_{2,d}}, \quad v_{2,d} = \int I(\| x \| \le 1)dx.$$

Note that we apply the same amount of smoothing h in every direction; in practice, one can use a bandwidth matrix H and the quantity $K\left(\frac{x-X_i}{h}\right)$ becomes $K(H^{-1}(x - X_i))$.

Intuitively, KDE has the effect of smoothing out each data point into a smooth bump, whose the shape is determined by the kernel function $K(x)$. Then, KDE sums over all these bumps to obtain a density estimator. At regions with many observations, because there will be many bumps around, KDE will yield a large value. On the other hand, for regions with only a few observations, the density value from summing over the bumps is low, because only have a few bumps contribute to the density estimate.

Figure 2 presents examples of KDE in the 1D case. There are six observations, as indicated by the black lines. We smooth these observations into bumps (red bumps) and sum over all of the bumps to form the final density estimator (the blue curve). In R, many packages are equipped with programs for computing KDE; see [3] for a listing.

Remark 2.1 (Adaptive smoothing): The amount of smoothing can depend on the location x [4] or the data point X_i [5]. In the former case, we use $h = h(x)$ so KDE becomes $\widehat{p}_n(x) = \frac{1}{nh^d(x)} \sum_{i=1}^n K\left(\frac{x-X_i}{h(x)}\right)$, which is referred to as the balloon estimator [6]. In the latter case, we use $h = h_i = h(X_i)$ for the ith data points with the resulting density estimate being $\widehat{p}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} K\left(\frac{x-X_i}{h_i}\right)$, which is referred to as the sample smoothing estimator [6]. For more details regarding adaptive smoothing, we refer the readers to Section 6.6 of [7].

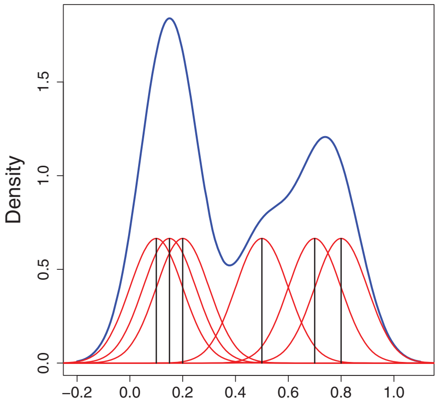

Figure 2. A 1D illustration of how KDE is constructed. There are six observations, located at the positions indicated by black lines. We then smooth these observations into small bumps (red bumps) and the sum them to obtain the density estimate (blue curve). (colour online.)

2.1. Convergence rate

To measure the errors of KDE, we consider three types of errors: the pointwise error, uniform error, and mean integrated square error (MISE). The pointwise error is the simplest error and is related to the confidence interval (Section 3.1). The uniform error has many useful theoretical properties since it measures the uniform deviation of the estimator and can be used to bound other types of errors. The uniform error is related to the confidence band and the geometric (and topological) features; see Section 3.2 and Section 4. The MISE (actually it is a risk measurement of the estimator) is generally used in bandwidth selection (Section 2.3), because it measures the overall performance of the estimator and is related to the mean square error.

Pointwise error. For a given point x, the pointwise error of KDE is the difference between KDE $\widehat{p}_n(x)$ and $p(x)$, the true density function evaluated at x. Let $\nabla^2 p = \sum_{\ell=1}^{d} \frac{\partial^2 p}{\partial x_\ell^2}$ be the Laplacian of the function p. Under smoothness conditions [7–9],

$$\begin{aligned}
\widehat{p}_n(x) - p(x) &= \underbrace{\mathbb{E}\big(\widehat{p}_n(x)\big) - p(x)}_{B_h(x)} + \underbrace{\widehat{p}_n(x) - \mathbb{E}\big(\widehat{p}_n(x)\big)}_{\mathcal{E}_n(x)} \\
&= O\big(h^2\big) + O_P\left(\sqrt{\frac{1}{nh^d}}\right), \\
B_h(x) &= \frac{h^2}{2}\sigma_K^2 \nabla^2 p(x) + o\big(h^2\big), \\
\mathcal{E}_n(x) &= \sqrt{\frac{\mu_K \cdot p(x)}{nh^d}} \cdot Z_n(x) + o_P\left(\sqrt{\frac{1}{nh^d}}\right),
\end{aligned} \tag{2}$$

where $Z_n(x)\vec{D}N(0,1)$ and $\sigma_K^2 = \int \|x\|^2 K(x)dx$, $\mu_K = \int K^2(x)dx$ are constants depending only on the kernel function K. Thus, when $h \to 0$ and $nh^d \to \infty$, $\widehat{p}_n(x)\vec{P}p(x)$, i.e.KDE $\widehat{p}_n(x)$ is a consistent estimator of $p(x)$. Equation (2) presents the decomposition of the (pointwise) estimation error of KDE in terms of the bias $B_h(x)$ and the stochastic variation $\mathcal{E}_n(x)$. This decomposition will be used frequently in deriving other errors and constructing the confidence regions.

Uniform error. Another error metric is the uniform error (also known as the L_∞ error), the maximal difference between $\widehat{p}_n$ and p: $\sup_x|\widehat{p}_n(x) - p(x)|$. According to empirical process theory [10–13], the uniform error

$$\sup_x |\widehat{p}_n(x) - p(x)| = O\big(h^2\big) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right) \tag{3}$$

under mild conditions (see [11,12] for more details). The error rate in (3) and the pointwise error rate in (2) differ only in the stochastic variation portion and the difference is at the rate $\sqrt{\log n}$. The presence of an extra $\sqrt{\log n}$ in the uniform error rate is a very common phenomenon in nonparametric estimation owing to empirical process theory. The uniform error has many useful theoretical properties [14–17], because it provides a uniform control of the estimation error over the entire support.

MISE. The MISE is one of the most well-known error measurements [7,8] among all of the error measures used in KDE. The MISE is defined as $\int \mathbb{E}\left(\left(\widehat{p}_n(x) - p(x)\right)^2\right)dx$. Thus, the MISE measures the L_2 risk of KDE. Under regularity conditions, the MISE

$$\begin{aligned}\int \mathbb{E}\left(\left(\widehat{p}_n(x) - p(x)\right)^2\right)dx &= \int B_h^2(x)dx + \int \mathrm{Var}(\mathcal{E}_n(x))dx \\ &= \frac{h^4}{4}\sigma_K^4 \int |\nabla^2 p(x)|^2 dx + \frac{\mu_K}{nh^d} + o(h^4) + o\left(\frac{1}{nh^d}\right).\end{aligned} \tag{4}$$

The MISE can be viewed as the mean square error of KDE. The dominating term $\frac{h^4}{4}\sigma_K^4 \int |\nabla^2 p(x)|^2 dx + \frac{\mu_K}{nh^d}$ is called the asymptotic mean integrated square error (AMISE). Equation (4) shows that the error (risk) of KDE can be decomposed in to a bias component, $\frac{h^4}{4}\sigma_K^4 \int |\nabla^2 p(x)|^2 dx$, and a variance component $\frac{\mu_K}{nh^d}$ together with small corrections. This decomposition is known as the bias-variance tradeoff [8] and is very useful in practice because we can choose the smoothing bandwidth h by optimizing this error. If we ignore smaller order terms and use the AMISE, the minimal error occurs when we choose

$$h_{\mathrm{opt}} = \left(\frac{4\mu_K}{\sigma_K^4 \int |\nabla^2 p(x)|^2 dx \cdot \frac{1}{n}}\right)^{\frac{1}{d+4}} \tag{5}$$

which leads to the optimal MISE

$$\int \mathbb{E}\left(\left(\widehat{p}_{n,\mathrm{opt}}(x) - p(x)\right)^2\right)dx = \inf_{h>0} \int \mathbb{E}\left(\left(\widehat{p}_n(x) - p(x)\right)^2\right)dx = O\left(n^{-\frac{2}{d+4}}\right).$$

Equation (5) will be a key result in choosing the smoothing bandwidth (Section 2.3). Note that in practice, people generally select the smoothing bandwidth by minimizing the MISE rather than other errors because (i) it is a risk function that does not depend on any particular sample, (ii) it measures the overall estimation error rather than putting too much weight on a small portion of the support (i.e. it is more robust to small perturbations), and (iii) it has useful theoretical behaviors, including the expression of the bias-variance tradeoff and the connection to the mean square error.

Remark 2.2 (Boundary bias): When the density function is discontinuous, the bias of KDE at the discontinuities will be of the order $O(h)$ rather than $O(h^2)$, and this bias is called the boundary bias [7,8]. In practice, one can use the boundary kernel to reduce the boundary bias (see, e.g. Chapter 6.2.3.5 in [7]).

2.2. Derivative estimation

KDE can be used to estimate the derivative of the density function. This is often called *density derivative estimation* [18,19]. The idea is simple: we use the derivative of KDE as an estimator of the corresponding derivative of the density function. Let $[\beta] = (\beta_1, \ldots, \beta_d)$ be a multi-index (each β_ℓ is a non-negative integer and $|[\beta]| = \sum_{\ell=1}^{d} \beta_d$). Define $D^{[\beta]} =$

$\frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}} \cdots \frac{\partial^{\beta_d}}{\partial x_d^{\beta_d}}$ to be the $[\beta]$-th order partial derivative operator. For instance, $[\beta] = (1, 3, 0, \ldots, 0)$ implies $D^{[\beta]} = \frac{\partial}{\partial x_1}\frac{\partial^3}{\partial x_2^3}$ and $|[\beta]| = 4$. Then, under smoothness assumptions [19],

$$D^{[\beta]}\widehat{p}_n(x) - D^{[\beta]}p(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+2|[\beta]|}}}\right). \tag{6}$$

That is, the (MISE or pointwise) error rate of gradients of KDE is $O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+2}}}\right)$ and the error rate of second derivatives (Hessian matrix) of KDE is $O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right)$. Similarly to $B_h(x)$ and $\mathcal{E}_n(x)$ in the density estimation, there are explicit formulas for the bias and stochastic variation of density derivative estimation; see [19] for more details. Some examples of using gradient and second derivative estimation can be found in [14,20–24].

2.3. Bandwidth selection

How to choose the smoothing bandwidth for KDE is a classical research topic in nonparametric statistics. This problem is often known as *bandwidth selection*. Figure 3 shows KDE's with different amounts of smoothing of the same data-set. When h is too small (left panel), there are many wiggles in the density estimate. When h is too large (right panel), we smooth out important features. When h is at the correct amount (middle), we can see a clear picture of the underlying density.

Common approaches to bandwidth selection include the rule of thumb [25], least square cross-validation, [26–29], biased cross-validation, [30], and plug-in method [31,32]. Roughly speaking, the core idea behind all of these methods is to minimize the AMISE, the dominating quantity in the MISE (4), or other similar error measurements. Different bandwidth selectors can be viewed as different estimators to the AMISE, and h is chosen by minimizing the AMISE estimator. Overviews and comparisons of the existing methods can be found in [33,34], page 135–137 in [8], and Chapter 6.5 in [7].

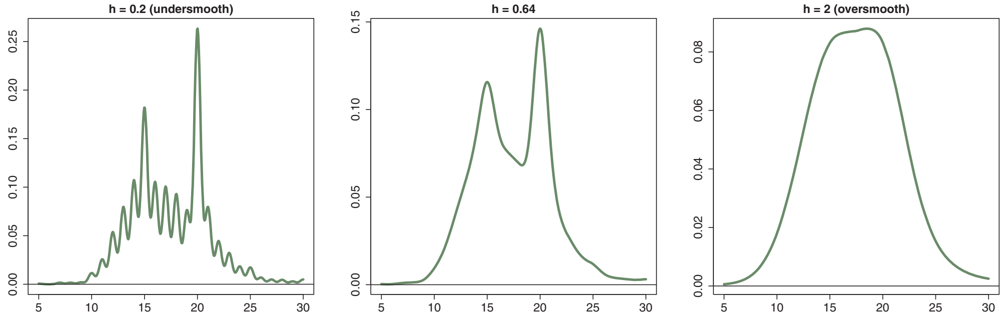

Figure 3. Smoothing bandwidth and KDE. We use the same data as in the left panel of Figure 1. We display KDE using three different amounts of smoothing. The left panel is the case of undersmoothing: we choose an excessively small bandwidth h. The middle panel is the case of the correct amount of smoothing, which is chosen according to the default rule in R. The right panel is the case of oversmoothing: the chosen h is too large.

While most of the literature focuses on the univariate case, [21] provides a generalization of all of the above methods to the multivariate case and also generalizes the AMISE criterion into density derivative estimation. In R, one can use packge 'ks[1]' or 'kedd[2]' to choose smoothing bandwidths for both density estimation and density derivative estimation. Note that the ks package is applicable to multivariate data.

In addition to the above approaches, [35] proposes a method, known as the Lepski's approach [36,37], that treats the bandwidth selection problem as a model selection problem and proposes a new criterion for selecting the smoothing bandwidth. One feature of Lepski's approach is that the selected bandwidth enjoys many statistical optimalities [35].

Remark 2.3 (Kernel selection): In contrast to bandwidth selection, the choice of kernel function does not play an important role in KDE. The effect of the kernel function on the estimation error is just a constant shift (via σ_K and μ_K in equation (2)), and the difference is generally very small among common kernel functions (see, e.g. page 72 of [8] and Section 6.2.3 in [7]), so most of the literature ignore this topic.

3. Confidence intervals and confidence bands

Confidence regions of the density function are random intervals $C_{1-\alpha}(x)$ derived from the sample such that $C_{1-\alpha}(x)$ covers the true value of $p(x)$ with probability at least $1-\alpha$. Based on this notion, there are two common types of confidence regions:

- *Confidence interval*: for a given x, the set $C_{1-\alpha}(x)$ satisfies

$$P(p(x) \in C_{1-\alpha}(x)) \geq 1-\alpha.$$

- *Confidence band*: the interval $C_{1\,-\,\alpha}(x)$ satisfies

$$P(p(x) \in C_{1-\alpha}(x) \quad \forall\, x \in \mathbb{K}) \geq 1-\alpha.$$

Namely, confidence intervals are confidence regions with only local coverage and confidence bands are confidence regions with simultaneous coverage. If a confidence interval/band has only coverage $1-\alpha+o(1)$, it will be called an *asymptotically valid* $1-\alpha$ confidence interval/band.

For simplicity, we first ignore the bias between $\widehat{p}_n(x)$ and $p(x)$ by assuming $p(x)=\mathbb{E}\left(\widehat{p}_n(x)\right)$ in Sections 3.1 and 3.2 (i.e. we assume $B_h(x)=0$). We will discuss strategies for handling the bias in Section 3.3.

3.1. Confidence intervals

For a given point x, by Equation (2),

$$\sqrt{nh^d}\left(\widehat{p}_n(x)-\mathbb{E}\left(\widehat{p}_n(x)\right)\right)=\sqrt{nh^d}\mathcal{E}_n\vec{d}N\left(0,\sigma_p^2(x)\right), \tag{7}$$

where $\sigma^2{}_p(x) = \mu_K \cdot p(x)$. Equation (7) implies that a straight-forward approach to construct a confidence band is to use asymptotic normality with a variance estimator.

Method 1: Plug-in approach. A simple method is replacing $p(x)$ in the asymptotic variance by its estimator $\widehat{p}_n(x)$, leading to the following $1 - \alpha$ confidence interval of $p(x)$:

$$C_{1-\alpha,\mathsf{PI}}(x) = \left[\widehat{p}_n(x) - z_{1-\alpha/2}\sqrt{\frac{\mu_K \cdot \widehat{p}_n(x)}{nh^d}},\quad \widehat{p}_n(x) + z_{1-\alpha/2}\sqrt{\frac{\mu_K \cdot \widehat{p}_n(x)}{nh^d}}\right]. \tag{8}$$

We call this method the "plug-in method" because we plug-in the variance estimator to construct a confidence interval. When $h \to 0$, $nh^d \to \infty$, $\widehat{p}_n(x)$ is a consistent estimator of $p(x)$. As a result,

$$P\left(\mathbb{E}\left(\widehat{p}_n(x)\right) \in C_{1-\alpha,\mathsf{PI}}(x)\right) = 1 - \alpha + o(1).$$

Method 2: Bootstrap and plug-in approach. An alternative method is to estimate the asymptotic variance using the bootstrap [38]. In more detail, we use the empirical bootstrap (also known as the nonparametric bootstrap or Efron's bootstrap, which is to sample the original data with replacement) to generate bootstrap sample $X^*{}_1$, ..., $X^*{}_n$. Then, we apply KDE to the bootstrap sample, resulting in a bootstrap KDE $\widehat{p}_n^*(x)$. When we repeat the bootstrap B times, we then have B bootstrap KDEs $\widehat{p}_n^{*(1)}(x), \ldots, \widehat{p}_n^{*(B)}(x)$. Let

$$\widehat{\sigma}^2_{p,\mathsf{BT}}(x) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\widehat{p}_n^{*(j)}(x) - \bar{p}_n^*(x)\right)^2,$$

where $\bar{p}_n^*(x) = \frac{1}{B}\sum_{j=1}^{B}\widehat{p}_n^{*(j)}(x)$ is the sample average of the bootstrap KDE's. Namely, $\widehat{\sigma}^2_{p,\mathsf{BT}}(x)$ is the sample variance of the B bootstrap KDE's evaluated at x. A bootstrap $1 - \alpha$ confidence interval is

$$C_{1-\alpha,\mathsf{BT+PI}}(x) = \left[\widehat{p}_n(x) - z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{p,\mathsf{BT}}(x),\quad \widehat{p}_n(x) + z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{p,\mathsf{BT}}(x)\right]. \tag{9}$$

Because the bootstrap variance estimator $\widehat{\sigma}^2_{p,\mathsf{BT}}(x)$ converges to $\frac{\sigma_p^2(x)}{nh^d}$ in the sense that

$$\frac{\widehat{\sigma}^2_{p,\mathsf{BT}}(x)}{\sigma_p^2(x)/(nh^d)}\vec{P}1,$$

the bootstrap variance estimator is consistent, so the confidence interval will also be consistent:

$$P\left(\mathbb{E}\left(\widehat{p}_n(x)\right) \in C_{1-\alpha,\mathsf{BT+PI}}(x)\right) = 1 - \alpha + o(1).$$

Method 3: Bootstrap approach. In addition to the above methods, one can use a fully bootstrapping approach to construct a confidence interval without using asymptotic normality. Let $\widehat{p}_n^{*(1)}(x), \ldots, \widehat{p}_n^{*(B)}(x)$ be bootstrap KDE's as in the previous method. We define a pointwise deviation of a bootstrap KDE by

$$\Delta_1(x) = |\widehat{p}_n^{*(1)}(x) - \widehat{p}_n(x)|, \ldots, \Delta_B(x) = |\widehat{p}_n^{*(B)}(x) - \widehat{p}_n(x)|.$$

Then we compute the $1-\alpha$ quantile of the empirical CDF of $\Delta_1(x)$, ..., $\Delta_B(x)$:

$$c_{1-\alpha,\mathsf{BT}}(x) = \widehat{G}_x^{-1}(1-\alpha), \quad \widehat{G}_x(t) = \frac{1}{B}\sum_{j=1}^{B} I\left(\Delta_j \le t\right).$$

A $1-\alpha$ confidence interval of $p(x)$ is

$$C_{1-\alpha,\mathsf{BT}}(x) = \left[\widehat{p}_n(x) - c_{1-\alpha,\mathsf{BT}}(x), \quad \widehat{p}_n(x) + c_{1-\alpha,\mathsf{BT}}(x)\right]. \tag{10}$$

Because the distribution of $|\widehat{p}_n^{*}(x) - \widehat{p}_n(x)|$ approximates the distribution of $|\widehat{p}_n(x) - \mathbb{E}\left(\widehat{p}_n(x)\right)|$, this confidence interval is also asymptotically valid, i.e.

$$P\left(\mathbb{E}\left(\widehat{p}_n(x)\right) \in C_{1-\alpha,\mathsf{BT}}(x)\right) = 1 - \alpha + o(1).$$

3.1.1. Example: Confidence intervals

In Figure 4, we compare the three approaches of constructing a confidence interval using the NACC Uniform Data-Set, as described in the Introduction (Section 1) and the left panel of Figure 1. The left panel is the 95% confidence interval of each point using the plug-in approach (method 1); the middle panel is the 95% confidence interval from the plug-in and bootstrap approach (method 2); the right panel is the 95% confidence interval from the bootstrap approach (method 3).

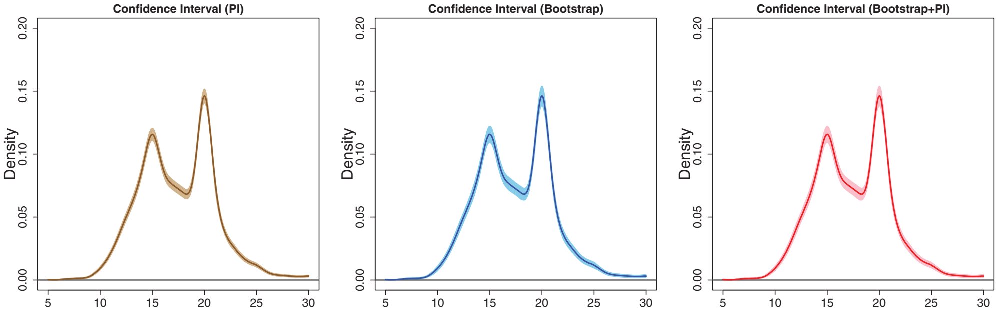

Figure 4. 95% confidence intervals from KDE. We use the same data as in the left panel of Figure 1. Left: We obtain confidence intervals using the plug-in approach (method 1 in Section 3.1). Middle: We construct confidence intervals using the plug-in approach with the bootstrap (method 2 in Section 3.1). Right: We build confidence intervals using the bootstrap approach (method 3 in Section 3.1). The three confidence regions are nearly the same, although they are constructed using different approaches. Note that there are two problems with these confidence regions. First, since we ignore the bias, the actual coverage might be substantially less than the nominal coverage 95%. Second, because they are confidence intervals, the we only have pointwise coverage. Thus, even though the actual coverage is guaranteed, these regions might not cover the entire actual density function.

Essentially, the three confidence intervals are very similar; in particular, the confidence intervals from the method 2 and 3 are nearly identical. The interval from method 1 is slightly smaller than the other two.

While all of these intervals are valid for each given point, there is no guarantee that they will cover the entire density function *simultaneously*. In the next section, we introduce methods of constructing confidence bands (confidence regions with simultaneous coverage).

3.2. Confidence bands

Now, we present methods of constructing confidence bands. The key idea is to approximate the distribution of the uniform error $\sup_x|\widehat{p}_n(x) - p(x)|$ and then convert it into a confidence band. To be more specific, let $G(t) = P\left(\sup_x|\widehat{p}_n(x) - p(x)| < t\right)$ be the CDF of the uniform error, and let $\overline{c}_{1-\alpha} = G^{-1}(1-\alpha)$ be the $1-\alpha$ quantile. Then it can be shown that the set

$$\overline{C}(x) = \left[\widehat{p}_n(x) - \overline{c}_{1-\alpha},\ \widehat{p}_n(x) + \overline{c}_{1-\alpha}\right]$$

is a confidence band, i.e.

$$P\left(p(x) \in \overline{C}(x)\ \ \forall x \in \mathbb{K}\right) = 1 - \alpha.$$

Therefore, as long as we have a good approximation of the distribution $G(t)$, we can convert the approximation into a confidence band.

Method 1: Plug-in approach. An intuitive approach is to derive the asymptotic distribution of $\sup_x|\widehat{p}_n(x) - p(x)|$ directly and then invert it into a confidence band. [39,40] proved that the uniform loss converges to an extreme value distribution in the sense that

$$P\left(\sqrt{-2\log h}\left(\sqrt{nh^d}\sup_x \frac{|\widehat{p}_n(x) - \mathbb{E}\left(\widehat{p}_n(x)\right)|}{\sqrt{p(x)\mu_K}} - d_n\right) < t\right) \to e^{-2e^{-t}}, \tag{11}$$

where $d_n = O\left(\sqrt{-2\log h}\right)$ is a quantity depending only on n, h and the kernel function K. [40] provided an exact expression for the quantity d_n. Let $E_{1-\alpha} = -\log\left(-\frac{\log\alpha}{2}\right)$ be the $1-\alpha$ quantile of the right-hand-side CDF. Define

$$c_{1-\alpha} = \sqrt{\frac{p(x)\mu_K}{nh^d}}\left(d_n + \frac{E_{1-\alpha}}{\sqrt{-2\log h}}\right).$$

Then, by Equation (11), $\sup_x|\widehat{p}_n(x) - \mathbb{E}\left(\widehat{p}_n(x)\right)|$ falls within $[0, c_{1-\alpha}]$ with probability at least (asymptotically) $1-\alpha$. To construct a confidence band, we replace the quantity $p(x)$ in $c_{1-\alpha}$ with a plug-in estimate from KDE, leading to

$$c_{1-\alpha,\mathsf{PI}} = \sqrt{\frac{\widehat{p}_n(x)\mu_K}{nh^d}}\left(d_n + \frac{E_{1-\alpha}}{\sqrt{-2\log h}}\right).$$

Then a $1-\alpha$ confidence band will be

$$C^{\dagger}_{1-\alpha,\mathsf{PI}}(x) = \left[\widehat{p}_n(x) - c_{1-\alpha,\mathsf{PI}},\quad \widehat{p}_n(x) + c_{1-\alpha,\mathsf{PI}}\right]. \tag{12}$$

Although Equation (12) is an asymptotically valid confidence band, the convergence to the extreme value distribution in Equation (11) is very slow [41]. Thus, we need a huge sample size to guarantee that the confidence band from (12) is asymptotically valid. To resolve this problem, we use the bootstrap.

Method 2: Bootstrap approach. The key element of how the bootstrap works is that the uniform error can be approximated accurately by the supremum of a Gaussian process [42–44]. In more detail, there exists a tight Gaussian process $B_n(x)$ such that

$$\sup_t \left| P\left(\sqrt{nh^d} \sup_x |\widehat{p}_n(x) - \mathbb{E}(\widehat{p}_n(x))| < t\right) - P\left(\sup_x |B_n(x)| < t\right)\right| = o(1). \tag{13}$$

Moreover, the difference between the bootstrap KDE and the original KDE also has a similar convergent result [45–47]:

$$\sup_t \left| P\left(\sqrt{nh^d} \sup_x |\widehat{p}^*_n(x) - \widehat{p}_n(x)| < t | X_1, \ldots, X_n\right) - P\left(\sup_x |B_n(x)| < t\right)\right| = o_P(1), \tag{14}$$

where $B_n(x)$ is the same Gaussian process as the one in Equation (13). Thus, the distribution of $\sup_x|\widehat{p}_n(x) - \mathbb{E}(\widehat{p}_n(x))|$ will be approximated by the distribution of its bootstrap version $\sup_x|\widehat{p}^*_n(x) - \widehat{p}_n(x)|$. As a result, the bootstrap quantile of uniform error converges to the quantile of the actual uniform error, thereby proving that the bootstrap confidence band is asymptotically valid.

Here is the formal construction of a bootstrap confidence band. Let $\widehat{p}^{*(1)}_n(x), \ldots, \widehat{p}^{*(B)}_n(x)$ be the bootstrap KDE's. We define the uniform deviation of the bootstrap KDE's by

$$\Delta_1 = \sup_x |\widehat{p}^{*(1)}_n(x) - \widehat{p}_n(x)|, \ldots, \Delta_B = \sup_x |\widehat{p}^{*(B)}_n(x) - \widehat{p}_n(x)|.$$

Then, we compute the $1-\alpha$ quantile of the empirical CDF of Δ_1, ..., Δ_B as

$$c_{1-\alpha,\mathsf{BT}} = \widehat{G}^{-1}_{\mathbb{K}}(1-\alpha), \quad \widehat{G}_{\mathbb{K}}(t) = \frac{1}{B}\sum_{j=1}^{B} I(\Delta_j(x) \le t).$$

A $1-\alpha$ confidence band will be

$$C^{\dagger}_{1-\alpha,\mathsf{BT}}(x) = \left[\widehat{p}_n(x) - c_{1-\alpha,\mathsf{BT}},\quad \widehat{p}_n(x) + c_{1-\alpha,\mathsf{BT}}\right]. \tag{15}$$

By Equations (13) and (14), when $B \to \infty$,

$$P\left(\mathbb{E}(\widehat{p}_n(x)) \in C^{\dagger}_{1-\alpha,\mathsf{BT}}(x) \ \ \forall x \in \mathbb{K}\right) = 1 - \alpha + o(1).$$

Namely, the set $C^{\dagger}_{1-\alpha,\mathsf{BT}}(x)$ is an asymptotically valid $1-\alpha$ confidence band.

3.2.1. Example: confidence bands

Figure 5 presents confidence bands using KDE. The left panel shows the confidence band from the bootstrap approach introduced in the previous section. Compared to the confidence intervals in Figure 4, the confidence bands are wider because we need to control the coverage simultaneously for every point.

However, the confidence band in the left panel of Figure 5 (and all of the confidence intervals in Figure 4) has a serious problem–the coverage guarantee is for the expected value of KDE $\mathbb{E}\left(\widehat{p}_n(x)\right)$ rather than for the true density function p. Unless we undersmooth KDE (choose h converging at a faster rate than $O(n^{-\frac{1}{d+4}})$), the confidence band shows undercoverage. We will discuss this topic in more detail in Section 3.3.

To construct an (asymptotically valid) confidence band with h being at rate $O(n^{-\frac{1}{d+4}})$, we use the debiased estimator introduced in [48]. The details are provided in Section 3.3.3 and Figure 6. The right panel of Figure 5 shows a confidence band from the debiased KDE approach. Although the confidence band is wider, the coverage is guaranteed for such a confidence band.

3.3. Handling the bias

In the previous section, we ignored the bias in KDE. However, the bias could be a severe problem in reality because it systematically shifted our confidence interval/band so the actual coverage is below the nominal coverage. Here we discuss strategies to handle bias.

3.3.1. Ignoring the bias

A simple strategy is to ignore the bias and focus on inferring the expectation of KDE $p_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right)$. p_h is called the smoothed or mollified density function [14,22,49]. As

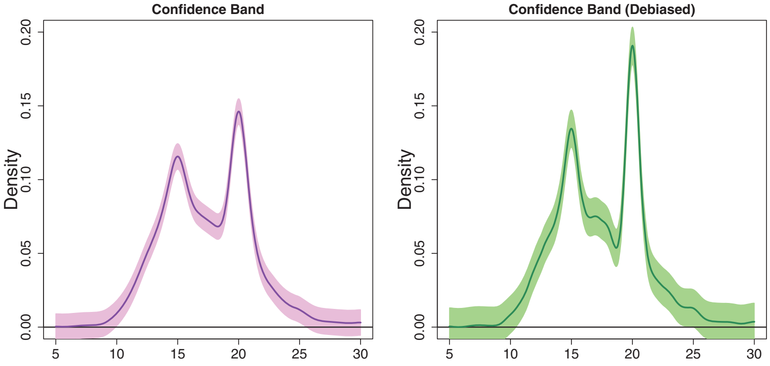

Figure 5. 95% confidence bands from KDE via the bootstrap. This is the same data-set as in the left panel of Figure 1. Left: The confidence band from bootstrapping the uniform error of KDE (method 2 of Section 3.2). Although this is a confidence band, we ignore the bias in constructing the confidence band so the actual coverage could be below the nominal coverage. Right: The confidence band from bootstrapping the uniform error of the debiased KDE (method proposed in Figure 6). Because we are using the debiased KDE, the density curve is slightly different from the blue curve in the left panel. Although the confidence band is wider than the confidence in the left, this confidence band has coverage guaranteed when the smoothing bandwidth is chosen at rate $O(n^{-\frac{1}{d+4}})$. Note that there are possible approach to narrowing down the size of this confidence band; we refer the reader to [48].

Confidence Band of Density Function.

(1) Choose the smoothing bandwidth h by a standard approach such as the rule of thumb or cross-validation [106–108].

(2) Compute the bias-corrected KDE (with $b = h$ in equation (16))

$$\widetilde{p}_n(x) = \widehat{p}_n(x) - \frac{h^2}{2}\sigma_K^2 \nabla^2 \widehat{p}_n(x).$$

(3) Bootstrap the original sample for B times and compute the bootstrap bias-corrected KDE

$$\widetilde{p}_n^{*(1)}(x), \cdots, \widetilde{p}_n^{*(B)}(x).$$

(4) Compute the quantile

$$\widetilde{c}_{1-\alpha} = \widehat{F}^{-1}(1-\alpha), \quad \widehat{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I\left(\|\widetilde{p}_n^{*(j)} - \widetilde{p}_n\|_\infty < t\right).$$

(5) Output the confidence band

$$\widetilde{C}^{\dagger}_{1-\alpha,\mathsf{BT}}(x) = [\widetilde{p}_n(x) - \widetilde{c}_{1-\alpha}, \widetilde{p}_n(x) + \widetilde{c}_{1-\alpha}]$$

Figure 6. A confidence band of the density function from bootstrapping the debiased KDE [48]. This confidence band is asymptotically valid and is compatible with most of the bandwidth selectors introduced in Section 2.3.

long as the kernel function K is smooth, p_h will also be smooth. Moreover, p_h exists even when the distribution function is singular (in this case, the population density p does not exist). For inferring geometric or topological features, p_h might be a better parameter of interest because structures in p_h generally represent salient structures of p [14] and many topological structures of p_h will be similar to those of p when h is small [17,22,24]. If we switch our target to p_h, we have to make it clear that this is a confidence region of p_h rather than of p when we report our confidence regions.

3.3.2. Undersmoothing

Undersmoothing is a very common approach to handle bias in KDE (and other nonparametric approaches). Recall from Equation (2):

$$\begin{aligned}\widehat{p}_n(x) - p(x) &= B_h(x) = \mathcal{E}_n(x)\\ &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right).\end{aligned}$$

The bias is $B_h(x) = O(h^2)$, and the stochastic variation is the $\mathcal{E}_n(x) = O_P\left(\sqrt{\frac{1}{nh^d}}\right)$ term. If now we take $h \to 0$ such that $h^2 = o\left(\sqrt{\frac{1}{nh^d}}\right)$, then the stochastic variability dominates the errors. Therefore, we can ignore the bias term and use the method suggested in Sections 3.1 and 3.2.

Note that $h^2 = o(\sqrt{\frac{1}{nh^d}})$ is equivalent to $nh^{d+4} \to 0$, which corresponds to choosing a smaller smoothing bandwidth than the optimal smoothing bandwidth ($h_{\mathrm{opt}} \asymp n^{-\frac{1}{d+4}}$) that balances the bias and the variance (this is why it is called undersmoothing). Although the undersmoothing provides a valid construction of confidence regions, such a choice of bandwidth implies that the size of the confidence band is larger than the optimal size because we are inflating the variance to eliminate the bias. Some references to undersmoothing can be found in [50–56].

3.3.3. Bias-corrected and oversmoothing

An alternative approach to construct a valid confidence band is to correct the bias of KDE explicitly; this approach is known as the bias-corrected method and the resulting KDE is called the bias-corrected KDE. Recall from Equation (2) that the bias in KDE is

$$\mathbb{E}\left(\widehat{p}_n(x)\right) - p(x) = \frac{h^2}{2}\sigma_K^2 \nabla^2 p(x) + o\left(h^2\right).$$

Thus, we can correct the bias by estimating $\nabla^2 p(x)$. The quantity $\nabla^2 p(x)$ can be estimated by $\nabla^2 \widehat{p}_b(x)$, where

$$\widehat{p}_b(x) = \frac{1}{nb^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right)$$

is KDE using smoothing bandwidth b. Recall from Section 2.2 that the second derivative estimator has an error rate

$$\nabla^2 \widehat{p}_b(x) - \nabla^2 p(x) = O\left(b^2\right) + O_P\left(\sqrt{\frac{1}{nb^{d+4}}}\right). \tag{16}$$

Thus, to obtain a consistent estimator of $\nabla^2 p(x)$, we have to choose another smoothing bandwidth b, and this smoothing bandwidth needs to be larger than the optimal bandwidth $h_{\mathrm{opt}} = O(n^{-\frac{1}{d+4}})$. Because the choice of b corresponds to oversmoothing KDE, this approach is also called the 'oversmoothing' method.

Using $\widehat{p}_b(x)$, the bias-corrected KDE is

$$\tilde{p}_n(x) = \widehat{p}_n(x) - \frac{h^2}{2}\sigma_K^2 \nabla^2 \widehat{p}_b(x).$$

When $\nabla^2 \widehat{p}_b(x)$ is a consistent estimator of $\nabla^2 p(x)$, the pointwise error rate is

$$\tilde{p}_n(x) - p(x) = o\left(h^2\right) + o_P\left(h^2\right) + O_P\left(\sqrt{\frac{1}{nh^d}}\right),$$

so the dominating quantity, $O_P(\sqrt{\frac{1}{nh^d}})$, is the stochastic variation in $\tilde{p}_n(x)$. Thus, the confidence regions can be constructed by replacing $\widehat{p}_n(x)$ by $\tilde{p}_n(x)$ in equations (8), (9), (10), and (15). An incomplete list of literature of the bias-corrected approach is as follows[57–65].

[66] proposes a plug-in method of constructing a confidence interval with $b = h$. Although the choice $b = h$ does not lead to a consistent estimate of the second derivative, the bias will be pushed into the next order because the bias-corrected estimator can be viewed as a higher order kernel function (see page 157 in [7]). Thus, since the dominating term in the estimation error is the stochastic variation, we can construct an asymptotically valid confidence interval using the plug-in approach.

[48] further generalizes the idea of [66] to confidence bands by bootstrapping the bias-corrected kernel density estimator with $b = h$. Figure 6 summarizes the procedure of constructing a confidence band using the approach in [48]. The resulting confidence band, $\tilde{C}^{\dagger}_{1-\alpha,\mathsf{BT}}(x)$, has the following property:

$$P\Big(p(x) \in \tilde{C}^{\dagger}_{1-\alpha,\mathsf{BT}}(x) \;\; \forall x \in \mathbb{K}\Big) = 1 - \alpha + o(1)$$

when $h = h_{\mathsf{opt}} = O\Big(n^{-\frac{1}{d+4}}\Big)$. Namely, $\tilde{C}^{\dagger}_{1-\alpha,\mathsf{BT}}(x)$ is an asymptotically valid $1 - \alpha$ confidence band when we pick h under the optimal rate. The right panel of Figure 5 shows an example of this confidence band. Although this approach generally leads to a wider confidence band, this confidence band has asymptotically $1 - \alpha$ coverage whereas the confidence band in the left panel of Figure 5 has undercoverage.

Remark 3.1 (Calibration): In addition to the above methods, another possible approach is to choose a corrected coverage of confidence regions; this approach is called 'calibration' and is related to the work in [67–70]. The principal idea is to investigate the effect of the bias on the coverage of the confidence band and then choose a conservation quantile to guarantee the nominal coverage of the resulting confidence regions.

4. Geometric and topological features

KDE can be used to estimate not only the underlying density function but also geometric (and topological) structures related to the density. To be more precise, many geometric (and topological) features of $\widehat{p}_n$ converges to the corresponding structures of p, and hence, we can use a structure of $\widehat{p}_n$ as the estimator of that structure of p.

Because geometric and topological structures generally involve the gradient and Hessian matrix of the density function, we define some notations here. We define $g(x) = \nabla p(x)$ to be the gradient of the density and $H(x) = \nabla\nabla p(x)$ to be the Hessian matrix of the density. Moreover, we also define $\lambda_1(x) \geq \cdots \geq \lambda_d(x)$ to be the largest to the smallest eigenvalues of $H(x)$ and $v_1(x), \ldots, v_d(x)$ to be the corresponding eigenvectors.

4.1. Local modes

A well-known geometric feature of the density function is its (global) mode. Actually, when Parzen introduced KDE, he mentioned the use of the mode of KDE to estimate the mode of the density function [1]. The asymptotic distribution and confidence sets of the mode were later discussed in [71,72].

We can extend the definition of the (global) mode to a local sense and define the local modes:

$$\mathcal{M} = \{x : g(x) = 0, \lambda_1(x) < 0\}.$$

Namely, $\mathcal{M}$ is the collection of points for which the density function is locally maximized. A natural estimator of $\mathcal{M}$ is a plug-in from KDE [73,74]:

$$\widehat{\mathcal{M}} = \{x : \widehat{g}_n(x) = 0, \widehat{\lambda}_1(x) < 0\},$$

where $\widehat{g}_n(x)$ and $\widehat{\lambda}_1(x)$ are KDE version of $g(x)$ and $\lambda_1(x)$. Under mild assumptions, $\widehat{\mathcal{M}}$ is a consistent estimator of $\mathcal{M}$ [73,74]. Note that one can use the mean shift algorithm [75–77] to compute the estimator $\widehat{\mathcal{M}}$ numerically.

Note that one can use the local modes to cluster data points; this is called mode clustering [74,78,79] or mean-shift clustering [75–77]. The left panel of Figure 7 shows a case of estimated local modes (black boxes) and mode clustering using the mean-shift algorithm. In R, one can use the library 'LPCM[3]' to compute the estimator $\widehat{\mathcal{M}}$ and perform mode clustering.

4.2. Level sets

Level sets are regions for which the density value is equal to or above a particular level. Given a level λ, the λ-level set [80,81] is

$$L_\lambda = \{x : p(x) \geq \lambda\}.$$

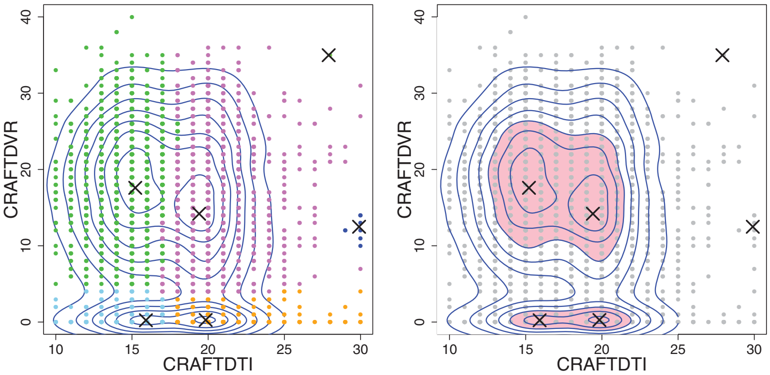

Figure 7. Estimating geometric features using KDE. This is the same data-set as the right panel of Figure 1. Left: Density local modes (black crosses) and mode clustering. The colored points describe the clusters that the respective subject belong to. Mode clustering uses the gradient flow to partition data points into clusters. Thus, each cluster (points with the same color) has a local mode as its representative. Right: Density contours (blue), local modes (black crosses), and a density level set (pink area). A density level set is just a region containing points whose density values are greater than or equal to a particular level. Thus, it will contain regions within some contour lines (blue curves). (colour online.)

A natural estimator of L_λ is the plug-in estimate from KDE:

$$\widehat{L}_\lambda = \{x : \widehat{p}_n(x) \geq \lambda\}.$$

The pink area in the right panel of Figure 7 is one instance of the density level set.

There is substantial statistical literature discussing different types of convergence of $\widehat{L}_\lambda$; see [49,80–82] and the references therein. [83] and [14] propose procedures for constructing confidence sets of L_λ through bootstrapping the estimator $\widehat{L}_\lambda$. Note that a visualization tool[4]for multivariate level sets is proposed in [14].

4.3. Ridges

Another interesting geometric structures are ridges [22,24,84] of the density functions. Formally, ridges are defined as follows. Let $V(x) = [v_2(x) \cdots v_d(x)] \in \mathbb{R}^{d\times(d-1)}$ be the matrix consisting of the second eigenvector to the last eigenvector. A density ridge is defined as

$$\mathcal{R} = \{x : V(x)V(x)^T g(x) = 0, \lambda_2(x) < 0\}.$$

Intuitively, any point $x \in \mathcal{R}$ is a local mode in the subspace spanned by $v_2(x)$, ..., $v_d(x)$. Thus, if we move away from $\mathcal{R}$ in the subspace, the density value decreases, which is the characteristic attribute of a ridge.

To estimate $\mathcal{R}$, we again use the plug-in from KDE:

$$\widehat{\mathcal{R}} = \left\{x : \widehat{V}(x)\widehat{V}(x)^T \widehat{g}(x) = 0, \widehat{\lambda}_2(x) < 0\right\},$$

where $\widehat{V}(x)$ and $\widehat{\lambda}_2(x)$ are KDE versions of $V(x)$ and $\lambda_2(x)$ respectively. The convergence rate and topological characteristics were discussed in [24]. [22] and [84] both studied the asymptotic theory, and [22] further proposed methods of constructing confidence sets of $\mathcal{R}$. [86] introduced the subspace-constrained mean shift (SCMS) algorithm to compute $\widehat{\mathcal{R}}$. The red curves in the right panel of Figure 7 are estimated ridges from the SCMS algorithm.

4.4. Morse–Smale complex

The Morse–Smale complex [87] of a density function p is a partition of the entire support $\mathbb{K}$ based on the density gradient flow. For any point $x \in \mathbb{K}$, we define a gradient ascent flow $\pi_x(t)$ such that

$$\pi_x^{'}(t) = g(\pi_x(t)), \quad \pi_x(0) = x.$$

Namely, $\pi_x(t)$ is a flow starting at x such that we move along the orientation of the density gradient ascent. By Morse theory [88,89], such a flow converges to a destination that is one of the critical points (points where $g(x) = 0$) when the density function is smooth.

Similarly, we define a gradient descent flow $\gamma_x(t)$ such that

$$\gamma_x^{'}(t) = -g(\gamma_x(t)), \quad \gamma_x(0) = x.$$

In a similar manner to $\pi_x(t)$, $\gamma_x(t)$ starts at x but now the flow moves by following density gradient descent. Again, by Morse theory, such a flow converges also to one of the critical points, but not to the same point as the destination of $\pi_x(t)$. Based on the destinations of $\pi_x(t)$ and $\gamma_x(t)$, we partition the entire support $\mathbb{K}$ into different regions; points within the same region share the same destination for both the gradient ascent flow and the gradient descent flow. This partition is called the Morse–Smale complex.

To estimate the Morse–Smale complex of p, we use the Morse–Smale complex of $\widehat{p}_n$. [20] and [90] studied the convergence of gradient flows and the Morse–Smale complex of $\widehat{p}_n$ and proved the statistical consistency of these geometric features. [90] further proposed to use the Morse–Smale complex to visualize a multivariate density function.[5] Note that one can use the R package 'msr[6]' that to perform data analysis with the Morse–Smale complex (see [91–93] for more details).

4.5. Cluster trees

Cluster trees (also known as density trees [94,95]) are tree-structured objects summarizing the structure of the underlying density function. The left panel of Figure 8 provides an example of a cluster tree corresponding to KDE in the right panel of Figure 2 (also the same data-set as in Figures 7 and 8). A cluster tree is constructed as follows. Recall that $L_\lambda = \{x: p(x) \geq \lambda\}$ is the density level set at the level λ. When the level λ is too large, L_λ will

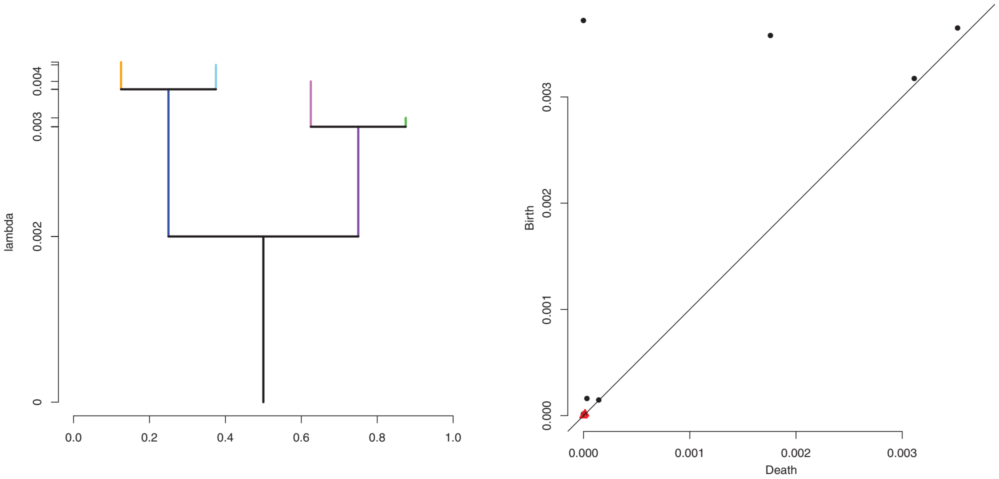

Figure 8. Estimating topological features of KDE. This is the same data-set as in the right panel of Figure 1. Left: Cluster tree, a tree structure representing KDE. The four leaves correspond to the the four high-density local modes in Figure 7 (other local modes are tiny so the cluster tree algorithm ignores them). Right: The persistent diagram. The top four black dots denote the four persistent topological features (four connected components) of KDE, which are created by the four high-density local modes in Figure 7. There are several structures in the bottom left corner; they correspond to the topological noises in constructing a persistent diagram. For more details, we refer the reader to [85].

be empty, because no region has a density value above such level. When we gradually decrease λ from a large number, at some levels (when λ hits the density value of local modes), a new connected component will be created; this corresponds the creation of a connected component. Moreover, at particular levels, two or more connected components will merge (generally at the density value of local minima or saddle points); this corresponds to the elimination of a connected component. Cluster tree uses a tree structure to summarize the creation and elimination of connected components at different levels. Since a cluster tree always live in a 2D plane, it is an excellent tool for visualizing a multivariate density function. For more details, we refer to the review paper in [85].

The above defines a cluster tree using the underlying population density p. In practice, we can construct an estimated cluster tree using KDE $\widehat{p}_n$. Convergence of the cluster tree estimator was studied in [15,96–98]. [17] provides a procedure for constructing confidence sets of cluster trees based on bootstrapping KDE $\widehat{p}_n$. In R, one can use the package 'TDA[7]' to construct a tree estimator of KDE.

4.6. Persistent diagram

A persistent diagram [85,99,100] is a diagram summarizing the topological features of a density function p. The construction of a persistent diagram is very similar to that of a cluster tree, but now we focus on not only the connected components but also higher-order topological structures, such as loops and voids (see [16,85] for a more details).

There are several means of estimating a persistent diagram. A natural approach is to use the persistent diagram of KDE $\widehat{p}_n$. For such an estimator, the stability theorem in [100] together with the uniform convergence in Section 2.1 are sufficient to prove the convergence of the estimated persistent diagram toward the population persistent diagram. For statistical inference, [16] proposes a bootstrap approach over KDE to construct a confidence set. In practice, one can use R package 'TDA[8]' to construct the persistent diagram of KDE.

The right panel of Figure 8 shows an example of the persistent diagram of connected components (zeroth-order topological features) and loops (first-order topological features) of KDE described in the right panel of Figure 2. At the top of the figure, the four dots indicate the existence of the four high-density local modes. In the bottom left regions, the black dots and red triangles are the topological noises representing low-density local modes (black dots) and low-density 'loops' structures (red triangles; first-order topological features). Note that the two low-density local modes (black crosses) in the right part of both panels of Figure 7 are topological noises corresponding to the two black dots in the bottom left corner of the persistent diagram.

5. Estimating the CDF

KDE can also be used to estimate a CDF [101]. The estimator is simple; we just integrate KDE (for simplicity, here we consider the univariate case):

$$\widehat{F}_n(x) = \int_{-\infty}^{x} \widehat{p}_n(y)dy. \tag{17}$$

Convergence of such estimators was extensively analysed soon after their introduction [102–107].

Similarly to the pointwise error rate, one can show that

$$\widehat{F}_n(x) - F(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{n}}\right) + O_P\left(\sqrt{\frac{h}{n}}\right)$$

(see the derivation in [101] and [107]). Again, the first quantity $O(h^2)$ is related to the bias. The other two quantities are related to the stochastic variation. Under such a rate, the optimal smoothing bandwidth will be $h^* \approx n^{-1/3}$, which leads to the error rate

$$\widehat{F}_n(x) - F(x) = O_P\left(\sqrt{\frac{1}{n}}\right),$$

the same as using the empirical CDF. Note that as long as $h = O(n^{-1/4})$, we will obtain the square root error rate.

To construct a confidence band of $F(x)$ via $\widehat{F}_n(x)$, one can use the uniform central limit theorem proposed by [108] to relate the uniform loss to the supremum of a Gaussian process and then use either the limiting distribution or the bootstrap. Note that to apply the result in [108], one have to undersmooth (see Theorem 4 and Section 4.1.1 in [108]) the data or use a higher order kernel function (see Remark 7 and Corollary 2 in [108]).

5.1. ROC curve

KDE can also be applied to estimate and infer the receiver operating characteristic (ROC) curve [109]. In the setting of the ROC curve, we observed two samples. The first is the sample of healthy subjects, whose responses are $X_1, \ldots, X_n$ from an unknown density P. The other is the sample of diseased individuals, whose response are $Y_1, \ldots, Y_m$ from an unknown density G. Consider a simple rule of classification based on choosing a cutoff point s such that we classify an individual as diseased if its response value is larger than s, otherwise it is classified as a healthy individual.

For such a rule, the sensitivity is $SE(s) = 1 - G(s)$, the probability of detecting a diseased subject. We also define the specificity $SP(s) = F(s)$, the probability of successfully assigning a healthy subject to the healthy group. Then, the ROC curve is defined as the plotting of the true positive fraction $SE(s)$ versus the false positive fraction $1 - SP(s)$, or equivalently, as plotting the function

$$ROC(t) = 1 - G\left(F^{-1}(1 - t)\right).$$

A recent review on ROC curves can be found in [110].

A classical nonparametric approach of estimating $ROC(t)$ is to plug-in the empirical CDF estimator for both F and G [111]. As an alternative, one can use the integrated KDEs of both samples to estimate the ROC curve [112–114] as Equation (17). This is often called a smoothed estimator of the ROC curve because the resulting ROC curve estimator is generally a smooth curve.

To construct confidence bands of an ROC curve, most methods propose using the plug-in estimate from the empirical distribution and constructing the confidence band by the bootstrap [115–117]. A formal proof of the theoretical validity of such a bootstrap approach is provided in [114,118,119]. Note that can also construct a confidence band by bootstrapping the smoothed ROC curve estimator and using the method proposed in Section 3.2 to construct a confidence band.

6. Conclusion and open problems

In this tutorial, we reviewed KDE's basic properties and its applications in estimating structures of the underlying density function. For readers who would like to learn more about different varieties of KDE, we recommend [8] and [7]. Because this is a tutorial, we ignore many advanced topics such as the minimax theory and adaptation. An introduction of these theoretical properties can be found in [9].

Although KDE has been widely studied since its introduction in the 1960s, there are still open problems that deserve further investigation. Here we briefly discuss some open problems related to the materials in this tutorial.

- *Confidence bands of other KDE-type estimators.* In addition to estimating a probability density function and its related structures, the idea of kernel smoothing can be applied to estimate a regression, hazard, or survival function. Moreover, in casual inference, we might be interested in the difference between the regression/hazard/survival function from the control group and that from the treatment group as a characteristic of the treatment effect. One example is the conditional average treatment effect [120–123]. Although in this tutorial we have seen methods of constructing confidence bands of density functions, how to construct a (asymptotically) valid confidence band of these functions remains an open question.
- *Multidimensional problems.* When the dimension of the data d is large, KDE poses several challenges. First, KDE (and most other nonparametric density estimators) suffers severely from the so-called *curse of dimensionality*: The optimal convergence rate $O\left(n^{-\frac{2}{d+4}}\right)$ is very slow when d is large, and this slow convergence rate cannot be improved [9,124] unless we assume extra smoothness. One way to solve this problem is to find density surrogates that can be estimated easily and to switch our parameter of interest to a density surrogate. However, this rises the question of what the correct surrogates and the corresponding estimators are, and this still remains unclear. Another issue of KDE in multi-dimensions is visualization. When $d > 3$, we can no longer see the entire KDE, and therefore we must use visualization tools to explore our density estimates. However, it is still unclear how to choose a visualization tool in practice.
- *More about geometric/topological structures.* In Section 4, we saw that several useful geometric and topological structures can be estimated by the corresponding structures of KDE. However, we do not yet fully understand the behavior of these estimators. For instance, how to choose the smoothing bandwidth that optimally estimate these structures is unclear. Handling this issue may require generalizing the concept of the MISE to the set estimator [125] and choosing the smoothing bandwidth that minimizes such an error measurement. In addition to bandwidth selection, uniform

inference remains an open question for these structures. Although there are methods of constructing confidence sets of most of these structures, it is unclear whether the resulting confidence sets are uniform for a collection of density functions. Moreover, theoretical optimality, such as the minimax theory, remains unclear for several of these structures, presenting another set of open questions in the study of KDE.

Notes

1. https://cran.r-project.org/web/packages/ks/index.html
2. https://cran.r-project.org/web/packages/kedd/index.html
3. https://cran.r-project.org/web/packages/LPCM/index.html
4. R source code: https://github.com/yenchic/HDLV
5. R source code: https://github.com/yenchic/Morse_Smale
6. https://cran.r-project.org/web/packages/msr/index.html
7. https://cran.r-project.org/web/packages/TDA/index.html
8. https://cran.r-project.org/web/packages/TDA/index.html

Acknowledgments

We thank two referees for the very useful suggestions. We also thank Gang Chen, Aurelio Uncini, and Larry Wasserman for useful comments.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Yen-Chi Chen is an assistant professor in the Department of Statistics, a data science fellow in the eScience Institute, and a statistician in the National Alzheimer's Coordinating Center at the University of Washington. His research focuses on nonparametric statistics, cluster analysis, topological data analysis, and applications in various fields.

ORCID

Yen-Chi Chen http://orcid.org/0000-0002-4485-306X

References

[1] Parzen E. On estimation of a probability density function and mode. Ann Math Stat. 1962;33 (3):1065–1076.
[2] Beekly DL, Ramos EM, Lee WW, et al. The national alzheimer's coordinating center (nacc) database: the uniform data set. Alzheimer Dis Assoc Disord. 2007;21(3):249–258.
[3] Deng H, Wickham H. Density estimation in r. Available from: http://vita.had.co.nz/papers/density-estimation.pdf, 2011.
[4] Loftsgaarden DO, Quesenberry CP. A nonparametric estimate of a multivariate density function. Ann Math Stat. 1965;36(3):1049–1051.
[5] Breiman L, Meisel W, Purcell E. Variable kernel estimates of multivariate densities. Technometrics. 1977;19(2):135–144.

[6] Terrell GR, Scott DW. Variable kernel density estimation. Ann Stat. 1992;20(3):1236–1265.

[7] Scott DW. Multivariate density estimation: theory, practice, and visualization. Hoboken (NJ): Wiley; 2015.

[8] Wasserman L. All of nonparametric statistics. New York: Springer; 2006.

[9] Tsybakov AB. Introduction to nonparametric estimation. New York (NY): Springer; 2009.

[10] Yukich J. Laws of large numbers for classes of functions. J Multivariate Anal. 1985;17 (3):245–260.

[11] Giné E, Guillou A. Rates of strong uniform consistency for multivariate kernel density estimators. In journal: Annales de l'Institut Henri Poincare (B) Probability and Statistics, Vol. 38. Elsevier; 2002. p. 907–921.

[12] Einmahl U, Mason DM. Uniform in bandwidth consistency of kernel-type function estimators. Ann Stat. 2005;33(3):1380–1403.

[13] Rao BP. Nonparametric functional estimation. Harahan (LA): Academic Press; 2014.

[14] Chen YC, Genovese CR, Wasserman L. Density level sets: asymptotics, inference, and visualization. *arXiv:1504.05438*, 2015c.

[15] Chen YC. Generalized cluster trees and singular measures. *arXiv preprint arXiv:1611.02762*, 2016.

[16] Fasy BT, Lecci F, Rinaldo A, et al. Confidence sets for persistence diagrams. Ann Stat. 2014;42(6):2301–2339.

[17] Jisu K, Chen YC, Balakrishnan S, et al. Statistical inference for cluster trees. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Advances in neural information processing systems. Neural Information Processing Systems (NIPS); 2016. p. 1831–1839.

[18] Stoker TM. Smoothing bias in density derivative estimation. J Am Stat Assoc. 1993;88 (423):855–863.

[19] Chacón JE, Duong T, Wand M. Asymptotics for general multivariate kernel density derivative estimators. Stat Sin. 2011;21:807–840.

[20] Arias-Castro E, Mason D, Pelletier B. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. J Mach Learn Res. 2016;17(43):1–28.

[21] Chacón JE, Duong T. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. Electron J Stat. 2013;7:499–532.

[22] Chen YC, Genovese C R, Wasserman L. Asymptotic theory for density ridges. Ann Stat. 2015b;43(5):1896–1928.

[23] Genovese CR, Perone-Pacifico M, Verdinelli I, et al. On the path density of a gradient field. Ann Stat. 2009;37(6A):3236–3271.

[24] Genovese CR, Perone-Pacifico M, Verdinelli I, et al. Nonparametric ridge estimation. Ann Stat. 2014;42(4):1511–1545.

[25] Silverman BW. Density estimation for statistics and data analysis. Boca Raton (FL): Chapman and Hall; 1986.

[26] Rudemo M. Empirical choice of histograms and kernel density estimators. Scand J Stat. 1982;9(2):65–78.

[27] Bowman A. An alternative method of cross-validation for the smoothing of density estimates. Biometrika. 1984;71(2):353–360.

[28] Bowman AW, Azzalini A. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Vol. 18. Oxford University Press; 1997.

[29] Stone CJ. An asymptotically optimal window selection rule for kernel density estimates. Ann Stat. 1984;12(4):1285–1297.

[30] Scott DW, Terrell GR. Biased and unbiased cross-validation in density estimation. J Am Stat Assoc. 1987;82(400):1131–1146.

[31] Woodroofe M. On choosing a delta-sequence. Ann Math Stat. 1970;41(5):1665–1671.

[32] Sheather S, Jones C. A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc (Statistical Methodology). 1991;53(3):683–690.

[33] Jones MC, Marron JS, Sheather SJ. A brief survey of bandwidth selection for density estimation. J Am Stat Assoc. 1996;91(433):401–407.

[34] Sheather SJ. Density estimation. Stat Sci. 2004;19(4):588–597.

[35] Goldenshluger A, Lepski O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. Ann Stat. 2011;39(3):1608–1632.

[36] Goldenshluger A, Lepski O. Universal pointwise selection rule in multivariate function estimation. Bernoulli. 2008;14(4):1150–1190.

[37] Lepski O, Goldenshluger A. Structural adaptation via lp-norm oracle inequalities. Probab. Theory Relat Fields. 2009;126(1–2):47–71.

[38] Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7(1):1–26.

[39] Bickel P, Rosenblatt M. On some global measures of the deviations of density function estimates. Ann Stat. 1973;1(6):1071–1095.

[40] Rosenblatt M. On the maximal deviation of k-dimensional density estimates. Ann Probab. 1976;4(6):1009–1015.

[41] Hall P. On convergence rates of suprema. Probab Theory Relat Fields. 1991;89(4):447–455.

[42] Neumann MH. Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. Ann Stat. 1998;26(5):2014–2048.

[43] Chernozhukov V, Chetverikov D, Kato K. Comparison and anti-concentration bounds for maxima of gaussian random vectors. Probab Theory RelatFields. 2014b;162(1–2):47–70.

[44] Chernozhukov V, Chetverikov D, Kato K. Gaussian approximation of suprema of empirical processes. Ann Stat. 2014c;42(4):1564–1597.

[45] Chernozhukov V, Chetverikov D, Kato K. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann Stat. 2013;41(6):2786–2819.

[46] Chernozhukov V, Chetverikov D, Kato K. Anti-concentration and honest, adaptive confidence bands. Ann Stat. 2014a;42(5):1787–1818.

[47] Chernozhukov V, Chetverikov D, Kato K. Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. Stoch Process Appl. 2016.

[48] Chen YC. Nonparametric inference via bootstrapping the debiased estimator. *arXiv preprint arXiv:1702.07027*, 2017.

[49] Rinaldo A, Wasserman L. Generalized density clustering. Ann Stat. 2010;38(5):2678–2722.

[50] Bjerve S, Doksum KA, Yandell BS. Uniform confidence bounds for regression based on a simple moving average. Scand J Med Sci. 1985;12(2):159–169.

[51] Hall P. On bootstrap confidence intervals in nonparametric regression. Ann Stat. 1992a;20 (2):695–711.

[52] Hall P, Owen AB. Empirical likelihood confidence bands in density estimation. J Comput Graph Stat. 1993;2(3):273–289.

[53] Chen SX. Empirical likelihood confidence intervals for nonparametric density estimation. Biometrika. 1996;83(2):329–341.

[54] Neumann MH, Polzehl J. Simultaneous bootstrap confidence bands in nonparametric regression. J Nonparametric Stat. 1998;9(4):307–333.

[55] Chen SX, Qin YS. Confidence intervals based on local linear smoother. Scand J Stat. 2002;29 (1):89–99.

[56] McMurry TL, Politis DN. Bootstrap confidence intervals in nonparametric regression with built-in bias correction. Stat Probab Lett. 2008;78(15):2463–2469.

[57] Härdle W, Bowman AW. Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. J Am Stat Assoc. 1988;83(401):102–110.

[58] Hardle W, Marron J. Bootstrap simultaneous error bars for nonparametric regression. Ann Stat. 1991;19(2):778–796.

[59] Hall P. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann Stat. 1992b;20(2):675–694.

[60] Eubank RL, Speckman PL. Confidence bands in nonparametric regression. J Am Stat Assoc. 1993;88(424):1287–1301.

[61] Sun J, Loader CR. Simultaneous confidence bands for linear regression and smoothing. Ann Stat. 1994;22(3):1328–1345.

[62] Härdle W, Huet S, Jolivet E. Better bootstrap confidence intervals for regression curve estimation. Stat J Theor Appl Stat. 1995;26(4):287–306.

[63] Neumann MH. Automatic bandwidth choice and confidence intervals in nonparametric regression. Ann Stat. 1995;23(6):1937–1959.

[64] Xia Y. Bias-corrected confidence bands in nonparametric regression. J R Stat Soc (Statistical Methodology). 1998;60(4):797–811.

[65] Härdle W, Huet S, Mammen E, et al. Bootstrap inference in semiparametric generalized additive models. Econometric Theory. 2004;20(2):265–300.

[66] Calonico S, Cattaneo MD, Farrell MH. On the effect of bias estimation on coverage accuracy in nonparametric inference. *arXiv preprint arXiv:1508.02973*, 2015.

[67] Beran R. Prepivoting to reduce level error of confidence sets. Biometrika. 1987;74(3):457–468.

[68] Hall P. On the bootstrap and confidence intervals. Ann Stat. 1986;14(4):1431–1452.

[69] Loh WY. Calibrating confidence coefficients. J Am Stat Assoc. 1987;82(397):155–162.

[70] Hall P, Horowitz J. A simple bootstrap method for constructing nonparametric confidence bands for functions. Ann Stat. 2013;41(4):1892–1921.

[71] Romano JP. Bootstrapping the mode. Ann Inst Stat Math. 1988a;40(3):565–586.

[72] Romano JP. On weak convergence and optimality of kernel density estimates of the mode. Ann Stat. 1988b;16(2):629–647.

[73] Chazal F, Fasy BT, Lecci F, et al. Robust topological inference: distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

[74] Chen YC, Genovese CR, Wasserman L. A comprehensive approach to mode clustering. Electron J Stat. 2016;10(1):210–241.

[75] Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory. 1975;21(1):32–40.

[76] Cheng Y. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell. 1995;17(8):790–799.

[77] Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. IEEE Trans Patt Anal Mach Intell. 2002;24(5):603–619.

[78] Chacón JE. A population background for nonparametric density-based clustering. Stat Sci. 2015;30(4):518–532.

[79] Azizyan M, Chen Y-C, Singh A, et al. Risk bounds for mode clustering. *arXiv preprint arXiv:1505.00482*, 2015.

[80] Polonik W. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. Ann Stat. 1995;23(3):855–881.

[81] Tsybakov AB. On nonparametric estimation of density level sets. Ann Stat. 1997;25(3):948–969.

[82] Rinaldo A, Singh A, Nugent R, et al. Stability of density-based clustering. J Mach Learn Res. 2012;13(1):905–948.

[83] Mammen E, Polonik W. Confidence regions for level sets. J Multivariate Anal. 2013;122:202–214.

[84] Qiao W, Polonik W. Theoretical analysis of nonparametric filament estimation. Ann Stat. 2016;44(3):1269–1297.

[85] Wasserman L. Topological data analysis. *arXiv preprint arXiv:1609.08227*, 2016.

[86] Ozertem U, Erdogmus D. Locally defined principal curves and surfaces. J Mach Learn Res. 2011;12(Apr):1249–1286.

[87] Banyaga A. Lectures on Morse homology. Vol. 29. Netherlands: Springer Science & Business Media; 2004.

[88] Morse M. Relations between the critical points of a real function of n independent variables. Trans Am Math Soc. 1925;27(3):345–396.

[89] Morse M. The foundations of a theory of the calculus of variations in the large in m-space (second paper). Trans Am Math Soc. 1930;32(4):599–631.

[90] Chen YC, Genovese CR, Wasserman L. Statistical inference using the Morse-Smale complex. *arXiv preprint arXiv:1506.08826*, 2015d.

[91] Gerber S, Bremer PT, Pascucci V, et al. Visual exploration of high dimensional scalar functions. IEEE Trans Vis Comput Graph. 2010;16(6):1271–1280.
[92] Gerber S, Potter K. Data analysis with the Morse-Smale complex: the msr package for r. J Stat Softw. 2011;50(2):1–22.
[93] Gerber S, Rübel O, Bremer PT, et al. Morse–Smale regression. J Comput Graph Stat. 2013;22 (1):193–214.
[94] Stuetzle W. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J Classif. 2003;20(1):025–047.
[95] Klemelä J. Smoothing of multivariate data: density estimation and visualization. Vol. 737. Hoboken (NJ): Wiley; 2009.
[96] Balakrishnan S, Narayanan S, Rinaldo A, et al. Cluster trees on manifolds. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in neural information processing systems. Neural Information Processing Systems (NIPS); 2013. p. 2679–2687.
[97] Chaudhuri K, Dasgupta S, Kpotufe S, et al. Consistent procedures for cluster tree estimation and pruning. IEEE Trans Inf Theory. 2014;60(12):7900–7912.
[98] Eldridge J, Belkin M, Wang Y. Beyond hartigan consistency: merge distortion metric for hierarchical clustering. In: Lawrence N, Reid M, editors. Proceedings of the 28th Conference on Learning Theory. Machine Learning Research; 2015. p. 588–606.
[99] Edelsbrunner H, Morozov D. Persistent homology: theory and practice. In: Proceedings of the European Congress of Mathematics. Zürich: European Mathematical Society; 2012. p. 31–50.
[100] Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. Discrete Comput Geom. 2007;37(1):103–120.
[101] Nadaraya EA. Some new estimates for distribution functions. Theory Probab Appl. 1964;9 (3):497–500.
[102] Winter B. Strong uniform consistency of integrals of density estimators. Can J Stat. 1973;1 (1–2):247–253.
[103] Reiss RD. Nonparametric estimation of smooth distribution functions. Scand J Stat. 1981;8 (2):116–119.
[104] Fernholz LT. Almost sure convergence of smoothed empirical distribution functions. Scand J Stat. 1991;18(3):255–262.
[105] Yukich J. Weak convergence of smoothed empirical processes. Scand J Stat. 1992;21(4):271–279.
[106] Mack Y. Remarks on some smoothed empirical distribution functions and process. Bull Inf Cybern. 1984;21(1–2):29–35.
[107] Azzalini A. A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika. 1981;68(1):326–328.
[108] Giné E, Nickl R. Uniform central limit theorems for kernel density estimators. Probab Theory Relat Fields. 2008;141(3–4):333–387.
[109] McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Mak. 1984;4(2):137–150.
[110] Demidenko E. Confidence intervals and bands for the binormal ROC curve revisited. J Appl Stat. 2012;39(1):67–79.
[111] Hsieh F, Turnbull BW. . Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Ann Stat. 1996;24(1):25–40.
[112] Zou KH, Hall W, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Stat Med. 1997;16(19):2143–2156.
[113] Zhou XH, Harezlak J. Comparison of bandwidth selection methods for kernel smoothing of ROC curves. Stat Med. 2002;21(14):2045–2055.
[114] Hall P, Hyndman RJ, Fan Y. Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika. 2004;91(3):743–750.
[115] Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med. 1994;13(5–7):499–508.

[116] Macskassy S, Provost F. Confidence bands for ROC curves: methods and an empirical study. In: López De Mántaras R, Saitta L, editors. Proceedings of the First Workshop on ROC Analysis in AI. Frontiers in Artificial Intelligence and Applications; 2004.

[117] Moise A, Clément B, Ducimetière P, et al. Comparison of receiver operating curves derived from the same population: a bootstrapping approach. Comput Biomed Res. 1985;18(2):125–131.

[118] Bertail P, Clémençcon SJ, Vayatis N. On bootstrapping the ROC curve. In: Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, editors. Advances in neural information processing systems. Neural Information Processing Systems (NIPS); 2009. p. 137–144.

[119] Horváth L, Horváth Z, Zhou W. Confidence bands for roc curves. J Stat Plan Inference. 2008;138(6):1894–1904.

[120] Abrevaya J, Hsu Y-C, Lieli RP. Estimating conditional average treatment effects. J Bus Econ Stat. 2015;33(4):485–505.

[121] Hsu YC. Consistent tests for conditional treatment effects. The Econ J. 2017;20(1):1–22.

[122] Lee S, Whang YJ. Nonparametric tests of conditional treatment effects. Unpublished manuscript. 2009.

[123] Ma Y, Zhou XH. Treatment selection in a randomized clinical trial via covariate-specific treatment effect curves. Stat Methods Med Res. 2014;26(1):124–141.

[124] Stone CJ. Optimal global rates of convergence for nonparametric regression. Ann. Stat. 1982;10(4):1040–1053.

[125] Chen YC, Genovese CR, Ho S, et al. Optimal ridge detection using coverage risk. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in neural information processing systems. Neural Information Processing Systems (NIPS); 2015a. p. 316–324.