An Empirical Study of Adaptive Kernel Density Estimation in Detecting Distributional Overlap

Chao Chen[1]

Supervisor(s): Jesse Krijthe [1], Rickard Karlsson [1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Chao Chen
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Rickard Karlsson, Frans Oliehoek

Abstract

Given data from an observational study or a randomized experiment, the positivity assumption must hold in order to draw causal relations between the treatment and outcome. However, there is shortage of automatic tools that verify compliance with the positivity assumption. We present tools that uses adaptive and standard kernel density estimation (KDE) methods for validating the assumption. Our empirical analysis of the methods offers insight into when a KDE method can and can not reliably be applied to verify positivity in datasets.

1 Introduction

Estimating the causal effect of an intervention on an outcome is a frequently performed task. For example, in marketing, we might want to estimate the effect of product placement on consumer purchasing behaviour. Likewise, in epidemiology, the effect of a new insulin drug on diabetic patients may need to be investigated. If, to this end, an ideal randomized experiment is performed, the observed outcome can be fully attributed to the treatment due to exchangeability between the treated and untreated [7] . Exchangeability implies that if the intervention of the treated and untreated group were swapped, the same average causal effects would be discovered [17]. Thus, we can conclude that the association between the observed treatment and observed outcome is in fact a causal relationship [7]. However, for questions such as how therapy affects depression symptoms or how inflation affects spending behaviour, it is unethical or impractical to administer randomized experiments [17]. Evidently, many scientific studies are based on observational data and much of our knowledge is derived from observational studies [7].

A limitation of observational studies is we can not ensure random treatment assignment [7]. For example, observational data may show that therapy is associated with more severe depressive symptoms but it hides the fact that patients with a severe prognosis may be more likely to be offered therapy. Thus, the association does not imply a causal relationship. This does not mean that causal inference from observational data is in vain. Observational studies can be treated as *conditional randomized experiments* if the identifiability assumptions, consistency, conditional exchangeability (unconfoundedness), and positivity, hold [4] [17]. However, the identifiability assumptions are not guaranteed in observational studies and require validation. In our research, we will focus on verifying the positivity assumption in datasets because contrary to the other assumptions, positivity can be verified from data [7].

The positivity assumption requires that every stratum of the population to be assigned to every treatment group. The assumption can be verified by measuring overlap in the covariate distributions of the groups [15] or by checking the propensity score of the subjects [17]. With an estimate of the covariate distributions, we can find or detect the absence of a region where both the distributions have support. Causal relations can then be inferred for subjects whose covariates lie in this region. Without data from the full population, assumptions about the underlying distribution, when invalid, may exacerbate the density estimate. Thus, in this paper, we will estimate overlap by means of non-parametric density estimation, which makes no prior assumptions about the underlying distribution.

Our goal is to analyse the performance of adaptive KDE methods in overlap estimation whereas majority of literature used the standard (non-adaptive) kernel density estimation. The main challenge of standard KDE is efficiently choosing the bandwidth h that minimises the Mean Integrated Squared Error (MISE) of the estimation [14]. Adaptive KDE methods address this challenge by automatically selecting the bandwidth h and varying the bandwidth of each kernel based on its local density [13]. As adaptive KDE methods were found to yield better density estimates when data is sparse [14], which is common in high dimensional data, we will explore two adaptive KDE methods, *adaptive KDE* (aKDE) and *variable KDE* (vKDE), in overlap estimation. Thus, the main research question is: *How do adaptive kernel density estimation methods compare to the classical kernel density estimation method in estimating overlap?*. To answer this question, the subquestions that will be answered are:

1. (*SQ1*) In what scenarios (datasets) does adaptive KDE outperform standard KDE, if at all and vice versa?
2. (*SQ2*) What properties of the method allows it to perform comparatively better?
3. (*SQ3*) Is there a relationship between the Mean Integrated Squared Error (MISE) and IoU (Intersection over Union) of the KDE methods?

Our contribution is three-fold. First, we provide an empirical analysis of standard and adaptive KDE methods in overlap estimation in 1D and 2D settings. Second, we identify scenarios where each of the methods fail and use that information to propose a set recommendations that guide the selection of KDE methods in overlap estimation. Finally, we demonstrate how the performance of a KDE method in density estimation and overlap estimation are related.

In the remainder of this paper, we will provide a formal problem description of overlap and non-parametric density estimation (Section 2), discuss related work in applying density estimation for overlap estimation (Section 3) and will outline the methodologies used (Section 4). Section 5 will present and explain the experimental results that answer the research questions. Section 6 will address the reproducibility of the results and Section 7 will summarise the work.

2 Background

2.1 Positivity through Overlap

Positivity requires the conditional probability for every subject to be part of any treatment group to be strictly between zero and one [18]. This requirement is intuitive: if all subjects received the same treatment (e.g. psychotherapy), it would be infeasible to measure the causal effect of the treatment on the outcome (e.g. depressive symptoms). In a dichotomous experiment, where the covariates of the subjects $i = 1, 2, \ldots m$ are denoted by $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and treatment denoted by

$T = \{0, 1\}$, the positivity assumption can be phrased as follows:

$$\forall X_i \in \mathcal{X}, t \in T : 0 < P(T = t \mid X = X_i) < 1. \quad (1)$$

The quantity $P(T = t \mid X = X_i)$ is also referred to as the propensity score and can be estimated directly through propensity score analysis [10].

An equivalent way to verify the positivity assumption is to find overlap (common support) in the covariate distributions of different treatment groups. In the dichotomous setting, positivity is satisfied when there is overlap between $P(X \mid T = 0)$ and $P(X \mid T = 1)$. The overlap region is a set $B \in \mathcal{X}$ of covariate values shared by both treatment groups. For a threshold $\epsilon \in (0, 1)$, the overlap region can be defined as:

$$\mathcal{B}_\epsilon = \{X_i \in \mathcal{X}; \forall t \in T \mid P(X = X_i \mid T = t) > \epsilon\} \quad (2)$$

Causal relationships can only be determined for subjects in the overlapping region since it is the region where the positivity assumption holds.

2.2 Density Estimation

Density estimation refers to building an estimate of a probability density function given samples from the underlying distribution [13]. Parametric density estimation assumes that the samples are drawn from a known type of distribution (e.g. Gaussian) and constructs the estimate of the density function by estimating the parameters. Nonparametric density estimation does not assume the parametric family of the samples [13]. Our scope is limited to nonparametric density estimation.

Nonparametric density estimation techniques are used because in many real-world settings, the underlying type of distribution that the samples are drawn from is unknown and possibly complex. We focused on kernel density estimation methods which typically take the form [14]:

$$\hat{f}(\mathbf{t}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{t} - \mathbf{X}_i}{h}\right), \quad (3)$$

In (3), $\hat{f}(\mathbf{t})$ is the estimated probability density function evaluated at test point $\mathbf{t}$. $K : \mathbb{R}^d \to \mathbb{R}$ is kernel function that is centered at 0 and integrates to 1. $\mathbf{X}_1, \ldots \mathbf{X}_n$ are random samples from an unknown distribution p. $\mathbf{t}$ are the test points of the estimation. h is the bandwidth or the smoothing parameter, which approaches 0 as n increases. In (3), h is fixed but there are many adaptive schemes which vary h in order to decrease the error in estimation of f.

Adaptive KDE methods are a class of methods which vary the bandwidth h based on the location of the test point or the sample point [14]. If h is determined based on the test point, the KDE method produces a *balloon estimator*, which take the form:

$$\hat{f}(\mathbf{t}) = \frac{1}{nh(\mathbf{t})^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{t} - \mathbf{X}_i}{h(\mathbf{t})}\right). \quad (4)$$

In (4), the bandwidth $h(\mathbf{t})^d$ is a function of the test point $\mathbf{t}$. If the bandwidth h depends on the sample points in the KDE method, the resulting estimator is termed a *sample smoothing estimator*. The bandwidth is a function $h(\mathbf{X}_i)^d$ of the sample point as shown:

$$\hat{f}(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(\mathbf{X}_i)^d} K\left(\frac{\mathbf{t} - \mathbf{X}_i}{h(\mathbf{X}_i)}\right) \quad (5)$$

The estimate $\hat{f}(\mathbf{t})$ by balloon estimators are not always densities [14] since $\hat{f}(\mathbf{t})$ may not integrate to one. Sample smoothing estimators always produce densities as long as K is a density. Moreover, in 1-d data, balloon estimators showed no significant improvement from fixed kernel methods and its efficiency was close to 0.0 [14]. Therefore, we have chosen to focus on sample smoothing estimators as they tend to have smaller mean integrated squared error (MISE) when the sample size is small [14], which is common in domains where overlap estimation is useful.

3 Related Work

Overlap estimation has been extensively studied in multiple domains, such as epidemiology [17] [18], psychology [11], economics [1] [9], and ecology [16] [5], where measuring causal effects is relevant [12] [10]. As of writing, the overlap index (OVI) and Bhattacharyya coefficient (BC) are the most common metrics for distributional overlap when a density estimate is available. For two density functions f_p and f_q, the OVI and BC between them are defined as (6) and (7), respectively.

$$\text{OVI}(f_p, f_q) = \int_{\mathbb{R}^d} \min\{f_p, f_q\}\, d\mathbf{x}. \quad (6)$$

$$\text{BC}(f_p, f_q) = \int_{\mathbb{R}^d} \sqrt{f_p f_q}\, d\mathbf{x} \quad (7)$$

A standard approach to estimating overlap uses the standard KDE method to construct an estimation of the covariate distributions of two groups and subsequently use these estimates to compute the OVI [11] [3]. Anderson *et al.* have shown that, under assumptions about the kernel and underlying distribution, the error in the estimation of the OVI using the standard KDE method will approach a Gaussian distribution centered at zero [1] as the sample size increases. They assume that (i) the kernel must have compact support, (ii) the densities approximated must have finite expectation and variance, (iii) undersmoothed bandwidths ($h \in (n^{-1/2}, n^{-1/4})$), and (iv) samples must be independent and identically distributed (i.i.d.). In our research, assumptions (i), (ii), and (iv) are satisfied but (iii) is violated by the variable and adaptive methods. Winner *et al.* [16] have used the auto-correlated KDE (AKDE) method to derive a confidence interval for BC, an alternative metric to the OVI. Their method accounts for different sources of bias by propagating the bias in the AKDE estimate to the estimate of the BC. However, density estimation does not always have a role in computing the amount of overlap. Fu *et al.* [6] have used a distribution-free approach to derive an upper bound to the OVI. Circumventing the challenge of setting kernel parameters, Johno *et al.* [8] have resided to a decision tree based strategy to recover an estimate for the OVI. Moreover, Oberst *et al.* [10] reduced estimating the overlap to a classification problem and produced an algorithm which outputs interpretable descriptions of regions of local overlap.

However, a limitation of the OVI and BC is that it acts more as a similarity metric rather than a measure of where the distributions overlap. OVI and BC fail to fully capture the notion of overlap: OVI = 1 and BC = 1 may not hold even if the covariate distributions of interest are fully overlapping (see Figure 10b). Alternatively, the KDE methods should be evaluated based on the Intersection over Union (IoU) (8) value with the true overlap region. The overlap region is defined as (2) and $\mathcal{B}_\epsilon$ is the true overlap region and $\hat{\mathcal{B}}_\epsilon$ is the estimated overlap region.

$$\text{IoU}(\mathcal{B}_\epsilon, \hat{\mathcal{B}}_\epsilon) = \frac{|\mathcal{B}_\epsilon \cap \hat{\mathcal{B}}_\epsilon|}{|\mathcal{B}_\epsilon \cup \hat{\mathcal{B}}_\epsilon|} \tag{8}$$

Thus, our work will focus on the Intersection over Union (IoU) value of the true and estimated overlap as it quantifies the ability of the KDE methods to identify the region of overlap.

4 Method

In this section, we will describe the KDE methods implemented to be used for overlap estimation (Section 4.1). The metrics we will use to measure the performance of a KDE method in overlap estimation (Section 4.2). Finally, we will outline the datasets used in the experiments (Section 4.3).

4.1 Kernel Density Estimation

Standard KDE The standard KDE serves as a baseline to the adaptive methods. In producing an estimate, the kernel width h is identical for all kernels. h is selected according to the scheme to minimise the approximate mean integrated squared error (AMISE) proposed by Silverman [13]. `scikit`'s [1] implementation of standard KDE with bandwidth selection according to Silverman's *rule of thumb* [13] is used in the experiments.

$$h = 0.9 \min\left(\hat{\sigma}, \frac{\text{IQR}}{1.34}\right) n^{-\frac{1}{5}} \tag{9}$$

The density estimate $\hat{f}(\mathbf{t})$ is computed according to (3) with kernel width set according to (9). $\hat{\sigma}$ is the empirical standard deviation of the sample points. IQR is the interquartile range of the sample points. The bandwidth h will grow according to the variance in the sample points.

vKDE vKDE differs from the standard KDE by allowing the width of the kernels to vary from one point to another [13] based on its distance from other sample points. When a point $\mathbf{X}_i$ is very far away from other sample points, it may have been sampled from a region of low density or be an outlier, and the kernel placed on the sample point should be flatter. The vKDE method uses the sample point's distance from its neighbours as an indicator of its local density.

$$\hat{f}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{hd_{i,k}} K\left(\frac{\mathbf{t} - \mathbf{X}_i}{hd_{i,k}}\right) \tag{10}$$

where $d_{i,k}$ is the distance from sample point $\mathbf{X}_i$ to the k-th nearest sample point. h is the kernel width and $d_{i,k}$ is the scaling parameter of the kernel. If $d_{i,k}$ is large, it is far away from most samples and $d_{i,k}$ is small when it is close to most samples. For a sample point X_i in 1-d, the Gaussian kernel can be described as $\mathcal{N}(X_i, h{d_{i,k}}^2)$. For a sample point $\mathbf{X}_i$ in n-d, the Gaussian kernel can be described as $\mathcal{N}(\mathbf{X}_i, \mathbf{\Sigma}_{i,k})$ where $\mathbf{\Sigma}_{i,k}$ is an identity matrix scaled by $h{d_{i,k}}^2$. k has been selected according to literature as $k = \sqrt{n}$, exactly as k is selected in k-nearest neighbour algorithms [13].

aKDE aKDE accounts for the local density of a sample point in the density estimation through a two stage procedure. A pilot estimate (initial estimate) is computed to estimate the local density of the sample points. The pilot estimate is then used to determine bandwidth factors by which the kernels are scaled in the adaptive estimate. Obtaining the pilot estimate can be done through any density estimation method as the method's sensitivity to the pilot estimate can be controlled [2].

1. Find a pilot estimate $\tilde{f}(\mathbf{t})$ that is positive for all sample points $\tilde{f}(\mathbf{X}_i) > 0$.
2. Define a bandwidth factor $\lambda_i = \{\tilde{f}(\mathbf{X}_i)/g\}^{-\alpha}$, where $\log g = n^{-1}\sum \log \tilde{f}(\mathbf{X}_i)$. α is the sensitivity parameter with $0 \leq \alpha \leq 1$.
3. Compute the *adaptive kernel estimate* $\hat{f}(\mathbf{t})$ according to (11).

$$\hat{f}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d \lambda_i^d} K\left(\frac{\mathbf{t} - \mathbf{X_i}}{h\lambda_i}\right) \tag{11}$$

α determines how strongly the pilot estimate will affect the final estimate. A large sensitivity parameter will cause the estimation to be sensitive to the variations in the pilot estimate and leads to a greater difference in the kernel width within the sample. Setting $\alpha = \frac{1}{d}$ will ensure that the number of observations within a scaled kernel is approximately the same throughout the density. Thus, in the experiments, we have set $\alpha = \frac{1}{d}$. The individual kernels are defined analogously to the kernels for vKDE, where it would be a Gaussian kernel $\mathcal{N}(\mathbf{X}_i, h{\lambda_i}^2)$ in 1D and $\mathcal{N}(\mathbf{X}_i, \mathbf{\Sigma}_i)$ with $\mathbf{\Sigma}_i = h{\lambda_i}^2 I$.

4.2 Metrics

Intersection-over-Union The IoU is computed as shown in (8). For synthetic datasets, the true overlap region $\mathcal{B}_\epsilon$ is computed using the `scipy`[2] library. For distributions where the probability density function exists, `scipy` can be used to compute the true values of the density over a domain, which is subsequently used to derive the true overlap region. The estimated overlap region $\hat{\mathcal{B}}_\epsilon$ is computed using the density estimations produced by the KDE methods. As ϵ is kept constant at 0.05 for all experiments, the subscript ϵ will be omitted from now on. Moreover, IoU = 1 when $|\mathcal{B}_\epsilon| = 0$ and $|\hat{\mathcal{B}}_\epsilon| = 0$ hold.

[1] https://scikit-learn.org/stable/

[2] https://scipy.org/

False Positive Rate False positive rate (FPR) is the proportion of negative samples (points in non-overlap region) that is classified as a positive sample (points in overlap region) (12). In initial experiments, we found that the kernel bandwidth of the adaptive methods were relatively large compared to the standard method. This resulted in the adaptive methods to overestimate overlap region. A high FPR is malignant in overlap estimation as it can lead to falsely inferring causal effects. We aim to identify properties of such scenarios by measuring the FPR for the distributions described in Section 4.3.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{12}$$

Mean Integrated Squared Error Majority of research on KDE have focused on optimizing the bandwidth to minimise the MISE (13) of the estimation. Work on how effective the KDE methods are in finding the overlap is limited. Moreover, without the true density at hand, it is infeasible to quantify the performance, such as with the IoU, of the overlap estimation method. Knowing the relationship between the MISE and the IoU will allow us to leverage knowledge of the MISE. For example, if the MISE of a KDE method is known certain settings, we can use the MISE as a proxy for the methods performance in estimating overlap.

$$\text{MISE} = \mathbb{E}\left[\int_{\mathbb{R}^d} \left(\hat{f}(\mathbf{t}) - f(\mathbf{t})\, d\mathbf{t}\right)^2\right] \tag{13}$$

4.3 Datasets

We compared the performance of the KDE methods in overlap estimation in both synthetic and natural settings. For synthetic settings, 5 two-class datasets, each representing different scenarios, with 200 to 400 data points in each class were generated. Visualisations of the datasets can be found in subsection A.1. For the natural settings, the Iris dataset consisting of measurements from 3 types of Iris flowers will be used. Since the true overlap is not known in the Iris dataset, the performance of the methods will be assessed qualitatively.

Figure 10a shows a 1D Gaussian dataset where the two classes are partially overlapping at $\epsilon = 0.05$. The boundaries of one class is fully contained in the other class in the dataset shown in Figure 10b. In Figure 10c, a Gamma distributed class and a Gaussian distributed class are shown. Multivariate Gaussians are shown in Figure 10d. Bimodal distributions in 2D will be investigated through the dataset in Figure 10e, where each class is a Gaussian mixture model consisting of two components that are weighted equally.

5 Results

We perform an empirical analysis on aKDE, vKDE, and standard KDE using their IoU, TPR, FPR, and MISE in 1D and 2D settings. First, to answer *SQ1*, we measure the performance of all KDE methods on the synthetic datasets and compare the results in Section 5.1. Second, to answer *SQ2*, we delve into specific scenarios where at least one KDE method fails in Section 5.2. Then, to understand the relationship between the MISE and IoU, we look at how they vary together over all the datasets as variance in the datasets increase in Section 5.3 to answer *SQ3*. Finally, we conclude the results with recommendations on the application of the methods.

For every set of parameters for a dataset, an experiment is repeated 20 times. The results are averaged the runs and the uncertainty is represented using the standard deviation of the runs.

5.1 Comparison of Methods

1D Datasets For the dataset shown in Figure 10a, σ of both classes were varied to check the methods' performance in case of partial overlap. In the scenario that the classes share the same center $\mu = 0$ (Figure 10b), σ was increased for only one of the classes while the other distribution was fixed at $\mathcal{N}(0, 0.25)$. The case of no overlap is investigated using the dataset shown in Figure 10c. μ of the Gaussian distributed class is decreased to move it closer to the Gamma distributed class, resulting in overlap when $\mu \leq 3.5$.

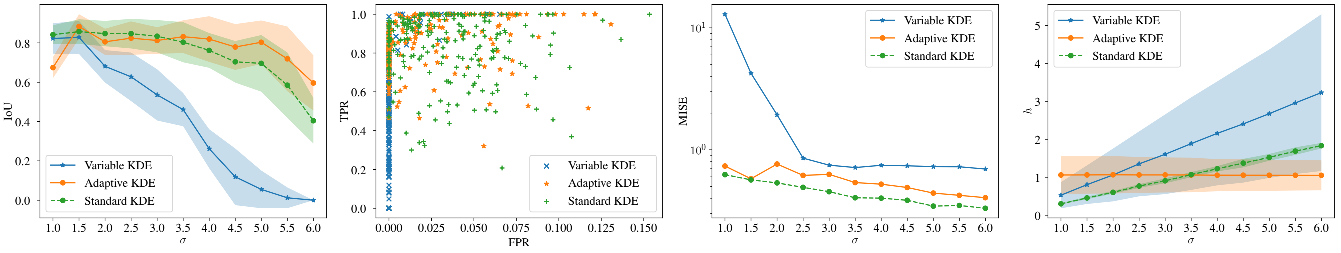
For all KDE methods, IoU is lower when σ is high as shown in Figure 1a and Figure 11a. When σ increases, given a fixed sample size, density estimation becomes less accurate locally because there are less samples per region. Consequently, the estimation becomes less smooth and has more peaks, which can cause a region to be incorrectly classified as in or outside of the overlap region. Despite local variations of the density estimate, MISE does not increase (Figure 1c, Figure 11c) because it is global measure and the magnitude of the local errors are small.

vKDE fails when σ is large because the kernel sizes grow disproportionately (Figure 1d, Figure 11d), resulting in a density estimate that is too flat. Consequently, the height density estimate may be below the threshold ϵ which leads to low FPR and TPR.

In overlap estimation for dataset in Figure 10b, IoU of aKDE is comparatively low because it underestimates the density for $\mathcal{N}(0, 0.25)$ (the class whose parameters are fixed in experiments shown in Figure 11). As shown on Figure 11d, the average kernel bandwidth h is near 1 when $\sigma = 0.5$ of the true density. Thus, the density estimate of $\mathcal{N}(0, 0.25)$ by aKDE has higher density at the tails of the distribution, which increases the estimated size of the overlap when σ is increased for $\mathcal{N}(0, \sigma^2)$. Likewise, IoU of vKDE plummets when its kernel bandwidth h grows larger than the bandwidth of aKDE.

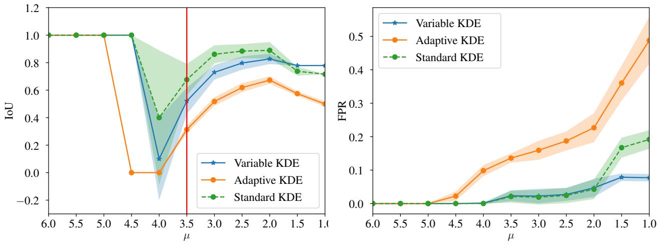
In Figure 2a, the IoU of all methods drop significantly when overlap is identified. However, all KDE methods identify the overlap too early and thus their IoU drops significantly at $\mu = 3$. The IoU remains low due to the errors in the density estimation of $\Gamma(1, 2)$. $\Gamma(1, 2)$ is defined for $x > 0$, but the KDE methods will produce estimates with tails extending to values $x < 0$. Thus, in this scenario, errors in the density estimate causes errors in the overlap estimation. In addition, it is a scenario where vKDE outperforms aKDE as vKDE does not have the tendency to produce long tails in the density estimate.

2D Datasets For dataset in Figure 10d, the covariance matrix shared both classes is a scaled identity matrix σI and the metrics are measured as σ increases. Overlap estimation for bimodal distributions is investigated by varying the co-



(a) IoU (b) ROC (c) MISE (d) Kernel Size h

Figure 1: Overlap estimation for $\mathcal{N}(0,\sigma^2)$ and $\mathcal{N}(3,\sigma^2)$ using 200 samples per class. The significant drop in IoU of vKDE Figure 1a can be attributed to its kernels growing too large in Figure 1d.



(a) IoU (b) FPR

Figure 2: Estimation of overlap between $\Gamma(1,2)$ and $\mathcal{N}(\mu,1)$ using 200 samples per class. The IoU of all methods are initially high as $|\mathcal{B}|=0$. As μ decreases, the two distributions overlap and errors in estimating $\Gamma(1,2)$ begins to affect the IoU. The μ at which the true distributions begin to overlap is indicated by the red bar.

variance matrix of each component of the Gaussian mixture (Figure 10e). Each component has the covariance matrix σI.

In 2D settings, vKDE fails. First, the vKDE's IoU shown in Figure 3a decreases rapidly as σ increases and its IoU remains low in Figure 4a. In Figure 4a, the sudden increase in IoU at $\sigma=2$ is due to vKDE correctly detecting the absence of overlap. Second, as shown in Figure 3b, vKDE tends to have high FPR (up to 0.125) while FPR of standard KDE is bounded at 0.05 and FPR of aKDE is bounded at 0.1. Furthermore, vKDE's low FPR for the Gaussian mixture (Figure 5a) comes at the expense of TPR close to 0 (Figure 5b). The TPR and FPR of vKDE is shown together in Figure 4b where most data points are clustered in a region of low TPR and high FPR. vKDE's performance in 2D can be explained by its kernel bandwidth h: h is consistently larger than the variance σ (Figure 3d, Figure 4d). This results in overestimation of the boundary of the density when variance is low (high FPR) and underestimation of the density near the center when variance is low (low TPR). Finally, all three methods produce false positives when $\sigma=0.25$ due to overestimating the boundaries of the density.

5.2 Properties of the Methods

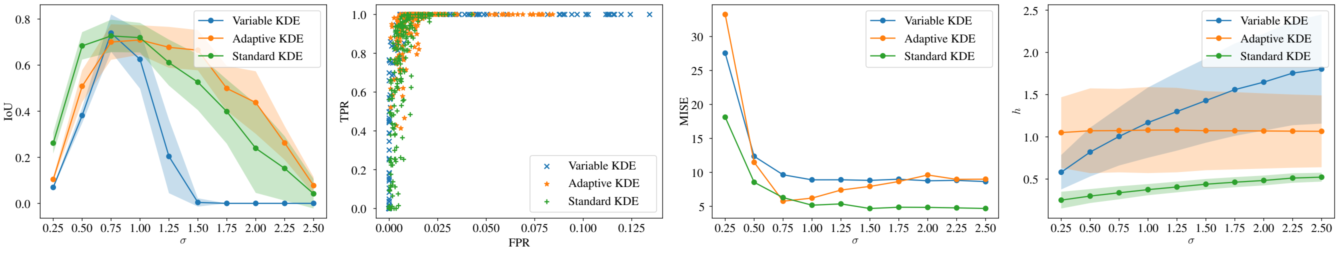
In this section, we demonstrate the properties of each method by highlighting the extreme settings of where one method fails and the others do not. In addition, we address a setting, the Iris dataset, where all of the methods fail and note the sensitivity of the IoU to the threshold value ϵ.

vKDE fails In Figure 1a, the IoU for vKDE decreases rapidly compared to that of aKDE and standard KDE. To understand this phenomenon, we investigated the behaviour of the methods when estimating overlap between $\mathcal{N}(0,25)$ and $\mathcal{N}(3,25)$ using 200 samples per class.

The low IoU can not be attributed to a high MISE. The red density can be denoted by f_1 and the blue density can be denoted by f_2. The MISE of vKDE is 0.66, of aKDE is 0.53 and of standard is 0.36. The difference in their MISE does not scale to the difference in their IoU. Moreover, standard KDE, with the lowest MISE, does not have the highest IoU. When σ is very high, the density in all regions is relatively low. Thus, oscillations in the density estimation correspond to very small errors that do not greatly increase the MISE, which explains why standard KDE does not have a high IoU despite a lower MISE. However, the oscillation does significantly affect the overlap estimation. A local decrease in the density estimate $\hat{f}_1$ or $\hat{f}_2$ below can cause a region to be considered not part of the overlap region as observed in Figure 6c. Conversely, a slight local increase in the density estimate can also cause a region to be incorrectly classified as in the overlap region.

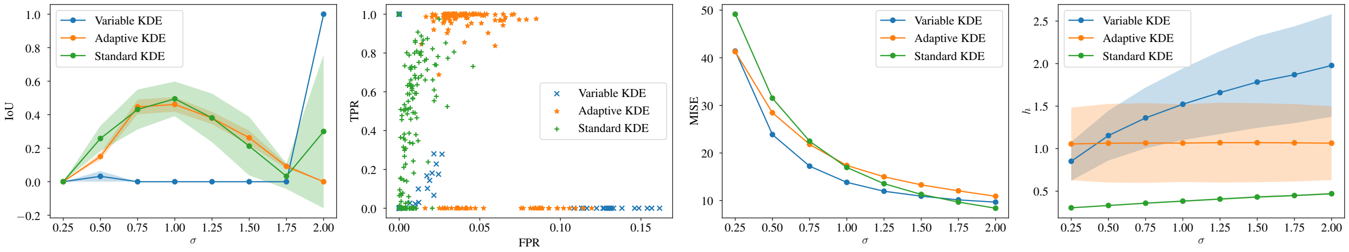
vKDE has a low IoU as it significantly underestimates the densities, causing $\hat{f}_1$ and $\hat{f}_2$ to fall mostly below the threshold $\epsilon=0.05$. This is due to the scaling of the kernels by the sample point's distance to its neighbours $d_{j,k}$ as shown in (10). When σ is very high, the distance between sample points are large on average, resulting in large kernel bandwidths h Figure 1d. The density estimates are thus too flat. The performance of vKDE can be improved by normalizing the distance between neighbours. Furthermore, note that the performance is sensitive to the threshold level ϵ, since the methods would be less sensitive to oscillations in the density estimate at a lower ϵ.

aKDE fails aKDE tendency to produce smooth estimates can fall short. For dataset Figure 10c, the IoU of aKDE drops significantly when μ approaches 4. The IoU drops because at $\mu=4$, the distributions do not yet overlap but due to the overestimation of aKDE, overlap is found nevertheless. The FPR is further exacerbated when μ of the Gaussian distribution decreases to the center of the Gamma distribution as shown in



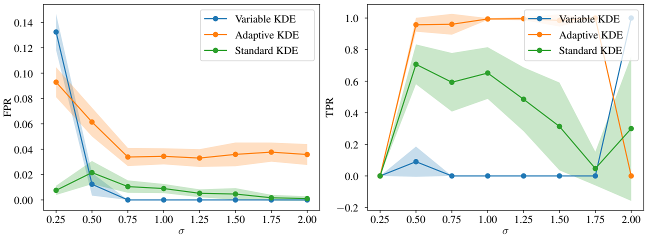
(a) IoU (b) ROC (c) MISE (d) Kernel

Figure 3: Performance of KDE methods in estimating overlap between $\mathcal{N}([0,1],\sigma I)$ and $\mathcal{N}([0,2],\sigma I)$ using 200 samples per class.



(a) IoU (b) ROC (c) MISE (d) Kernel Size h

Figure 4: Performance of KDE methods in estimating overlap between for Gaussians mixtures with 400 samples per class.



(a) FPR of Gaussian mixture (b) TPR of Gaussian mixture

Figure 5: TPR and FPR of KDE methods on Gaussian mixture model shown in Figure 10e

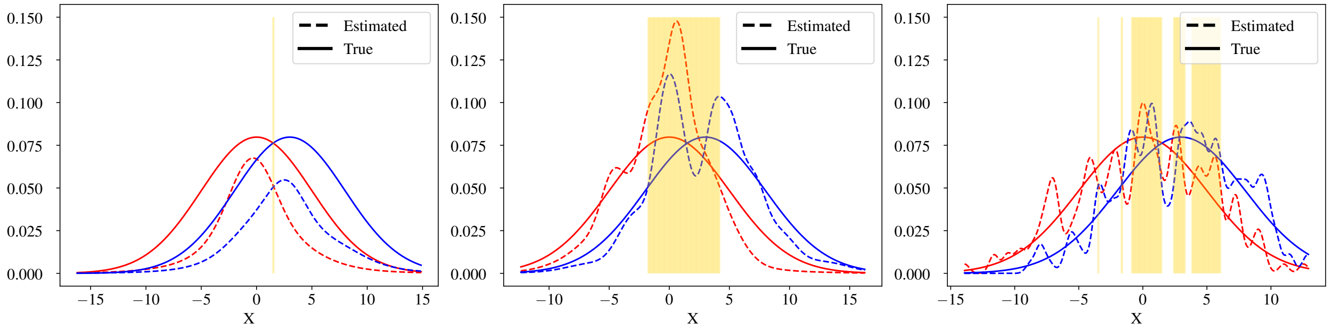
Figure 2b ($\mu = 1$). This is investigated in Figure 7 using $\Gamma(1,2)$ and $\mathcal{N}(1,1)$ with 500 samples from each distribution.

As shown in Figure 7, the FPR of aKDE is very high as its estimate of the Gamma distribution is very flat and overestimating the density in regions where the $\Gamma(1,2)$ already drops. In addition, the estimate of $\Gamma(1,2)$ by aKDE resembles an estimate of a Gaussian distribution, giving rise to a fat tail in the estimation that further increases FPR.

All methods fail In the case of estimating overlap for the Iris dataset, the performance of the methods can only be examined qualitatively. Inspired by Oberst *et al.* in [10], we estimated the overlap between Versicolor and Setosa in their sepal lengths and sepal widths was measured Figure 8a. The overlap region identified appears coherent with the data and is comparable to the region found in [10]. However, standard KDE, along with aKDE and vKDE, fail to estimate overlap between Virginica and Setosa given their sepal lengths and widths, producing many false positives. The classes shown in Figure 8b are linearly separable and thus $|\mathcal{B}_{0.1}| = 0$. The estimated overlap is region falls between the boundaries of the two classes for all three methods as shown in Figure 8b and Figure 13. The kernels of the methods are too large given the variance in the dataset, resulting in overestimating the boundaries of the distribution and the overlap. Nevertheless, it is a possibility that though not shown in Figure 8b, the joint distribution between sepal length and sepal width of Virginica and Setosa do overlap but the dataset available is too small.

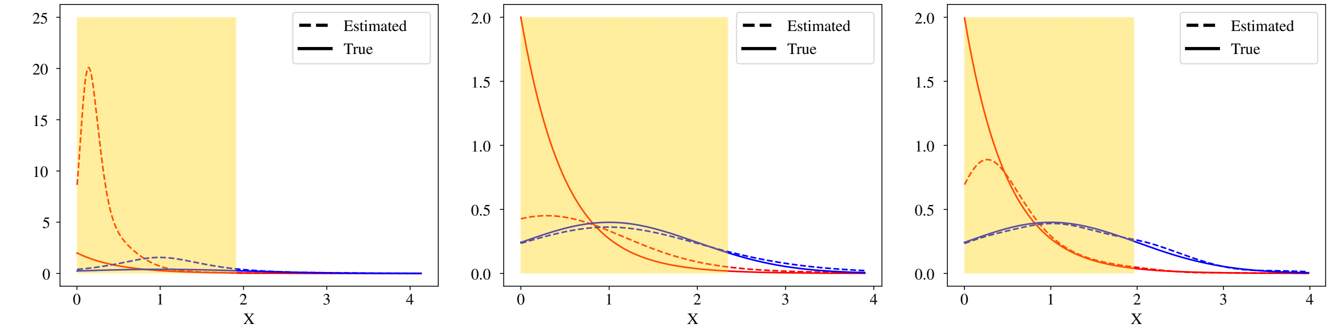
5.3 Relationship between MISE and IoU

The MISE of a density estimate is not a reliable proxy for the IoU. The IoU depends on the threshold ϵ, the distribution of interest, and the local errors of the density estimate in the region overlap, while MISE is a global metric of error. As shown in Figure 1, Figure 11, Figure 3, and Figure 4, stabilization of the MISE does not guarantee stagnation of the IoU. Nevertheless, we made the following observations for the synthetic datasets: (i) when variance is high, both the MISE and IoU tends to be low, (i) and in 2D, when variance is low, MISE tends to be high and IoU is low. (i) can be observed on Figure 14, Figure 15, Figure 9, and Figure 16 where the darker points, representing high variance, are clustered in regions of low MISE and low IoU. The results align with conclusions drawn in Section 5.1: when variance is high, the underlying density is flat and the errors in the estimation do not contribute significantly to the MISE. Rather, the low IoU can be attributed to the methods' tendency to overesti-



(a) vKDE (IoU = 0.02) (b) aKDE (IoU = 0.90) (c) Standard KDE (IoU = 0.55)

Figure 6: Overlap estimation by all three KDE methods with $n = 200$ samples. The estimated overlap region between $\mathcal{N}(0, 25)$ and $\mathcal{N}(3, 25)$ is shown in yellow. Though the MISE of all three methods do not differ significantly, their IoU's do. The IoU for vKDE is low due to underestimation of the densities: a significant portion of its density estimate falls below the threshold $\epsilon = 0.05$.



(a) vKDE (FPR = 0.03) (b) aKDE (FPR = 0.24) (c) Standard KDE (FPR = 0.05)

Figure 7: Estimation of overlap between $\Gamma(1, 2)$ and $\mathcal{N}(1, 1)$ using $n = 500$ samples. Overlap region is shown in yellow. aKDE has high a FPR as it produces an estimation of the Gamma distribution that is too flat. The flat estimation can be attributed to its large kernel bandwidth h and generally smooth estimates.

mate the boundaries of the density. (ii) can be observed on Figure 9 and Figure 16 where there is a cluster of datapoints in the region of high MISE and low IoU.

5.4 Summary of Results

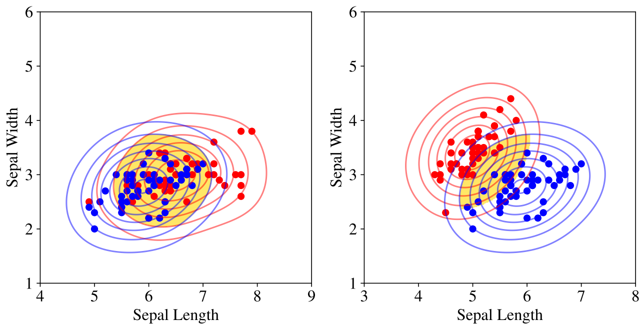
We conclude the results with a summary of the properties and recommendations for the methods:

- Standard KDE is recommended in general when false positives are considered problematic. Incorrectly identifying regions of overlap may lead to incorrectly drawing causal relations. Standard KDE has shown comparatively low FPR and high IoU in many settings explored in this paper.
- aKDE is recommended when the variance of the distribution is high. In such cases, overlap estimation can be sensitive to oscillations in the density estimate and aKDE has the advantage of producing smooth estimates at small sample sizes.
- vKDE is generally not recommended for overlap estimation for datasets with dimensionality above 1. Without tuning k, which is used to determine the scaling parameter $d_{j,k}$ of a kernel, the density estimate by vKDE can be erroneous, and k can not be tuned in practice. Moreover, using a rule of thumb for k does not yield promising results. However, there is the possibility that an additional scaling parameter can alleviate the method's sensitivity to k.

6 Responsible Research

6.1 Ethical considerations

Observational studies are prevalent in the field of epidemiology, medicine, economics, and ecology. Overlap estimation is employed on data from observational studies and randomized experiments to determine whether the positivity assumptions holds. Our tool can be used by researchers to verify the positivity assumption before drawing causal inferences between treatment and observed outcome. Therefore, it is critical to not overpromise the performance of overlap estimation methods and identify scenarios where they fail so that causal conclusions can not be drawn incorrectly. Furthermore, clear



(a) Viriginica and Versicolor

(b) Setosa and Versicolor

Figure 8: Overlap estimation for Iris data with standard KDE. Overlap region at $\epsilon = 0.1$ is indicated in yellow. Standard KDE fails to identify overlap between Setosa and Versicolor using sepal width and sepal length, but succeeds in overlap estimation between Virginica and Versicolor.

documentation of the code and a usable interface further reduces the possibility for erroneous use of the tool. We also provide plotting tools that facilitate interpretation of results from data.

6.2 Reproducibility

Code and datasets required to reproduce the results are publicly available along with the paper. The Iris dataset is also publicly available on the UCI machine learning repository [3]. All experiments have been repeated 20 times to minimise the impact of randomness on our results. Furthermore, the results from experiments were manually verified by visually inspecting the overlapping region found and comparing density estimations to the true density. Finally, though seeds were not used in the pseudo-random generators for the synthetic datasets, we have verified that the patterns observed and conclusions drawn are not affected.

7 Conclusions and Future Work

Overlap estimation of covariate distributions can be applied to detect violations of the positivity assumption. We estimate the region of overlap by means of adaptive and standard KDE methods. The aKDE, vKDE, and standard KDE have been compared on their IoU, FPR, and TPR for synthetic datasets and on qualitative assessments for real-world datasets. Our empirical analysis of the KDE methods in overlap estimation unveils their underlying properties. The extreme settings where they fail guide us on when a method should be selected for identifying overlap.

Future work can investigate how the performance of the methods will change as the threshold ϵ is varied. Moreover, we recommend analysis of the KDE methods for real-world datasets of higher dimensions and multi-modal distributions as they better typify the challenges that may occur in overlap estimation.

[3] https://archive.ics.uci.edu/

References

[1] Gordon Anderson, Oliver Linton, and Yoon-Jae Whang. Nonparametric estimation and inference about the overlap of two distributions. *Journal of Econometrics*, 171(1):1–23, 2012.

[2] Leo Breiman, William Meisel, and Edward Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

[3] Traci E Clemons and Edwin L Bradley. A nonparametric measure of the overlapping coefficient. *Computational Statistics & Data Analysis*, 34(1):51–61, 2000.

[4] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

[5] Carolyn A. Eckrich, Shannon E. Albeke, Elizabeth A. Flaherty, R. Terry Bowyer, and Merav Ben-David. rkin: Kernel-based method for estimating isotopic niche size and overlap. *Journal of Animal Ecology*, 89(3):757–771, 2020.

[6] Hao Fu, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. An upper bound for the distribution overlap index and its applications, 2023.

[7] Robins JM Hernán MA. *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton, 2023.

[8] Hisashi Johno and Kazunori Nakamoto. Decision tree-based estimation of the overlap of two probability distributions, 2022.

[9] María del Pilar Ester Arroyo Lopez. *Measuring the extent of overlap by using multivariate discriminant analysis*. PhD thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey, 1997.

[10] Michael Oberst, Fredrik D. Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush R. Varshney. Characterization of overlap in observational studies, 2020.

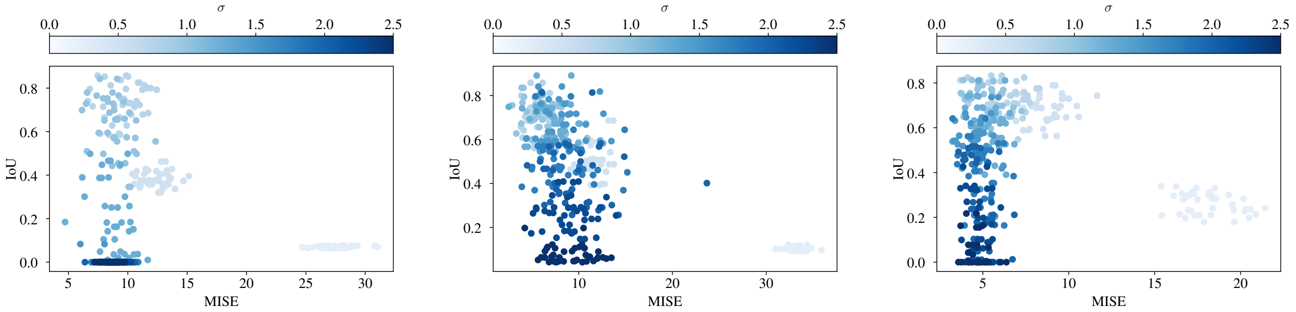
[11] Massimiliano Pastore and Antonio Calcagni. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, 10, 2019.

[12] Miriam Seoane Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, and João Santos. A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. *Information Fusion*, 89:228–253, 2023.

[13] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[14] George R. Terrell and David W. Scott. Variable Kernel Density Estimation. *The Annals of Statistics*, 20(3):1236 – 1265, 1992.

[15] Daniel Westreich and Stephen R. Cole. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677, 02 2010.



(a) vKDE

(b) aKDE

(c) Standard KDE

Figure 9: The relationship between MISE and IoU of dataset in Figure 10d. The shade of the data points indicate the variance of the dataset from which the MISE and IoU were measured.

[16] Kevin Winner, Michael J. Noonan, Christen H. Fleming, Kirk A. Olson, Thomas Mueller, Daniel Sheldon, and Justin M. Calabrese. Statistical inference for home range overlap. *Methods in Ecology and Evolution*, 9(7):1679–1691, 2018.

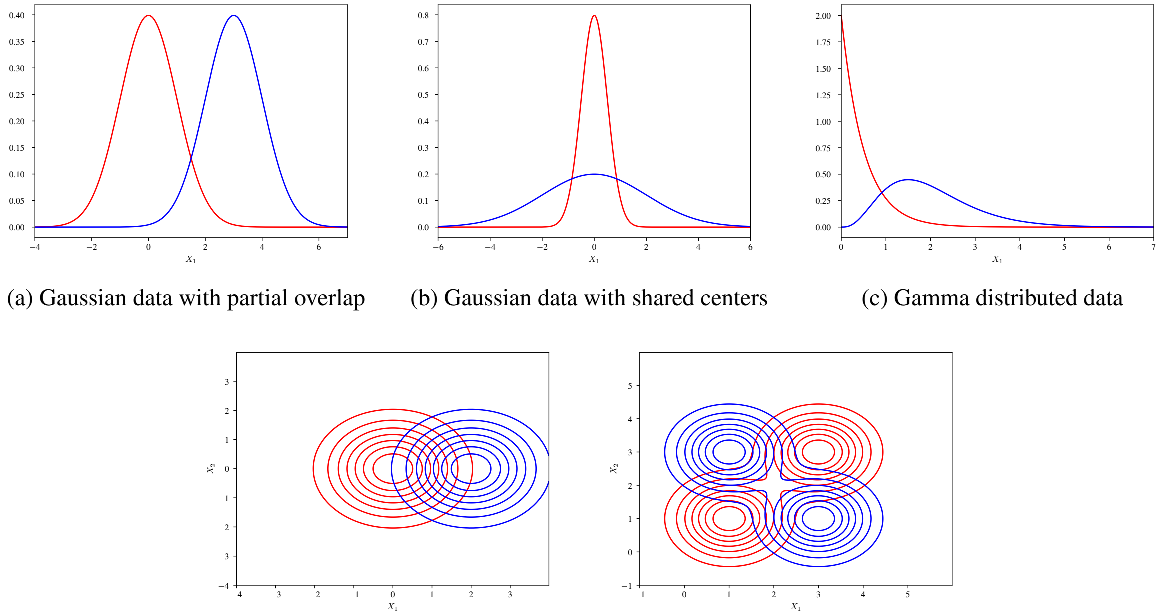
[17] Angela Yaqian Zhu. *Causal Inference Methods For Addressing Positivity Violations And Bias In Observational And Cluster-Randomized Studies*. PhD thesis, University of Pennsylvania, 2022.

[18] Yaqian Zhu, Rebecca A. Hubbard, Jessica Chubak, Jason Roy, and Nandita Mitra. Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and Drug Safety*, 30(11):1471–1485, 2021.

A Appendix

A.1 Datasets

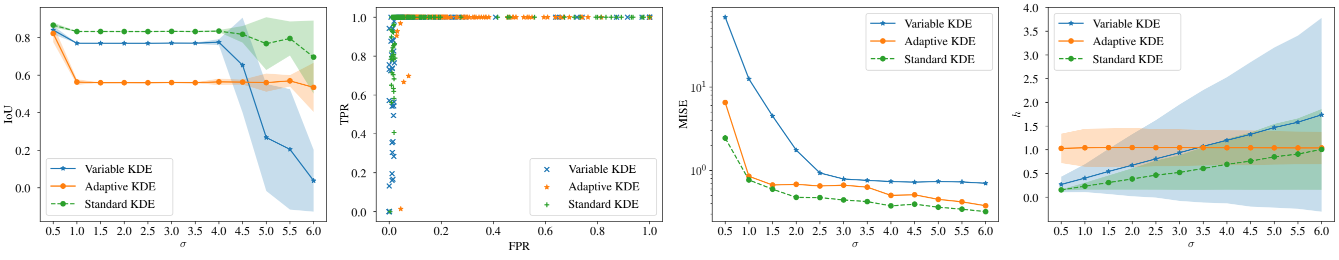
A.2 Additional Experimental Results



(a) Gaussian data with partial overlap (b) Gaussian data with shared centers (c) Gamma distributed data

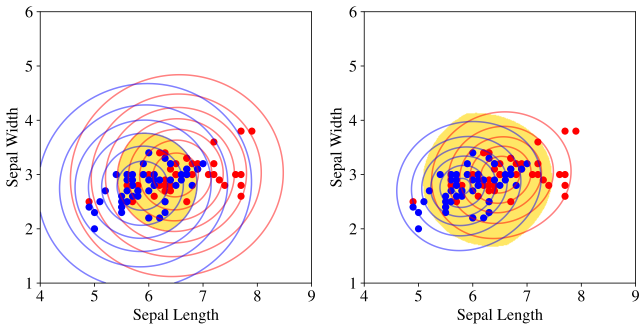
(d) Multivariate Gaussians (e) Mixture Gaussians

Figure 10: True distributions of synthetic datasets in 1D and 2D



(a) IoU (b) ROC (c) MISE (d) Kernel Size h

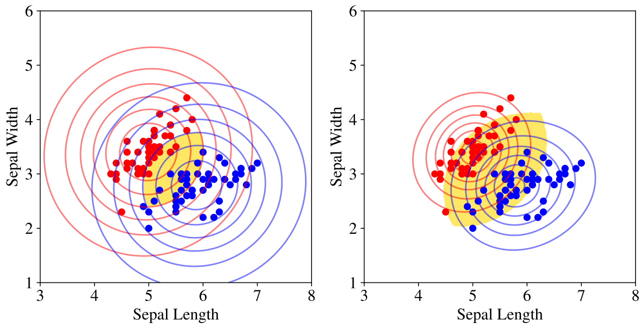
Figure 11: Overlap estimation for $\mathcal{N}(0, 0.25)$ and $\mathcal{N}(0, \sigma)$. The IoU is relatively stable until σ grows too large. The IoU of adaptive starts out low and remains relatively low because its average kernel size is too large for $\mathcal{N}(0, 0.25)$. The average kernel had width of 1 while variance of one of the distributions was only 0.5.



(a) Adaptive KDE

(b) Variable KDE

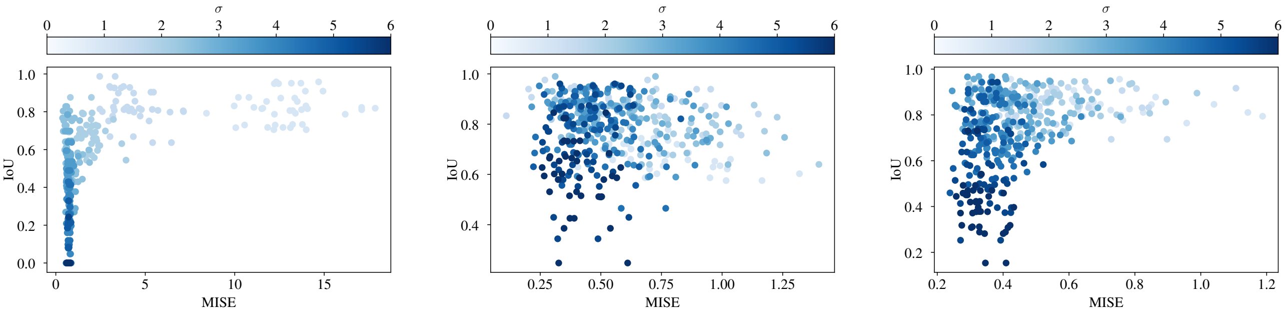
Figure 12: Success case: overlap estimation for the Iris data between Iris Virginica and Iris Versicolor on their sepal width and sepal length.



(a) Adaptive KDE

(b) Variable KDE

Figure 13: Fail case: overlap estimation for the Iris data. The overlap region between Iris Setosa and Iris Versicolor for the features sepal length and sepal width.

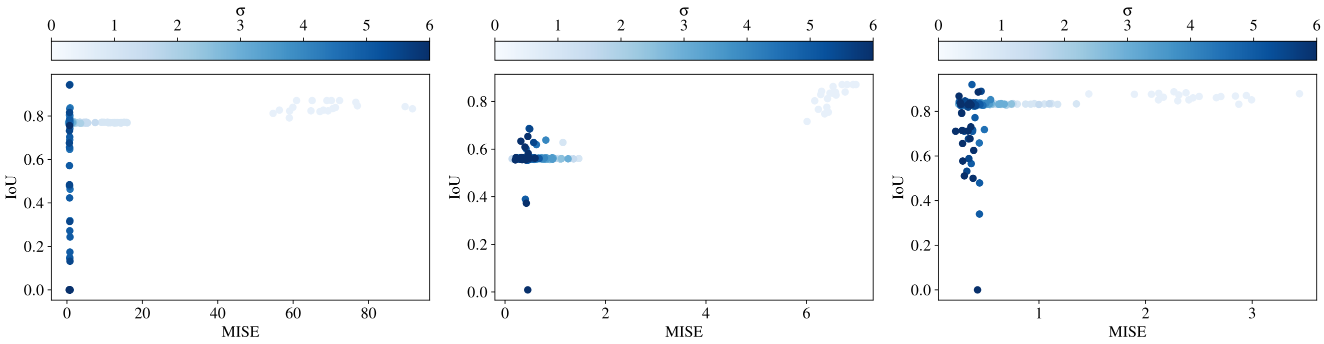


(a) Variable KDE

(b) Adaptive KDE

(c) Standard KDE

Figure 14: 1D Gaussian overlap

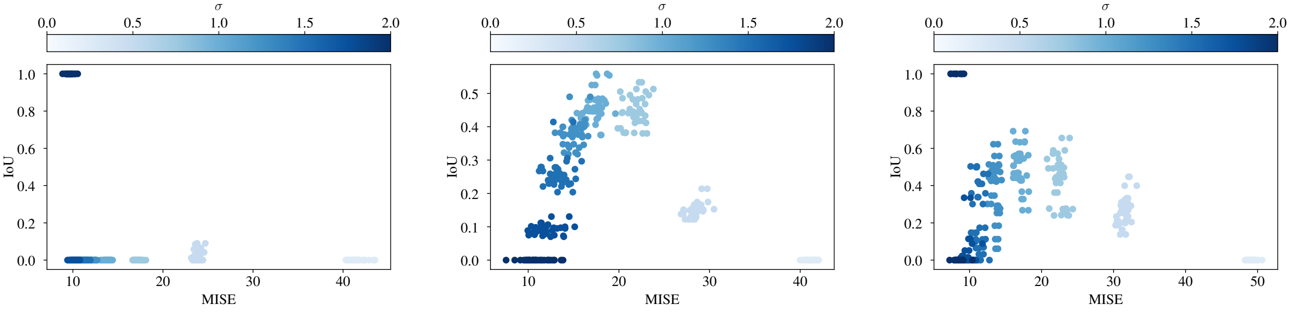


(a) Variable KDE

(b) Adaptive KDE

(c) Standard KDE

Figure 15: MISE and IoU relation for shared centers Gaussians



(a) Variable KDE

(b) Adaptive KDE

(c) Standard KDE

Figure 16: 2D Mixture Gaussian overlap