JIRSS (2023)
Vol. 22, No. 01, pp 1-27
DOI: 10.22034/JIRSS.2024.713811

A Comparison between Parametric and non-parametric Density Estimation based on LCV Variance Estimation

Abdolreza Sayyareh [1]

[1] Department of Computer Sciences and Statistics, K.N.Toosi University of Technology, Tehran, Iran.

Received: 06/07/2021, Accepted: 07/09/2023, Published online: 15/06/2024

Abstract. When the parametric model does not hold, and we cannot fit a parametric model to the data, the true density may be estimated non-parametrically, as in the case of a kernel estimate. The purpose of this paper is to present a comparison between parametric and non-parametric models. The parametric investigation contains Vuong's test, and tracking interval based on the known maximum likelihood estimation theory. The presented non-parametric analysis involves kernel density estimation. Modified differences of Kullback-Leibler criteria between two rival models and Vuong's test, have been considered. In this circumstance, we address the problem of cross-validation estimation of variance for Kullback-Leibler divergence between the true but unknown density and its kernel estimator. A simulation study and data analysis have shown that the parametric density is a more realistic estimate of the data generating density.

Keywords. Akaike Information Criterion, Kernel Density, Likelihood Cross-validation, Time Series, Tracking Interval, Vuong's Test.

MSC: 62E17, 62F03, 62F25.

1 Introduction

In practice, we want to learn about the phenomenon behind our data. Thus, we propose a parametric or non-parametric model. There is uncertainty about selecting a nested or non-nested parametric family for data and estimating its parameters. On the other hand, a non-parametric density estimator does not tell us much more than the data. If we believe the model is approximately correct, we apply parametric inference. But,

Corresponding Author: Abdolreza Sayyareh (asayyareh@kntu.ac.ir).

it is possible the data do not fit well to any member of the family. In that case, it is better to use a kernel density estimator, also known as Parzen's window, which is a non-parametric density estimator and will learn the shape of the density from the data naturally.

In this circumstance, choosing the smooth bandwidth is a classic research topic in non-parametric statistics, known as bandwidth selection. When the bandwidth is too small, there are many wiggles in the density estimate. When the bandwidth is too large, we smooth out important features see Stone (1984) and Silverman (1986). Overviews and comparisons of the existing methods for bandwidth selection can be found in Scott (2015). In contrast to bandwidth selection, the choice of kernel function does not play an essential role in non-parametric density estimation.

In statistical model selection, we encounter two types of errors. The first error is caused by modeling, and the second error is done by estimation of the parameter vector, say θ, where we also encounter what is called the estimation error, namely the bias, and variance. Let R denotes the overall risk, R(M) denotes the risk of modeling or misspecification risk, and R(E) denote the risk of estimation or statistical risk. Then, we can define:

$$\text{Overal Risk= Risk of modeling+ Risk of estimation.}$$

Although the statistical risk, can be rather estimated, the misspecification risk generally cannot.

Statistical inference assumes that there is the true, generally unknown density $h(\cdot)$ as the data generating density. Without any assumptions, $h(\cdot)$ is just a nonnegative, integrable function with an integral equal to one. Statisticians aims to approach $h(\cdot)$.

We say a family of densities or models, $\mathcal{G}$, is well-specified if $h \in \mathcal{G}$ and is misspecified otherways. Therefore, in a parametric case, it is well specified if there exists a $\beta^* \in B$ such that $g^{\beta^*} \in \mathcal{G}$. Sometimes, we consider two or more models, namely rival models. In general,

$$\mathcal{G} = \{g^{\beta}(\cdot); \beta \in B\},$$

and

$$\mathcal{F} = \{f^{\gamma}(\cdot); \gamma \in \Gamma\},$$

are to rival models to estimate $h(\cdot)$. We say, these two models are compatible. So, we faced with the amount of risk. If the model is well-specified, the R(M) is zero.

The loss using $g^{\beta}(y)$ in place of $h(y)$ for Y a random variable which is independent and identically distributed from random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is the log-likelihood ratio $\log \frac{h(\mathbf{y})}{g^{\beta}(\mathbf{y})}$ and its expectation is defined as Kullback-Leibler, $\mathcal{KL}$, divergence:

$$\begin{aligned}\mathcal{KL}(h, g) &= E_h\left\{\log \frac{h(Y)}{g^{\beta}(Y)}\right\} \\ &= E_h\{\log h(Y)\} - E_h\{\log g^{\beta}(Y)\}.\end{aligned}$$

Denote by β_0 the value that minimizes $\mathcal{KL}(h, g)$. If $h(y)$ is well-specified, we have $\beta_0 = \beta^*$. For this value of β, we have $E_h\left\{\log \frac{h(Y)}{g^{\beta_0}(Y)}\right\}$ as the amount of $\mathcal{KL}$ risk. This risk is a misspecification risk, the risk of selection of an unsuitable approximation for $h(y)$. The standard method for estimating β_0 is by maximum likelihood estimation. In this method, the estimator is any value of β amounts to solving,

$$\max_{\beta \in B} \prod_{i=1}^{n} g^{\beta}(y_i) \quad \text{subject to} \quad \beta \in B.$$

As it is known, the maximum likelihood estimation problem is equivalent to,

$$\min_{\beta \in B} \quad -\int_{\Re} \log g^{\beta}(y) dG_n(y) \quad \text{subject to} \quad \beta \in B,$$

where $G_n(y)$ is the empirical distribution function related to the assumed family. Second, if, for whatever reason, a parametric model for $h(\cdot)$ is not available, we are faced with a non-parametric estimation problem. Without any assumptions, $h(\cdot)$ is just a nonnegative, integrable function with an integral equal to one. Thus, we are dealing with an infinite dimensional problem, and infinitely many parameters are required to describe $h(\cdot)$.

Various approaches for kernel estimation proposed that they can be divided into three classes. The first class is to learn the kernel parameters given a parameterized kernel function, see Stone (1974) and Hastie et al. (2001) for one of the most commonly used kernel parameter selection methods, the cross-validation technique. For classification, as Fisher had proposed, Xiong et al. (2005) and Wang et al. (2008) define such a criterion, which maximizes the between-class scatter and minimizes the within-class scatter in the kernel space.

In the second class of approaches, a kernel matrix is directly learned without prespecifying a kernel function, and a positive semi-definiteness constraint has to be imposed. This approach needs to specify a known parameterized kernel function. In fact, given the observed data, one usually does not have prior knowledge for which kernel function should be selected, see Liu et al. (2009) and Yeung et al. (2007).

In the third class, some new kernel functions are proposed for the problem available. It has shown that the popularly used kernels, such as the Gaussian kernel, have successful performance in some applications, but they have some limitations, see Jaakkola et al. (1999). Moreno et al. (2003) have developed a Kullback-Leibler divergence-based kernel for the use of multimedia applications.

Hjort and Jones (1996) developed a non-parametric density estimator. They defined a local kernel-smoothed likelihood function which, for each observation, can be used to estimate the best local parametric approximation to the true density, They showed that, when the bandwidth used is large enough, this amounts to full likelihood parametric density estimation, and otherwise, the method is essentially non-parametric.

An intrinsic aspect of non-parametric density estimation is that there is no goodness of fit tests.

The common approaches to bandwidth selection include the rule of thumb, least square cross-validation, and plug-in method, see Goldenshluger (2011). Lacour et al. (2017) explained some general ideas about the calibration issue of estimator selection methods, which has become a crucial issue in non-parametric estimation used in penalized log likelihood estimation and pairwise comparison. They presented a new approach for bandwidth selection for kernel density estimation.

Section 2 contains the general aspects of the model selection idea. In section 3 we have brought the theoretical results about the new introduced variance, and we have extended Vuong's test, and the tracking interval to compare a parametric and a non-parametric rival models. Sections 4 and 5 have considered simulation and real data analysis to verify the theoritical result. Section 6 contains conclusions.

2 Model Selection

The most important risk function is based on the known risk, Kullback-Leibler's(1959) divergence. Maximum likelihood estimators, use of some information criteria like the Akaike criterion, can be grounded on the Kullback-Leibler's divergence. For a well-specified model misspecification risk is zero. On the other hand, the value of β_0 is unknown, and we should search for a $g^{\beta}(y)$ close to $h(y)$, that is, one that minimizes an estimator of $\mathcal{KL}$ risk. Words, we should estimate it, say $\hat{\beta}_n$ the maximum likelihood estimator of β, which minimizes an estimator of $\mathcal{KL}(h, g)$,

$$\begin{aligned}\mathcal{KL}(h, g^{\hat{\beta}_n}) &= E_h\left\{\log\frac{h(Y)}{g^{\hat{\beta}_n}(Y)}\right\}\\ &= E_h\{\log h(Y)\} - E_h\{\log g^{\hat{\beta}_n}(Y)\}. \end{aligned} \tag{2.1}$$

The other kind of risk is realated to the divergence between $g^{\beta_0}(y)$ and $g^{\hat{\beta}_n}(y)$. This risk is a statistical risk, and we denote it by $E_h\left\{\log\frac{g^{\beta_0}(Y)}{g^{\hat{\beta}_n}(Y)}\right\}$. In fact, we want to compute the risk of using the estimated density $g^{\hat{\beta}_n}(y)$ instead of the true but unknown $h(y)$. From (2.1), the expected Kullback-Leibler, $E\mathcal{KL}$, risk is given by,

$$\begin{aligned} E\{\mathcal{KL}(h, g^{\hat{\beta}_n})\} &= E\mathcal{KL}(h, g^{\hat{\beta}_n})\\ &= E_h\left\{\log\frac{h(Y)}{g^{\hat{\beta}_n}(Y)}\right\} = E_h\left\{\log\frac{h(Y)}{g^{\beta_0}(Y)}\right\} + E_h\left\{\log\frac{g^{\beta_0}(Y)}{g^{\hat{\beta}_n}(Y)}\right\}\\ &= E\mathcal{KL}(h, g^{\beta_0}) + E\mathcal{KL}(g^{\beta_0}, g^{\hat{\beta}_n})\\ &= \mathcal{KL}(h, g^{\beta_0}) + E\mathcal{KL}(g^{\beta_0}, g^{\hat{\beta}_n}). \end{aligned} \tag{2.2}$$

As we have expected, the overall risk is the sum of the misspecification risk and the statistical risk and expected Kullback-Leibler refers to the two folded expectations on Y, and **Y**, because the second term on the right-hand side contains $\hat{\beta}_n$ which is a function of **Y**. All expectations are on $h(\cdot)$. In (2.2), the first term in the right is bise and the second term is variance in model selection. Consider the case where both models are well-specified, so the R(M) is zero, and we have to estimate the R(E).

Let $\hat{h}(\cdot)$ is kernel estimate of $h(\cdot)$, where for bandwidth w and kernel K we have,

$$\hat{h}(y|w) = (nw)^{-1} \sum_{j=1}^{n} K\left\{\frac{y - Y_j}{w}\right\}.$$

where $K(u)$ is a symmetric function, with $\int K(u)du = 1$ and $K(0) = 1$. w is bandwidth or smoothing parameter. The kernel function $K(w^{-1}(y - Y_j))$ with center y, gives more weight to sample points near y.

Here, the loss function is $\log \frac{h(y)}{\hat{h}(y)}$ and the $E\mathcal{KL}$ risk for $h(y)$ and $\hat{h}(y)$ is given by,

$$E\mathcal{KL}(h, \hat{h}) = E_h\left\{\log \frac{h(Y)}{\hat{h}(Y|w)}\right\}. \tag{2.3}$$

We may write the $E\mathcal{KL}(h, \hat{h})$ based on the bias and variance. To do this note that:

$$\log \frac{h(Y)}{\hat{h}(Y|w)} = \log \frac{h(Y)}{E_h\{\hat{h}(Y|w)\}} + \log \frac{E_h\{\hat{h}(Y|w)\}}{\hat{h}(Y|w)},$$

where the expectation is taken on **Y**. Thus (2.3) becomes,

$$\begin{aligned} E\mathcal{KL}(h, \hat{h}) &= E\left\{\log \frac{h(Y)}{E\{\hat{h}(Y|w)\}} + \log \frac{E\{\hat{h}(Y|w)\}}{\hat{h}(Y|w)}\right\} \\ &= E_h\left\{\log \frac{h(Y)}{E\{\hat{h}(Y|w)\}}\right\} + E_h\left\{\log \frac{E\{\hat{h}(Y|w)\}}{\hat{h}(Y|w)}\right\}, \end{aligned} \tag{2.4}$$

which we could show (2.4) as,

$$E\mathcal{KL}(h, \hat{h}) = \int_{\mathfrak{R}} h(y) \log \frac{h(y)}{E_h\{\hat{h}(y|w)\}} dy + \int_{\mathfrak{R}} h(y) E\left[\log\left\{\frac{E_h\{\hat{h}(y|w)\}}{\hat{h}(y|w)}\right\}\right] dy.$$

The expected $\mathcal{KL}$ risk has an expansion as:

$$E\{\mathcal{KL}(h, g^{\hat{\beta}_n})\} = E_h\{\log h(Y)\} - E_h\{n^{-1} \sum_{i=1}^{n} \log g^{\hat{\beta}_n}(Y_i)\} + n^{-1} Tr(I_g^{-1} J_g) + o_p(n^{-1}), \tag{2.5}$$

where $E_h\{n^{-1} \sum_{i=1}^{n} \log g^{\hat{\beta}_n}(Y_i)\}$ is an estimatior of $E_h\{\log g^{\hat{\beta}_n}(Y)\}$, $I_g = -E_h\{\frac{\partial^2 \log g^{\beta}(y)}{\partial \beta^2}\}$ and $J_g = E_h\{[\frac{\partial \log g^{\beta}(y)}{\partial \beta}]|_{\beta_0}\} E_h\{[\frac{\partial \log g^{\beta}(y)}{\partial \beta}|_{\beta_0}]^T\}$. Takeuchi information criterion, TIC, follows

from this expansion if replacing $E_h\{n^{-1}\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)\}$ by $n^{-1}\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)$, multiplying by 2n and deleting the constant term $E_h\{\log h(Y)\}$. Thus,

$$TIC = -2\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i) + 2Tr(I_g^{-1}J_g).$$

In a well-specified case we know $I_g = J_g$ and $Tr(I_g^{-1}J_g) = Tr(I_p) = p$ where I_p is the identity matrix and p is the dimension of B, the space of β. Thus Akaike information criterion, AIC, Akaike (1973), is given by,

$$AIC = -2\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i) + 2p.$$

Introducing AIC for well-specified models is prior to TIC which introduced in 1976. AIC is known as an estimator for the expected $\mathcal{KL}$ risk. The kernel estimate of $h(\cdot)$ has an estimator which introduced by Habbema et al. (1974). As we saw, the expected $\mathcal{KL}$ between $h(\cdot)$ and $\hat{h}(\cdot)$ is given by,

$$E\mathcal{KL}(h,\hat{h}) = E_h\{\log h(Y)\} - E_h\{\log \hat{h}(Y)\}.$$

The first term is irrelevant for optimization over $\hat{h}$. So, we approximate the expectation in the second term with the empirical version of the integral,

$$E_h\{\log \hat{h}(Y)\} \approx \frac{1}{n}\sum_{i=1}^{n}\log \hat{h}(Y_i).$$

If we replace the kernel density estimate in the empirical version of the relevant part of expected $\mathcal{KL}$, we have:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\log \hat{h}(y_i) &= \frac{1}{n}\sum_{i=1}^{n}\log\left((nw)^{-1}\sum_{j=1}^{n}K\left\{\frac{y_i - Y_j}{w}\right\}\right)\\ &= -\log nw + \frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{j=1}^{n}K\left\{\frac{y_i - Y_j}{w}\right\}\right).\end{aligned}$$

For very small $h(\cdot)$, we have $K\left\{\frac{y_i - Y_j}{w}\right\}$ unless $y_i = y_j$. So,

$$\frac{1}{n}\sum_{i=1}^{n}\log \hat{h}(y_i) \approx -\log nw + \log K(0),$$

which tends to infinity as w goes to zero. To maximize the likelihood of a kernel density estimate, we have to make the bandwidth as small as possible. Based on the limit, using the Dirac delta density function, the density is given by,

$$\hat{h}(y) = \frac{1}{n}\sum_{j=1}^{n}\delta(y - y_j),$$

which is an unrestricted maximum likelihood estimate of the distribution. This approximation exactly is what we request from the likelihood idea because we have used the raw empirical cumulative distribution function, and the approximation is an unrestricted maximum likelihood estimate of the density. One way out of it is cross validated log-likelihood to select a bandwidth. The likelihood cross-validation, LCV, see Marron(1985), to estimate w is,

$$CV(w) = n^{-1} \sum_{i=1}^{n} \log \hat{h}_i(Y_i|w),$$

where $\hat{h}_i(Y_i|w) = \{(n-1)w\}^{-1} \sum_{j \neq i} K\left\{\frac{y-Y_j}{w}\right\}$. The bandwidth $\hat{w}_0$ which maximizes $CV(w)$ does not depend on n^{-1} or $(n-1)^{-1}$ in the definition of $\hat{h}_i(.)$. We denote $E\mathcal{KL}(h, \hat{h})$ by $E\mathcal{KL}_{n-1}(h, \hat{h})$ to emphesize that we have considered sample size equal to $n-1$ instead of n in definition of $\hat{h}$. From (2.3) we have,

$$E\mathcal{KL}_{n-1}(h, \hat{h}) = E_h \{\log h(Y)\} - E_h \{CV(w)\}. \tag{2.6}$$

Habbema et al.(1974) selects the bandwidth w that maximizes $LCV(w) = \prod_{i=1}^{n} h_i(Y_i|w)$. For kernel estimation, this is equivalent to minimizing,

$$LCV(w) = -\sum_{i=1}^{n} \log \hat{h}(Y_i) - \sum_{i=1}^{n} \log\left(1 - \frac{k(0)}{nw\hat{h}(Y_i)}\right).$$

The asymptotically optimal bandwidth is given by

$$w_{opt} = \left(\frac{c_1(k)}{nc_0(k)^2 \int_{-\infty}^{\infty} h^{(2)}(y)} dy\right)^{1/5},$$

where $c_0(k) = \int_{-\infty}^{\infty} v^2 k(v) dv$ and $c_1(k) = \int_{-\infty}^{\infty} k^2(v) dv$.

An Akaike style criterion based on cross-validation is given by,

$$AIC_{cv} = \sum_{i=1}^{n} \log \hat{h}(Y_i) - \sum_{i=1}^{n} \frac{k(0)}{nw\hat{h}(Y_i)},$$

see Loader(1999). To compair $g^{\hat{\beta}_n}(\cdot)$ and $\hat{h}(\cdot)$ as parametric estimators and kernel estimators of true but unknown data generating density $h(.)$ we have considered the differences of the expected $\mathcal{KL}$ risks as defined in the minimization of $E\mathcal{KL}_{n-1}(h, \hat{h})$ is equal to maximization the relevant part in $\mathcal{KL}$ risk, which is here $E\{CV(w)\}$. Define the difference between $E\mathcal{KL}(h, g^{\hat{\beta}_n})$ and $E\mathcal{KL}(h, \hat{h})$ by $\Delta(g^{\hat{\beta}_n}, \hat{h})$,

$$\Delta(g^{\hat{\beta}_n}, \hat{h}) = E\mathcal{KL}(h, g^{\hat{\beta}_n}) - E\mathcal{KL}(h, \hat{h}).$$

From (2.5) and (2.6) we have:

$$\begin{aligned}
&\Delta(g^{\hat{\beta}_n}, \hat{h}) \\
&= E_h\{\log h(Y)\} - E_h\{n^{-1}\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)\} + n^{-1}Tr(I_g^{-1}J_g) + o_p(n^{-1}) \\
&- \{E_h\{\log h(Y)\} - E_h\{CV(w)\}\} \\
&= -E_h\{n^{-1}\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)\} + n^{-1}Tr(I_g^{-1}J_g) + o_p(n^{-1}) + E_h\{CV(w)\} \\
&= -n^{-1}[E_h\{\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)\} - Tr(I_g^{-1}J_g)] + E_h\{CV(w)\} + o_p(n^{-1})
\end{aligned}$$

Thus we have,

$$-n^{-1}\left[E_h\left\{\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)\right\} - Tr(I_g^{-1}J_g)\right] + E_h\{CV(w)\} = \Delta(g^{\hat{\beta}_n}, \hat{h}) + o_p(n^{-1}). \tag{2.7}$$

Using Takeuchi and Akaike information criteria as the estimators of $E\mathcal{KL}(h, g^{\hat{\beta}_n})$ for the two first terms in (2.7) and $\hat{E}_h\{CV(w)\} = n^{-1}\sum_{i=1}^{n}\log\hat{h}_i(Y_i|w)$ as a estimator for $E_h\{CV(w)\}$. We have two estimations for $\Delta(g^{\hat{\beta}_n}, \hat{h})$, namely $D_{TIC}(g^{\hat{\beta}_n}, \hat{h})$ and $D_{AIC}(g^{\hat{\beta}_n}, \hat{h})$, respectively as,

$$\begin{aligned}
D_{TC}(g^{\hat{\beta}_n}, \hat{h}) &= \frac{1}{2}n^{-1}TIC(g^{\hat{\beta}_n}) + CV(w) \\
&= -n^{-1}\left[\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i) - Tr(I_g^{-1}J_g)\right] + CV(w),
\end{aligned}$$

and

$$\begin{aligned}
D_{AC}(g^{\hat{\beta}_n}, \hat{h}) &= \frac{1}{2}n^{-1}AIC(g^{\hat{\beta}_n}) + CV(w) \\
&= -n^{-1}\left[\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i) - p\right] + CV(w).
\end{aligned}$$

From (2.7), we may write $\Delta(g^{\hat{\beta}_n}, \hat{h}) + o_p(n^{-1})$ as following,

$$\Delta(g^{\hat{\beta}_n}, \hat{h}) + o_p(n^{-1}) = E_h\left\{-\frac{1}{n}\sum_{i=1}^{n}\log\frac{g^{\hat{\beta}_n}(Y_1)}{\hat{h}_i(Y_i|w)}\right\} + \frac{1}{n}Tr(I_g^{-1}J_g).$$

Note that,

$$D_{TC}(g^{\hat{\beta}_n}, \hat{h}) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{g^{\hat{\beta}_n}(Y_i)}{\hat{h}_i(Y_i|w)} + \frac{1}{n}Tr(I_g^{-1}J_g), \tag{2.8}$$

and

$$D_{AC}(g^{\hat{\beta}_n}, \hat{h}) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{g^{\hat{\beta}_n}(Y_i)}{\hat{h}_i(Y_i|w)} + \frac{p}{n}. \tag{2.9}$$

The relations (2.7) and (2.9) suggest that $\sqrt{n}(D_{AC}(g^{\hat{\beta}_n}, \hat{h}) - \Delta(g^{\hat{\beta}_n}, \hat{h}))$ as a random variable, has zero mean and the variance that we denote by ω_*^2, where ω_*^2 is equal to

$$\omega_*^2 = var\left\{\log \frac{g^{\beta}(Y)}{h(Y|w)}\right\},$$

where Y is a random variable dependent and with the same distribution as $Y_1, \ldots, Y_n$ and has an estimator as,

$$\hat{\omega}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\log \frac{g^{\hat{\beta}_n}(Y_i)}{\hat{h}(Y_i|w)}\right\}^2 - \left\{\frac{1}{n}\sum_{i=1}^{n} \log \frac{g^{\hat{\beta}_n}(Y_i)}{\hat{h}(Y_i|w)}\right\}^2. \tag{2.10}$$

(2.10) is a natural estimator for ω_*^2. We can use (2.8) instead of $D_{AC}(g^{\hat{\beta}_n}, \hat{h})$ with small modifications.

Confidence Set

Confidence regions of $h(\cdot)$ are random intervals, say, $C_{1-\alpha}(y)$ derived from observations, and cover the true value of $h(y)$ with probability at least $1-\alpha$. For a given point y,

$$\sqrt{nw}(\hat{h}(y) - E\{\hat{h}(Y)\}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(y)),$$

where $\sigma^2(y) = \int K^2(u)du.h(y)$. This equation implies that an approach to construct a confidence band is to use asymptotic normality with a variance estimator. A simple approach is replacing $h(y)$ by $\hat{h}(y)$, $\hat{\sigma}^2(y) = \int K^2(u)du.\hat{h}(y)$. An alternative method is to estimate the asymptotic variance using the bootstrap.

In the next section, we compute the underlying variance with a different approach based on the likelihood cross-validation, LCV, suitable for constructing the tracking interval.

3 Main Result

The unknown density has to be approximated using a process known as density estimator. There are three known approaches histogram, parametric probability density estimation involves selecting a usual parametric distribution for data, and non-parametric density estimation like kernel density estimation. Here we consider the nested and non-nested parametric density estimator and the kernel density estimation.

$\Delta(g^{\hat{\beta}_n}, \hat{h})$ can be decomposed as the sum of the difference of the R(M) and the difference of R(E). So,

$$\Delta(g^{\hat{\beta}_n}, \hat{h}) = \Delta_{R(M)}(g^{\hat{\beta}_n}, \hat{h}) + \Delta_{R(E)}(g^{\hat{\beta}_n}, \hat{h}).$$

On the other hand, $\omega_*^2 \approx var(\log g^{\beta}(Y)) + var(CV(w))$, which has an estimator as,

$$\hat{\omega}_{nc}^2 \approx \hat{var}(\log g^{\beta}(Y)) + \hat{var}(CV(w)). \tag{3.1}$$

Set $V = var(\log g^{\beta}(Y))$. So,

$$V = E\{[\log g^{\beta}(Y)]^2\} - E^2\{\log g^{\beta}(Y)\},$$

has an estimator as,

$$\hat{V}n = n^{-1}\sum_{i=1}^{n}[\log g^{\hat{\beta}_n}(Y_i)]^2 - [n^{-1}\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)]^2. \tag{3.2}$$

The vaiance of $CV(w)$ will be computed based on the variance of $\hat{h}(.)$.

$$\begin{aligned}
\hat{var}(CV(w)) &= var\left(n^{-1}\sum_{i=1}^{n}\log\hat{h}_i(Y_i|w)\right)\\
&= var\left(n^{-1}\sum_{i=1}^{n}\log\left\{\{(n-1)w\}^{-1}\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}\right)\\
&= var\left(n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}\right)\\
&= n^{-2}\sum_{i=1}^{n}var\left(\log\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}\right),
\end{aligned}$$

Based on the delta method,

$$var\left(\log\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}\right) \approx \frac{var\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}}{\left\{E\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}\right\}^2}.$$

Thus,

$$\begin{aligned}
\hat{var}(CV(w)) &\approx n^{-2}\sum_{i=1}^{n}\frac{var\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}}{\left[E\left\{\sum_{j\neq i}K\left\{\frac{y-Y_i}{w}\right\}\right\}\right]^2}\\
&= n^{-2}\sum_{i=1}^{n}\frac{\sum_{j\neq i}var\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}}{\left[\sum_{j\neq i}E\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}\right]^2}\\
&= n^{-2}n\frac{n-1}{(n-1)^2}\frac{var\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}}{\left[E\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}\right]^2}\\
&= \frac{1}{n(n-1)}\frac{var\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}}{\left[E\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}\right]^2}.
\end{aligned} \tag{3.3}$$

But,

$$
\begin{aligned}
var\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\} &= E\left\{\left[K\left\{\frac{y-Y_i}{w}\right\}\right]^2\right\} - \left\{E\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}\right\}^2 \\
&= \int_{y_1} K^2\left\{\frac{y-y_1}{w}\right\} f(y_1)\,dy_1 - \left\{\int_{y_1} K\left\{\frac{y-y_1}{w}\right\} f(y_1)\,dy\right\}^2 \\
&= \int_v wK^2(v)f(y-vw)\,dv - \left\{\int_v wK(v)f(y-vw)\,dv\right\}^2 \\
&= w\int_v K^2(v)f(y-vw)\,dv - w^2\left\{\int_v K(v)f(y-vw)\,dv\right\}^2 \\
&= w\int_v K^2(v)[f(y) - vwf^{(1)}(y) + O(w^2)]\,dv \\
&- w^2\left\{\int_v K(v)[f(y) - vwf^{(1)}(y) + \frac{1}{2}v^2w^2f^{(2)}(y) + O(w^3)]\,dv\right\}^2 \\
&= w\left[f(y)\int_v K^2(v)\,dv - wf^{(1)}(\zeta)\int_v vK^2(v)\,dv + O(w^2)\right] \\
&- w^2\left\{f(y)(1) - wf^{(1)}(y)(0) + \frac{1}{2}w^2f^{(2)}(y)\int_v K(v)v^2\,dv + O(w^3)\right\}^2 \\
&= w\left[f(y)\int_v K^2(v)\,dv - O(w\int_v |v|K^2(v)\,dv) + O(w^2)\right] \\
&- w^2\left\{f(y) + \frac{1}{2}w^2f^{(2)}(y)\int_v K(v)v^2\,dv + O(w^3)\right\}^2, \qquad (3.4)
\end{aligned}
$$

where

$$\frac{1}{3!}w^3\left|\int f^{(3)}(\bar{y})v^3k(v)\,dv\right| \le Cw^3\left|\int v^3k(v)\,dv\right| = O(w^3),$$

and C is a constant. By symmetry condition on v we have $\int_v vK(v)\,dv = 0$.

$$\left[E\left\{K\left\{\frac{y-Y_i}{w}\right\}\right\}\right]^2 = O(w^2).$$

From (3.3) and (3.4) we ave:

$$
\begin{aligned}
&\hat{var}(CV(w)) \\
&\approx \frac{1}{n(n-1)}\frac{wf(y)\int_v K^2(v)\,dv + O(w^3) - w^2\left\{f(y) + \frac{1}{2}w^2f^{(2)}(y)\int_v K(v)v^2\,dv + O(w^3)\right\}^2}{w^2\left\{f(y) + \frac{1}{2}w^2f^{(2)}(y)\int_v K(v)v^2\,dv + O(w^3)\right\}^2} \\
&= \frac{1}{n(n-1)}\frac{wf(y)\int_v K^2(v)\,dv + O(w^3)}{w^2\left\{f(y) + \frac{1}{2}w^2f^{(2)}(y)\int_v K(v)v^2\,dv + O(w^3)\right\}^2} - 1.
\end{aligned}
$$

Since $f(.)$ is a probability density, integrating on numerator and denumerator both over y, gives the simple approximation. Thus we have:

$$\begin{aligned} v\hat{a}r(CV(w)) &\approx \frac{1}{n(n-1)}\left[\frac{w\int_v K^2(v)\,dv + O(w^3)}{w^2\{1+O(w^3)\}^2} - 1\right] \\ &= \frac{1}{n(n-1)}\left[\frac{\int_v K^2(v)\,dv + O(w^2)}{w} - 1\right]. \end{aligned} \tag{3.5}$$

Subestuting (3.2) and (3.5) in (3.1) we have:

$$\begin{aligned} \hat{\omega}^2_{nc} &\approx \hat{V}_n + v\hat{a}r(CV(w)) \\ &= n^{-1}\sum_{i=1}^{n}[\log g^{\hat{\beta}_n}(Y_i)]^2 - [n^{-1}\sum_{i=1}^{n}\log g^{\hat{\beta}_n}(Y_i)]^2 \\ &+ \frac{1}{n(n-1)}\left[\frac{\int_v K^2(v)\,dv + O(w^2)}{w} - 1\right]. \end{aligned} \tag{3.6}$$

Thus (3.1) and (3.6) are two competing estimators for ω^2_*.

3.1 Vuong's Test

Vuong's (1989) test is constructed to test $(g^{\beta}(.))_{\beta\in B}$ against $(f^{\gamma}(.))_{\gamma\in\Gamma}$, using the $\mathcal{KL}$ divergence as a closeness measure. The focus of this approach is to test the hypothesis that the models under consideration are equally close to the true unknown model, h. The null hypothesis of the Vuong's test is

$$\mathcal{H}_0 : \mathcal{E}_h\left\{\log\frac{g_Y^{\beta_0}(Y)}{f_Y^{\gamma_0}(Y)}\right\} = 0,$$

against

$$\mathcal{H}_1 : \mathcal{E}_h\left\{\log\frac{g_Y^{\beta_0}(Y)}{f_Y^{\gamma_0}(Y)}\right\} > 0, \quad \text{or} \quad \mathcal{H}_1 : \mathcal{E}_h\left\{\log\frac{g_Y^{\beta_0}(Y)}{f_Y^{\gamma_0}(Y)}\right\} < 0.$$

Define, $\mathcal{LR}_n^{g/f} = n^{-1}\log\frac{g_Y^{\hat{\beta}_n}(Y)}{f_Y^{\hat{\gamma}_n}(Y)}$, under $\mathcal{H}_0$

$$\mathcal{V}_n^0 = \frac{\mathcal{LR}_n^{g^{\hat{\beta}_n}/f^{\hat{\gamma}_n}}}{\sqrt{n}\hat{\omega}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where

$$\hat{\omega}^2_n = \frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{g_Y^{\hat{\beta}_n}(Y)}{f_Y^{\hat{\gamma}_n}(Y)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{g_Y^{\hat{\beta}_n}(Y)}{f_Y^{\hat{\gamma}_n}(Y)}\right]^2. \tag{3.7}$$

We know that $\hat{\omega}_n$ is a consistent estimator for $\omega_* = \left\{ \mathcal{V}ar_h \left[\log \frac{g_Y^{\beta_0}(Y)}{f_Y^{\gamma_0}(Y)} \right] \right\}^{1/2}$.

Here we substitute $\hat{h}$ for $f_Y^{\gamma_0}(Y)$, to get the Vuong's test for comparison $g_Y^{\beta_0}(.)$ and $\hat{h}(.)$ as two rival models. To provide a simple test for model selection, one chooses a critical value c from the standard normal distribution for some significance level. If the value of the statistic $\mathcal{V}_n^0$ is higher than c then one rejects the null hypothesis in favor of $(g^{\beta}(.))_{\beta \in B}$ being better than $(f^{\gamma}(.))_{\gamma \in \Gamma}$. If $\mathcal{V}_n^0$ is smaller than $-c$ then one rejects the null hypothesis in favor of $(f^{\gamma}(.))_{\gamma \in \Gamma}$ being better than $(g^{\beta}(.))_{\beta \in B}$. Finally if $|\mathcal{V}_n^0| < c$ then one cannot discriminate between the two competing models.

Sayyareh (2012) studied the results of Vuong's test, Cox's test, Akaike's information criterion, Bayesian information criterion, Kullback information criterion and biased corrected Kullback information criterion and the ability of these tests to discriminate between non-nested models, based on stochastic and parametric simulations.

3.2 Tracking Interval

Since the value of Akaike information criteria has no direct interpretation, normalization of a difference of Akaike information criteria for estimating the difference of expected Kullback-Leibler divergences between maximum likelihood estimators of the distributions of two different models is proposed. The variability of this statistic, can estimated. Thus an interval can be constructed which contains the true difference of expected Kullback-Leibler divergences. Commenges et al (2008) proposed the tracking interval. Panahi and Sayyareh (2014) obtained the tracking interval under Type II hybrid censoring scheme.
In this subsection, we obtain the tracking interval based on the two rival models. Based on the normality of $\sqrt{n}(D_{AC}(g^{\hat{\beta}_n}, \hat{h}) - \Delta(g^{\hat{\beta}_n}, \hat{h}))$, for $g^{\beta} \neq h$, we have

$$\frac{\sqrt{n}(D_{AC}(g^{\hat{\beta}_n}, \hat{h}) - \Delta(g^{\hat{\beta}_n}, \hat{h}))}{\hat{\omega}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

So we can compute the tracking interval as

$$n^{-1}\mathcal{LR}_n^{g/h} - Z_{\alpha/2}\hat{\omega}_n < \triangle(g^{\hat{\beta}}, \hat{h}) < n^{-1}\mathcal{LR}_n^{g/h} + Z_{\alpha/2}\hat{\omega}_n,$$

where $\hat{\gamma}^B$ and $\hat{\beta}^B$ converge to γ_* and β_* in probability, respectively, and the logarithm function is a continuous function. This is not a usual confidence interval because $\triangle(f^{\hat{\gamma}_n}, g^{\hat{\beta}_n})$ changes with n. This interval has the property

$$P_h(A_n < \triangle(g^{\hat{\beta}}, \hat{h}) < B_n) \longrightarrow 1 - \alpha,$$

where $A_n = n^{-1}\mathcal{LR}_n^{g/h} - Z_{\alpha/2}\hat{\omega}_n$, $B_n = n^{-1}\mathcal{LR}_n^{g/h} + Z_{\alpha/2}\hat{\omega}_n$, $1 - \Phi(z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and Φ is the standard normal distribution. When the tracking interval contains zero, it means two competing models are equivalent, otherwise it selects one of them as the optimal.

Note

In a comparison between $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$, we note that the variance component for model selection based on the likelihood functions of two general models as P^θ and P^η is given by $\sigma_k^2 = var\{\log P^\theta - \log P^\nu\}$. Based on (3.7), an estimation of the variance of Vuong's statistic(usual variance) is $\hat{\omega}_n^2 = \frac{1}{n}\sum_{i=1}^n \left[\log \frac{P^{\hat{\theta}}}{P^{\hat{\eta}}}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^n \log \frac{P^{\hat{\theta}}}{P^{\hat{\eta}}}\right]^2$.

If we set $P^\eta = h(y|w)$ as a kernel density, based on (3.7), we should compute the variance of Vuong's statistics for $\log P^\theta - \log \hat{h}(\cdot|w)$. For the kernel density estimator, $\hat{h}(\cdot)$, based on the equations (2.6) and (3.1), we found the variance of Vuong's statistics as, $\sigma_n^2 = \hat{\omega}_n^2 = \frac{1}{n}\sum_{i=1}^n \left\{\log \frac{g^{\hat{\beta}_n}(Y_i)}{\hat{h}(Y_i|w)}\right\}^2 - \left\{\frac{1}{n}\sum_{i=1}^n \log \frac{g^{\hat{\beta}_n}(Y_i)}{\hat{h}(Y_i|w)}\right\}^2$ and the LCV variance(proposed variance) as $\sigma_{nc}^2 = \hat{\omega}_{nc}^2 = \hat{var}(\log g^\beta(Y)) + \hat{var}(CV(w))$. Theoretical comparison between $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ is not straightforward, because we replace a nonparametric rival density with a parametric rival model and compute the variance of a test statistic. For empirical comparison we should use another parametric model instead the nonparametric density. But, it is the idea discussed in the theory of Vuong's test.

In the next two sections, we use both variances to compare the nested and non-nested rival models by simulation and real data study.

4 Simulation

We performed a simulation study to decide whether the two rival models, the parametric and non-parametric models, are equivalent or not. We have considered the Weibull(α, β) distribution, say W, as the true model and Gamma(η, θ) density, G, and $\hat{h}$ with Gaussian kernel as two rival models. We generated 10^4 replications from the above models for $n = 30, 100$. We set $(\alpha, \beta) = (1.9, 3.1)$ and $w = 0.2, 0.4, 0.5, 0.6, 0.8$. Using R, one can use package *ks* or *kedd* to choose smoothing bandwidths. Table 1 shows the estimated variances, $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ for log ratio of rival models. For each w, the usual variance of log ratio for the parametric model against the non-parametric model is greater than the variance of log ratio for two parametric models and, also, the usual estimation of the variance of the log ratio of parametric models against non-parametric model are always greater than the computed variance based on CV, $\hat{\omega}_n^2 > \hat{\omega}_{nc}^2$.

Tables 2 and 3 contain the results of the simulation for computing Vuong's test and the tracking interval based on $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ and their lengths(below the intervals). We have considered $\Delta(G, \hat{h})$, $\Delta(W, \hat{h})$ and $\Delta(G, W)$. We need to $\mathcal{LR}^{a/b}$ and the penalty term in the tracking interval, so we use the $\mathcal{LR}^{a/b}$ in the tables. We considered i.i.d sample of size $n = 30, 100$ from the Weibull$(1.9, 3.1)$ distribution, as the true model. The rival models are Weibull, W, gamma, G, and the kernel density estimator, $\hat{h}(\cdot)$. Since model W is well-specified, the misspecification error, R(M), is zero. We have estimated the three rival models using the data which has generated from Weibull$(1.9, 3.1)$. For each replication, we computed the maximum likelihood estimators, kernel density estimator,

AIC, and other required statistics. The optimum bandwidth is $w_{opt} = 0.4$. The results for $w = 0.2, 0.4, 0.5, 0.6, 0.8$ are summarized in Tables 2 and 3. For two variances, $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$, Vuong's statistics and tracking intervals are computed.

The first and second columns indicate that the parametric rival models are better than the kernel density estimator for describe the data. The Weibull and gamma models always for $n = 30$ are equivalent. The related tracking interval based upon the usual variance except for $w = 0.2, 0.4$ agrees with Vuong's test. At $n = 30$ and $w = 0.5$, the tracking interval, based on the usual variance, preferred the gamma model against the Weibull model. Notably, these results for comparison between the Weibull and gamma models are not dependent on the bandwidth values.

Table 1: Variance of log likelihood ratio of rival models for n=30, 100

	w	log ratio	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
		$\mathcal{LR}^{G/\hat{h}}$	1.6237	0.4439
	0.8	$\mathcal{LR}^{W/\hat{h}}$	1.6902	0.4008
		$\mathcal{LR}^{W/G}$	0.0263	
		$\mathcal{LR}^{G/\hat{h}}$	1.1449	0.5542
	0.6	$\mathcal{LR}^{W/\hat{h}}$	1.2145	0.5121
		$\mathcal{LR}^{W/G}$	0.0232	
		$\mathcal{LR}^{G/\hat{h}}$	1.8491	0.5463
$n = 30$	0.5	$\mathcal{LR}^{W/\hat{h}}$	1.7259	0.4431
		$\mathcal{LR}^{W/G}$	0.0352	
		$\mathcal{LR}^{G/\hat{h}}$	0.9535	0.3162
	0.4	$\mathcal{LR}^{W/\hat{h}}$	0.9697	0.3116
		$\mathcal{LR}^{W/G}$	0.0165	
		$\mathcal{LR}^{G/\hat{h}}$	2.5669	0.3588
	0.2	$\mathcal{LR}^{W/\hat{h}}$	2.5329	0.3541
		$\mathcal{LR}^{W/G}$	0.0283	
		$\mathcal{LR}^{G/\hat{h}}$	1.2789	0.6234
	0.8	$\mathcal{LR}^{W/\hat{h}}$	1.1094	0.3812
		$\mathcal{LR}^{W/G}$	0.0808	
		$\mathcal{LR}^{G/\hat{h}}$	1.5402	0.5094
	0.6	$\mathcal{LR}^{W/\hat{h}}$	1.5534	0.4571
		$\mathcal{LR}^{W/G}$	0.0309	
		$\mathcal{LR}^{G/\hat{h}}$	1.2744	0.3973
$n = 100$	0.5	$\mathcal{LR}^{W/\hat{h}}$	1.1838	0.3359
		$\mathcal{LR}^{W/G}$	0.0177	
		$\mathcal{LR}^{G/\hat{h}}$	1.91399	0.3583
	0.4	$\mathcal{LR}^{W/\hat{h}}$	1.0919	0.3097
		$\mathcal{LR}^{W/G}$	0.0240	
		$\mathcal{LR}^{G/\hat{h}}$	1.5751	0.4508
	0.2	$\mathcal{LR}^{W/\hat{h}}$	1.6138	0.4653
		$\mathcal{LR}^{W/G}$	0.0323	

Table 2: The numeric results of the Vuong's test and the tracking interval based on $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ for $n = 30$. The numbers below the intervals are the lengths of the intervals

w	Δ	Vuong's statistics		Tracking interval	
		$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
0.8	$\mathcal{LR}^{G/\hat{h}}$	3.0089	5.7542	(-1.9167, -1.0044) 0.9123	(-1.6992, -1.2223) 0.4759
	$\mathcal{LR}^{W/\hat{h}}$	2.9981	6.1568	(-1.9492, -1.0188) 0.9304	(-1.7105, -1.2044) 0.5061
	$\mathcal{LR}^{W/G}$	-0.3923		(-0.0813, 0.0348)	
0.6	$\mathcal{LR}^{G/\hat{h}}$	2.7066	3.8902	(-1.5302, -0.7644) 0.7658	(-1.4137, -0.8809) 0.5328
	$\mathcal{LR}^{W/\hat{h}}$	2.8881	4.1399	(-1.5659, -0.7771) 0.7888	(-1.4276, -0.8759) 0.5517
	$\mathcal{LR}^{W/G}$	-0.4355		(-0.0787, 0.0403)	
0.5	$\mathcal{LR}^{G/\hat{h}}$	3.3726	6.2051	(-2.2599, -1.2869) 0.9730	(-2.0378, -1.5088) 0.5290
	$\mathcal{LR}^{W/\hat{h}}$	3.3224	6.5569	(-2.1625, -1.2223) 0.9402	(-1.9306, -1.4067) 0.5239
	$\mathcal{LR}^{W/G}$	-1.1801		(0.0137, 0.1481)	
0.4(opt)	$\mathcal{LR}^{G/\hat{h}}$	2.0010	3.1036	(-1.0972, -0.3984) 0.6988	(-0.9490, -0.5466) 0.4024
	$\mathcal{LR}^{W/\hat{h}}$	2.9831	3.5284	(-1.1765, -0.4718) 0.7047	(-1.0239, -0.5583) 0.4656
	$\mathcal{LR}^{W/G}$	-1.6259		(-0.1224, -0.0303)	
0.2	$\mathcal{LR}^{G/\hat{h}}$	2.1611	5.7767	(-1.9606, -0.8141) 1.1465	(-1.6018, -1.1729) 0.4289
	$\mathcal{LR}^{W/\hat{h}}$	2.3828	6.2957	(-2.0632, -0.9325) 1.1307	(-1.7119, -1.2249) 0.4870
	$\mathcal{LR}^{W/G}$	-1.7996		(-0.1707, -0.0503)	

For $n = 100$ and bandwidth $w = 0.5$, Vuong's statistic is equal to -0.2655, and the tracking interval based on the usual variance is $(-0.0756, 0.0235)$ which shows that the Weibull and gamma models are equivalent, but we emphasize that this result does not depend on the bandwidth, $w = 0.5$. For the value of w when $n = 100$ we prefer the Weibull density in favor of gamma density. The lower and upper limits of tracking intervals of parametric models and kernel density estimator for all w's are negative, so, we conclude that the parametric rival distributions are better candidates than the non-parametric model to estimate the true density, see Figures 1 and 2 are accordant with the results of Tables 2 and 3 for $n = 30, 100$. It shows that, based on our simulation the kernel density estimator is not superior to the parametric model to estimate the true model.

Table 3: The nemeric results of Vuong's test and tracking interval based on $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ for $n = 100$. The numbers below the intervals are the length of intervals

w	Δ	Vuong's statistics		Tracking interval	
		$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
0.8	$\mathcal{LR}^{G/\hat{h}}$	6.3462	9.0896	(-1.6820, -1.2387) 0.4433	(-1.6151, -1.3056) 0.3095
	$\mathcal{LR}^{W/\hat{h}}$	7.1456	12.1903	(-1.7367, -1.3238) 0.4129	(-1.6512, -1.3787) 0.2725
	$\mathcal{LR}^{W/G}$	2.2290		(-0.0256, -0.0142)	
0.6	$\mathcal{LR}^{G/\hat{h}}$	6.0043	10.4409	(-1.7597, -1.2732) 0.4865	(-1.6563, -1.3766) 0.2797
	$\mathcal{LR}^{W/\hat{h}}$	5.98028	11.0237	(-1.7610, -1.2725) 0.4885	(-1.6493, -1.3591) 0.2902
	$\mathcal{LR}^{W/G}$	2.0089		(-0.0447, -0.0341)	
0.5	$\mathcal{LR}^{G/\hat{h}}$	3.3458	5.9922	(-1.0075, -0.5649) 0.4426	(-0.9098, -0.6626) 0.2472
	$\mathcal{LR}^{W/\hat{h}}$	3.6994	6.9438	(-1.0491,-0.6226) 0.4265	(-0.9494, -0.6881) 0.2613
	$\mathcal{LR}^{W/G}$	-0.2655		(-0.0756, 0.0235)	
0.4(opt)	$\mathcal{LR}^{G/\hat{h}}$	5.2794	9.4139	(-1.3698, -0.9512) 0.4186	(-1.2779, -1.0431) 0.2348
	$\mathcal{LR}^{W/\hat{h}}$	5.6532	10.6149	(-1.4194, -1.0098) 0.4096	(-1.3237, -1.0692) 0.2545
	$\mathcal{LR}^{W/G}$	2.4962		(-0.0845, -0.0238)	
0.2	$\mathcal{LR}^{G/\hat{h}}$	5.0686	9.4747	(-1.5545, -1.0625) 0.4920	(-1.4401, -1.1769) 0.2632
	$\mathcal{LR}^{W/\hat{h}}$	5.1072	9.5113	(-1.5826, -1.0848) 0.4978	(-1.4675, -1.1758) 0.2917
	$\mathcal{LR}^{W/G}$	2.1046		(-0.0606, -0.0099)	

For all selected values of w, the length of intervals based on the proposed variance is shorter than the length of intervals based on the usual variance. A larger sample size or lower variability will result in the tracking interval with a smaller margin of error. A smaller sample size or a higher variability will result in a wider tracking interval with a larger margin of error. The level of confidence also affects the interval width. In consequence, we fixed the level of confidence at 0.95 a value that makes the length of the tracking interval large enough. Also, we considered the sample size $n = 30$, a moderately small value, and $n = 100$ a moderatly large value. As we saw, the length of the tracking intervals based on the proposed variance is shorter than the tracking intervals based on the usual variance, this seems to be the case because the proposed variance is empirically smaller than the usual variance. In a comparison of the parametric rival models, for $n = 30$, Vuong's test proposes the equality of the Weibull and gamma models, but the tracking intervals suggest the Weibull model.

Various measures have been studied of the discrepancy of the kernel density estimator, $\hat{h}$, from the true density, $h(\cdot)$. By considering a single point, say y, a natural measure is the mean square error, mse, denoted by,

$$mse(\hat{h}) = E\{(\hat{h}(y) - h(y))^2\}.$$

A Monte Carlo estimation of mse and Kolmogorov-Smirnov, K.S., statistic for simulated data is computed, see Tables 4 and 5. As we see, for $w = 0.8, 0.6, 0.5, 0.4, 0.2$ the mse for parametric models against the kernel density estimator is too small. On the other hand, the K.S. statistic indicates the equality of the parametric rival models and the kernel density estimator to describe the data but rejects the equality of Weibull and gamma models.

Table 4: mse for estimated parametric and kernel density estimator, $n = 30$

		Weibull	gamma	kernel model
mse	$w = 0.8$	0.0293	0.2135	0.4306
	$w = 0.6$	0.0608	0.0114	0.0882
	$w = 0.5$	0.0549	0.0036	0.1548
	$w = 0.4$	0.01319	0.1761	0.0614
	$w = 0.2$	0.0567	0.0059	0.7971
K.S. p-value	$w = 0.8$	0.1350	0.3925	0.0346
	$w = 0.6$	0.5941	0.5941	0.0346
	$w = 0.5$	0.9988	0.3929	0.1350
	$w = 0.4$	0.2391	0.5941	0.0709
	$w = 0.2$	0.3929	0.2391	0.0025

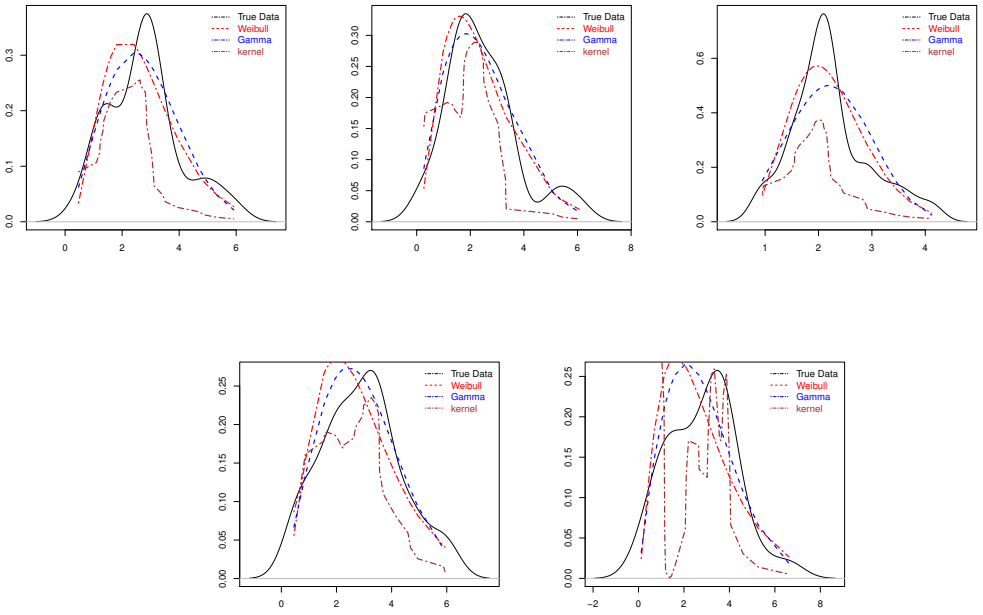

Figure 1: Graph of assumed density and rival models in the simulation for n=30: left upper, $w = 0.8$, right upper, $w = 0.6$, left lower, $w = 0.5$, right lower, $w = 0.4$ and bottom middle figure, $w = 0.2$

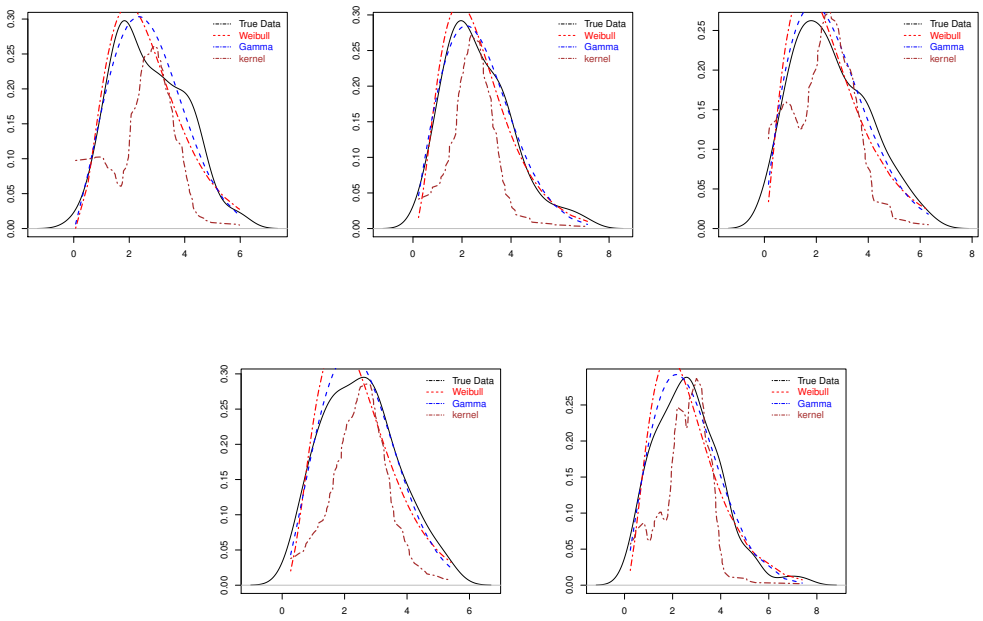

Figure 2: Graph of assumed density and rival models in the simulation for n=100: left upper, $w = 0.8$, right upper, $w = 0.6$, left lower, $w = 0.5$, right lower, $w = 0.4$ and bottom middle figure, $w = 0.2$

Table 5: mse for estimated parametric and kernel density estimator, $n = 100$

		Weibull	gamma	kernel model
mse	$w = 0.8$	0.0074	0.0052	0.1301
	$w = 0.6$	0.0232	0.0545	0.0466
	$w = 0.5$	0.0018	0.0525	0.0723
	$w = 0.4$	0.0032	0.0007	0.1236
	$w = 0.2$	0.0239	2.003910^{-6}	0.0834
K.S. p-value	$w = 0.8$	0.4676	0.4676	1.2×10^{-12}
	$w = 0.6$	0.3667	0.8127	4.95×10^{-7}
	$w = 0.5$	0.5806	0.5806	1.701×10^{-6}
	$w = 0.4$	0.5706	0.8127	4.705×10^{-6}
	$w = 0.2$	0.9062	0.8127	3.729×10^{-5}

5 Real Data

Our example appertain to the comparison of possible models of association of the Europe oil prices, Brent. This data can be found in https://fred.stlouisfed.org/series. The Brent dataset consists of daily returns of the Europe oil prices with the sample extending from May 1987 to March 2021 for a total of n = 8577 observations. The series

VB is obtained by substituting the series Brent in function $Q(X)$,

$$Q(X) = X/10,$$

The dataset VB describes the information on oil volatility. The descriptive statistics of our datasets are given in 6 which shows that the series of the dataset has a mean that is different from zero. The series VB has positive skewness. Also, both are characterized by heavy tails since they have negative sample excess kurtosis. The hypothesis of normality is accepted for all series since $P - value > 0.05$.

Table 6: Descriptive Statistics for Empirical Series

series	n	$\bar{x}$	$\hat{\sigma}$	$\mathcal{S}$	$\mathcal{K}$
VB	8577	4.6389	10.324	0.8869	-0.3589

Where, $\mathcal{S}$ denotes the sample skewness, $\mathcal{K}$ denotes the sample excess kurtosis.

Real Data Study 1

As the simulation part, we consider three rival models, Weibull, Gamma, and kernel density estimator. The estimated variances, $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ for log ratio of true and rival models are given in 7. For each w, the usual estimation of the variance of the log ratio of parametric models against non-parametric model is always greater than the computed variance based on CV, $\hat{\omega}_n^2 > \hat{\omega}_{nc}^2$. Also, the usual variance of log likelihood ratio for the parametric model against the non-parametric model is greater than the variance of log ratio for two parametric models and the gamma distribution has greater variance than the Weibull distribution against the kernel density estimator using Gaussian kernel, $\hat{h}(\cdot)$.

The values of Vuong's statistic and the tracking intervals are given in Table 8. The results show that the Weibull and gamma models are better than the kernel density estimator. In the third row, for each w, the Weibull model is prefered to the gamma model. The optimum bandwidth is $w_{opt} = 0.4$.

For the real data, we estimated the mse and the Kolmogorov-Smirnov, K.S., statistic. see Table 9. For $w = 0.8, 0.6, 0.5, 0.4, 0.2$ the mse for parametric models in comparison to the kernel density estimator is too small. The K.S. test proposes to reject the suitability of Weibull, gamma, and kernel density models, because, the p-value is always less than 2.2×10^{-16}. The mse for the Weibull model and gamma model has small values relative to the mse for the kernel density estimator, which suggests using parametric density to describe the data. Figure 3 confirms the results of the test and tracking interval.

Table 7: Variance of the log likelihood function for ratio of the rival models, $w = 0.2, 0.4, 0.5, 0.6, 0.8$

w	log ratio	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
0.8	$\mathcal{LR}^{G/\hat{h}}$	1.5423	0.4967
	$\mathcal{LR}^{W/\hat{h}}$	1.6018	0.4087
	$\mathcal{LR}^{W/G}$	0.0069	
0.6	$\mathcal{LR}^{G/\hat{h}}$	1.3845	0.4967
	$\mathcal{LR}^{W/\hat{h}}$	1.4271	0.4087
	$\mathcal{LR}^{W/G}$	0.0069	
0.5	$\mathcal{LR}^{G/\hat{h}}$	1.3293	0.4967
	$\mathcal{LR}^{W/\hat{h}}$	1.3630	0.4087
	$\mathcal{LR}^{W/G}$	0.0069	
0.4	$\mathcal{LR}^{G/\hat{h}}$	1.3029	0.4967
	$\mathcal{LR}^{W/\hat{h}}$	1.3272	0.4087
	$\mathcal{LR}^{W/G}$	0.0069	
0.2	$\mathcal{LR}^{G/\hat{h}}$	1.4030	0.4967
	$\mathcal{LR}^{W/\hat{h}}$	1.4053	0.4087
	$\mathcal{LR}^{W/G}$	0.0069	

Table 8: Results of the Vuong's test and the tracking interval for the real data. The numbers below the intervals are the length of intervals

w	Δ	Vuong's statistics		Tracking interval	
		$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
0.8	$\mathcal{LR}^{G/\hat{h}}$	43.0161	75.7985	(-1.1474, -1.0943) 0.0531	(-1.1343, -1.1054) 0.0289
	$\mathcal{LR}^{W/\hat{h}}$	41.0701	81.3024	(-1.1158, -1.0636) 0.0522	(-1.1028, -1.0735) 0.0293
	$\mathcal{LR}^{W/G}$	17.2741		(-0.0285, -0.0319)	
0.6	$\mathcal{LR}^{G/\hat{h}}$	38.1107	63.6272	(-0.9644, -0.9160) 0.0484	(-0.9547, -0.9257) 0.0290
	$\mathcal{LR}^{W/\hat{h}}$	36.3307	67.8850	(-0.9345, -0.8854) 0.0491	(-0.9231, -0.8938) 0.0293
	$\mathcal{LR}^{W/G}$	17.2741		(-0.0285, -0.0319)	
0.5	$\mathcal{LR}^{G/\hat{h}}$	35.1418	57.4889	(-0.8733, -0.8259) 0.0474	(-0.8641, -0.8351) 0.0290
	$\mathcal{LR}^{W/\hat{h}}$	33.4689	61.1183	(-0.8433, -0.7954) 0.0479	(-0.8325, -0.8032) 0.0293
	$\mathcal{LR}^{W/G}$	17.2741		(-0.0285, -0.0319)	
0.4	$\mathcal{LR}^{G/\hat{h}}$	32.0547	51.9167	(-0.7908, -0.7439) 0.0469	(-0.7818, -0.7528) 0.0290
	$\mathcal{LR}^{W/\hat{h}}$	30.5091	54.9757	(-0.7608, -0.7134) 0.0474	(-0.7502, -0.7209) 0.0293
	$\mathcal{LR}^{W/G}$	17.2741		(-0.0285, -0.0319)	
0.2	$\mathcal{LR}^{G/\hat{h}}$	26.6327	44.7602	(-0.6860, -0.6374) 0.0486	(-0.6762, -0.6472) 0.0290
	$\mathcal{LR}^{W/\hat{h}}$	23.3946	47.0866	(-0.6358, -0.6071) 0.0287	(-0.6446, -0.6153) 0.0293
	$\mathcal{LR}^{W/G}$	17.2741		(-0.0285, -0.0319)	

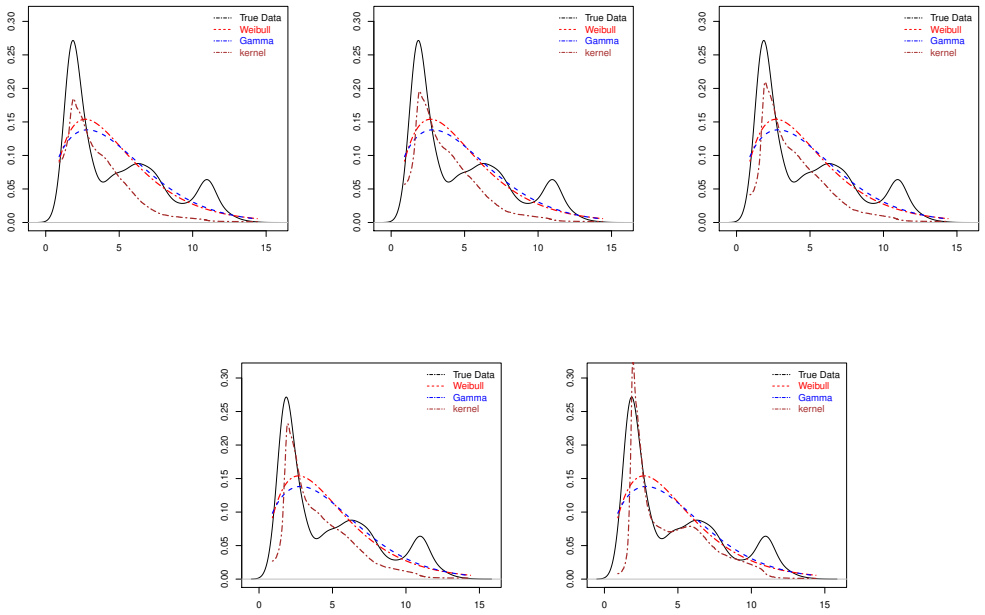

Figure 3: Graph of Europe oil prices(true), kernel density estimate and the parametric rival models: left upper, $w = 0.8$, right upper, $w = 0.6$, left lower, $w = 0.5$, right lower, $w = 0.4$ and bottom middle figure, $w = 0.2$

Table 9: The mse values for real data

	Weibull	gamma	kernel model
w=0.8	0.0033	0.0034	7.3620
w=0.6	0.0012	0.0003	3.3539
w=0.5	0.0004	0.0023	2.0988
w=0.4	0.0007	0.0010	1.2121
w=0.2	6.47×10^{-6}	0.0001	0.2495

Real Data Study 2

We again consider the Europe oil prices, Brent. The rival models are first-order autoregressive(AR) model $x_t = \phi x_{t-1} + \epsilon_t$ with normal error $N(\mu_\epsilon, \sigma_\epsilon^2)$ and the kernel density estimator as two completely non-nested models. For the kernel model, we consider the Gaussian and Epanechnikov kernels, but we reported only the Gaussian kernel result because both kernels have the same results.

In Figure 4, the correlation diagrams, and the partial correlation of the data are shown. According to the partial correlation diagram, it can be seen that the data follow the first-order autoregressive model. The ACF graph, shows non-stationarity in autore-

gressive model. Generally, we consider the non-stationarity implied by a unit root in an associated auotoregressive polynomial. For non-stationary autoregressive models with i.i.d errors, weak consistency of the order estimators studied by researchers, and variants of AIC for general non-stationary stochastic models used. Table 13 shows the mean square errors for the AR model are small, thus the non-stationarity does not affect the prediction. It considered that the parameters of the autoregressive model are estimated using the maximum likelihood method and for the kernel density estimator we consider the Gaussian kernel. The estimated values are shown in Table 10.

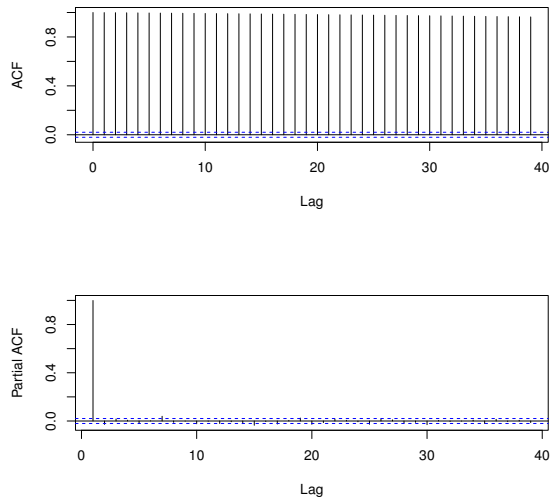

Figure 4: *ACF*, *PACF* graph. The partial correlation shows non-stationarity and the partial correlation diagram, shows that the data follow the AR(1) model

Table 10: Estmated values of parameters in the autoregressive model

ϕ	μ_ϵ	σ^2_ϵ
0.9992	0.0041	0.0149

The estimated variances, $\hat{\omega}^2_n$ and $\hat{\omega}^2_{nc}$ for log ratio of the two rival models are given in Table 11. For each $w = 0.8, 0.6, 0.5, 0.4, 0.2$, the usual estimation of the variance of the log ratio of the kernel density estimator against the first-order autoregressive model is always greater than the computed variance based on CV, $\hat{\omega}^2_n > \hat{\omega}^2_{nc}$. The values of Vuong's statistics and the tracking intervals are given in Table 12. The results show that the first-order autoregressive model is better than the kernel density estimator with the Gaussian (and epanechnikov) kernel to describe the data.

Table 11: Variance of the log likelihood ratio of the rival models for $w = 0.2, 0.4, 0.5, 0.6, 0.8$

w	log ratio	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
0.8	$\mathcal{LR}^{\hat{h}/AR}$	4.5960	3.7299
0.6	$\mathcal{LR}^{\hat{h}/AR}$	4.4086	3.7299
0.5	$\mathcal{LR}^{\hat{h}/AR}$	4.3309	3.7299
0.4	$\mathcal{LR}^{\hat{h}/AR}$	4.2758	3.7299
0.2	$\mathcal{LR}^{\hat{h}/AR}$	4.3044	3.7299

Table 12: The numeric results of the Vuong's test and the tracking interval based on $\hat{\omega}_n^2$ and $\hat{\omega}_{nc}^2$ for the real data. The numbers below the intervals are the length of intervals

w	Δ	Vuong's statistics		Tracking interval	
		$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$	$\hat{\omega}_n^2$	$\hat{\omega}_{nc}^2$
0.8	$\mathcal{LR}^{\hat{h}/AR}$	-163.6503	-181.6607	(-7.3999, -7.3118) 0.0881	(-7.3995, -7.3164) 0.0831
0.6	$\mathcal{LR}^{\hat{h}/AR}$	-163.0067	-177.2166	(-7.2193, -7.1330) 0.0863	(-7.2158, -7.1365) 0.0793
0.5	$\mathcal{LR}^{\hat{h}/AR}$	-162.3798	-174.9749	(-7.1283, -7.0428) 0.0855	(-7.1252, -7.0459) 0.0793
0.4	$\mathcal{LR}^{\hat{h}/AR}$	-161.5233	-172.9391	(-7.0458, -6.9608) 0.0850	(-7.0429, -6.9636) 0.0793
0.2	$\mathcal{LR}^{\hat{h}/AR}$	-158.5478	-170.3210	(-6.9403, -6.8550) 0.0853	(-6.9373, -6.8579) 0.0794

Mean square error, mse, and Kolmogorov-Smirnov, K.S., statistics for the real data are computed, see Table 13. As we see, for $w = 0.8, 0.6, 0.5, 0.4, 0.2$ the mse for the AR model against the kernel density estimator is too small. On the other hand, the K.S. statistic indicates the AR model is preferred to the kernel density estimator.

Table 13: mse and KS statistics for estimated first-order autoregressive model, and kernel density estimator

		AR model	Kernel model
mse	$w = 0.8$	7.6335×10^{-6}	7.3620
	$w = 0.6$	7.6335×10^{-6}	3.3539
	$w = 0.5$	7.6335×10^{-6}	2.0988
	$w = 0.4$	7.6335×10^{-6}	1.2121
	$w = 0.2$	7.6335×10^{-6}	0.2495
K.S. p-value	$w = 0.8$	0.9671	2.2×10^{-16}
	$w = 0.6$	0.9671	2.2×10^{-16}
	$w = 0.5$	0.9671	2.2×10^{-16}
	$w = 0.4$	0.9671	2.2×10^{-16}
	$w = 0.2$	0.9671	2.2×10^{-16}

Figure 5 confirms the results of the test and the tracking interval. The real data graph is overlap with the graph of the AR model.

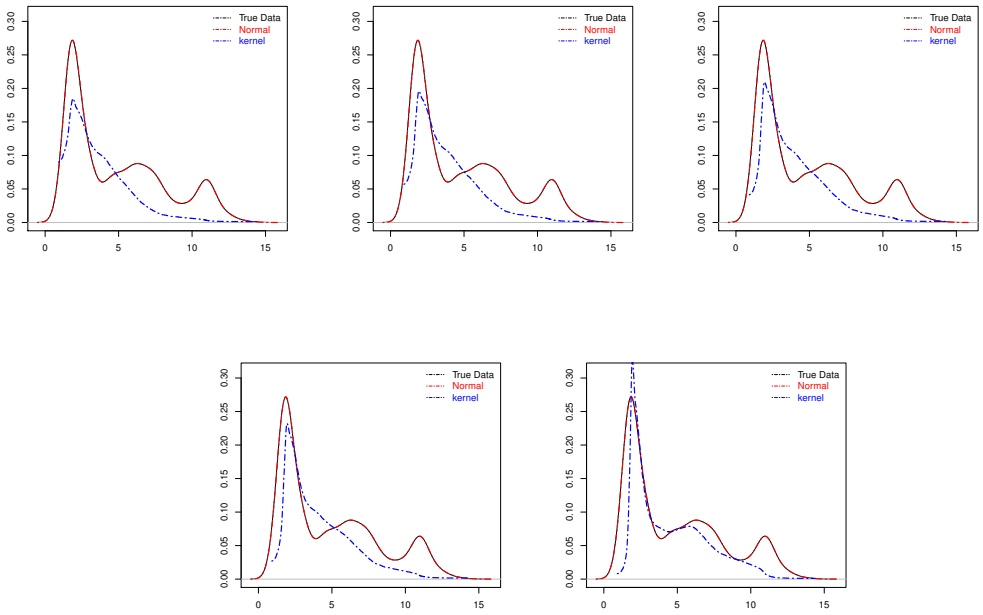

Figure 5: Graph of the Europe oil prices(true), the AR model, and the kernel density estimator: left upper, $w = 0.8$, right upper, $w = 0.6$, left lower, $w = 0.5$, right lower, $w = 0.4$ and bottom middle figure, $w = 0.2$

6 Conclusion

This paper develops the parametric and non-parametric density estimation. Our attention was on the known risk function, the expected Kullback-Leibler divergence criterion. Based on the known theory in model selection, in order to use Vuong's test and the tracking interval for both parametric rival models and Kernel density estimator, we have considered two types of variances, one based on traditional and the other based on new cross-validation approaches. The simulation and real data analysis have shown that to describe data with unknown density, the parametric density is preferred to the kernel density estimator. For parametric models, we have considered nested and non-nested models. The risk of modeling for a non-nested model is null and all error is due to the estimation. Comparison of the parametric and non-parametric models in this paper is, in fact, a comparison between the likelihood density estimation and the likelihood cross-validation. It is known that if the rival parametric model does not work, the kernel density estimator is a suitable alternative. As we know, in literature, it has been shown that the kernel's shape is not important. In fact, what matters is the kernel bandwidth which determines the level of smoothness. Here we have considered

the optimum bandwidth, which had the same results as other bandwidths.

When the parametric model does not hold and we can not fit a parametric model to the real data, the true density may be estimated non-parametrically. On the other hand, a finite mixture model as a combination of parametric and non-parametric estimation is possible. So, in addition, we should estimate the mixing parameter. Here we have done a comparison between a parametric and a non-parametric rival model, but a comparison between semiparametric and parametric or non-parametric models is possible.

Acknowledgement

The author would like to thank the respected Editor and two referees for their helpful comments and suggestions.

References

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle. *In the Second International Symposium on Information Theory*, eds. by B.N. Petrov and F. Csake, Akademiai Kiado, Hungary, 267-281.

Commenges, D., Sayyareh, A., Letenneur, L., Guedj, J. and Bar-Hen, A. (2008), Estimating a difference of Kullback-Leibler risks using a normalized difference of AIC. *The Annals of Applied Statistics*, **2**(3), 1123-1142.

Goldenshluger, A., and Lepski, O. (2011), Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality.

Jaakkola, T., Diekhans, M., and Haussler, D. (1999), Using the fisher kernel method to detect remote protein homologies. In Proc. *Internation Conference on Intelligent Systems for Molecular Biology*, 149-158.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*. Springer-Verlag (2nd Edition), New York, NY.

Habbema, J. D. F., Hermans, J. and van der Broek, K. (1974), A stepwise discrimination program using density estimation. In Bruckman, G.(ed), Compstat. Vienna: Physica Verlag, 100-110.

Hjort, N. L., and Jones, M. C. (1996), Locally parametric nonparametric density estimation. *The Annals of Statistics*, 1619-1647.

Liu, J., Chen, J., Chen, S. and Ye. J. (2009), Learning the optimal neighborhood kernel for classification. *In International Joint Conference on Artificial Intelligence*, Pasadena, California.

Lacour, C., Massart, P., and Rivoirard, V. (2017), Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, **79**, 298-335.

Loader, C. R. (1999), Bandwidth selection: clasic or plug-in. The Annals of Statistics. **27**(2), 415-438.

Moreno, P. J., P., Ho, P. and Vasconcelos, N. (2003), A kullback-leibler divergence based kernel for svm classification in multimedia applications. In Advances in Neural Information Processing Systems.

Marron, J. S. (1985), An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Annals of statistic*. **13**(3), 1011-1023.

Panahi, H., and Sayyareh, A. (2014), Tracking interval for type II hybrid censoring scheme. *Journal of The Iranian Statistical Society*, **13**(2), 187-208.

Sayyareh, A. (2012), Inference after separated hypotheses testing: an empirical investigation for linear models. *Journal of Statistical Computation and Simulation*, **82**(9), 1275-1286.

Scott, D. W. (2015), *Multivariate density estimation: theory, practice, and visualization*. John Wiley and Sons.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, (Vol. 26). CRC press.

Stone, C. J. (1984), An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 1285-1297.

Stonem M. (1974), Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

Vuong, Q. H. (1989), Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**(2), 307-333.

Wang, L., Chan, K.L., Xue, P. and Zhou, L.P. (2008), A kernel-induced space selection approach to model selection in klda. IEEE Trans. *Neural Networks*, **19**, 2116-2131.

Yeung, D., Chang, H. and Dai. G. (2007), Learning the kernel matrix by maximizing a kfd-based class separability criterion. *Pattern Recognition*, **40**, 2021-2028.

Xiong, H., Swamy, M.N.S., and Ahmad, M.O. (2005), Optimizing the kernel in the empirical feature space. *IEEE Transactions on Neural Networks*, **16**(2), 460-474.