Computational Statistics and Data Analysis 89 (2015) 51–71

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Improving cross-validated bandwidth selection using subsampling-extrapolation techniques

Qing Wang [a,*], Bruce G. Lindsay [b]

[a] Department of Mathematics and Statistics, Williams College, Williamstown, MA, USA

[b] Department of Statistics, Pennsylvania State University, University Park, PA, USA

HIGHLIGHTS

- A two-stage subsampling-extrapolation bandwidth selection procedure is proposed.
- An automatic nested cross-validation method is developed to select the subsample size.
- The extrapolated bandwidth selectors achieve a smaller mean square error.
- The second-order extrapolated bandwidth selector has a relative convergence rate $n^{-1/4}$.

ARTICLE INFO

Article history:
Received 16 July 2013
Received in revised form 19 February 2015
Accepted 3 March 2015
Available online 16 March 2015

Keywords:
Bandwidth selection
Cross-validation
Extrapolation
L^2 distance
Nonparametric kernel density estimator
Subsampling

ABSTRACT

Cross-validation methodologies have been widely used as a means of selecting tuning parameters in nonparametric statistical problems. In this paper we focus on a new method for improving the reliability of cross-validation. We implement this method in the context of the kernel density estimator, where one needs to select the bandwidth parameter so as to minimize L^2 risk. This method is a two-stage subsampling-extrapolation bandwidth selection procedure, which is realized by first evaluating the risk at a fictional sample size m ($m \leq$ sample size n) and then extrapolating the optimal bandwidth from m to n. This two-stage method can dramatically reduce the variability of the conventional unbiased cross-validation bandwidth selector. This simple first-order extrapolation estimator is equivalent to the rescaled "bagging-CV" bandwidth selector in Hall and Robinson (2009) if one sets the bootstrap size equal to the fictional sample size. However, our simplified expression for the risk estimator enables us to compute the aggregated risk without any bootstrapping. Furthermore, we developed a second-order extrapolation technique as an extension designed to improve the approximation of the true optimal bandwidth. To select the optimal choice of the fictional size m given a sample of size n, we propose a nested cross-validation methodology. Based on simulation study, the proposed new methods show promising performance across a wide selection of distributions. In addition, we also investigated the asymptotic properties of the proposed bandwidth selectors.

1. Introduction

Cross-validation methodology has long been a popular method for selecting tuning parameters in non and semiparametric models. However, it has also been criticized for its high variability and its corresponding tendency to overfit the data.

* Correspondence to: 18 Hoxsey Street, Williamstown, MA 01267, USA. Tel.: +1 413 597 4960.
E-mail address: qww1@williams.edu (Q. Wang).

http://dx.doi.org/10.1016/j.csda.2015.03.005

This paper develops new methods for the improvement of the conventional cross-validation procedures. It is based on a blending of U-statistic estimation and asymptotic theory. These new methods are realized by estimating the cross-validation risk with small training sets, then extrapolating the results to the desired sample size. The extrapolation step requires some asymptotic theory, but only the rate of convergence, not any unknown constants. We will show that such a two-stage procedure can dramatically reduce the high variability and overfitting that is the major liability of the conventional unbiased cross-validation.

We view our results as part of the following paradigm: when one is estimating nonparametrically a statistical property of samples of target size m, such as the risk inherent in using a particular model, then one can do a much more accurate estimation when the target m is much smaller than the actual sample size n. The intuition is that there are many, many more subsamples of size $n/2$, say, than there are subsamples of size n or $n-1$.

To motivate our extrapolation methodology, we will here show how it works when used in risk estimation in the context of nonparametric kernel density estimation. In the process we will also show that for this problem the risk function for arbitrary m is surprisingly simple. In particular, cross-validation estimation at an arbitrary training sample size of m does not require repeated subsampling at size m, thereby greatly speeding up and improving accuracy of the methods we propose. We believe this to be a major new insight in the kernel density estimation literature.

To simplify notation, consider a univariate random variable $X \in \mathcal{R}$. In statistical practice, we often know little about the underlying distribution of X which is crucial in exploratory or inferential analysis (Silverman, 1986). So, our main task is to estimate the unknown density function $f(x)$ based on a set of observations. In this paper, we focus on the nonparametric kernel density estimator (Fix and Hodges, 1951). Given an i.i.d. sample of size n, $\mathcal{X}_n = (X_1, \ldots, X_n)$, the kernel density estimator at x is defined for a kernel K as

$$\hat{f}_h(x \mid \mathcal{X}_n) = n^{-1} \sum_{i=1}^{n} K_h(X_i - x) \quad (x \in \mathcal{R}), \tag{1.1}$$

where $h > 0$ is called the bandwidth parameter. Here $K_h(t) = h^{-1}K(t/h)$ and function K is the kernel function. As the choice of K does not greatly affect the density estimation (Hardle et al., 1994), throughout this paper we consider a commonly used location kernel function, the Gaussian kernel.

$$K_h(x - x_0) = (h\sqrt{2\pi})^{-1} e^{-(x-x_0)^2/(2h^2)} \sim N(x_0, h^2). \tag{1.2}$$

However, our proposed methodologies do not depend on the choice of K, and the theoretical results in this paper will be stated in terms of an arbitrary symmetric kernel function K of order r ($r \geq 2$). For the definition of the order of a kernel function, please see Turlach (1993).

Although one has free choice of the kernel function in a density estimator, the choice of the bandwidth h is generally viewed as much more crucial. In order to select the optimal smoothing parameter h, we need to evaluate how closely $\hat{f}_h$ can approximate f for a given data set. Most bandwidth selectors are based on first choosing a risk function that measures the error made in using a particular bandwidth h. One can then estimate the risk function for a given data set and choose the bandwidth that minimizes the empirical risk. Such bandwidth selectors are referred to as data-driven methods.

The main result of this paper is to propose a two-stage subsampling-extrapolation bandwidth selection procedure. This work is closely related to the rescaled bagging cross-validation method of Hall and Robinson (2009) and the partitioned cross-validation method of Marron (1987). Recent work involving bagging and subsampling in problems other than kernel density estimation includes Meinshausen and Buhlmann (2010) and Shah and Samworth (2012). Unlike the bandwidth selectors discussed in Park and Marron (1990) and Sheather and Jones (1991), which are based on asymptotic theory, our proposed methodology is a hybrid of the cross-validation method and the asymptotic theory. As such it does not require the estimation of $R(f'')$ or a third-stage estimation of $R(f''')$. (By convention, we denote $R(g) = \int g^2(x)dx$ for any given function g.) Hence, it is more straightforward to implement than plug-in estimators. Most importantly, it can be used in a wide variety of problems where plug-in methodology is not available.

We present an extensive simulation study in Section 4.1 to compare the proposed methods with the conventional cross-validation estimator. It will be seen that our bandwidth selectors achieve a smaller expected integrated square error that is much closer to the theoretical optimum than the standard cross-validation. Moreover, a comparison of the proposed methods to indirect cross-validation (Savchuk et al., 2011, 2010; Mammen et al., 2012) can be found in Section 4.2. In addition, we compare our methods to the asymptotic selection of the subsample size m that was described in Marron (1987).

2. U-statistic estimate of L^2 risk

In this section, we will derive a simple U-statistic form estimator for the risk that arises from L^2 distance. It is a new representation for the unbiased risk estimator and enables us to calculate the aggregated risk at subsamples of size m ($m \leq n$) much more efficiently than the repeated bootstrapping done in Hall and Robinson (2009) or the partitioning method used in Marron (1987).

2.1. L^2 distance-based assessment

Define the integrated square error (ISE) as

$$\mathrm{ISE}(h) = \int \left\{\hat{f}_h(x \mid \mathcal{X}_n) - f(x)\right\}^2 dx. \tag{2.1}$$

If one wants to evaluate $\hat{f}_h$ over all possible samples of size n, one can consider the mean integrated squared error (MISE), also known as the L^2 risk.

$$\mathrm{Risk}_{L^2}(h, n) = \mathrm{MISE}(h) = E_{\mathcal{X}_n}\left[\int \left\{\hat{f}_h(x \mid \mathcal{X}_n) - f(x)\right\}^2 dx\right]. \tag{2.2}$$

By Fubini's theorem, the risk in (2.2) can be decomposed in the following fashion: $\mathrm{Risk}_{L^2}(h, n) = \int E_{\mathcal{X}_n}\{f^2(x) - 2f(x)\hat{f}_h(x) + \hat{f}_h^2(x \mid \mathcal{X}_n)\}dx$. Furthermore, one can omit terms independent of h and focus on the relative risk, denoted as $R_{L^2}(h, n) = -2E_{\mathcal{X}_n}\{\int f(x)\hat{f}_h(x \mid \mathcal{X}_n)dx\} + E_{\mathcal{X}_n}\{\int \hat{f}_h^2(x \mid \mathcal{X}_n)dx\}$. It can be shown by simple algebra that the first expectation in $R_{L^2}(h, n)$ can be represented as $E\{K_h(X_1 - X_2)\}$. Moreover, the second expectation equals $E\{(K_h * K_h)(X_1 - X_2)\} + n^{-1}E\{(K_h * K_h)(0) - (K_h * K_h)(X_1 - X_2)\}$, where $*$ is the convolution operator.

If we denote

$$A_h(X_1, X_2) = (K_h * K_h)(X_1 - X_2) - 2K_h(X_1 - X_2), \tag{2.3}$$

$$B_h(X_1, X_2) = (K_h * K_h)(0) - (K_h * K_h)(X_1 - X_2), \tag{2.4}$$

then

$$R_{L^2}(h, n) = E\{A_h(X_1, X_2)\} + n^{-1}E\{B_h(X_1, X_2)\}. \tag{2.5}$$

Note that the *only* dependence on sample size n on the right hand side of (2.5) occurs in the multiplier n^{-1}. In the case of Gaussian kernel, $(K_h * K_h)(X_1 - X_2) = K_{\sqrt{2}h}(X_1 - X_2)$ and $(K_h * K_h)(0) = 1/(2h\sqrt{\pi})$. One can also denote $\mathrm{MISE}(h)$ as $\int \mathrm{bias}^2(\hat{f}_h(t))dt + \int \mathrm{Var}(\hat{f}(t))dt$. Then, $E\{B_h(X_1, X_2)\} = n\int \mathrm{Var}(\hat{f}_h(t))dt$ corresponds to the integrated variance, and $E\{A_h(X_1, X_2)\} = \int \mathrm{bias}^2(\hat{f}_h(t))dt - \int f(t)^2 dt$ is the relative integrated squared bias.

Following the footsteps of Ray and Lindsay (2008) and Lindsay and Liu (2009), we propose to estimate the risk at sample sizes m that may be much smaller than the actual size n. Our $m < n$ paradigm motivates us to hypothesize that these estimators will have much lower variability. Our simulations verify this. We also know that the cross-validation criterion for bandwidth selection tends to choose the bandwidths that are too small, and so overfit the density (Loader, 1999). As we will show, the new methodology particularly avoids the overfitting problem by reducing the chances of selecting a small bandwidth. Throughout this paper, we use m to represent the subsample size, which we might also call the *fictional sample size*. As we will see, it is closely related to the training sample size in cross-validation.

Denote the relative risk evaluated at fictional size m as

$$R_{L^2}(h, m) = E\{A_h(X_1, X_2)\} + m^{-1}E\{B_h(X_1, X_2)\}. \tag{2.6}$$

This formula gives an important insight into the risk estimation problem. The only place the fictional size m appears is as a coefficient. If we take the derivative of $R_{L^2}(h, m)$ (2.6) with respect to h and set it equal to zero, we thereby identify h as a potential minimum to the risk for that m. We can invert this thinking and solve $R'_{L^2}(h, m) = 0$ with h fixed, thereby finding the m that leads to optimization of the risk.

$$m^*(h) = -\frac{d}{dh}E\{B_h(X_1, X_2)\} \Big/ \frac{d}{dh}E\{A_h(X_1, X_2)\}. \tag{2.7}$$

The uniqueness of the solution shows that for each h there is exactly one m for which it is potentially optimal. In particular, if K is the Gaussian kernel and f is the standard normal, we have

$$m^*(h) = \frac{\{(h^2+1)(h^2+2)\}^{3/2} - h^3(h^2+2)^{3/2}}{2\sqrt{2}h^3(h^2+1)^{3/2} - h^3(h^2+2)^{3/2}}. \tag{2.8}$$

If we desire the optimal h for a particular fixed m_0, we solve the inverse problem $m_0 = m^*(h_{\mathrm{opt}})$ for h_{opt}. This normal theory m^* curve (2.8) will be examined later in light of the asymptotic theory (see Fig. 1). For now we note that in the normal example m^* is decreasing in h. It follows that the optimal bandwidth h for each m is a decreasing function of m.

Remark 1. For difficult densities, the theoretical curve $m^*(h)$ is not monotonic, and so many methods based on asymptotics are likely to fail at some sample sizes. For example, if the curve has the shape $\smile$, with two regions of decrease separated by a region of increase, then for some values of m there will be three solutions in h to $m = m^*(h)$. These will correspond to two local minima to the risk curve and the local maximum in-between. The central region in which $m^*(h)$ is increasing corresponds to values of h that are never optimal for any value of m. For an example of this, see Fig. 6, where we show the $m^*(h)$ curve for the claw density that will be discussed in Section 4.1.

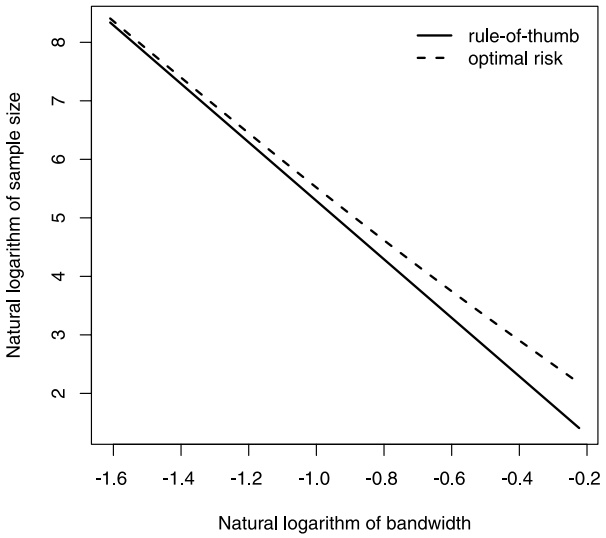

Fig. 1. Sample size against bandwidth when f is $N(0, 1)$ on the log–log scale.

2.2. An unbiased estimate for L^2 risk

Because the relative L^2 risk at fictional size m, $R_{L^2}(h, m)$ (2.6), involves the unknown density function f, we cannot use it directly to find the optimal bandwidth. In practice, one needs to first estimate the risk based on a set of observations and then select an estimated optimal bandwidth by minimizing the estimated risk score. A straightforward (and unbiased) estimation for $R_{L^2}(h, m)$ is by constructing a U-statistic based on a kernel function of size two.

Define

$$U_{L^2,m} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \psi_{L^2,h}(x_i, x_j), \tag{2.9}$$

where $\psi_{L^2,h}(x_1, x_2) = A_h(x_1, x_2) + m^{-1}B_h(x_1, x_2)$ is a symmetric kernel function of size two, with A_h and B_h defined earlier in (2.3) (2.4). Because a general U-statistic is a function of the order statistics, $U_{L^2,m}$ (2.9) therefore is the best unbiased estimator for the relative risk in this nonparametric context (Fraser, 1954).

Note that $U_{L^2,m}$ is equivalent to the unbiased cross-validation formula when $m = n$. That is, both of them are unbiased estimates for the relative MISE and are functions of the order statistics (modulo terms that do not depend on h). In addition, the un-rescaled bagging cross-validation bandwidth selector proposed by Hall and Robinson (2009) is actually nothing more than the bandwidth selector based on (2.9) if one makes the bootstrap size equal to m. However, the simple expression for our U-statistic risk estimator enables us to compute the aggregated risk much more efficiently than bootstrapping. First, we can generally compute the complete U-statistics, being of order two, much more efficiently than subsampling subsets of size m. Secondly, the calculations can be done for all m at once, in effect.

If we minimize (2.9) over h by setting the derivative in h to zero, we have

$$\hat{m}^*(h) = -\sum_{1\le i<j\le n} \frac{d}{dh} B_h(x_i, x_j) \Big/ \sum_{1\le i<j\le n} \frac{d}{dh} A_h(x_i, x_j). \tag{2.10}$$

This equation describes the dependence structure between m and h in minimizing the estimated L^2 risk for a given sample of size n. In practice, one can construct a U-statistic form estimate for the risk at a fictional sample size m. Then, by minimizing the U-statistic risk estimate one can obtain the optimal bandwidth choice for a given value of m. We call this the *simple subsampling bandwidth selector*. We note that the empirical curve $\hat{m}^*(h)$ is not necessarily strictly decreasing in h. If this happens, root selection rules must be applied.

Remark 2. It is easy to plot the empirical curve $\hat{m}^*(h)$ for any particular data set. If the empirical curve is not monotonic decreasing in the region of m values of interest, we would recommend against using extrapolation or plug-in methods.

3. Bandwidth selection procedures

In this section we will examine several ways to use $m < n$ risk estimation to improve performance in L^2 risk minimization. They will proceed in order of increasing sophistication in their use of asymptotic theory.

3.1. Using risk at m to select bandwidth

The simplest way to use the reduced variability of the risk estimator for smaller values of m is to pick the bandwidth h selected based on $U_{L^2,m}(h)$. In addition to reducing variance in h estimation, however, one is introducing bias, and so one must consider the trade-off that occurs between bias and variance.

We carried out a simulation study that indicated that as m decreased the optimal bandwidth h_m became larger (see the 2012 Pennsylvania State University Ph.D. thesis of Q. Wang which is available electronically from Pennsylvania State University library). The average integrated square error decreased as one used fictional size $m < n$ until m reached a certain threshold, beyond which further improvement in MISE could not be achieved. For instance, in the standard normal case half-sampling with $m = n/2$ gave us the smallest simulated MISE, with an improvement of about 12% over $m = n$ for samples of size $n = 100$. This conclusion agrees with the result in Hall and Robinson (2009).

Although using a fictional size $m < n$ can help to reduce the variability of the bandwidth selector and therefore achieve a smaller MISE, it turns out there are simple techniques to reduce this bias without increasing variance. We will show next in Section 3.2 how to do this.

3.2. First-order extrapolation in selecting h

We now propose a two-stage, subsampling-extrapolation, approach in bandwidth selection. We will first develop a first-order extrapolated bandwidth selector. Motivated by the rule-of-thumb criterion, the two pieces we combine are the estimated optimal bandwidth at $m < n$ and the rate of convergence of the estimator as $n \to \infty$. Later we will offer further refinements to this method. We will then provide a simulation comparison of all methods.

Recall the U-statistic form risk estimator, $U_{L^2,m}$ (2.9), computed based on squared distance and evaluated at a fictional sample size m. We denote the corresponding simple subsampling bandwidth selector as $\hat{h}_{L^2,m}$, which minimizes the U risk estimate at size m. Although $\hat{h}_{L^2,m}$ $(m < n)$ was less variable than $\hat{h}_{L^2,n}$, it tended to be biased larger than the optimal bandwidth choice h_{opt} at sample size n. This is due to the fact that $m^*(h)$ (2.7) is decreasing in h, and we have evaluated the risk at a size m less than the original sample size n.

Our goal is to take advantage of the small variability of the risk estimate at fictional size $m < n$ and also try to remove the incurred bias in $\hat{h}_{L^2,m}$ by referring to the asymptotic relationship between m and h on the log–log scale. We can motivate our approach using the following well-known theory. If the density f is known, the optimal bandwidth for minimizing the asymptotic MISE based on an order-2 kernel can be written as

$$h_{\text{opt}}(m) = m^{-1/5} C(f),$$

where $C(f)$ is a constant depending on f. A typical rule-of-thumb bandwidth selector estimates the constant $C(f)$ from the data in some way. For example, if we assume that both the true distribution and the kernel function are Gaussian, we then have $\hat{h}_{\text{rot}} = 1.06\hat{\sigma} m^{-1/5}$, where $\hat{\sigma}$ is an estimate for the population standard deviation.

We have derived an explicit formula for finding $m^*(h)$, the value of m for which h is optimal. This asymptotic formula can be rearranged to say that, in the limit as m gets large, this relationship can be represented as

$$\log m = \mathbb{C} - 5 \log \hat{h}_{\text{rot}}. \tag{3.1}$$

Here $\mathbb{C}$ is a constant independent of the bandwidth $\hat{h}_{\text{rot}}$. This equation represents a straight-line relationship with slope -5 on the log–log scale. One may ask whether this simple linear relationship is a good approximation for the exact relationship between $\log m^*(h)$ and $\log h$ found in (2.7). Fig. 1 displays the comparison between the rule-of-thumb criterion and the optimal risk criterion on the log–log scale when the underlying distribution f is the standard normal, and the kernel function K is Gaussian. It can be seen that for a fixed value of m the rule of thumb always yields a smaller bandwidth than the one given by the exact risk curve, but their left hand asymptotes match.

Our derivation of (3.1) from the rule-of-thumb selection rule was heuristic. We therefore show more formally that this relationship is valid for any arbitrary smooth kernel function K and density function f. The following lemma verifies this statement. For proof, please see Appendix.

Lemma 1. *Assume K is a smooth symmetric kernel function of order r $(r \geq 2)$, and f is a probability density function that is $(2r-1)$th order differentiable. By minimizing the L^2 risk, we can obtain the explicit expression of the fictional sample size m for which h would be optimal:*

$$m^*(h) = -\frac{d}{dh} E\{B_h(X_1, X_2)\} \Big/ \frac{d}{dh} E\{A_h(X_1, X_2)\},$$

where $A_h(x_1, x_2)$ and $B_h(x_1, x_2)$ are defined in (2.3) and (2.4). Moreover, it can be shown that

$$\log m^*(h) + (2r+1) \log h \to \text{constant}, \quad \text{as } h \to 0.$$

And,

$$d \log m^*(h) / (d \log h) \to -(2r+1), \quad \text{as } h \to 0.$$

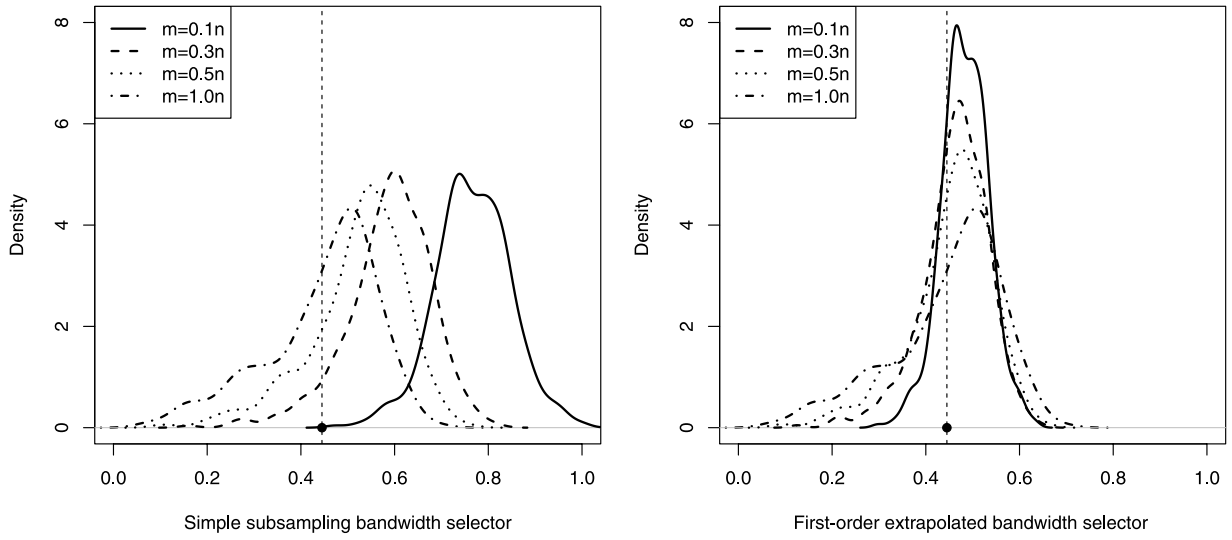

Fig. 2. Sampling distributions of simple subsampling and first-order extrapolated bandwidth selectors. The vertical dashed line marks the actual optimal bandwidth.

According to Lemma 1, the optimal risk curve approaches the rule-of-thumb straight line on log–log scale as $\log h \to -\infty$. In particular, when one considers a Gaussian kernel, the order of the kernel function r is 2. In this case, we would have $d \log m^*(h)/(d \log h) \to -5$ as $h \to 0$. This confirms that the asymptotic slope of the optimal risk curve in Fig. 1 is indeed -5, the same as the slope of the rule-of-thumb straight line.

As a result, if one knows the bandwidth selected for m, one simple way to remove the bias introduced by using $m < n$ is to extrapolate the bandwidth selected at size m to the optimal value at n based on the approximate linear relationship between $\log m^*$ and $\log h$. We summarize the two-stage bandwidth selection procedure based on linear extrapolation as follows:

1. *Subsampling stage*: Construct the U-statistic estimate for the L^2 risk at a fictional size m ($m \leq n$) and obtain the subsampling bandwidth selector, denoted as $\hat{h}_{L^2,m}$.

$$\hat{h}_{L^2,m} = \arg\min_{h>0} U_{L^2,m}(h). \tag{3.2}$$

2. *Extrapolation stage*: Extrapolate $\hat{h}_{L^2,m}$ to an approximation for $\hat{h}_{L^2,n}$ by referring to the approximate linear relationship between $\log m^*$ and $\log h$ as discussed in Lemma 1. This gives the estimator

$$\hat{h}_1 = (m/n)^{1/5}\, \hat{h}_{L^2,m}. \tag{3.3}$$

We call $\hat{h}_1$ the *first-order extrapolated bandwidth selector*. Intuitively $\hat{h}_1$ should have low bias when m is close to n and when $(\log h, \log m^*(h))$ relationship in minimizing the L^2 risk resembles a straight line. The fact that $\hat{h}_1$ is a shrunken version of $\hat{h}_{L^2,m}$ means that it has less variance and so, as we shall see, the variability of the bandwidth selector is reduced significantly compared with the traditional cross-validation bandwidth selector.

For any particular density f there will exist an optimal choice of the fictional size m such that it optimizes the trade-off between bias and variance of the bandwidth selector. Fig. 2 shows the density curves for the sampling distributions of the simple subsampling bandwidth selector $\hat{h}_{L^2,m}$ and the first-order extrapolated bandwidth selector $\hat{h}_1$ at different fictional sample sizes m. These density curves were plotted based on drawing $R = 500$ samples of size $n = 100$ from the standard normal distribution. This plot demonstrates graphically how the first-order extrapolation corrected for the bias incurred by using $m < n$ while simultaneously reducing variability over standard cross-validation ($m = n$). In particular, it greatly decreased the selection of values of h that were "too small", corresponding to overfitting. Later in Table 4 it will be seen that the first order extrapolation improved the MISE efficiency ratio, $\text{MISE}_{\text{opt}}/\text{EISE}(\hat{h}_1)$, of the standard cross-validation from 64% to over 80%. Here EISE stands for expected integrated square error. The use of EISE rather than MISE in assessing bandwidth selectors was suggested by Jones (1991). In our simulation section we will determine the optimal value of $p = m/n$ for a number of sampling distributions. It will be shown there that the optimal value of p varies somewhat based on the smoothness of the density, but $p = 0.3$ worked well for many. One possible strategy for the extrapolation estimator is to use a fixed value of p regardless of the data. We will compare this strategy with a few others in the simulation section.

3.3. Second-order extrapolation in selecting h

The first-order extrapolated bandwidth selector introduced in Section 3.2 is based on an approximate linear relationship between $\log m^*(h)$ and $\log h$. We have shown in Lemma 1 that the true optimal risk curve approaches the rule-of-thumb straight line as h goes to 0. That is, the linear approximation is most accurate when h is fairly small. Therefore, we wonder whether we could improve the approximation and seek a more accurate relationship between $\log m^*(h)$ and $\log h$ that would be useful for smaller values of n.

Notice from Fig. 2 that the first-order extrapolated bandwidth tends to be biased, more and more so as the range of extrapolation increases. As the optimal risk curve of $m^*(h)$ is decreasing in h, we propose to consider a second-order correction. We start by noticing that the explicit expression for $m^*(h)$ can be written as

$$m^*(h) = \frac{(1/2\sqrt{\pi})h^{-2} + 2h\mu_2(\phi)C_{02} + o(h)}{4h^3\mathbb{C}_1 + 6h^5\mathbb{C}_2 + o(h^5)},$$

where $\mu_j(\phi) = \int x^j\phi(x)dx$ and ϕ is a Gaussian kernel, and $C_{ij} = \int f^{(i)}(x)f^{(j)}(x)dx$ for $i, j \in \mathbb{Z}$ with the assumption that $f^{(0)}(x) = f(x)$. In addition, $\mathbb{C}_1 = \mu_4(K)C_{04}/24 + (\mu_2(\phi))^2C_{22}/8$ and $\mathbb{C}_2 = (3/96)\mu_2(\phi)\mu_4(\phi)C_{24}$.

To find the correct second order expansion, we need to take this expansion to the next term. We use the approximation $\log(1 + x) \approx x$ for x close to 0, and write

$$\begin{aligned}
\log\left\{-\frac{d}{dh}E(B_h(X_1, X_2))(2h^2\sqrt{\pi})\right\} &= \log\{1 + 2h^3\mu_2(\phi)C_{02}(2\sqrt{\pi}) + o(1)\} \\
&\approx 2h^3\mu_2(\phi)C_{02}(2\sqrt{\pi}) \\
\log\left\{\frac{d}{dh}E(A_h(X_1, X_2))(4h^3\mathbb{C}_1)^{-1}\right\} &= \log\{1 + (3\mathbb{C}_2/2\mathbb{C}_1)h^2 + o(1)\} \\
&\approx h^2(3\mathbb{C}_2/2\mathbb{C}_1).
\end{aligned}$$

Therefore, the overall magnitude of the second-order error on the log scale is h^2.

We then assume

$$m^*(h) \approx \tilde{m}(h) = \mathbb{C}_0h^{-5}e^{ah^2}, \tag{3.4}$$

where $\mathbb{C}_0$ is a constant and $a \in \mathcal{R}$. Note that the adjustment term e^{ah} goes to 1 as $h \to 0$ but is strictly bigger than 1 if $h > 0$ and $a > 0$.

We propose to approximate parameter a by matching $m^*(h)$ and $\tilde{m}(h)$ at two values of h and solving for the unknowns. Let h_0 and c_0h_0 ($c_0 > 1$) be two chosen bandwidths. The two equations

$$\begin{aligned}
m^*(h_0) &= \mathbb{C}_0h_0^{-5}e^{ah_0} \\
m^*(c_0h_0) &= \mathbb{C}_0(c_0h_0)^{-5}e^{c_0ah_0}
\end{aligned}$$

then provide a way to solve for the unknown parameter a. In particular,

$$\hat{a} = \frac{1}{h_0^2(c_0^2 - 1)}\log\left\{c_0^5\frac{m^*(c_0h_0)}{m^*(h_0)}\right\} \geq 0.$$

Then, for any unknown h, formula (3.4) can be represented as

$$m^*(h) = m^*(h_0)\frac{m^*(h)}{m^*(h_0)} \approx m^*(h_0)\,(h/h_0)^{-5}\,e^{\hat{a}(h-h_0)} := m^{**}(h).$$

We propose to invert the curve m^{**} determined by the last approximation to estimate an optimal bandwidth for any particular m including $m = n$. We note that the inversion relationship can be expressed as an explicit correction to the log–log linear relationship.

$$\log m^{**}(h) = \log m^*(h_0) - 5\,(\log h - \log h_0) + \hat{a}(h^2 - h_0^2). \tag{3.5}$$

As seen in (3.5), m^{**} is in fact just an exponential curve fitted through the two points $(\log h_0, \log m^*(h_0))$ and $(\log(c_0h_0), \log m^*(c_0h_0))$ and with slope -5.

Fig. 3 illustrates the theoretical curves of $(\log h, \log m(h))$ relationship when f is the standard normal. The solid line is based on the rule-of-thumb criterion; the dashed curve represents the exact relationship between $\log m^*(h)$ and $\log h$ in minimizing the L^2 risk (2.8); the dotted curve displays the relationship of the second-order extrapolation based on formula (3.5), using $c_0 = 2$ and $h_0 = 0.5$ as an illustration.

It can be clearly seen that the second-order extrapolation curve was surprisingly close to the true optimal risk curve for the standard normal case. It should provide less bias than the first-order extrapolation, but it could add variability. In other words, we might expect the second-order extrapolation method to outperform the linear extrapolation bandwidth selector for cases where the density function is fairly smooth.

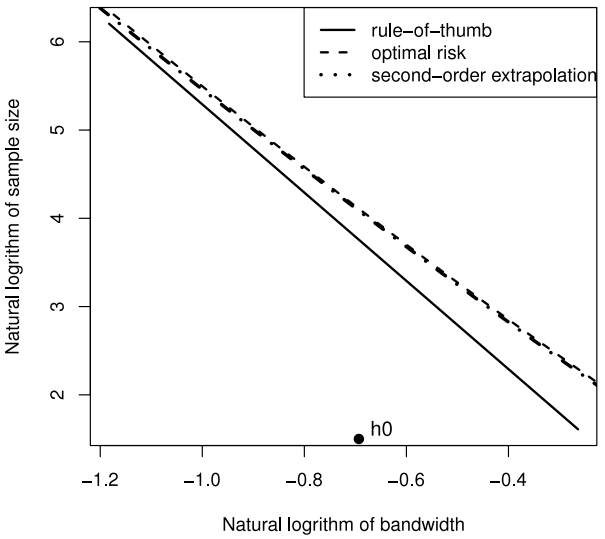

Fig. 3. Sample size against bandwidth on the log–log scale when f is standard normal. Here, $h_0 = 0.5$ as an illustration.

In our implementation of this second-order extrapolation in the simulation, we chose h_0 in (3.5) to be the subsampling bandwidth selector $\hat{h}_{L^2,m}$ at the fictional size $m = pn$. As a result, $m^*(h_0) \approx m$. We also used $c_0 = 2$, and estimated m^* by using Eq. (2.10):

$$m^*(c_0 h_0) \approx \hat{m}^*(c_0 \hat{h}_{L^2,m}) = -\frac{\sum_{1\le i<j\le n} \frac{d}{dh} B_{c_0\hat{h}_{L^2,m}}(X_i, X_j)}{\sum_{1\le i<j\le n} \frac{d}{dh} A_{c_0\hat{h}_{L^2,m}}(X_i, X_j)}.$$

Parameter a was then estimated by $\hat{a} = \{\hat{h}^2_{L^2,m}(c_0^2 - 1)\}^{-1} \log\{c_0^5 \hat{m}^*(c_0\hat{h}_{L^2,m})/m\}$.

Under this scheme, given a sample of size n, the optimal bandwidth at n, as extrapolated from m ($m < n$) based on Eq. (3.5) is the root of the following score function, denoted as $\hat{h}_2$. We call it the *second-order extrapolated bandwidth selector*.

$$\log n = \log m - 5\left(\log h - \log \hat{h}_{L^2,m}\right) + \hat{a}(h^2 - \hat{h}^2_{L^2,m}). \tag{3.6}$$

Eq. (3.5) indicates that $m^{**}(h)$ is bigger than $\mathbb{C}_0 h^{-5}$, so we would expect $\hat{h}_2$ to be smaller than the first-order extrapolation bandwidth selector $\hat{h}_1$ for a given value of m.

To illustrate the advantage of the second-order extrapolation in comparison with the first-order extrapolation, we revisited the numerical study presented in Fig. 2 but now implemented the second-order extrapolation technique. In Fig. 4 we compared the simulated density plots of the first-order and second-order bandwidth selectors when the fictional sample size $m = 0.1n$, $0.3n$, $0.5n$ or $1.0n$. It is clearly seen that for small values of m, the improvement of the bandwidth selector based on second-order extrapolation is noticeable.

3.4. Selecting the optimal fictional size: nested cross-validation

We will see in our simulation section that the optimal fictional size for both first and second-order extrapolation depends on the choice of $p = m/n$. In real problems one cannot determine the optimal choice of the fictional size $m = pn$ that minimizes the risk at n. As a result, when implementing the two-stage bandwidth selection procedure, one needs to either fix the choice of m prior to bandwidth selection, or implement an automatic, data-driven method to pick the best choice of m based on a data set. Previous literature, such as Hall and Robinson (2009), Shah and Samworth (2012), and Shao (1993), suggest to use $m \le n/2$. However, in their papers there does not exist an automatic selection method for picking the best choice of m in the subsampling (or bagging) procedure. Here we propose a nested cross-validation methodology in selecting the optimal fictional sample size m that overcomes the drawback of choosing fictional size m subjectively. This method seems to perform consistently well across a wide selection of distributions (Section 4.1). We note that this two-layer cross-validation is made computationally feasible by our U-statistic estimation of the bandwidth selector curve.

Let $p = m/n$ be the fixed proportion of data used in the fictional sample. We consider a cross-validation strategy for selecting p. Let $\hat{f}_{p,n^*}$ be the estimator of the density based on extrapolation (first or second-order) for a data set of size n^* by subsampling of size $m = pn^*$. As we show below, from a sample of size n we can estimate the L^2 risk of $\hat{f}_{p,n^*}$ unbiasedly as

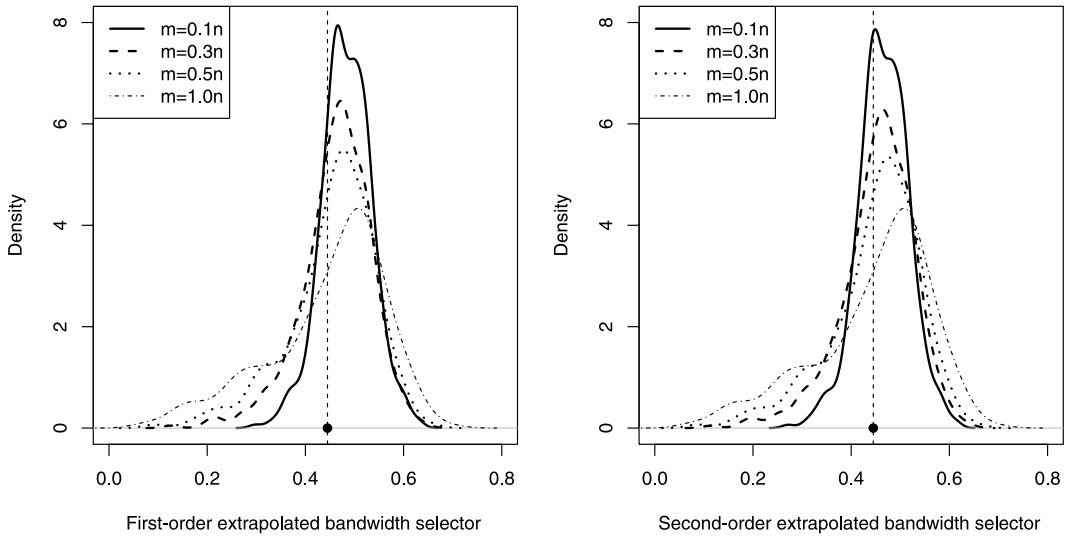

Fig. 4. Sampling distributions of first-order (left panel) and second-order (right panel) extrapolated bandwidth selectors when f is Standard Normal. The vertical dashed line marks the actual optimal bandwidth.

a function of p provided that $n^* < n$. We again acknowledge we will get more stable estimation when n^* is not close to n, and so use $n^* = n/2$ in our simulation. We choose the value of p that minimizes the estimated risk at size n^*. We then use this selected value of p to create the extrapolation estimator on the full sample of n data points.

Denote the data set of size n as $\mathcal{X}_n = (X_1, \ldots, X_n)$. We can take a random subsample S of size $n^* < n$, say $S = (X_1^*, \ldots, X_{n/2}^*)$ with $n^* = n/2$, without replacement out of $\mathcal{X}_n$. We then consider a grid of p values, i.e. p_j $(1 \le j \le J)$. For each choice of p and a subsample S, one can realize the subsampling-extrapolation bandwidth selector by first finding the optimal bandwidth at size pn^* and then extrapolating it from pn^* to n^*. We denote the extrapolated bandwidth at size n^* as $\hat{h}_{p,S}$. Note that $\hat{h}_{p,S}$ is dependent on both the proportion choice p and the subsample S.

We want to consider the L^2 risk of using bandwidth $\hat{h}_{p,S}$ at sample size n^* as a function of p and then determine the optimal choice of p at size n^*. We denote the risk at size n^* as

$$R_{L^2}(p, n^*) \propto E\left\{\int \hat{f}_{p,n^*}(x)^2 dx\right\} - 2E\left\{\int f(x)\hat{f}_{p,n^*}(x)dx\right\}, \tag{3.7}$$

where $\hat{f}_{p,n^*}(x) = n^{*-1}\sum_{i=1}^{n^*} K_{\hat{h}_p}(x - X_i)$, and $\hat{h}_p$ is the bandwidth extrapolated from pn^* to n^*.

Since the density estimator $\hat{f}_{p,n^*}(x)$ depends on both the random subsample S and the choice of p, the risk function $R(p, n^*)$ cannot be estimated by the standard U estimator in (2.9). However, one can estimate the first expectation in (3.7) unbiasedly by $\binom{n}{n^*}^{-1}\sum_S\left\{n^{*-2}\sum_i\sum_j(K_{\hat{h}_p} * K_{\hat{h}_p})(x_i - x_j)\right\}$. The second expectation in (3.7) can be estimated unbiasedly by

$$\binom{n}{n^*}^{-1}\sum_S\left\{(n - n^*)^{-1}\sum_{X_i\notin S}\hat{f}_{p,n^*}(X_i \mid S)\right\},$$

where S represents a subset of size n^* taken out of $\mathcal{X}_n$.

In practice, one can repeatedly draw subsamples of size n^* out of $\mathcal{X}_n$, denoted as $S_1, \ldots, S_B$. For a particular value of p, we denote the first-order extrapolated bandwidth at size n^* as $\hat{h}_{p,S_1}, \ldots, \hat{h}_{p,S_B}$. The estimated risk of using proportion p at size n^* is then

$$\hat{R}_{L^2}(p, n^*) \propto \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{n^{*2}}\sum_i\sum_j(K_{\hat{h}_{p,S_b}} * K_{\hat{h}_{p,S_b}})(x_i - x_j)\right\} - \frac{2}{B}\sum_{b=1}^{B}\left\{\frac{1}{n - n^*}\sum_{X_i\notin S_b}\hat{f}_{p,n^*}(X_i \mid S_b)\right\}. \tag{3.8}$$

There exists a choice of p that minimizes $\hat{R}_{L^2}(p, n^*)$ which is the optimal proportion p at size n^*. We might hope that the best p for n^* also yields satisfactory performance at sample size n. If the optimal choice of p does not largely rely on the subsample size n^*, then it would be reasonable to identify the optimal choice of p for n based on a smaller size n^* using the nested cross-validation methodology.

Remark 3. Marron (1987) presents a method that is closely related to bagging cross-validation which he calls partitioned cross-validation. Marron (1987) shows that the proportion of subsample, $p = m/n$, achieves its optimal value at

Table 1
Density functions considered in the simulation study.

Distribution	Probability density function
Standard normal	$N(0, 1)$
t(2)	$t(2)$
Mixture 1	$0.5N(-1.5, 1) + 0.5N(1.5, 1)$
Mixture 2	$0.5N(0, 1) + 0.5N(0, 0.1)$
Mixture 3	$0.5N(0, 1) + 0.5N(0, 0.01)$
Ten-fold mixture	$0.1\sum_{i=1}^{10} N(10i - 5, 1)$
Claw density	$0.5N(0, 1)+0.1\sum_{i=0}^{4} N(i/2-1, 0.01)$

$p_{\text{Marron}} = (\sigma/C_1)^{-5/4}n^{-3/8}$, where

$$C_1 = \frac{\left\{\left[\int K^2(x)dx\right]^{3/5}\right\}\left\{\int x^4K(x)dx\right\}\left\{\left[\int (f^{(2)}(x))^2dx\right]^2\right\}}{20\left\{\left[\int x^2K(x)dx\right]^{11/5}\right\}\left\{\left[\int (f^{(2)}(x))^2dx\right]^{8/5}\right\}},$$

$$\text{and } \sigma^2 = \frac{8\left\{\int (f(x))^2dx\right\}\left\{\int [K * (K-L)(x) - (K-L)(x)]^2\,dx\right\}}{25\left\{\left[\int K^2(x)dx\right]^{7/5}\right\}\left\{\left[\int x^2K(x)dx\right]^{6/5}\right\}\left\{\left[\int (f^{(2)}(x))^2dx\right]^{3/5}\right\}}.$$

Here, function L is defined as $L(x) = -xK'(x)$, and $*$ is the convolution operator. The constant $\int [K*(K-L)(x)-(K-L)(x)]^2dx$ in the definition of σ^2 can be obtained with the help of Corollary 6.4.1. in Aldershof et al. (1995). We refer to Marron's formula as MCV. Based on Marron's formula, p goes to zero as n goes to infinity. To compute the optimal partition size, one needs to estimate the integrated square derivatives of the unknown density function using two-stage estimators such as those proposed in Jones and Sheather (1991). We will compare Marron's formula with our nested cross-validation proposal through simulation studies in Section 4.1.

4. Simulation studies

We have now developed four possible methods for bandwidth selection based on subsampling plus extrapolation: we have the first- and second-order extrapolations carried out at a fixed, pre-chosen p, plus the two extrapolation methods using p chosen from a nested cross-validation. In order to investigate the performance of the proposed two-stage bandwidth selection procedures, we conducted the following simulation studies.

4.1. Comparison to unbiased cross-validation

We consider $R = 500$ random samples of size n ($n = 100$ or 200) drawn independently from a certain distribution. A list of the seven distributions under consideration can be found in Table 1, among which there are bimodal/multimodal, outlying, and heavy-tailed distributions. A particularly difficult density is the claw, with 5 extreme spikes, each with only 10% of the data. As we will see later in Table 4, it will create an outlier in our results. Other literature that consider the claw density and notice its unusual behavior include Marron and Wand (1992) and Loader (1999).

For each subset taken out of $\mathcal{X}_n$, we consider ten possible values of $p = m/n$, i.e. $p = 0.1, 0.2, \ldots, 1.0$. In selecting the optimal choice of p, we first compare the performance between the nested cross-validation with $n^* = n/2$ and Marron's formula (see Remark 3). We compute percent of times each p value is selected out of the 500 replications based on each method. Tables 2 and 3 reveal that the selection of p is quite noisy for $n = 100$, but does improve for $n = 200$. In making comparisons between the two methods, one should notice that Marron's optimal p is not exactly the minimizer of Eq. (3.7) but the asymptotic minimizer of $\text{MSE}(p) = E\{(\hat{h}_p - h_{\text{MISE}})^2\}$, with h_{MISE} being the minimizer of MISE. Regardless of the different target functions for minimization, Marron's formula has less variation than our nested method, and so might be expected to show somewhat better performance. We will investigate this point in our simulation section.

Next, we conduct an empirical investigation of how well one could estimate the optimal p using a nested cross-validation with $n^* = n/2$. The optimal proportion $p_{\text{opt},n}$ that is dependent on sample size n, is the minimizer of Eq. (3.7) with $n^* = n$. For several samples from each distribution, we plot the empirical expected integrated square error (EISE) curve as a function of p based on formula (3.8) in order to see whether we could do a good job at estimating the risk as a function of p. In Fig. 5 each of the panels displays three empirical curves (dashed curve) based on three random samples of size $n = 100$, taken from the corresponding distribution, and the theoretical true risk curve (solid curve) at size $n = 100$. (The curves for t(2) distribution are omitted here, as they show similar pattern as in the normal case.) Although the empirical curves are quite variable, the shape of each curve seems to follow the truth in most of the cases except the claw density. Thus, we will use the nested cross-validation methodology in the following discussions.

We then compare the performance of different bandwidth selectors in terms of the efficiency ratio, $\text{MISE}_{\text{opt}}/\text{EISE}(\hat{h}_p)$, where EISE stands for expected integrated square error. We use the exact MISE(h) formula in Theorem 2.1 of Marron and

Table 2
Percentages of optimal p selections out of $R = 500$ replications.

$R = 500$, $n^* = n/2$	$p = 0.1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Standard normal										
$p_{\text{opt},n} = 0.1$	Nested cross-validation ($n = 100$)									
Percent (%)	92.8	3.4	1.0	1.2	0.2	0.6	0.0	0.4	0.0	0.4
$p_{\text{opt},n} = 0.1$	Marron's Formula ($n = 100$)									
Percent (%)	0.0	53.6	43.8	2.6	0.0	0.0	0.0	0.0	0.0	0.0
$p_{\text{opt},n} = 0.1$	Nested cross-validation ($n = 200$)									
Percent (%)	93.0	3.8	1.0	0.2	0.2	0.2	0.0	0.2	0.2	1.2
$p_{\text{opt},n} = 0.1$	Marron's Formula ($n = 200$)									
Percent (%)	7.4	91.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mixture 1										
$p_{\text{opt},n} = 0.3$	Nested cross-validation ($n = 100$)									
Percent (%)	45.6	8.6	14.8	12.0	8.8	4.6	2.0	0.6	1.6	1.4
$p_{\text{opt},n} = 0.3$	Marron's Formula ($n = 100$)									
Percent (%)	0.2	1.6	17.2	80.8	0.2	0.0	0.0	0.0	0.0	0.0
$p_{\text{opt},n} = 0.3$	Nested cross-validation ($n = 200$)									
Percent (%)	18.2	18.8	30.8	17.6	7.2	2.6	1.6	1.2	0.6	1.4
$p_{\text{opt},n} = 0.3$	Marron's Formula ($n = 200$)									
Percent (%)	0.0	3.2	96.6	0.0	0.0	0.0	0.0	0.0	0.0	0.2
Mixture 2										
$p_{\text{opt},n} = 0.2$	Nested cross-validation ($n = 100$)									
Percent (%)	22.6	39.6	19.8	7.2	4.2	2.4	1.0	1.0	0.4	1.8
$p_{\text{opt},n} = 0.2$	Marron's Formula ($n = 100$)									
Percent (%)	0.0	0.4	43.4	46.6	8.6	0.6	0.4	0.0	0.0	0.0
$p_{\text{opt},n} = 0.2$	Nested cross-validation ($n = 200$)									
Percent (%)	21.0	47.6	17.4	8.2	2.0	1.6	0.8	0.6	0.4	0.4
$p_{\text{opt},n} = 0.2$	Marron's Formula ($n = 200$)									
Percent (%)	0.0	29.4	69.2	1.4	0.0	0.0	0.0	0.0	0.0	0.0

Table 3
Percentages of optimal p selections out of $R = 500$ replications.

$R = 500$, $n^* = n/2$	$p = 0.1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Mixture 3										
$p_{\text{opt},n} = 0.3$	Nested cross-validation ($n = 100$)									
Percent (%)	0.2	7.8	26.0	28.0	15.0	9.8	4.0	3.8	2.0	3.6
$p_{\text{opt},n} = 0.3$	Marron's Formula ($n = 100$)									
Percent (%)	0.0	0.0	2.0	85.2	12.6	0.2	0.0	0.0	0.0	0.0
$p_{\text{opt},n} = 0.2$	Nested cross-validation ($n = 200$)									
Percent (%)	0.0	19.4	40.8	19.8	10.2	4.6	1.4	0.8	1.2	1.8
$p_{\text{opt},n} = 0.2$	Marron's Formula ($n = 200$)									
Percent (%)	0.0	95.6	4.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claw density										
$p_{\text{opt},n} = 1.0$	Nested cross-validation ($n = 100$)									
Percent (%)	72.2	10.6	4.2	1.8	0.8	0.4	1.0	0.4	0.8	7.8
$p_{\text{opt},n} = 1.0$	Marron's Formula ($n = 100$)									
Percent (%)	0.0	0.0	0.6	3.2	11.4	32.0	36.2	10.8	5.4	0.4
$p_{\text{opt},n} = 1.0$	Nested cross-validation ($n = 200$)									
Percent (%)	32.6	10.0	2.8	0.2	0.0	0.2	1.4	1.6	5.0	46.2
$p_{\text{opt},n} = 1.0$	Marron's Formula ($n = 200$)									
Percent (%)	0.0	0.0	1.4	14.4	55.0	24.0	5.2	0.0	0.0	0.0
Ten-fold										
$p_{\text{opt},n} = 0.6$	Nested cross-validation ($n = 100$)									
Percent (%)	58.4	0.0	0.0	0.2	9.4	17.0	10.8	3.2	0.2	0.8
$p_{\text{opt},n} = 0.6$	Marron's Formula ($n = 100$)									
Percent (%)	0.0	0.0	0.0	1.2	14.6	31.2	29.0	18.4	5.6	0.0
$p_{\text{opt},n} = 0.4$	Nested cross-validation ($n = 200$)									
Percent (%)	0.0	0.0	0.2	13.2	46.2	28.6	8.6	2.6	0.0	0.6
$p_{\text{opt},n} = 0.4$	Marron's Formula ($n = 200$)									
Percent (%)	0.0	0.0	0.4	22.0	51.2	25.0	1.4	0.0	0.0	0.0

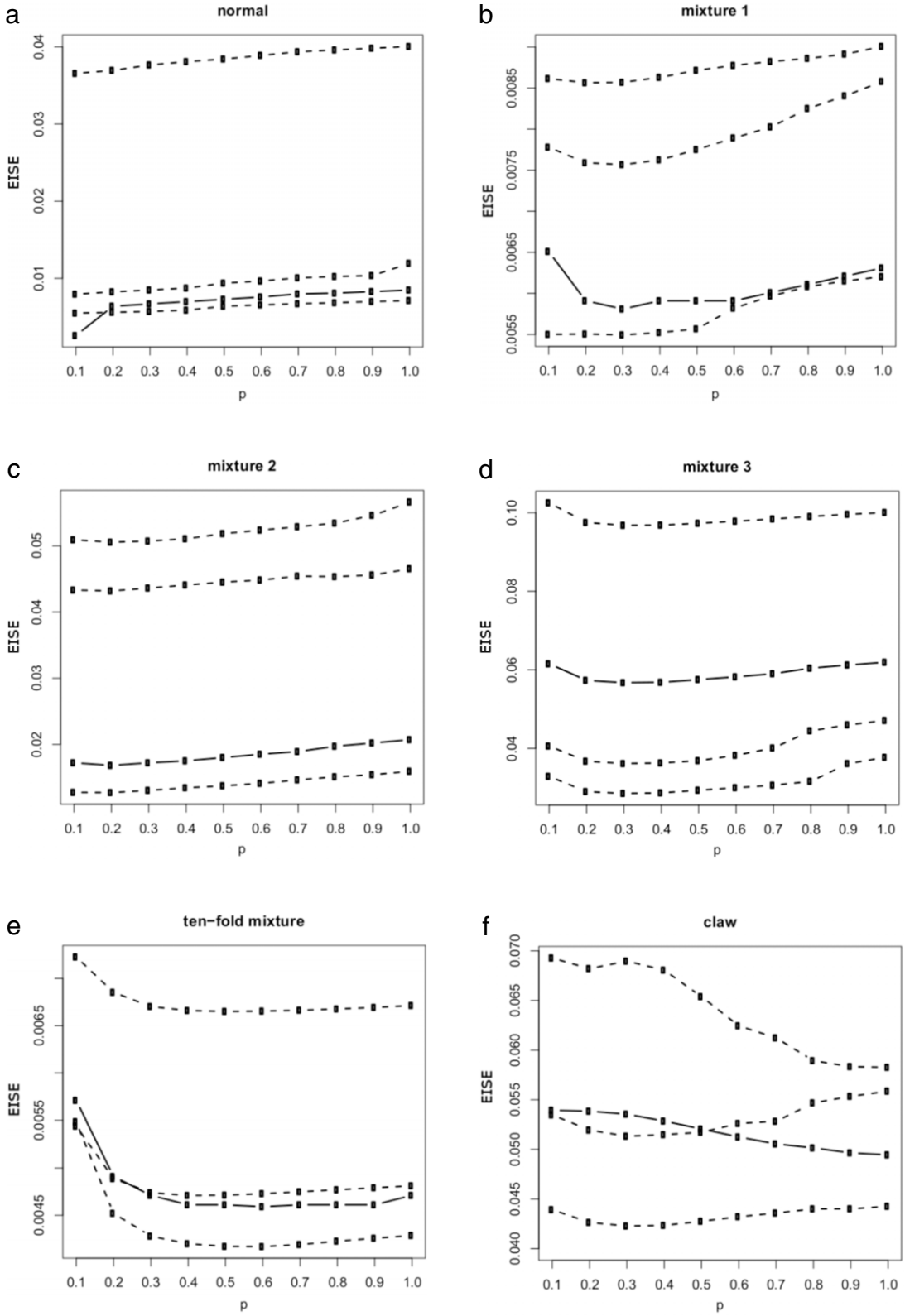

Fig. 5. Comparison of empirical curves of EISE(p) (dashed curves) and the theoretical truth (solid curve).

Table 4

Comparison of efficiency ratio, $\text{MISE}_{\text{opt}}/\text{EISE}(\hat{h}_p)$, where the mixtures are in order of the cross-validation efficiency. Boldface indicates the best of the five competing methods.

$n = 100$	CV	EX1			EX2		EX1	EX2
		MCV	NCV	Fixed p	NCV	Fixed p	Optimal p	Optimal p
Normal	63.7%	80.9%	81.5%	80.8% $p = 0.3$	**84.1**%	83.3% $p = 0.2$	87.4% $p = 0.1$	88.2% $p = 0.1$
t(2)	69.8%	82.8%	80.5%	82.6% $p = 0.3$	81.1%	**83.2**% $p = 0.2$	84.5% $p = 0.1$	85.0% $p = 0.1$
Mixture 2	68.5%	81.8%	79.5%	82.6% $p = 0.3$	83.5%	**84.8**% $p = 0.2$	84.4% $p = 0.2$	84.8% $p = 0.2$
Mixture 1	78.7%	84.3%	84.8%	**85.5**% $p = 0.3$	80.6%	84.8% $p = 0.2$	85.5% $p = 0.3$	85.2% $p = 0.3$
Mixture 3	80.1%	87.7%	87.8%	87.8% $p = 0.3$	88.2%	**88.4**% $p = 0.2$	87.5% $p = 0.2$	88.4% $p = 0.2$
Ten-fold	94.3%	**98.3**%	86.9%	94.3% $p = 0.3$	94.4%	97.2% $p = 0.2$	96.3% $p = 0.6$	98.9% $p = 0.5$
Claw	**74.7**%	72.9%	68.2%	68.5% $p = 0.3$	68.9%	68.8% $p = 0.2$	74.7% $p = 1.0$	74.7% $p = 1.0$
MISE	CV	EX1			EX2		EX1	EX2
$n = 200$		MCV	NCV	Fixed p	NCV	Fixed p	Optimal p	Optimal p
Normal	68.0%	86.6%	83.8%	83.8% $p = 0.3$	**89.9**%	86.1% $p = 0.2$	90.0% $p = 0.1$	90.0% $p = 0.1$
t(2)	73.9%	86.2%	83.8%	86.3% $p = 0.3$	86.0%	**86.7**% $p = 0.2$	87.9% $p = 0.1$	88.5% $p = 0.1$
Mixture 2	67.3%	86.3%	84.0%	86.1% $p = 0.3$	86.5%	**87.4**% $p = 0.2$	87.1% $p = 0.2$	88.5% $p = 0.1$
Mixture 1	73.7%	88.6%	80.1%	**93.3**% $p = 0.3$	91.4%	89.1% $p = 0.2$	93.5% $p = 0.4$	89.5% $p = 0.3$
Mixture 3	77.4%	87.1%	86.1%	87.1% $p = 0.3$	87.2%	**87.8**% $p = 0.2$	87.2% $p = 0.2$	87.8% $p = 0.2$
Ten-fold	93.6%	98.8%	95.3%	95.3% $p = 0.3$	99.2%	**99.5**% $p = 0.2$	95.7% $p = 0.4$	99.6% $p = 0.4$
Claw	**82.1**%	69.1%	61.7%	52.0% $p = 0.3$	48.4%	46.6% $p = 0.2$	82.1% $p = 1.0$	82.1% $p = 1.0$

Wand (1992) to compute the theoretical optimal bandwidth for normal mixtures, and approximate the theoretical optimal bandwidth for t(2) by simulation. Notice that in practice without replication of size-n samples, the optimal choice of p at size n is not obtainable. We first determine how the optimal choice of p varies over our sampling distributions. In Table 4 EX1 and EX2 stand for first- and second-order extrapolations respectively. In the last two columns of Table 4, one can see the relative efficiency ratio that can be attained when one uses the optimal value of p for that density. The optimal p seems to vary somewhat over the densities. One can also see that the second-order extrapolation, with optimal p, seems to do better than first-order extrapolation. After comparing the EISE over a grid of values of p, we choose to use $p = 0.3$ for the fixed-p first-order extrapolation and $p = 0.2$ for the fixed-p second-order extrapolation; these values seem to yield satisfactory results across a wide range of distributions.

Then, we focus on the comparison between the four proposed bandwidth selectors, i.e. the first- and second-order extrapolated bandwidth selectors with pre-chosen p or with nested cross-validation, and the unbiased cross-validated bandwidth selector. In addition, we also include the comparison between Marron's formula (see Remark 3) and nested cross-validation in selecting the optimal p in the context of first-order extrapolation. Because Marron (1987) only focuses on the study of first-order extrapolation, the comparison between Marron's optimal p and the nested cross-validation method is only fair for the first-order extrapolated bandwidth selector $\hat{h}_1$. We use MCV to stand for Marron's formula and NCV to stand for nested cross-validation. In formula (3.8) $\hat{h}_p$ is set to be the first-order extrapolated bandwidth for EX1 and the second-order extrapolated bandwidth for EX2. The realization of the second-order extrapolated bandwidth selector is based on setting $h_0 = \hat{h}_{L^2,m}$ and $c_0 = 2$.

Our first observation from Table 4 is that the results from the claw density are distinctly different from the rest. This density is considered in Loader (1999) to demonstrate that plug-in methodologies could perform very poorly relative to conventional cross-validation. Van Es (1992) shows that the relative rate of convergence of ordinary cross-validation bandwidth selector is faster for non-smooth cases, such as in the case of claw density. It is noted in Marron and Wand (1992) that the true MISE curve of the claw density has local minima when the sample size $n \leq 53$ (see Fig. 6). That is, $m^*(h)$ function has the $\smile$ shape in the region of m of interest as seen in Fig. 7. It is clear from this plot why extrapolation of this curve can work so poorly at some sample sizes. Only when n increases to above 100, does the MISE curve start to have an obvious

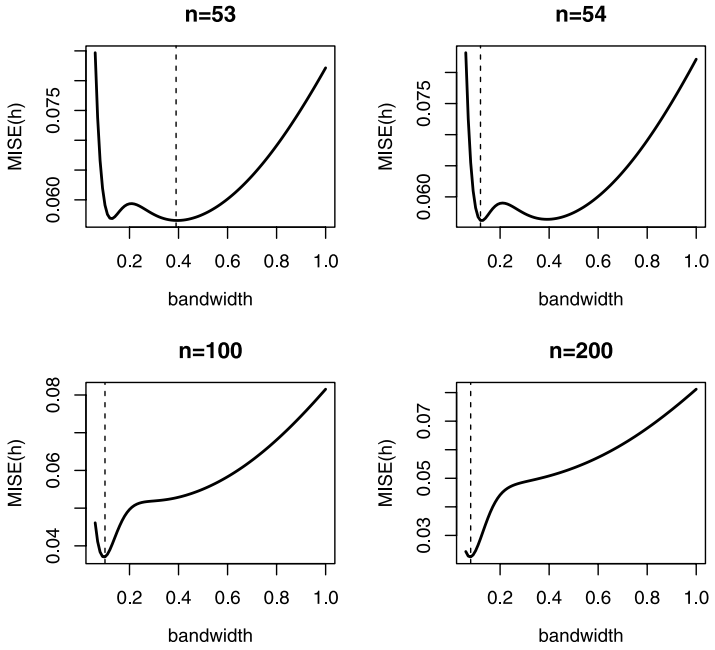

Fig. 6. The theoretical curve of MISE(h) for claw density at different sample sizes n. The dashed line in each panel marks the global minima of MISE(h).

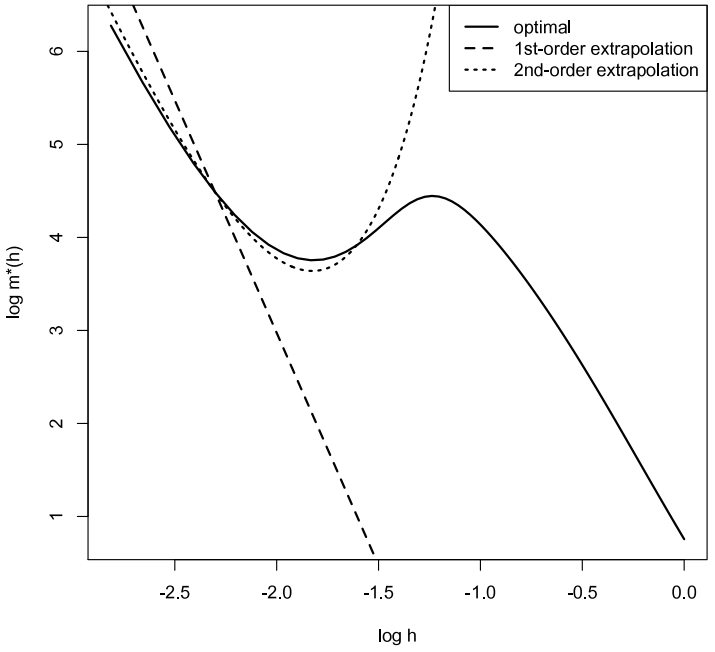

Fig. 7. The theoretical and extrapolated curves of $m^*(h)$ in the case of claw density on the log–log scale. The extrapolated point for the first-order extrapolation is $h_0 = 0.1$ (-2.3 on log scale); the extrapolated points for the second-order extrapolation is $h_0 = 0.1$ and 0.2 (-2.3 and -1.6 on log scale).

global minima. In this example, we have found that the efficiency ratio for the plug-in bandwidth selector of Sheather and Jones (1991) is as small as 9.9% for $n = 100$ and 6.1% for $n = 200$. Our extrapolation methods do perform much better than plug-in methods in the case of claw density. Moreover, a user who follows our advice not to use extrapolation methods or plug-in when $\hat{m}^*(h)$ is non-monotonic in the region of interest would end up using conventional cross-validation most of the time.

If we ignore the claw density, we can make the following general observations: If we compare the methods that use first-order extrapolation, it is clear that for both sample sizes $n = 100$ and $n = 200$, the fixed-p method yields slightly better

Table 5
Comparison to indirect cross-validation based on $\hat{E}\{ISE(\hat{h})/ISE(\hat{h}_0)\}$.

Standard normal						
n	UCV	ICV	EX1		EX2	
			fixed $p = 0.3$	opt. $p = 0.1$	fixed $p = 0.2$	opt. $p = 0.1$
100	2.4670	1.7218	1.6478	1.4171	1.7018	1.4434
n	UCV	ICV	EX1		EX2	
			fixed $p = 0.3$	opt. $p = 0.1$	fixed $p = 0.2$	opt. $p = 0.1$
250	1.9159	1.4757	1.4637	1.3062	1.4186	1.3238
Bimodal normal mixture						
n	UCV	ICV	EX1		EX2	
			fixed $p = 0.3$	opt. $p = 0.3$	fixed $p = 0.2$	opt. $p = 0.3$
100	1.6995	1.3614	1.3667	1.3667	1.3827	1.3640
n	UCV	ICV	EX1		EX2	
			fixed $p = 0.3$	opt. $p = 0.3$	fixed $p = 0.2$	opt. $p = 0.2$
250	1.5160	1.2874	1.2453	1.2453	1.2331	1.2331

results than nested cross-validation. The performance of Marron's formula is similar to our proposed nested cross-validation in the context of first-order extrapolation at sample size $n = 100$, but MCV shows a small systematic superiority over NCV at $n = 200$. However, the realization of Marron's formula involves complex formulas as defined in Remark 3. Overall, we could obtain over 80% efficiency in all cases. The fixed-p second-order extrapolation is almost a clear winner over the fixed-p first-order extrapolation. Moreover, the nested cross-validation using second-order extrapolation gives very similar performance as the fixed-p second order extrapolation. Both seem promising bandwidth selection tools.

In conclusion, we note that fixed-p first-order extrapolation is a computationally inexpensive way to improve upon standard cross-validation when $m^*(h)$ is reasonably behaved. For additional refinements, one can apply the second-order extrapolation.

4.2. Comparison to indirect cross-validation

Savchuk et al. (2011, 2010) and Mammen et al. (2012) discuss a modification of the unbiased cross-validation (UCV) method, called *indirect cross-validation* (ICV), that aims to reduce the large variability of UCV bandwidth selector with the help of a selection kernel.

Given a selection kernel of the following form

$$L(u; \alpha, \sigma) = (1 + \alpha)\phi(u) - \frac{\alpha}{\sigma}\phi(u/\sigma),$$

indirect cross-validation is to first find the UCV bandwidth selector based on L, say $\hat{h}_L$, and then rescale it to a bandwidth using Gaussian kernel ϕ. We denote the latter bandwidth selector as $\hat{h}_\phi$. They argue that the conventional cross-validation works more stably with a more complicated kernel function than a second-order Gaussian kernel in bandwidth selection. The relationship between $\hat{h}_L$ and $\hat{h}_\phi$ is approximated based on expressions of their asymptotic optimal bandwidth, which can be written as

$$\hat{h}_\phi = \left(\frac{R(\phi)\mu_{2L}^2}{R(L)\mu_{2\phi}^2}\right)^{1/5} \hat{h}_L,$$

where the rescaler only depends on the selection kernel.

Following a referee's suggestion, we will show below a numerical comparison between our proposed subsampling-extrapolation bandwidth selectors and the ICV bandwidth selector. To be comparable to the results in Savchuk et al. (2011, 2010), we consider the same simulation setting and choices of distribution functions: we consider $R = 1000$ independent samples of size n ($n = 100$ and 250) generated randomly from standard normal or a bimodal normal mixture $0.5N(-1, 4/9) + 0.5(1, 4/9)$. Notice that their bimodal normal mixture is essentially the same as our Mixture 1 shown in Table 1. In addition, we compute $\hat{E}\{ISE(\hat{h})/ISE(\hat{h}_0)\}$ as the measure to evaluate the performance of each method. This measure is used in Savchuk et al. (2011, 2010) to evaluate the performance of the ICV bandwidth selector. It can be seen in Table 5 that our proposed bandwidth selectors perform comparably, or even better, than the indirect cross-validation bandwidth selector.

Remark 4. Although indirect cross-validation bears similarity with our subsampling-extrapolation method, we use extrapolation techniques differently. Their method is to shift the optimal risk curve based on a selection kernel, say m_L^*, to

one that is based on a second-order Gaussian kernel, say m^*_ϕ, at sample size n. In comparison, our method is to move along the same m^*_ϕ curve from a subsample size m to the original sample size n. It seems possible to implement extrapolation techniques in indirect cross-validation. That is, one can implement subsampling-extrapolation method on the risk curve based on a selection kernel, then shift the selected bandwidth to one that is based on a Gaussian kernel. However, the investigation of this possibility is beyond the scope of this paper.

5. Asymptotic properties

5.1. First-order extrapolation

The asymptotic properties of the first-order extrapolated bandwidth selector $\hat{h}_1$ is easy to obtain, since it is a simple multiple of the subsampling bandwidth selector. When $p = m/n$ is considered as a fixed constant, Hall and Robinson (2009) show that the relative convergence rate of $\hat{h}_1$ is of order $n^{-1/10}$. Although this is the same rate as for the UCV bandwidth selector, the asymptotic variance in $\hat{h}_1$ is reduced by using m less than n. More specifically, the asymptotic variance can be written as $(m/n)^{4/5}$ $(\mathbb{C}_1 + (m/n)\mathbb{C}_2)$, where both $\mathbb{C}_1$ and $\mathbb{C}_2$ are constants. The asymptotic variance can be reduced by a factor of $(m/n)^{4/5}$ to say the least. If $\mathbb{C}_1$ is much smaller than $\mathbb{C}_2$, the reduction could be of a factor of $(m/n)^{9/5}$. When one considers half-sampling, the reduction in the asymptotic variance is around 50%.

Marron (1987) considers the case when p is dependent on sample size n. Marron (1987) shows that the optimal choice of p that minimizes AMISE is of order $n^{-3/8}$ (see Remark 3). With this optimal value of p the relative rate of convergence of $\hat{h}_1$ can achieve $n^{-1/4}$, as shown in Eq. (3.3) in Marron (1987). We have verified that the result in Marron (1987) agrees with Hall and Robinson (2009) when p is considered constant.

5.2. Second-order extrapolation

We have noticed from numerical results that the second-order extrapolated bandwidth selector does improve over the first-order extrapolation method. Here we want to investigate whether $\hat{h}_2$ has better asymptotic properties than $\hat{h}_1$.

Denote $g(h) = \log(m/n) - 5(\log h - \log \hat{h}_{L^2,m}) + a(h^2 - \hat{h}^2_{CV,m})$, where $\hat{h}_2$ is the solution of $g(h) = 0$. Using Taylor series to expand $g(h)$ around $\hat{h}_1$, we have

$$\hat{h}_2 \approx \hat{h}_1 - \frac{a}{5} p^{-2/5}(1 - p^{2/5})\hat{h}_1^3.$$

Because $a > 0$ and $0 < p < 1$, $\hat{h}_2$ reduces the positive bias inherited in $\hat{h}_1$.

When p is a fixed constant, it is easy to see that the relative convergence rate of $\hat{h}_2$ is the same as $\hat{h}_1$. When p is dependent on n, one can follow the proof in Marron (1987) and show that the optimal value of p, as a function of n, that minimizes the asymptotic mean square error of $\hat{h}_2$ is of order $n^{-3/8}$. With this choice of p, the relative convergence rate of $\hat{h}_2$ can achieve $n^{-1/4}$. (Please see Appendix for more details.)

In short, the relative rate of convergence for the second-order extrapolated bandwidth selector is the same as the first-order extrapolation method in both cases that p is a fixed constant and p is dependent on n. However, we have illustrated that it can improve upon the first-order extrapolation methodology and reduce the positive bias in finite sample scenarios.

6. Discussion and future work

This paper has been focusing on the discussion of one particular scenario, risk estimation and bandwidth selection in a kernel density estimator. However, the proposed subsampling-extrapolation procedures can be easily generalized to other important problems in which one seeks an optimal smoothing parameter as long as some basic asymptotic results about rates of convergence are known. In addition, the subsampling-extrapolation technique could be in theory extended to other interesting applications, such as variance estimation and quantile estimation; we will study these applications in another paper. In particular, when considering the estimation of a U-statistic variance, the application of extrapolation techniques can enable one to relax the restriction of the kernel size needed in the unbiased variance estimator of a general U-statistic devised in Wang and Lindsay (2014).

From another aspect, the two-stage bandwidth selection procedure can be easily applied to cases where the risk is evaluated based on the Kullback–Leibler distance. We have preliminary simulation result showing that similar conclusions hold when one changes the loss function from L^2 to Kullback–Leibler distance. The asymptotic properties of Kullback–Leibler risk (also called likelihood cross-validation) are quite complex, as discussed in Hall (1987). van Es (1991) found the rate of convergence of a likelihood cross-validation bandwidth selector for a bounded kernel function K over the unit interval, a more general result for any arbitrary kernel does not seem present.

There are a number of possible ways to tune our methods. For example, we used $n^* = n/2$ for nested cross-validation without further inspection. We did not try to tune the risk estimation for p, even though the selection mechanism showed

some significant bias at $n = 100$. Notice from Tables 2 and 3 that at sample size $n = 100$, nested cross-validation tends to over-select small values of p. One possible solution to avoid selecting small p is to implement the "1-SE rule" (Breiman et al., 1984) in nested cross-validation based on a reasonable estimation of the standard error of the risk estimator. According to the 1-SE rule, one would choose the largest p value that leads to a risk estimate $\hat{R}_{L^2}(p, n^*)$ no more than one standard error above the minimum risk score. We have investigated the application of jackknife variance estimator in the 1-SE rule. However, due to the large positive bias of jackknife variance estimator, there was no gain in applying the 1-SE rule in this scenario. In addition, as seen in Fig. 5 many of the empirical risk curves are fairly flat, indicating that choosing a smaller p value in the stage of nested cross-validation may not have a large effect on the final bandwidth selector.

Acknowledgments

This research was partially supported by the National Science Foundation (DMS 0714839). We want to thank the three anonymous reviewers for their valuable comments that lead to a much improved version of the manuscript.

Appendix A. Proof for Lemma 1

Consider an arbitrary kernel function K of order $r \geq 2$. Recall that the relative L^2 risk, denoted as $\mathrm{R}_{L^2,n} = E_{\mathcal{X}_n}\left\{\int \hat{f}_h(x)^2 dx\right\} - 2E_{\mathcal{X}_n}\left\{\int f(x)\hat{f}_h(x)dx\right\}$.

Notice that the first expectation can be written as

$$E_{\mathcal{X}_n}\left\{\int \hat{f}_h(x)^2 dx\right\} = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left\{\int K(u)K\left(u - \frac{X_i - X_j}{h}\right)du\right\}.$$

Denote $\bar{K}(v) = (K * K)(v) = \int K(u)K(u - v)du$, and $\bar{K}_h(v) = \frac{1}{h}\bar{K}\left(\frac{v}{h}\right)$. We then have

$$\begin{aligned} E_{\mathcal{X}_n}\left\{\int \hat{f}_h(x)^2 dx\right\} &= \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left\{\bar{K}\left(\frac{X_i - X_j}{h}\right)\right\} \\ &= \frac{\mathbb{C}}{nh} + \frac{n-1}{n} E_{\mathcal{X}_2}\{\bar{K}_h(X_1, X_2)\}, \end{aligned}$$

where $\mathbb{C} = \bar{K}(0) = \int K(u)^2 du$ is a non-zero constant that is independent of h.

Because $E_{\mathcal{X}_n}\left\{\int f(x)\hat{f}_h(x)dx\right\} = E_{\mathcal{X}_2}\{K_h(X_1, X_2)\}$, we then have

$$\mathrm{R}_{L^2,n}(h) = \left[E\{\bar{K}_h(X_1, X_2)\} - 2E\{K_h(X_1, X_2)\}\right] + \frac{1}{n}\left[\mathbb{C}/h - E\{\bar{K}_h(X_1, X_2)\}\right].$$

If we want to evaluate the risk at a fictional size m which may be smaller than the original sample size n, the closed form expression of $m^*(h)$ for which h would be optimal is

$$m^*(h) = -\frac{\frac{d}{dh}\left[\mathbb{C}/h - E\{\bar{K}_h(X_1, X_2)\}\right]}{\frac{d}{dh}\left[E\{\bar{K}_h(X_1, X_2)\} - 2E\{K_h(X_1, X_2)\}\right]}.$$

We claim that:

1. The order of the new kernel $\bar{K} = K * K$ is the same as the order for K.
2. Denote $\mu_r(L) = \int u^r L(u)du$ $(r \geq 2)$ for a function L. We have:

$$\begin{cases} \text{If } 0 \leq j < r, & \text{then } \mu_j(\bar{K}) = 0. \\ \text{If } r \leq j < 2r, & \text{then } \mu_j(\bar{K}) = 2\mu_j(K). \\ \text{If } j \geq 2r, & \text{then } \mu_j(\bar{K}) > 2\mu_j(K). \end{cases}$$

Proof for Claim 1:

Notice that

$$\int x^j \bar{K}(x)dx = \int K(u)\left\{\int (u + y)^j K(y)dy\right\}du \quad (j \in \mathcal{N}).$$

The inner integral is non-zero only if $\int y^j K(y)dy \neq 0$. If K is of order r, then $\bar{K}$ must be of order r as well.

Proof for Claim 2:

$$\begin{aligned}\mu_j(\bar{K}) &= \int K(u)\left\{\int (u+y)^j K(y)dy\right\}du = \int K(u)\left\{\sum_{i=0}^{j}\binom{k}{i}u^i\int y^{j-i}K(y)dy\right\}du\\ &= \sum_{i=0}^{j}\binom{j}{i}\left\{\int u^{j-i}K(u)du\right\}\left\{\int y^i K(y)dy\right\}\\ &= \sum_{i=0}^{j}\binom{j}{i}\mu_{j-i}(K)\mu_i(K).\end{aligned}$$

Because $\mu_{j-i}(K)\neq 0$ if and only if $j-i\geq r$, and $\mu_i(K)\neq 0$ if and only if $i\geq r$, we have

- If $0\leq j<r$, then $\mu_{j-i}(K)=0$ and $\mu_i(K)=0$ for all $0\leq i\leq j$.
- If $r\leq j<2r$, then $\mu_{j-i}(K)\neq 0$ and $\mu_i(K)\neq 0$ only if $i=0$ or $i=j$. Therefore, $\mu_j(\bar{K})=2\mu_j(K)$.
- If $j\geq 2r$, then $\mu_j(\bar{K})\geq 2\mu_j(K)+\binom{j}{k}\mu_{j-k}(K)\mu_k(K)$. In particular, when $j=2r$, we have

$$\mu_{2r}(\bar{K}) = 2\mu_r(K)+\binom{2r}{r}\{\mu_r(K)\}^2.$$

Notice that $E\{K_h(X_1,X_2)\}=\int (K_h*f)(x)f(x)dx$, where

$$\begin{aligned}(K_h*f)(x) &= \int h^{-1}K\left(\frac{u-x}{h}\right)f(u)du = \int K(z)f(hz+x)dz\\ &= \int K(z)\left\{f(x)+f'(x)hz+\cdots+f^{(r)}\frac{(hz)^r}{r!}+o(h^r)\right\}dz\\ &= f(x)+\frac{h^r\mu_r(K)}{r!}f^{(r)}(x)+\cdots+\frac{h^{2r-1}\mu_{2r-1}(K)}{(2r-1)!}f^{(2r-1)}(x)+o(h^r).\end{aligned}$$

Therefore,

$$E\{K_h(X_1,X_2)\} = \int f^2(x)dx+\frac{h^r\mu_r(K)}{r!}E\{f^{(r)}\}+\cdots+\frac{h^{2r-1}\mu_{2r-1}(K)}{(2r-1)!}E\{f^{(2r-1)}\}+O(h^{2r}),$$

where K is of order r and is symmetric.

In addition, based on the results in Claims 1 and 2 we also have

$$\begin{aligned}E\{\bar{K}_h(X_1,X_2)\} &= \int f^2(x)dx+\frac{h^r\mu_r(\bar{K})}{r!}E(f^{(r)})+\cdots+\frac{h^{2r-1}\mu_{2r-1}(\bar{K})}{(2r-1)!}E(f^{(2r-1)})+O(h^{2r})\\ &= \int f^2(x)dx+2\left\{\frac{h^r\mu_r(K)}{r!}E(f^{(r)})+\cdots+\frac{h^{2r-1}\mu_{2r-1}(K)}{(2r-1)!}E(f^{(2r-1)})\right\}+O(h^{2r}).\end{aligned}$$

We have

$$\begin{aligned}E\{A_h(X_1,X_2)\} &= E\{\bar{K}_h(X_1,X_2)\}-2E\{K_h(X_1,X_2)\} = -\int f^2(x)dx+O(h^{2r}),\\ E\{B_h(X_1,X_2)\} &= \mathbb{C}/h-E\{\bar{K}_h(X_1,X_2)\}\\ &= \mathbb{C}/h-\int f^2(x)dx-2E(f^{(r)})\left(h^r\mu_r(\bar{K})\right)/r!+O(h^{r+2}).\end{aligned}$$

Thus, $m^*(h)=\{\mathbb{C}/h^{2r+1}+(2/h^r(r-1)!)\,\mu_r(K)E(f^{(r)})+O(h^{-r+2})\}/O(1)$, which implies that $h^{2r+1}m^*(h)=\{\mathbb{C}+(2/(r-1)!)h^{r+1}\mu_r(K)E(f^{(r)})+O(h^{r+3})\}/O(1)\to\mathbb{C}$ as $h\to 0$ ($\mathbb{C}\neq 0$). In other words,

$$\log m^*(h)+(2r+1)\log h\to\text{constant},\quad \text{as } h\to 0.$$

The other statement can be verified easily based on the above result.

Appendix B. The derivation of second-order extrapolation

The explicit expression of $m^*(h)$ for a h that yields the optimal MISE is

$$m^*(h) = -\frac{d}{dh}E\{B_h(X_1, X_2)\} \Big/ \frac{d}{dh}E\{A_h(X_1, X_2)\},$$

where

$$A_h(x_1, x_2) = K_{\sqrt{2}h}(x_1, x_2) - 2K_h(x_1, x_2),$$
$$B_h(x_1, x_2) = 1/(2h\sqrt{\pi}) - K_{\sqrt{2}h}(x_1, x_2).$$

Without loss of generality, assuming a Gaussian kernel function i.e. $K_h(x_1, x_2) = (1/h)\phi((x_1 - x_2)/h)$, we have

$$\begin{aligned} E\{K_{\sqrt{2}h}(x_1, x_2)\} &= \iint K_{\sqrt{2}h}(x_1, x_2)dF(x_1)dF(x_2) \\ &= \iint \left\{\int K_h(x_1, x)K_{h^2}(x_2, x)dx\right\} dF(x_1)dF(x_2) \\ &= \int \left\{\int K_h(x_1, x)f(x_1)dx_1\right\} \left\{\int K_h(x_2, x)f(x_2)dx_2\right\} dx. \end{aligned}$$

Notice that

$$\begin{aligned} \int K_h(x_1, x)f(x_1)dx_1 &= \int \phi(t)f(x + th)dt \\ &= \int \phi(t)\left\{f(x) + thf'(x) + \cdots + \frac{h^4}{4!}f^{(4)}(t) + o(h^4)\right\} dt \\ &= f(x) + \frac{h^2}{2!}f^{(2)}(x)\mu_2(\phi) + \frac{h^4}{4!}f^{(4)}x\mu_4(\phi) + o(h^4) \end{aligned}$$

where $\mu_k(\phi) = \int t^k\phi(t)dt$ for $k \geq 2$ and $\phi(t)$ is a Gaussian kernel that is symmetric and of order 2.

Therefore,

$$\begin{aligned} E\{K_{\sqrt{2}h}(X_1, X_2)\} &= \int \left\{f(x) + \frac{h^2}{2!}f^{(2)}(x)\mu_2(\phi) + \frac{h^4}{4!}f^{(4)}x\mu_4(\phi) + o(h^4)\right\}^2 dx \\ &= \int f(x)^2dx + h^2\mu_2(\phi)C_{02} + \frac{h^4}{12}\mu_4(\phi)C_{04} + \frac{h^4}{4}\{\mu_2(\phi)\}^2C_{22} + \frac{h^6}{24}\mu_2(\phi)\mu_4(\phi)C_{24} + o(h^6) \end{aligned}$$

where $C_{ij} = \int f^{(i)}(x)f^{(j)}(x)dx$ for $i, j \in \mathbb{Z}$, assuming $f^{(0)}(x) = f(x)$.

Similarly, we have

$$\begin{aligned} E\{K_h(X_1, X_2)\} &= \int f(x)^2dx + \frac{h^2}{2}\mu_2(\phi)E(f^{(2)}) + \frac{h^4}{48}\mu_4(\phi)C_{04} + \frac{h^4}{16}\{\mu_2(\phi)\}^2C_{22} \\ &\quad + \frac{h^6}{192}\mu_2(\phi)\mu_4(\phi)C_{24} + o(h^6). \end{aligned}$$

As a result,

$$E\{A_h(X_1, X_2)\} = -\int f(x)^2dx + h^4\left[\frac{C_{04}\mu_4(\phi)}{24} + \frac{C_{22}\{\mu_2(\phi)\}^2}{8}\right] + \frac{3h^6}{96}\mu_2(\phi)\mu_4(\phi)C_{24} + o(h^6).$$

We denote it as $E\{A_h(X_1, X_2)\} = -\int f(x)^2dx + \mathbb{C}_1h^4 + \mathbb{C}_2h^6 + o(h^6)$. And,

$$E\{B_h(X_1, X_2)\} = 1/(2h\sqrt{\pi}) - \int f(x)^2dx - h^2\mu_2(\phi)E(f^{(2)}) + o(h^2).$$

Thus, we have

$$m^*(h) = \frac{(1/2\sqrt{\pi})h^{-2} + 2h\mu_2(\phi)C_{02} + o(h)}{4h^3\mathbb{C}_1 + 6h^5\mathbb{C}_2 + o(h^5)}.$$

Appendix C. Estimation of the expected integrated squared error as a function of p

Recall that the MISE of using a fixed bandwidth h is defined as

$$\begin{aligned}\mathrm{MISE}(h) &= E_{\mathcal{X}_n}\left[\int \{f(x)-\hat{f}_h(x)\}^2 dx\right] \\ &= \int f(x)^2 dx + E_{\mathcal{X}_n}\left[\int \hat{f}_h(x)^2 dx\right] - 2E_{\mathcal{X}_n}\left[\int f(x)\hat{f}_h(x)dx\right].\end{aligned}$$

Consider a random bandwidth selector based on two-stage bandwidth selection procedures, say $\hat{h}_p$. We want to consider MISE as a function of p (denoted as EISE(p)), where p is the proportion of the data in the subsampling stage, and then evaluate the mean integrated squared error of using proportion p in the two-stage bandwidth selection procedure.

Assume that the true underlying distribution is a normal mixture with k different normal components, i.e. $f \sim \sum_{j=1}^{k} \omega_j N(\mu_j, \sigma_j^2)$. We denote

$$f(x) = \sum_{j=1}^{k} \omega_j K_{\sigma_j}(x-\mu_j),$$

where $\sum_j \omega_j = 1$ and $K_{\sigma_j}(x-\mu_j) = \phi\left(\frac{x-\mu_j}{\sigma_j}\right)$.

The constant $\int f(x)^2 dx$ in the mean integrated squared error is easy to compute when the underlying distribution is a normal mixture. It can be shown that

$$\int f(x)^2 dx = \int \left(\sum_{j=1}^{k} \omega_j K_{\sigma_j}(x-\mu_j)\right)^2 dx = \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j K_{\sqrt{\sigma_i^2+\sigma_j^2}}(\mu_i-\mu_j).$$

When considering nonparametric kernel density estimator with Gaussian kernel function, the estimated density at x with extrapolated bandwidth $\hat{h}_p$ is

$$\hat{f}_{p,n^*}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\hat{h}_p}(x-X_i).$$

Thus, given a choice of p and a sample of size n i.e. $\mathcal{X}_n = (X_1, \ldots, X_n)$, we have

$$\int \hat{f}_{p,n^*}(x)^2 dx = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{\sqrt{2}\hat{h}_p}(x_i-x_j)$$

$$\int f(x)\hat{f}_{p,n^*}(x)dx = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \omega_j K_{\sqrt{\sigma_j^2+\hat{h}_p^2}}(x_i-\mu_j).$$

Therefore, the EISE, as a function of p, can be estimated by simulation based on R random samples of size n in the following way:

$$\widehat{EISE}(p) = \frac{1}{R}\sum_{\mathcal{X}_n}\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{\sqrt{2}\hat{h}_p}(x_i-x_j) - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\omega_j K_{\sqrt{\sigma_j^2+h^2}}(x_i-\mu_j)\right\}.$$

Appendix D. Relative rate of convergence of $\hat{h}_2$ with optimal p

Because $\hat{h}_{CV,m} \sim (np)^{-1/5}$, we have $\hat{h}_2 \sim p^{1/5}\hat{h}_{CV,m} - (a_0/5)n^{-3/5}p^{-2/5} + o(n^{-3/5}p^{-2/5})$ where a_0 is a constant. From Theorems 1 and 2 in Marron (1987), we have

$$\begin{aligned}\mathrm{bias}(\hat{h}_2) &= \mathbb{C}n^{-3/5}p^{2/5} + o(n^{-3/5}p^{-2/5}) \\ \mathrm{Var}(\hat{h}_2) &= \sigma^2 n^{-3/5}p^{4/5}\end{aligned}$$

where $\mathbb{C}$ and σ^2 are constants, independent of sample size n. Thus, the asymptotic mean square error of $\hat{h}_2$ is

$$\mathrm{AMSE}(\hat{h}_2) = \sigma^2 n^{-3/5}p^{4/5} + (\mathbb{C}n^{-3/5}p^{-2/5})^2.$$

The choice of p that minimizes AMSE is $p_{\mathrm{opt}} = (\mathbb{C}/\sigma^2)^{5/8}n^{-3/8} = O(n^{-3/8})$. With this choice of p, it is easy to see that the relative rate of convergence of $\hat{h}_2$ is of order $n^{-1/4}$.

References

Aldershof, B., Marron, J.S., Park, B.U., Wand, M.P., 1995. Facts about the Gaussian probability density function. Appl. Anal. 59 (1), 289–306.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Monterey, CA.
Fix, E., Hodges, J.L., 1951. Discriminatory analysis, nonparametric discrimination: consistency properties, Report Number 4, Project Number 21-49-004. USAF School of Aviation Medicine, Randolph Field, Texas.
Fraser, D.A.S., 1954. Completeness of order statistics. Canad. J. Math. 6, 42–45.
Hall, P., 1987. On Kullback–Leibler loss and density estimation. Ann. Statist. 15 (4), 1491–1519.
Hall, P., Robinson, A.P., 2009. Reducing variability of crossvalidation for smoothing-parameter choice. Biometrika 96 (1), 175–186.
Hardle, W.K., Muller, M., Sperlich, S., Werwatz, A., 1994. Nonparametric and Semiparametric Models. Springer.
Jones, M.C., 1991. The roles of ISE and MISE in density estimation. Statist. Probab. Lett. 11, 511–514.
Jones, M.C., Sheather, S.J., 1991. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Statist. Probab. Lett. 11, 511–514.
Lindsay, B.G., Liu, J., 2009. Model assessment tools for a model false world. Inst. Math. Stat.-Stat. Sci. 24, 303–318.
Loader, C.R., 1999. Bandwidth selection: classical or plug-in? Ann. Statist. 27 (2), 415–438.
Mammen, E., Miranda, M.D.M., Nielsen, J.P., Sperlich, S., 2012. A comparative study of new cross-validated bandwidth selectors for kernel density estimation. http://arxiv.org/abs/1209.4495.
Marron, J.S., 1987. Partitioned cross-validation. Econometric Rev. 6, 271–283.
Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.
Meinshausen, N., Buhlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B 72, 417–473.
Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. J. Amer. Statist. Assoc. 85.
Ray, S., Lindsay, B.G., 2008. Model selection in high-dimensions: a quadratic-risk based approach. J. R. Stat. Soc. Ser. B 70, 95–118.
Savchuk, O.Y., Hart, J.D., Sheather, S.J., 2011. An empirical study of indirect cross-validation. In: Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger. World Scientific Publishing, Singapore, pp. 288–308.
Savchuk, O.Y., Hart, J.D., Sheather, S.J., 2010. Indirect cross-validation for density estimation. J. Amer. Statist. Assoc. 105 (489), 415–423.
Shah, R.D., Samworth, R.J., 2012. Variable selection with error control: another look at stability selection. J. R. Stat. Soc. Ser. B 74, 1–26.
Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88 (422), 486–494.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B 53, 683–690.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. In: Monographs on Statistics and Applied Probability, Chapman & Hall, London.
Turlach, B.A., 1993. Bandwidth selection in kernel density estimation: a review. Discussion Paper 9307, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.
Van Es, B., 1991. Likelihood cross-validation bandwidth selection for nonparametric kernel density estimators. Nonparametr. Stat. 1, 83–110.
Van Es, B., 1992. Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. Ann. Statist. 20 (3), 1647–1657.
Wang, Q., Lindsay, B.G., 2014. Variance estimation of a general U-statistic with application to cross-validation. Statist. Sinica 24 (3), 1117–1141.