University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 125

Asymptotic Optimality of Likelihood Based Cross-Validation

Mark J. van der Laan[*] Sandrine Dudoit[†]
Sunduz Keles[‡]

[*]Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley

[‡]Division of Biostatistics, School of Public Health, University of California, Berkeley

Asymptotic Optimality of Likelihood Based Cross-Validation

Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles

Abstract

Likelihood-based cross-validation is a statistical tool for selecting a density estimate based on n i.i.d. observations from the true density among a collection of candidate density estimators. General examples are the selection of a model indexing a maximum likelihood estimator, and the selection of a bandwidth indexing a nonparametric (e.g. kernel) density estimator. In this article, we establish asymptotic optimality of a general class of likelihood based cross-validation procedures (as indexed by the type of sample splitting used, e.g. V-fold cross-validation), in the sense that the cross-validation selector performs asymptotically as well (w.r.t. to the Kullback-Leibler distance to the true density) as an optimal benchmark model selector which depends on the true density. Crucial conditions of our theorem are that the size of the validation sample converges to infinity, which excludes leave-one-out cross-validation, and that the candidate density estimates are bounded away from zero and infinity. We illustrate these asymptotic results and the practical performance of likelihood based cross-validation for the purpose of bandwidth selection with a simulation study.

1 Introduction

Density estimation arises in important and common problems in the statistical literature. As discussed below, bandwidth selection in kernel density estimation, selecting the number of components in mixture models, and variable selection in regression (e.g., logistic and linear regression with normal error), are three examples of problems that involve explicitly or implicitly some form of density estimation.

Let $X_1, \ldots, X_n$ be n independent and identically distributed (i.i.d.) random variables with distribution P and corresponding density f with respect to a dominating measure μ. Let $f_k(\cdot \mid P_n)$ be an estimator of f, $k = 1, \ldots, K(n)$, where P_n denotes the empirical distribution function. For example, $f_k(\cdot \mid P_n)$ can be the maximum likelihood estimator of f according to a model $\mathcal{M}_k$, that is,

$$f_k(\cdot \mid P_n) = \max{}^{-1}_{f \in \mathcal{M}_k} \int \log(f(x)) dP_n(x).$$

A fundamental and practical problem is the selection of a $\hat{k}$ in such a manner that $f_{\hat{k}}(\cdot \mid P_n)$ converges to the true density f optimally. For mixture modeling, $\mathcal{M}_k$ could denote the mixture model with k components. In the case of variable selection in regression, $\mathcal{M}_k$ could be a model for the conditional density of a continuous (regression with normal error) or discrete (multinomial regression) outcome Y, given a set of covariates Z, corresponding with a regression model $\mu_k(Z)$ for the conditional mean $E(Y \mid Z)$. Here k could index a particular set of variables in the regression model. Alternatively, in the regression context, k could index a forward selection algorithm $f_k(\cdot \mid P_n)$ applied to the empirical distribution P_n which stops after having selected k variables. For kernel density estimation, the parameter k could correspond to the bandwidth of the kernel density estimator.

Implicit in this selection problem is the notion of distance between two distributions. Here, we focus on the Kullback-Leibler divergence as a measure of distance between two densities. The Kullback-Leibler divergence between densities f and g is defined as

$$DKL(f, g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x) d\mu(x)$$

and has the following two basic properties: $DKL(f, g) \geq 0$ and $DKL(f, g) = 0$ if and only if $f = g$ a.s. Ideally, given P_n, one seeks $f_k(\cdot \mid P_n)$ that is closest to the true f. With the Kullback-Leibler criterion, one would choose

$$\tilde{k}_n \equiv \min^{-1}_{k\in\{1,\ldots,K(n)\}} DKL(f, f_k) = \min^{-1}_{k\in\{1,\ldots,K(n)\}} - \int \log(f_k(x \mid P_n))dP(x). \tag{1}$$

However, P is unknown. One could envisage using the empirical distribution, P_n, in place of the true P but this could lead to over-fitting. Instead, we turn to cross-validation. In this setting, the learning sample $X_1, \ldots, X_n$ is split (repeatedly) at random into two sets, a training set and a validation set. A density f_k is estimated for each $k \in \{1, \ldots, K(n)\}$ using the training set only and the empirical distribution for the validation set is used in place of the true P in the distance criterion.

Leave-one-out likelihood cross-validation in density estimation is discussed in Silverman (1986) who refers to Stone (1974a) and Geisser (1975) for its general applicability to model fitting as well. Silverman (1986) refers to Scott & Factor (1981) to indicate that for densities with infinite support this leave-one-out likelihood cross-validation method for bandwidth selection in density estimation is sensitive to outliers, and to Schuster & Gregory (1981) to point out that leave-one-out cross-validation can result, in fact, into inconsistent density estimators under non-pathological conditions. Stone (1984) provides an asymptotically optimal bandwidth selection rule for kernel density estimation, which has a leave-one out cross-validation interpretation.

Recent work on (V-fold or Monte-Carlo) cross-validated likelihood methods for choosing the number of components in mixture models is found in Smyth (2000) and Pavlic & van der Laan (2003). In particular, the simulation studies of Pavlic & van der Laan (2003) showed that likelihood based cross-validation performed well compared to common approaches based on validity functionals such as Akaike's information criterion (Akaike (1973), Bozdogan (2000)), Bayesian Information criterion BIC (Schwartz (1978)) or Minimum description length (Rissanen (1978), see Hansen & Yu (2001), for an overview) and ICOMP (Bozdogan (1993)).

Likelihood based cross-validation covers in particular squared error-loss cross-validation for prediction. Specifically, let $\mathcal{M}_k$ be a regression model $Y = \mu_k(Z) + N(0, \sigma^2)$, with μ_k ranging over a family of curves indexed by k, and let $f_k(X \mid P_n)$ be the corresponding (least squares estimator (i.e., maximum likelihood estimator) . There is a rich literature on leave-one-out cross-validation in nonparametric univariate regression. For example, Silverman (1984) proposes a fast approximation of the leave-one out cross-validation method in spline regression. We refer to Härdle (1993) for an overview on the leave-one-out cross-validation method in kernel regression. In particular, Härdle & Marron (1985a) and Härdle & Marron (1985b) prove that leave-one-out cross-validation is asymptotically optimal for choosing the smoothing parameter in nonparametric kernel regression (see page 158, Härdle (1993)). In the general prediction literature involving covariate and model selection cross-validation is commonly used for estimation of the risk for squared error loss (e.g., Breiman et al. (1984), Breiman (1996), Burman (1989), Shao (1993), Zhang (1993)), Hastie et al. (2001), Ripley (1996), Stone (1974b), Stone (1977)). The main procedures include: leave-one-out cross-validation, V-fold cross-validation (i.e., random division of the learning set into V mutually exclusive and exhaustive sets), Monte Carlo cross-validation (i.e., repeated random splits of the learning set into a training and a validation set), and the bootstrap. Györfi et al. (2002) recently proved that for bounded outcomes, the single-split cross-validation for the squared error loss function is asymptotically optimal in selecting predictors based on the training sample in the same sense as in our Theorem 1 below.

This article considers general likelihood based cross-validation procedures and establishes a similar result to that of Györfi et al. (2002). Theorem 1 and its Corollary show that, under general conditions on P, the cross-validation selector for k is asymptotically optimal, in the sense that it performs as well as a benchmark selector based on the true underlying distribution P. We illustrate this asymptotic result and the practical performance of likelihood based cross-validation for the purpose of bandwidth selection with a simulation study.

2 Method and Results

2.1 Framework.

To formalize the cross-validated likelihood method, we introduce a binary random vector $S_n \in \{0,1\}^n$, independent of P_n. A realization of S_n defines a particular split of the sample of n observations into a training sample $\{i \in \{1,\ldots,n\} : S_n(i) = 0\}$ and a validation sample $\{i \in \{1,\ldots,n\} : S_n(i) = 1\}$. Let P^1_{n,S_n}, P^0_{n,S_n} be the empirical distributions of the validation and training samples, respectively. Let the proportion $p = \sum_{i=1}^n S_n(i)/n \in (0,1)$ of observations in the validation sample be constant (but possibly depend on

n). We define the cross-validated likelihood criterion as:

$$\hat{\theta}_{n(1-p)}(k) = -E_{S_n} \int \log \left(f_k(x \mid P^0_{n,S_n}) \right) dP^1_{n,S_n}(x).$$

This criterion defines an optimal choice $\hat{k}$ given by

$$\hat{k} = \min{}^{-1}_{k \in \{1,\ldots,K(n)\}} \hat{\theta}_{n(1-p)}(k).$$

We note that different choices of the random variable S_n cover many types of cross-validation such as V-fold cross-validation, Monte-Carlo (repeated random splits) cross-validation, and resampling (bootstrap) cross-validation. The latter corresponds with resampling n observations with replacement from the original data set and setting $S_{n,i}$ equal to the number of times the observation i is sampled. In this case, P^0_{n,S_n}, P^1_{n,s_n} denote the empirical distributions of the resampled observations, and the excluded observations, respectively. Our proof of Theorem 1 below straightforwardly generalizes to random p, and therefore our results apply to bootstrap cross-validation as well.

To obtain a benchmark for the selected $\hat{k}$ we also define

$$\tilde{\theta}_{n(1-p)}(k) = -E_{S_n} \int \log \left(f_k(x \mid P^0_{n,S_n}) \right) dP(x)$$

and its minimizer

$$\tilde{k} = \min{}^{-1}_{k \in \{1,\ldots,K(n)\}} \tilde{\theta}_{n(1-p)}(k).$$

Note that $\tilde{k}$ corresponds to an optimal selector since it indexes the minimizer over k of the expectation over S_n of the Kullback-Leibler distance of the density estimator $f_k(\cdot \mid P^0_{n,S_n})$ based on the training sample to the true distribution P:

$$k \rightarrow E_{S_n} \int \log \left(\frac{f(x)}{f_k(x \mid P^0_{n,S_n})} \right) dP(x).$$

If necessary, we will also refer to $\tilde{k}$ as $\tilde{k}_{n(1-p)}$ to distinguish it from the minimizer $\tilde{k}_n$ for the whole sample of n observations, as defined in (1), of

$$k \rightarrow \tilde{\theta}_n(k) = -\int \log \left(f_k(x \mid P_n) \right) dP(x).$$

The theorem below shows that asymptotically, the cross-validation selector $\hat{k}$ performs as well as the optimal benchmark selector $\tilde{k}_{n(1-p)}$ in the sense that the ratio $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}_{n(1-p)}) - \theta_{opt})$ of mean conditional Kullback-Leibler distances converges to 1. The theorem implies also an interesting result for the Hellinger distance between $f_{\hat{k}}(\cdot \mid P_{n(1-p)})$ and the true density f since the Kullback-Leibler distance bounds, in particular, the Hellinger distance (see e.g. van der Vaart (1998), page 62):

$$\int (\sqrt{f}(x) - \sqrt{g}(x))^2 d\mu(x) \leq \int \log\left(\frac{f(x)}{g(x)}\right) f(x) d\mu(x).$$

Finally, we define the minimum of $g \to -\int \log(g(x))dP(x)$ among all densities g:

$$\theta_{opt} = -\int \log(f(x))dP(x).$$

Note that $\tilde{\theta}_{n(1-p)}(\hat{k}) \geq \tilde{\theta}_{n(1-p)}(\tilde{k}) \geq \theta_{opt}$.

Before we state the theorem we will present two regression examples. We refer to our simulation study in section 3 for a detailed treatment of a bandwidth selection example in kernel density estimation.

Example 1 (Regression for continuous outcome) Suppose $X = (Y, Z)$, where Y is a continuous outcome and Z is a vector of covariates. Given a regression model $\mu_k(Z \mid \beta_k)$ for the conditional mean $E(Y \mid Z)$, let

$$\mathcal{M}_k = \left\{ f_k(Y; Z \mid \beta_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y - \mu_k(Z \mid \beta_k))^2}{2\sigma^2}\right) : \beta_k \right\}$$

be the Gaussian regression model for the conditional density of Y, given Z: that is, $Y = \mu_k(Z \mid \beta_k) + N(0, \sigma^2)$. Suppose $f_k(\cdot \mid P_n) = f_k(\cdot \mid \beta_k(P_n))$ is the maximum likelihood estimator according to this model, where $\beta_k(P_n)$ is the corresponding maximum likelihood estimator of β_k. Then $\beta_k(P_n) = \min_{\beta_k}^{-1} \sum_{i=1}^n (Y_i - \mu_k(Z_i \mid \beta_k))^2$ is the least squares estimator. In addition, we have that up till a multiplicative and additive constant

$$\hat{\theta}_{n(1-p)}(k) = E_{S_n} \int (y - \mu_k(z \mid \beta_k(P^0_{n,S_n})))^2 dP^1_{n,S_n}(y, z)$$

is the standard residual sum of squares of the predictor $\mu_k(z \mid \beta_k(P^0_{n,S_n}))$ based on the training sample over the validation sample, averaged across all S_n-specific sample splits. Consequently, $\hat{k} = \min_k{}^{-1}\hat{\theta}_{n(1-p)}(k)$ denotes the squared-error loss cross-validation selector. Finally, we note that up till a multiplicative and additive constant

$$\tilde{\theta}_{n(1-p)}(k) = E_{S_n} \int (y - \mu_k(z \mid \beta_k(P^0_{n,S_n})))^2 dP(y, z)$$

is the average over S_n of the true conditional risk of the predictor $\mu_k(z \mid \beta_k(P^0_{n,S_n}))$ based on the S_n-specific training sample, so that $\tilde{k}$ indexes the predictor with minimal true conditional risk.

Example 2 (Logistic regression) Suppose $X = (Y, Z)$, where Y is a Bernoulli random variable, Z is a vector of covariates. Given a regression model $\mu_k(Z \mid \beta_k)$ for the conditional mean $E(Y \mid Z) = P(Y = 1 \mid Z)$, let

$$\mathcal{M}_k = \left\{ f_k(Y; Z \mid \beta_k) = \mu_k(Z \mid \beta_k)^Y \{1 - \mu_k(Z \mid \beta_k)\}^{1-Y} : \beta_k \right\}.$$

Suppose $f_k(\cdot \mid P_n) = f_k(\cdot \mid \beta_k(P_n))$ is the maximum likelihood estimator according to this model, where $\beta_k(P_n)$ is the corresponding maximum likelihood estimator of β_k. In this case $\hat{\theta}_{n(1-p)}(k)$ equals

$$E_{S_n} \int y \log\{\mu_k(z \mid \beta_k(P^0_{n,S_n}))\} + (1-y) \log\{1 - \mu_k(z \mid \beta_k(P^0_{n,S_n}))\} dP^1_{n,S_n}(y, z)$$

and $\tilde{\theta}_{n(1-p)}(k)$ equals

$$= E_{S_n} \int y \log\{\mu_k(z \mid \beta_k(P^0_{n,S_n}))\} + (1-y) \log\{1 - \mu_k(z \mid \beta_k(P^0_{n,S_n}))\} dP(y, z).$$

2.2 Finite sample result and asymptotic implications.

We will now present our main result.

Theorem 1 *Suppose that there exist $\epsilon > 0$ and $L < \infty$ so that $\epsilon < f_k(X \mid P_n) < L$ a.s. for all $k \in \{1, \ldots, K(n)\}$. Let $M_1 = 2\log(L/\epsilon)$ and $M_2 = 4L/\epsilon$.*

For any $\delta > 0$ we have

$$E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq (1+2\delta) \left\{ E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} \right\} + 2c(M_1, M_2, \delta) \frac{1 + \log(K(n))}{np},$$

where

$$c(M_1, M_2, \delta) = 2(1+\delta)^2 \left(\frac{M_1}{3} + \frac{M_2}{\delta} \right).$$

This finite sample result has the following asymptotic implications: If

$$\frac{\log(K(n))}{(np)\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \to 0 \text{ for } n \to \infty, \qquad (2)$$

then

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \rightarrow 1 \textit{ for } n \rightarrow \infty.$$

Similarly, if

$$\frac{\log(K(n))}{(np)\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \rightarrow 0 \textit{ in probability for } n \rightarrow \infty, \tag{3}$$

then

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \rightarrow 1 \textit{ in probability for } n \rightarrow \infty.$$

The final convergence in probability statement follows from the fact that, given a sequence of random variables $X_1, X_2, \ldots,$ $E \mid X_n \mid= O(g(n))$ for a positive function $g(n)$ implies $X_n = O_P(g(n))$, which itself is a directconsequence of Markov's inequality. We note that our conditions for the asymptotic optimality statements exclude leave-one out cross-validation, since it is required that the validation sample size np converges to infinity.

Proof. We have

$$\begin{aligned}
0 &\leq \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\
&= -E_{S_n} \int \log\left(\frac{f_{\hat{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x) \\
&= -E_{S_n} \int \log\left(\frac{f_{\hat{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x) \\
&\quad +(1+\delta)E_{S_n} \int \log\left(\frac{f_{\hat{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP^1_{n,S_n}(x) \\
&\quad -(1+\delta)E_{S_n} \int \log\left(\frac{f_{\hat{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP^1_{n,S_n}(x) \\
&\leq -E_{S_n} \int \log\left(\frac{f_{\hat{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x) \\
&\quad +(1+\delta)E_{S_n} \int \log\left(\frac{f_{\hat{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP^1_{n,S_n}(x) \\
&\quad -(1+\delta)E_{S_n} \int \log\left(\frac{f_{\tilde{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP^1_{n,S_n}(x) \\
&= -(1+2\delta)E_{S_n} \int \log\left(\frac{f_{\tilde{k}}(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x) + T_{n,\hat{k}} + R_{n,\tilde{k}},
\end{aligned}$$

where

$$
\begin{aligned}
T_{n,k} &= (1+\delta)E_{S_n}\int \log\left(\frac{f_k(x \mid P^0_{n,S_n})}{f(x)}\right) d(P^1_{n,S_n}-P)(x) \\
&\quad +\delta E_{S_n}\int \log\left(\frac{f_k(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x)
\end{aligned}
$$

and

$$
\begin{aligned}
R_{n,k} &= -(1+\delta)E_{S_n}\int \log\left(\frac{f_k(x \mid P^0_{n,S_n})}{f(x)}\right) d(P^1_{n,S_n}-P)(x) \\
&\quad +\delta E_{S_n}\int \log\left(\frac{f_k(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x).
\end{aligned}
$$

Thus

$$
0 \leq E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq (1+2\delta)\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\} + ET_{n,\hat{k}} + ER_{n,\tilde{k}}.
$$

In the sequel we will show that $\max(ET_{n,\hat{k}}, ER_{n,\tilde{k}}) \leq c(M_1,M_2,\delta)\frac{1+\log(K(n))}{np}$, which then completes the proof of the finite sample statement. The asymptotic implications are direct corollaries of this finite sample result.

We represent $T_{n,k}$ and $R_{n,k}$ as $T_{n,k} = E_{S_n}T_{n,k}(S_n)$ and $R_{n,k} = E_{S_n}R_{n,k}(S_n)$, respectively. We introduce the following notation for the relevant random variables

$$
\begin{aligned}
\tilde{H}_k &\equiv \int \log\left(\frac{f_k(x \mid P^0_{n,S_n})}{f(x)}\right) dP(x) \leq 0 \\
\bar{H}_k &\equiv \int \log\left(\frac{f_k(x \mid P^0_{n,S_n})}{f(x)}\right) dP^1_{n,S_n}(x).
\end{aligned}
$$

Note that

$$
\begin{aligned}
T_{n,k}(S_n) &= -(1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] + \delta\tilde{H}_k \\
R_{n,k}(S_n) &= -(1+\delta)\left[\bar{H}_k - \tilde{H}_k\right] + \delta\tilde{H}_k.
\end{aligned}
$$

This gives us for $s \geq 0$

$$
\begin{aligned}
Pr(T_{n,\hat{k}}(S_n) \geq s \mid P^0_{n,S_n}, S_n) &= Pr\left(-(\tilde{H}_{\hat{k}} - \bar{H}_{\hat{k}}) \geq \frac{1}{1+\delta}\left\{s - \delta\tilde{H}_{\hat{k}}\right\} \mid P^0_{n,S_n}, S_n\right) \\
&\leq K(n) \max_{k\in\{1,\ldots,K(n)\}} Pr\left(-(\tilde{H}_k - \bar{H}_k) \geq \frac{1}{1+\delta}\left\{s - \delta\tilde{H}_k\right\} \mid P^0_{n,S_n}, S_n\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& Pr(R_{n,\tilde{k}}(S_n) \geq s \mid P^0_{n,S_n}, S_n) \\
& \leq K(n) \max_{k \in \{1,\ldots,K(n)\}} Pr\left((\tilde{H}_k - \bar{H}_k) \geq \frac{1}{1+\delta} \left\{ s - \delta \tilde{H}_k \right\} \mid P^0_{n,S_n}, S_n \right).
\end{aligned}
$$

We now proceed bounding $Pr\left(\pm(\tilde{H}_k - \bar{H}_k) \geq \frac{1}{1+\delta} \left\{ s - \delta \tilde{H}_k \right\} \mid P^0_{n,S_n}, S_n \right)$, by using Bernstein's inequality, which we state here as a lemma for ease of reference. A proof is given in Lemma A.2, p. 564 in Györfi et al. (2002).

Lemma 1 Bernstein's inequality. *Let Z_i, $i = 1, \ldots, n$, be independent real valued random variables such that $Z_i \in [a, b]$ with probability one. Let $0 < \sum_{i=1}^n VAR(Z_i)/n \leq \sigma^2$. Then, for all $\epsilon > 0$,*

$$
Pr\left(\frac{1}{n} \sum_{i=1}^n (Z_i - EZ_i) > \epsilon \right) \leq \exp\left(-\frac{1}{2} \frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3} \right).
$$

This implies

$$
Pr\left(\frac{1}{n} \mid \sum_{i=1}^n (Z_i - EZ_i) \mid > \epsilon \right) \leq 2\exp\left(-\frac{1}{2} \frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3} \right).
$$

Conditional on P^0_{n,S_n}, S_n, we consider the random variable

$$
Z_k = -\log\left(\frac{f_k(X \mid P^0_{n,S_n})}{f(X)} \right),
$$

and let Z_{ki}, $i = 1, \ldots, np$, be the np i.i.d. copies of Z_k corresponding to X_i, given $S_n(i) = 1$. Note that $\bar{H}_k = -1/np \sum_{i=1}^{np} Z_{ki}$ and $\tilde{H}_k = -E(Z_k \mid P^0_{n,S_n}, S_n)$ so that $\tilde{H}_k - \bar{H}_k = 1/np \sum_{i=1}^{np} Z_{ki} - E(Z_k \mid P^0_{n,S_n}, S_n)$ represents a centered empirical mean of i.i.d. random variables. We will apply Bernstein's inequality to this centered empirical mean and exploit the following special property of Z_k to obtain an $\exp(-nps/c)$ tail probability instead of the usual $\exp(-nps^2/c)$ for some $c < \infty$. This will show that the centered empirical mean converges at an np rate instead of the usual $(np)^{0.5}$.

Lemma 2 *We have*

$$
\sigma_k^2 \equiv VAR(Z_k \mid P^0_{n,S_n}, S_n) \leq M_2 E(Z_k \mid P^0_{n,S_n}, S_n) = -M_2 \tilde{H}_k.
$$

Proof of Lemma. Note $EZ_k^2 = \int \log^2(f_k(x)/f(x))f(x)d\mu(x)$ and $EZ_k = -\int \log(f_k(x)/f(x))f(x)d\mu(x)$, where we use the short-hand notation f_k for $f_k(\cdot \mid P^0_{n,S_n})$. Firstly, (van der Vaart (1998), page 62) provides the following relation between the quadratic Hellinger distance and Kullback-Leibler distance for two densities f, g w.r.t. a dominating measure μ:

$$\int(\sqrt{g} - \sqrt{f})^2 d\mu \leq -\int \log(g/f) f d\mu.$$

This is shown as follows: Since $\log(x) \leq 2(\sqrt{x} - 1)$ we have

$$\begin{aligned}\int \log(g/f) f d\mu &\leq 2\int(\sqrt{g/f} - 1) f d\mu \\ &= 2\int \sqrt{g}\sqrt{f} d\mu - 2 \\ &= -\int(\sqrt{g} - \sqrt{f})^2 d\mu,\end{aligned}$$

where we used at the last equality that $-2 = -\int(\sqrt{f}^2 + \sqrt{g}^2)d\mu$. Secondly, we have

$$\int \log^2(g/f) f d\mu \leq 4 \parallel \frac{f}{\min(f,g)} \parallel_\infty \int(\sqrt{g} - \sqrt{f})^2 d\mu,$$

where the supremum is taken over a support of X. This is shown as follows: Applying $\log(x) \leq 2(\sqrt{x} - 1)$ to $\log(g/f)$ and $\log(f/g)$ yields:

$$\mid \log(g/f) \mid \leq 2\frac{\mid \sqrt{f} - \sqrt{g} \mid}{\min(\sqrt{f}, \sqrt{g})}.$$

Thus

$$\int \log^2(g/f) f d\mu \leq 4 \parallel \frac{f}{\min(f,g)} \parallel_\infty \int(\sqrt{f} - \sqrt{g})^2 d\mu.$$

Combining the two inequalities proves the lemma. □

We now proceed as follows. From Lemma 2 we have $-\tilde{H}_k \geq \sigma^2/M_2$, where $M_2 = 4L/\epsilon$. Thus,

$$\begin{aligned}
& Pr\left(-(\tilde{H}_k - \bar{H}_k) \geq \tfrac{1}{1+\delta}\left\{s - \delta \tilde{H}_k\right\}\Big|\, P^0_{n,S_n}, S_n\right) \\
& = Pr\left(E(Z_k \mid P^0_{n,S_n}, S_n) - \tfrac{1}{np}\textstyle\sum_{i=1}^n Z_{k,i} \geq \tfrac{1}{1+\delta}\left[s + \delta E(Z_k \mid P^0_{n,S_n}, S_n)\right]\Big|\, P^0_{n,S_n}, S_n\right) \\
& \leq Pr\left(E(Z_k \mid P^0_{n,S_n}, S_n) - \frac{1}{np}\sum_{i=1}^n Z_{k,i} \geq \frac{1}{1+\delta}\left[s + \delta \frac{\sigma_k^2}{M_2}\right]\Bigg|\, P^0_{n,S_n}, S_n\right) \\
& \leq \exp\left[-\frac{np}{2}\frac{1}{(1+\delta)^2}\frac{(s+\delta\sigma_k^2/M_2)^2}{\sigma_k^2 + \frac{M_1}{3(1+\delta)}(s+\delta\sigma_k^2/M_2)}\right],
\end{aligned}$$

where we applied Bernstein's inequality to the centered empirical mean $1/np\sum_i Z_{k,i} - E(Z_k \mid P^0_{n,S_n}, S_n)$, where we note that $\mid Z_k \mid < \log(L/\epsilon)$ so that we can set $b - a = M_1 = 2\log(L/\epsilon)$. The same bound applies to $Pr\left((\tilde{H}_k - \bar{H}_k) \geq \tfrac{1}{1+\delta}\left\{s - \delta\tilde{H}_k\right\}\Big|\, P^0_{n,S_n}, S_n\right)$.

We now note that for $s \geq 0$

$$\begin{aligned}
\frac{(s+\delta\sigma_k^2/M_2)^2}{\sigma_k^2 + \frac{M_1}{3(1+\delta)}(s+\delta\sigma_k^2/M_2)} & \geq \frac{(s+\delta\sigma_k^2/M_2)}{\frac{\sigma_k^2}{s+\delta\sigma_k^2/M_2} + \frac{M_1}{3}} \geq \frac{(s+\delta\sigma_k^2/M_2)}{\frac{M_2}{\delta} + \frac{M_1}{3}} \\
& \geq \frac{s}{\frac{M_2}{\delta} + \frac{M_1}{3}},
\end{aligned}$$

which is independent of k. This shows that

$$Pr(T_{n,\hat{k}}(S_n) \geq s \mid P^0_{n,S_n}, S_n) \leq K(n)\exp\left[-\frac{np}{c(M_1, M_2, \delta)}s\right]$$

with $c(M_1, M_2, \delta) = 2(1+\delta)^2(M_1/3 + M_2/\delta)$. The same bound applies to $Pr(R_{n,\tilde{k}}(S_n) \geq s \mid P^0_{n,S_n}, S_n)$.

Since the bound is independent of P^0_{n,S_n}, S_n, this provides us also with

$$\begin{aligned}
Pr(T_{n,\hat{k}}(S_n) \geq s) & \leq K(n)\exp\left[-\frac{np}{c(M_1, M_2, \delta)}s\right] \\
Pr(R_{n,\tilde{k}}(S_n) \geq s) & \leq K(n)\exp\left[-\frac{np}{c(M_1, M_2, \delta)}s\right].
\end{aligned}$$

Thus for each $u > 0$ we have

$$ET_{n,\hat{k}} = ET_{n,\hat{k}}(S_n)$$

$$
\begin{aligned}
&\leq \quad EI(T_{n,\hat{k}}(S_n) \geq 0)T_{n,\hat{k}}(S_n) \\
&= \quad \int_0^\infty Pr(T_{n,\hat{k}}(S_n) > s)ds \\
&\leq \quad u + \int_u^\infty K(n) \exp\left[-\frac{np}{c(M_1, M_2, \delta)} s\right] ds.
\end{aligned}
$$

The minimum is attained at $u = c(M_1, M_2, \delta)\log(K(n))/np$ and is given by $c(M_1, M_2, \delta)(\log(K(n))+1)/np$. Similarly, $ER_{n,\tilde{k}} \leq c(M_1, M_2, \delta)(\log(K(n))+1)/np$. This completes the proof of the theorem. □

2.3 Asymptotic optimality

Theorem 1 provides a finite sample bound for the expected value of $\tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k})$, which compares the performance of the cross-validated selector $\hat{k}$ to the benchmark $\tilde{k}$ in terms of the conditional Kullback-Leibler distances. $\tilde{\theta}_{n(1-p)}(k)$ based on $n(1-p)$ training observations. This bound is used to prove that the ratio $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to one, or equivalently hat $\tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to zero.

However, one would like the cross-validated selector $\hat{k}$ to perform as well as a benchmark selector $\tilde{k}_n$ based on the whole sample of size n, rather than only $n(1-p)$ as above. The following is an immediate corollary of Theorem 1, which relates $\tilde{\theta}_{n(1-p)}(\hat{k})$ to that of a benchmark selector based on n observations, $\tilde{\theta}_n(\tilde{k}_n)$. In this corollary, we use the notation $p = p_n$ to emphasize the dependence of the validation set proportion p on n. It shows that if $p = p_n$ converges slowly enough to zero when the sample size n converges to infinity, then, given a mild condition (4) below, the wished asymptotic optimality of the selector $\hat{k}$ follows.

Corollary 1 *Suppose that there exist $\epsilon > 0$ and $L < \infty$ so that $\epsilon < f_k(X \mid P_n) < L$ a.s. for all $k \in \{1, \ldots, K(n)\}$.*

If $p = p_n \to 0$, (3) holds, and for $n \to \infty$

$$
\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \quad \to \quad 1 \quad \textit{in probability} \tag{4}
$$

then

$$\frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \quad \rightarrow \quad 1 \quad \textit{in probability.} \tag{5}$$

A sufficient condition for (4) to hold is that

$$\left(n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right), (n(1-p_n))^\gamma \left(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}\right)\right) \overset{D}{\Rightarrow} (Z, Z)$$

for some $\gamma > 0$ and random variable Z with $Pr(Z > a) = 1$ for some $a > 0$. In particular, if $Pr(S_n = s) = 1$ for some $s \in \{0,1\}^n$ (i.e., single split cross-validation), then it suffices to assume $n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right) \overset{D}{\Rightarrow} Z$ for some $\gamma > 0$ and $Pr(Z > a) = 1$ for some $a > 0$.

Proof of Corollary. Firstly, note that

$$\frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \quad \rightarrow \quad 1$$

by Theorem 1. This proves the first statement of the corollary. We now show that (4) holds under the given sufficient condition. Define

$$\begin{aligned} Z_{1,n} &= n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right) \\ Z_{2,n} &= (n(1-p_n))^\gamma \left(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}\right) \end{aligned}$$

If $(Z_{1,n}, Z_{2,n}) \overset{D}{\Rightarrow} (Z, Z)$ then by the continuous mapping theorem we have $\frac{Z_{1,n}}{Z_{2,n}} \rightarrow 1$. However, note that

$$\frac{Z_{1,n}}{Z_{2,n}} = \frac{1}{(1-p_n)^\gamma} \frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}}.$$

Thus, if $p_n \rightarrow 0$, then we have

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \rightarrow 1,$$

and thus (4) holds. If there is only one split i.e. $P(S_n = s) = 1$ for some s, then $Z_{1,n} = Z_{2,\frac{n}{1-p_n}}$, and hence $Z_{1,n} \overset{D}{\Rightarrow} Z$ implies $(Z_{1,n}, Z_{2,n}) \overset{D}{\Rightarrow} (Z, Z)$. This completes the proof. □

An important and practical issue is the impact of the cross-validation proportion p on the estimators $\tilde{\theta}_{n(1-p)}(k)$ in relation to $\tilde{\theta}_n(k)$. The following discussion provides some intuition regarding the behavior of $\tilde{\theta}_{n(1-p)}(k)$ compared to $\tilde{\theta}_n(k)$ for a a density estimator $f_k(\cdot \mid P_n)$ based on the entire empirical distribution P_n. One can argue that, due to the expectation w.r.t. S_n in the definition of $\tilde{\theta}_{n(1-p)}(k)$, for each fixed $p \in (0,1)$, the first order linear approximation of $\tilde{\theta}_{n(1-p)}(k) - \tilde{\theta}_n(k)$ equals zero. This is formalized by the following argument. Let $\theta_k = -\int \log(f_k(x \mid P))dP(x)$ be the parameter corresponding withthe "estimator" $\tilde{\theta}_n(k) = \int \log(f_k(x \mid P_n))dP(x)$. Suppose

$$\tilde{\theta}_n(k) - \theta_k = \frac{1}{n}\sum_{i=1}^{n} IC_k(X_i \mid P) + R_k(P_n, P)$$

for some function $IC_k(\cdot \mid P)$ of X and remainder term $R_k(P_n, P)$. Application of this expansion to $\int \log(f_k(x \mid P^0_{n,S_n}))dP(x)$ and taking the expectation w.r.t. S_n yields

$$\tilde{\theta}_{n(1-p)}(k) - \theta_k = E_{S_n}\frac{1}{n(1-p)}\sum_{i=1}^{n} IC_k(X_i \mid P) I(S_n(i) = 0) + E_{S_n} R_k(P^0_{n,S_n}, P).$$

Now, we note that, the first term on the right-hand side actually equals $\frac{1}{n}\sum_{i=1}^{n} IC_k(X_i \mid P)$. Consequently,

$$\tilde{\theta}_n(k) - \tilde{\theta}_{n(1-p)}(k) = R_k(P_n, P) - E_{S_n} R_k(P^0_{n,S_n}, P).$$

In words, the difference between $\tilde{\theta}_n(k)$ and $\tilde{\theta}_{n(1-p)}(k)$ is driven by the second order terms. Due to this fact, that is, even for a fixed $p \in (0,1)$, $\tilde{\theta}_{n(1-p)}(k)$ can be viewed as a decent approximation of $\tilde{\theta}_n(k)$, one expects that the sensitivity of the likelihood cross-validation selector $\hat{k}(p)$ to the choice of p (i.e., the choice of distribution for S_n), is significantly less than it would be for single split cross-validation.

In fact, in our bandwidth selection simulation study we have the remarkable practical result that, for each choive of V defining the V-fold likelihood based cross-validation procedure (e.g., $V = 2$, which corresponds with $p = 0.5$), $(E\tilde{\theta}_n(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}) \approx 1$ for large sample size.

3 Simulation for bandwidth selection

3.1 Fixed p optimality result.

In this subsection, we illustrate the result of Theorem 1 in the context of kernel density estimation with a simulation study. In particular, the studied likelihood based cross-validation method is used to choose the optimal bandwidth in a density estimation problem using a gaussian kernel. The gaussian kernel density estimate for a sample $x_1, \cdots, x_n$ is given by

$$\hat{f}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right),$$

where $K(.)$ is the standard normal density function and b is the bandwidth of this kernel. We generated 20 replicate data sets from the standard normal distribution enforcing the compact support in the interval $[-2, 2]$ at each of the following six samples sizes: $n =$ 50, 100, 200, 400, 800, 1600. $K(n) = 100$ different bandwidth values b are generated from the interval $[0.02, 2]$ so that the difference between any two consecutive bandwidth values is 0.02. We set the proportion of the validation set to $p = 0.1$ and perform 10-fold likelihood based cross-validation to select the optimal bandwidth. For this choice of the kernel, the integral θ_{opt} is given by

$$\theta_{opt} = -\int_{-2}^{2} \log\left(\frac{\phi(x)}{\Phi(2) - \Phi(-2)}\right) \frac{\phi(x)}{\Phi(2) - \Phi(-2)} dx$$

where $\phi(x)$ and $\Phi(x)$ denote the density and the cumulative distribution function of the standard normal distribution, respectively. We performed the simulations in `R` and used the `R`-function `integrate()` to compute θ_{opt} and $\tilde{\theta}_{n(1-p)}(.)$ with numerical integration. Figure 1 shows the ratio $\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta opt}$ for each of the 20 replicate data sets at each of the six sample sizes. As predicted by Theorem 1, we observe from this plot that this ratio converges to 1 in probability as n increases . In Table 1 we report $\frac{\hat{E}\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\hat{E}\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta opt}$ at each sample size where $\hat{E}\tilde{\theta}_n(.)$ is the averaged $\tilde{\theta}_n(.)$ over 20 replicate data sets. To visualize this convergence result for a single data set as its size increases, we plot in Figure 2 the true density versus the kernel density estimate using the bandwidth selected by the likelihood based cross-validation method.

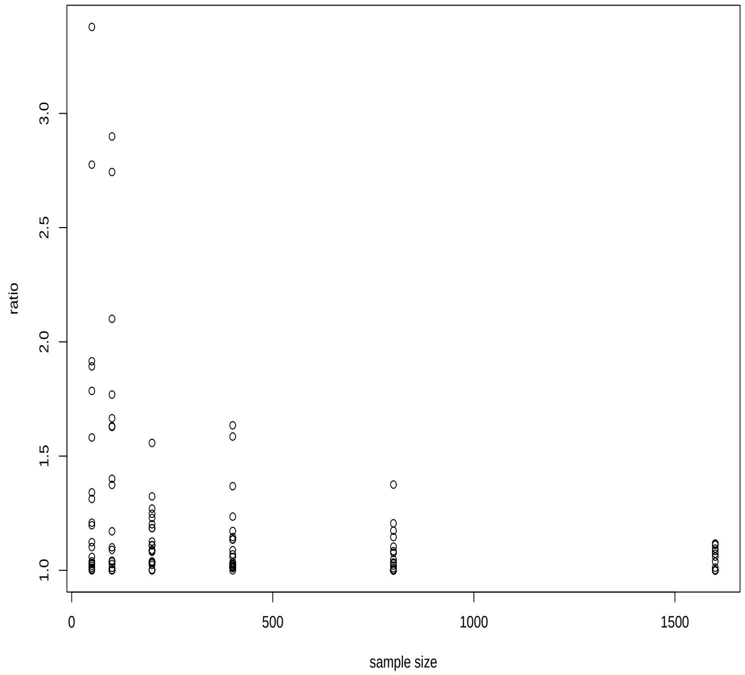

Figure 1: *Illustration of* $\frac{\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}} \rightarrow 1$: The ratios corresponding to various sample sizes are reported for 20 replicate data sets.

3.2 Sensitivity to p.

In this subsection we investigate the effect of p with a simulation. We have

$$\hat{k}(p) = \min\nolimits^{-1}_{k\in\{1,\cdots,K(n)\}}\hat{\theta}_{n(1-p)}(k),$$

for a given p. For the k-th bandwidth value b_k the true conditional risk based on n observations is given by

$$\tilde{\theta}_n(k) = -\int \log \hat{f}_{b_k}(x)f(x)dx,$$

where the kernel density estimate $\hat{f}_{b_k}(x)$ uses all of the n observations. Then, $\hat{k}(p)$ for $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$ are computed for

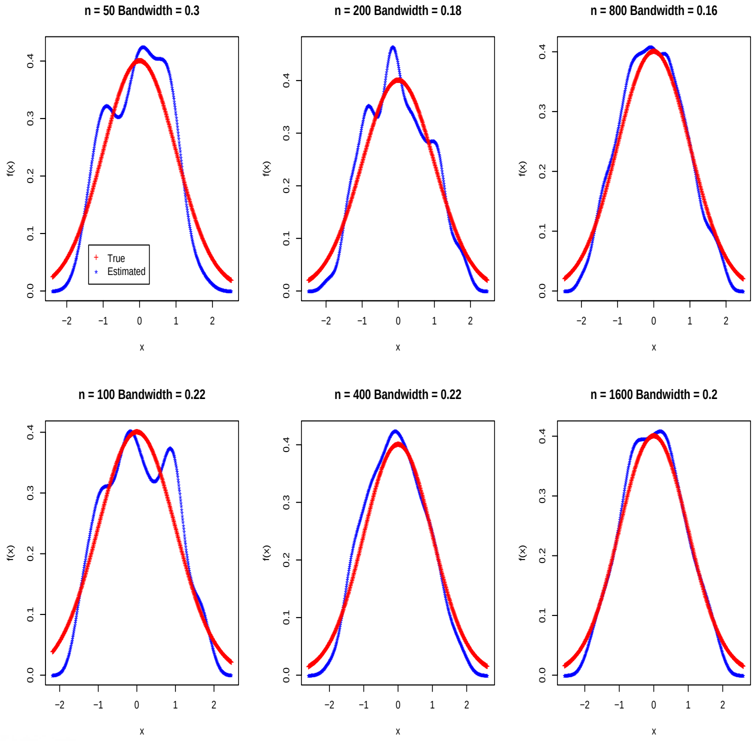

Figure 2: *True density versus estimated density based on a single data set*: The optimal bandwidth is selected by the 10-fold likelihood based cross validation.

n	50	100	200	400	800	1600
$\frac{\hat{E}\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}}{\hat{E}\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}}$	1.542497	1.400015	1.150882	1.139386	1.068780	1.033064

Table 1: $\frac{\hat{E}\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}}{\hat{E}\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}}$ based on 20 replicate data sets at each of the six different sample sizes.

	n					
p	50	100	200	400	800	1600
0.05	1.493594	1.465201	1.168274	1.115338	1.089441	1.047685
0.1	1.531736	1.391971	1.144236	1.136916	1.075563	1.048454
0.15	1.577241	1.473550	1.118831	1.117599	1.076197	1.061919
0.20	1.518429	1.417260	1.120498	1.100698	1.065835	1.064060
0.25	1.302580	1.443560	1.111674	1.182325	1.060759	1.100572
0.30	1.430726	1.388704	1.148916	1.119423	1.080356	1.083632
0.35	1.238741	1.414966	1.076628	1.093445	1.092477	1.112602
0.40	1.477980	1.617694	1.200306	1.123990	1.091412	1.091008
0.45	1.411283	1.483116	1.090528	1.142125	1.134810	1.143657
0.50	1.320979	1.398095	1.099359	1.136470	1.146952	1.167325

Table 2: *V-fold likelihood based cross validation:* $\frac{\hat{E}\tilde{\theta}_n(\hat{k}(p))-\theta_{opt}}{\hat{E}\tilde{\theta}_n(\tilde{k})-\theta opt}$ based on 20 replicate data sets at six different sample sizes for each $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$.

10 replicate data sets at each of the six different sample sizes: $n =$ 50, 100, 200, 400, 800, 1600. Note that each choice of p corresponds to approximately a $1/p$-fold cross validation scheme. Table 2 reports $\frac{\hat{E}\tilde{\theta}_n(\hat{k}(p))-\theta_{opt}}{\hat{E}\tilde{\theta}_n(\tilde{k})-\theta_{opt}}$ based on 20 replicate data sets at each of the six different sample sizes. It is evident from this table that the likelihood based cross-validation procedure is performing equally well with any choice of p. We also report the same quantity obtained performing likelihood based cross-validation with single split using various p-values in Table 3. As we commented in subsection 2.1, the likelihood based cross-validation procedure with single split seems to be sensitive to the choice of p.

	n					
p	50	100	200	400	800	1600
0.05	21.778985	30.591547	5.366258	3.488738	2.147304	1.287172
0.1	4.969151	8.139912	3.709904	2.105173	1.948626	1.291611
0.15	1.972465	5.234631	2.283455	1.831317	1.628340	1.153562
0.20	1.836114	10.036376	2.465654	1.377272	1.370639	1.093183
0.25	2.495359	4.262036	1.246727	1.232388	1.209813	1.092931
0.30	2.260952	4.298054	1.410498	1.149826	1.215430	1.123646
0.35	1.553013	3.862468	1.511450	1.111143	1.165148	1.151871
0.40	1.446852	1.615702	1.276998	1.123451	1.146859	1.113719
0.45	1.583617	1.757668	1.263186	1.170124	1.112150	1.133443
0.50	1.333555	2.193936	1.258745	1.164263	1.149889	1.175700

Table 3: *Single split likelihood based cross-validation:* $\frac{\hat{E}\tilde{\theta}_n(\hat{k}(p))-\theta_{opt}}{\hat{E}\tilde{\theta}_n(\tilde{k})-\theta_{opt}}$ based on 20 replicate data sets at six different sample sizes for each $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$ using a single split.

4 Discussion

In this article 1) for a fixed $p \in (0,1)$, we have established asymptotic equivalence of the likelihood cross-validation selector $\hat{k}(p)$ and the benchmark selector $\tilde{k}_{n(1-p)}$, and 2) for a sequence $p = p_n$ converging to zero slowly enough with sample size n, we showed asymptotic equivalence of $\hat{k}(p_n)$ and the optimal selector $\tilde{k}_n$. Here we use the notation $\hat{k}(p)$ to stress the dependence of the selector $\hat{k}$ on p We also argued, and illustrated this in our simulation study, that in many applications the asymptotic performance of $\hat{k}(p)$ for fixed p could be relatively insensitive to the choice p. In future research we plan to study the sensitivity to p in more detail and develop, and test a proposal for a data adaptive choice $\hat{p}$.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. Petrov & F. Csaki, eds. Budapest: Academiai Kiado.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse fisher information matrix. In *Information and Classification*, O. Opitz, B. Lausen & R. Klar, eds. Heidelberg: Springer Verlag.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology* **44**, 62–91.

Breiman, L. (1996). Out-of-bag estimation. Tech. rep., Department of Statistics, U.C. Berkeley.

Breiman, L., Friedman, J., Olshan, R. & Stone, C. (1984). *Classification and Regression Trees*. Monterey: Wadsworth & Brooks/Cole.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validati on and the repeated learning testing methods. *Biometrika* **76**, 503–514.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**, 320–328.

Györfi, L., Kohler, M., A., K. & Walk, H. (2002). *A Distribution-Free Theory of Nanparametric Regression*. New York: Springer-Verlag.

Hansen, M. & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association* **96**, 746–774.

Härdle, W. (1993). *Applied Nonparametric Regression*. Cambridge University Press.

Härdle, W. & Marron, J. (1985a). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72**, 481–484.

Härdle, W. & Marron, J. (1985b). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics* **13**, 1465–1481.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2001). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer Verlag.

Pavlic, M. & van der Laan, M. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data Analysis* **41**, 413–428.

Ripley, B. D. (1996). *Pattern recognition and neural networks.* Cambridge, New York: Cambridge University Press.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica* **14**, 465–471.

Schuster, E. & Gregory, C. (1981). On the non-consistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the interface*, W. Eddy, ed.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Scott, D. & Factor, L. (1981). Monte carlo study of three data-based nonparametric density estimators. *Journal of the American Statistical Association* **76**, 9–15.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Silverman, B. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *Journal of the American Statistical Association* **79**, 584–589.

Silverman, B. (1986). *Density Estimation for Statistics and Data analysis.* Chapman & Hall.

Smyth, P. (2000). Model selection of probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **10**, 63–72.

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics* **12**, 1285–1297.

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

STONE, M. (1974b). Cross-validatory choice and assessment of statistics predictions. *Journal of the Royal Statistical Society, Series B* **36**, 111–147.

STONE, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64**, 29–35.

VAN DER VAART, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

ZHANG, P. (1993). Model selection via multifold cross-validation. *Annals of Statistics* **21**, 299–313.