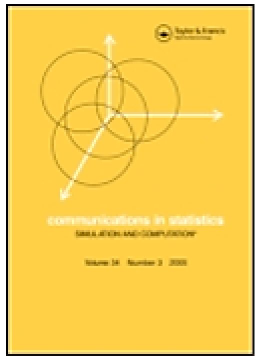
Estimating prediction error: Cross-validation vs accumulated prediction error

Jenny Häggström, Xavier de Luna



Estimating prediction error: Cross-validation vs accumulated prediction error

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2009-0244.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	22-Jan-2010
Complete List of Authors:	Häggström, Jenny; Umeå University, Dept of Statistics de Luna, Xavier; Umeå University, Dept of Statistics
Keywords:	Local polynomial regression, Non-parametric regression, Out-of-sample validation, Smoothing parameter
Abstract:	We study the validation of prediction rules such as regression models and classification algorithms through two out-of-sample strategies, cross-validation and accumulated prediction error. We use the framework of Efron (1983) where measures of prediction errors are defined as sample averages of expected errors and show through exact finite sample calculations that cross-validation and accumulated prediction error yield different smoothing parameter choices in non-parametric regression. The difference in choice does not vanish as sample size increases.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
APEHaggstromdeLuna.zip	

Estimating prediction error: Cross-validation vs accumulated prediction error.

Jenny Häggström and Xavier de Luna*

Department of Statistics, Umeå University, SE-90187 Umeå, Sweden

January 22, 2010

Abstract

We study the validation of prediction rules such as regression models and classification algorithms through two out-of-sample strategies, cross-validation and accumulated prediction error. We use the framework of Efron (1983) where measures of prediction errors are defined as sample averages of expected errors and show through exact finite sample calculations that cross-validation and accumulated prediction error yield different smoothing parameter choices in non-parametric regression. The difference in choice does not vanish as sample size increases.

Keywords: Local polynomial regression; Non-parametric regression; Out-of-sample validation; Smoothing parameter.

*Corresponding address: Department of Statistics, Umeå University, , SE-90187 Umeå, Sweden. E-mail: xavier.deluna@stat.umu.se Tel: +46 90 7865559. Fax: +46 90 7866614.

1 Introduction

This paper is concerned with the validation of general prediction rules for a variable of interest y. We consider situations where a prediction rule $\hat{\mu}^n$ for y is obtained using a sample of size n, $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. Typical examples of prediction rules are regression and classification applications. As an illustration of the former, Figure 1 (left panel) displays 221 observations from a light detection and ranging (LIDAR) experiment, where the horizontal axis is the distance, x, travelled by laser-emitted light before it is reflected back to its source and the vertical axis is the logarithm of the ratio, y, of received light from two laser sources, one of which had a frequency equal to the resonance frequency of mercury. LIDAR experiments are used to detect such chemical compounds in the atmosphere; see Ruppert et al. (2003, Sec. 2.7) for more details. Together with the data two regression curves are displayed (plain and dashed line), which are prediction rules of y given x.

As Efron (2004) pointed out there are two main schools to approach the problem of assessing the performance of different prediction rules for a given variable y: out-of-sample methods and covariance penalty methods. The latter class of methods is model based, i.e. parametric statistical models for the data generating mechanism are assumed, while the former is non-parametric and allow for the comparison of prediction rules obtained with different inferential frameworks and/or different modelling strategies. Although covariance penalty methods may be more efficient when the assumptions made hold (e.g., Efron, 2004), we consider in this paper out-of-sample validation which is more widely applicable. The evaluation of prediction rules is concerned with the question: "what is the generalizability of a given prediction rule?" A common approach to validation arises from a translation of this question to, here quoting Efron (2004): (Question 1) "We wonder how well [a prediction rule] will predict a future dataset [of same size n] independently generated by the same mechanism that produced $\mathbf{y}$." Under certain conditions (see Section 2), this statement justifies the use of a cross-validation (CV, Stone, 1974) criterion

$$\frac{1}{n}\sum_{i=1}^{n}\mathscr{L}(y_i, \tilde{\mu}_i^{n-1}), \tag{1}$$

where $\mathscr{L}$ is a loss function and $\tilde{\mu}_i^{n-1}$ is the prediction rule $\hat{\mu}^n$ obtained with a sample where observation y_i has been omitted.

Another approach to measuring the generalizability of a prediction rule arises from asking: (Question 2) How good is a prediction rule at predicting n observations arriving sequentially? As described in Section 2, this question justifies the use of the accumulated prediction error criterion (APE) (due to Dawid (1984) and Rissanen (1986))

$$\frac{1}{n-m}\sum_{i=m+1}^{n}\mathscr{L}(y_i,\breve{\mu}_i^{i-1}), \tag{2}$$

where $\breve{\mu}_i^{i-1}$ is the prediction rule $\hat{\mu}^n$ obtained with the sub-sample $\mathbf{y}^{i-1}=(y_1,y_2,\ldots,y_{i-1})'$ and m is the size of the sub-sample on which the first prediction rule is obtained.

CV is widely used, particularly in non-parametric applications, and the study of its properties has received much attention in the literature. The application of APE methods is less spread, except for time series situations (e.g., Sjöstedt, 2000, de Luna and Genton, 2005, Wagenmakers, Grünwald, and Steyvers, 2006), where the time ordering of the observed units is used in accumulating prediction errors in APE. However, APE is also applicable in situations where units are exchangeable (no natural ordering exists) as was advocated by Dawid (1984) and later on by de Luna and Skouras (2003), where APE is shown to be consistent in discriminating between model selection strategies.

In this paper we focus on the APE criterion and use the framework of Efron (1983) to make comparisons with CV. This allows us to shed some new light on the properties of these criteria. In particular, the APE criterion is an unbiased estimate of a measure of prediction error which is different from the one used to justify CV. The use of CV and APE to select the smoothing parameter in nonparametric regression is further studied and we show, for instance, that CV and APE yield different choices of smoothing parameters, where the difference does not vanish as the sample size increases.

The paper is organized as follows. The next section presents the two different measures of prediction errors which can be used to justify the use of CV and APE respectively. Section 3 focuses on linear prediction rules in regression situations, where out-of-sample methods are typically used to select a smoothing parameter. We review some asymptotics on the use of CV for this purpose and note that few results are available for APE. We then give finite sample expressions for the measures of prediction errors estimated by CV and APE. These results allows us in Section 4 to present numerical experiments to study and compare CV and APE. The paper is concluded in Section 5.

2 Measures of prediction error: new data versus sequential principle

Validation methods can be justified as estimators of measures of prediction errors. In this respect, a measure justified by Question 1 mentioned above is (e.g., Efron, 2004)

$$\mathrm{Err}^{new}=\frac{1}{n}\sum_{i=1}^{n}E\{\mathscr{L}(y_i^0,\hat{\mu}_i^n)|\mathbf{y}\}, \tag{3}$$

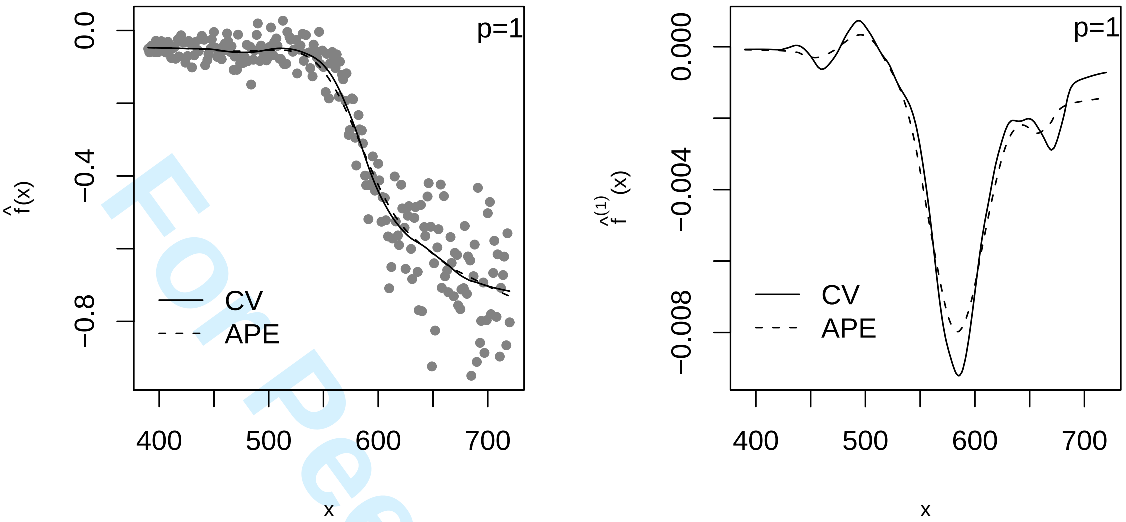


Figure 1: Left hand panel: LIDAR data with two loess fit obtained with $h = 0.23$ (plain line, CV choice) and with $h = 0.31$ (dashed line, APE choice). Right hand panel: Fitted first derivatives corresponding to the curve fitted on the left hand panel.

where $y_1^0, y_2^0, \ldots, y_n^0$ is a new independent sample generated by the same mechanism as $\mathbf{y}$. For instance, with a linear prediction rule (see Section 3.1 below) and squared error loss, $\mathscr{L}(y, \mu) = (y - \mu)^2$, it can be shown (Hastie and Tibshirani, 1990) that the CV criterion (1) yields an approximately unbiased estimate of Err^{new}, $E(\frac{1}{n}\sum_{i=1}^n (y_i - \tilde{\mu}_i^{n-1})^2) \approx E(\mathrm{Err}^{new})$.

A measure of prediction error based on Question 2 is:

$$\mathrm{Err}^{seq} = \frac{1}{n-m} \sum_{i=m+1}^{n} E\{\mathscr{L}(y_i, \breve{\mu}_i^{i-1})|\mathbf{y}^{i-1}\}. \tag{4}$$

An advantage of this prediction rule is that an unbiased estimator of $E(\mathrm{Err}^{seq})$ is readily available in wide generality by using the APE criterion (2): by the law of iterated expectations $E\{\frac{1}{n-m}\sum_{i=m+1}^n \mathscr{L}(y_i, \breve{\mu}_i^{i-1})\} = E(\mathrm{Err}^{seq})$.

3 CV and APE applied to nonparametric regression

3.1 Linear prediction rules

We consider situations where $\mathbf{y}$ is observed at fixed design points $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$. Then, prediction rules are typically constructed by regressing y on x, for example with linear predictors of the form

$$\hat{\boldsymbol{\mu}}^n_{\mathbf{z}} = \mathbf{S}(\mathbf{z})\mathbf{y},$$

where $\mathbf{S}(\mathbf{z})$ is an $n_z \times n$ matrix with n_z the dimension of the design vector $\mathbf{z}$ at which predictions are made. For $\mathbf{z} = \mathbf{x}$, $\hat{\boldsymbol{\mu}}^n_{\mathbf{x}}$ contains fitted values (i.e. prediction of the response at the observed design points). When parametric linear models are fitted with ordinary least squares we have $\mathbf{S}(\mathbf{x}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which is called the hat matrix, and where $\mathbf{X} = (\mathbf{1}\ \mathbf{x})$ with $\mathbf{1}$ an n-length vector of 1:s.

Examples of nonparametric linear prediction rules include kernel smoothers, cubic smoothing splines and local polynomial regression (see e.g., Schimek, 2000, and references therein). We focus on the latter in this paper to illustrate the use of CV and APE. Local polynomial regression with weights assigned by the tricube kernel, called *loess* in Cleveland and Devlin (1988), consists of fitting a polynomial of degree p at a design point z using only the part of the data that is deemed to be sufficiently close to the target. The fit, at z, is

$$\hat{\mu}^n_z(h) = \mathbf{e}_1'(\mathbf{Z}'\mathbf{W}_{h,z}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}_{h,z}\mathbf{y}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$, a $(p+1)$-length vector,

$$\mathbf{Z} = \begin{pmatrix} 1 & (x_1 - z) & \ldots & (x_1 - z)^p \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - z) & \ldots & (x_n - z)^p \end{pmatrix}$$

and

$$\mathbf{W}_{h,z} = \mathrm{diag}(K\big((x_1 - z)/b_1(h)\big)/b_1(h), \ldots, K\big((x_n - z)/b_n(h)\big)/b_n(h)).$$

The tricube kernel, $K(\cdot)$, is defined as

$$K(u) = \left\{ \begin{array}{ll} \frac{70}{81}(1 - |u|^3)^3, & \text{if } |u| < 1 \\ 0, & \text{if } |u| \geq 1 \end{array} \right\},$$

and, assuming no ties,

$$b_j(h) = \text{the } (hn)\text{:th nearest (in Euclidean distance) to } z \text{ among the } x_j\text{:s for } x_j \neq z,\ h \in [1/n, 1].$$

The parameter h is typically called smoothing parameter and in this case is the proportion of observations being used to produce the local fit. Thus, if h=1 all observations are used and the fit is global, while if h=0.2 then the $0.2n$ observations closest to z are used to produce $\hat{\mu}_z^n(h)$. This implies that if h is held constant for $i = 1, \ldots, n$ a constant number of observations are used for the local fits while $b_j(h)$, the size of the local neighborhood, varies with the data points x_j.

An illustrative example is provided in Figure 1. On the left hand panel, the LIDAR data described in the introduction is plotted together with two fitted curves with `loess` ($p = 1$) in R (R Development Core Team, 2008), one obtained with $h = 0.23$ (dashed) and one with $h = 0.31$. The difference between the two fit is most obvious by looking at the corresponding fitted first derivatives (right hand panel). We can see that a larger value for h yields a less wiggly behaviour. Out-of-sample validation is often used to select a relevant value for the smoothing parameter. The values used in Figures 1 correspond to the choices made with CV and APE ($m = 50$). The larger value for h, and hence the smoother curve, is obtained with APE. Note that the derivative is of main interest in this application since its proportional to mercury concentration at distance x, see Ruppert et al. (1997).

3.2 Choice of smoothing parameter

While CV is routinely used to choose a relevant smoothing parameter, the use of APE is rare. As a consequence there are many theoretical results available on the properties of CV as a method to select smoothing parameters, while these are scarce for APE. We make a review of these results (selective for CV and up to our knowledge exhaustive for APE).

3.2.1 Reviewing some asymptotics on the use of CV and APE

In the literature concerning smoothing parameter selection, the optimal smoothing parameter is often defined (using the squared error loss) either as

$$\hat{h}_0 = \arg\min_h \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{\mu}_i^n(h))^2$$

or

$$h_0 = \arg\min_h \frac{1}{n}\sum_{i=1}^{n} E[(\mu_i - \hat{\mu}_i^n(h))^2],$$

where $\mu_i = E(y_i)$, which is allowed to vary at different design points x_i. Note that throughout the design $\mathbf{x}$ is considered as non-random. Since $\mathrm{Err}^{new}(h)$ can be expressed as

$$\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{\mu}_i^n(h))^2,$$

where $\sigma^2 = Var(y_i)$, $\hat{h}_0$ and h_0 are the minimizers of $\mathrm{Err}^{new}(h)$ and $E(\mathrm{Err}^{new}(h))$, respectively. Consequently the asymptotic performance of smoothing parameter selectors is often measured in terms of the rate of convergence of the resulting $\hat{h}$ to one of the optimal smoothing parameters defined above. The relative asymptotic success of CV depends on if one considers $\hat{h}$ as an estimator of $\hat{h}_0$ or h_0. From Hall and Johnstone (1992), in a setting where the prediction rule is a kernel estimator and the smoothing parameter representing a local neighborhood of constant size, we have that

$$\left(\frac{h_{CV} - h_0}{h_0}\right) = O_p(n^{-1/10}),$$

where h_{CV} is the CV smoothing parameter, i.e. chosen by minimizing (1). This is a relatively slow rate of convergence, since there exist plug-in selectors for which the relative error is of $O_p(n^{-1/2})$ (albeit requiring strong smoothness assumptions on f (Hall and Johnstone, 1992)). In the case of approximating $\hat{h}_0$ CV does perform somewhat better. The relative error is still of $O_p(n^{-1/10})$ but in this situation h_{CV} is optimal in the sense that there is no empirical smoothing parameter for which the relative error can be reduced below $n^{-1/10}$; see Hall and Johnstone (1992). Similar results for selection of a constant size neighborhood and selection of a smoothing parameter of the k nearest neighbor type (i. e equivalent to h in Section 3.1) in local linear regression can be found in Li and Racine (2004) and Ouyang, Li, and Li (2006), respectively.

For APE there are no results qualitatively comparable to those reviewed above. Modha and Masry (1998) gave, however, rates of convergence for the integrated mean-squared errors in estimating non-parametrically the regression functions, showing that CV and APE achieved the same rates. Finally, de Luna and Skouras (2003) showed a consistency result for APE holding under weak assumptions. Loosely, their results say that APE will eventually (as the sample size grows) choose the prediction strategy that has lowest $E(\text{Err}^{seq})$. This result is limited to the comparison of a finite number of strategies and, therefore, it does not apply to the selection of a smoothing parameter. On the other hand, APE could consistently choose between a collection of prediction strategies defined by using different linear smoothers associated with different bandwidth selection criteria.

3.3 Finite sample properties

Explicit finite sample expressions for prediction errors can be obtained for linear prediction rules allowing for direct comparison, without the need to use asymptotic approximations. Thus, assuming independently distributed observations, a fixed design vector $\mathbf{x}$ and using the squared error loss, we have

$$
\begin{aligned}
\mathrm{Err}^{new}(h) &= \frac{1}{n}\sum_{i=1}^{n} E\{(y_i^0 - \hat{\mu}_i^n)^2|\mathbf{y}\} \\
&= \frac{1}{n}\sum_{i=1}^{n} E\Big\{\Big(y_i^0 - \sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}y_j\Big)^2\Big|\mathbf{y}\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} E\Big\{\Big(y_i^0 - \mu_i + \mu_i - \sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}y_j\Big)^2\Big|\mathbf{y}\Big\} \\
&= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(\mu_i - \sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}y_j\Big)^2, \\
E(\mathrm{Err}^{new}(h)) &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} E\Big\{\Big(\mu_i - \sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}y_j\Big)^2\Big\} \\
&= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(\mu_i^2 - 2\mu_i\sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}E(y_j) + Var(\{\mathbf{S}_h(\mathbf{x})\}_{ij}y_j) \qquad (5)\\
&\quad + \Big\{E\Big(\sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}y_j\Big)\Big\}^2\Big) \\
&= \sigma^2 + \frac{\sigma^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(\mu_i - \sum_{j=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ij}\mu_j\Big)^2 \qquad (6)\\
&\qquad (7)
\end{aligned}
$$

where $\{\mathbf{S}_h(\mathbf{x})\}_{ij}$ is the (ij):th entry in the $n \times n$ smoothing matrix. Analogous derivations yield

$$
\begin{aligned}
\mathrm{Err}^{seq}(h) &= \sigma^2 + \frac{1}{n-m}\sum_{i=m+1}^{n}\Big(\mu_i - \sum_{j=1}^{i-1}\{\mathbf{S}_h^{i-1}(\mathbf{x}^i)\}_{ij}y_j\Big)^2, \\
E(\mathrm{Err}^{seq}(h)) &= \sigma^2 + \frac{\sigma^2}{n-m}\sum_{i=m+1}^{n}\sum_{j=1}^{i-1}\{\mathbf{S}_h^{i-1}(\mathbf{x}^i)\}_{ij}^2 \qquad (8)\\
&\quad + \frac{1}{n-m}\sum_{i=m+1}^{n}\Big(\mu_i - \sum_{j=1}^{i-1}\{\mathbf{S}_h^{i-1}(\mathbf{x}^i)\}_{ij}\mu_j\Big)^2, \qquad (9)
\end{aligned}
$$

where $\{\mathbf{S}_h^{i-1}(\mathbf{x}^i)\}_{ij}$ is the (ij):th entry in the $i \times (i-1)$ smoothing matrix based on the design for sub-sample $i = 1, 2, \dots, i-1$ with $\mathbf{x}^i = (x_1, x_2, \dots, x_i)$.

When the data generating mechanism is known (in particular μ_i and σ^2) these prediction errors can be computed for different values of h and compared. Such numerical illustrations are provided in the next section giving some insights on how these two measures of prediction errors differ.

4 Numerical experiments

We have several objectives in this section. First we want to illustrate numerically that the minimization of $E(\mathrm{Err}^{new})$ and $E(\mathrm{Err}^{seq})$ may yield different optimal smoothing parameters, where the difference does not vanish with increasing sample sizes, thereby showing that both measures of prediction errors are not measuring the same thing. Secondly, we want to study the finite sample properties of smoothing parameters obtained by estimating $E(\mathrm{Err}^{new})$ and $E(\mathrm{Err}^{seq})$ with CV and APE respectively.

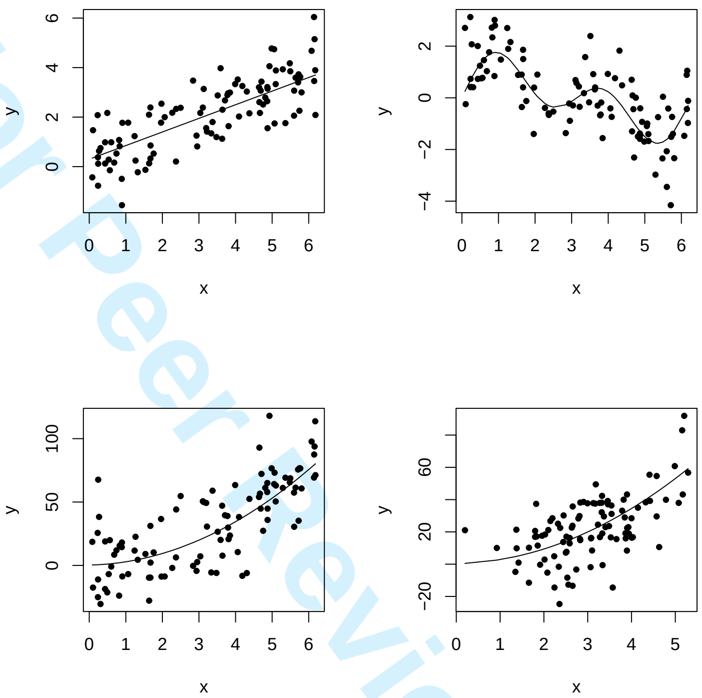


Figure 2: Regression functions and 100 generated observations: f_1 and f_2 from left to right in the upper panel, f_3 with $x_i \sim \mathcal{U}$ and f_3 with $x_i \sim \mathcal{N}$ from left to right in the lower panel.

4.1 Design of the experiments

The following data generating mechanisms are considered:

$$y_{k,i} = f_k(x_i) + \epsilon_{k,i}, \quad k = 1, 2, 3, \ i = 1, \ldots, n, \ n = 100, 200, 500, 1000$$

where $f_1(x_i) = 0.3 + 0.55x_i$, $f_2(x_i) = \sin(x_i) + \sin(2x_i)$ and $f_3(x_i) = 0.3 + 0.55x_i + 2x_i^2$, either $x_i \sim \mathcal{U}(0, 2\pi)$ or $x_i \sim \mathcal{N}(\pi, 1)$, x_i is generated only once since it is considered fixed. $\epsilon_{k,i} \sim \mathcal{N}(0,\mathrm{Var}(f_k(x_i)))$, thereby ensuring that the signal-to-noise ratio equals one. For $x_i \sim \mathcal{U}(0, 2\pi)$, $\mathrm{Var}(f_1(x_i)) = (0.3025/3)\pi^2 \approx 0.9952$, $\mathrm{Var}(f_2(x_i)) = 1$ and $\mathrm{Var}(f_3(x_i)) = (0.3025/3)\pi^2 + (4.4/3)\pi^3 + (256/45)\pi^4 \approx 600.6206$. For $x_i \sim \mathcal{N}(\pi, 1)$, $\mathrm{Var}(f_3(x_i)) = (8.3025/3) + 4.4\pi + 16\pi^2 \approx 180.0392$.

We consider local linear regression (`loess` with $p = 1$ in R, (R Development Core Team, 2008)) and local quadratic regression (`loess` with $p = 2$) estimators. For (2), (4) and (9) we fix $m = 50$. **In order to investigate if the behavior of APE becomes more similar to CV if m is allowed to increase with n, as suggested by Ing (2007), we also let $m = n/2$ for $n = 200, 1000$, f_1, f_2, $p = 1, 2$ with $x_i \sim \mathcal{U}$.** We call this criterion APE$_\delta$. Moreover, an arbitrary ordering of the observations is used to accumulate prediction errors. Using a non-random ordering should be avoided when observations are exchangeable since, for instance, accumulating prediction errors for observations ordered with increasing x value would imply that the predictions made are always outside the support of the fitted regression function. We also include in the study the popular model based covariance penalty criteria AIC (Akaike, 1974), $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^n(h))^2 + \frac{2}{n}\hat{\sigma}^2\sum_{i=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ii}$, and BIC (Schwarz, 1978), $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^n(h))^2 + \frac{log(n)}{n}\hat{\sigma}^2\sum_{i=1}^{n}\{\mathbf{S}_h(\mathbf{x})\}_{ii}$ for comparison. $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i^n(h^*))^2}{n - \sum_{i=1}^{n}(2\{\mathbf{S}_{h^*}(\mathbf{x})\}_{ii} - \sum_{j=1}^{n}\{\mathbf{S}_{h^*}^2(\mathbf{x})\}_{ij})}$ with $h^* = 0.13$. When looking for optimal h values we restrict attention to the span $[0.1, 1]$. Figure 4 displays data simulated with the design described.

The design of this study is chosen to be able to distinguish two types of situations: cases where the true regression function $f(x)$ is a polynomial function of order smaller or equal to the order p used in loess (correct specification case), and cases where $f(x)$ is not a polynomial function or it is a polynomial function of order larger than p (misspecification case).

4.2 Results

Results are based on 1000 replicates. For all the simulated situations we show results on the minimal values obtained for the different measures of prediction errors and their estimates (Tables 1-2 and Figures 2-3, 6-7; tables and Figures 6-7 in the appendix), as well as for the argument of these minima, i.e. the h values selected (Tables 3-4 and Figures 4-5, 8-9; tables and Figure 8-9 in the appendix).

We first comment on the minima obtained for $E(\mathrm{Err}^{new})$ and $E(\mathrm{Err}^{seq})$ and their respective smoothing parameters. We can note that for designs with polynomial regression functions, the minima of $E(\mathrm{Err}^{new})$ and $E(\mathrm{Err}^{seq})$ obtained by using a value of p such that we are in the correct specification case are smaller, for n large enough, than the minima obtained with lower value of p (such that we are in the misspecification case, Table 2 f_3) and the minima obtained with higher values of p (Table 1, f_1). Estimates of the minimum of $E(\mathrm{Err}^{new})$ and $E(\mathrm{Err}^{seq})$ obtained with CV and APE seem to be comparable with a slight systematic advantage for CV. This difference may be due to minima of Err^{seq} having a larger variability than the minima of Err^{new}.

We focus now on Table 3-4 summarized in Figures 4-5, 8-9. A striking result in Table 3 and 4 is the fact that the difference between the values for h minimizing $E(\mathrm{Err}^{new})$ and $E(\mathrm{Err}^{seq})$ increases with the sample size in all misspecification situations. This shows that there is an intrinsic difference between the two measures of prediction errors. In the correct specification cases we observe that the selected values of h approach one as the sample size grows, while they tend to zero in the misspecification cases. It can be noted that APE yields h values closer to the ideal value one in the former situations. On the other hand, CV is slightly less biased (as an estimator of h minimizer of $E(\mathrm{Err}^{new})$) than APE (as an estimator of h minimizer of $E(\mathrm{Err}^{seq})$). However, APE yields h values which are less variable than those obtained with CV and this results in lower MSE for APE, in both types of situations, when samples are large enough.

The discrepancy, in selected values of h, between APE_δ and CV decreases with n. This is in accordance with the theoretical results by Ing (2007) mentioned in Section 4.1.

Finally, while AIC behaves similarly to CV (note that they have been shown to be asymptotically equivalent in parametric situations, (Stone (1977); Shao (1997)), BIC yields systematically higher values than APE. This is to be put in contrast with results showing that BIC and APE are asymptotically equivalent for linear time series models (Ing, 2007). Thus, while AIC may be a reasonable substitute for CV in the situations studied, BIC cannot be used if the interest lies in minimizing $E(\mathrm{Err}^{seq})$.

5 Discussion

Cross-validation and accumulated prediction error are two methods used in order to validate and compare different prediction rules without making parametric assumptions. In this paper, we highlight that these two methods are estimating two different measures of prediction error, answering thereby different questions.

In order to compare CV and APE we use the framework developed by Efron (1983) to study CV and give its counterpart for APE. This allows us to show that when the prediction rules are linear smoothers, and the above methods are used to select the amount of smoothing (smoothing parameter) then the prediction errors estimated by CV and APE lead to different optimal smoothing parameters. Moreover, the latter difference does not vanish as the sample increases. This result is important since it shows that the choice between CV and APE does matter.

Our comparative results are exact for finite samples and do not rely on asymptotic approximations. Moreover, the non-equivalence result of CV and APE in our non-parametric context is in accordance with previous results obtained with parametric prediction rules; see de Luna and Skouras

(2003) and references therein.

Acknowledgments

We acknowledge the financial support of the Swedish Research Council (grant 70246501, the Ageing and Living Condition Program and the Swedish Initiative for Microdata Research in the Medical and Social Sciences). The simulation results obtained were run on facilities made available by the High Performance Computing Center North at Umeå University.

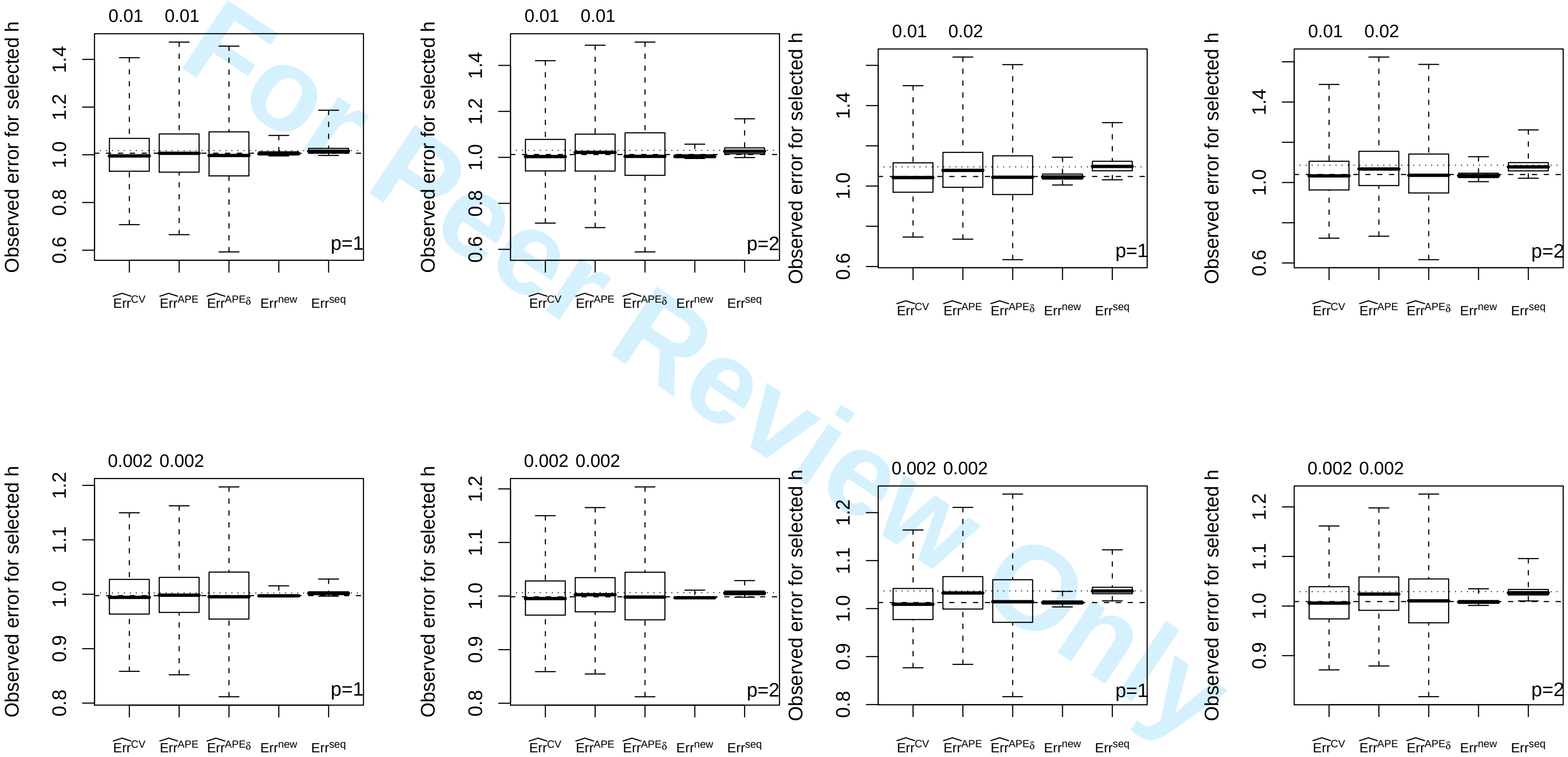


Figure 3: Boxplots for results from Table 1 in the appendix. Values of (1), (2), (3) and (4) for the minimizing h:s. First row $n = 200$, second row $n = 1000$. First two columns $\mu_i = f_1(x_i)$, last two columns $\mu_i = f_2(x_i)$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates $E(Err^{new})$ and $E(Err^{seq})$ for the minimizing smoothing parameter, respectively.

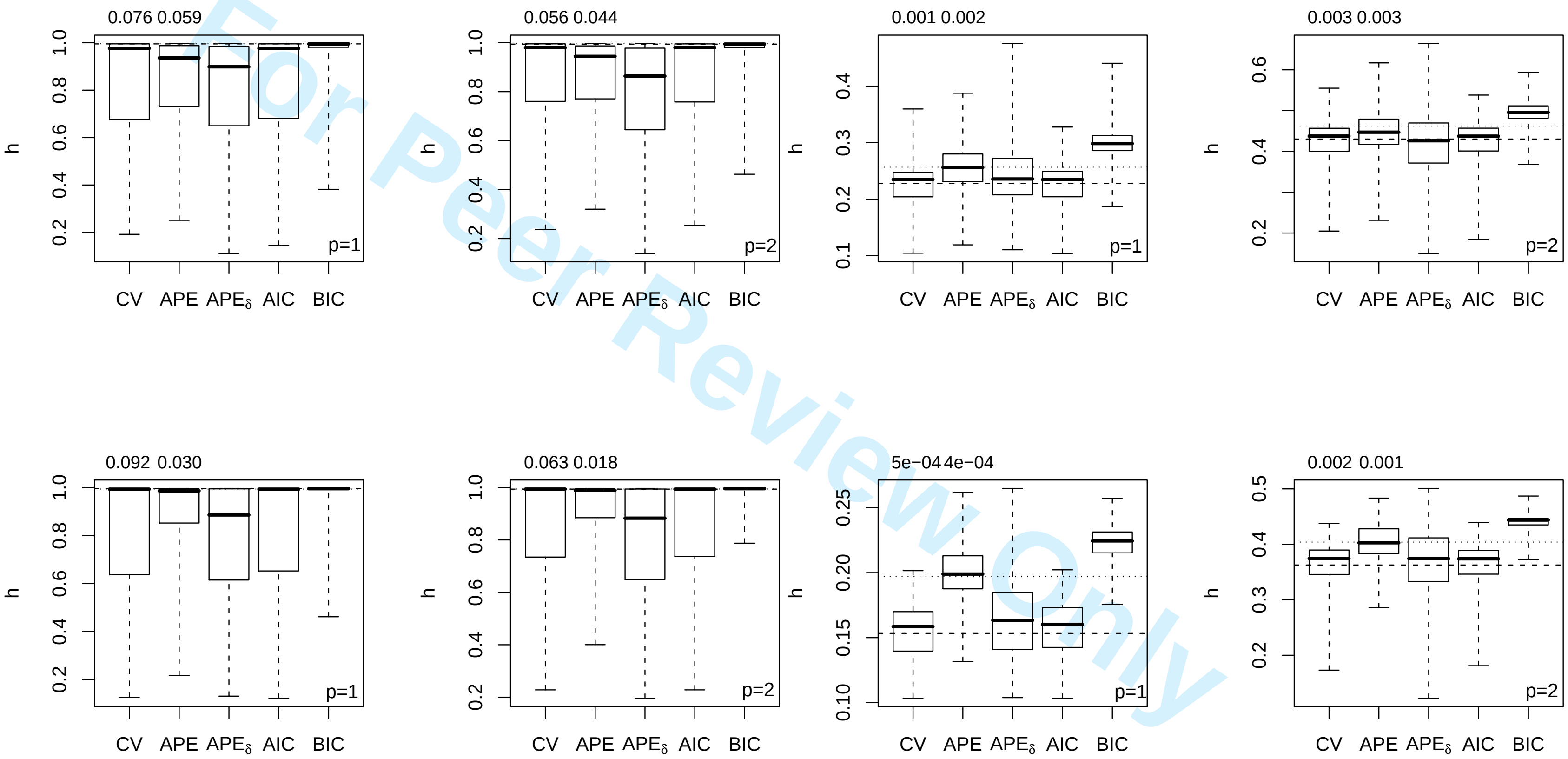


Figure 4: Boxplots for results from Table 3 in the appendix. Values of h minimizing (1), (2), AIC and BIC. First row $n = 200$, second row $n = 1000$. First two columns $\mu_i = f_1(x_i)$, last two columns $\mu_i = f_2(x_i)$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates the smoothing parameter minimizing $E(Err^{new})$ and $E(Err^{seq})$, respectively.

Appendix

Additional figures and tables with results discussed in Section 4.2 are included in the sequel.

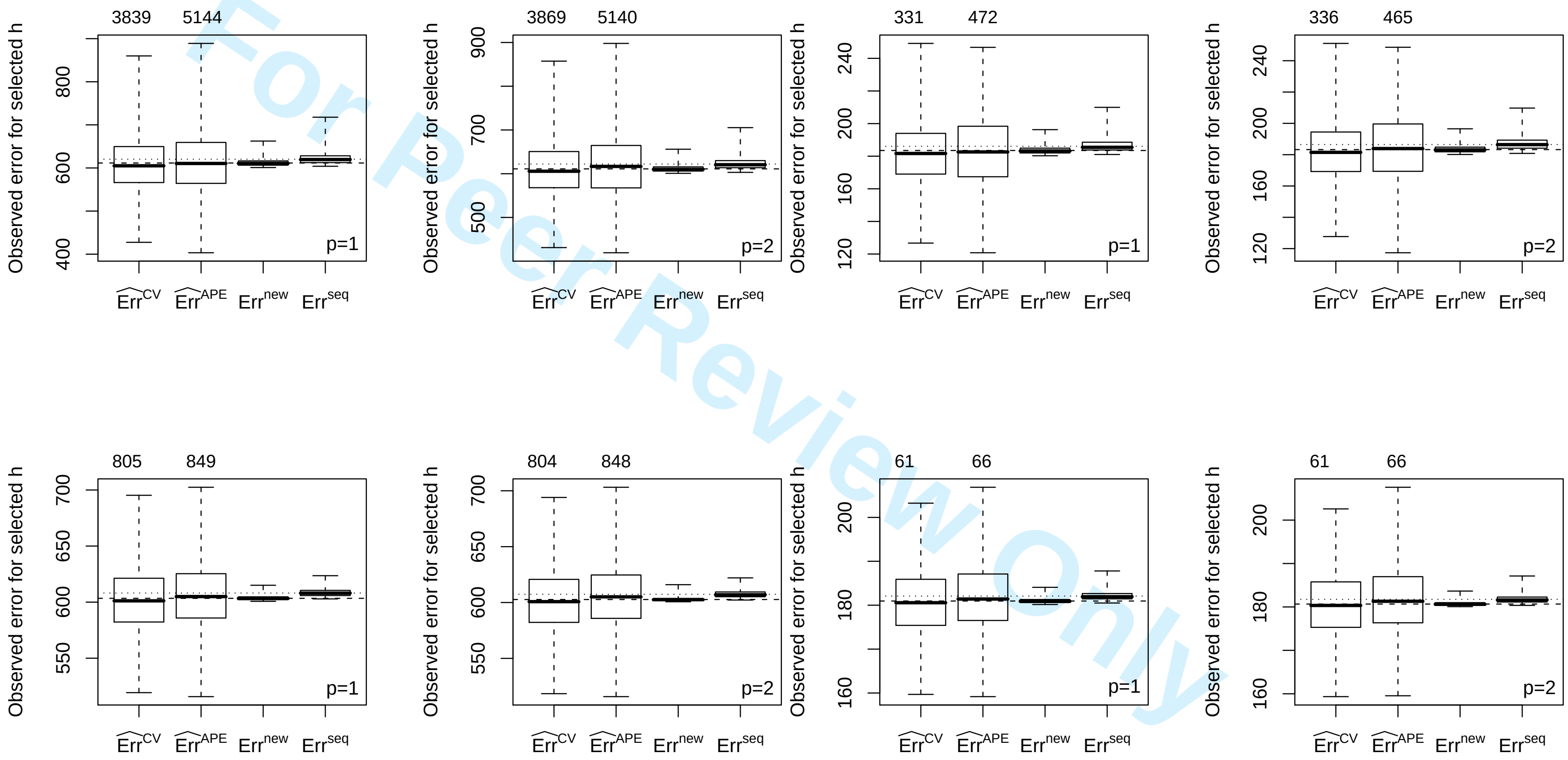


Figure 5: Boxplots for results from Table 2. Values of (1), (2), (3) and (4) for the minimizing h:s . First row $n = 200$, second row $n = 1000$. First two columns $\mu_i = f_3(x_i)$ with $x_i \sim \mathcal{U}$, last two columns $\mu_i = f_3(x_i)$ with $x_i \sim \mathcal{N}$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates $E(Err^{new})$ and $E(Err^{seq})$ for the minimizing smoothing parameter, respectively.

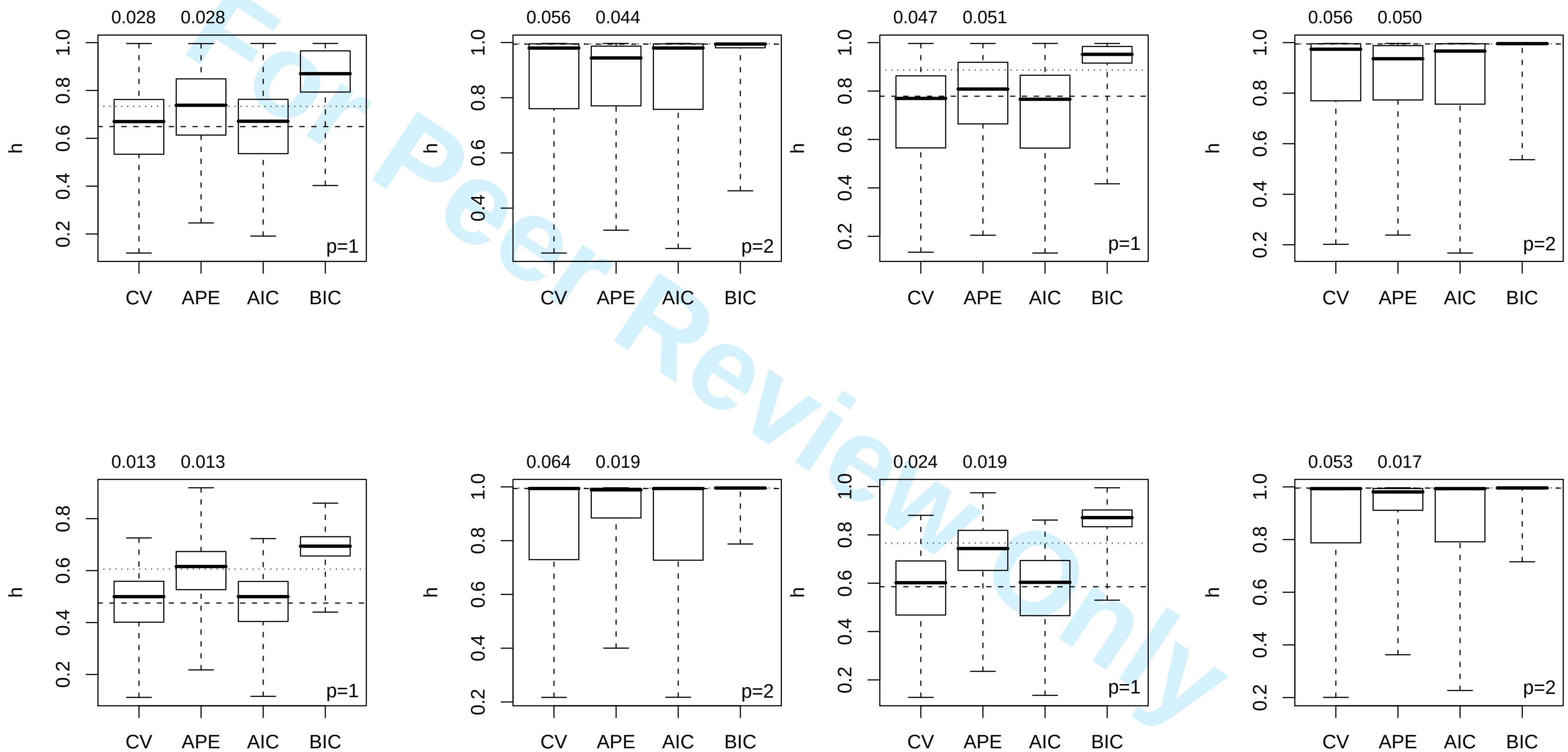


Figure 6: Boxplots for results from Table 4. Values of h *minimizing (1), (2), AIC and BIC. First row* $n = 200$*, second row* $n = 1000$*. First two columns* $\mu_i = f_3(x_i)$ *with* $x_i \sim \mathcal{U}$*, last two columns* $\mu_i = f_3(x_i)$ *with* $x_i \sim \mathcal{N}$*. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates the smoothing parameter minimizing* $E(Err^{new})$ *and* $E(Err^{seq})$*, respectively.*

17

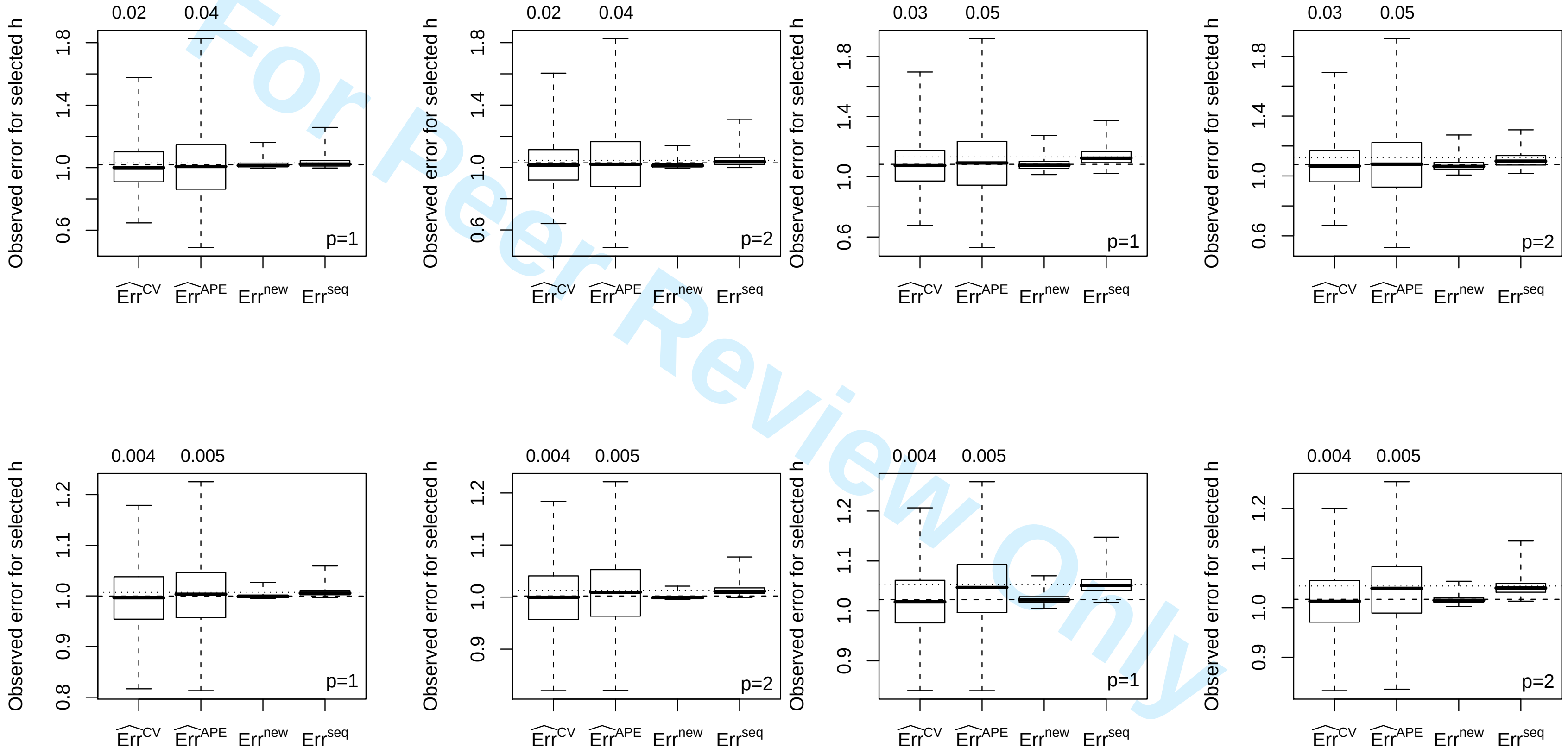


Figure 7: Boxplots for results from Table 1. Values of (1), (2), (3) and (4) for the minimizing h:s. First row $n = 100$, second row $n = 500$. First two columns $\mu_i = f_1(x_i)$, last two columns $\mu_i = f_2(x_i)$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates $E(Err^{new})$ and $E(Err^{seq})$ for the minimizing smoothing parameter, respectively.

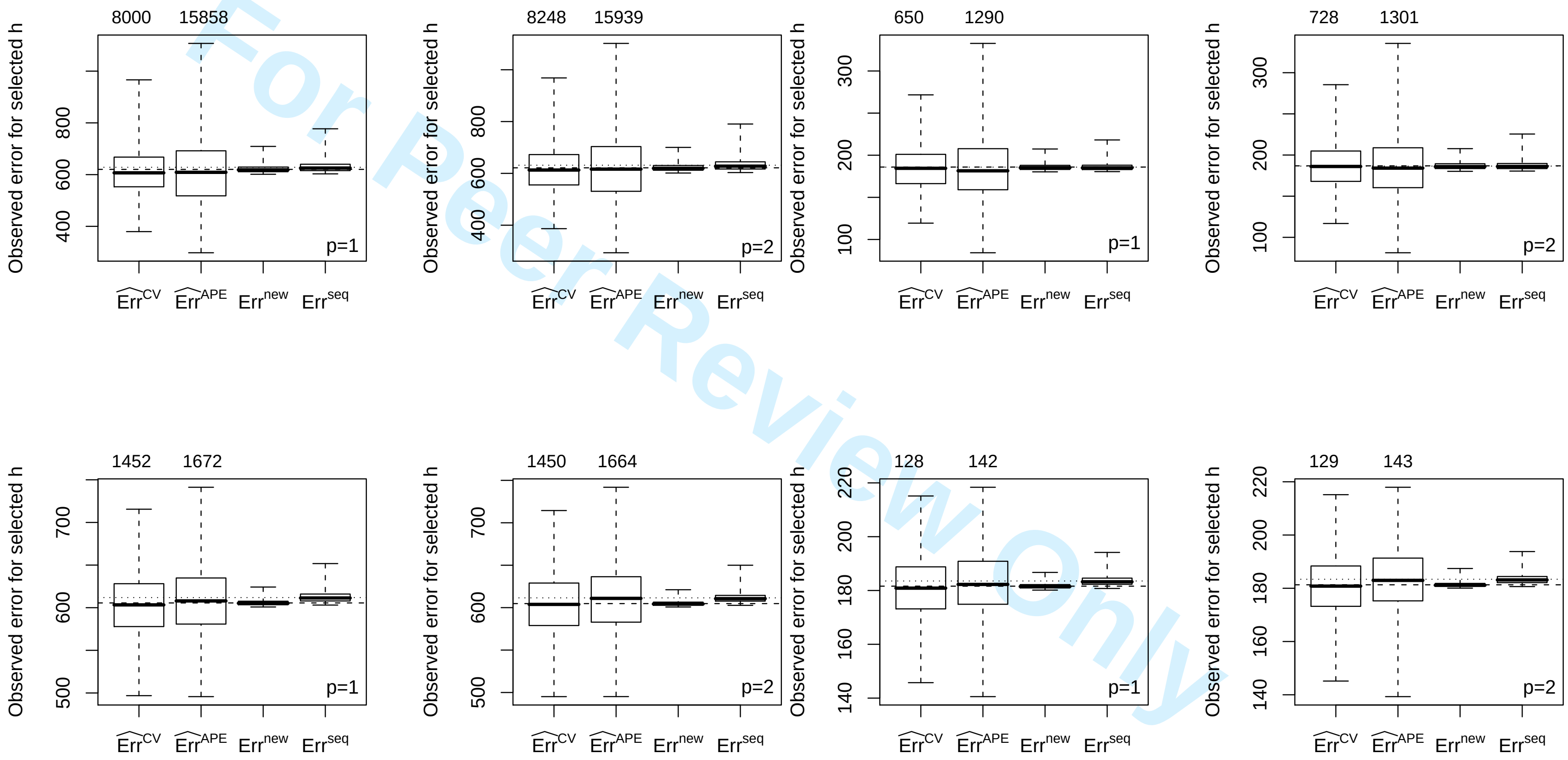


Figure 8: Boxplots for results from Table 2. Values of (1), (2), (3) and (4) for the minimizing h:s. First row $n = 100$, second row $n = 500$. First two columns $\mu_i = f_3(x_i)$ with $x_i \sim \mathcal{U}$, last two columns $\mu_i = f_3(x_i)$ with $x_i \sim \mathcal{N}$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates $E(Err^{new})$ and $E(Err^{seq})$ for the minimizing smoothing parameter, respectively.

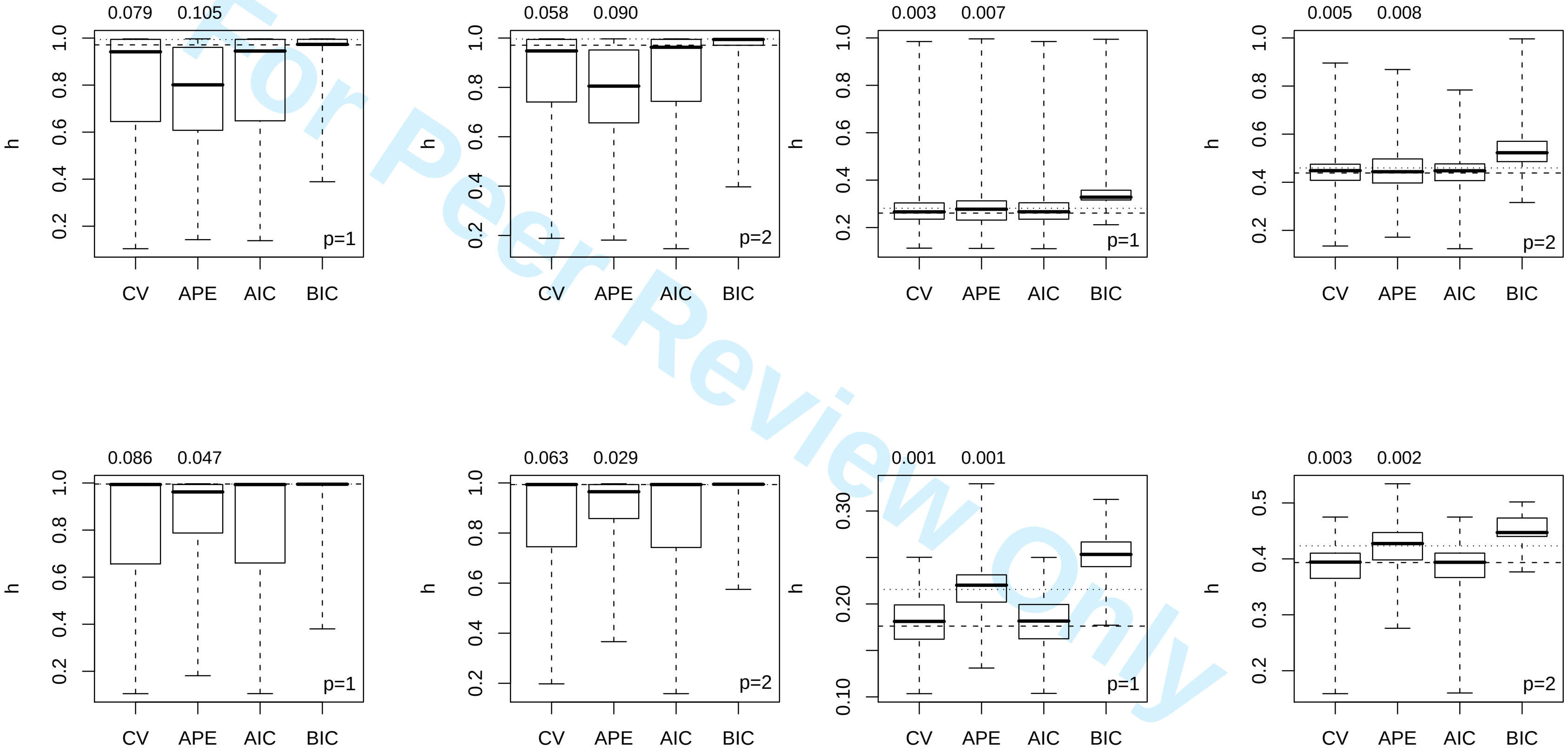


Figure 9: Boxplots for results from Table 3. Values of h minimizing (1), (2), AIC and BIC. First row $n = 100$, second row $n = 500$. First two columns $\mu_i = f_1(x_i)$, last two columns $\mu_i = f_2(x_i)$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates the smoothing parameter minimizing $E(Err^{new})$ and $E(Err^{seq})$, respectively.

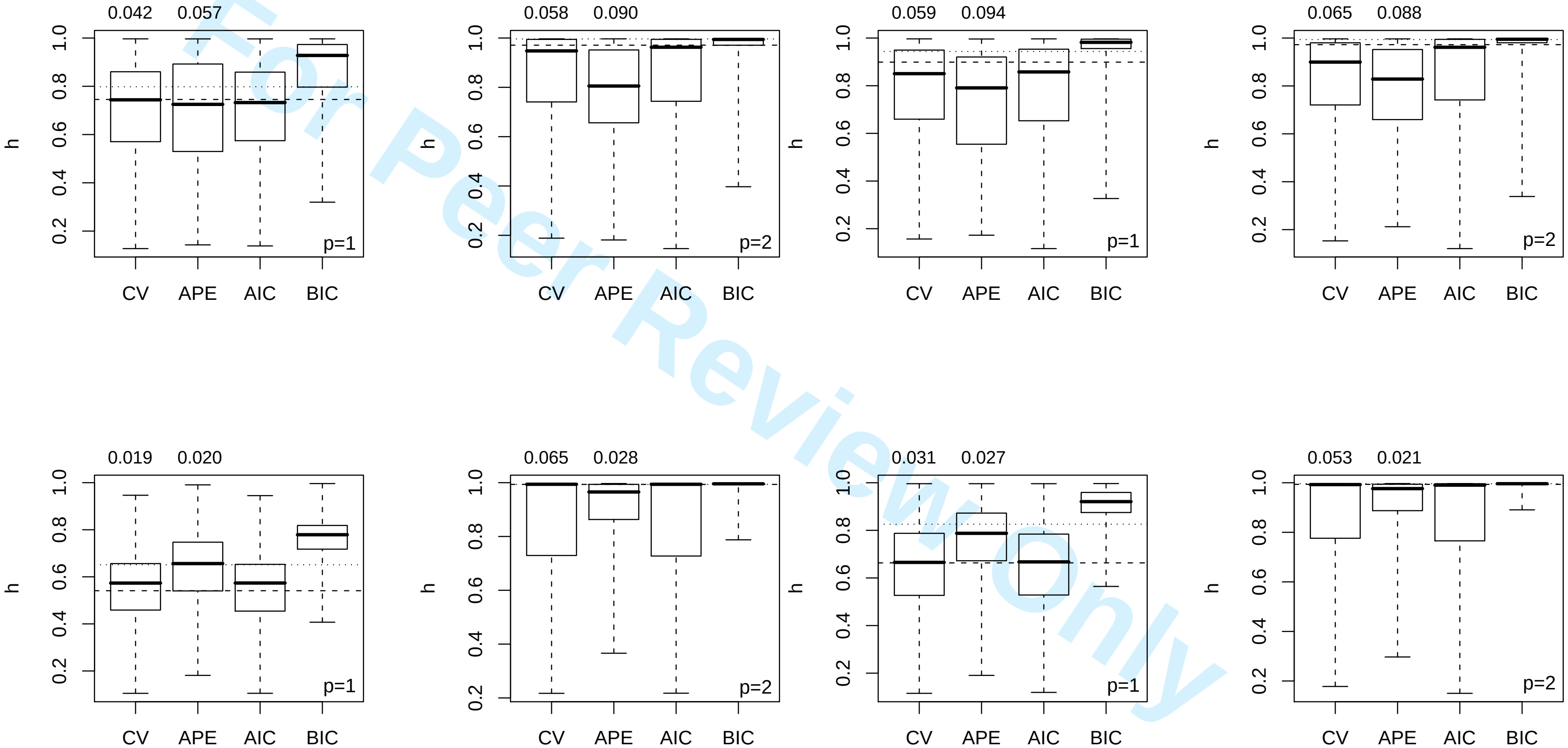


Figure 10: Boxplots for results from Table 4. Values of h minimizing (1), (2), AIC and BIC. First row $n = 100$, second row $n = 500$. First two columns $\mu_i = f_3(x_i)$ with $x_i \sim \mathcal{U}$, last two columns $\mu_i = f_3(x_i)$ with $x_i \sim \mathcal{N}$. MSE for (1) and (2) are given on top. The dashed and dotted lines indicates the smoothing parameter minimizing $E(Err^{new})$ and $E(Err^{seq})$, respectively.

Table 1: *Mean values of (1), (2), (3), (4) and values of (6) and (9), for the minimizing h:s, for f_1, and f_2 with $x_i \sim \mathcal{U}$. Sample standard deviation in parenthesis and bias (difference between the mean of the minima of (1) or (2) and the minimum of (6) or (9)) in brackets.*

n	$\widehat{\text{Err}}^{CV}$	$\widehat{\text{Err}}^{APE}$	Err^{new}	Err^{seq}	$E(\text{Err}^{new})$	$E(\text{Err}^{seq})$
f_1, $x \sim U$, $p = 1$						
100	1.01 (0.15) [-0.01]	1.01 (0.21) [-0.02]	1.02 (0.02)	1.03 (0.03)	1.02	1.03
200	1.00 (0.10) [-0.004]	1.01 (0.12) [-0.01]	1.01 (0.01)	1.02 (0.02)	1.01	1.02
500	1.00 (0.06) [-0.002]	1.00 (0.07) [-0.003]	1.00 (0.01)	1.01 (0.01)	1.00	1.01
1000	1.00 (0.05) [-0.002]	1.00 (0.05) [-0.002]	1.00 (0.002)	1.00 (0.004)	1.00	1.00
f_1, $x \sim U$, $p = 2$						
100	1.03 (0.15) [-0.004]	1.03 (0.21) [-0.02]	1.02 (0.02)	1.05 (0.04)	1.03	1.05
200	1.01 (0.10) [-0.003]	1.03 (0.12) [-0.005]	1.01 (0.01)	1.03 (0.02)	1.01	1.03
500	1.00 (0.06) [-0.002]	1.01 (0.07) [-0.002]	1.00 (0.004)	1.01 (0.01)	1.00	1.01
1000	1.00 (0.05) [-0.002]	1.00 (0.05) [-0.002]	1.00 (0.002)	1.01 (0.01)	1.00	1.01
f_2, $x \sim U$, $p = 1$						
100	1.08 (0.16) [-0.002]	1.10 (0.22) [-0.03]	1.08 (0.04)	1.13 (0.06)	1.08	1.13
200	1.05 (0.11) [-0.002]	1.08 (0.13) [-0.01]	1.05 (0.02)	1.10 (0.04)	1.05	1.10
500	1.02 (0.06) [-0.003]	1.05 (0.07) [-0.01]	1.02 (0.01)	1.05 (0.02)	1.02	1.05
1000	1.01 (0.05) [-0.002]	1.03 (0.05) [-0.003]	1.01 (0.005)	1.04 (0.01)	1.01	1.03
f_2, $x \sim U$, $p = 2$						
100	1.07 (0.16) [-0.001]	1.09 (0.22) [-0.03]	1.07 (0.04)	1.11 (0.05)	1.08	1.12
200	1.04 (0.11) [-0.003]	1.07 (0.12) [-0.01]	1.04 (0.02)	1.08 (0.03)	1.04	1.09
500	1.01 (0.06) [-0.002]	1.04 (0.07) [-0.005]	1.02 (0.008)	1.04 (0.02)	1.02	1.04
1000	1.01 (0.05) [-0.002]	1.03 (0.05) [-0.003]	1.01 (0.004)	1.03 (0.01)	1.01	1.03

Table 2: *Similar to Table 1 but for f_3 with $x_i \sim \mathcal{U}$ and $x_i \sim \mathcal{N}$.*

n	$\widehat{\text{Err}}^{CV}$	$\widehat{\text{Err}}^{APE}$	Err^{new}	Err^{seq}	$E(\text{Err}^{new})$	$E(\text{Err}^{seq})$
f_3, $x \sim U$, $p = 1$						
100	615 (89) [-6]	613 (125) [-15]	622 (15)	631 (21)	620	628
200	608 (62) [-3]	614 (71) [-6]	613 (8)	622 (12)	612	620
500	604 (38) [-2]	609 (41) [-3]	606 (4)	613 (6)	606	612
1000	602 (28) [-1]	606 (29) [-2]	604 (2)	609 (3)	603	608
f_3, $x \sim U$, $p = 2$						
100	619 (91) [-2]	622 (126) [-10]	624 (16)	633 (22)	622	632
200	609 (62) [-2]	619 (72) [-3]	612 (8)	624 (14)	611	622
500	605 (38) [-0.2]	611 (41) [-1]	605 (3)	612 (6)	605	612
1000	602 (28) [-1]	606 (29) [-1]	603 (2)	608 (3)	603	607
f_3, $x \sim N$, $p = 1$						
100	184 (25) [-2]	184 (36) [-2]	186 (4)	186 (5)	186	186
200	182 (18) [-2]	184 (22) [-2]	184 (2)	187 (4)	184	186
500	181 (11) [-0.5]	183 (12) [-1]	182 (1)	184 (2)	182	184
1000	181 (8) [-0.2]	182 (8) [-0.3]	181 (1)	182 (1)	181	182
f_3, $x \sim N$, $p = 2$						
100	187 (27) [0.2]	186 (36) [-1]	187 (5)	187 (5)	187	187
200	182 (18) [-1]	185 (22) [-1]	184 (2)	187 (4)	183	187
500	181 (11) [-0.2]	183 (12) [-0.1]	181 (1)	184 (2)	181	183
1000	181 (8) [-0.2]	182 (8) [-0.1]	181 (1)	182 (1)	181	182

Table 3: *Mean of the h:s minimizing (1), (2), AIC, BIC, (6)and (9) for f_1 and f_2 with $x_i \sim \mathcal{U}$. Sample standard deviation in parenthesis and bias (difference between the mean of the minimizing h:s and h minimizing (6) or (9)) in brackets.*

n	CV	APE	AIC	BIC	$E(\mathrm{Err}^{new})$	$E(\mathrm{Err}^{seq})$
f_1, $x \sim U$, $p = 1$						
100	0.807	0.756	0.810	0.955	0.971	0.994
	(0.228)	(0.221)	(0.229)	(0.086)		
	[-0.164]	[-0.237]				
200	0.833	0.839	0.834	0.974	0.995	0.996
	(0.224)	(0.187)	(0.222)	(0.075)		
	[-0.162]	[-0.156]				
500	0.828	0.869	0.830	0.988	0.996	0.996
	(0.240)	(0.175)	(0.237)	(0.045)		
	[-0.168]	[-0.127]				
1000	0.816	0.903	0.817	0.991	0.996	0.994
	(0.244)	(0.147)	(0.244)	(0.035)		
	[-0.180]	[-0.090]				
f_1, $x \sim U$, $p = 2$						
100	0.836	0.771	0.839	0.970	0.971	0.996
	(0.200)	(0.199)	(0.209)	(0.063)		
	[-0.135]	[-0.225]				
200	0.860	0.861	0.858	0.982	0.994	0.996
	(0.195)	(0.160)	(0.197)	(0.049)		
	[-0.134]	[-0.134]				
500	0.856	0.897	0.854	0.993	0.994	0.994
	(0.209)	(0.141)	(0.210)	(0.022)		
	[-0.138]	[-0.097]				
1000	0.856	0.924	0.855	0.995	0.994	0.994
	(0.209)	(0.116)	(0.210)	(0.010)		
	[-0.138]	[-0.070]				
f_2, $x \sim U$, $p = 1$						
100	0.267	0.280	0.266	0.358	0.261	0.281
	(0.053)	(0.085)	(0.055)	(0.113)		
	[0.006]	[-0.001]				
200	0.224	0.255	0.225	0.300	0.228	0.257
	(0.037)	(0.041)	(0.037)	(0.027)		
	[-0.004]	[-0.001]				
500	0.179	0.218	0.179	0.252	0.176	0.216
	(0.028)	(0.026)	(0.028)	(0.017)		
	[0.003]	[0.002]				
1000	0.154	0.200	0.157	0.223	0.153	0.197
	(0.022)	(0.020)	(0.022)	(0.011)		
	[0.001]	[0.003]				
f_2, $x \sim U$, $p = 2$						
100	0.438	0.449	0.435	0.533	0.439	0.460
	(0.069)	(0.090)	(0.071)	(0.070)		
	[-0.001]	[-0.010]				
200	0.423	0.447	0.422	0.494	0.430	0.462
	(0.055)	(0.056)	(0.056)	(0.030)		
	[-0.008]	[-0.015]				
500	0.380	0.422	0.380	0.455	0.393	0.423
	(0.051)	(0.040)	(0.051)	(0.021)		
	[-0.013]	[-0.001]				
1000	0.364	0.403	0.364	0.438	0.363	0.404
	(0.041)	(0.032)	(0.041)	(0.015)		
	[0.001]	[-0.001]				

Table 4: *Similar to Table but 3 but for f_3 with $x_i \sim \mathcal{U}$ and $x_i \sim \mathcal{N}$.*

n	CV	APE	AIC	BIC	$E(\text{Err}^{new})$	$E(\text{Err}^{seq})$
f_3, $x \sim U$, $p = 1$						
100	0.704 (0.201) [-0.042]	0.699 (0.217) [-0.099]	0.701 (0.202)	0.882 (0.120)	0.745	0.798
200	0.646 (0.169) [-0.003]	0.717 (0.167) [-0.017]	0.648 (0.169)	0.862 (0.114)	0.649	0.734
500	0.548 (0.136) [0.007]	0.640 (0.142) [-0.012]	0.546 (0.138)	0.765 (0.092)	0.541	0.651
1000	0.474 (0.115) [-0.001]	0.601 (0.113) [-0.005]	0.474 (0.114)	0.690 (0.064)	0.475	0.606
f_3, $x \sim U$, $p = 2$						
100	0.836 (0.200) [-0.135]	0.771 (0.199) [-0.225]	0.839 (0.209)	0.970 (0.063)	0.971	0.996
200	0.860 (0.195) [-0.134]	0.861 (0.160) [-0.135]	0.858 (0.197)	0.982 (0.049)	0.994	0.996
500	0.851 (0.213) [-0.142]	0.900 (0.140) [-0.094]	0.849 (0.215)	0.993 (0.023)	0.994	0.994
1000	0.854 (0.211) [-0.140]	0.924 (0.116) [-0.072]	0.853 (0.212)	0.995 (0.010)	0.994	0.996
f_3, $x \sim N$, $p = 1$						
100	0.778 (0.211) [-0.121]	0.733 (0.222) [-0.210]	0.774 (0.218)	0.955 (0.077)	0.899	0.944
200	0.705 (0.203) [-0.074]	0.769 (0.192) [-0.117]	0.704 (0.204)	0.935 (0.071)	0.779	0.887
500	0.647 (0.175) [-0.017]	0.763 (0.150) [-0.063]	0.646 (0.175)	0.911 (0.064)	0.663	0.826
1000	0.575 (0.156) [-0.011]	0.724 (0.132) [-0.042]	0.577 (0.156)	0.863 (0.062)	0.585	0.766
f_3, $x \sim N$, $p = 2$						
100	0.818 (0.203) [-0.155]	0.774 (0.201) [-0.219]	0.839 (0.204)	0.978 (0.055)	0.972	0.994
200	0.859 (0.194) [-0.135]	0.8543 (0.173) [-0.141]	0.852 (0.198)	0.990 (0.030)	0.995	0.995
500	0.867 (0.192) [-0.126]	0.916 (0.122) [-0.080]	0.863 (0.196)	0.995 (0.007)	0.993	0.996
1000	0.871 (0.193) [-0.124]	0.926 (0.111) [-0.071]	0.872 (0.192)	0.995 (0.011)	0.996	0.996

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*, 596–610.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: the prequential approach (with discussion). *Journal of the Royal Statistical Society Series A 147*, 278–292.

de Luna, X. and M. G. Genton (2005). Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica 15*, 547–568.

de Luna, X. and K. Skouras (2003). Choosing a model selection strategy. *Scandinavian Journal of Statistics 30*, 113–128.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association 78*, 316–331.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association 99*, 619–642.

Hall, P. and I. Johnstone (1992). Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society B 54*, 475–530.

Hastie, T. and R. J. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall, London.

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics 35*, 1238–1277.

Li, Q. and J. Racine (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica 14*, 485–512.

Modha, D. S. and E. Masry (1998). Prequential and cross-validated regression estimation. *Machine Learning 33*, 5–39.

Ouyang, D., D. Li, and Q. Li (2006). Cross-validation and non-parametric k-nearest-neighbour estimation. *Econometrica Journal 9*, 448–471.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rissanen, J. (1986). Order of estimation by accumulated prediction errors. *Journal of Applied Probability 23A*, 55–61.

Ruppert, D., M. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

Ruppert, D., M. P. Wand, U. Holst, and O. Hössjer (1997). Local polynomial variance function estimation. *Technometrics 39*, 262–273.

Schimek, M. G. (Ed.) (2000). *Smoothing and Regression: Approaches, Computation and Application.* John Wiley & sons, New York.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica 7*, 221–264.

Sjöstedt (2000). Resampling m-dependent random variables with applications to forecasting. *Scandinavian Journal of Statistics 27*, 543–562.

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

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

Wagenmakers, E.-J., P. D. Grünwald, and M. Steyvers (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology 50*, 149–166.