Cross-Validatory Bandwidth Selections For Regression Estimation Based on Dependent Data*

Qiwei Yao and Howell Tong

Institute of Mathematics and Statistics
University of Kent at Canterbury
Canterbury, Kent CT2 7NF, U.K.

Abstract

We suggest a simple and fast method to determine the bandwidth in kernel regression. The method can be viewed as a generalized cross-validation. We have proved asymptotic optimality of the proposed bandwidth selector under the assumption that the observations are strictly stationary and ρ-mixing. Simulation has been conducted to compare the performance of various cross-validation bandwidth selectors applied to dependent data, which shows that the ordinary cross-validation method is quite stable in regression estimation with random design even when the data are highly correlated.

Some key words: bandwidth; cross-validation; kernel estimation; locally linear regression; ρ-mixing.

Abbreviated Title: The cross-validation bandwidth selection

*Research partially supported by the EC Human Capital and Mobility Contract ERB CHRX-CT 94-0963.

1 Introduction

The goal of this paper is two-fold. First, we propose a new method to determine the bandwidth in kernel regression. Second, a simulation study has been conducted to compare the performance of various versions of cross-validation bandwidth selectors in regression estimation with dependent data.

Of great importance in nonparametric kernel regression is the choice of the smoothing parameter h, also called the bandwidth. Suppose that we have n observations. Once h is specified, the number of data that lie within one bandwidth of a given point is of the size nh. Effectively, only these observations are used to estimate the regression function at this point. Intuitively, a large bandwidth will lead to a very smooth estimator which inevitably incurs large bias. On the other hand, a small bandwidth might reduce the bias, but the variability of the estimated curve could be large since only a few data are used in the estimation. A good choice of h should be a tradeoff between these two types of drawback. Therefore, at the core of most methods for selecting h is the minimization of the mean squared error of the estimator (in an appropriate way). All the methods discussed in this paper share this characteristic.

The choice of the smoothing parameter will always be influenced by the purpose for which the parameter is to be used. For example, a good bandwidth for estimating an unknown curve is not necessarily good for prediction (cf. Silverman 1986, Härdle 1990). Nevertheless, an automatically selected bandwidth is often a useful starting point. The most frequently used technique for automatic bandwidth selection is the cross validation.

In this paper, we propose a new method to select the bandwidth as follows. We use the first m sample points to estimate the unknown function and choose the bandwidth such that the estimator gives the minimum mean squared predictive errors (MSPE) for the last $(n-m)$ sample points. Then we *reduce* this preliminary bandwidth appropriately when applied to the entire sample. The appropriate reduction (to be described in §2.2 below) will be based on some theoretical considerations, in a manner similar to Marron's (1987) partitioned cross-validation approach. We shall prove that the bandwidth selected in the proposed way converges to the theoretical optimal bandwidth (see Theorem 1 in §2.3 below). Perhaps the most obvious drawback of the method is that the data have not been used in the most efficient way, compared with the ordinary cross-validation method. On the other hand, it saves considerable computing time relative to other cross-validation methods. Numerical examples show that in terms of the integrated squared error (ISE) of the estimator curve, the proposed method is not as appealing as the "leave-$(2l+1)$-out" versions of cross-validation for $l = 0, 1, \ldots$, due to the afore-mentioned inefficiency of the former in using the data. However, it gives competitive performance in prediction (see §3 below).

On the other hand, it has been well documented that if the observations are dependent, the cross validation will not always produce good bandwidths. For instance, if the errors of the model are positively correlated, the cross-validation will produce small bandwidths which result in rough kernel estimates of regression functions (See, for example, Altman 1990, Hart 1991, 1994). In order to cope with possible dependence among data, it has been suggested that the cross validation, if retained, will require some modification. For example, the modification could simply be the "leave-($2l$+1)-out" version of cross-validation (cf. Hart and Vieu 1990, Chu and Marron 1991, Härdle and Vieu 1992). Intuitively, the stronger is the dependence, the larger is the value of l needed. However, a further scrutiny into the above observation is needed. For regression models, such as models with fixed designs in which the nearest neighbours in the *time space* of an observed point are also its nearest neighbours in the *state space*, leaving more than one out in the cross validation leads to a significant improvement when the data are dependent (cf. Altman 1990, Chu and Marron 1991, Hart 1991 and 1994). However, for general regression models in which the regressors are random, we argue that it is unclear whether a "leave-($2l$+1)-out" version of cross-validation method is still appealing (see Remark 2 in §3). Simulation studies indicate that the ordinary cross-validation method is reasonably safe to use.

The paper is organized as follows. §2 deals with the bandwidth selector. Although we restrict our discussion to the locally linear regression, the method can be readily applied to other kernel regression estimations. §3 reports the results of simulation studies on two stochastic regression models and a time series model. Some asymptotic results and their proofs are given in §4.

2 Bandwidth selectors

2.1 Locally linear regression

Suppose that $\{X_t,\ Y_t\}$ is a strictly stationary discrete-time process and both X_t and Y_t are scalar. It is of interest to estimate the regressive function $f(x) = \mathrm{E}\{Y_1 | X_1 = x\}$. In the case that $X_t = Y_{t-1}$, $f(.)$ is the autoregressive function for the time series $\{Y_t\}$. Given the observations $\{(X_t,\ Y_t);\ 1 \leq t \leq n\}$, one of the conventional nonparametric estimators of $f(x)$ is the Nadaraya-Watson kernel regression estimator, which can be viewed as the minimizer of the following least squares problem

$$\sum_{t=1}^{n} \{Y_t - a\}^2 K\left(\frac{X_t - x}{h}\right),$$

where $K(.)$ is a kernel function and $h > 0$ is the bandwidth. However, if the derivative of f at the point x exists, by Taylor's expansion, we have

$$f(z) \approx f(x) + \dot{f}(x)(z - x).$$

This suggests the locally linear regression estimator: $\hat{f}_{n,h}(x) = \hat{a}$, where $(\hat{a}, \hat{b})$ minimize

$$\sum_{t=1}^{n} \{Y_t - a - b(X_t - x)\}^2 K\left(\frac{X_t - x}{h}\right).$$

As a by-product, $\hat{b}$ is a natural estimator for $\dot{f}(x)$. It has been pointed out that the locally linear regression method has various advantages over other popular kernel methods, e.g. the Nadaraya-Watson method (see, for example, Fan 1992). Further, Fan *et al.* (1993) proposed the use of locally polynomial fitting if the estimators for the derivatives of $f(.)$ are also of interest.

Let $\hat{\beta}_n = (\hat{f}_{n,h}(x), \hat{\dot{f}}_{n,h}(x))^{\tau}$. The least squares theory gives

$$\hat{\beta} = (X^{\tau} W X)^{-1} X^{\tau} W Y, \tag{2.1}$$

where $Y = (Y_1, \ldots, Y_n)^{\tau}$, $W = \text{diag}(K(\frac{X_1 - x}{h}), \ldots, K(\frac{X_n - x}{h}))$, and X is an $n \times 2$ matrix with $(1, X_i - x)$ as the i-th row. More specifically,

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{t=1}^{n} W_{n,1}\left(\frac{X_t - x}{h}, x\right) Y_t, \quad \hat{\dot{f}}_{n,h}(x) = \frac{1}{nh^2} \sum_{t=1}^{n} W_{n,2}\left(\frac{X_t - x}{h}, x\right) Y_t, \tag{2.2}$$

where

$$W_{n,1}(t, x) = (1, 0) S_n^{-1}(x)(1, t)^{\tau} K(t), \tag{2.3}$$

$$W_{n,2}(t, x) = (0, 1) S_n^{-1}(x)(1, t)^{\tau} K(t),$$

and $S_n(x)$ is a 2×2 matrix with the (i, j)-th element $s_{i+j-2}(x)$, and

$$s_j(x) = \frac{1}{nh} \sum_{t=1}^{n} \left(\frac{X_t - x}{h}\right)^j K\left(\frac{X_t - x}{h}\right). \tag{2.4}$$

2.2 Bandwidth selection

Before introducing a new method for choosing bandwidths for locally linear regression with dependent data, let us look at the theoretical optimal bandwidth which minimizes the Mean Squared Error (MSE) of the estimator $\hat{f}_n(x)$. It follows from Theorem 2 in Section 4 below that

$$\mathrm{E}\{\hat{f}_{n,h}(x) - f(x)\}^2 \approx \frac{h^4}{4} \sigma_0^4 \{\ddot{f}(x)\}^2 + \frac{1}{nhp(x)} \sigma^2(x) \int K^2(u) \mathrm{d}u,$$

where $\sigma_0^2 = \int u^2 K(u) \mathrm{d}u$, $\sigma^2(x) = \mathrm{Var}(Y_1 | X_1 = x)$, $p(x)$ is the marginal density function of X_1. To minimize the above approximate expression of MSE, the bandwidth should be chosen as a function of x and should be equal to

$$h_n(x) = \frac{1}{n^{1/5}} \left(\frac{\sigma^2(x) \int K^2(u) \mathrm{d}u}{p(x) \sigma_0^4 \{\ddot{f}(x)\}^2} \right)^{1/5}. \tag{2.5}$$

The above bandwidth cannot be directly applied in practice because it depends on various unknown functions. However, it does indicate that a reasonable bandwidth is such that

$$h \propto n^{-\frac{1}{5}}, \tag{2.6}$$

which motivates the following proposal for choosing h.

We split the sample into two pieces $\{(X_t, Y_t),\ 1 \le t \le m\}$, and $\{(X_t, Y_t),\ m < t \le n\}$. We estimate $f(.)$ using the first m observations. The estimator given by (2.2) is denoted as $\hat{f}_{m,h}(.)$. We choose h such that $\hat{f}_{m,h}(.)$ gives the best prediction for Y_t for $m < t \le n$ in the sense that $h = \hat{h}_m$ minimizes

$$\mathrm{ECV}_m(h) = \frac{1}{n-m} \sum_{t=m+1}^{n} \{Y_t - \hat{f}_{m,h}(X_t)\}^2 w(X_t), \tag{2.7}$$

over H_m, where $w(.)$ is a weight function, and

$$H_m = [a m^{-\frac{1}{5} - \varepsilon_0},\ b m^{-\frac{1}{5} + \varepsilon_0}], \tag{2.8}$$

where $0 < a < b < \infty$, and $\varepsilon_0 \in (0, \frac{1}{150})$ are some constants. In the light of (2.6), for the estimator $\hat{f}_{n,h}$ which is based on the whole sample, we use the bandwidth

$$\hat{h}_n = \hat{h}_m \left(\frac{m}{n} \right)^{1/5}. \tag{2.9}$$

The above approach could still be viewed as a generalization of the cross validation. We leave out the last $n - m$ observations for validation. In fact, the proposed method is fast. For example, the ordinary cross-validation method entails $n(n-1)$ kernel evaluations while the proposed method requires only $m(n-m)$ kernel evaluations. For $m = \frac{2}{3}n$, it is over 4 times faster than the ordinary cross-validation method.

Equation (2.9) gives a constant bandwidth over the sample space, which can be sufficient if $f(.)$ is not very ‘wiggly’. Clearly such a bandwidth will fail to do a good job if the unknown curve has a rather complicated structure; this is usually the case in multi-step prediction, especially when the underlying model (skeleton) is chaotic. In order to capture the complexity of such a curve, a variable bandwidth is necessary. The following modification to (2.7) and (2.9) can be applied for this purpose. Suppose we want to estimate f at x. First we estimate $f(.)$ using the first m sample points with bandwidth

$$\hat{h}_m(x) = \text{argmin}_{h \in H_m} \frac{1}{n-m} \sum_{t=m+1}^{n} \{Y_t - \hat{f}_{m,h}(X_t)\}^2 w_m(X_t - x), \tag{2.10}$$

where $\hat{f}_{m,h}$ is given as in (2.2) but with m instead of n, $w_m(.)$ is a weight function, and $w_m(x) \to 0$ as $m \to \infty$ for any $|x| \neq 0$. It is intuitively clear that we may set

$$\hat{h}_n(x) = \hat{h}_m(x) \left(\frac{m}{n}\right)^{1/5} \tag{2.11}$$

as the bandwidth for estimating $f(x)$ when using the whole sample. We shall justify this later. Further spatial adaption can be carried out in the spirit of Fan and Gijbels (1995).

2.3 Asymptotical optimality

We proceed with the case of a variable bandwidth $\hat{h}_n(x)$ of (2.11), since it is technically more involved. We make some remarks on the case of a constant bandwidth $\hat{h}_n$ of (2.9) whenever appropriate.

Note that the best (pointwise) prediction for Y_t based on X_t is $f(X_t) = \text{E}\{Y_t|X_t\}$. To justify the above approach, we compare the $\hat{h}_n(x)$ with the bandwidth which minimizes the average squared errors of the *fictitious* post samples $\{(X_t, Y_t),\ t = n+1, \ldots, n+m'\}$

$$M_n(x,h) = \frac{1}{m'} \sum_{t=n+1}^{n+m'} \{\hat{f}_{n,h}(X_t) - f(X_t)\}^2 w_n(X_t - x). \tag{2.12}$$

If m'/n converges to a positive constant, Theorem 3 (ii) (in §4 below) shows that

$$M_n(x,h) \sim \frac{h^4}{4} \sigma_0^4 \{\ddot{f}(x)\}^2 p(x) + \frac{1}{nh} \sigma^2(x) \int K^2(u) \text{d}u. \tag{2.13}$$

The minimizer of the RHS of the above expression is $h_n(x)$ given as in (2.5). The following theorem shows that the proposed method yields an estimate which is asymptotically equivalent to $h_n(x)$ in an appropriate sense.

Theorem 1. Let $w_n(.) = b^{-1} w(./b)$, where $w(.)$ is a density function with a bounded support, and $b^{-1} = O(n^{\varepsilon_0/2})$ as $n \to \infty$. Assume m is chosen in such a way that both m/n and $(n-m)/n$ converge to some positive constants as $n \to \infty$. Then, under conditions (A1) – (A6) given in §4 below, for $x \in \{p(x) > 0\}$,

$$\frac{1}{h_n(x)} |\hat{h}_n(x) - h_n(x)| \xrightarrow{\mathcal{P}} 0,$$

where $\hat{h}_n(x)$ is given as in (2.11).

Proof of Theorem 1. We notice that

$$M_m(x,h) = \frac{1}{n-m} \sum_{t=m+1}^{n} \{\hat{f}_{m,h}(X_t) - f(X_t)\}^2 w_m(X_t - x). \tag{2.14}$$

By Theorem 3 (i), $M_m(x,h)$ has a similar asymptotic expansion as (2.13). The minimizer of the asymptotic expansion is $h_m(x) \equiv (n/m)^{1/5} h_n(x)$. By (2.11) and (2.5), we only need to prove that

$$m^{\frac{1}{5}} |\hat{h}_m(x) - h_m(x)| \xrightarrow{\mathcal{P}} 0, \tag{2.15}$$

where $\hat{h}_m(x)$ is defined as in (2.10). Let

$$D(x,h) = \frac{(m^{\frac{1}{5}} h)^4}{4} \sigma_0^4 \{\ddot{f}(x)\}^2 p(x) + \frac{1}{m^{\frac{1}{5}} h} \sigma^2(x) \int K^2(u) \mathrm{d}u.$$

For any fixed x, $D(x,h)$ has a unique minimum value at $h_m(x)$ over $h \in H_m$, where H_m is defined as in (2.8). Further, $D_0 \equiv D(x, h_m(x))$ is a positive constant independent of m. It is also easy to see that for any given $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$D(x,h) \geq D_0 + \delta, \quad \text{for all } m^{\frac{1}{5}} |h - h_m(x)| \geq \varepsilon, \tag{2.16}$$

and $\delta > 0$ is also independent of m.

Now, suppose that $m^{\frac{1}{5}} |\hat{h}_m(x) - h_m(x)|$ does not converge to 0 in probabiltiy. Then there exists a subsequence of $\{m\}$, say $\{m'\}$, for which

$$P\{(m')^{\frac{1}{5}} |\hat{h}_{m'}(x) - h_{m'}(x)| \geq \varepsilon\} > \eta > 0,$$

for all $m' > m_0' > 0$, where η and m_0' are constants. It follows from (2.16) that for all $m' > m_0'$,

$$P\left\{ \frac{D(x, \hat{h}_{m'}(x))}{D(x, h_{m'}(x))} \geq 1 + \frac{\delta}{D_0} \right\} > \eta.$$

Note that $M_m(x,h) = m^{-4/5} D(x,h)\{1 + o_p(1)\}$ uniformly for $h \in H_m$ (cf. Theorem 3 (i) in §4 below). Therefore, for all sufficiently large m',

$$P\left\{ \frac{M_m(x, \hat{h}_{m'}(x))}{M_m(x, h_{m'}(x))} \geq 1 + \frac{\delta}{D_0} \right\} > \eta/2 > 0,$$

which contradicts Theorem 4 in §4 below. Hence, $m^{\frac{1}{5}} |\hat{h}_{m'}(x) - h_{m'}(x)| \xrightarrow{\mathcal{P}} 0$.

Remark 1. For $\hat{h}_n$ given in (2.9), if we choose $w(.)$ as a density function with a bounded support contained in the support of $p(.)$, then it can be proved that under the conditions of Theorem 1, $h_n^{-1}(\hat{h}_n - h_n) \xrightarrow{\mathcal{P}} 0$, where

$$h_n = \frac{1}{n^{1/5}} \left\{ \frac{\int K^2(u)\mathrm{d}u \int \sigma^2(x)w(x)\mathrm{d}x}{\sigma_0^4 \int \{\ddot{f}(x)\}^2 p(x)w(x)\mathrm{d}x} \right\}^{\frac{1}{5}}. \tag{2.17}$$

3 Simulation studies

Three simulated models have been used to illustrate the finite sample behaviour of the newly proposed method, and also the 'leave-k-out' versions of the cross-validation method, simply referred to as the CV(k) method in the discussion below, for $k = 1, 3, 5, 7$, and 9. The first model has the independent observations from the regressor but auto-correlated errors. The second model has the auto-correlated observations from the regressor but independent errors. The third model is a nonlinear AR(1) process. We compare all the above selectors with the bandwidth h_{opt} which minimizes the integrated squared error

$$\mathrm{ISE}(h) = \int \{\hat{f}_{n,h}(x) - f(x)\}^2 w(x)\mathrm{d}x, \tag{3.1}$$

for simulated data, where $\hat{f}_{n,h}(.)$ is defined as (2.2), and $w(.)$ is a weight function. We choose $w(.)$ as the indicator function of the (estimated) 95% inner sample space of X_1. The Gaussian kernel is used throughout the calculation. The mean squared predictive errors for 100 post sample points based on

$$\mathrm{MSPE}(h) = \frac{1}{100}\sum_{t=1}^{100}\{Y_{n+t} - \hat{f}_{n,h}(X_{n+t})\}^2$$

are also calculated for each of the data-driven bandwidths. We always set $n = 300$ for the sample size, $m = 200$ for the estimation of $\hat{h}_n$ given in (2.9). For each model, we replicate the Monte Carlo experiment 100 times. The selected bandwidths and the corresponding values of both ISE and MSPE are presented in box-plots.

Remark 2. The examples indicate the following interesting features.

(a) Hart and Vieu (1990) claimed that the ordinary cross-validation CV(1) is robust to moderate amounts of dependence in the data, but *some improvement* can be obtained by taking CV($2l+1$) with $l \geq 1$ when the data are sufficiently highly dependent. However, for all the examples shown below, the CV($2l+1$) methods with $l \geq 1$ do not furnish any systematic improvement over CV(1) in terms of both the ISE and the MSPE even though the data are *sufficiently highly* *dependent* (see Figures 2, 3, 5, 6, 8 and 9). It is easy to see that when the regressor X_t is random, data points $X_{t\pm l}$ for $l = 1, 2, \ldots$ are not likely to be nearest neighbours of X_t. Note that only the data points near X_t in space are effectively used in estimating $f(x)$ at $x = X_t$. Therefore, it does not make any significant difference whether we leave out $X_{t\pm l}$ ($l = 1, 2$, and etc.) or not, unless the neighbours in the time space coincide with the neighbours in the state space such as some low-frequency time series data. This explains why in most of the following examples, the CV$(2l+1)$ methods with different values of l perform roughly the same. For the regression models with fixed design $X_t = t/n$, Chu and Marron (1991) has developed an interesting central limit theorem which explains how the effect of dependence on cross-validation is alleviated as the value of l is increased. However, their theory does not apply to our setup. For example, the asymptotic expansion (3.1) of Chu and Marron (1991) is no longer valid when X_t is random (cf. (A.25) and (A.26) of Härdle and Vieu 1992, and also (4.3) and (2.13)).

(b) The cross-validation does not over-smooth the negatively correlated data (see Figures 1(d,e) and 4(d,e)). The situation again seems different from the fixed designed models (cf. Altman 1990, Chu and Marron 1991, for example). On the other hand, our results seem to follow the tendency that cross-validation methods undersmooth rough realizations (eg. Figure 7(b,c)) and oversmooth relatively smoother samples (eg. Figure 7(a)). (Cf. Hall and Johnstone 1992.)

(c) For the given sample size $n = 300$, the new method (2.11) leads to the largests averages of the ISE values (see Figures 2, 5 and 8). The large variation of (2.11) and its corresponding ISE and MSPE is due to the inefficiency in the use of the data. However, it provides a competitive performance in prediction (see Figures 3, 6 and 9).

Example 1. We start with the exponential model

$$Y_t = 2.3 \exp\{-\frac{1}{2} X_t^2\} + 0.25 \epsilon_t, \tag{3.2}$$

where $\{X_t\}$ is a sequence of independent and standard normal random variables, ϵ_t is from an AR(1) process, i.e.

$$\epsilon_t = \rho \epsilon_{t-1} + \sqrt{1 - \rho^2} e_t,$$

and e_t, $t \geq 1$, are independent $N(0,1)$ random variables, for each t, e_t is independent of $\{X_k,\ k \leq t\}$, and $\rho \in (-1, 1)$ is a constant. It is easy to see that $\epsilon_t \sim N(0,1)$. The simulation is carried out for five different values of ρ : 0.9, 0.5, 0, -0.5, and -0.9. The boxplots of the bandwidths selected by the new method (2.9) and the CV(k) methods ($k = 1, 3, 5, 7$, and 9), together with h_{opt}, in the 100 replications are presented in Fig.1. Figs.2 and 3 are the boxplots of the corresponding the

ISE values and the MSPE values. Fig.1 indicates clearly that the cross-validation method CV(k), even with large values of k, tends to give smaller bandwidths than h_{opt}. The bandwidth selected by (2.9) is closer to h_{opt}. However, its variation is larger too.

(Fig.1 — Fig.3 are about here)

Example 2. Consider another exponential model

$$Y_t = 2.3 \exp\{-\frac{1}{2} X_t^2\} + 0.25\epsilon_t, \tag{3.3}$$

where $\{\epsilon_t\}$ is a sequence of independent and standard normal random variables, X_t is from the AR(1) process

$$X_t = \rho X_{t-1} + \sqrt{1-\rho^2} e_t,$$

and $\{e_t,\ t \geq 1\}$ are independent $N(0,1)$ random variables, for each t, e_t is independent of $\{X_k,\ k \leq t\}$, and $\rho \in (-1,1)$ is a constant. Obviously, $X_t \sim N(0,1)$. A Monte Carlo experiment with 100 replications was carried out for each of five different values of ρ. Fig.4(a) suggests that CV(k) with $k > 1$ chooses the bandwidth which is closer to h_{opt} than that selected by CV(1) when $\rho = 0.9$. However, the situation is opposite when $\rho = -0.9$ (see Fig.4(e)).

(Fig.4 — Fig.6 are about here)

Example 3. Let us consider the quadratic model

$$X_{t+1} = 0.23 X_t (16 - X_t) + 0.3\epsilon_t, \tag{3.4}$$

where $\{\epsilon_t\}$ is a sequence of *i.i.d* random variables, and ϵ_1 has a standard normal distribution truncated on the interval [-12, 12]. We consider three cases: $Y_t = X_{t+p}$ for $p = 1, 2, 3$. For each case, a Monte Carlo experiment with 100 replications was conducted. The results are reported in Fig.5 and Fig.6. Compared with h_{opt}, all the methods yield a larger bandwidth when $p = 1$ and a smaller bandwidth when $p = 2, 3$.

(Fig.7 — Fig.9 are about here)

4 Asymptotical properties

Theorems 2, 3 and 4 stated in this section are not only useful for §2 but also of independent interest.

We use the same notation as in §2. To discuss the asymptotic properties, we need the following assumptions. We denote by $g(.|x)$ the conditional density function of Y_1 given $X_1 = x$, and by $p(.)$ the marginal density function of X_1. We use c to denote a generic constant which may be different at different places.

(A1) Let $\psi(x) = \int y^2 g(y|x)dy$. The marginal density function p and ψ have continuous first derivatives.

(A2) $\mathrm{E}Y_1^8 < \infty$, and $f(x) = \mathrm{E}\{Y_1|X_1 = x\}$ has continuous third derivative and $|\ddot{f}(x)| \leq c < \infty$ for all $x \in \{p(x) > 0\}$.

(A3) The joint density of the distinct elements of $(X_1, Y_1, X_k, Y_k)\,(k > 0)$ is bounded by a constant independent of k.

(A4) The strictly stationary process $\{(X_i, Y_i), i \geq 1\}$ is ρ-mixing, i.e.

$$\rho_j \equiv \sup_{i\geq 1}\{\sup_{U\in\mathcal{F}_1^i, V\in\mathcal{F}_{i+j}^\infty} |\mathrm{Corr}(U,\ V)|\} \to 0,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \ldots, j\}(j \geq i)$. Further, we assume that $\sum_{k=1}^\infty \rho_k < \infty$

(A5) $K(.)$ is a bounded symmetric density function with a bounded support in $\mathbf{R}$. Further $\int xK(x)dx = 0$, $\int x^2K(x)dx = \sigma_0^2 > 0$, and $|K(x) - K(y)| \leq c|x - y|$ for any $x, y \in \mathbf{R}$.

(A6) For ε_0 given in (2.8), there exists $d_n \to \infty$ as $n \to \infty$, such that $d_n = o(n^{\frac{1}{5}-4\varepsilon_0})$, and $\sum_{k=d_n}^n \rho_k = o(n^{-\frac{1}{5}-5\varepsilon_0})$.

The condition of bounded support of kernel function is imposed for the brevity of proofs, which can be removed at the expense of longer proofs. In particular, Gaussian kernels are allowed. The assumption that the process is ρ-mixing is also for technical convenience. In fact, an autoregressive process satisfying some mild conditions is ρ-mixing (cf. Section 4.4 of Győrfi *et al* 1989). More detailed discussion on different mixing conditions can also be found in Bradley (1986). Condition (A6) is not the weakest possible either.

Theorem 2. Suppose that conditions (A1) – (A5) hold. Then for $h \in H_n$ and $x \in \{p(x) > 0\}$,

$$\sqrt{nh}\{\hat{f}_{n,h}(x) - f(x) - \frac{1}{2}h^2\sigma_0^2\ddot{f}(x)\} \xrightarrow{d} N(0, p^{-1}(x)\sigma^2(x)\int K^2(u)\mathrm{d}u),$$

where $\sigma_0^2 = \int u^2 K(u)\mathrm{d}u$, $\sigma^2(x) = \mathrm{Var}(Y_1|X_1 = x)$.

Theorem 3. Let $w_n(.) = b^{-1}w(./b)$, where $w(.)$ is a density function with a bounded support, and $b^{-1} = O(n^{\varepsilon_0/2})$ as $n \to \infty$. Suppose that conditions (A1) – (A6) hold, and $x \in \{p(x) > 0\}$.

(i) Assume m is chosen in such a way that both m/n and $(n-m)/n$ converge to some positive constants as $n \to \infty$. Then, for $h \in H_m$ uniformly,

$$M_m(x,h) = \frac{h^4}{4}\sigma_0^4\{\ddot{f}(x)\}^2 p(x) + \frac{1}{mh}\sigma^2(x)\int K^2(u)\mathrm{d}u + o_p(h^4 + \frac{1}{mh}), \tag{4.1}$$

(ii) Assume m' is chosen in such a way that m'/n converges to a positive constant. Then, for $h \in H_n$ uniformly,

$$M_n(x,h) = \frac{h^4}{4}\sigma_0^4\{\ddot{f}(x)\}^2 p(x) + \frac{1}{nh}\sigma^2(x)\int K^2(u)\mathrm{d}u + o_p(h^4 + \frac{1}{nh}), \tag{4.2}$$

where $M_m(x,h)$ and $M_n(x,h)$ are defined in (2.14) and (2.12) respectively.

Remark 3. Similar to Theorem 3, we can prove that if $w(.)$ is a density function with a compact support contained in the support of $p(.)$, then

$$\begin{aligned} M_n(h) &\equiv \frac{1}{m'}\sum_{t=n+1}^{n+m'}\{\hat{f}_{n,h}(X_t) - f(X_t)\}^2 w(X_t) \\ &= \frac{h^4}{4}\sigma_0^4\int\{\ddot{f}(x)\}^2 p(x)w(x)\mathrm{d}x + \frac{1}{nh}\int\sigma^2(x)w(x)\mathrm{d}x\int K^2(u)\mathrm{d}u + o_p(h^4 + \frac{1}{nh}). \quad (4.3)\end{aligned}$$

Theorem 4. Let $w_n(.) = b^{-1}w(./b)$, where $w(.)$ is a density function with a bounded support, and $b^{-1} = O(n^{\varepsilon_0/2})$ as $n \to \infty$. Assume m is chosen in such a way that both m/n and $(n-m)/n$ converge to some positive constants, as $n \to \infty$. Then, under conditions (A1) – (A6),

$$\frac{M_m(x, \hat{h}_m(x))}{\inf_{h\in H_m} M_m(x,h)} \xrightarrow{\mathcal{P}} 1,$$

where $M_m(x,h)$ is defined as in (2.14) and H_m is defined as in (2.8).

Theorem 2 is a corollary of Theorem 1 of Yao and Tong (1996). We now prove Theorems 3 and 4 in a sequence of lemmas.

Lemma 1. Assume that conditions (A1) – (A5) hold. Then for any compact subset $G \subset R$,

$$\sup_{x\in G, h\in H_n} |s_j(x) - \mathrm{E}\{s_j(x)\}| \xrightarrow{\mathcal{P}} 0, \quad j = 0, 1, 2,$$

where $s_j(.)$ is defined as in (2.4).

Proof. To indicate the dependence of s_j on h, we write $s_j(x,h) = s_j(x)$ when necessary.

For any $x \in G$, $h \in H_n$ and $\varepsilon > 0$, it follows from Tchebychev inequality that

$$\begin{aligned}
& P\{|s_j(x) - \mathrm{E}s_j(x)| \geq \varepsilon\} \\
\leq\ & \frac{1}{n^2h^2\varepsilon^2}\sum_{i=1}^{n}\mathrm{E}\left\{\left(\frac{X_i - x}{h}\right)^{2j} K\left(\frac{X_i - x}{h}\right)\right\}^2 \\
+\ & \frac{2}{nh^2\varepsilon^2}\sum_{k=1}^{n-1}\left|\mathrm{E}\left\{\left(\frac{X_1 - x}{h}\right)^{j}\left(\frac{X_k - x}{h}\right)^{j} K\left(\frac{X_1 - x}{h}\right) K\left(\frac{X_k - x}{h}\right)\right\}\right. \\
-\ & \left.\left\{\mathrm{E}\left[\left(\frac{X_1 - x}{h}\right)^{j} K\left(\frac{X_1 - x}{h}\right)\right]\right\}^2\right| \\
\leq\ & \frac{c}{nh} + \frac{2}{nh^2\varepsilon^2}\mathrm{E}\left\{\left(\frac{X_1 - x}{h}\right)^{2j} K^2\left(\frac{X_1 - x}{h}\right)\right\}\sum_{k=1}^{n-1}\rho_k \\
\leq\ & cn^{-\frac{4}{5}+\varepsilon_0} \equiv \pi_n. \qquad (4.4)
\end{aligned}$$

We cover $G \times H_n$ by a finite number of open balls B_k centered at $(x_k, h^{(k)})$, $k = 1, \ldots, l_n$, in such a way that

$$G \times H_n \subset \cup_{k=1}^{l_n} B_k, \qquad l_n = O(n^{3/5+5\varepsilon_0}(\log n)^2),$$

and for $(x,h) \in B_k$,

$$|x - x_k| + |h - h^{(k)}| \leq n^{-2(1/5+\varepsilon_0)}/\log n.$$

Hence, for $(x,h) \in B_k$,

$$\begin{aligned}
& \left|\frac{X_t - x}{h} - \frac{X_t - x_k}{h^{(k)}}\right| \leq \left|\frac{X_t - x}{h} - \frac{X_t - x_k}{h}\right| + \left|\frac{X_t - x_k}{h} - \frac{X_t - x_k}{h^{(k)}}\right| \\
=\ & \frac{1}{h}|x - x_k| + \frac{|X_t - x_k|}{hh^{(k)}}|h - h^{(k)}| \\
\leq\ & \max\left\{c, \frac{|X_t - x_k|}{h}\right\} n^{1/5+\varepsilon_0}\{|x - x_k| + |h - h^{(k)}|\} \\
\leq\ & \max\left\{c, \frac{|X_t - x_k|}{h}\right\} n^{-(1/5+\varepsilon_0)}/\log n.
\end{aligned}$$

Therefore, by condition (A5), we have that

$$\begin{aligned}
& \left|\frac{1}{h}\left(\frac{X_t - x}{h}\right)^{j} K\left(\frac{X_t - x}{h}\right) - \frac{1}{h^{(k)}}\left(\frac{X_t - x_k}{h^{(k)}}\right)^{j} K\left(\frac{X_t - x_k}{h^{(k)}}\right)\right| \\
& \leq \frac{c}{h}\left|\frac{X_t - x}{h} - \frac{X_t - x_k}{h^{(k)}}\right| + c\left|\frac{1}{h} - \frac{1}{h^{(k)}}\right| = O((\log n)^{-1}). \qquad (4.5)
\end{aligned}$$

The above limit holds uniformly for all t. Consequently,

$$P\{\sup_{x \in G, h \in H_n} |s_j(x) - \mathrm{E}s_j(x)| \geq \varepsilon\}$$

$$
\begin{aligned}
&= P\{\sup_{x\in G, h\in H_n} |s_j(x,h) - \mathrm{E}s_j(x,h)| \geq \varepsilon\} \\
&= P\{\max_{1\leq k\leq l_n} |s_j(x_k, h^{(k)}) - \mathrm{E}s_j(x_k, h^{(k)})| + O((\log n)^{-1}) \geq \varepsilon\} \\
&\leq l_n \cdot \pi_n + o(1) \to 0, \qquad (4.6)
\end{aligned}
$$

and the limit follows from (4.4). The proof is completed.

Lemma 2. Assume that conditions (A1) – (A6) holds. Then uniformly for $h \in H_n$ and x in any compact subset of $\{p(x) > 0\}$,

(i) $\frac{1}{nhp(x)} \sum_{t=1}^n \epsilon_t^2 K^2(\frac{X_t - x}{h}) \xrightarrow{\mathcal{P}} \sigma^2(x) \int K^2(u)\mathrm{d}u$, where $\epsilon_t = Y_t - f(X_t)$,

(ii) $\frac{1}{nhp(x)} \sum_{t=1}^n (\frac{X_t - x}{h})^2 K(\frac{X_t - x}{h}) \xrightarrow{\mathcal{P}} \sigma_0^2$.

The proof of Lemma 2 is similar to that of Lemma 1 and is omitted here.

Lemma 3. Assume that the conditions of Theorem 3 hold. Then uniformly for $h \in H_m$ and $x \in \{p(x) > 0\}$,

(i) $\frac{h}{m(n-m)} \sum_{k=m+1}^n \sum_{t=1}^m p^{-1}(X_k)\ddot{f}(X_k)\epsilon_t K(\frac{X_k - X_t}{h}) w_n(X_k - x) = o_p(m^{-(\frac{4}{5}+4\varepsilon_0)})$,

(ii) $\frac{1}{m^2(n-m)h^2} \sum_{k=m+1}^n \sum_{1\leq i<j\leq n} p^{-2}(X_k)\epsilon_i\epsilon_j K(\frac{X_k - X_i}{h}) K(\frac{X_k - X_j}{h}) w_n(X_k - x)$
$= o_p(m^{-(\frac{4}{5}+4\varepsilon_0)})$,

(iii) $\frac{h^2}{(n-m)} \sum_{t=m+1}^n \epsilon_t \ddot{f}(X_t) w_n(X_t - x) = o_p(m^{-(\frac{4}{5}+4\varepsilon_0)})$,

(iv) $\frac{1}{m(n-m)h} \sum_{t=m+1}^n \sum_{i=1}^m \epsilon_t p^{-1}(X_t) K\left(\frac{X_i - X_t}{h}\right) w_n(X_t - x) = o_p(m^{-(\frac{4}{5}+4\varepsilon_0)})$.

Proof. We only show (i). The proofs of the rest are similar in principle, although more technical details will be involved in the proof of (ii).

Let $I_1(h)$ denote the LHS of the equality in (i). By the definition of ρ-mixing condition,

$$
\begin{aligned}
|\mathrm{E}I_1(h)| &\leq \frac{h}{m(n-m)} \sum_{k=m+1}^n \sum_{t=1}^m \rho_{k-t} \{\mathrm{Var}(\epsilon_t)\}^{\frac{1}{2}} \\
&\quad \cdot \left\{\mathrm{Var}\left[\ddot{f}(X_k)p^{-1}(X_k)w_n(X_k - x)\mathrm{E}\left\{K\left(\frac{X_k - X_t}{h}\right)\middle| X_k\right\}\right]\right\}^{\frac{1}{2}} \\
&\leq \frac{ch^2}{m(n-m)b} \sum_{k=m+1}^n \sum_{t=1}^m \rho_{k-t} \leq \frac{ch^2}{n(n-m)} \sum_{k=1}^{m\vee(n-m)} k\rho_k = o(m^{-(\frac{4}{5}+4\varepsilon_0)}). \qquad (4.7)
\end{aligned}
$$

We decompose the variance into four parts as follows.

$$
\mathrm{Var}(I_1(h)) = \left\{ \sum_{m<k=l\leq n} \sum_{1\leq i=j\leq m} + \sum_{m<k=l\leq n} \sum_{\substack{1\leq i,j\leq m \\ i\neq j}} + \sum_{\substack{m<k,l\leq n \\ k\neq l}} \sum_{1\leq i=j\leq m} + \sum_{\substack{m<k,l\leq n \\ k\neq l}} \sum_{\substack{1\leq i,j\leq m \\ i\neq j}} \right\}
$$

$$\begin{aligned}
&\frac{h^2}{m^2(n-m)^2}\mathrm{Cov}\left\{\ddot{f}(X_k)p^{-1}(X_k)\epsilon_i K\left(\frac{X_k-X_i}{k}\right)w_n(X_k-x),\right.\\
&\qquad\left.\ddot{f}(X_l)p^{-1}(X_l)\epsilon_j K\left(\frac{X_l-X_j}{k}\right)w_n(X_l-x)\right\}\\
\equiv\;& I_{11}+I_{12}+I_{13}+I_{14}. \qquad\qquad (4.8)
\end{aligned}$$

Note that

$$\begin{aligned}
|I_{14}| \;\le\;& \frac{2h^2}{m^2(n-m)^2}\sum_{\substack{m<k,l\le n\\ k\ne l}}\sum_{\substack{1\le i<j\le m\\ i\ne j}}\left\{\left|\mathrm{E}\left(\ddot{f}(X_k)\ddot{f}(X_l)p^{-1}(X_k)p^{-1}(X_l)\epsilon_i\epsilon_j K\left(\frac{X_k-X_i}{h}\right)\right.\right.\right.\\
&\left.\left.\cdot\, K\left(\frac{X_l-X_j}{h}\right)w_n(X_k-x)w_n(X_l-x)\right)\right|\\
&+\left|\mathrm{E}\left(\ddot{f}(X_k)p^{-1}(X_k)\epsilon_i K\left(\frac{X_k-X_i}{h}\right)w_n(X_k-x)\right)\right|\\
&\left.\cdot\left|\mathrm{E}\left(\ddot{f}(X_l)p^{-1}(X_l)\epsilon_j K\left(\frac{X_l-X_j}{h}\right)w_n(X_l-x)\right)\right|\right\}\\
\le\;& \frac{2h^2}{m^2(n-m)^2}\sum_{\substack{m<k,l\le n\\ k\ne l}}\sum_{\substack{1\le i<j\le m\\ i\ne j}}\rho_{j-i}\{\mathrm{Var}(\epsilon_i)\}^{\frac{1}{2}}\left(\mathrm{Var}\left\{\mathrm{E}\left[K\left(\frac{X_k-X_i}{h}\right)\middle| X_k\right]\right.\right.\\
&\left.\left.\cdot\,\ddot{f}(X_k)\ddot{f}(X_l)p^{-1}(X_k)p^{-1}(X_l)\epsilon_j K\left(\frac{X_l-X_j}{h}\right)w_n(X_k-x)w_n(X_l-x)\right\}\right)^{\frac{1}{2}}\\
&+\frac{ch^3}{m^2(n-m)^2}\sum_{\substack{m<k,l\le n\\ k\ne l}}\sum_{\substack{1\le i<j\le m\\ i\ne j}}\rho_{k-i}\{\mathrm{Var}(\epsilon_i)\}^{\frac{1}{2}}\\
&\cdot\left(\mathrm{Var}\left\{\ddot{f}(X_k)p^{-1}(X_k)w_n(X_k-x)\mathrm{E}\left[K\left(\frac{X_k-X_i}{h}\right)\middle| X_k\right]\right\}\right)^{\frac{1}{2}}\\
\le\;& \frac{ch^{3\frac{1}{2}}}{m^2(n-m)^2 b}\sum_{\substack{m<k,l\le n\\ k\ne l}}\sum_{\substack{1\le i<j\le m\\ i\ne j}}\rho_{j-i}+\frac{ch^4}{m^2(n-m)^2 b^{1/2}}\sum_{\substack{m<k,l\le n\\ k\ne l}}\sum_{\substack{1\le i<j\le m\\ i\ne j}}\rho_{k-i}\\
=\;& O\left(\frac{h^{3\frac{1}{2}}}{mb}\right)+O\left(\frac{h^4}{mb^{1/2}}\right).
\end{aligned}$$

On the other hand, it is easy to see that $I_{11}=O(m^{-2}h^3b^{-1})$, $I_{12}=O(m^{-1}h^4b^{-1})$, and $I_{13}=O(m^{-1}h^4b^{-1})$. It follows from (4.8) that

$$\mathrm{Var}(I_1(h))=O\left(\frac{h^{3\frac{1}{2}}}{mb}\right)=o\left(m^{-3\varepsilon_0}m^{-2(\frac{4}{5}+4\varepsilon_0)}\right)\equiv\pi_m$$

for all $h\in H_m$, where the last equality follows from the fact that $\varepsilon_0<\frac{1}{150}$, and $b^{-1}=O(n^{\varepsilon_0/2})$.

We cover H_m by a finite number of open intervals B_k centred at $h^{(k)}$, $k=1,\ldots,l_m$, in such a way that

$$H_m\subset\cup_{k=1}^{l_m}B_k,\qquad l_m=O(m^{2\varepsilon_0}\log m),$$

and for $h\in B_k$,

$$|h-h^{(k)}|\le m^{-1/5-\varepsilon_0}/\log m.$$

By (A5), for $h \in B_k$,

$$\left|hK\left(\frac{X_k - X_t}{h}\right) - h^{(k)}K\left(\frac{X_k - X_t}{h^{(k)}}\right)\right| \le h\left|K\left(\frac{X_k - X_t}{h}\right) - K\left(\frac{X_k - X_t}{h^{(k)}}\right)\right|$$
$$+K\left(\frac{X_k - X_t}{h^{(k)}}\right)|h - h^{(k)}| \le c\frac{|X_k - X_t|}{h^{(k)}}|h - h^{(k)}| + c|h - h^{(k)}| \le cm^{-1/5-\varepsilon_0}/\log m.$$

Consequently, for any $h \in B_k$ $(1 \le k \le l_m)$,

$$|I_1(h) - I_1(h^{(k)})| \le cm^{-1/5-\varepsilon_0}/\log m \frac{1}{m(n-m)} \sum_{k=m+1}^{n}\sum_{t=1}^{m} p^{-1}(X_k)|\ddot{f}(X_k)\epsilon_t|w_n(X_k - x),$$

which is independent of k and is of the order $O_p(m^{-1/5-\varepsilon_0}/\log m)$.

For any $\varepsilon > 0$, let $\varepsilon(m) = \varepsilon(h^4 + \frac{1}{mh})$. Then,

$$\begin{aligned}
& P\left\{\sup_{h\in H_m} |I_1(h) - \mathrm{E}I_1(h)| \ge \varepsilon(m)\right\} \\
\le\ & P\left\{\max_{1\le k\le l_m} |I_1(h^{(k)}) - \mathrm{E}\{I_1(h^{(k)})\}| + \max_{1\le k\le l_m}\sup_{h\in B_k} |I_1(h) - I_1(h^{(k)})| + o(m^{-1/5-\varepsilon_0}) > \varepsilon(m)\right\} \\
\le\ & \sum_{k=1}^{l_m} P\left\{|I_1(h^{(k)}) - \mathrm{E}\{I_1(h^{(k)})\}| > \varepsilon(m)/2\right\} + P\left\{o_p(m^{-1/5-\varepsilon_0}) > \varepsilon(m)/2\right\} \\
\le\ & 4\{\varepsilon(m)\}^{-2}\sum_{k=1}^{l_m} \mathrm{Var}\{I_1(h^{(k)})\} + o(1) \to 0.
\end{aligned}$$

Now, (i) follows from (4.7) immediately. The proof is completed.

Proof of Theorem 3. Let $G \subset \{p(x) > 0\}$ be a compact subset, and $h \in H_m$. It follows immediately from Lemma 1 that for $x \in G$ uniformly

$$S_m^{-1}(x) \xrightarrow{\mathcal{P}} p^{-1}(x)\begin{pmatrix} 1 & 0 \\ 0 & \sigma_0^2 \end{pmatrix}^{-1} = p^{-1}(x)\begin{pmatrix} 1 & 0 \\ 0 & \sigma_0^{-2} \end{pmatrix}.$$

(See (2.4)). It follows from (2.1) and (2.3) that

$$\hat{f}_{m,h}(x) = \frac{1}{mhp(x)}\sum_{t=1}^{m} K\left(\frac{X_t - x}{h}\right) Y_t\{1 + o_p(1)\},$$

uniformly for $x \in G$. Similar to (2.1), we have $\hat{\beta} - \beta = (X^\tau WX)^{-1}X^\tau W(Y - X\beta)$. Note that $K(.)$ has a bounded support. We have that for $x \in G$, uniformly,

$$\begin{aligned}
\hat{f}_{m,h}(x) - f(x) &= \left\{\frac{1}{mhp(x)}\sum_{t=1}^{m} K\left(\frac{X_t - x}{h}\right)\{Y_t - f(x) - \dot{f}(x)(X_t - x)\}\right\}\{1 + o_p(1)\} \\
&= \left\{\frac{1}{mhp(x)}\sum_{t=1}^{m} K\left(\frac{X_t - x}{h}\right)\{\epsilon_t + \frac{1}{2}\ddot{f}(x)(X_t - x)^2 + o_p(h^2)\}\right\}\{1 + o_p(1)\} \\
&= \left\{\frac{1}{mhp(x)}\sum_{t=1}^{m} \epsilon_t K\left(\frac{X_t - x}{h}\right) + \frac{1}{2}\ddot{f}(x)h^2\sigma_0^2\right\}\{1 + o_p(1)\}, \qquad (4.9)
\end{aligned}$$

the last equality being a consequence of Lemma 2 (ii). By Lemma 2 (i),

$$\begin{aligned}
\{\hat{f}_{m,h}(x) - f(x)\}^2 = &\left\{ \frac{h^4}{4}\{\ddot{f}(x)\}^2\sigma_0^4 + \frac{1}{m^2h^2p^2(x)}\sum_{t=1}^{m}\epsilon_t^2 K^2\left(\frac{X_t - x}{h}\right) \right. \\
&+ \frac{\sigma_0^2 h}{mp(x)}\ddot{f}(x)\sum_{t=1}^{m}\epsilon_t K\left(\frac{X_t - x}{h}\right) \\
&\left. + \frac{1}{m^2h^2p^2(x)}\sum_{1\le i<j\le m}\epsilon_i\epsilon_j K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right)\right\}\{1 + o_p(1)\} \\
= &\left\{ \frac{h^4}{4}\{\ddot{f}(x)\}^2\sigma_0^4 + \frac{1}{mhp(x)}\sigma^2(x)\int K^2(u)\mathrm{d}u + \frac{\sigma_0^2 h}{mp(x)}\ddot{f}(x)\sum_{t=1}^{m}\epsilon_t K\left(\frac{X_t - x}{h}\right) \right. \\
&\left. + \frac{1}{m^2h^2p^2(x)}\sum_{1\le i<j\le m}\epsilon_i\epsilon_j K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right)\right\}\{1 + o_p(1)\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
M_m(x,h) \quad = \quad &\left\{ \frac{h^4\sigma_0^4}{4}\frac{1}{n-m}\sum_{k=m+1}^{n}\{\ddot{f}(X_k)\}^2 w_n(X_k - x) \right. \\
+ \quad &\frac{1}{mh}\int K^2(u)\mathrm{d}u\frac{1}{n-m}\sum_{k=m+1}^{n}\frac{\sigma^2(X_k)}{p(X_k)}w_n(X_k - x) \\
+ \quad &\frac{\sigma_0^2 h}{m(n-m)}\sum_{k=m+1}^{n}\sum_{t=1}^{m}\ddot{f}(X_k)\epsilon_t p^{-1}(X_k) K\left(\frac{X_t - X_k}{h}\right) w_n(X_k - x) \\
+ \quad &\frac{1}{m^2(n-m)h^2}\sum_{k=m+1}^{n}\sum_{1\le i<j\le m}\epsilon_i\epsilon_j p^{-2}(X_k) K\left(\frac{X_i - X_k}{h}\right) \\
&\left. \cdot K\left(\frac{X_j - X_k}{h}\right) w_n(X_k - x)\right\}\{1 + o_p(1)\}.
\end{aligned}$$

By Lemma 3 (i) and (ii), the last two terms on the RHS of the above expression are of the order $o_p(h^4 + \frac{1}{mh})$. The asymptotic expansion (4.1) follows from the Ergodic Theorem immediately.

The proof of (4.2) is similar and omitted here.

Proof of Theorem 4. Let

$$CV_m(x,h) = \frac{1}{n-m}\sum_{t=m+1}^{n}\{Y_t - \hat{f}_{m,h}(X_t)\}^2 w_n(X_t - x).$$

Similar to Härdle and Vieu (1992), in order to prove Theorem 4 it is sufficient to show that

$$\sup_{h,h'\in H_m}\frac{|M_m(x,h) - M_m(x,h') + CV_m(x,h') - CV_m(x,h)|}{M_m(x,h)} \xrightarrow{\mathcal{P}} 0. \qquad (4.10)$$

For each m, let $h_m, h_m' \in H_m$ be the maximizers of the LHS of the above expression. Note that for any h,

$$CV_m(x,h) - M_m(x,h) = \frac{1}{n-m}\sum_{t=m+1}^{n}\epsilon_t^2 w_n(X_t - x)$$

$$+\frac{2}{n-m}\sum_{t=m+1}^{n}\epsilon_t\{f(X_t)-\hat{f}_{m,h}(X_t)\}w_n(X_t-x)$$

Hence,

$$\begin{aligned}&|M_m(x,h)-M_m(x,h')+CV_m(x,h')-CV_m(x,h)|\\ \le\ &\frac{2}{n-m}\left|\sum_{t=m+1}^{n}\epsilon_t\{f(X_t)-\hat{f}_{m,h}(X_t)\}w_n(X_t-x)\right|\\ +\ &\frac{2}{n-m}\left|\sum_{t=m+1}^{n}\epsilon_t\{f(X_t)-\hat{f}_{m,h'}(X_t)\}w_n(X_t-x)\right|.\end{aligned}$$

By (4.9),

$$\begin{aligned}\frac{1}{n-m}\sum_{t=m+1}^{n}\epsilon_t\{f(X_t)-\hat{f}_{m,h_m}(X_t)\}w_n(X_t-x)&=\frac{h_m^2\sigma_0^2}{2(n-m)}\sum_{t=m+1}^{n}\epsilon_t\ddot{f}(X_t)w_n(X_t-x)\\ +\frac{1}{m(n-m)h_m}\sum_{t=m+1}^{n}\sum_{i=1}^{m}\epsilon_t p^{-1}(X_t)K\left(\frac{X_i-X_t}{h_m}\right)w_n(X_t-x).\end{aligned}$$

It follows from Lemma 3 (iii) and (iv) that

$$|M_m(x,h_m)-M_m(x,h_m')+CV_m(x,h_m')-CV_m(x,h_m)|=o_p(m^{-(\frac{4}{5}+4\varepsilon_0)}).$$

On the other hand, by (4.1),

$$M_m(x,h_m)=O_p(h_m^4+\frac{1}{mh_m})=O_p(m^{-(\frac{4}{5}+4\varepsilon_0)}).$$

Therefore, (4.10) holds. The proof is completed.

References

Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, **85**, 749-759.

Chu, C.K. and Marron, J.S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, **19**, 1906-1918.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, **87**, 998-1004.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaption. *J. Roy. Statist. Soc.* **B**, **57**, 371-394.

Fan. J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. Technical Report, Dept. of Statist., Univ. of North Carolina.

Farmer, J.D. and Sidorowich, J.J. (1987). Predicting chaotic time series. *Phys. Rev. Lett.*, **59**, 845-848.

Győrfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Non-parametric Curve Estimation from Time Series.* Springer-Verlag, Berlin.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *J. Roy. Statist. Soc.*, **B**, **54**, 475-530.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series. *J. Time Series Anal.*, **13**, 209-232.

Hart, J.D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc.*, **B**, **53**, 173-187.

Hart, J.D. (1994). Automated kernel smoothing of dependent data by using time series crossing-validation. *J. Roy. Statist. Soc.*, **B**, **56**, 529-542.

Hart, J.D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, **18**, 873-890.

Marron, J.S. (1987). Partitioned cross-validation. *Econometric Rev.*, **6**, 271-284.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: a nonparametric approach. *J. Nonparametric Statist*, **6**, 273-292.

Captions

Fig.1. The boxplots of the selected bandwidths by h_{opt}, the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.2) with (a) $\rho = 0.9$, (b) $\rho = 0.5$, (c) $\rho = 0$, (d) $\rho = -0.5$, and (e) $\rho = -0.9$.

Fig.2. The boxplots of the ISEs corresponding to h_{opt}, the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.2) with (a) $\rho = 0.9$, (b) $\rho = 0.5$, (c) $\rho = 0$, (d) $\rho = -0.5$, and (e) $\rho = -0.9$.

Fig.3. The boxplots of the MSPEs corresponding to the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.2) with (a) $\rho = 0.9$, (b) $\rho = 0.5$, (c) $\rho = 0$, (d) $\rho = -0.5$, and (e) $\rho = -0.9$.

Fig.4. The boxplots of the selected bandwidths by h_{opt}, the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.3) with (a) $\rho = 0.9$, (b) $\rho = 0.5$, (c) $\rho = 0$, (d) $\rho = -0.5$, and (e) $\rho = -0.9$.

Fig.5. The boxplots of the ISEs corresponding to h_{opt}, the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.3) with (a) $\rho = 0.9$, (b) $\rho = 0.5$, (c) $\rho = 0$, (d) $\rho = -0.5$, and (e) $\rho = -0.9$.

Fig.6. The boxplots of the MSPEs corresponding to the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.3) with (a) $\rho = 0.9$, (b) $\rho = 0.5$, (c) $\rho = 0$, (d) $\rho = -0.5$, and (e) $\rho = -0.9$.

Fig.7. The boxplots of the selected bandwidths by h_{opt}, the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.4) with (a) $p = 1$, (b) $p = 2$, and (c) $p = 3$.

Fig.8. The boxplots of the ISEs corresponding to h_{opt}, the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.4) with (a) $p = 1$, (b) $p = 2$, and (c) $p = 3$.

Fig.9. The boxplots of the MSPEs corresponding to the new method (2.9), and CV(k) for $k = 1, 3, 5, 7$ and 9 for model (3.4) with (a) $p = 1$, (b) $p = 2$, and (c) $p = 3$.

Figure 1(a)

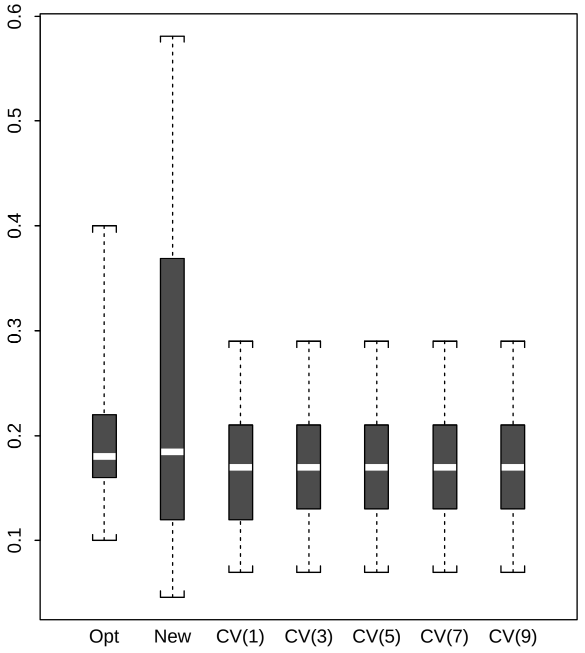


Figure 1(b)

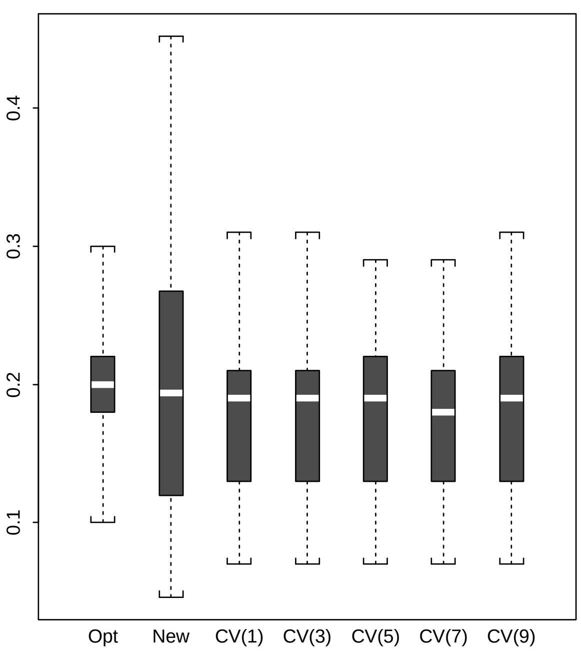


Figure 1(c)

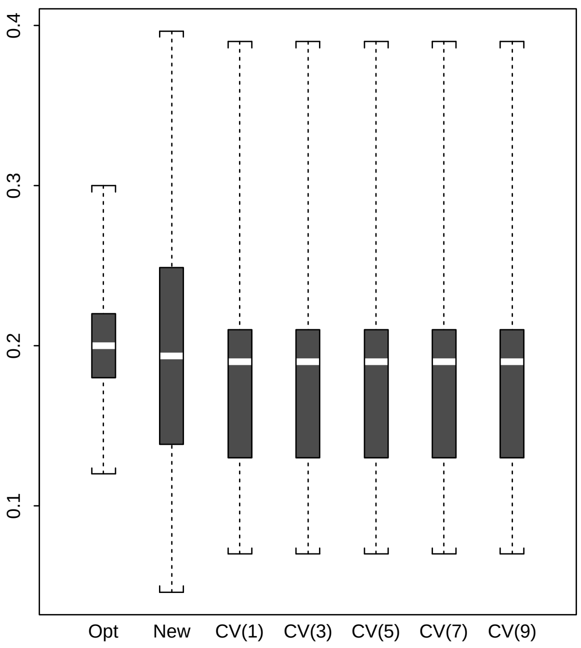


Figure 1(d)

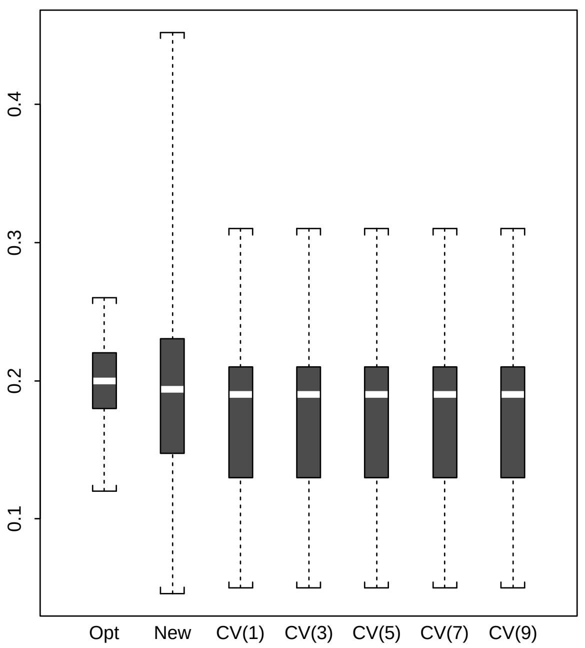
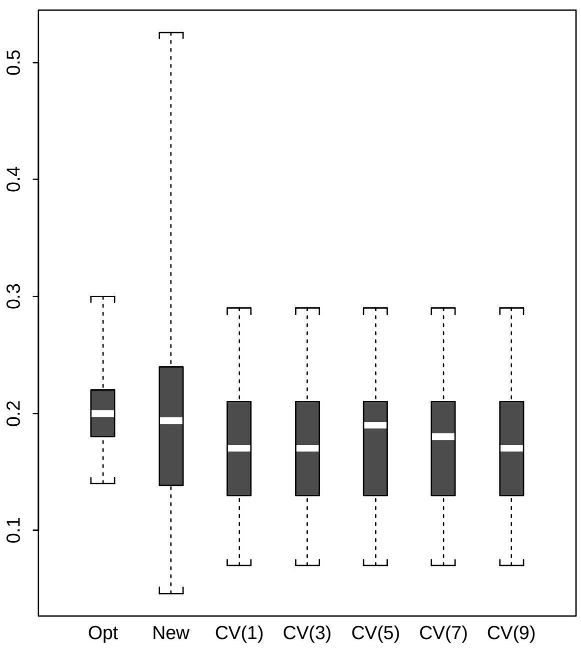


Figure 1(e)



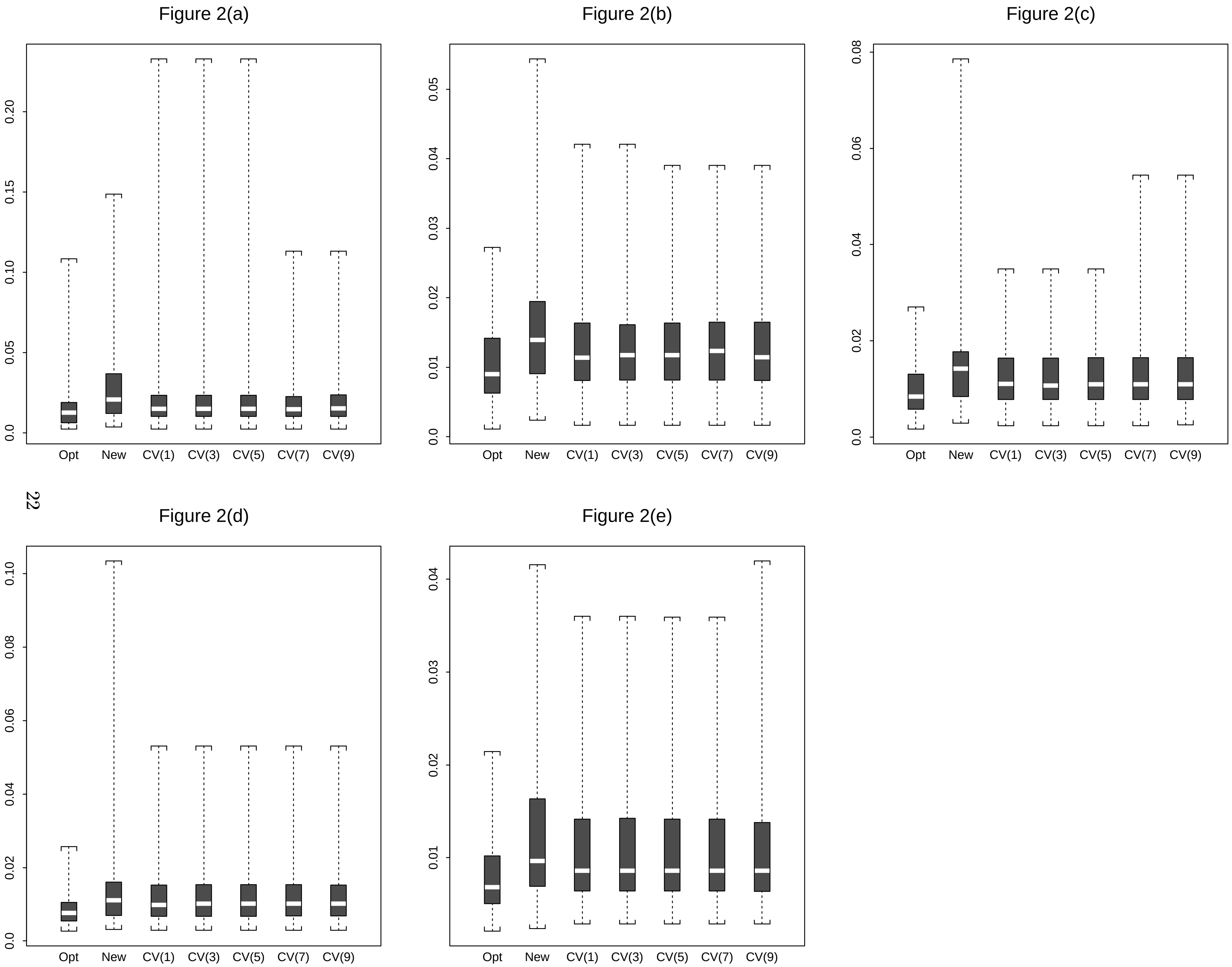


Figure 2(a)

Figure 2(b)

Figure 2(c)

Figure 2(d)

Figure 2(e)

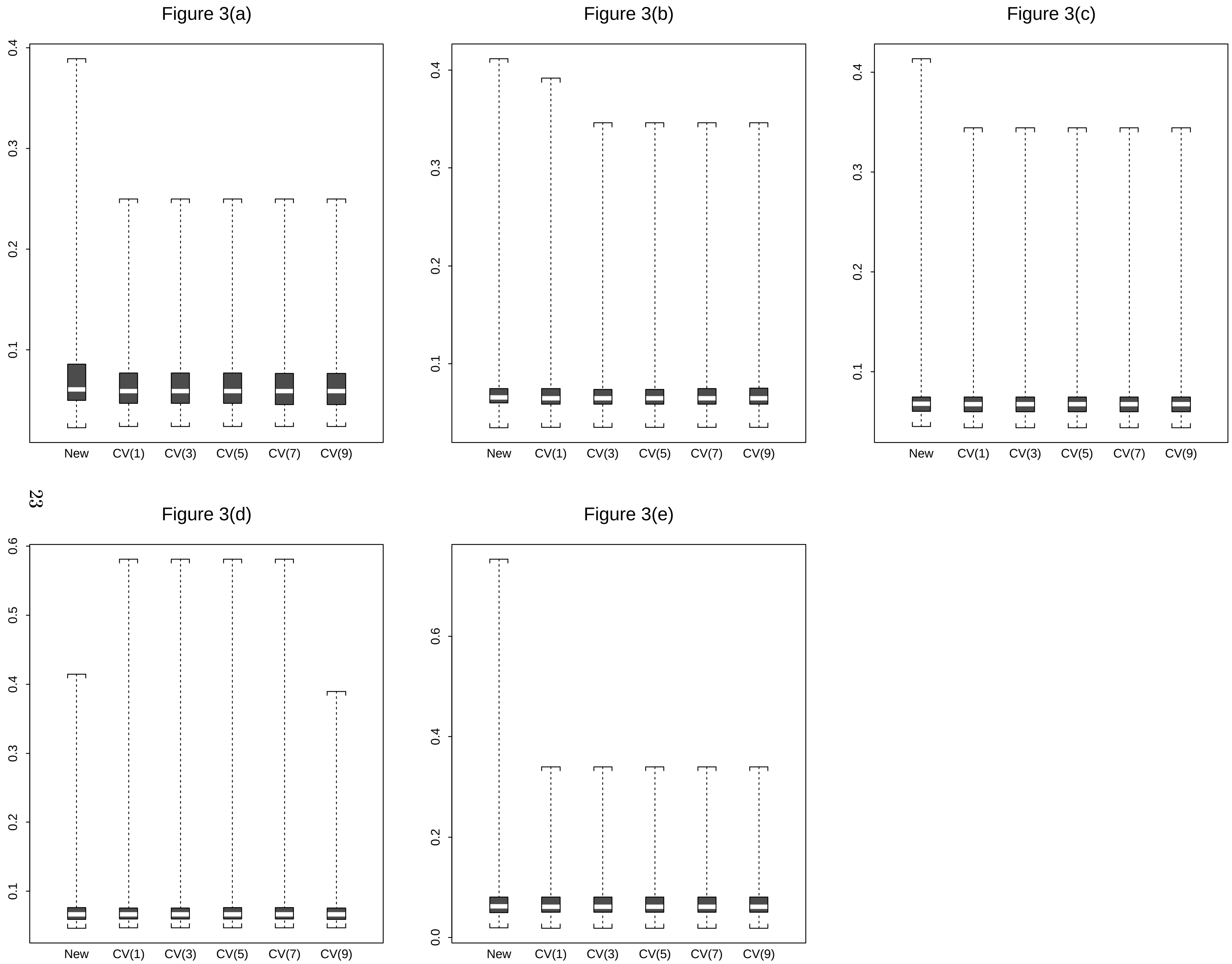


Figure 3(a)

Figure 3(b)

Figure 3(c)

Figure 3(d)

Figure 3(e)

Figure 4(a)

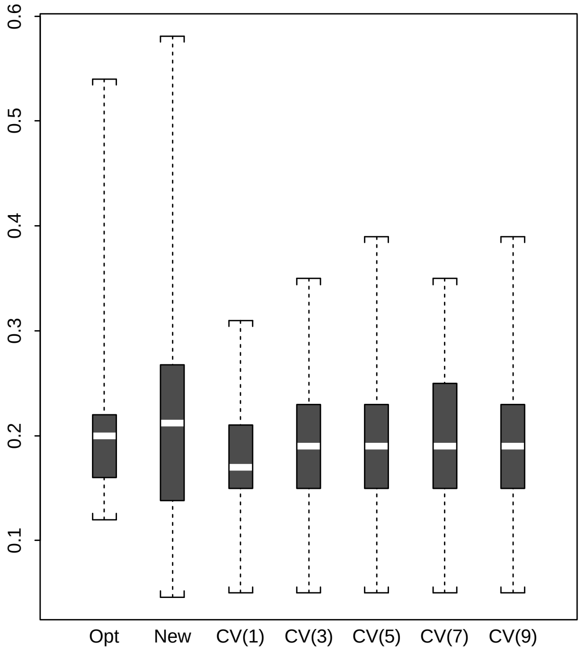


Figure 4(b)

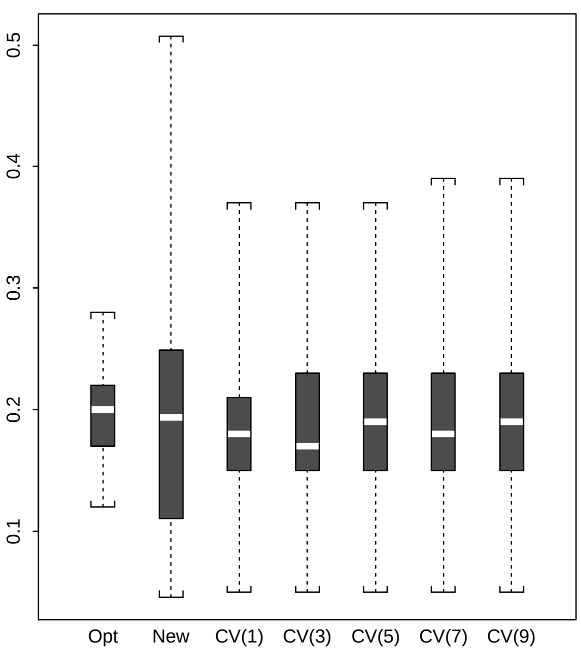


Figure 4(c)

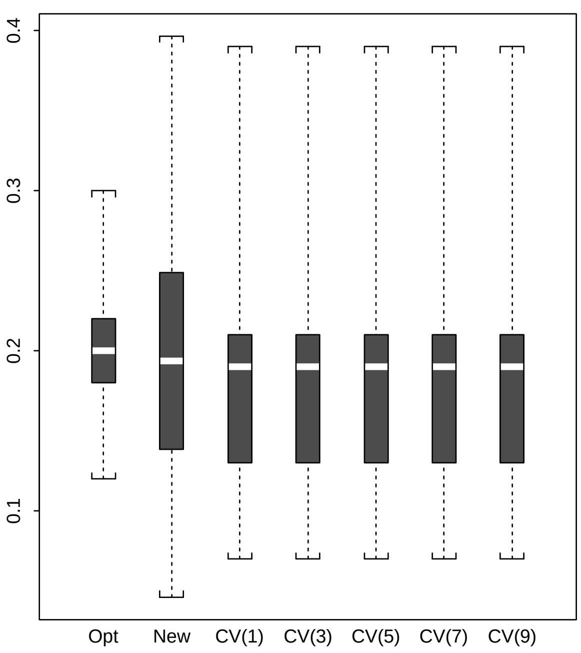


Figure 4(d)

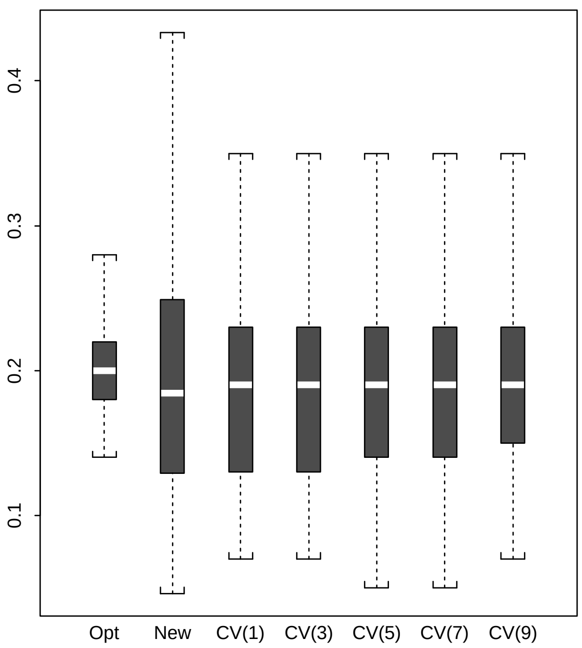


Figure 4(e)

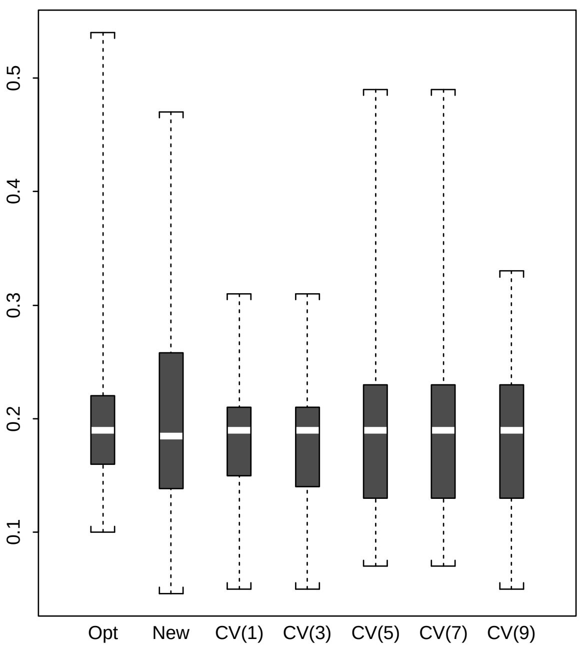


Figure 5(a)

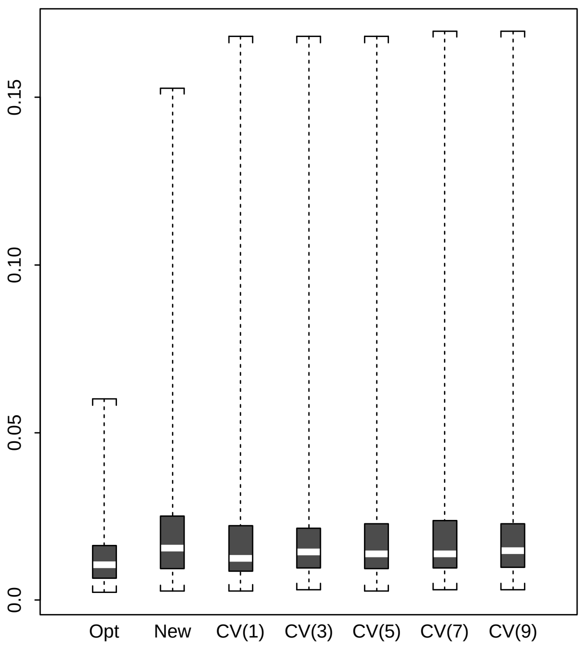


Figure 5(b)

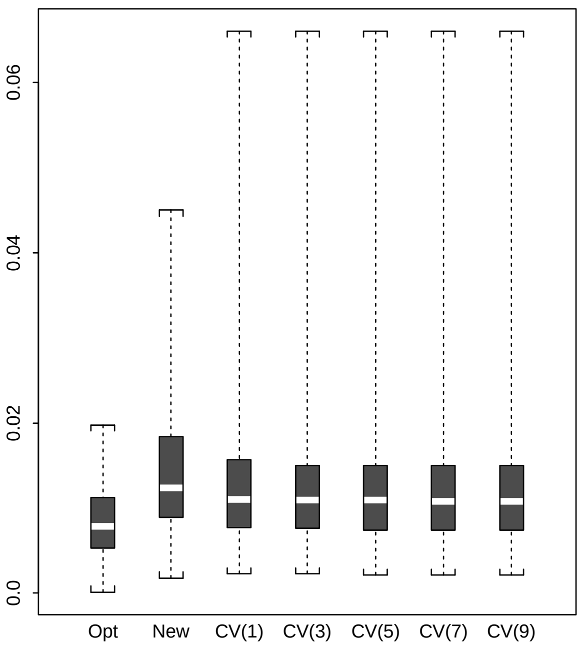


Figure 5(c)

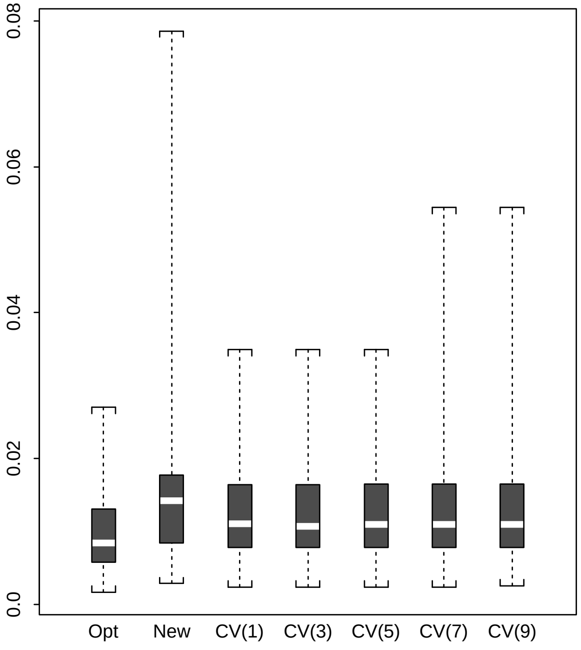


Figure 5(d)

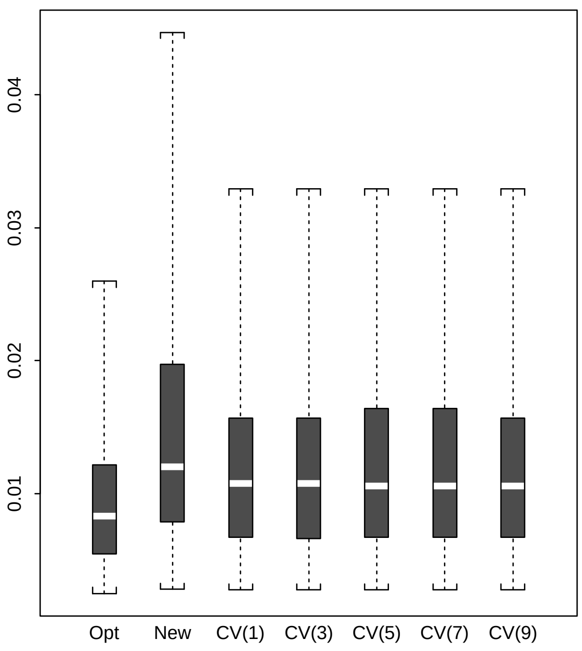


Figure 5(e)

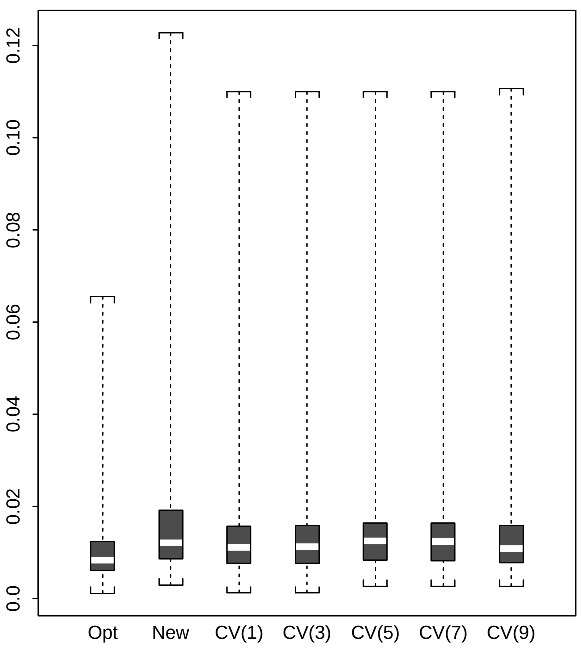


Figure 6(a)

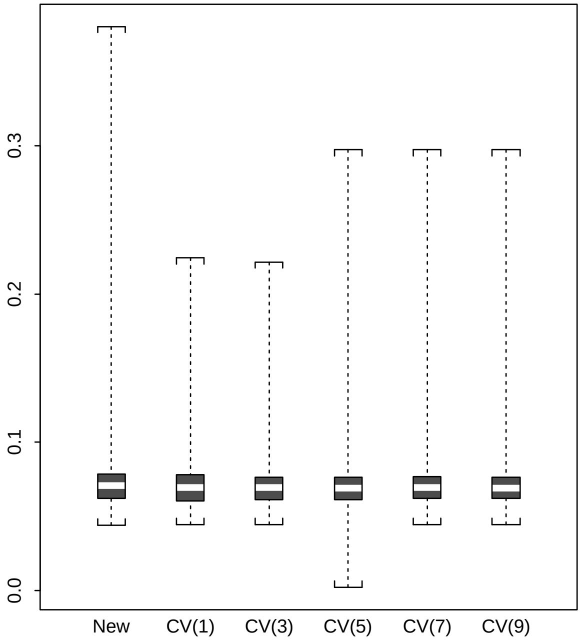


Figure 6(b)

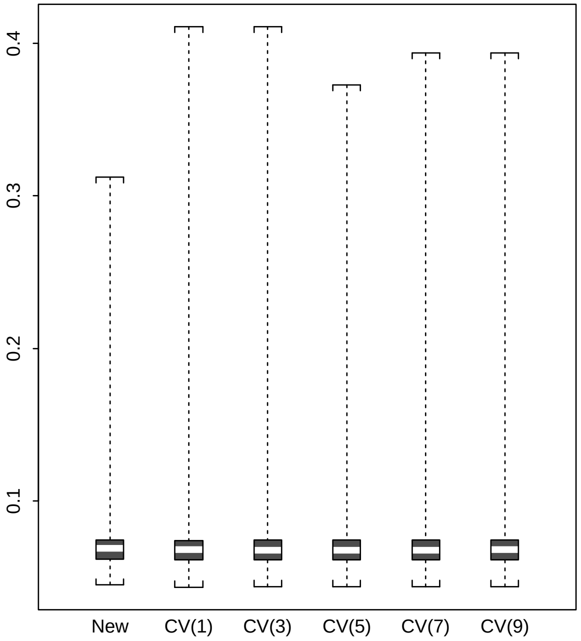


Figure 6(c)

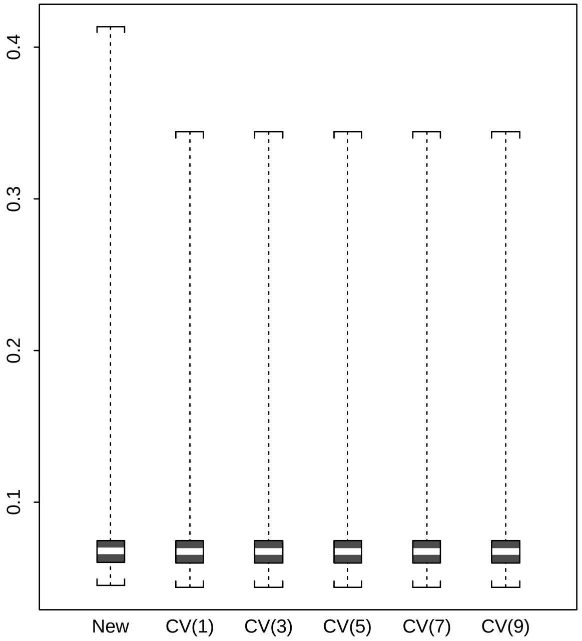


Figure 6(d)

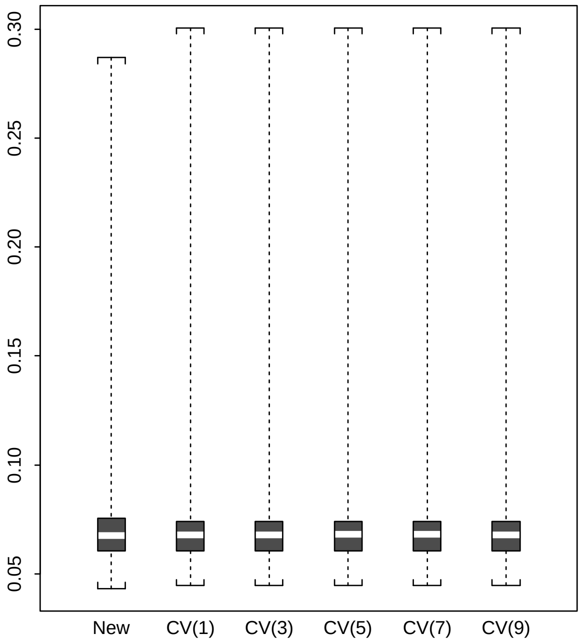


Figure 6(e)

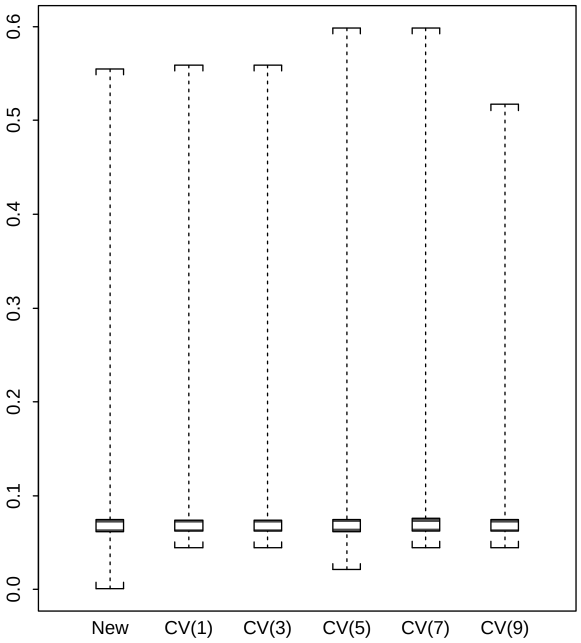


Figure 7(a)

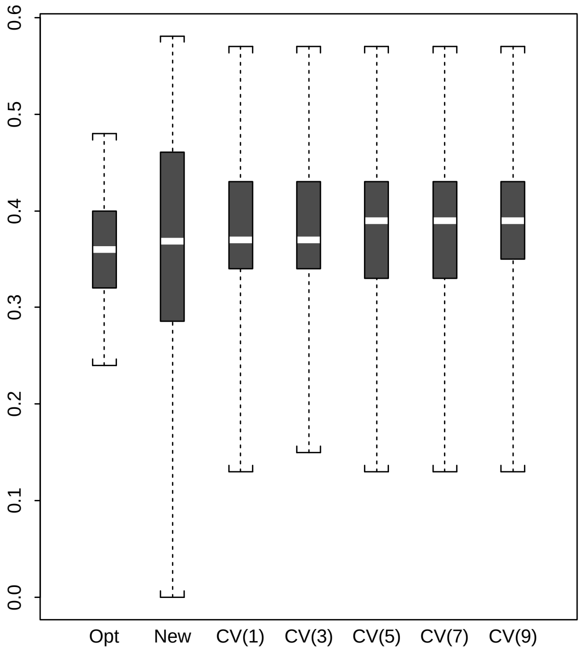


Figure 7(b)

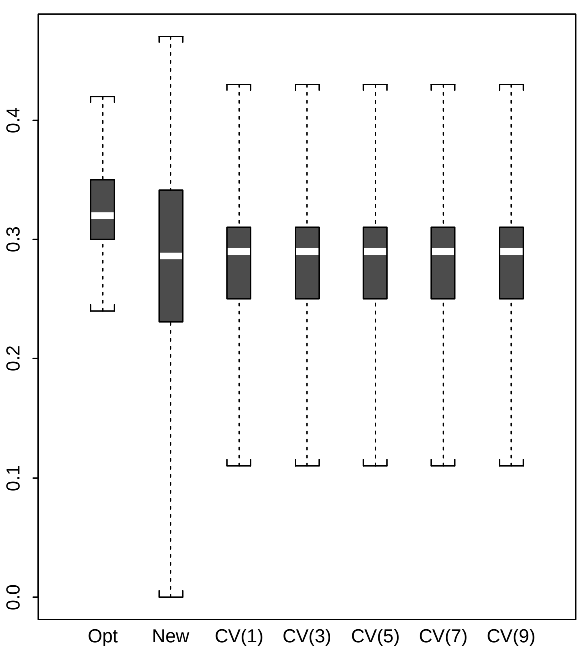


Figure 7(c)

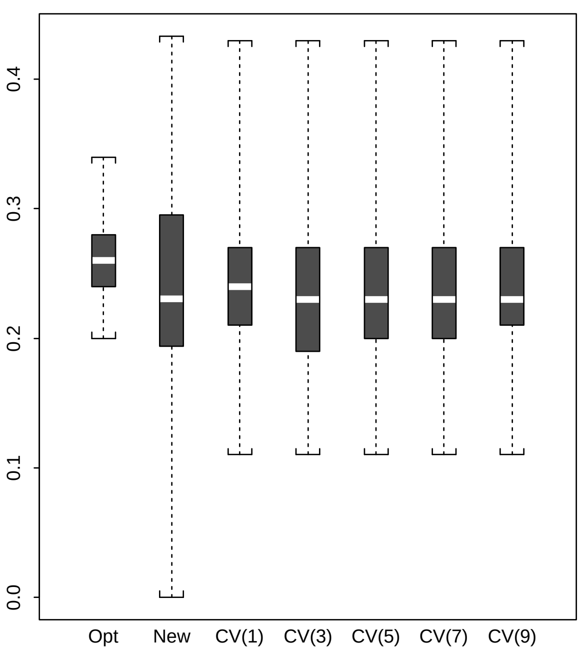


Figure 8(a)

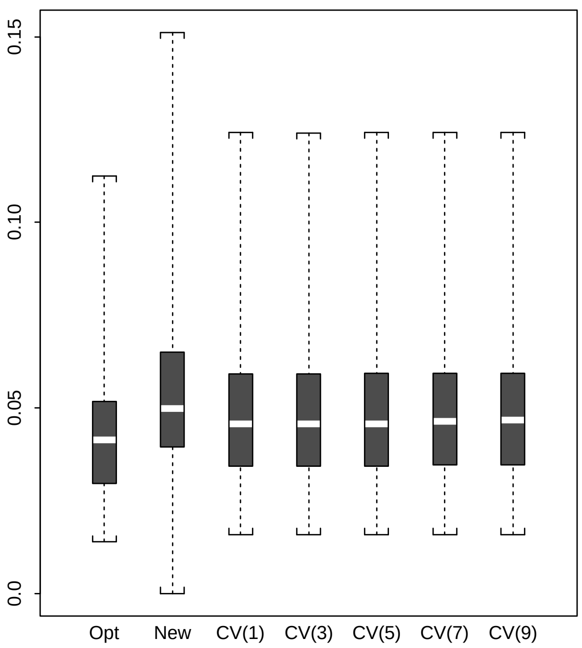


Figure 8(b)

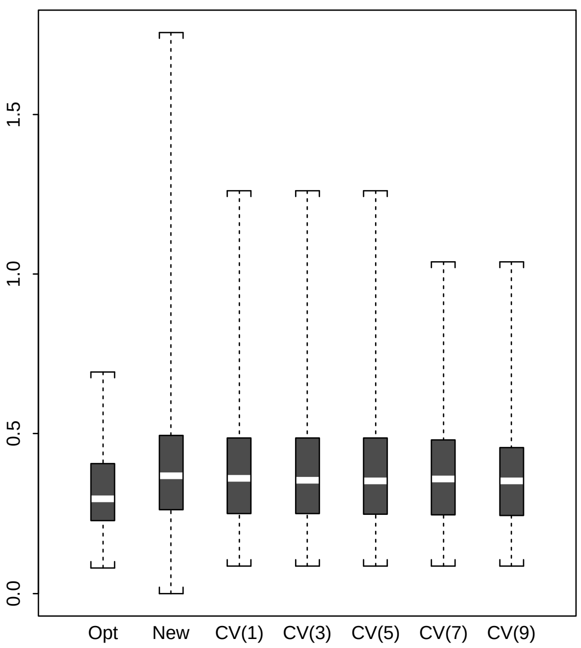
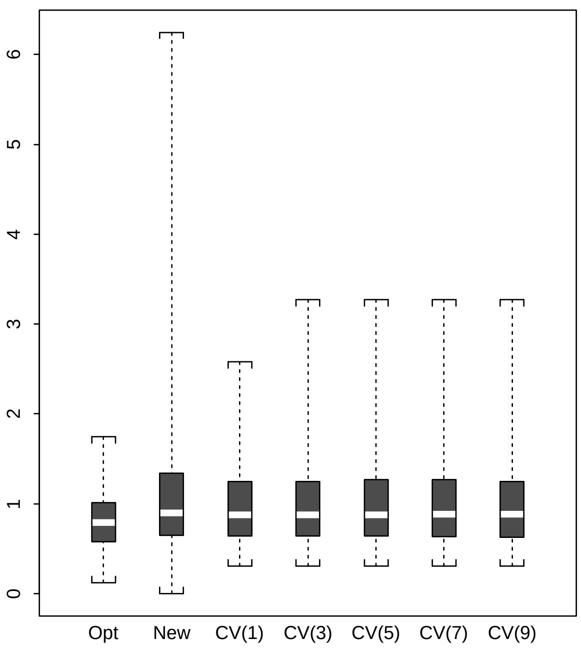


Figure 8(c)



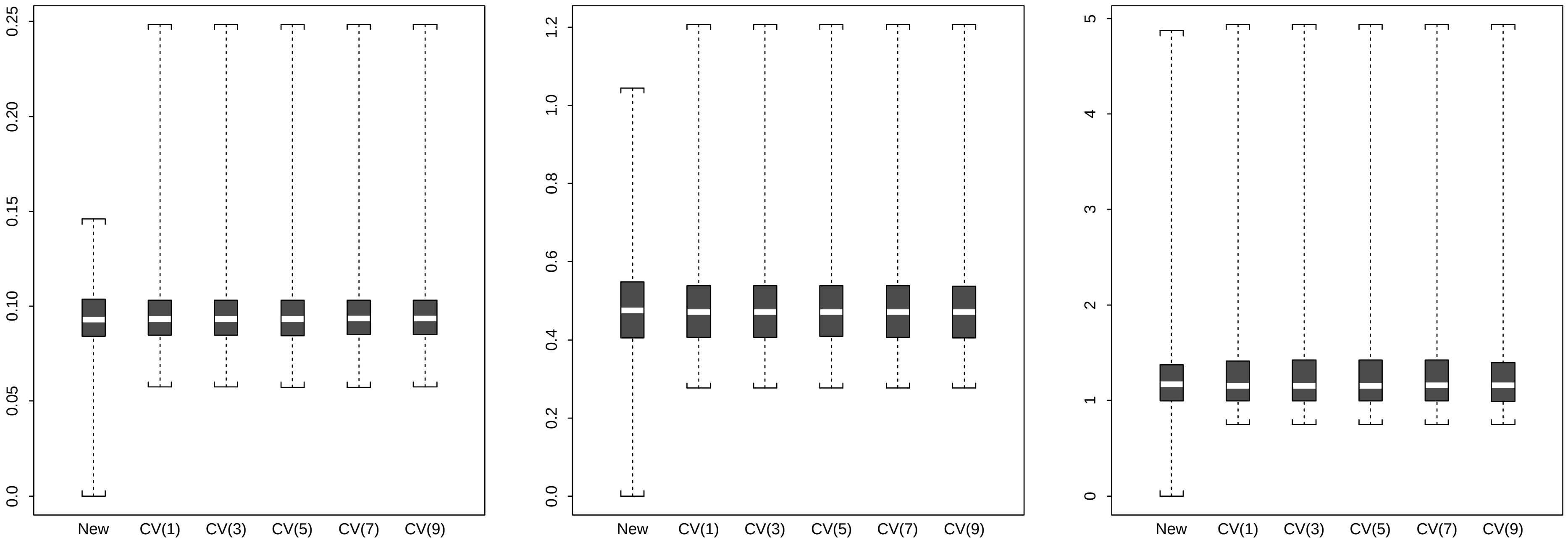


Figure 9(a)

Figure 9(b)

Figure 9(c)