cvmgof: an R package for Cramér-von Mises goodness-of-fit tests in regression models

Romain Azaïs, Sandie Ferrigno, Marie-José Martinez

▶ **To cite this version:**

Romain Azaïs, Sandie Ferrigno, Marie-José Martinez. cvmgof: an R package for Cramér-von Mises goodness-of-fit tests in regression models. Journal of Statistical Computation and Simulation, 2022, 92 (6), pp.1246-1266. 10.1080/00949655.2021.1991346 . hal-03101612

HAL Id: hal-03101612
https://hal.science/hal-03101612

Submitted on 7 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

cvmgof: an R package for Cramér-von Mises goodness-of-fit tests in regression models

R. Azaïs[a] and S. Ferrigno[b] and M.-J. Martinez[c]

[a] Inria, École Normale Supérieure de Lyon, France; [b] Université de Lorraine, CNRS, Inria, IECL, Nancy, France; [c] Univ. Grenoble Alpes, CNRS, LJK, Grenoble, France

ARTICLE HISTORY
Compiled January 7, 2021

ABSTRACT
Many goodness-of-fit tests have been developed to assess the different assumptions of a (possibly heteroscedastic) regression model. Most of them are "directional" in that they detect departures from a given assumption of the model. Other tests are "global" (or "omnibus") in that they assess whether a model fits a dataset on all its assumptions. We focus on the task of choosing the structural part of the regression function because it contains easily interpretable information about the studied relationship. We consider 2 nonparametric "directional" tests and one nonparametric "global" test, all based on generalizations of the Cramér-von Mises statistic. To perform these goodness-of-fit tests, we develop the `R` package `cvmgof` providing an easy-to-use tool for practitioners, available from the Comprehensive R Archive Network (CRAN). The use of the library is illustrated through a tutorial on real data. A simulation study is carried out in order to show how the package can be exploited to compare the 3 implemented tests.

KEYWORDS
Goodness-of-fit test; Cramér-von Mises statistic; nonparametric regression; model check; bandwidth; wild bootstrap; regression function

1. Introduction

Let consider $(X, Y) \in \mathbb{R}^2$ and a regression model to predict the value of Y from that of X,

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

where $m(\cdot)$ is the regression function (also referred to as the link function), $\sigma^2(\cdot)$ the variance function and ε the random error term. In addition to assumptions about the functional form of the regression function and the variance function, this model requires that the random error term is additive, independent of X, and often assumed to follow the Gaussian distribution $N(0, 1)$.

Methods to assess how well a model fits a set of observations fall under the banner of goodness-of-fit tests (see D'Agostino and Stephens [6] for a review of goodness-of fit techniques). In this paper, we focus on nonparametric goodness-of-fit tests that have

CONTACT S. Ferrigno. Email: sandie.ferrigno@univ-lorraine.fr

been developed to assess the different assumptions in models as (1). Most of them are "directional" in that they detect departures from mainly one given assumption of the model. For example, Alcalá *et al.* [2] and Van Keilegom *et al.* [26] have proposed tests to assess the assumption about the functional form of $m(\cdot)$; Dette [9] and Liero [20] have proposed tests to assess the homoscedasticity assumption, i.e., $\sigma(\cdot) \equiv \sigma$; Heuchenne and Van Keilegom [16] have developed a goodness-of-fit test for the form of the distribution of the error ε. However, when several directional tests are applied to the same model, each test requires the correctness of other assumptions. The assessment of the overall validity of the model may become a difficult problem to be solved. In particular, the practitioner faces the well-known multiple testing problem.

Other tests are "global" in that they assess whether a model fits a dataset on all its assumptions. For example, Andrews [4] has developed a global test based on a nonparametric estimator of the joint cumulative distribution function whereas Ducharme and Ferrigno [10] have proposed a test based on a nonparametric estimator of the conditional cumulative distribution function. Another global test developed by Zheng [28] is based on a nonparametric estimator of the conditional density. If the global test is not significant, one can consider using the model in practice. However, when the null hypothesis is rejected, it does not specify exactly where the difference is occurred and it can not be easy to determine which aspects of the null hypothesis are wrong.

In this paper, we focus on the task of choosing the structural part $m(\cdot)$ in models as (1). It gets most attention because it contains easily interpretable information about the relationship between X and Y. To validate the form of the regression function, we restrict ourselves to 3 tests based on a generalization of the Cramér-von Mises statistics [27]: the first 2 tests are the directional tests developed by Alcalá *et al.* [2] and Van Keilegom *et al.* [26] while the 3rd one is the global test proposed by Ducharme and Ferrigno [10].

To perform goodness-of-fit tests, several `R` packages have been developed providing an easy-to-use tool for many users. Package `goftest` [12] implements Cramér-von Mises and Anderson-Darling tests of goodness-of-fit for continuous univariate distributions with known parameters. Package `fgof` [17] implements classical empirical distribution function goodness-of-fit tests for one sample data with 2 bootstrap methods: multiplier bootstrap and parametric bootstrap. Another recent package is `goft` [14] which performs several goodness-of-fit tests for probability distributions with unknown parameters that are used in statistical modelling, such as the multivariate and univariate Gaussian distributions.

This paper is devoted to the presentation of the `R` package `cvmgof`, available from the Comprehensive R Archive Network (CRAN), which is – to the best of our knowledge – the first to implement the 3 aforementioned goodness-of-fit tests based on a generalization of the Cramér-von Mises statistic [2,10,26]. The library uses wild bootstrap [8] (particularly adapted to heteroscedastic regression models as (1)) to compute the critical test value, as well as an optimal choice of the bandwidth under the null hypothesis. `cvmgof` has been developed to be easy-to-use and perform each of the tests in one line of code with default parameters, while the output provides both the value of the test statistic, the decision of the test, the p-value, and the estimated optimal bandwidth (if applicable). The practitioner can also easily compare the test procedures with different kernel functions, bootstrap distributions, numbers of bootstrap replicates, or bandwidths.

The article is organized as follows. Section 2 is devoted to the presentation of preliminary concepts: the Cramér-von Mises statistic, wild bootstrap method and local

polynomial estimation. The 3 goodness-of-fit tests as well as bandwidth selection are discussed in Section 3. A tutorial on the `R` package `cvmgof` is given in Section 4. In Section 5, a simulation study is carried out to show how the package can be used to quantitatively compare the 3 test methods, with different parameters, in terms of statistical significance and power function. Section 6 is dedicated to concluding remarks.

2. Preliminaries

2.1. Cramér-von Mises test statistic

In the 3 tests of goodness-of-fit studied in this paper, the test statistic is a distance between a nonparametric estimator of the tested function and its form under the null hypothesis. There are many distance measures applicable to distribution functions. Like in [15,25], we restrict ourselves to a L^2-distance called Cramér-von Mises distance [5,27]. Consider a dataset $(Y_i)_{1\leq i\leq n}$ of real-valued independent and identically distributed (i.i.d.) variables with cumulative distribution function (c.d.f.) $F(\cdot)$. One aims to test the form of the c.d.f.,

$$H_0\colon F(\cdot) = F_0(\cdot) \quad \text{vs} \quad H_1\colon F(\cdot) \neq F_0(\cdot).$$

In this context, a nonparametric estimator of $F(\cdot)$ is the empirical c.d.f. defined, for any $y \in \mathbb{R}$, by

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i\leq y\}}.$$

The Cramér-von Mises test statistic related to test under consideration is given by

$$T_n = n\int \Big(\widehat{F}_n(y) - F_0(y)\Big)^2 w(y)F_0(dy),$$

where $w(\cdot)$ is a weight function satisfying regularity assumptions. It should be noted that if one chooses $w(\cdot) = (F_0(\cdot)(1-F_0(\cdot)))^{-1}$, we obtain the so-called Anderson-Darling statistic [3]. The 3 goodness-of-fit tests investigated in the present article are based on Cramér-von Mises statistics with selected weight functions equal to one.

The statistical testing procedure with significance α (one often takes $\alpha = 0.05$ or $\alpha = 0.01$) consists in comparing the value of the test statistic T_n to the $(1-\alpha)$-quantile of its distribution under H_0: if T_n is larger than the $(1-\alpha)$-quantile, H_0 can be considered unlikely and H_0 is rejected. Otherwise, H_0 can be accepted since the test fails in proving its unlikelihood. This requires to identify the distribution of the test statistic under H_0, which is often done through the approximation by the asymptotic regime $n \to \infty$. Nevertheless, the asymptotic distribution of T_n under H_0 often involves unknown parameters that need to be estimated to apply the procedure. For these reasons, a solution often considered in the literature is to compare the value of the test statistic with its empirical distribution obtained from a bootstrap sample, which does not require to estimate the distribution of T_n under H_0.

2.2. Wild bootstrap

In this subsection, we briefly present a wild bootstrap method for implementing a test on the regression function. Wild bootstrap has been shown to be particularly adapted to heteroscedastic models as (1). We refer the interested reader to [7].

Let $(X_i, Y_i)_{1\leq i\leq n}$ be a set of independent data satisfying model (1). From this dataset, we assume that we have a strategy to compute a nonparametric estimate $\widehat{m}_n(\cdot)$ of the regression function $m(\cdot)$ as well as a statistic T_n for testing the form of $m(\cdot)$,

$$H_0\colon m(\cdot) = m_0(\cdot) \quad \text{vs} \quad H_1\colon m(\cdot) \neq m_0(\cdot). \tag{2}$$

We introduce the notation $\epsilon_i = \sigma(X_i)\varepsilon_i$. The regression error ϵ_i can be estimated by

$$\widehat{\epsilon}_i = Y_i - \widehat{m}_n(X_i),$$

Roughly speaking, the idea of wild bootstrap is to build new variables of interest Y_i^b by modifying the additive error ϵ_i. More precisely, let $(U_i)_{1\leq i\leq n}$ be a sequence of i.i.d. variables with mean 0 and variance 1. For instance U_i may be distributed according to the Gaussian distribution, the Rademacher distribution [21], or the law proposed by Mammen in [22],

$$\mathbb{P}\left(U_i = -\frac{\sqrt{5}-1}{2}\right) = \frac{\sqrt{5}+1}{2\sqrt{5}} \qquad \text{and} \qquad \mathbb{P}\left(U_i = \frac{\sqrt{5}+1}{2}\right) = 1 - \frac{\sqrt{5}+1}{2\sqrt{5}}.$$

From $(U_i)_{1\leq i\leq n}$, we obtain a new dataset $(X, Y^b) = (X_i, Y_i^b)_{1\leq i\leq n}$ as

$$Y_i^b = \widehat{m}_n(X_i) + U_i\,\widehat{\epsilon}_i.$$

It should be remarked that neither the mean nor the variance of the regression error have been changed.

We replicate this procedure B times and obtain B datasets $(X, Y^b)_{1\leq b\leq B}$. From each bootstrap dataset (X, Y^b), we compute the nonparametric estimate $\widehat{m}_n^b(\cdot)$ of $m(\cdot)$ as well as the statistic T_n^b to test $m(\cdot) = \widehat{m}_n(\cdot)$, which measures the distance between $\widehat{m}_n(\cdot)$ and $\widehat{m}_n^b$. We now compare T_n to the distribution of the bootstrap statistic T_n^b. The test is rejected if the test statistic T_n is greater than the $(1-\alpha)$-quantile of the empirical distribution of the T_n^b's.

2.3. Local polynomial estimation

This subsection is dedicated to a short presentation of local polynomial estimation methods. An overview on these techniques for regression and variance functions can be found in [11].

We first focus on the problem of estimating the regression function $m(\cdot)$ by this type of methods, notably used in [2,26]. It will be useful to remark that $m(x) = \mathbb{E}(Y|X = x)$. If $m(\cdot)$ is smooth enough, the $(p+1)$-order Taylor expansion of $m(z)$ in a neighbourhood of a fixed point x can be expressed as:

$$m(z) \approx m(x) + (z-x)m^{(1)}(x) + \ldots + (z-x)^p \frac{m^{(p)}(x)}{p!},$$

with $m^{(\nu)}(x)$ the ν^{th} derivative function of $m(\cdot)$ evaluated in x. It should be noted that $m(z)$ minimizes in $r(z)$ the expression $\mathbb{E}((Y-r(z))^2\,|X=z)$. Consequently, one can estimate $m(x)$ through the following least squares problem,

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left[Y_i - \sum_{\nu=0}^{p} \beta_\nu (X_i - x)^\nu \right]^2 K\left(\frac{X_i - x}{h}\right), \tag{3}$$

where $K(\cdot)$ is a kernel function, and $h>0$ is a bandwidth that controls the size of the local neighbourhood. The solution of this optimization problem is explicit and given by

$$\widehat{\beta} = (\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{W} \mathbf{Y},$$

where $\mathbf{X}$ is the $n\times(p+1)$-matrix defined by $\mathbf{X}_{i,j} = (X_i - x)^j$, $\mathbf{W}$ is the $n\times n$ diagonal matrix defined by $W_{i,i} = K\left(\frac{X_i - x}{h}\right)$, and $\mathbf{Y}$ is the n column vector of the Y_i's.

Component $\widehat{\beta}_\nu$ of $\widehat{\beta}$ is an estimate of $\dfrac{m^{(\nu)}(x)}{\nu!}$. In particular, $\widehat{\beta}_0$ is an estimate of $m(x)$. Thus, $\widehat{m}_n(x)$ can be defined as

$$\widehat{m}_n(x) = \widehat{\beta}_0 = \sum_{i=1}^{n} W_{n,p}\left(\frac{X_i - x}{h}\right) Y_i,$$

with

$$W_{n,p}(t) = (1,0,\ldots,0)(\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-1}(1, ht, \ldots, (ht)^p)^\intercal K(t).$$

It can be remarked that setting $p=0$ in the above gives back Nadaraya-Watson estimator (also called constant adjustment) notably used in [26],

$$\widehat{m}_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}. \tag{4}$$

With $p=1$, we obtain the local linear estimator of $m(x)$ used in [2],

$$\widehat{m}_n(x) = \frac{r_{n,2}(x) f_{n,0}(x) - r_{n,1}(x) f_{n,1}(x)}{r_{n,0}(x) r_{n,2}(x) - r_{n,1}^2(x)}, \tag{5}$$

with, for $j\in\{0,1,2\}$,

$$\begin{aligned} r_{n,j}(x) &= \sum_{i=1}^{n} (X_i - x)^j K\left(\frac{X_i - x}{h}\right), \\ f_{n,j}(x) &= \sum_{i=1}^{n} (X_i - x)^j K\left(\frac{X_i - x}{h}\right) Y_i. \end{aligned}$$

Local polynomial methods have been adapted to the conditional distribution function $F(y|x) = \mathbb{E}(\mathbb{1}_{\{Y\leqslant y\}}|X=x)$ in [10]. For $y\in\mathbb{R}$ and some z such as $F(y|z)$ is a

regular enough function, the $(p+1)$-order Taylor expansion of $F(y|z)$ in the neighbourhood of a fixed point x can be expressed as

$$F(y|z) \approx F(y|x) + (z-x)F^{(1)}(y|x) + \ldots + (z-x)^p \frac{F^{(p)}(y|x)}{p!},$$

with $F^{(\nu)}(y|x)$ the ν^{th} derivative function of $F(y|z)$ with respect to z, evaluated in x. Note also that $F(y|z)$ minimizes in $r(y|z)$ the expression $\mathbb{E}\left((\mathbb{1}_{\{Y\leq y\}} - r(y|z))^2 | Z = z\right)$. This leads to the same least squares problem as in local polynomial regression estimation where Y_i is replaced by $\mathbb{1}_{\{Y_i\leq y\}}$. Following the same reasoning, we obtain

$$\widehat{F}_n(y|x) = \sum_{i=1}^{n} W_{n,p}\left(\frac{X_i - x}{h}\right)\mathbb{1}_{\{Y_i\leq y\}}. \tag{6}$$

With $p=1$, we obtain the local linear estimator of $F(y|x)$ used in [10].

3. Cramér-von Mises goodness-of-fit tests for regression models

3.1. *Alcalá* et al. *test*

The goodness-of-fit test developed by Alcalá *et al.* in [2] verifies the form (2) of the regression function $m(\cdot)$ in the model (1). The test statistic is defined as

$$T_n = n\sqrt{h}\int \left(\widehat{m}_n(x) - m_0(x)\right)^2 dx,$$

where $\widehat{m}_n(\cdot)$ is the local linear estimator of the regression function given in (5).

A theoretical study of the asymptotic behaviour of T_n under H_0 is presented in [2, Theorem 2.1]. In our investigation, we will apply the test via the wild bootstrap procedure presented in Subsection 2.2, which does not present any difficulty.

3.2. *Van Keilegom* et al. *test*

The goodness-of-fit test investigated by Van Keilegom *et al.* in [26] verifies the form (2) of the regression function $m(\cdot)$ in the model (1). The keystone is that the null hypothesis holds if and only if the random variables $\varepsilon = (Y - m(X))/\sigma(X)$ and $\varepsilon_0 = (Y - m_0(X))/\sigma(X)$ have the same distribution. The test consists in estimating and comparing these 2 distributions. We define the nonparametric residuals by

$$\widehat{\varepsilon}_i = \frac{Y_i - \widehat{m}_n(X_i)}{\widehat{\sigma}_n(X_i)}$$

where $\widehat{m}_n(\cdot)$ is the Nadaraya-Watson estimator of the regression function (4) and

$$\widehat{\sigma}_n^2(\cdot) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i^2}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} - \widehat{m}_n^2(\cdot) \tag{7}$$

is the Naradaya-Watson estimator of the variance function $\sigma^2(\cdot)$. This leads to the following estimate of the c.d.f. of ε,

$$\widehat{F}_{\varepsilon}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\widehat{\varepsilon}_i \leq y\}}.$$

On the other hand, the distribution of ε_0 can be estimated through

$$\widehat{F}_{\varepsilon_0}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\widehat{\varepsilon}_{0,i} \leq y\}},$$

where the $\widehat{\varepsilon}_{0,i} = (Y_i - m_0(X_i))/\widehat{\sigma}_n^2(X_i)$'s are the estimated residuals under H_0. In this framework, the test statistic is given by

$$T_n = n \int \left(\widehat{F}_{\varepsilon}(y) - \widehat{F}_{\varepsilon_0}(y)\right)^2 d\widehat{F}_{\varepsilon_0}(y).$$

The behaviour of the test statistic is studied from a theoretical point of view in [26, Corollary 3.2 (null hypothesis) and Theorem 3.2 (alternative hypothesis)]. The application of wild bootstrap is straightforward (see Subsection 2.2). We refer the reader to [26, 4. Simulations and data analysis] for a different bootstrap approach based on the residual distribution.

3.3. Ducharme and Ferrigno test

The goodness-of-fit test developed in [10] tests the form of the conditional c.d.f. under model (1),

$$H_0 : F(\cdot|\cdot) = F_0(\cdot|\cdot) \quad \text{vs} \quad H_1 : F(\cdot|\cdot) \neq F_0(\cdot|\cdot).$$

The conditional c.d.f. involves both the link m and variance σ^2 functions. In other words, the test concerns the whole model (1) and not only the form of the link function. For this reason, the test is said global or omnibus. The test statistic under consideration is

$$T_n = n\sqrt{h} \int \int \left(\widehat{F}_n(y|x) - F_0(y|x)\right)^2 F_0(dy|x)dx$$

where $\widehat{F}_n(y|x)$ is the local linear estimator of the conditional c.d.f. given by (6) with $p = 1$. In practice, T_n can be calculated as $T_n = n\sqrt{h}\int t_n(x)dx$, with

$$\begin{aligned} t_n(x) &= \frac{1}{3} + \sum_{i=1}^{n-1} \widehat{F}_n^2(Y_{(i)}|x)\left(F_0(Y_{(i+1)}|x) - F_0(Y_{(i)}|x)\right) \\ &- \sum_{i=1}^{n-1} \widehat{F}_n(Y_{(i)}|x)\left(F_0^2(Y_{(i+1)}|x) - F_0^2(Y_{(i)}|x)\right) \\ &+ \left(F_0^2(Y_{(n)}|x) - F_0(Y_{(n)}|x)\right), \end{aligned}$$

where $Y_{(i)}$ denotes the i^{th} order statistic.

One of the main differences with the 2 other tests under consideration in this paper is that the definition of the test statistic does not require an estimate of the regression function but an estimate of the conditional c.d.f.

The behaviour of the test statistic is studied from a theoretical point of view in [10, Theorem 3.1 (null hypothesis) and Theorem 4.1 (alternative hypothesis)]. The application of wild bootstrap is straightforward (see Subsection 2.2). To be consistent, we propose to implement the wild bootstrap using the estimate of the regression function naturally obtained from the one of the conditional c.d.f. at stake in this test,

$$\begin{aligned} \widehat{m}_n(x) &= \int y \, d\widehat{F}_n(y|x) \\ &= Y_{(n)}\widehat{F}_n\left(Y_{(n)}|x\right) - Y_{(1)}\widehat{F}_n\left(Y_{(1)}|x\right) - \sum_{i=1}^{n-1}\left(Y_{(i+1)} - Y_{(i)}\right)\, \widehat{F}_n\left(Y_{(i)}|x\right). \quad (8) \end{aligned}$$

3.4. Choice of the bandwidth parameter

The bandwidth h is a crucial parameter in nonparametric estimation since it determines the degree of smoothing of the estimator and therefore in particular its complexity. This choice is critical, as under or over smoothing can reduce precision. An important literature has been devoted to the choice of the bandwidth in this context [1,11,13,18,19].

Theoretical optimal bandwidth can be obtained in nonparametric estimation of the regression function by minimizing the asymptotic weighted mean squared error [11] and in nonparametric estimation of the conditional distribution function by minimizing the asymptotic weighted integrated mean squared error [13]. These quantities, however, depend on unknown parameters that can be estimated in practice by cross-validation [19] or plug-in methods [18].

In the context of goodness-of-fit tests based on nonparametric estimation of the error distribution [1], the bandwidth parameter is chosen such that the empirical risk of the nonparametric test is closest to the significance level. This approach is of great interest because it is based on the accuracy of the test (in terms of risk) and not on the quality of the estimation of the link function (that only represents an intermediate step in the whole test procedure). Nevertheless, this results in a heavy computational cost. In this work, we propose to take into account the fact that the practitioner aims to compare the link function of the data to a given specific form, i.e., make use of the null hypothesis as in [1], but with a lower computational cost. The bandwidth parameter is selected such that the nonparametric estimation of the regression function (within the test under consideration) is the nearest to the tested regression function under the null hypothesis,

$$h^\star = \arg\min_h \|\widehat{m}_n^0 - m_0\|_2^2,$$

where $\widehat{m}_n^0$ has been computed from a virtual dataset of size n of the form (X_i, Y_i^0) using the same X_i's as in the dataset under consideration and where the Y_i^0's were simulated under H_0. $\widehat{m}_n^0$ is obviously the nonparametric estimator of the link function considered in the test at stake, i.e., (5) for Alcala *et al.* test and (4) for Van Keilegom *et*

al. test. The situation is a bit different for Ducharme and Ferrigno test, since the test statistic does not use a nonparametric estimator of the link function. In this case, we use the nonparametric estimator of the link function (8) deduced from the local linear estimator of the conditional cumulative distribution function investigated within this test.

It should be remarked that, when h tends to infinity in (3), the argument of the minimum $\widehat{\beta}$ does not depend on x, which implies that the estimate of the link function is polynomial with degree p. Consequently, when the link function under H_0 is a polynomial of the same degree p, the method presented above tends to select a bandwidth as large as possible. In other words, this bandwidth selection algorithm should not be used with linear null hypothesis for Alcala *et al.* and Ducharme and Ferrigno tests.

4. The cvmgof package

The purpose of this section is to present a tutorial on `cvmgof`, an `R` package developed to perform the 3 Cramér-von Mises goodness-of-fit tests for regression models of the form (1) presented in Section 3, namely Alcala *et al.* test [2] (abbreviated in `acgm`, color code: red), Van Keilegom *et al.* test [26] (`vkgmss`, color code: green), and Ducharme and Ferrigno test [10] (`df`, color code: blue). The library is available from the Comprehensive R Archive Network (CRAN) following this link: `https://CRAN.R-project.org/package=cvmgof`. Once it has been downloaded and installed, the library can be loaded in the `R` session with:

```
library(cvmgof)
```

In order to illustrate the functionalities of the package, we consider the `Boston` dataset provided in the `MASS` package, that collects 506 data of dimension 14 related to housing values in the suburbs of Boston, MA, USA. References can be found in the documentation of the library. In this tutorial, we arbitrarily restrict ourselves to the link between 2 of the 14 variables: `MEDV`, i.e., the median value of owner-occupied homes in $1000's, and `LSTAT`, i.e., the lower status of the population. The 2 variables of the dataset under consideration can be obtained as:

```
library(MASS)
X = Boston$medv
Y = Boston$lstat
```

We estimate the link function that connects the variables `X` and `Y` using polynomial local estimators considered in this paper. First, we use the estimator at stake in Alcala *et al.* test:

```
xgrid = seq(5,50,by=0.1)
lf_acgm = acgm.linkfunction.estim(xgrid, X, Y, bandwidth=7.5)
```

The link function is estimated along the x-axis through the regular grid defined by `xgrid`. The bandwidth value is arbitrarily set to 7.5 by the user. The default kernel is Epanechnikov. Similarly, one can estimate the link function using methods developed in Van Keilegom *et al.* and Ducharme and Ferrigno tests with different bandwidths or kernel functions:

```
lf_vkgmss = vkgmss.linkfunction.estim(xgrid, X, Y, 7.5,
  kernel.function=kernel.function.quart)
lf_df = df.linkfunction.estim(xgrid, X, Y, 12.5)
```

The following piece of code displays the dataset as well as the 3 estimates above:

```
plot(X,Y,pch='+',xlim=c(5,50),ylim=c(0,40),
  col='gray',xlab='MEDV',ylab='LSTAT')
lines(xgrid,lf_acgm,lwd=3,col='red')
lines(xgrid,lf_vkgmss,lwd=3,col='dark green')
lines(xgrid,lf_df,lwd=3,col='blue')
```

We aim to test if the values `Y` are connected to `X` through a polynomial of order 2, which parameters are estimated below:

```
reg = lm(Y~I(X)+I(X^2))
lf_poly2 = reg$coefficients[1]+reg$coefficients[2]*xgrid
  +reg$coefficients[3]*xgrid^2
lines(xgrid,lf_poly2,lwd=3,lty=3)
```

Consequently, we shall test the form of the link function (2) with

$$m_0(x) = 39.04764268 - 1.71502726\,x + 0.02068709\,x^2. \tag{9}$$

The last code line adds the estimated polynomial of order 2 to the plot (see Fig. 1). We need to define the link function under H_0 to perform Alcala *et al.* and Van Keilegom *et al.* tests, which can be done as:

```
lf.H0 = function(x){
  reg$coefficients[1]+reg$coefficients[2]*x+reg$coefficients[3]*x^2
}
```

Alcala *et al.* test with significance value 0.05 and optimal bandwidth is easily performed in one line of code:

```
set.seed(pi)
test_acgm = acgm.test.bootstrap(X, Y, lf.H0, 0.05, bandwidth='optimal',
  bootstrap=c(50,'Mammen'))
```

Even if the dataset is not random here, the bootstrap method adds some randomness in the procedure, which is fixed here by arbitrarily setting the seed to π. Consequently, if one runs several times the same code, one should obtain identical results. One can also obtain the optimal bandwidth under H_0 (with the method described in Subsection 3.4) as well as the test statistic step by step:

```
set.seed(pi)
Y.H0 = lf.H0(X) + rnorm(length(X), mean=0, sd=sd(residuals(reg)))
hopt_acgm = acgm.bandwidth.selection.linkfunction(X, Y.H0, lf.H0)
acgm_stat = acgm.statistics(X, Y, lf.H0, hopt_acgm)
```

It should be noted that the generation of virtual data under H_0 requires to estimate the variance of the residuals (in a homoscedastic way here). Van Keilegom *et al.* test can be applied in the same way:

```
set.seed(pi)
test_vkgmss = vkgmss.test.bootstrap(X, Y, lf.H0, 0.05, bandwidth='optimal',
  bootstrap=c(25,'Rademacher'))
```

Here the bootstrap parameters have been set by the user: 25 wild bootstrap replicates using the Rademacher distribution. Default value is 50 replicates with Mammen distribution. Ducharme and Ferrigno test deals with the conditional cumulative distribution function but the estimation of the optimal bandwidth is done from the related link function as for the 2 other tests (see Subsection 3.4). Consequently, the link function under H_0 must be passed as an argument to obtain the optimal bandwidth under H_0:

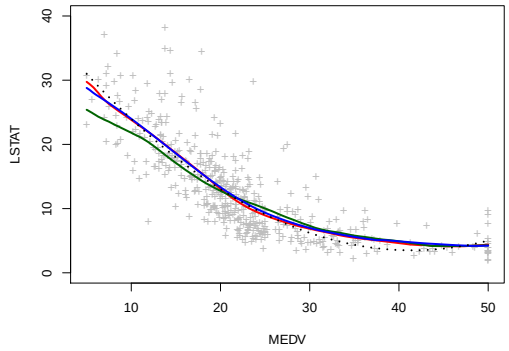

Figure 1. `LSTAT` vs. `MEDV` in `Boston` dataset, local polynomial estimates of the link function (`acgm` in red, `vkgmss` in green, and `df` in blue), and link function under H_0 (black dashed line).

	`acgm`	`vkgmss`	`df`
dec.	accept	accept	accept
$h^\star$	8.98	4.01	8.23
p-val.	0.46	1	0.18
stat.	117339	0.08	620.75

Table 1. Results obtained from the 3 tests `acgm`, `vkgmss` and `df` when testing (2) with m_0 defined in (9).

```
hopt_df = df.bandwidth.selection.linkfunction(X, Y.H0, lf.H0)
```

The following script defines the conditional cumulative distribution function under H_0 and runs the test with the optimal bandwidth obtained at the previous step:

```
cond_cdf.H0 = function(x,y){
  out = matrix(0, nrow=length(x), ncol=length(y))
  for (i in 1:length(x)){
    out[i,] = pnorm(y-lf.H0(x[i]), 0, sd(residuals(reg)))
  }
  out
}
set.seed(pi)
test_df = df.test.bootstrap(X, Y, cond_cdf.H0, 0.05, hopt_df,
  integration.step=0.1)
```

The integration step can be set by the user, here to 0.1, while the default value is 0.01. Finally, the results obtained from the 3 tests are contained in the objects `test_acgm`, `test_vkgmss`, and `test_df`, and gathered in Tab. 1. They all conclude that H_0 can be accepted.

5. Numerical experiments

In this section, we show how the `cvmgof` package can be used to compare quantitatively and qualitatively the behaviour of the 3 tests, from different models (homoscedastic in Subsection 5.1, heteroscedastic in Subsection 5.2, and heteroscedastic with undisclosed variance function in Subsection 5.3) with varying sample sizes, using different kernels and resampling procedures of wild bootstrap.

5.1. Homoscedastic polynomial model

Regression model. In this first experiment, the model under consideration is defined as

$$Y = aX^2 + 5X + \varepsilon,$$

where X is uniformly distributed on the interval $[0, 1]$, $\varepsilon \sim N(0, 1)$, and X and ε are independent. The null hypothesis consists in the parametric model $a = 5$, i.e.,

$$H_0 \text{: } m(x) = 5x^2 + 5x.$$

The parameter a will be used to quantify the deviation from the null hypothesis and will go from 1 to 10 in the numerical results below.

Experimentations. The objective of this simulation study is to test H_0 through the exhaustive list of combinations of parameters, i.e., from the 3 tests, from 3 kernel functions (Epanechnikov, Gaussian and quartic kernels), and from the 3 wild bootstrap procedures given in Subsection 2.2 (each with 50 bootstrap replicates). The bandwidth is selected as a function of the data and of H_0 as explained Subsection 3.4. In addition, we aim to compare the results from 3 different sample sizes: $n = 50$, 100, and 200. For each combination of factors, the experiments have been replicated 50 times.

Numerical results. The empirical results are organized as follows:

- Figs. 2, 3, and 4: power functions as a function of parameter a, estimated from 50 simulated samples. Each figure presents the results for the 3 tests, 3 sample sizes, and 3 kernel functions, but for a unique wild bootstrap procedure (Mammen wild bootstrap in Fig. 2, Rademacher wild bootstrap in Fig. 3, and Gaussian wild bootstrap in Fig. 4).
- Figs. 5, 6, and 7: boxplots of optimal bandwidths as a function of parameter a, each of them being estimated from 50 simulated samples. Each figure presents the results for the 3 tests and 3 kernel functions, but for a unique sample size ($n = 50$ in Fig. 5, $n = 100$ in Fig. 6, and $n = 200$ in Fig. 7).
- Figs. 8 and 9: 50 estimates of the link function from $n = 200$ data (under the null hypothesis $a = 5$ in Fig. 8 and under the alternative hypothesis $a = 8$ in Fig. 9) for the 3 tests and 3 kernel functions.

It should be noted that the results presented in Figs. 5, 6, 7, 8, and 9 do not involve the notion of bootstrap.

Discussion on test accuracy from Figs. 2, 3, and 4.

- First of all, it can be remarked that, as expected, the larger the sample size, the more accurate the estimated power functions. The results obtained from a sample size equal to 50 are not satisfying regardless of the test, the kernel adjustment and the wild bootstrap method used, while the 3 tests perform really well only from a sample size equal to 200. In particular, the Ducharme and Ferrigno global test requires a sample size of 200 to perform qualitatively as well as the 2 directional tests.
- From this model, we can not observe any clear effect of the kernel function on the test results.
- The Mammen and Gaussian distributions in wild bootstrap resampling lead, for the 3 tests, to very good results when $n = 200$ compared to the Rademacher wild bootstrap. More precisely, the empirical risk under H_0 obtained using the Rademacher wild bootstrap method is further from the theoretical risk of 5% than from the 2 other distributions.
- With the Gaussian bootstrap method, the Van Keilegom *et al.* test yields less accurate power functions than with the 2 other bootstrap distributions, and than from the 2 other tests as well. As a consequence, it seems that Gaussian resampling is not adapted to this test procedure.

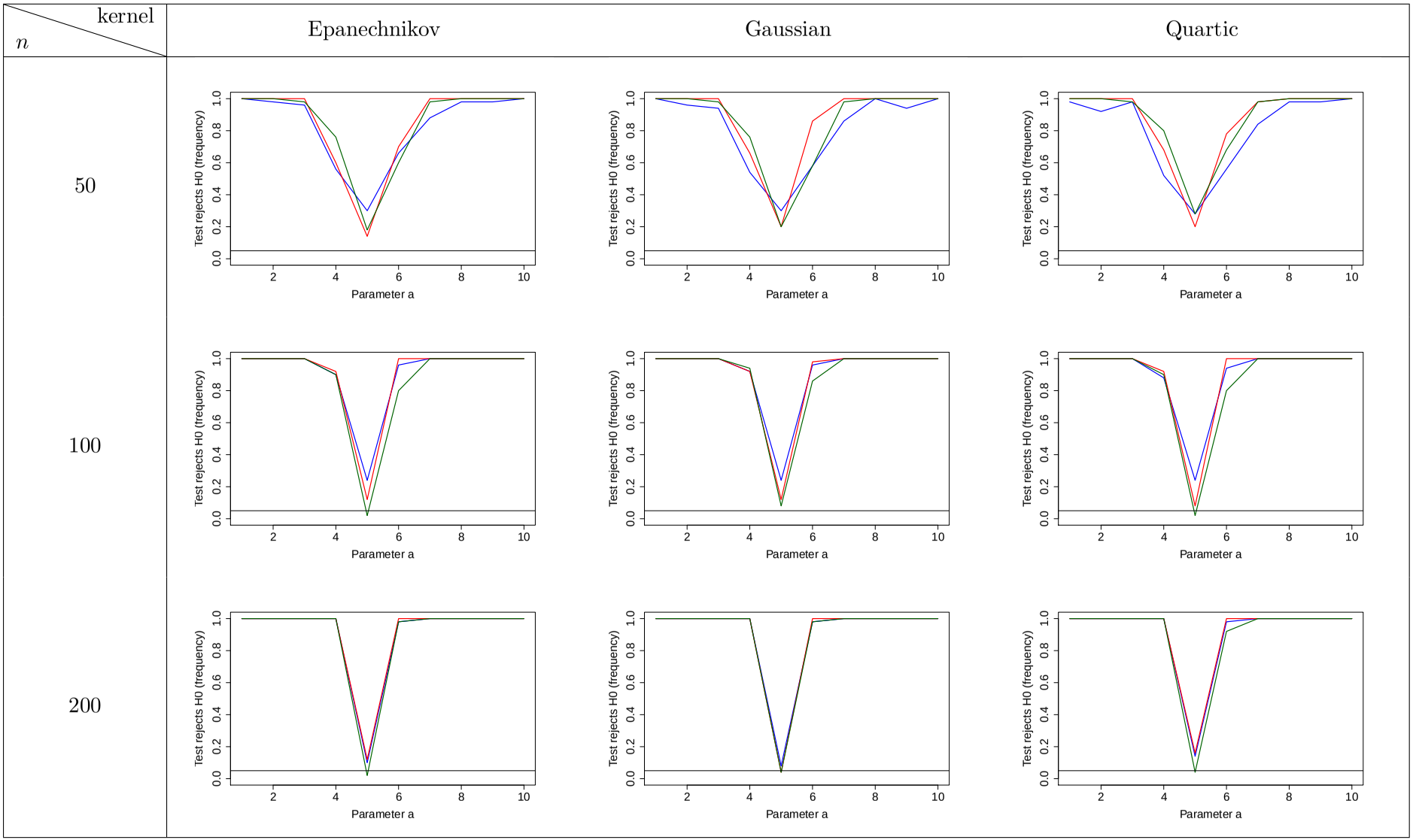

Figure 2. Empirical power functions of the 3 tests under consideration (Alcalá *et al.* test in red, Van Keilegom *et al.* test in green, and Ducharme and Ferrigno test in blue) computed from 50 simulated samples. Mammen wild bootstrap resampling has been applied.

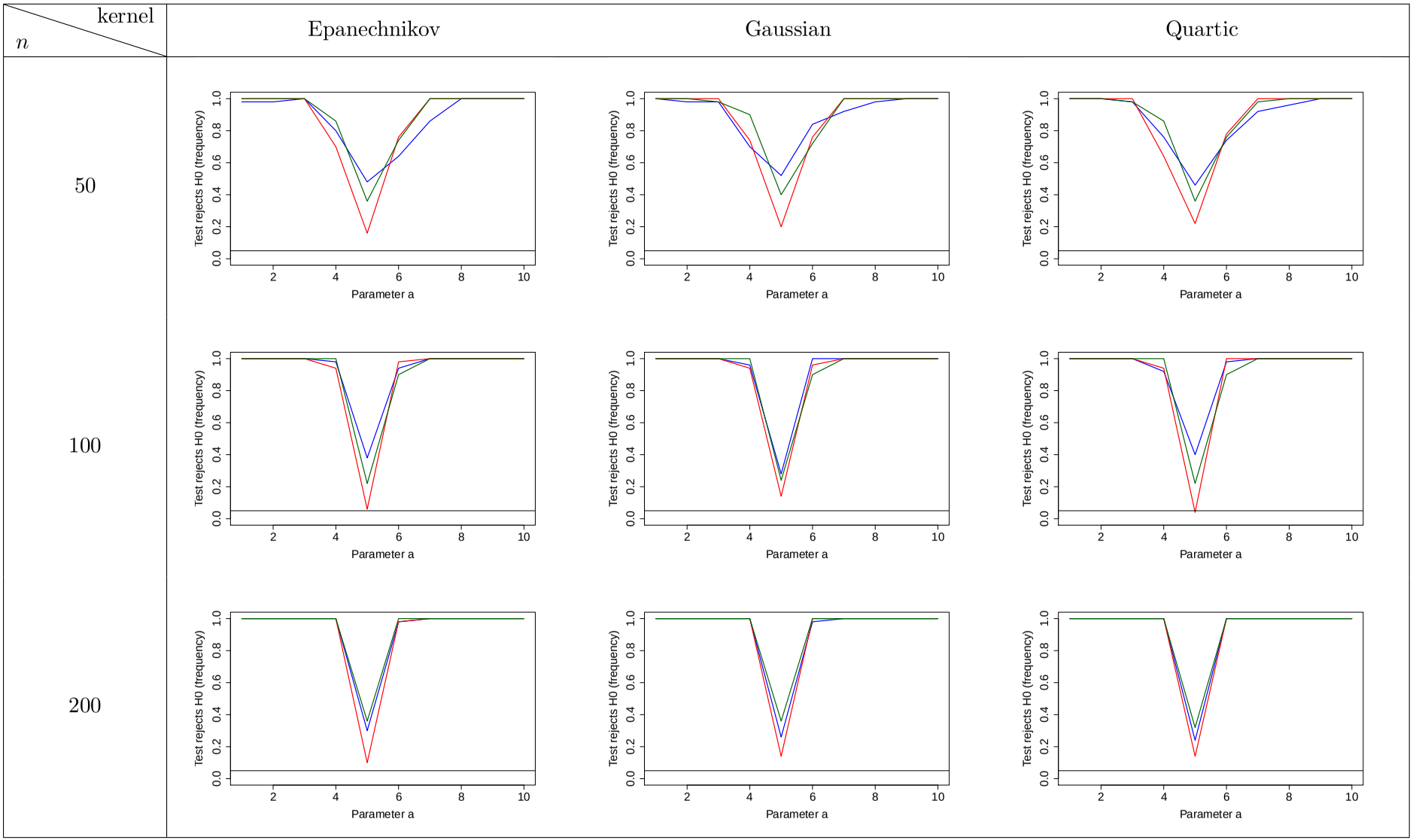

Figure 3. Empirical power functions of the 3 tests under consideration (Alcalá *et al.* test in red, Van Keilegom *et al.* test in green, and Ducharme and Ferrigno test in blue) computed from 50 simulated samples. Rademacher wild bootstrap resampling has been applied.

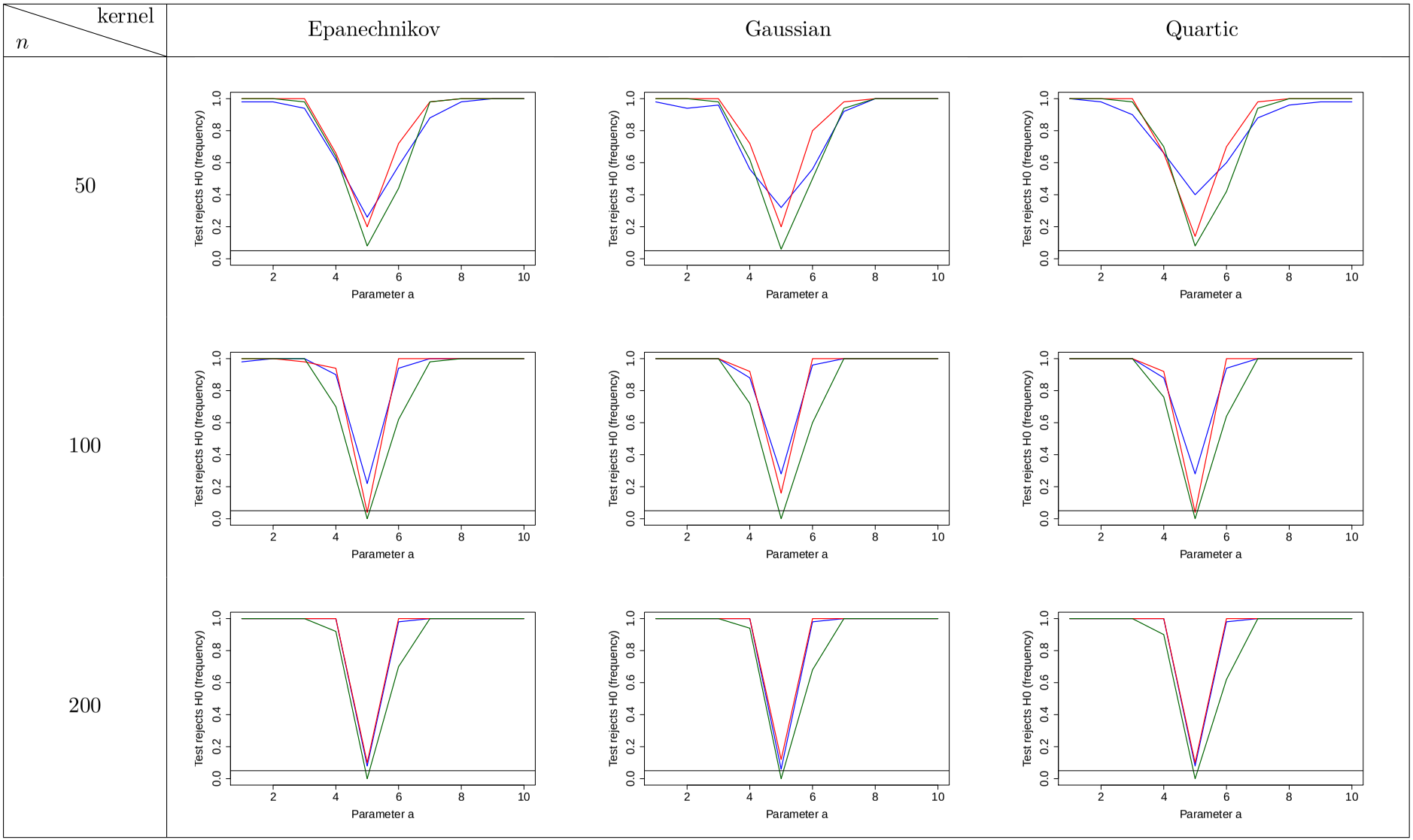

Figure 4. Empirical power functions of the 3 tests under consideration (Alcalá *et al.* test in red, Van Keilegom *et al.* test in green, and Ducharme and Ferrigno test in blue) computed from 50 simulated samples. Gaussian wild bootstrap resampling has been applied.

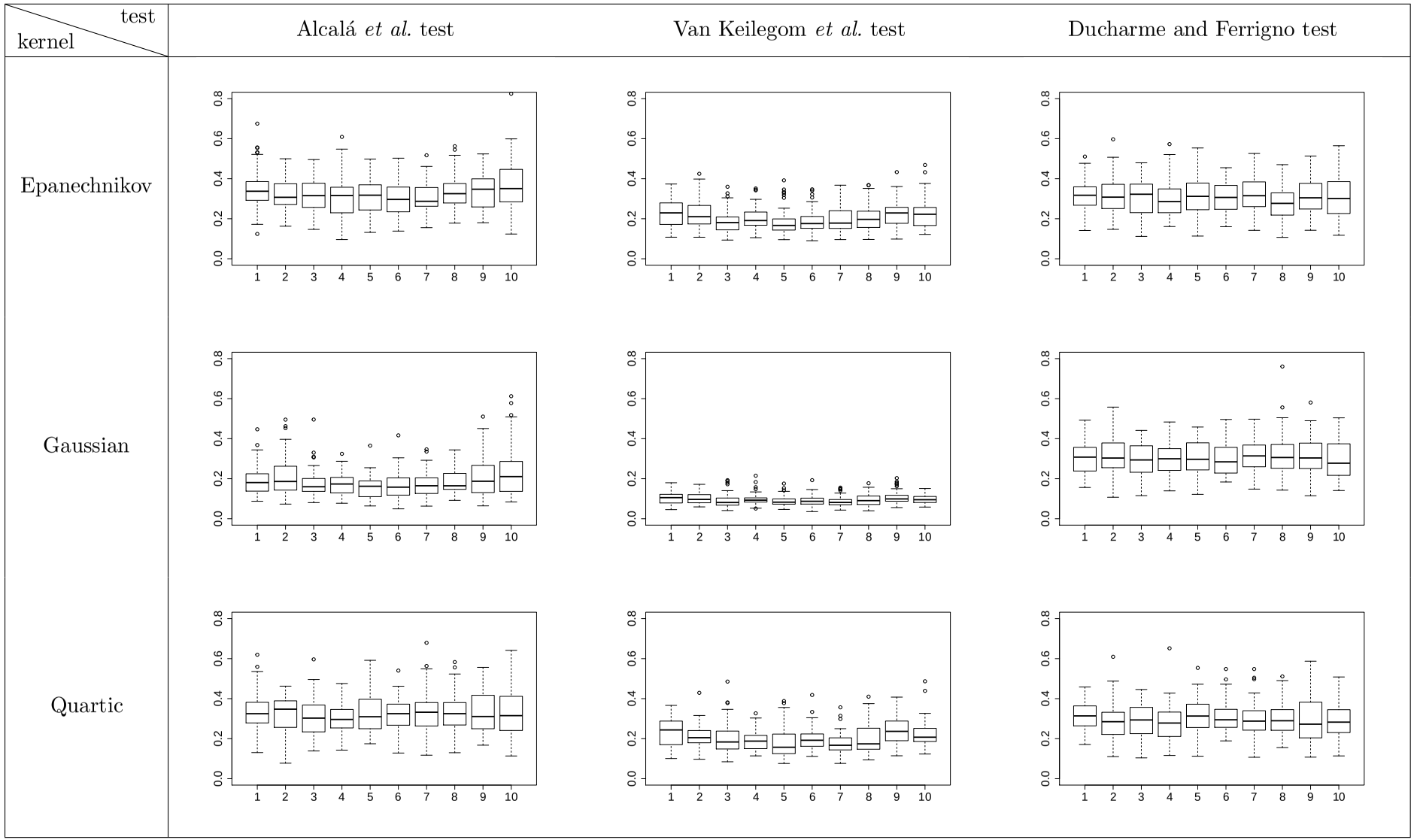

Figure 5. Boxplots of optimal bandwidth parameters as a function of parameter a, each of them being estimated from 50 samples of size $n = 50$, for the 3 tests and 3 kernel functions.

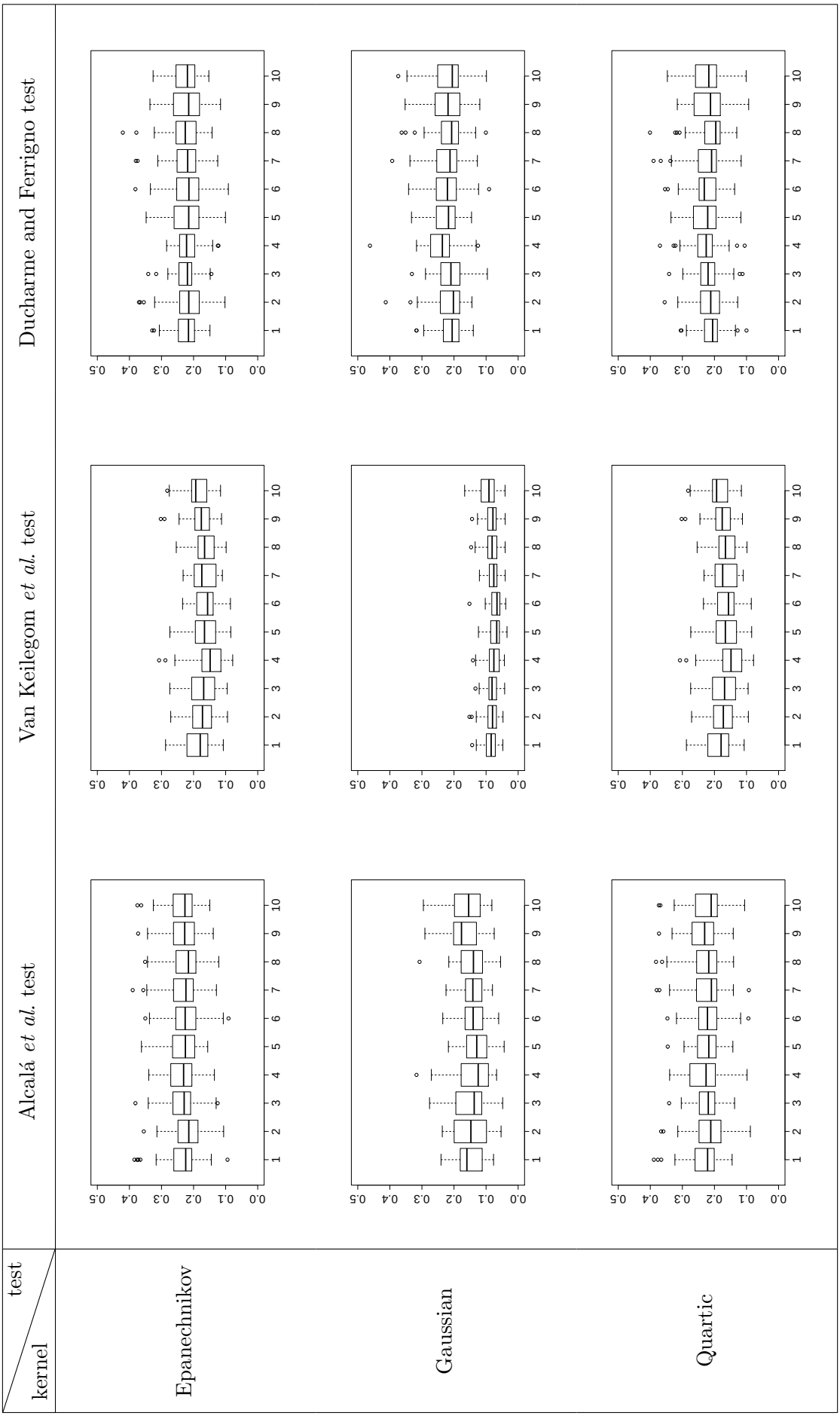

Figure 6. Boxplots of optimal bandwidth parameters as a function of parameter a, each of them being estimated from 50 samples of size $n = 100$, for the 3 tests and 3 kernel functions.

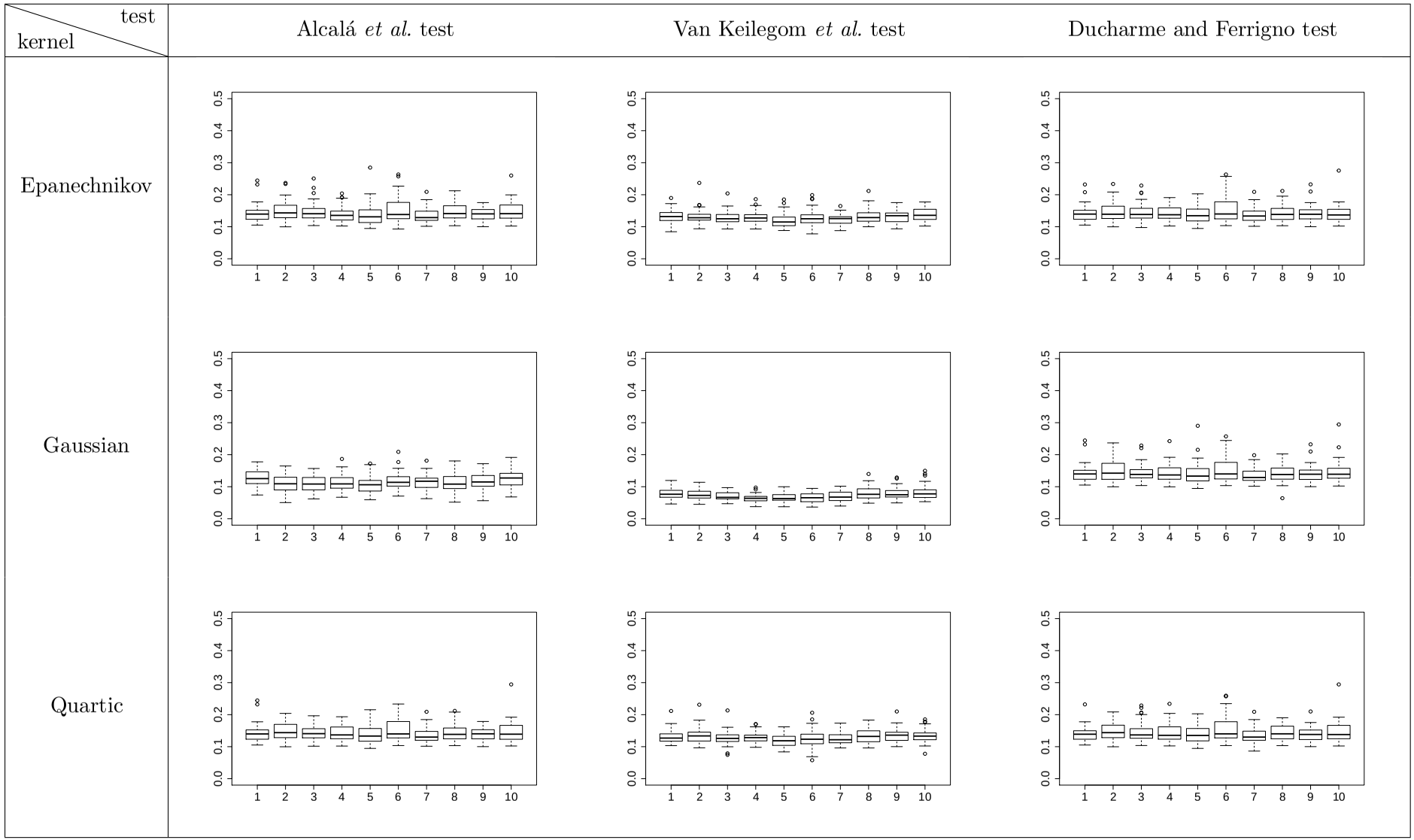

Figure 7. Boxplots of optimal bandwidth parameters as a function of parameter a, each of them being estimated from 50 samples of size $n = 200$, for the 3 tests and 3 kernel functions.

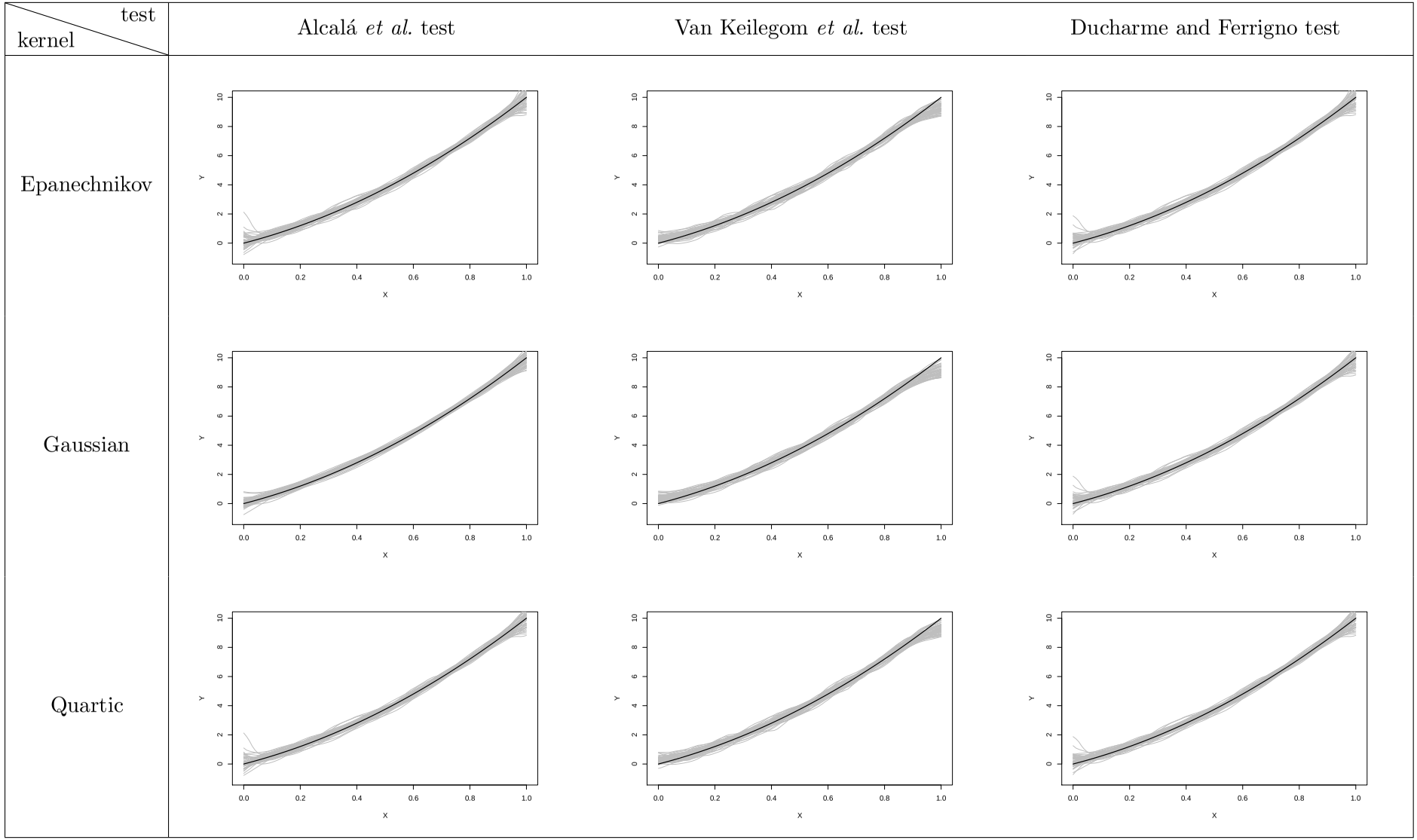

Figure 8. 50 estimates of the link function under the null hypothesis $a = 5$ from the 3 tests and 3 kernel functions and from samples of size $n = 200$.

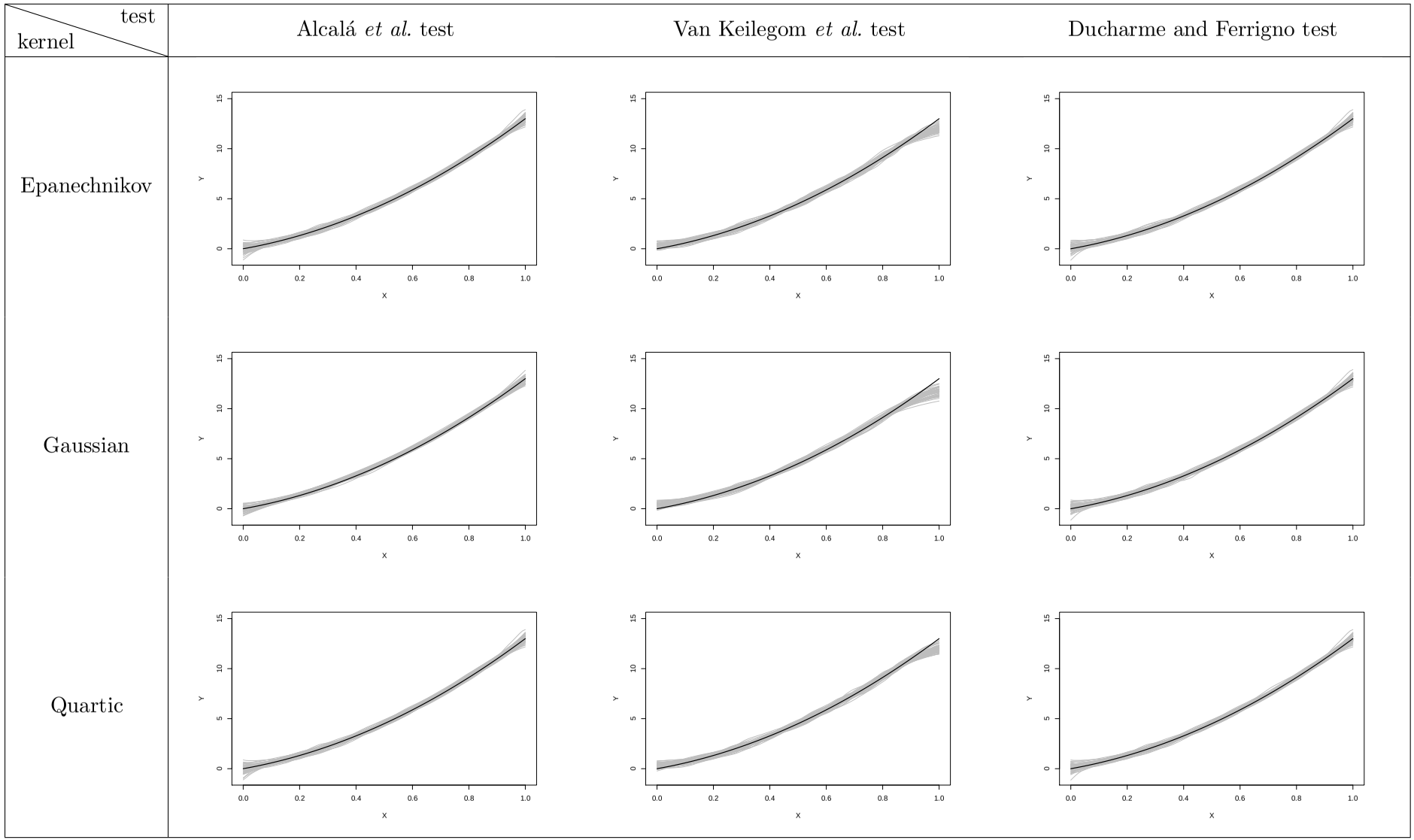

Figure 9. 50 estimates of the link function under the alternative hypothesis $a = 8$ from the 3 tests and 3 kernel functions and from samples of size $n = 200$.

Discussion on bandwidth selection from Figs. 5, 6, and 7.

- We observe that, as predicted by the theory, the bandwidth parameter values tend to become smaller as the sample size increases. In addition, the variability seems to decrease as well. (It should be noted that, for the sake of readability of boxplots, the y-scale is not the same in Fig. 5 and in Figs. 6 and 7.)
- We remark that the optimal value of the bandwidth depends on the test and on the kernel function. For instance, for Van Keilegom *et al.* test, the bandwidths as well as their variability are much smaller when using the Gaussian kernel adjustment compared to the 2 other kernel functions.
- From samples of size $n = 50$, the bandwidth and its variability seem to depend on the value of parameter a except for Ducharme and Ferrigno test. The effect fades when working with samples of size $n = 100$ or 200

Discussion on link function estimates from Figs. 8 and 9. The optimal bandwidth procedure makes the practitioner able to get very good estimates of the regression function whether under the null hypothesis or under the alternative hypothesis chosen. It should be noted that the adjustment curves have more irregularity for the Van Keilegom *et al.* test. Indeed, the estimation in Alcalá *et al.* and Ducharme and Ferrigno tests is made from the local linear method, in contrast to the Van Keilegom *et al.* test for which the estimate is of Nadaraya-Watson type.

5.2. Heteroscedastic polynomial model

The model under question in this 2nd experiment is given by

$$Y = aX^2 + 5X + (1 + 3X)\varepsilon,$$

where X is uniformly distributed on the interval $[0, 1]$, $\varepsilon \sim N(0, 1)$, and ε and X are independent. The link function is the same as the one of Subsection 5.1 but the model is heteroscedastic with a standard deviation going from 1 (when $X = 0$) to 4 (when $X = 1$). As above, the null hypothesis corresponds to $a = 5$, and the value of a will be used to quantify the deviation from the null hypothesis (between 1 and 10).

For this empirical study, we focus on the comparison of the empirical power functions of the 3 tests with the Epanechnikov kernel and the wild bootstrap procedure proposed by Mammen with 50 replicates. The experiments have been replicated 50 times from samples of size $n = 100$ and $n = 200$. The numerical results are presented in Fig. 10.

Ducharme and Ferrigno global test competes with the 2 other directional tests, in particular with Van Keilegom *et al.* test from 200 data. Under the alternatives, the probabilities of rejecting the null hypothesis obtained from Alcalá *et al.* directional test are slightly better than those obtained by Ducharme and Ferrigno and Van Keilegom *et al.* tests. However, Ducharme and Ferrigno test is better than the 2 other methods in terms of risk under H_0.

It should be noted that the empirical power functions are, for the 3 tests less accurate from the heteroscedastic model than from the homoscedastic model of the previous section, in particular in the area of alternatives close to H_0, i.e., $a = 5 \pm 3$. In addition, we observe a slight inversion since the best test was Van Keilegom *et al.* with a suitable choice of parameters.

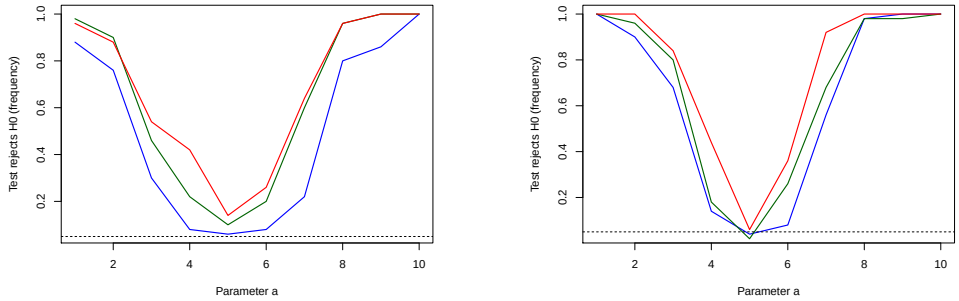

Figure 10. Empirical power functions of the 3 tests under consideration (Alcalá *et al.* test in red, Van Keilegom *et al.* test in green, and Ducharme and Ferrigno test in blue solid line) computed from 50 simulated samples of size 100 (left panel) and 200 (right panel). Epanechnikov kernel and Mammen wild bootstrap resampling have been applied.

5.3. Heteroscedastic polynomial model with undisclosed variance

The model considered in this 3rd numerical experiment is given by

$$Y = a + \sin(X) + \begin{cases} 0.2\,\varepsilon & \text{if} \quad X < 5, \\ 2\,\varepsilon & \text{if} \quad X \geq 5, \end{cases}$$

where X is uniformly distributed on the interval $[0, 10]$, $\varepsilon \sim N(0, 1)$, and ε and X are independent. This model is only piecewise-homoscedastic. The null hypothesis corresponds to a null value of the intercept, i.e., $a = 0$. We aim to compare the 3 tests under consideration in this paper but without assuming that the variance function is disclosed to the practitioner. As a consequence, it is required to estimate it in a nonparametric way before applying the Ducharme and Ferrigno test. We deliberately choose a difficult case with a discontinuous variance function.

We estimate the variance as a smooth function of X through the estimated residuals using the Nadaraya-Watson estimator (7) already presented in Subsection 3.2. This is nevertheless relevant if the nonsmooth property of the variance is not known by the user.

We use the Epanechnikov kernel and the wild bootstrap procedure proposed by Mammen with 50 bootstrap samples, and perform the 3 tests as well as the Ducharme and Ferrigno test with estimated variance function from 100 samples of sizes 100 and 200. The results in terms of empirical power functions are presented in Fig. 11.

Ducharme and Ferrigno test is better than the 2 other methods in terms of empirical risk under H_0, both with disclosed and estimated variance. On the other hand, the best empirical power functions are the ones obtained from Alcalá *et al.* and Van Keilegom *et al.* tests, while the 3rd test achieves less satisfying results around H_0, in particular as expected when the variance is estimated. Even if the global test can be applied when the variance function is unknown by estimating it in a nonparametric way, this illustrates that the results are significantly degraded. Taking into account that, in that case, the practitioner does not want to test the form of the variance, the 2 directional tests should be preferred.

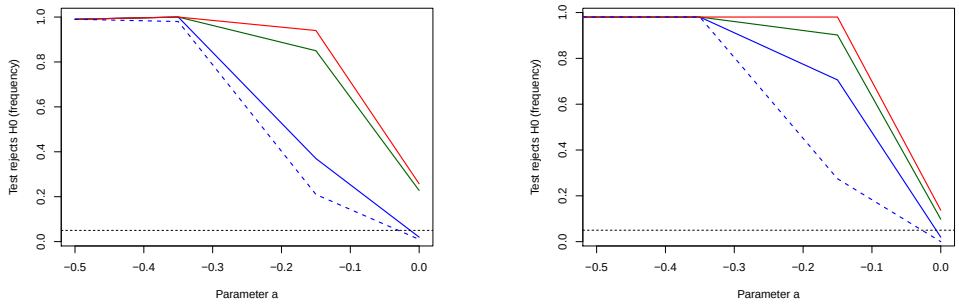

Figure 11. Empirical power functions of the 3 tests under consideration (Alcalá *et al.* test in red, Van Keilegom *et al.* test in green, Ducharme and Ferrigno test in blue solid line, and Ducharme and Ferrigno test with estimated variance function in blue dotted line) computed from 100 simulated samples of size 100 (left panel) and 200 (right panel). Epanechnikov kernel and Mammen wild bootstrap resampling have been applied.

6. Concluding remarks

In this paper, we have focused on the task of choosing the structural part of the regression function using 3 goodness-of-fit tests based on Cramér-von Mises statistic: Alcalá *et al.*, Van Keilegom *et al.*, and Ducharme and Ferrigno tests. The latter is global, meaning that it can assess whether a model fits a dataset on all its assumptions, while the 2 other tests can only verify the functional form of the link function.

To perform these 3 tests, we have developed the `R` package `cvmgof`. A simulation study has been carried out to compare the behaviour of the 3 procedures, from 3 different models, under the same conditions including the sample size, the kernel adjustment, the wild bootstrap procedure, and the selection method of the bandwidth. We highlighted that not all the combinations of parameters are adapted to the 3 tests. In other words, the parameters must be selected carefully, depending on the test chosen and on the data, and `cvmgof` can help the practitioner to do this. In addition, the global test is able to compete with 2 directional tests, especially to detect the null hypothesis. Under the alternatives, and with a suitable set of parameters, Van Keilegom *et al.* test is better than the 2 others from data generated under the homoscedastic model, while Alcalá *et al.* test provides the best results under the heteroscedastic model.

To complete this work, it would be interesting to assess the other assumptions of a regression model such as the functional form of the variance or the additivity of the random error term. It should be noted that this can already be done using Ducharme and Ferrigno test implemented in `cvmgof` since it is a global test. However, it would be relevant to compare the results obtained from Ducharme and Ferrigno test with the ones obtained from other directional tests, e.g., [16,20], especially developed to assess one of these specific assumptions.

The implementation of these directional tests would enrich `cvmgof` package and offer a complete easy-to-use tool for validating regression models. Moreover, the assessment of the overall validity of the model when using several directional tests could be compared with that done when using only a global test. In particular, the well-known problem of multiple testing could be discussed by comparing the results obtained from multiple test procedures with those obtained when using a global test strategy.

Finally, another perspective of this work would be to develop a similar tool for other statistical models widely used in practice such as generalized linear models [23,24].

References

[1] M G Akritas and I Van Keilegom. Nonparametric estimation of the residual distribution. Scand J Stat, 28:549–568, 2001.

[2] J T Alcalá, J A Cristóbal, and W González Manteiga. Goodness-of-fit test for linear models based on local polynomials. Statistics & Probability Letters, 42(1):39–46, 1999.

[3] T W Anderson and D A Darling. A test of goodness of fit. Journal of the American Statistical Association, 49(268):765–769, 1954.

[4] D W K Andrews. A conditional Kolmogorov test. Econometrika, 65:1097–1128, 1997.

[5] H Cramér. On the composition of elementary errors. Skandinavisk Aktuarietidskrift, 11:13–74, 1928.

[6] R B D'Agostino and M A Stephens. Goodness-of-fit techniques. Marcel Dekker, Inc, New York, 1999.

[7] H Davidson and T Flachaire. The wild bootstrap, tamed at last. Journal of Econometrics, 146:162–169, 2008.

[8] R Davidson and E Flachaire. The wild bootstrap, tamed at last. Journal of Econometrics, 146:162–169, 2008.

[9] H Dette. A consistent test for heteroscedasticity in nonparametric regression based on kernel method. Journal of Statistical Planning and Inference, 103:311–329, 2002.

[10] G R Ducharme and S Ferrigno. An omnibus test of goodness-of-fit for conditional distributions with applications to regression models. Journal of Statistical Planning and Inference, 142:2748–2761, 2012.

[11] J Fan and I Gijbels. Local polynomial modelling and its applications. Chapman & Hall, London, 1996.

[12] J Faraway, G Marsaglia, J Marsaglia, and A Baddeley. goftest: Classical goodness-of-fit tests for univariate distributions, 2017. R package version 1.1-1.

[13] S Ferrigno and G R Ducharme. Un choix de fenêtre optimal en estimation polynomiale locale de la fonction de répartition conditionnelle. C. R. Acad. Sci. Paris, Ser.I, 346:83–86, 2008.

[14] E González-Estrada and J N Villaseñor. An R package for testing goodness of fit: goft. Journal of Statistical Computation and Simulation, 88(4):726–751, 2018.

[15] W Hardle and E Mammen. Comparing nonparametric versus parametric regression fits. The Annals of Statistics, 21(4):1926–1947, 1993.

[16] C Heuchenne and I Van Keilegom. Goodness-of-fit tests for the error distribution in nonparametric regression. Computational Statistics & Data Analysis, 54(8):1942–1951, 2010.

[17] I Kojadinovic and J Yan. fgof: Fast Goodness-of-fit Test, 2012. R package version 0.2-1.

[18] J Kolácek. Plug-in method for nonparametric regression. Computational Statistics, 23(Issue 1):63–78, 2008.

[19] Q Li and J Racine. Cross-validated local linear nonparametric regression. Statistica Sinica, 14:485–512, 2004.

[20] H Liero. Testing homoscedasticity in nonparametric regression. Journal of Nonparametric Statistics, 15(1):31–51, 2003.

[21] R.Y Liu. Bootstrap procedure under some non-i.i.d. models. Annals of Statistics, 16:1696–1708, 1988.

[22] E Mammen. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist., 21(1):255–285, 1993.

[23] P McCullagh and JA Nelder. Generalized Linear Models, Second Edition. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[24] MC Rodriguez-Campos, W Gonzalez-Manteiga, and R Cao. Testing the hypothesis of a generalized linear regression model using nonparametric regression estimation. Journal of Statistical Planning and Inference, 67(1):99 – 122, 1998.

[25] W Stute, W Gonzàlez Manteiga, and M Presedo Quindimil. NN goodness of fit test for

linear models. Journal of Statistical Planning and Inference, 53:75–92, 1996.

[26] I Van Keilegom, W González Manteiga, and C Sánchez Sellero. Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. Test, 17:401–415, 2008.

[27] R von Mises. On the foundations of probability and statistics. The Annals of Mathematical Statistics, 12:195–205, 1941.

[28] J X Zheng. A consistent test of conditional parametric distributions. Econometric Theory, 16:667–669, 2000.