Data Driven Score Tests for a Homoscedastic Linear Regression Model: the Construction and Simulations

Tadeusz Inglot, Teresa Ledwina

Abstract: We propose new tests for testing the validity of a semiparametric random-design linear regression model. The construction consists of several steps. First, we follow the classical idea of overfitting and replace the basic problem by a series of auxiliary subproblems. Next, to test whether extra terms are significant we construct a counterpart of classic score statistic. Finally, we combine the solution with smoothing methods providing guidelines to choose the right subproblem. This leads to data driven score tests for the initial testing problem. Under the null model our construction is asymptotically distribution free, as shown in Inglot and Ledwina [19], [20]. We illustrate the result by a small simulation study. We also compare the finite sample performace of our tests with the recent solution introduced by Guerre and Lavergne [9], as well as to Cramér-von Mises type construction. The simulation experiment indicates the very good performance of the proposed tests.

MSC 2000: 62G10, 62J05
Key words: Cramér-von Mises test, efficient score, hypothesis testing, data driven test, linear regression, selection rule, semiparametric inference, smoothing methods

1 Introduction

The problem of verifying the linear structure of a regression function is central in applied statistics. Therefore, it is not surprising that there is an extensive literature on several possible solutions under a variety of different restrictions. For the set-up which is considered in this paper, current literature is dominated by two, not very much different, ideas in approaching the problem. The first one exploits old Cramér-von Mises, Kolmogorov-Smirnov and other similar solutions. For some evidence see Stute [31], Stute et al. [32] and [33], Diebolt and Zuber [6], Koenker and Xiao [25], Khmaladze and Koul [24] e.g. The second one relies on comparing some parametric and nonparametric fits, cf. Kozek [26], Härdle and Mammen [10], Horovitz and Spokoiny [12], Zhang and Dette [34], Guerre and Lavergne [9], to mention few of them. Some different ideas are exploited in Dette nad Munk [4] and Dette [5]. For some more specialized situations, mostly focused on the fixed

The second named author is deeply indebted to A. Schick for insightful discussion in Oberwolfach, September 2003. The discussion resulted, in particular, in sharpening the previous version of present Theorem 2.2. The authors are grateful to J. Mielniczuk for comments. The programming work was done by A. Janic under support from the KBN 5 P03A 03020 grant. Her kind co-operation is gratefully acknowledged. The computations were partly done at the Institute of Mathematics and Informatics of Wrocław University of Technology. We especially thank M. Kaczmarz, W. Rutkowski and M. Wyłomański for help and advice in this regard. This research was partially supported by grant 5 P03A 03020 from KBN.

design and Gaussian errors, some new ideas were introduced recently. For some evidence see Baraud et al. [1] and Fan and Huang [7]. For further references, again mostly concerning the fixed design situation, see Hart [11]. For some comparison of few of these solutions with adequate data driven score test see Inglot and Ledwina [18],[21].

The purpose of this paper is to propose and investigate some new tests of fit for the problem described below. The solutions shall be derived via elaborating on some ideas sketched in Choi et al. [3] and matching the result with the approach proposed in Ledwina [27] and extended in Inglot, Kallenberg and Ledwina [13], among others.

Let $Z = (X, Y)$ denote a random vector in $I \times R,\ I = [0, 1]$. We would like to verify the null hypothesis H_0 asserting

$$Y = \beta[v(X)]^T + \epsilon, \tag{1.1}$$

where X and ϵ are independent, $E\epsilon = 0,\ E\epsilon^2 < \infty,\ \ \beta \in R^q$ is a vector of unknown real valued parameters while $v(x) = (v_1(x), ..., v_q(x))$ is a vector of known functions. The symbol ^{T} denotes transposition. All vectors are considered to be row vectors.

We start with the classical idea of overfitting and reducing the verification of (1.1) to testing whether extra terms are significant. More precisely, given a fixed k, we embed our null model (1.1) into the following auxiliary model

$$\texttt{M(k)} \qquad Y = \theta[u(X)]^T + \beta[v(X)]^T + \epsilon, \tag{1.2}$$

which satisfies the following assumptions

$<$ **M1** $>$

$u(x) = (u_1(x), ..., u_k(x)),\ v(x) = (v_1(x), ..., v_q(x)),\ x \in I$, and the measurable functions $u_1, ..., u_k, v_1, ..., v_q$ are bounded and linearly independent; $\theta \in R^k,\ \beta \in R^q$ are unknown parameters;
X has an unknown density g with respect to the Lebesgue measure λ supported on I;
ϵ has an unknown density f with respect to the Lebesgue measure λ on R. The density f satisfies $E_f\epsilon = 0,\ \tau = E_f\epsilon^2$ and $0 < \tau < \infty$;
X and ϵ are independent.

At the first step we construct appropriate score test statistic [precisely: efficient score statistic], for the given fixed k, for testing $H_0(k) : \theta = 0$ against $\theta \neq 0$ in `M(k)` satisfying $< M1 >$ and some further regularity conditions $< M2 >$ and $< M3 >$. An efficient score vector along with its appropriate estimator play the central role in this construction. Section 2.1 presents the efficient score vector while Section 2.2 contains some general class of estimators of this entity. Theorem 2.2, the basic result of Section 2.3, gives conditions under which the influence of nuisance parameters β, f, g on the limiting behaviour of the related score statistic is asymptotically negligible. Section 2.4 contains the next step of our construction, i.e. incorporating into the score statistic a score-based selection rule for determining the dimension k.

We discuss two selection rules and call the resulting statistics the data driven score tests. Section 3 presents the results of our simulation study. The simulation results show that these data driven constructions possess two fundamental advantages of efficient score statistics. Namely, for moderate sample sizes the critical values are stable for a variety of nuisance parameters, while empirical powers are high, considerably exceeding those of the best existing solutions of the problem.

2 Data driven score tests

Before we introduce the test statistics, we present a series of auxiliary constructions and results.

2.1 Efficient score vector for testing $\theta = 0$ in $\mathtt{M(k)}$

A general result for score vectors in some large class of regression models is given in Schick [30]. For completeness, in Inglot and Ledwina [15] some existing results on score vectors in the model $\mathtt{M(k)}$ were reproved and results on efficient score vectors for testing (1.2) were derived. This paper also concerns more general heteroscedastic case. Below we quote some of these results.

In the case under consideration, in addition to the basic model assumptions $< M1 >$ we need the following ones

$< \mathbf{M2} >$ $f'(y)$ exists for all $y \in R$ and $J = J(f) = \int_R [f'(y)]^2/f(y)\lambda(dy) < \infty$,

$< \mathbf{M3} >$ $g > 0 \quad \lambda$ - a.e.

Under these three assumptions the efficient score vector for testing $H_0(k) : \theta = 0$ in $\mathtt{M(k)}$, calculated at $z = (x, y)$, is of the form

$$\ell^*(z) = -\{[f'/f](\varepsilon)\}[\tilde{u}(x) - \tilde{v}(x)\mathbf{V}^{-1}\mathbf{M}] + \tau^{-1}\varepsilon[m_1 - m_2\mathbf{V}^{-1}\mathbf{M}],$$

where

$$\varepsilon = y - v(x)\beta^T,\ m_1 = E_g u(X),\ m_2 = E_g v(X),\ m = (m_1, m_2),$$

$$\tilde{w}(x) = (\tilde{u}(x), \tilde{v}(x)),\ \tilde{u}(x) = u(x) - m_1,\ \tilde{v}(x) = v(x) - m_2,$$

while $\mathbf{M}$ and $\mathbf{V}$ are blocks in

$$\mathbf{W} = \begin{pmatrix} \mathbf{U} & \mathbf{M}^T \\ \mathbf{M} & \mathbf{V} \end{pmatrix} = \frac{1}{4}\left\{ J E_g[\tilde{w}(X)]^T[\tilde{w}(X)] + \frac{1}{\tau} m^T m \right\}.$$

Note that, due to $< M3 >$, $\mathbf{W}$ is positive definite [cf. Remark C.13 in Inglot and Ledwina [15]].

2.2 Efficient score statistic and a general result

We introduce the additional notation

$$\vartheta = (\sqrt{g}, \sqrt{f}), \quad \eta = (\beta, \vartheta) \text{ and } \ell^*(z;\eta) = \ell^*(z).$$

Moreover, P_η^n denotes the joint distribution of $Z_1, ..., Z_n$ under the null model (1.1).

Finally set $\mathbf{W}^{11} = (\mathbf{U} - \mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}$, $\mathbf{L} = 4^{-1}\mathbf{W}^{11}$ and define

$$W_k(\eta) = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \ell^*(Z_i;\eta)\right] \mathbf{L} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \ell^*(Z_i;\eta)\right]^T.$$

From $< M1 > - < M3 >$, Corollaries C.16, C.18 and Remark C.13 of Inglot and Ledwina [15], e.g., under the null hypothesis $H_0(k)$, $\mathbf{L}$ is positive definite and it holds that

$$E_\eta \ell^*(Z;\eta) = 0, \quad \left\{E_\eta[\ell^*(Z;\eta)]^T[\ell^*(Z;\eta)]\right\}^{-1} = \mathbf{L}, \quad W_k(\eta) \xrightarrow{\mathcal{D}} \chi_k^2, \tag{2.1}$$

where χ_k^2 denotes a random variable from the central chi-square distribution with k degrees of freedom.

Define

$$W_k(\hat{\eta}) = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}^*(Z_i;\hat{\eta})\right] \hat{\mathbf{L}} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}^*(Z_i;\hat{\eta})\right]^T, \tag{2.2}$$

where $\hat{\ell}^*(\bullet;\hat{\eta})$ is an estimator of $\ell^*(\bullet;\eta)$, while $\hat{\mathbf{L}}$ is an estimator of $\mathbf{L}$.

Finally, let $||\bullet||$ denote the Euclidean norm of a given vector. The relation (2.1) and a simple argument yield the following result.

Proposition 2.1. *Assume the null hypothesis $H_0(k) : \theta = 0$ is true and the assumptions $< M1 >, < M2 >$ and $< M3 >$ are fulfilled. Suppose that $\hat{\mathbf{L}}$ is a consistent estimator of $\mathbf{L}$ and the estimator $\hat{\ell}^*(\bullet;\hat{\eta})$ satisfies the following condition*

$$P_\eta^n\left(\left|\left|\sum_{i=1}^n [\hat{\ell}^*(Z_i;\hat{\eta}) - \ell^*(Z_i;\eta)]\right|\right| \geq \delta\sqrt{n}\right) \to 0 \text{ as } n \to \infty \text{ for every } \delta > 0. \tag{2.3}$$

Then for the test statistic $W_k(\hat{\eta})$ defined in (2.2) it holds that

$$W_k(\hat{\eta}) \xrightarrow{\mathcal{D}} \chi_k^2, \quad \text{as } n \to \infty.$$

$W_k(\hat{\eta})$ is an efficient score statistic for testing $H_0(k)$ in `M(k)`. As said before, we shall abbreviate this name to score statistic. Choi et al. [3] used the name efficient test statistic for such a construction.

2.3 Some class of estimators $\hat{\ell}^*$ of ℓ^* satisfying (2.3)

We follow some well established ideas. On one hand, our construction is obviously linked to the approach of Bickel [2], Example 3. On the other, it incorporates the very useful contribution of Schick [29] showing that using only a small fraction of the sample to estimate the score function, as in Bickel [2] , can be avoided.

Suppose $Z_1, ..., Z_n$ are i.i.d. vectors obeying (1.2). Note that, as usual in the score test theory, all considerations below are done under the assumption $\theta = 0$.

Take $\zeta = \left[\frac{n}{2}\right]$ and divide $Z_1, ..., Z_n$ into two parts $Z_1, ..., Z_\zeta$ and $Z_{\zeta+1}, ..., Z_n$. In order to clearly show an important feature of our construction, we shall, for a moment, display in formulas the expectation m as if it were the next nuisance parameter. Additionally set $<1> = \{1, ..., \zeta\}$, $<2> = \{\zeta+1, ..., n\}$. The superscript (j), $j = 1, 2$, appearing below, indicates from which part of the sample we estimate the related quantity.

The basic structure of $\hat{\ell}^*$ at the observed points $Z_1, ..., Z_n$ is as follows

$$\hat{\ell}^*(Z_i; \hat{\eta}) = \begin{cases} \ell^*(Z_i; \hat{\beta}_*^{(2)}, \hat{g}^{(2)}, \hat{f}^{(2)}, \hat{m}^{(1)}), & \text{if } i \in <1>, \\ \ell^*(Z_i; \hat{\beta}_*^{(1)}, \hat{g}^{(1)}, \hat{f}^{(1)}, \hat{m}^{(2)}), & \text{if } i \in <2>, \end{cases}$$

where

$$\hat{m}_1^{(1)} = \frac{1}{\zeta} \sum_{i \in <1>} u(X_i), \;\; \hat{m}_1^{(2)} = \frac{1}{n-\zeta} \sum_{i \in <2>} u(X_i), \;\; \tilde{u}^{(j)}(\bullet) = u(\bullet) - \hat{m}_1^{(j)}, \;\; j = 1, 2,$$

$\hat{m}_2^{(1)}, \hat{m}_2^{(2)}$ and $\tilde{v}^{(j)}(\bullet)$, $j = 1, 2$, are defined analogously, while $\hat{\beta}_*^{(j)}$ is a discretized version of a $\sqrt{n}$ - consistent estimator $\hat{\beta}^{(j)}$ of β, based on the jth part of the sample.

The specific form of $\hat{m}^{(j)}$, together with the fact that in the construction of $\hat{\ell}^*$ only the estimators $\hat{m}^{(j)}$ are matched to Z_i with i from $<j>$ and the requirements for $\sqrt{n}$ - consistency of an estimator for β are the strongest requirements on estimators we imposed in the construction. When estimating other quantities there is a lot of freedom, as seen from Theorem 2.2, below.

To write the form of the estimators $\hat{\ell}^*(Z_i; \hat{\eta})$, $i \in <j>$, $j = 1, 2$, explicitly denote by $\hat{\mathbf{V}}^{(j)}, \hat{\mathbf{M}}^{(j)}$, $\hat{\tau}^{(j)}, \hat{\tau}^{(j)} > 0 -$ a.e. and $\widehat{[f'/f]}^{(j)}$ the related estimators of the appropriate quantities. Note that having these estimators, we do not need to estimate the density g itself. Set $\hat{\varepsilon}_i^{(j)} = Y_i - v(X_i)[\hat{\beta}_*^{(j)}]^T$. For $i \in <1>$ we have

$$\hat{\ell}(Z_i; \hat{\eta}) = -\widehat{[f'/f]}^{(2)} \left(\hat{\varepsilon}_i^{(2)}\right) \left[\tilde{u}^{(1)}(X_i) - \tilde{v}^{(1)}(X_i)[\hat{\mathbf{V}}^{(2)}]^{-1}\hat{\mathbf{M}}^{(2)}\right]$$
$$+ \frac{1}{\hat{\tau}^{(2)}} \left[\hat{\varepsilon}_i^{(2)}\right] \left[\hat{m}_1^{(1)} - \hat{m}_2^{(1)}[\hat{\mathbf{V}}^{(2)}]^{-1}\hat{\mathbf{M}}^{(2)}\right],$$

and for $i \in <2>$ the definition is analogous.

Theorem 2.2. *Suppose that under the null distribution P_η^n for $j = 1, 2$ the following hold : $\hat{\beta}^{(j)}$ are $\sqrt{n}$ - consistent estimators of β, while $\hat{\tau}^{(j)}, \hat{\mathbf{V}}^{(j)}$ and $\hat{\mathbf{M}}^{(j)}$*

are consistent estimators of $\tau, \mathbf{V}$ *and* $\mathbf{M}$*, respectively. Moreover, assume that the estimators* $\widehat{[f'/f]}^{(j)}$, $j = 1, 2$, *of* f'/f *are consistent in the* L_2 *norm, i.e. for every* $\delta > 0$

$$P_\eta^n \left(\int_R \left(\widehat{[f'/f]}^{(j)}(y) - [f'/f](y) \right)^2 f(y)\lambda(dy) > \delta \right) \to 0 \quad as \quad n \to \infty.$$

Then the estimator $\hat{\ell}^*$ *of* ℓ *defined above satisfies the condition (2.3) of Proposition 2.1.*

A proof of Theorem 2.2 is given in Inglot and Ledwina [19],[20].

2.4 Determining k in $W_k(\hat{\eta})$ by some score-based selection rules

We now consider a nested family of auxiliary models $\mathtt{M(k)}$, $k = 1, ..., d$, where d is fixed but otherwise arbitrary. Following the construction proposed in Ledwina [27], as e.g. in Kallenberg and Ledwina [22] we define score-based selection rule $S1$ as follows

$$S1 = \min\{1 \le k \le d : W_k(\hat{\eta}) - k \log n \ge W_s(\hat{\eta}) - s \log n, \;\; s = 1, ..., d\}.$$

The rule $S1$ mimics the Schwarz BIC criterion. Since the penalty $s \log n$ is relatively heavy, $S1$ is well suited to detect low dimensional models $\mathtt{M(k)}$. In contrast, the rule

$$A1 = \min\{1 \le k \le d : W_k(\hat{\eta}) - 2k \ge W_s(\hat{\eta}) - 2s, \;\; s = 1, ..., d\},$$

imitating the Akaike AIC criterion, is expected to work well when high dimensional disturbances $\mathtt{M(k)}$ of the null model $\mathtt{M(0)}$: $Y = \beta[v(X)]^T + \epsilon$ are present. Based on our experience and some previous research, the following "intermediate" solution was proposed and discussed in Inglot and Ledwina [17]: use $A1$ when the distribution of the data at hand is very distinct from the null model and $S1$ otherwise. To provide a threshold defining which rule should be applied, we propose to consider the magnitude of the estimated standardized components of the efficient score vector. More precisely, in the present set-up, under the assumptions and notation of Proposition 2.1, set $(\mathcal{Y}_1, ...\mathcal{Y}_k) = \left[n^{-1/2} \sum_{i=1}^n \hat{\ell}^*(Z_i; \hat{\eta}) \right] \hat{\mathbf{L}}^{1/2}$. Then, obviously, $W_k(\hat{\eta}) = ||(\mathcal{Y}_1, ..., \mathcal{Y}_k)||^2$. Following the discussion presented in Inglot and Ledwina [17], we propose to use the following penalty in this problem

$$\pi(s, n, p) = \begin{cases} s \log n, & \text{if } \max_{1 \le t \le d} |\mathcal{Y}_t| \le \sqrt{p \log n}, \\ 2s, & \text{if } \max_{1 \le t \le d} |\mathcal{Y}_t| > \sqrt{p \log n}, \end{cases} \tag{2.5}$$

where p is some fixed positive number. This strategy leads to the following refined selection rule

$$T1 = \min\{1 \le k \le d : W_k(\hat{\eta}) - \pi(k, n, p) \ge W_s(\hat{\eta}) - \pi(s, n, p), \;\; s = 1, ..., d\}.$$

It is evident that small p's result in $T1$ being in practice equivalent to $A1$, while large p's lead to $T1$ being very similar to $S1$. "Moderate" values of p give a meaningful "switching effect".

For $n \geq 8$, $S1 \leq T1 \leq A1$. Moreover, since under the null model $(\mathcal{Y}_1, ..., \mathcal{Y}_k) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}_k)$, then $P_\eta^n(T1 \neq S1) \to 0$ as $n \to \infty$. On the other hand, under H_0, for any $s \in \{2, ..., d\}$, $P_\eta^n(S1 = s) \leq P_\eta^n\left(W_s(\hat{\eta}) \geq (s-1)\log n\right)$. Hence, Proposition 2.1 yields

Proposition 2.3. *Under the null hypothesis $H_0 : Y = \beta[v(X)]^T + \epsilon$, the assumptions of Proposition 2.1 and $n \to \infty$, it holds that*

$$P_\eta^n(S1 > 1) \to 0, \quad W_{S1}(\hat{\eta}) \xrightarrow{\mathcal{D}} \chi_1^2, \quad \textit{and} \quad P_\eta^n(T1 > 1) \to 0, \quad W_{T1}(\hat{\eta}) \xrightarrow{\mathcal{D}} \chi_1^2.$$

3 Simulation study

Practical implementation of W_{S1} and W_{T1} requires some specification of the estimators appearing in (2.2) and (2.4). So, we shall first discuss this point.

3.1 Specification of estimators

We define W_{S1} and W_{T1} in the following way. The sample splitting scheme and estimators $m_i^{(j)}$, $i, j = 1, 2$, were applied according to the description in Section 2.3. The remaining parameters were estimated on the basis of the jth part of the sample, $j = 1, 2$, as follows. The components of $\hat{\beta}^{(j)}$ were ordinary least square estimators. The discretization was neglected in the simulations. $\hat{\tau}^{(j)}$ was the adjusted Rice [28] estimator. We estimated f'/f by $[\widetilde{f}^{(j)}]'/\widetilde{f}^{(j)}$, where $\widetilde{f}^{(j)}$ is the kernel estimator of f defined as follows

$$\widetilde{f}^{(1)}(y) = \gamma_\zeta + \frac{1}{\zeta\hat{\alpha}_\zeta^{(1)}} \sum_{i \in <1>} K\left(\frac{y - \hat{\varepsilon}_i^{(1)}}{\hat{\alpha}_\zeta^{(1)}}\right), \; \widetilde{f}^{(2)}(y) = \gamma_\zeta + \frac{1}{\zeta\hat{\alpha}_\zeta^{(2)}} \sum_{i \in <2>} K\left(\frac{y - \hat{\varepsilon}_i^{(2)}}{\hat{\alpha}_\zeta^{(2)}}\right),$$

where K is the standard Gaussian kernel, while $\gamma_\zeta = 0.0001$, $\hat{\alpha}_\zeta^{(j)} = (0.9)[\hat{\tau}^{(j)}]^{1/2}\zeta^{-1/7}$, $\hat{\varepsilon}_i^{(j)} = Y_i - v(X_i)[\hat{\beta}^{(j)}]^T$. To have some flexibility, we used a simple random bandwidth.

We estimated J in the first part of the sample by $\hat{J}^{(1)} = \zeta^{-1}\sum_{c \in <1>}\left\{[\widetilde{f}'^{(2)}/\widetilde{f}^{(2)}]\left(\hat{\varepsilon}_c^{(2)}\right)\right\}^2$ and in the second part of the sample by the analogous expression. We estimated $\mathbf{W}$ by $\hat{\mathbf{W}} = (\hat{\mathbf{W}}^{(1)} + \hat{\mathbf{W}}^{(2)})/2$, where $\hat{\mathbf{W}}^{(1)} = \frac{1}{4}\left\{\hat{J}^{(1)}\frac{1}{\zeta}\sum_{i \in <1>}[\tilde{w}^{(1)}(X_i)]^T[\tilde{w}^{(1)}(Xi)] + \frac{1}{\hat{\tau}^{(1)}}[\hat{m}^{(1)}]^T\hat{m}^{(1)}\right\}$ and $\hat{\mathbf{W}}^{(2)}$ is defined analogously. Finally, we considered the natural estimator $\hat{\mathbf{L}}$ of $\mathbf{L}$ given by $\hat{\mathbf{L}} = \frac{1}{4}(\hat{\mathbf{U}} - \hat{\mathbf{M}}^T\hat{\mathbf{V}}^{-1}\hat{\mathbf{M}})^{-1}$, where $\hat{\mathbf{M}}$, $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are blocks of $\hat{\mathbf{W}}$. Following our earlier considerations [cf. the discussion in Inglot and Ledwina [17]], we took $p = 2.4$ in (2.5). The choice of u_i's and d is given in Section 3.2.

Detailed proofs verifying the assumptions of Theorem 2.2 for this application are contained in Inglot and Ledwina [19].

3.2 Models used in the simulations

The scheme of our study matches those used in the papers by Baraud et al. [1], Diebolt and Zuber [6], Guerre and Lavergne [8],[9], as well as Horowitz and Spokoiny [12]. We considered the problem of testing $H_0: \; Y = 1 + 2X + \epsilon$. To construct W_{S1} and W_{T1} we considered $d = 10$ auxiliary models $\texttt{M(k)}$, which pertain to $u_i(x) = \cos([i+1]\pi x),\; i = 1, ..., 10$. We decided to consider the cosine system in view of the competitors we shall investigate. The statistic of Guerre and Lavergne is based on equispaced partitions, while the Cramér-von Mises test is tightly linked to cosine functions. Therefore, such a choice gives conditions for fair comparison.

We consider ϵ obeying one of three laws: Gaussian with 0 mean and standard deviation σ [$G(\sigma)$ for short], Laplace with 0 mean and standard deviation $\sqrt{2}/\varphi$ [$L(\varphi)$ for short] and normal mixture $(0.7)\phi(x - \mu/(0.7)) + (0.3)\phi(x + \mu/(0.3))$ [denoted $NM(\mu)$ in what follows], where ϕ is the $N(0,1)$ density function.

X was assumed to be independent of ϵ and obeying a beta distribution on [0,1]. Since changing, to some reasonable extent, the parameters of the beta distribution had no essential influence on the general picture, we restricted the presentation of results to the case where X is uniformly distributed.

The alternatives were defined by disturbing the pattern $1 + 2x$ [with each type of error: $G(\sigma),\;\; L(\varphi),\;\; NM(\mu)$] by one of the functions $r_l(x),\;\; l = 1, ..., 6$, where

$$r_1(x) = c \times \cos(o\pi x), \;\; c \in R, \;\; o = 2, 3, ...$$

$$r_2(x) = c \times L_s(x),\; c \in R,\; s = 2, 3, ... \;\; L_s - \text{sth normalized Legendre polynomial on } [0,1],$$

$$r_3(x) = c \times \frac{1}{t}\phi\left(\frac{x - 0.5}{t}\right), \;\; c \in R, \;\; t \in R_+, \;\; \phi - \text{the } N(0,1) \text{ density function},$$

$$r_4(x) = c \times (x - a)\mathbf{1}_{[a,1]}(x), \;\; c \in R, \;\; a \in (0,1),$$

$$r_5(x) = c \times \mathrm{arctg}[b(2x - 1)], \;\; c \in R, \;\; b \in (0, \infty),$$

$$r_6(x) = c \times \max\{\min\{(2x - 1)/(1 - 2a), 1\}, -1\}, \;\; c \in R, \;\; a \in (0, 1/2).$$

3.3 Empirical behaviour of test statistics under H_0

All simulation experiments presented in the paper were done for the same sample size $n = 300$ and $N = 10000$ Monte Carlo [MC] runs. Throughout we considered tests at the significance level $\alpha = 0.05$.

Let us start our discussion with some remarks on the behaviour of W_{S1} and W_{T1} under H_0. The asymptotic critical value of W_{S1} and W_{T1} is 3.841. In order to illustrate how the asymptotic theory works in the case of our implementation,

Table 1 presents simulated critical values of W_{S1} and W_{T1} under different error distributions.

Error distribution	**Parameter**	**Variance**	**Critical values**		
			W_{S1}	W_{T1}	**CvM**
$G(\sigma)$	0.25	0.063	5.91	6.11	27.88
	0.50	0.250	5.63	5.92	7.00
	0.75	0.563	5.83	6.04	3.22
	1.00	1.000	5.79	6.02	1.73
$L(\varphi)$	4.00	0.125	5.29	5.57	15.72
	2.00	0.500	5.27	5.50	3.86
	1.00	2.000	5.75	5.93	0.94
	0.50	8.000	5.61	5.82	0.23
$NM(\mu)$	0.20	1.191	5.94	6.08	1.52
	0.40	1.762	5.67	6.00	0.97
	0.60	2.714	5.81	6.05	0.63
	0.80	4.048	5.66	5.85	0.43

Table 1: Simulated critical values of W_{S1}, W_{T1} and CvM under the null model $Y = 1 + 2X + \epsilon$ with X uniform on [0,1] and different errors. Sample size $n = 300$. 5% significance level, $N = 10000$ MC runs.

As seen, an evident feature of the new procedures is that the critical values are very stable to changes of the error distributions and their parameters. Since the penalty in the selection rule $T1$ is slightly smaller, the respective critical values of the test W_{T1} are slightly larger. We would like to emphasize that critical values are also very stable with respect to the choice of d. Any choice of $d \geq 4$ gives practically the same simulated critical value. This follows from the fact that, under the null model and $n = 300$, in all the considered cases, the proportion of cases with $\{S1 = 1\}$ and $\{T1 = 1\}$ is in [0.97, 0.98] and [0.96, 0.97], respectively, and the remaining mass is mostly concentrated on dimensions 2 - 3. On the other hand, the simulated critical values are larger than the limiting values. This is a characteristic phenomenon for data driven tests, which was discussed in detail in some earlier papers. In the present set-up, to provide a practical way of generating critical values, one can apply the residual bootstrap, described e.g. on pp. 142 - 143 of Stute et al. [33]. We implemented this procedure in our simulation study and found that it works well. Anyway, to save simulation time, we present simulated powers for W_{S1} and W_{T1} in the case the averaged, from Table 1, critical values 5.68 and 5.91 are used. Finally, we would like to emphasize that the stability of critical values of the data driven tests with respect to the choice of d allows us to choose practically arbitrary $d \geq 4$. Enlarging d does not spoil empirical powers achieved for choices of smaller d's. Therefore, reasonable choice of d only depends on two factors: how complicated alternatives one likes to detect and how much time consuming calculations are reasonable in this context.

In our implementation of Guerre and Lavergne's [8],[9] solution we took binwidths in $\{2^{-2}, 2^{-3}, ..., 2^{-7}\}$, $c = 1.5$, $J_n = 6$ and used the adjusted Rice estimator for the variance of errors [cf. p. 17 in Guerre and Lavergne [8]]. In our simulation study we followed their prescription and applied the wild bootstrap with the two-point distribution for w_i given on p. 17 of Guerre and Lavergne [8]. We did $B = 400$ bootstrap replications and $N = 10000$ MC runs in each experiment. For simplicity, we shall denote the test statistic introduced by Guerre and Lavergne [8],[9] by $\hat{T}$.

To complete the picture, we also investigated the empirical behaviour of a transformed Cramér-von Mises statistic, which was developed in Stute et al. [32]. We shall denote this statistic by CvM. This transformation was introduced by Khmaladze [23] to remove the influence of nuisance parameters on the null distribution. The simulations reported in Table 1 show that the simulated critical values of CvM are highly unstable. A similar observations were made earlier in Diebolt and Zuber [6] and can be infered from Koenker and Xiao [25].

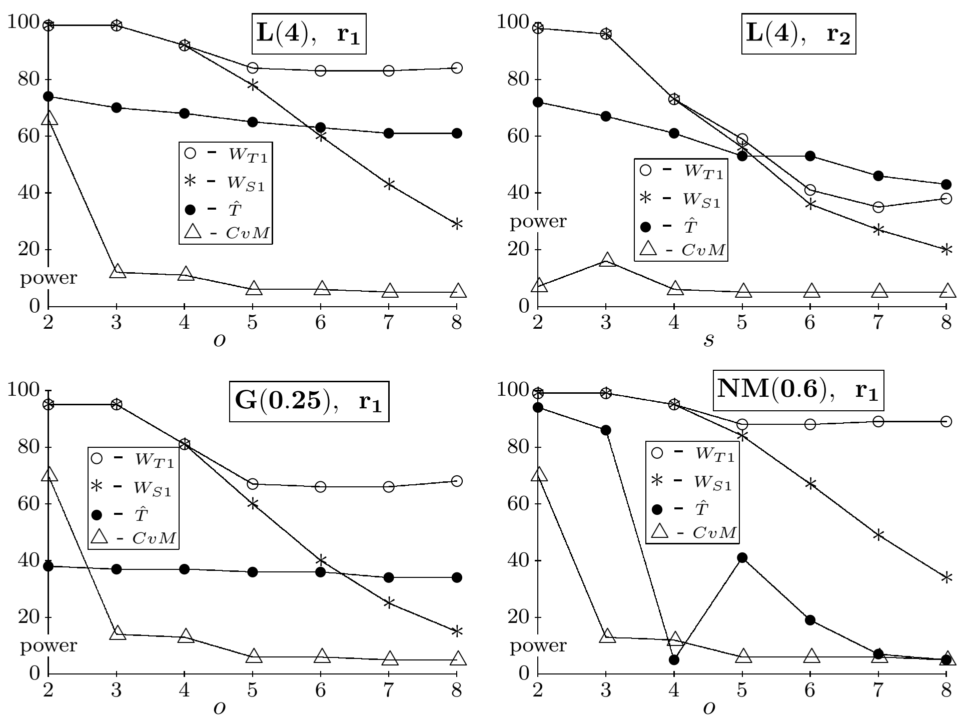

Figure 1: Simulated powers of tests based on W_{T1}, W_{S1}, $\hat{T}$ and CvM under the alternatives $Y = 1 + 2X + r_l(X) + \epsilon$, $l = 1, 2$, X uniform on $[0, 1]$ and different errors. Signal/noise 0.25. 5% nominal level, n=300, N=10000 MC runs

3.4 Empirical powers

In order to simulate powers of W_{S1} and W_{T1}, $\hat{T}$, as well as CvM, we used the averaged, bootstrap and "actual" critical values from Table 1, respectively.

A representative selection of simulation results is presented in Figures 1 and 2. In Figure 1 we show results of experiments which serve to understand the behaviour of test statistics when alternatives are oscillating, i.e. r_1 and r_2, given in Section 3.2, are taken into account. In all four cases considered there, the ratio signal/noise $= ||r_l||_2/\sqrt{\mathrm{Var}\epsilon}$, where $|| \bullet ||_2$ denotes the $L_2[0,1]$ norm, equals 0.25.

Figure 2 exhibits the behaviour of tests under more "smooth" deviations i.e. the disturbances r_3 - r_6.

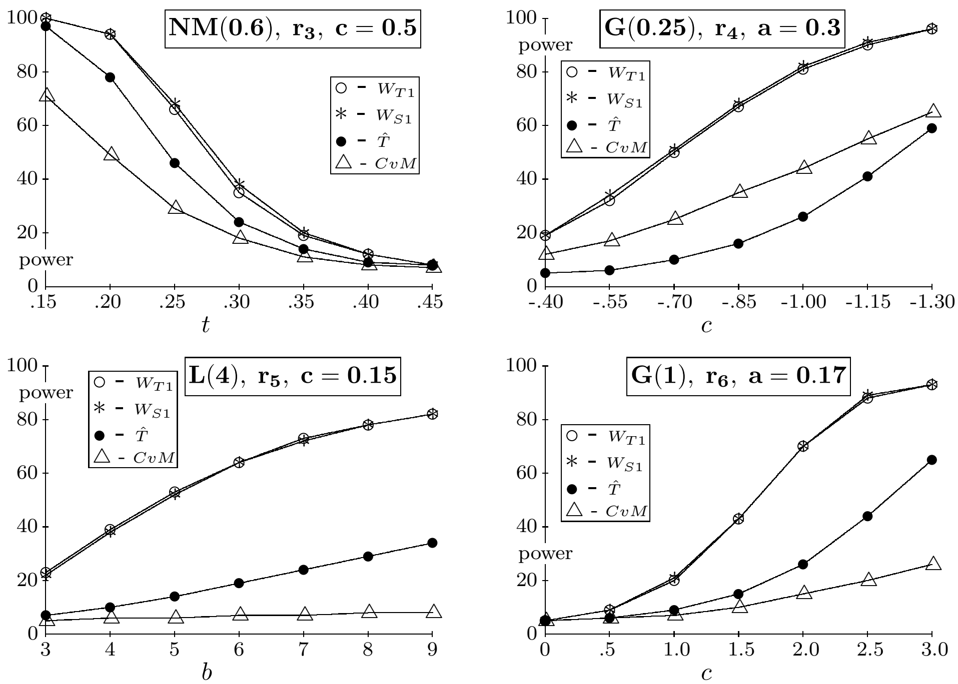

Figure 2: Simulated powers of tests based on W_{T1}, W_{S1}, $\hat{T}$ and CvM under the alternatives $Y = 1 + 2X + r_j(X) + \epsilon$, $l = 3, 4, 5, 6$, X uniform on $[0, 1]$ and different errors. 5% nominal level, n=300, N=10 000 MC runs

The simulation results confirm what might have been expected from our earlier experience.

As characteristic to data-driven Neyman tests with the Schwarz penalty, W_{S1} is powerful for "smooth" deviations from linear regression, while the minimax data-driven chi-square-type statistic $\hat{T}$ of Guerre and Lavergne [9] is more powerful for some extreme deviations, [such as highly oscillating alternatives, in particular].

However, Figure 1 [see the case $NM(0.6)$, r_1] shows that under large variance the $\hat{T}$ test has difficulty in detecting some high frequency oscillations. Under smaller μ and the same signal/noise ratio this drawback disappears. We also observed that $\hat{T}$ loses a lot of its power when the variance of the model error is small [cf. Figure 1, $G(0.25), r_1$] or the model is very close to the null model [cf. Figure 2]. Note also that in Guerre and Lavergne [8] it is shown that $\hat{T}$ compares favourably with the solution of Horovitz and Spokoiny [12]. The refined selection rule $T1$ works very well. In comparison to W_{S1}, one observes only a slight decrease in power of W_{T1} under low dimensional deviations and, simultaneously, a great gain in power under high dimensional alternatives. In all the cases considered, except $L(4)$ and r_2 with $o = 6,\ 7,\ 8$, the power of W_{T1} is larger than that of $\hat{T}$ and in many cases powers dramatically differ in favour of W_{T1}.

The behaviour of CvM is unsatisfactory. Obviously, the low power of CvM test is not due to the transformation, but follows from the nature of such a statistic.

4 Asymptotic comparisons

It is known that the Cramér-von Mises test is only capable of detecting very smooth deviations from the null model. Various known and recently studied aspects of its power behaviour are discussed in Inglot and Ledwina [14], [16], e.g. Guerre and Lavergne [9] have proved that their test is rate-optimal. It seems that this optimality notion does not have clear practical interpretation [cf. [16] for some illustration]. Our constructions would require a nonstandard analysis, as using Schwarz rule in some neibhourhood of the null model excludes standard approaches. Some more detailed comments on this issue are given in [14], e.g.

References

[1] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*,31:225–251, 2003.

[2] P. J. Bickel. On adaptive estimation. *Ann. Statist.*, 10:647–671, 1982.

[3] S. Choi, W. J. Hall, and A. Schick. Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Ann. Statist.*, 24:841–861, 1996.

[4] H. Dette, and A. Munk. A simple goodness-of-fit test for linear models under a random design assumption. *Ann. Inst. Statist. Math.*, 50:253–275, 1998.

[5] H. Dette. On a nonparametric test for linear relationships. *Statist. Probab. Lett.*, 46:307–316, 2000.

[6] J. Diebolt, and J. Zuber. On the half-sample method for goodness-of-fit in regression. Technical Report 00.1 Department of Mathematics and Statistics, La Trobe University, 2000.

[7] J. Fan, and L.-S. Huang. Goodness-of-fit tests for parametric regression models. *J. Amer. Statist. Assoc.*, 96:640–652, 2001.

[8] E. Guerre, and P. Lavergne. Data-driven rate-optimal specification testing in regression models. (`http://lavergne.pascal.free.fr`), 2003.

[9] E. Guerre, and P. Lavergne. Data-driven rate-optimal specification testing in regression models. *Ann. Statist.*, 33:840–870, 2005.

[10] W. Härdle, and E. Mammen. Comparing nonparametric versus parmetric regression fits. *Ann. Statist.*, 21:1926–1947,1993.

[11] J. D. Hart. *Nonparametric smoothing and lack-of-fit tests.* Springer, New York, 1997.

[12] J. L. Horowitz, V. G. and Spokoiny. An adaptive, rate optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69:599–631, 2001.

[13] T. Inglot, W. C. M. Kallenberg, T. and Ledwina. Data driven smooth tests for composite hypotheses. *Ann. Statist.*, 25:1222–1250, 1997.

[14] T. Inglot, and T. Ledwina. Intermediate approach to comparison of some goodness-of-fit tests. *Ann. Inst. Statist. Math.*, 53:810–834, 2001.

[15] T. Inglot, and T. Ledwina. Semiparametric regression: Hadamard differentiability and efficient score functions for some testing problems. Preprint 012, Institute of Mathematics, Wrocław University of Technology, 2003.

[16] T. Inglot, and T. Ledwina. On consistent minimax distinguishability and intermediate efficiency of Cramér-von Mises test. *J. Statist. Plann. Inference*, 124:453–474, 2004.

[17] T. Inglot, and T. Ledwina. Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra Appl.* [special issue], 2006.

[18] T. Inglot, and T. Ledwina. Asymptotic optimality of new adaptive test in regression model. *Ann. I. H. Poincaré*, in print, 2006.

[19] T. Inglot, and T. Ledwina. Data driven score tests of fit for a semiparametric homoscedastic linear regression model. Preprint 665, Institute of Mathematics of the Polish Academy of Sciences, 2006.

[20] T. Inglot, and T. Ledwina. Data driven score tests for homoscedastic linear regression model: asymptotic results. *Probab. Math. Statist.*, 26 [dedicated to the memory of K. Urbanik], to appear, 2006.

[21] T. Inglot, and T. Ledwina. Intermediate efficiency of some max-type statistics. *J. Statist. Plann. Inference*, in print, 2006.

[22] W. C. M. Kallenberg, and T. Ledwina. Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simul.*, 59:101–121, 1997.

[23] E. V. Khmaladze. Martingale approach in the theory of goodness-of-fit tests. *Theory Probab. Appl.*, 26:240–257, 1981.

[24] E. V. Khmaladze, and H. L. Koul. Martingale transforms goodness-of-fit tests in regression models. *Ann. Statist.*, 32:995–1034, 2004.

[25] R. Koenker, and Z. Xiao. Inference on the quantile regression process: appendices. (`http://www.econ.uiuc.edu/~roger/research/inference`), 2001.

[26] A. S. Kozek. A nonparametric test of fit of a parametric model. *J. Multivariate Anal.*, 37:66–75, 1991.

[27] T. Ledwina. Data driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.*, 89:1000–1005, 1994.

[28] J. Rice. Bandwidth choice for nonparametric regression. *Ann. Statist.*, 12:1215–1230, 1984.

[29] A. Schick. On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, 14:1139–1151, 1986.

[30] A. Schick. Efficient estimates in linear and nonlinear regression with heteroscedastic error. *J. Statist. Plann. Inference*, 58:371–387, 1997.

[31] W. Stute. Nonparametric model checks for regression. *Ann. Statist.*, 25:613–641, 1997.

[32] W. Stute, S. Thies, and L.-X. Zhu. Model checks for regression: an innovation process approach. *Ann. Statist.*, 26:1916–1934, 1998.

[33] W. Stute, W. G. Manteiga, and M. P. Quindimil. Bootstrap approximation in model checks in regression. *J. Amer. Statist. Assoc.*, 93:141–149, 1998.

[34] C. Zhang, and H. Dette. A power comparison between nonparametric regression tests. *Statist. Probab. Lett.*, 66:289–301, 2004.

Tadeusz Inglot: Polish Academy of Sciences and Wrocław University of Technology, Institute of Mathematics and Informatics, ul. Wybrzeże Wyspiańskiego 27, Wrocław, 50-370, Poland, inglot@im.pwr.wroc.pl

Teresa Ledwina: Polish Academy of Sciences, Institute of Mathematics, ul. Kopernika 18, Wrocław, 51-617, Poland, ledwina@impan.pan.wroc.pl