APPLICATIONES MATHEMATICAE
37,4 (2010), pp. 387–411

Teresa Ledwina (Wrocław)
Jan Mielniczuk (Warszawa)

VARIANCE FUNCTION ESTIMATION VIA MODEL SELECTION

Abstract. The problem of estimating an unknown variance function in a random design Gaussian heteroscedastic regression model is considered. Both the regression function and the logarithm of the variance function are modelled by piecewise polynomials. A finite collection of such parametric models based on a family of partitions of support of an explanatory variable is studied. Penalized model selection criteria as well as post-model-selection estimates are introduced based on Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) methods of estimation of the parameters of the models. The estimators are defined as ML or REML estimators in the models with dimensions determined by respective selection rules. Some encouraging simulation results are presented and consistency results on the solution pertaining to ML estimation for this approach are proved.

I. Introduction. Heteroscedastic regression models

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

with m and σ unknown, for which the variance of errors depends on explanatory variables, are commonly used in various fields including engineering, biology and economics. In some instances, estimation of the error variance function is of independent interest. Often, it is also important to use an estimator of variance to evaluate other related quantities, as e.g. in the case of Value at Risk (VaR) estimation. Moreover, for such models regression estimation methods account for heteroscedasticity by using some estimator of variance. A typical example is the Weighted Least Squares (WLS)

2010 *Mathematics Subject Classification*: 62J99, 62F12, 62G20.
Key words and phrases: Bayes Information Criterion, heteroscedastic regression, Kullback–Leibler distance, linear model, model selection, penalized likelihood estimator, variance function estimation.

DOI: 10.4064/am37-4-1

method in linear heteroscedastic regression. For a more detailed discussion of these aspects and some references to the vast literature on estimation for heteroscedastic models we refer to Fan and Yao (1998) and Ruppert et al. (2003).

Many developments concentrate mainly on fixed design or conditional set-up and study both parametric and nonparametric approaches. For a thorough discussion and some new developments for this case see e.g. Davidian and Carroll (1987), Müller and Stadtmüller (1993), Dette et al. (1998), Yuan and Wahba (2004) as well as Cai and Wang (2008).

The random design literature is mainly focused on some nonparametric approaches. The main observation underlying most of these approaches is that, provided X and ϵ are independent, the equality $E[\{Y - m(X)\}^2 \mid X = x] = \sigma^2(x)E\epsilon^2$ holds. Thus an estimator of σ^2 can be constructed by regressing squared residuals from a preliminary regression fit on explanatory variables. This is tantamount to a two-step procedure in which the regression function is estimated in the first step and the variance in the second. For examples exploiting kernel, local polynomial estimators and localized likelihoods see Carroll (1982), Silverman (1985), Müller and Stadtmüller (1993), Neumann (1994), Ruppert et al. (1997), Fan and Yao (1998), as well as Yu and Jones (2004). Some results (cf. Fan and Yao (1998) and Yu and Jones (2004)) indicate that the asymptotic behaviour of such variance estimators, at least in pointwise sense, is the same as if the regression function were known, i.e. as if the estimators were based on (unknown) squared errors.

Another idea is to model the mean and the variance in a flexible way and to combine modelling with some penalized likelihood approach. This idea, with penalization related to a degree of roughness of m and σ, was introduced and thoroughly discussed in Yau and Kohn (2003). See also Yuan and Wahba (2004) for related results.

For the random design case, we propose an approach based on model selection. The main idea is to simultaneously estimate the regression and the variance functions and to use penalization pertaining to complexity of the underlying models. To be specific, we concentrate on Gaussian errors and the case where the univariate explanatory variable is assumed to take values in a fixed finite interval $[a, b]$. X and ϵ are independent. We assume that the set of candidate models is finite and contains a correct model. The way we model the mean and the variance was inspired by some ideas developed for heteroscedastic fixed design linear models by Harvey (1976) and Verbyla (1993) as well as by solutions proposed in density estimation by Castellan (2003) and Birgé and Rozenholc (2006). Consequently, both the regression function and the logarithm of the variance function are modelled by piecewise polynomials. We study a finite collection of such parametric Gaussian models based on a family of partitions of

the support of the explanatory variable. We consider the Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) methods of estimation of the parameters of a given model or its appropriate transformation and introduce related penalized model selection criteria as well as post-model-selection estimates. The estimators are defined as ML or REML estimators in the models with dimensions determined by suitable selection rules. Motivated by appealing features of the Minimum Description Length (MDL) approach, we use one of the simplest forms of the two-part MDL principle which results in a penalty of the form appearing in the Bayes Information Criterion (BIC) introduced by Schwarz (1978). We shall refer to it as a BIC-type penalty. For the two-part MDL we refer to Lee (2001). Our theoretical development concentrates on ML estimation and a related post-model-selection estimator. We prove consistency of the selection rule from which mean square consistency of the proposed variance estimator follows immediately. We also consider the behaviour of that rule when the underlying distribution of (X, Y) is misspecified, i.e. it does not belong to any model on the list of models. The estimators pertaining to the REML method are used for comparison purposes in numerical experiments.

Simulation results show that our approach outperforms purely nonparametric approaches of Fan and Yao (1998) as well as of Yu and Jones (2004). This illustrates and supports some earlier beliefs and findings to the effect that for small and moderately large sample sizes it is better to use a suitably chosen parametric model than a nonparametric one. The model selection criterion that we use indeed implies that fitting parsimonious models is of primary interest.

Though the results of the present paper focus on Gaussian random regression with a one-dimensional explanatory variable, several extensions are possible. First, one can allow for errors having an arbitrary fixed density which satisfies certain smoothness conditions. The case of multivariate compactly supported predictors may be treated similarly with the aid of product orthonormal systems. Moreover, Borkowski and Mielniczuk (2010) recently applied this approach to the problem of conditional variance estimation in the case of a nonlinear autoregressive heteroscedastic process.

The paper is organized as follows. Details on the proposed family of models are contained in Secs. II.A and II.B. In Sec. II.C penalized likelihoods are introduced while some of the asymptotic properties are studied and discussed in Sec. II.D. Sec. II.E briefly discusses our numerical experiments. The proofs of the main results are given in Sec. III. Appendix I collects some analytical derivations while Appendix II provides some details on the numerical experiments.

II. Statistical framework and results

A. Parametric heteroscedastic regression models $\mathcal{P}_{kl}$. Assume that the explanatory random variable X takes values in a fixed interval $[a,b]$. Consider the equipartition of $[a,b]$ into k intervals, $k \in \{1,\dots,K\}$. Let M_{ck} denote the cth interval of the partition, $c = 1,\dots,k$. For fixed k and $s \in \mathbb{N}$, let $\Phi_{c0}(\cdot),\dots,\Phi_{c(s-1)}(\cdot)$ denote the Legendre polynomials of order $0,1,\dots,s-1$, respectively, transformed to the interval M_{ck}. Let

$$b_{sk} = (\beta_{10},\beta_{11},\dots,\beta_{1(s-1)},\beta_{20},\beta_{21},\dots,\beta_{2(s-1)},\dots,\beta_{k0},\beta_{k1},\dots,\beta_{k(s-1)})^T \in \mathbb{R}^{k\cdot s}, \tag{2}$$

where ^{T} stands for transposition. Given s and k, the regression function will be denoted by m_{sk} and parametrized as follows:

$$m_{sk}(x) = \sum_{c=1}^{k}\sum_{j=0}^{s-1} \beta_{cj}\Phi_{cj}(x)\mathbf{1}_{M_{ck}}(x), \qquad x \in [a,b], \tag{3}$$

where $\mathbf{1}_A$ denotes the indicator function of a set A. We shall employ a similar construction to parametrize the logarithm of the variance function. Namely, let S_{rl} denote the rth interval in the equipartition of $[a,b]$ into l intervals, $l \in \{1,\dots,L\}$. Define, for $t \in \mathbb{N}$,

$$e_{tl} = (\eta_{10},\eta_{11},\dots,\eta_{1(t-1)},\eta_{20},\eta_{21},\dots,\eta_{2(t-1)},\dots,\eta_{l0},\eta_{l1},\dots,\eta_{l(t-1)})^T \in \mathbb{R}^{l\cdot t} \tag{4}$$

and let

$$\sigma^2_{tl}(x) = \exp\Big\{\sum_{r=1}^{l}\sum_{j=0}^{t-1} \eta_{rj}\Phi_{rj}(x)\mathbf{1}_{S_{rl}}(x)\Big\}. \tag{5}$$

Note that $s,t \in \mathbb{N}$ are fixed throughout.

We now define a parametric heteroscedastic regression model. For any $B > 0$ let

$$\Theta_{kl} = \Theta^B_{kl} = \{\theta = (\theta_1,\dots,\theta_{s\cdot K+t\cdot L})^T : \theta^T = (b^T_{sk}, e^T_{tl}, 0,\dots,0), \ |\theta_i| \le B,\ i = 1,\dots,s\cdot K + t\cdot L\}, \tag{6}$$

where the vector (b^T_{sk}, e^T_{tl}) is appended by $s\cdot(K-k) + t\cdot(L-l)$ zeros. Observe that the parameter space Θ_{kl} is a compact subset of $\mathbb{R}^{s\cdot K+t\cdot L}$. Let f stand for the density of X and g denote the density of the standard normal distribution $N(0,1)$. Throughout the paper we assume that f is positive on $[a,b]$. For a fixed (k,l) we consider a parametric model $\mathcal{P}_{kl}$ of distributions of (X,Y) defined as follows:

$$\mathcal{P}_{kl} = \{P_\theta : \theta \in \Theta_{kl}\}, \quad \frac{dP_\theta}{d\lambda} = p_\theta(x,y), \tag{7}$$

where

$$p_\theta(x,y) = \frac{1}{\sigma_{tl}(x)} f(x) g\left(\frac{y - m_{sk}(x)}{\sigma_{tl}(x)}\right), \tag{8}$$

while λ is the Lebesgue measure on $[a,b]\times\mathbb{R}$, $m_{sk}(x)$ is given by (3) and $\sigma_{tl}(x)$ is given by (5). Putting it differently, for fixed (k,l), we consider (X,Y) such that

$$Y = m_{sk}(X) + \sigma_{tl}(X)\epsilon, \tag{9}$$

where X and ϵ are independent and $\epsilon \sim N(0,1)$. The model (9) can be thought of as an approximation of the nonparametric regression model $Y = m(X) + \sigma(X)\epsilon$, where m and σ are some unknown functions. However, the approximation effects are not studied analytically in our contribution.

We restrict our attention to the set of parametric models $\{\mathcal{P}_{kl}\}$ for $1 \le k \le K$ and $1 \le l \le L$. Finally, we set $\Theta = \bigcup_{k=1}^{K}\bigcup_{l=1}^{L}\Theta_{kl}$. The aim of the model selection and related post-model-selection estimation is to choose, given the data, a suitable partition and a suitable degree for the polynomial on each interval of the partition and then to consider likelihood-based estimators in the resulting parametric model. For these purposes some careful inspection of properties of the introduced family of models is useful.

B. Properties of the family $\mathcal{P}_{kl}$. (C1)–(C3) below show basic features of the parametrization.

(C1) $\mathcal{P}_{kl}\cap\mathcal{P}_{k'l'}\neq\emptyset$ for any (k,l), (k',l').

(C1) follows from the observation that e.g. distributions with constant regressions and variances belong to each $\mathcal{P}_{kl}$. It indicates that by giving a value of $\theta\in\Theta$ one does not identify the distribution of (X,Y). In fact, the lack of identifiability is due to a more profound difficulty pertaining to the introduced structures. For illustration consider the case $K=2$, $L=8$, $s=t=1$ and the vector $\theta=(1,0,0,2,2,2,0,0,0,0)$. For $k=1$ and $l=8$ this vector describes (9) with a constant regression function and the variance defined by $e_{18}=(0,0,2,2,2,0,0,0)$. However, when $k=2$ and $l=7$ this θ corresponds to (9) with a stepwise regression function, with discontinuity at $(a+b)/2$, and the variance defined via the vector e_{17}. So, the transformation $\theta\mapsto P_\theta$ is meaningless when considered on the whole Θ. However, we have the following property:

(C2) For fixed (k,l), let P_θ be a distribution pertaining to a density defined in (8). Then the mapping $\theta\mapsto P_\theta$ is 1-1 on Θ_{kl}.

(C2) follows from the observation that both conditional means and variances are piecewise continuous functions. Therefore given two different parameter values θ and θ' belonging to Θ_{kl} it is possible to choose a subset on which P_θ and $P_{\theta'}$ differ.

Assume now that for certain (k^0,l^0) and $\theta^0\in\Theta_{k^0l^0}$ we have $(X,Y)\sim P_{\theta^0}$. Then we obviously have

(C3) If $P_{\theta^0}\notin\mathcal{P}_{kl}$ then for any $\theta\in\Theta_{kl}$, $P_{\theta^0}\neq P_\theta$.

The next two properties concern the expected log-likelihoods. They are immediate consequence of (C2) and basic properties of entropy.

(C4) For $\theta \in \Theta_{k^0 l^0}$, the transformation $\theta \mapsto -\mathbb{E}_{P_{\theta^0}} \log p_\theta(X,Y)$ attains its unique minimum on $\Theta_{k^0 l^0}$ at $\theta = \theta^0$.

(C5) If $P_{\theta^0} \in \mathcal{P}_{k^0 l^0} \cap \mathcal{P}_{kl}$ then the transformation $\theta \mapsto -\mathbb{E}_{P_{\theta^0}} \log p_\theta(X,Y)$ attains its unique minimum on Θ_{kl} at a point θ^1 such that $P_{\theta^1} = P_{\theta^0}$.

The following statements describe some regularity properties of $\mathcal{P}_{kl}$.

(C6) For $i,j = 1,\dots,s\cdot K + t\cdot L$ and $\theta \in \Theta_{kl}$ the derivatives $(\partial/\partial\theta_i)p_\theta(x,y)$ and $(\partial^2/\partial\theta_i\partial\theta_j)p_\theta(x,y)$ exist P_θ-a.e. and are bounded by functions not depending on θ which are integrable with respect to Lebesgue measure on $[a,b]\times\mathbb{R}$.

(C7) For i,j and θ as in (C6) the derivatives $(\partial/\partial\theta_i)\log p_\theta(x,y)$ and $(\partial^2/\partial\theta_i\partial\theta_j)\log p_\theta(x,y)$ exist P_θ-a.e. and are bounded by functions not depending on θ which are integrable with respect to $P_{\theta'}$ for any $\theta' \in \Theta$.

(C8) We have

$$\mathbb{E}_{P_\theta}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log p_\theta(X,Y)\right] = -I(\theta),$$

where $I(\theta)$ is the information matrix pertaining to the distribution P_θ. Moreover, for θ^0 and θ^1 defined in (C4) and (C5), respectively, the matrices $I(\theta^0)$ and $I(\theta^1)$ are positive definite.

(C9) Let $h \in \mathbb{R}^{s\cdot K + t\cdot L}$ and let $\|\cdot\|$ denote the Euclidean norm of a vector or a matrix. Then

$$\lim_{\delta\to 0}\mathbb{E}_{P_\theta}\left[\sup_{\|h\|\le\delta}\left\|\frac{\partial^2}{\partial\theta\partial\theta^T}\log\frac{p_{\theta+h}(X,Y)}{p_\theta(X,Y)}\right\|\right] = 0.$$

The proofs of (C6)–(C9) are deferred to Appendix I.

Assume that (X_i,Y_i), $i=1,\dots,n$, are independent and for certain (k^0,l^0) and $\theta^0 \in \Theta_{k^0 l^0}$ we have $(X_i,Y_i) \sim P_{\theta^0}$, $i=1,\dots,n$. For given k,l and $\theta \in \Theta_{kl}$ consider

$$\mathcal{L}_n^{kl}(\theta) = \log\prod_{i=1}^n p_\theta(X_i,Y_i). \tag{10}$$

The last property concerns, possibly misspecified, maximized log-likelihood.

(C10) For any (k,l), the estimator $\hat\theta_{kl} = \arg\max_{\theta\in\Theta_{kl}} \mathcal{L}_n^{kl}(\theta)$ exists and is measurable (cf. White (1982), pp. 3 and 17).

C. Maximum Likelihood estimators within $\mathcal{P}_{kl}$ and some penalized selection criteria. Our basic solution is based on the standard log-likelihood function $\mathcal{L}_n^{kl}(\theta)$ as defined in (10). The estimator $\hat\theta_{kl}$, introduced

in (C10), is the Maximum Likelihood (ML) estimator of θ in the model $\mathcal{P}_{kl}$. It allows for simultaneous estimation of m_{sk} and σ^2_{tl}. As discussed in the introduction, this is in contrast to much more popular two-step procedures for which the mean function is estimated first and then we estimate the variance based on suitably defined residuals.

In order to find the best fitted model $\mathcal{P}_{kl}$ from our list of models and to define related post-model-selection estimators, a two-part MDL criterion will be employed. The pertaining penalty has the form

$$\text{pen}_n(b_{sk}, e_{tl}) = \frac{1}{2}(s \cdot k + t \cdot l) \log n. \tag{11}$$

As in our study the variance function estimation is of primary interest, we also consider an application of the Restricted Maximum Likelihood (REML) estimators of the parameters e_{tl} of the variance $\sigma_{tl}(x)$ in the model $\mathcal{P}_{kl}$ (cf. (5)–(7)). For a nice introduction to and discussion of the REML method we refer to Cressie and Lahiri (1993) and Verbyla (1990). Notice that REML estimators were developed to estimate variances and covariances in linear models with a given experimental matrix. Therefore, we describe the method in terms of the conditional distribution of $(Y_1, \dots, Y_n)$ given $X_1 = x_1, \dots, X_n = x_n$. For fixed (k, l), this conditional distribution corresponds to a Gaussian linear model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} m_{sk}(x_1) \\ \vdots \\ m_{sk}(x_n) \end{pmatrix} + [\Sigma(e_{tl})]^{1/2} \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{12}$$

where $\Sigma(e_{tl})$ is the $n \times n$ diagonal matrix with $\sigma^2_{tl}(x_1), \dots, \sigma^2_{tl}(x_n)$ on the diagonal (cf. (9)). Obviously, due to (3), $(m_{sk}(x_1), \dots, m_{sk}(x_n))^T = \mathcal{X} b_{sk}$ for a suitable $n \times sk$ matrix $\mathcal{X}$ and b_{sk} given by (2). Set $\mathcal{Y} = (Y_1, \dots, Y_n)^T$, $\mathcal{E} = (\epsilon_1, \dots, \epsilon_n)^T$ and $p = s \cdot k$. Then (12) takes the form

$$\mathcal{Y} = \mathcal{X} b_{sk} + [\Sigma(e_{tl})]^{1/2} \mathcal{E}. \tag{13}$$

Our focus is now on estimation of the vector of parameters e_{tl}. The idea behind the REML method is to apply the maximum likelihood principle to error contrasts rather than to the data themselves. The error contrasts are linear combinations of components of $(Y_1, \dots, Y_n)^T$ having 0 mean. Due to this, the influence of the mean vector of the original observations on the estimation process is reduced. Following Harville (1974), similarly to Cressie and Lahiri (1993), consider the particular linear transformation given by an $n \times (n-p)$ matrix Γ with $n-p$ linearly independent columns and such that

$$\Gamma^T \mathcal{X} = 0, \quad \Gamma^T \Gamma = I, \quad \Gamma \Gamma^T = I - \mathcal{X}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T,$$

where I is the $(n-p) \times (n-p)$ identity matrix. Set $U = \Gamma^T \mathcal{Y}$, $u = \Gamma^T y$

where $y = (y_1, \dots, y_n)^T$. Then the log-likelihood function, based on U, has the form

$$\mathcal{L}_U^{kl}(e_{tl}) = -\frac{n-p}{2}\log 2\pi + \frac{1}{2}\log|\mathcal{X}^T\mathcal{X}| - \frac{1}{2}\log|\mathcal{X}^T[\Sigma(e_{tl})]^{-1}\mathcal{X}| - \frac{1}{2}\log|\Sigma(e_{tl})| + \frac{1}{2}y^T\Pi(e_{tl})y,$$

where $|A|$ stands for the determinant of the matrix A, while

$$\Pi(e_{tl}) = [\Sigma(e_{tl})]^{-1} - [\Sigma(e_{tl})]^{-1}\mathcal{X}(\mathcal{X}^T[\Sigma(e_{tl})]^{-1}\mathcal{X})^{-1}\mathcal{X}^T[\Sigma(e_{tl})]^{-1}.$$

The REML estimator $\tilde{e}_{tl}$ of e_{tl} in $\mathcal{P}_{kl}$ (cf. also (6)) is defined by

$$\tilde{e}_{tl} = \arg\max_{(0,\dots,0,e_{tl}^T,0,\dots,0)\in\Theta_{kl}} \mathcal{L}_U^{kl}(e_{tl}).$$

There is considerable evidence available indicating that in some situations both finite-sample and asymptotic properties of REML estimators can be more appealing than those of ML estimators. For a related discussion we refer to Cressie and Lahiri (1993) and Jiang (1996). In particular, the distinction between the two classes is noticeable when p is large relative to n. On the other hand, due to dependence of components of U, introduced by the transformation Γ, the analysis of REML estimators is much more complex than the analysis of MLEs. Since our simulation study has not exhibited a significant superiority of REML-based procedures over MLE-based ones, our theoretical results are restricted to the ML approach.

For further purposes note that for inference based on $\mathcal{L}_U^{kl}(e_{tl})$, a BIC-type penalty has the form

$$\operatorname{pen}_n(e_{tl}) = \frac{1}{2}(t\cdot l)\log(n-p), \qquad p = s\cdot k. \tag{14}$$

This follows from the fact that one estimates by the ML method the $t\cdot l$-dimensional vector of parameters e_{tl} using the $(n-p)$-dimensional vector u.

D. Main asymptotic results. Our main result concerns the situation when the true distribution P of (X,Y) belongs to one (but not necessarily the only one) of the parametric families of the list, i.e. there exist (k^0,l^0) and $\theta^0 \in \Theta_{k^0l^0}$ such that $P = P_{\theta^0}$. Therefore, consider $\mathcal{P}_{k^0l^0}$ and $\mathcal{P}_{kl}$, $(k,l) \neq (k^0,l^0)$, and the corresponding penalized log-likelihood functions (cf. (10) and (11))

$$\mathcal{S}_0 = \sup_{\theta\in\Theta_{k^0l^0}} \mathcal{L}_n^{k^0l^0}(\theta) - \omega_0\log n, \quad \mathcal{S}_1 = \sup_{\theta\in\Theta_{kl}} \mathcal{L}_n^{kl}(\theta) - \omega_1\log n,$$

where $\omega_0 = 2^{-1}(s\cdot k^0 + t\cdot l^0)$ and $\omega_1 = 2^{-1}(s\cdot k + t\cdot l)$.

THEOREM 1. *Assume that $f(x) > 0$ for $x \in [a,b]$, and moreover $(X,Y) \sim P = P_{\theta^0}$ and $\theta^0 \in \Theta_{k^0l^0}$. If $P_{\theta^0} \in \mathcal{P}_{k^0l^0} \cap \mathcal{P}_{kl}$ and $s\cdot k + t\cdot l > s\cdot k^0 + t\cdot l^0$,*

or if $P_{\theta^0} \in \mathcal{P}_{k^0 l^0} \setminus \mathcal{P}_{kl}$, *then*

$$\lim_{n\to\infty} P(\mathcal{S}_0 > \mathcal{S}_1) = 1. \tag{15}$$

The proof of Theorem 1 is given in Sec. III.

REMARK 1. Theorem 1 states that, when using a penalty of the form $C \log n$ and with $n \to \infty$, there is an increasing tendency to select the most parsimonious model containing the true distribution. A rather obvious observation, based on the proof, is that the consistency result can be immediately generalized to other penalties which tend to infinity at a rate $o(n)$.

Consider now the selection procedure described as follows. Let $\hat{\theta}_{kl}$ be the ML estimator for $\theta \in \Theta_{kl}$ and define

$$(\hat{k}, \hat{l}) = \arg \max_{1\le k\le K, 1\le l\le L} \left\{ \mathcal{L}_n^{kl}(\hat{\theta}_{kl}) - \frac{1}{2}(s\cdot k + t\cdot l)\log n \right\}.$$

Moreover, still assuming that P belongs to at least one parametric model on the list, denote by $k^\diamond$ the integer $k = 1, \dots, K$ pertaining to the shortest vector b_{sk} in (2) describing the true regression function. Let $l^\diamond$ be defined analogously by the shortest description e_{tl} of the true variance function.

COROLLARY 1. *Suppose that the assumptions of Theorem* 1 *hold. Then*

$$\lim_{n\to\infty} P((\hat{k}, \hat{l}) = (k^\diamond, l^\diamond)) = 1.$$

Introduce now the post-model-selection estimator $\hat{\sigma}^2$ of the true variance function, related to the selection rule $(\hat{k}, \hat{l})$ by

$$\hat{\sigma}^2(x) = \hat{\sigma}^2_{t\hat{l}}(x) = \exp\Big\{ \sum_{r=1}^{l} \sum_{j=0}^{t-1} \hat{\eta}_{rj} \phi_{rj}(x) \mathbf{1}_{S_{rl}}(x) \Big\}. \tag{16}$$

REMARK 2. The form of $\hat{\sigma}^2(x)$ shows that the estimator is always positive and clearly based on the most adequate model on the list.

For $(k^\diamond, l^\diamond)$ defined above, put $\sigma^2_\diamond(x) = \sigma^2_{tl^\diamond}(x)$. Obviously, $\sigma^2_\diamond$ gives the most parsimonious description of the true variance function among those available on our list of models.

COROLLARY 2. *Under the assumptions of Theorem* 1,

$$\lim_{n\to\infty} E_P\{ \sup_{x\in[a,b]} [\hat{\sigma}^2(x) - \sigma^2_\diamond(x)]^2 \} = 0.$$

A justification of this result is provided in Sec. III.

REMARK 3. Corollary 2 states that $\hat{\sigma}^2$ is consistent in the mean-square sense. Obviously, Corollary 2 holds with $\sigma^2_\diamond(\cdot)$ replaced by any other valid

parametrization of the true variance function available on the list. Note that there is growing evidence showing that consistency of post-model-selection estimators may lead to superefficiency and therefore should be considered with caution. For a thorough discussion of this and related problems see e.g. Leeb and Pötscher (2008). On the other hand, recall that (C1) of Sec. II.B shows that identifiability of parameters is lost in the overall model $\{P_\theta : \theta \in \Theta\}$. In that case the conclusion of Sec. 4 of Pötscher (1991) is clearly in favour of consistent model selection. Moreover, Dukič and Peña (2005) indicate that using consistent selection rules in post-model-selection estimation is a good strategy if the number of competing models is relatively small and the submodels are essentially distinct.

Remark 4. There are many general results on consistency of selection rules: see Kohn (1983) and Sin and White (1996), for example. However, the lack of identifiability of our overall model discussed in connection with (C1) makes Kohn's results not immediately applicable to our case. In particular, an objective function l_n to be maximized, considered in his paper, is not well defined on the whole Θ in our setting. On the other hand, the level of generality of Sin and White (1996) forces a series of very involved assumptions which are clearly unnecessarily demanding for the models considered here. Therefore, in Sec. III we provide a detailed proof of Theorem 1 for the parametric list of models under study.

Below, for completeness, we also discuss an interesting generalization of Theorem 1 to a case where the observations are generated from a model outside the list under consideration. To state the result we first introduce the notion of Kullback–Leibler distance and a related concept of pseudo-true parameter.

For an arbitrary P with a density $p(x,y)$ with respect to the Lebesgue measure on $[a,b]\times\mathbb{R}$ consider the Kullback–Leibler distance $D(P\|P_\theta)$ between P and P_θ, where $P_\theta \in \mathcal{P}_{kl}$, given by

$$\begin{aligned} D(P\|P_\theta) &= \int p(x,y)\log\left(\frac{p(x,y)}{p_\theta(x,y)}\right)dx\,dy \\ &= \int p(x,y)\log p(x,y)\,dx\,dy \\ &\quad - \int p_1(x)\log f(x)\,dx - \int p(x,y)\log\frac{1}{\sigma_{tl}(x)}g\left(\frac{y-m_{sk}(x)}{\sigma_{tl}(x)}\right)dx\,dy, \end{aligned}$$

where $p_1(x) = \int p(x,y)\,dy$. It is easy to see that $D(P\|P_\theta)$ is a continuous function of θ on Θ_{kl}. As Θ_{kl} is compact, $\theta^*_{kl} = \arg\min_{\theta\in\Theta_{kl}} D(P\|P_\theta)$ exists although it may not be unique. Any θ^*_{kl} satisfying the last equality is called a *pseudo-true parameter value* (cf. Sawa (1978), p. 1276). Note that $\theta^*_{kl} = \arg\max_{\theta\in\Theta_{kl}} \mathbb{E}_P \log p_\theta(X,Y)$.

THEOREM 2. *Let $(X,Y) \sim P$, where P is dominated by the Lebesgue measure on $[a,b]\times\mathbb{R}$. Suppose that $E_P Y$ exists and is finite. Further assume that X has a density $f(x) > 0$ for $x \in [a,b]$ and $P \notin \mathcal{P}_{kl}$ for any (k,l). Moreover, assume that there exists a unique parametric model $\mathcal{P}_{k^*l^*}$ closest to P, so that $D(P\|P_{\theta^*_{k^*l^*}}) < D(P\|P_{\theta^*_{kl}})$ for any $(k,l) \neq (k^*,l^*)$, where θ^*_{ij}, $1 \le i \le K$, $1 \le j \le L$, is the pseudo-true value. Then*

$$\lim_{n\to\infty} P((\hat{k},\hat{l}) = (k^*,l^*)) = 1.$$

The proof of Theorem 2 is given in Sec. III. Note that we do not assume that $P_{\theta^*_{k^*l^*}}$ has to be unique.

REMARK 5. As mentioned earlier, we do not discuss the asymptotic behaviour of $\hat{\sigma}^2$ in the general case when the observations are generated by models outside the given list. Though undoubtedly instructive, such results are highly technical as a rule. Instead, in the next section we present results of a simulation study which show that the proposed class of estimators works nicely also in cases when the generating mechanism is not included in the list of models. Moreover, in the simulations we allow the dimension of the list of models to grow with n. This aspect was not included in our theoretical considerations. Our primary goal was to show a new flexible alternative to existing methods of solving this important and not easy problem. Overcoming technical difficulties was not our primary goal in this case.

E. Simulation results. Though the literature on variance function estimation is vast, comprehensive numerical studies concerning performance of various methods are sparse. Recently, Yu and Jones (2004) presented a relatively detailed study. Our experiments include the two regression models investigated in this paper. Moreover, we have considered the same sample sizes and we have used the same indices to assess the performance of the estimators.

More precisely, we considered two sample sizes $n = 100$ and $n = 500$ and the corresponding number of partitions $k \in \{2,3,4\}$, $l \in \{2,\dots,5\}$ if $n = 100$ and $k \in \{2,\dots,10\}$ and $l \in \{2,\dots,15\}$ if $n = 500$. We took $s = 3$ and $t = 1$ in both cases, i.e., regardless of the sample size, the regression function was modelled by piecewise second order Legendre polynomials on succeeding intervals and the logarithm of the variance function by stepwise functions. Both examples considered in Yu and Jones (2004) fall outside our class of models. We supplemented them with two further examples with the same property. In order to obtain a more complete picture we also included four examples in which at least one modelled function (mean or variance) belongs to the introduced class. A detailed description of the models con-

sidered is given in Appendix II. Here, in Figures 1–8, we simply plot the conditional means as well as standard deviations for each case considered and label them as r_j and sd_j, $j = 1, \ldots, 8$, respectively. The distribution of X is indicated in the captions of Tables 1–8. We use the notation $N_{[t]}(\mu, \vartheta^2)$ for the normal distribution with mean μ and variance ϑ^2, truncated to an appropriate interval $[a, b]$ on which X is supported. By $U[a, b]$ we denote the uniform distribution on $[a, b]$.

We studied two post-model-selection estimators $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ related to the above list of models, MLE and REML principles and the selection rules

$$(\hat{k}, \hat{l}) = \arg \max_{2 \le k \le K, 2 \le l \le L} \left\{ \mathcal{L}_n^{kl}(\hat{\theta}_{kl}) - \frac{1}{2}(s \cdot k + t \cdot l) \log n \right\},$$

$$\tilde{l} = \arg \max_{2 \le k \le K, 2 \le l \le L} \left\{ \mathcal{L}_U^{kl}(\tilde{e}_{tl}) - \frac{1}{2}(t \cdot l) \log(n - s \cdot k) \right\},$$

with $K(100) = 4$, $L(100) = 5$, and $K(500) = 10$, $L(500) = 15$, respectively. The estimators $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ have the structure given in (16), where the sum over l starts with $l = 2$ and respective pairs $(\hat{l}, \hat{e}_{tl})$ and $(\tilde{l}, \tilde{e}_{tl})$ are employed. In order to calculate the REML estimators the algorithm `remlscore` of Smyth (2002) from his R package `statmod` was used. We considered $k \ge 2$ and $l \ge 2$ for candidate models as `remlscore` failed to converge considerably more frequently for the simplest models with $k = 1$ or $l = 1$.

For comparison, we include related simulation results for the procedure introduced by Fan and Yao (1998), which is an application of local linear regression to the squared residuals, from regression fit, matched with some automatic bandwidth selection method. The C code of this procedure is available from the web site for Fan and Yao (2003). The resulting estimator is denoted by $\bar{\sigma}^2$.

The performance of any variance estimator in our experiment was evaluated using the empirical Integrated Standard Error (ISE). In the case of estimating σ^2 by $\hat{\sigma}^2$ this quantity equals $\mathrm{ISE} = n^{-1} \sum_{i=1}^{n} [\hat{\sigma}^2(X_i) - \sigma^2(X_i)]^2$, for other variance estimators considered it is defined analogously. For each model we also computed the empirical MISE, i.e. the average of ISE over simulation runs, the Standard Error (SE) of ISE as well as the Mean Integrated Variance (MIV) and the Integrated Squared Bias (ISB). For exact definitions of the last two quantities see Appendix II. It should be noted that, in contrast to our study, Yu and Jones (2004) calculated the standard error of MISE instead of the standard error of ISE. Therefore, in order to compare their results with ours, the values of SE reported in their Tables 2 and 3 have to be multiplied by c.a. 31.6, i.e. the square root of 1000, the number of simulation runs considered in their paper. Our results are based on 2000 samples generated for each parametric model.

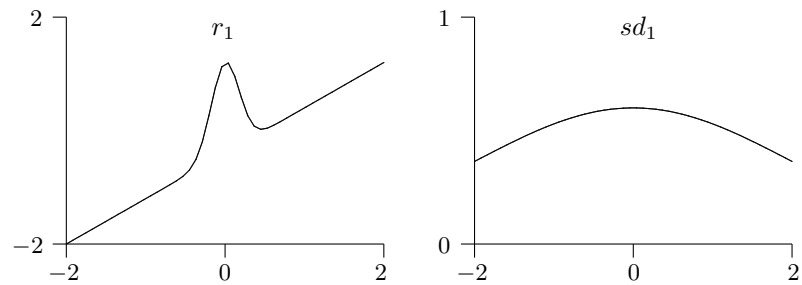

Fig. 1. Regression and standard deviation for Example 1

Table 1. Example 1: $X \sim N_{[t]}(0,1)$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	237.1	214.4	153.8	83.8	20.7	13.3	1.2	19.6
$\tilde{\sigma}^2$	45.2	33.2	9.7	36.2	16.4	6.5	5.2	11.3
$\hat{\sigma}^2$	40.3	23.0	10.0	35.9	17.6	6.5	5.3	11.4

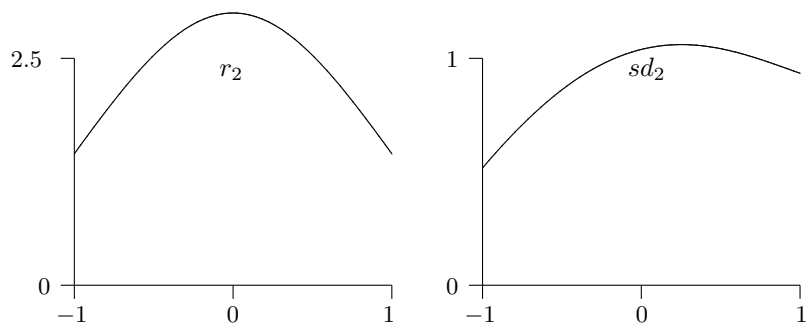

Fig. 2. Regression and standard deviation for Example 2

Table 2. Example 2: $N_{[t]}(0, [0.5]^2)$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	534.0	368.1	44.2	499.0	157.3	100	3.2	154.3
$\tilde{\sigma}^2$	407.1	338.7	74.9	336.4	146.0	60.2	35.9	110.3
$\hat{\sigma}^2$	400.9	249.5	75.0	336.4	153.3	60.6	36.0	110.3

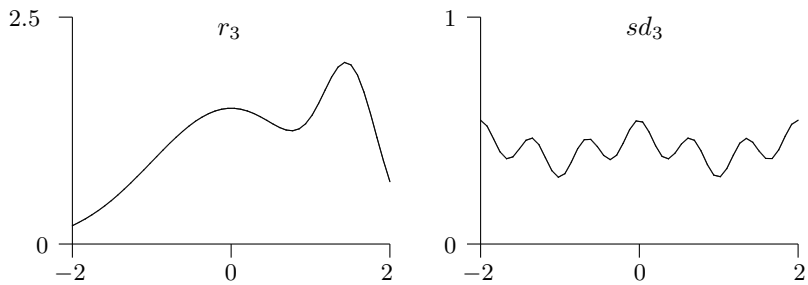

Fig. 3. Regression and standard deviation for Example 3

Table 3. Example 3: $X \sim U[-2, 2]$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	32.5	16.7	11.1	23.5	13.1	5.2	5.2	8.1
$\tilde{\sigma}^2$	23.2	11.5	13.4	10.6	15.7	2.1	13.0	2.9
$\hat{\sigma}^2$	23.2	9.1	13.4	10.6	16.0	2.0	13.0	2.9

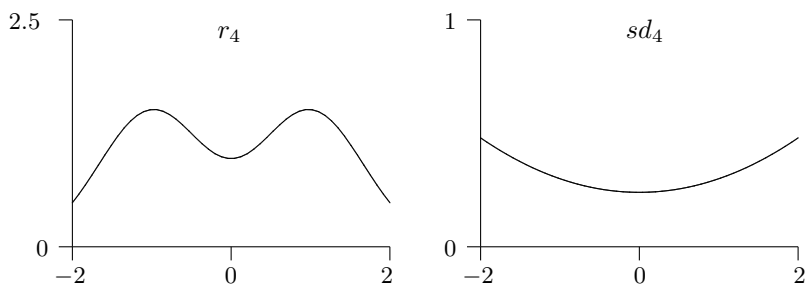

Fig. 4. Regression and standard deviation for Example 4

Table 4. Example 4: $X \sim N_{[t]}(0, 1)$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	8.8	6.7	2.8	6.9	2.9	2.1	0.3	2.6
$\tilde{\sigma}^2$	8.6	6.7	2.7	6.1	3.6	1.3	1.5	2.1
$\hat{\sigma}^2$	7.9	3.9	2.8	6.0	3.6	1.2	1.5	2.1

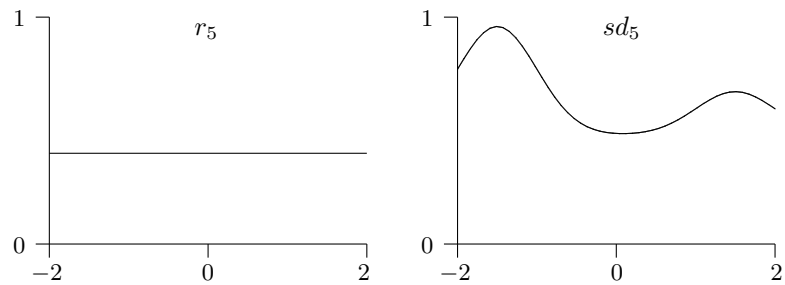

Fig. 5. Regression and standard deviation for Example 5

Table 5. Example 5: $X \sim N_{[t]}(0,1)$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	145.1	98.1	37.4	116.1	51.2	42.0	3.1	48.4
$\tilde{\sigma}^2$	168.5	136.4	30.5	140.1	58.7	24.6	17.8	40.6
$\hat{\sigma}^2$	149.4	82.0	30.5	141.4	59.5	20.4	18.3	39.6

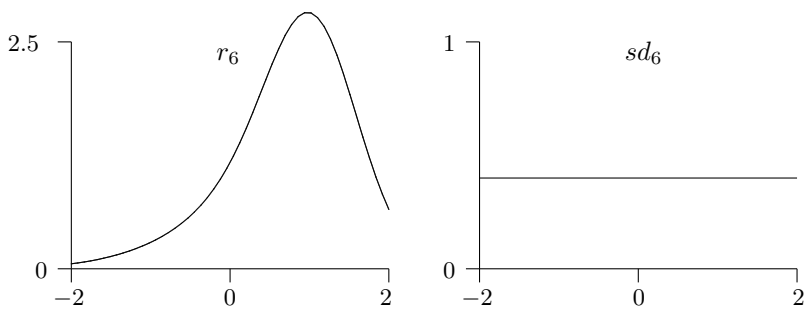

Fig. 6. Regression and standard deviation for Example 6

Table 6. Example 6: $X \sim N_{[t]}(0,1)$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	27.6	17.6	4.65	24.2	10.2	6.2	0.48	9.8
$\tilde{\sigma}^2$	14.3	16.7	0.02	14.2	2.4	2.6	0.02	2.4
$\hat{\sigma}^2$	14.5	13.3	0.03	14.0	3.2	3.0	0.01	2.5

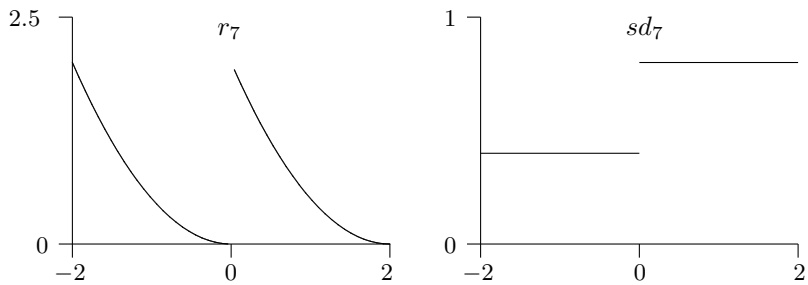

Fig. 7. Regression and standard deviation for Example 7

Table 7. Example 7: $X \sim U[-2, 2]$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	457.2	372.3	115.5	350.1	151.9	107.0	52.9	100.1
$\tilde{\sigma}^2$	115.9	165.9	4.5	119.7	18.0	25.3	2.0	20.0
$\hat{\sigma}^2$	120.8	135.9	4.5	119.7	23.5	29.0	2.0	20.0

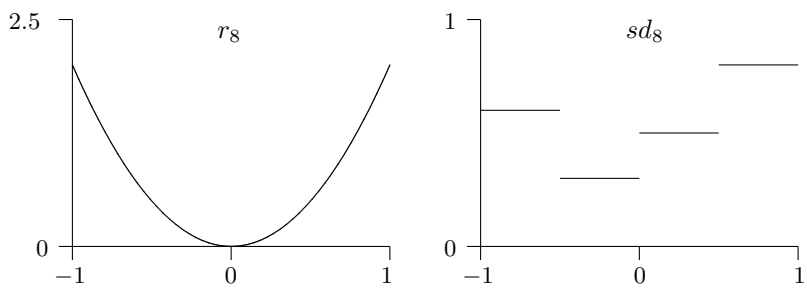

Fig. 8. Regression and standard deviation for Example 8

Table 8. Example 8: $X \sim U[-1, 1]$

	$n = 100$				$n = 500$			
	MISE	SE	ISB	MIV	MISE	SE	ISB	MIV
$\bar{\sigma}^2$	277.8	191.5	61.8	227.1	104.2	81.3	31.2	73.9
$\tilde{\sigma}^2$	185.1	176.9	15.1	185.1	28.9	26.2	9.3	36.1
$\hat{\sigma}^2$	172.0	144.2	15.0	184.7	26.8	26.9	9.3	36.1

Tables 1–8 give MISE and SE of ISE for the estimators in question. For better readability all actual values in the tables are multiplied, as in Yu and

Jones (2004), by 10^4. The tables are accompanied by Figures 1–8, where the respective regressions and standard deviations are plotted. As mentioned earlier, the first two cases presented in Tables 1 and 2 were studied in Yu and Jones (2004). An inspection of their Tables 2 and 3 shows that both $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ considerably outperform the seven variance function estimators considered there. They include, among others, besides the local likelihood estimator proposed by Yu and Jones, local linear smoothers based on squared residuals from various regression fits. For example, in the case of the first model for $n = 100$ the smallest MISE reported in the paper equals 139 with SE = 76.7 (after the above mentioned renormalization of SE). Also, for the Yu and Jones (2004) examples, it is seen that the post-model-selection estimators $\tilde{\sigma}^2$ and $\hat{\sigma}^2$ nicely compare with the Fan and Yao estimator $\bar{\sigma}^2$. For other cases we considered, a comparison of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ with $\bar{\sigma}^2$ is also encouraging.

The proposed estimators outperform the estimator $\bar{\sigma}^2$ both in MISE and SE in the cases when the variance is specified correctly (the last three examples), even if, as in Example 6, the regression function does not belong to the models considered. In such cases, when the sample size increases from $n = 100$ to $n = 500$, the decrease of MISE is more substantial for the post-model-selection estimators than for $\bar{\sigma}^2$. In Examples 4 and 5 the estimators $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ perform on a par with $\bar{\sigma}^2$ with respect to MISE whereas their SEs are significantly smaller. Only in Example 3 for $n = 500$, does $\bar{\sigma}^2$ noticeably outperform the proposed estimators in terms of MISE. Note that in Example 7, MISE of $\hat{\sigma}^2$ is more than 3.5 times smaller than that of $\bar{\sigma}^2$ for $n = 100$ and more than six times smaller for $n = 500$. It seems that the Fan and Yao estimator does not cope well with a situation when both the regression and the variance function are discontinuous at the same point. In general SEs for the post-model-selection estimators are smaller than SEs of $\bar{\sigma}^2$, indicating that the first two estimators behave more stably. In all examples considered $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ behave similarly with slightly smaller SEs for $\hat{\sigma}^2$ when $n = 100$. As a byproduct we obtained related goodness-of-fit measures for estimators of the regression function. Their examination reveals that with an exception of Example 7 the method of Fan and Yao (1998) is more precise than the post-model-selection estimation.

III. Proofs of the main results

A. Proof of Theorem 1. Consider first the case when $P = P_{\theta^0}$ and $P_{\theta^0} \in \mathcal{P}_{kl}$. Let $\theta^1 \in \Theta_{kl}$ be the unique point in Θ_{kl} such that $P_{\theta^0} = P_{\theta^1}$ and note that it satisfies (C5). Define the following statistics:

$$\mathcal{Q}^0 = \sup_{\theta \in \Theta_{k^0 l^0}} \mathcal{L}_n^{k^0 l^0}(\theta) - \mathcal{L}_n^{k^0 l^0}(\theta^0), \qquad \mathcal{Q}^1 = \sup_{\theta \in \Theta_{kl}} \mathcal{L}_n^{kl}(\theta) - \mathcal{L}_n^{kl}(\theta^1).$$

As $\mathcal{L}_n^{k^0l^0}(\theta^0) = \mathcal{L}_n^{kl}(\theta^1)$ we have

$$P(\mathcal{S}_0 > \mathcal{S}_1) = P(\mathcal{Q}^1 - \mathcal{Q}^0 < (\omega_1 - \omega_0)\log n).$$

Properties (C6)–(C9) imply that $\mathcal{P}_{k^0l^0}$ and $\mathcal{P}_{kl}$ satisfy the assumptions of Theorem 5.2.2 in Sen and Singer (1993). Note that although that theorem is stated for univariate random variables its obvious extension holds for the bivariate case. Thus, this result and (5.6.4) there imply that if $\hat{\theta}_{kl} = \arg\max_{\theta\in\Theta_{kl}} \mathcal{L}_n^{kl}(\theta)$ then $n^{1/2}(\hat{\theta}_{kl} - \theta^1) = O_{P_{\theta^1}}(1)$ and

$$\mathcal{Q}^1 = -n(\hat{\theta}_{kl} - \theta^1)^T I(\theta^1)(\hat{\theta}_{kl} - \theta^1) + o_{P_{\theta^1}}(1).$$

Analogous relations hold for $\mathcal{P}_{k^0l^0}$ when the data is generated from P_{θ^0}. Thus, as $P_{\theta^0} = P_{\theta^1}$, we get $\mathcal{Q}^1 - \mathcal{Q}^0 = O_{P_{\theta^0}}(1)$, from which (15) follows in this case since $\omega_1 > \omega_0$.

Consider now the case when $P = P_{\theta^0} \notin \mathcal{P}_{kl}$. For a fixed (u, v) and $\theta \in \Theta_{uv}$ let

$$\mathcal{R}_n^{uv}(\theta) = n^{-1}\sum_{i=1}^{n} \log p_\theta(X_i, Y_i).$$

Then we have

$$\mathcal{S}_0 - \mathcal{S}_1 = n[\sup_{\theta\in\Theta_{k^0l^0}} \mathcal{R}_n^{k^0l^0}(\theta) - \sup_{\theta\in\Theta_{kl}} \mathcal{R}_n^{kl}(\theta)] - (\omega_0 - \omega_1)\log n.$$

It follows from Theorem 2 in Jennrich (1969) that P_{θ^0}-a.e., as $n \to \infty$, $\mathcal{R}_n^{kl}(\theta) \to \mathbb{E}_{P_{\theta^0}} \log p_\theta(X, Y)$ uniformly in $\theta \in \Theta_{kl}$. For a fixed $\delta > 0$ set

$$A_i = \{\omega : \forall n \geq i,\, \forall \theta \in \Theta_{kl}\ \mathcal{R}_n^{kl}(\theta) \leq \lim_{m\to\infty} \mathcal{R}_m^{kl}(\theta) + \delta\}.$$

By uniform convergence we have $P_{\theta^0}(\bigcup_{i=1}^{\infty} A_i) = 1$. Since $\{A_i\}$ is increasing, we can choose N_1 such that $P_{\theta^0}(A_{N_1}) \geq 1 - \delta$. Let θ_{kl}^* be a pseudo-true value of the parameter defined above Theorem 2. For any $n \geq N_1$ and any $\theta \in \Theta_{kl}$, due to the definition of θ_{kl}^*, we have on A_{N_1}

$$\begin{aligned}(17)\qquad \mathcal{R}_n^{kl}(\theta) &\leq \lim_{m\to\infty} \mathcal{R}_m^{kl}(\theta) + \delta = \mathbb{E}_{P_{\theta^0}} \log p_\theta(X, Y) + \delta \\ &\leq \sup_{\theta\in\Theta_{kl}} \mathbb{E}_{P_{\theta^0}} \log p_\theta(X, Y) + \delta \\ &= \mathbb{E}_{P_{\theta^0}} \log p_{\theta_{kl}^*}(X, Y) + \delta.\end{aligned}$$

Consider now $\theta \in \Theta_{k^0l^0}$. Using again Jennrich's result we get

$$\mathcal{R}_n^{k^0l^0}(\theta) \geq \lim_{m\to\infty} \mathcal{R}_m^{k^0l^0}(\theta) - \delta = \mathbb{E}_{P_{\theta^0}} \log p_\theta(X, Y) - \delta$$

for any $n \geq N_2$ and all $\theta \in \Theta_{k^0l^0}$ on a set B_{N_2} of P_{θ^0}-probability greater than $1 - \delta$. By (C4), taking the supremum on both sides, we have on B_{N_2}, and for $n \geq N_2$,

$$\sup_{\theta\in\Theta_{k^0l^0}} \mathcal{R}_n^{k^0l^0}(\theta) \geq E_{P_{\theta^0}} \log p_{\theta^0}(X, Y) - \delta.$$

Thus for $n \geq \max(N_1, N_2)$, on the set $A_{N_1} \cap B_{N_2}$ having P_{θ^0}-probability at least $1 - 2\delta$,

$$\mathcal{S}_0 - \mathcal{S}_1 \geq n[D(P_{\theta^0} \| P_{\theta^*_{kl}}) - 2\delta] - (\omega_0 - \omega_1) \log n.$$

Hence for $2\delta < D(P_{\theta^0} \| P_{\theta^*_{kl}})$ we have, for n sufficiently large,

$$P(\mathcal{S}_0 > \mathcal{S}_1) = P_{\theta^0}(\mathcal{S}_0 > \mathcal{S}_1) \geq P_{\theta^0}(A_{N_1} \cap B_{N_2}) \geq 1 - 2\delta,$$

from which the conclusion follows as δ is an arbitrary positive number. ∎

REMARK 6. Observe that using again Jennrich's result on Θ_{kl} we obtain

$$\mathcal{R}^{kl}_n(\theta) \geq \lim_{m \to \infty} \mathcal{R}^{kl}_m(\theta) - \delta = \mathbb{E}_{P_{\theta^0}} \log p_\theta(X, Y) - \delta \tag{18}$$

for any $n \geq N$ and all $\theta \in \Theta_{kl}$ on the set of P_{θ^0}-probability greater than $1 - \delta$. Taking the supremum on both sides of (18) we have on this set, and for $n \geq N$,

$$\sup_{\theta \in \Theta_{kl}} \mathcal{R}^{kl}_n(\theta) \geq \mathbb{E}_{P_{\theta^0}} \log p_{\theta^*_{kl}}(X, Y) - \delta. \tag{19}$$

B. Proof of Corollary 2. By Corollary 1, it is enough to restrict attention to the case when $P = P_{\theta^0}$, $\theta^0 \in \Theta_{k^\diamond l^\diamond}$. By (C4), Theorem 2.2 of White (1982) implies that the ML estimator $\hat{e}_{k^\diamond l^\diamond}$ of $e_{k^\diamond l^\diamond}$ is consistent. From this and the boundedness of Legendre polynomials the conclusion follows. ∎

C. Proof of Theorem 2. The proof is similar to that of Theorem 1. Observe that the assumption implies that

$$\int p(x, y) \log(p(x, y)/p_{\theta^*_{k^* l^*}}(x, y))\, dx\, dy < \int p(x, y) \log(p(x, y)/p_{\theta^*_{kl}}(x, y))\, dx\, dy,$$

which is equivalent to

$$\int p(x, y) \log p_{\theta^*_{k^* l^*}}(x, y)\, dx\, dy > \int p(x, y) \log p_{\theta^*_{kl}}(x, y)\, dx\, dy. \tag{20}$$

Analogues of inequalities (19) and (17) imply that for any $\delta > 0$ there exists a set D of P-probability at least $1 - \delta$ and N such that for $n \geq N$ and $\omega \in D$,

$$\mathcal{R}^{k^* l^*}_n(\hat{\theta}_*) \geq \mathbb{E}_P \log p_{\theta^*_{k^* l^*}}(X, Y) - \delta$$

and

$$\mathcal{R}^{kl}_n(\hat{\theta}) \leq \mathbb{E}_P \log p_{\theta^*_{kl}}(X, Y) + \delta,$$

where $\hat{\theta}_*$ and $\hat{\theta}$ denote the ML estimators in $\Theta_{k^* l^*}$ and Θ_{kl}, respectively. Hence, with k^*, l^* replacing k^0, l^0 in $\mathcal{S}_0$ and ω_0,

$$\mathcal{S}_0 - \mathcal{S}_1 \geq n\Big[\int p(x, y) \log(p_{\theta^*_{k^* l^*}}(x, y)/p_{\theta^*_{kl}}(x, y))\, dx\, dy - 2\delta\Big] - (\omega_0 - \omega_1) \log n.$$

In view of (20), for sufficiently small δ,

$$\int p(x, y) \log(p_{\theta^*_{k^* l^*}}(x, y)/p_{\theta^*_{kl}}(x, y))\, dx\, dy - 2\delta > 0.$$

Since the closest parametric model is unique, the conclusion follows. ∎

Appendix I. Verification of analytical properties (C6)–(C9). Checking (C6) is routine and relies on the observations that a finite family of Legendre polynomials is uniformly bounded and it follows from (5) that for some positive C,

$$\min_{l\le L}\inf_{\theta\in\Theta_{kl}}\inf_{x\in[a,b]}\sigma^2_{tl}(x)\ge C>0.$$

For illustration, we check (C7) for the case of the second derivative of $\log p_\theta$ with respect to η_{vj} and $\eta_{v'j'}$. Indeed, it is easy to see that

$$\frac{\partial^2\log p_\theta(x,y)}{\partial\eta_{rj}\partial\eta_{r'j'}}=-\frac{1}{2}[\Phi_{rj}(x)\Phi_{r'j'}(x)][y-m_{sk}(x)]^2\frac{1}{\sigma^2_{tl}(x)}, \tag{21}$$

from which, reasoning as above, (C7) easily follows.

In order to check (C8) first note that, due to (C7), by the standard argument (cf. e.g. Sen and Singer (1993), p. 206), the formula for $I(\theta)$ holds. Moreover, since the integration and differentiation can be interchanged we get

$$-\frac{\partial^2}{\partial\theta\partial\theta^T}\{E_{P_\theta}[\log p_\theta(X,Y)]\}=I(\theta).$$

Therefore, if θ_0, defined in (C4), is an interior point of $\Theta_{k^0l^0}=\Theta^B_{k^0l^0}$, then $I(\theta^0)$ is positive definite. However, for $\theta^0\notin\operatorname{int}\Theta^B_{k^0l^0}$ we obviously have $\theta^0\in\operatorname{int}\Theta^{B^0}_{k^0l^0}$ for some $B^0>B$. Note also that, irrespective of the value of B, the properties (C2) and (C4) hold and thus $I(\theta^0)$ has to be positive definite. The same argument applies to $I(\theta^1)$, where θ^1 is the point defined in (C5).

Now we verify (C9). To ease notation consider the case $k=K$ and $l=L$ for which we abbreviate θ to $\theta=(b,e)$. Moreover, let $\theta+h=(b+h_1,e+h_2)$ and let $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ be vectors of Legendre polynomials pertaining to b and e, respectively. Then in view of (21),

$$\begin{aligned}\frac{\partial^2}{\partial\eta_{rj}\partial\eta_{r'j'}}\log\frac{p_{\theta+h}}{p_\theta}(X,Y)=&-\frac{1}{2}\frac{[\Phi_{rj}(X)\Phi_{r'j'}(X)][Y-(b+h_1)\circ\Phi_1(X)]^2}{\exp\{(e+h_2)\circ\Phi_2(X)\}}\\&+\frac{1}{2}\frac{[\Phi_{rj}(X)\Phi_{r'j'}(X)][Y-b\circ\Phi_1(X)]^2}{\exp\{e\circ\Phi_2(X)\}},\end{aligned}$$

where $\circ$ stands for the inner product in the relevant spaces. Set

$$\varepsilon=(Y-b\circ\Phi_1(X))/\exp(e\circ\Phi_2(X)/2).$$

The difference can be written as a sum of three terms

$$\begin{aligned}\frac{1}{2}[\Phi_{rj}(X)\Phi_{r'j'}(X)]\Bigg\{\Bigg(\varepsilon^2-\frac{\varepsilon^2}{\exp\{h_2\circ\Phi_2(X)\}}\Bigg)&-\frac{[h_1\circ\Phi_1(x)]^2}{\exp\{(e+h_2)\circ\Phi_2(X)\}}\\&+\frac{2\varepsilon h_1\circ\Phi_1(X)}{\exp\{(e/2+h_2)\circ\Phi_2(X)\}}\Bigg\}=:J_1+J_2+J_3.\end{aligned}$$

If $(X, Y) \sim P_\theta = P_{(b,e)}$ then $\varepsilon \sim N(0,1)$. In order to deal with J_1 observe that for h such that $\|h\| \le \delta$, with C and C' denoting generic constants and δ sufficiently small,

$$\sup_{\|h\|\le\delta} |J_1| \le C\varepsilon^2 \sup_{\|h\|\le\delta} |1 - \exp(-h_2 \circ \Phi_2(X))| \le C' \sup_{\|h\|\le\delta} \varepsilon^2 \sum_{i=1}^{L} |h_{2i}|$$
$$\le C'\varepsilon^2 L^{1/2}\delta^{1/2},$$

where $h_2 = (h_{21}, \dots, h_{2L})^T$ and the expected value of the above expression tends to 0 as $\delta \to 0$. The terms J_2 and J_3 and the second order derivatives with respect to b, as well as b and e, are dealt with similarly.

Appendix II. Some details on the simulation experiment

A. Description of examples. Each example is defined by a triple (r_i, sd_i, f_i), where r_i denotes the regression function, sd_i is the standard deviation while f_i stands for the density of the explanatory variable X. Below, we give a description of the regression functions r_1–r_8, the corresponding standard deviations sd_1–sd_8 and provide some information on the distribution of X.

EXAMPLE 1. $r_1(x) = x + 2\exp\{-16x^2\}$, $sd_1(x) = (0.5)\exp\{-x^2/8\}$, $x \in [-2,2]$ while $X \sim N(0,1)$, truncated to $[-2,2]$.

EXAMPLE 2. $r_2(x) = 2\cos x + \exp\{-x^2\}$, $sd_2(x) = (0.5)[2 + \sin x + \cos 2x]^{1/2}$, $x \in [-1,1]$ while $X \sim N(0, [0.5]^2)$, truncated to $[-1,1]$.

In Examples 4–6 we took $x \in [-2,2]$ and $X \sim N(0,1)$, truncated to $[-2,2]$. In Examples 3 and 7 we assumed $X \sim U[-2,2]$, while in Example 8 we considered $X \sim U[-1,1]$, where $U[a,b]$ denotes the uniform distribution on $[a,b]$.

The functions r_3, r_4 and r_6 are motivated by the list of Marron and Wand (1992) who considered them in the context of density estimation. They are expressed below as rescaled normal mixtures.

$$r_3 = 5\{\tfrac{3}{4}N(0,1) + \tfrac{1}{4}N(\tfrac{3}{2}, [\tfrac{1}{3}]^2)\},$$
$$r_4 = 5\{\tfrac{1}{2}N(-1, [\tfrac{2}{3}]^2) + \tfrac{1}{2}N(1, [\tfrac{2}{3}]^2)\},$$
$$r_6 = 5\{\tfrac{1}{5}N(0,1) + \tfrac{1}{5}N(\tfrac{1}{2}, [\tfrac{2}{3}]^2) + \tfrac{3}{5}N(\tfrac{13}{12}, [\tfrac{5}{9}]^2)\}.$$

Moreover,

$$sd_3(x) = 0.35 + (0.15)\{(0.3)\cos[\pi x] + (0.4)\sin[3(x - 0.175)\pi]\},$$
$$sd_4(x) = 0.2 + (0.05)x^2, \quad sd_6(x) = (0.4)\mathbf{1}_{[-2,2]}(x).$$

The remaining functions are as follows:

$$\begin{aligned} r_5 &= sd_6, \\ sd_5 &= 0.4 + (0.5)N(-1.5, [0.5]^2) + (0.2)N(1.5, [0.5]^2), \\ r_7(x) &= (0.5)\{x^2 \mathbf{1}_{[-2,0)}(x) + (2-x)^2 \mathbf{1}_{[0,2]}(x)\}, \\ sd_7(x) &= (0.4)\mathbf{1}_{[-2,0]}(x) + (0.8)\mathbf{1}_{(0,2]}(x), \\ r_8(x) &= 2x^2 \mathbf{1}_{[-1,1]}(x), \\ sd_8(x) &= (0.6)\mathbf{1}_{[-1,-1/2]}(x) + (0.3)\mathbf{1}_{(-1/2,0]}(x) \\ &\quad + (0.5)\mathbf{1}_{(0,1/2]}(x) + (0.8)\mathbf{1}_{(1/2,1]}(x). \end{aligned}$$

B. Goodness-of-fit measures. As a primary measure of performance of a variance estimator we considered its (empirical) Integrated Squared Error (ISE). We shall present this and the following notions for the case of $\hat{\sigma}^2$. Measures of goodness-of-fit for $\tilde{\sigma}^2$ and $\bar{\sigma}^2$ are defined in the same way. We have

$$\mathrm{ISE} = n^{-1} \sum_{i=1}^{n} [\hat{\sigma}^2(X_i) - \sigma^2(X_i)]^2.$$

Let ISE_j stand for the ISE of the jth repetition of the experiment, $j = 1, \ldots, M$. Then MISE and SE are defined as the empirical mean and the empirical standard deviation of $\{\mathrm{ISE}_j\}_{j=1}^{M}$. In all experiments we took $M = 2000$. We also calculated the Mean Integrated Variance (MIV) and Integrated Squared Bias (ISB), which are empirical means (over M runs) of the quantities

$$n^{-1} \sum_{i=1}^{n} [\hat{\sigma}^2(X_i) - \tilde{E}\hat{\sigma}^2(X_i)]^2 \quad \text{and} \quad n^{-1} \sum_{i=1}^{n} [\tilde{E}\hat{\sigma}^2(X_i) - \sigma^2(X_i)]^2, \tag{22}$$

respectively. In (22), $\tilde{E}\hat{\sigma}^2(X_i)$ denotes the estimator of the expected value of $\hat{\sigma}^2$ calculated at X_i, based on all M simulation results. Since the values of the explanatory variable differ in subsequent samples, the equality MISE = ISB + MIV holds only approximately.

C. Implementation issues. As discussed previously (cf. (13)), the model $\mathcal{P}_{kl}$ is the heteroscedastic regression model for which the regression function has the form $\mathcal{X}b_{sk}$ and the logarithm of the variance has the form $\mathcal{Z}e_{tl}$ for suitable $\mathcal{X}$ and $\mathcal{Z}$ of the dimensions $n \times (s \cdot k)$ and $n \times (t \cdot l)$, respectively. Every column of $\mathcal{X}$ contains values of the corresponding Legendre polynomial supported on a certain partition interval and calculated at sample points. Thus nonzero values in the column correspond to the sample points falling into the pertaining interval. In order for the matrix $\mathcal{X}$ (resp., $\mathcal{Z}$) to be of full column rank, at least s points (resp., t) have to fall into each partition interval. Recall that $s = 3$ and $t = 1$ in our experiments.

Both ML and REML estimators are found iteratively as zeros of the derivative of either the log-likelihood or restricted log-likelihood function. For ML estimation the maximal number of iterations (I_{max}) was set at $I_{max} = 100$ with a stop occurring when the absolute difference of the likelihood in two successive iterations did not exceed 10^{-6}. For REML estimation we used the analogous stopping criterion based on the restricted log-likelihood and $I_{max} = 200$. For each of the methods, the number of cases for which convergence failed or either $\mathcal{X}$ or $\mathcal{Z}$ was not of full column rank was less than 0.3% for each model considered. The exact form of the score vectors and information matrices for both methods are given in Verbyla (1993) and Smyth (2002).

Acknowledgments. The calculations reported in this paper were done at the Institute of Computer Science of the Polish Academy of Sciences. We specially thank Janusz Ćwik for his assistance with our numerical experiments. We are grateful to Przemysław Biecek for useful remarks and to Gwénaëlle Castellan and Yves Rozenholc for their preprints and to Andrzej Kozek for providing a copy of Verbyla (1990).

References

L. Birgé and Y. Rozenholc (2006), *How many bins should be put in a regular histogram*, ESAIM Probab. Statist. 10, 24–45.

P. Borkowski and J. Mielniczuk (2010), *Post-model-selection estimators of variance function for non-linear autoregression*, J. Time Series Anal. 31, 50–63.

T. T. Cai and L. Wang (2008), *Adaptive variance function estimation in heteroscedastic nonparametric regression*, Ann. Statist. 36, 2025–2054.

R. J. Carroll (1982), *Adapting to heteroscedasticity in linear models*, Ann. Statist. 10, 1224–1233.

G. Castellan (2003), *Density estimation via exponential model selection*, IEEE Trans. Inform. Theory 49, 2052–2060.

N. Cressie and S. N. Lahiri (1993), *The asymptotic distribution of REML estimators*, J. Multivariate Anal. 45, 217–233.

M. Davidian and R. J. Carroll (1987), *Variance function estimation*, J. Amer. Statist. Assoc. 82, 1079–1091.

H. Dette, A. Munk and T. Wagner (1998), *Estimating the variance in nonparametric regression—what is a reasonable choice?*, J. Roy. Statist. Soc. B 60, 751–764.

V. M. Dukič and E. A. Peña (2005), *Variance estimation in a model with Gaussian submodels*, J. Amer. Statist. Assoc. 100, 296–309.

J. Fan and Q. Yao (1998), *Efficient estimation of conditional variance functions in stochastic regression*, Biometrika 85, 645–660.

J. Fan and Q. Yao (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York.

A. C. Harvey, *Estimating regression models with multiplicative heteroscedasticity*, Econometrica 44, 461–465.

D. A. Harville (1974), *Bayesian inference for variance components using only error contrasts*, Biometrika 61, 383–385.

R. J. Jennrich (1969), *Asymptotic properties of non-linear least squares estimators*, Ann. Math. Statist. 40, 633–643.

J. Jiang (1996), *REML estimation: Asymptotic behaviour and related topics*, Ann. Statist. 24, 255–286.

R. Kohn (1983), *Consistent estimation of minimal subset dimension*, Econometrica 51, 367–376.

T. C. M. Lee (2001), *An introduction to coding theory and the two-part minimum description length principle*, Int. Statist. Rev. 69, 169–183.

H. Leeb and B. M. Pötscher (2008), *Model selection* in: Handbook of Financial Time Series, T. Andersen et al. (eds), Springer, Berlin.

J. S. Marron and M. P. Wand (1992), *Exact mean integrated squared error*, Ann. Statist. 20, 712–736.

H.-G. Müller and U. Stadtmüller (1993), *On variance function estimation with quadratic forms*, J. Statist. Plann. Inference 35, 213–231.

M. H. Neumann (1994), *Fully data-driven nonparametric variance estimators*, Statistics 25, 189-212.

B. M. Pötscher (1991), *Effects of model selection on inference*, Econometric Theory 7, 163–185.

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

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

T. Sawa (1978), *Information criteria for discriminating among alternative regression models*, Econometrica 46, 1273–1291.

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

P. K. Sen and J. M. Singer (1993), *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York.

B. W. Silverman (1985), *Some aspects of the spline smoothing approach to non-parametric regression curve fitting*, J. Roy. Statist. Soc. B 47, 1–52.

C.-Y. Sin and H. White (1996), *Information criteria for selecting possibly misspecified parametric models*, J. Econometrics 71, 207–225.

G. K. Smyth (2002), *An efficient algorithm for REML in heteroscedastic regression*, J. Comput. Graph. Statist. 11, 836–847.

A. P. Verbyla (1990), *A conditional derivation of residual maximum likelihood*, Austral. J. Statist. 32, 227-230.

A. P. Verbyla (1993), *Modelling variance heterogeneity: residual maximum likelihood and diagnostics* J. Roy. Statist. Soc. B 55, 493–508.

H. White (1982), *Maximum likelihood estimation of misspecified models*, Econometrica 50, 1–25.

P. Yau and R. Kohn (2003), *Estimation and variable selection in nonparametric heteroscedastic regression*, Statistics Comput. 13, 191–208.

K. Yu and M. C. Jones (2004), *Likelihood-based local linear estimation of conditional variance function*, J. Amer. Statist. Assoc. 99, 139–144.

M. Yuan and G. Wahba (2004), *Doubly penalized likelihood estimator in heteroscedastic regression*, Statist. Probab. Lett. 69, 11–20.

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

Jan Mielniczuk
Institute of Computer Science
Ordona 21
01-237 Warszawa, Poland
and
Warsaw University of Technology
Plac Politechniki 1
00-661 Warszawa, Poland
E-mail: miel@ipipan.waw.pl

Received on 13.5.2010 (2046)