Statistica Sinica **26** (2016), 925-954
doi:http://dx.doi.org/10.5705/ss.202014.0128

HETEROSCEDASTIC SEMIPARAMETRIC TRANSFORMATION MODELS: ESTIMATION AND TESTING FOR VALIDITY

Natalie Neumeyer, Hohsuk Noh and Ingrid Van Keilegom

University of Hamburg, Sookmyung Women's University and Université catholique de Louvain

Abstract: In this paper we consider a heteroscedastic transformation model of the form $\Lambda_\vartheta(Y) = m(X)+\sigma(X)\varepsilon$, where Λ_ϑ belongs to a parametric family of monotone transformations, $m(\cdot)$ and $\sigma(\cdot)$ are unknown but smooth functions, ε is independent of the d-dimensional vector of covariates X, $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = 1$. We consider the estimation of the unknown components of the model, ϑ, $m(\cdot)$, $\sigma(\cdot)$, and the distribution of ε, and we show the asymptotic normality of the proposed estimators. We propose tests for the validity of the model, and establish the limiting distribution of the test statistics under the null hypothesis. A bootstrap procedure is proposed to approximate the critical values of the tests. We carried out a simulation study to verify the small sample behavior of the proposed estimators and tests, and illustrate our method with a dataset.

Key words and phrases: Bootstrap, empirical distribution function, empirical independence process, local polynomial estimator, location-scale model, model specification, nonparametric regression, profile likelihood estimator.

1. Introduction

Assume we observe independent copies of a random vector (X, Y), where X represents a d-dimensional covariate and Y is a univariate response. One possibility is to analyze these data by fitting a non- or semiparametric regression model

$$Y = m(X) + \varepsilon, \text{ where } E[\varepsilon \mid X] = 0. \tag{1.1}$$

Doing so, often the conditional error distribution, given the covariate, still depends on X, so the dependency of the response Y on the covariate X goes beyond the first moment. If only the second moment is dependent on X one can fit a nonparametric location-scale model of the form

$$Y = m(X) + \sigma(X)\varepsilon, \text{ where } \varepsilon \perp X \text{ with } E[\varepsilon] = 0, \text{Var}(\varepsilon) = 1. \tag{1.2}$$

Here and throughout the paper $Z \perp X$ means that Z and X are stochastically independent. Such nonparametric location-scale models have been widely used, see e.g., Akritas and Van Keilegom (2001), Dette, von Lieres und Wilkau, and Sperlich (2005) or Hušková and Meintanis (2010), among many others. Note that the conditional normal distribution is always a special case because from $Y|X = x \sim N(m(x), \sigma^2(x))$ it follows that $\varepsilon \sim N(0, 1)$ does not depend on X. The general location-scale model (1.2) has several advantages over the unstructured model (1.1). First, the asymptotic analysis of statistical procedures often simplifies a lot. Further, the model allows us to estimate the error distribution with a parametric $\sqrt{n}$-rate, see Akritas and Van Keilegom (2001). Therefore the estimation of the conditional distribution of Y given X is much more efficient. Goodness-of-fit as well as other specification tests have been developed that specifically use the location-scale structure, see Section 2.4 in the recent review by González-Manteiga and Crujeiras (2013). When data (X, Y_1, Y_2) have been observed and one's interest lies in the dependence between Y_1 and Y_2, given X, under the location-scale structure the conditional copula of (Y_1, Y_2), given X, can not only be estimated with $\sqrt{n}$-rate, but also as precisely as if the errors were known, see Gijbels, Omelka, and Veraverbeke (2015).

The construction of valid resampling procedures is essential for most hypothesis tests in nonparametric regression. It is known that in heteroscedastic regression models simple residual bootstrap methods generally do not lead to valid procedures. Thus mostly the wild bootstrap is used, see Härdle and Mammen (1993) and Stute, González-Manteiga, and Presedo Quindimil (1998). However, Zhu, Fujikoshi, and Naito (2001) show that the wild bootstrap may fail if the conditional 4th moment of the error distribution depends on the covariate, while for the procedure considered there it works in the location-scale context. There are other cases where the wild bootstrap even fails in the location-scale model (1.2), see e.g., Neumeyer and Sperlich (2006). A (smooth or not smooth) heteroscedastic residual bootstrap often can be an alternative, see Neumeyer (2009a), and explicitly makes use of the location-scale structure.

Before application of model (1.2) a specification test should be conducted, a test for independence of ε and X. Such tests have been suggested by Einmahl and Van Keilegom (2008), Neumeyer (2009b), and Hlávka, Hušková, and Meintanis (2011). However, if those tests reject the null hypothesis, a remedy might be to transform the response Y by a suitable transformation Λ before fitting the location-scale model to the data (X, Y).

It is common in practice to transform the response variable before fitting a regression model to the data. The aim of the transformation is to reduce skewness or heteroscedasticity, or to induce normality. Often the transformation is chosen from a parametric class such as the Box-Cox power transformations introduced by Box and Cox (1964). Generalizations of this class were suggested by Bickel and Doksum (1981) and Yeo and Johnson (2000), among others. The parameter of the transformation in the class can be chosen data dependently by a profile likelihood approach, for instance. There is a huge literature on parametric transformation models and we refer to the monograph by Carroll and Ruppert (1988); see also the references in Fan and Fine (2013). Nonparametric estimation of the transformation in the context of parametric regression models has been considered by Horowitz (1996) and Zhou, Lin, and Johnson (2009), among others. Horowitz (2009) reviews estimation in transformation models with parametric regression in the cases where either the transformation or the error distribution, or both, are modeled nonparametrically. Linton, Sperlich, and Van Keilegom (2008) consider a parametric class of transformations, with the error distribution estimated nonparametrically and the regression function assumed to be additive. The aim of the transformation is to induce independence of the covariate and the error. Asymptotic normality of a profile likelihood estimator for the transformation parameter is proved. Heuchenne, Samb, and Van Keilegom (2015) consider a residual based empirical distribution function in the same model in order to estimate the error distribution. Recently, Colling and Van Keilegom (2015) considered goodness-of-fit tests for the regression function in a semiparametric transformation model, in which the transformation parameter is estimated by means of the profile likelihood estimator of Linton, Sperlich, and Van Keilegom (2008).

The aim of our paper is twofold. We generalize the results of Linton, Sperlich, and Van Keilegom (2008) by allowing heteroscedasticity. To this end in a parametric class of transformations we seek the one that leads to a nonparametric location-scale model of the form

$$\Lambda(Y) = m(X) + \sigma(X)\varepsilon, \text{ where } \varepsilon \perp X \text{ with } E[\varepsilon] = 0, \mathrm{Var}(\varepsilon) = 1, \tag{1.3}$$

where Λ denotes the transformation. The regression function m and variance function σ^2 are modeled fully nonparametrically, but analogous results can be obtained for semiparametric modeling. We estimate the transformation parameter by a profile-likelihood approach and prove asymptotic normality of the estimator. We investigate the performance of the estimator in a simulation study. Note that in the context of parametric regression, Zhou, Lin, and Johnson (2009) and Khan, Shin, and Tamer (2011) considered heteroscedastic transformation models.

We also propose, for the first time, a test for model validity in the context of transformation models with a parametric class of transformations and a non- (or semi-)parametric regression function. Mu and He (2007) consider estimation procedures in a transformation model with a linear quantile regression function and also suggest a test for model validity. In the general heteroscedastic case we suggest tests for the hypothesis of existence of some transformation Λ in the considered parametric class, such that the data fulfill model (1.3). The results can readily be modified to test whether such a model can hold with $\sigma \equiv 1$, i. e. a homoscedastic transformation model. Our test statistics are based on the difference between the estimated joint distribution of covariables and errors, and the product of the marginal distributions. A similar approach was used to test for validity of a location-scale model (without transformation) by Einmahl and Van Keilegom (2008). However, the estimation of the unknown transformation vastly complicates theoretical derivations. We show weak convergence of the estimated empirical process to a centered Gaussian process under the null hypothesis of model validity. As a by-product we obtain an expansion for the residual-based empirical distribution function that generalizes results by Heuchenne, Samb, and Van Keilegom (2015). Moreover, we discuss consistency of the proposed tests and demonstrate the finite sample properties of a bootstrap version of Kolmogorov-Smirnov and Cramér von Mises tests in a simulation study.

The rest of the paper is organized as follows. In Section 2 we define the profile likelihood estimator for the transformation parameter and show asymptotic normality. We further discuss estimation of the regression and variance function by local polynomial estimators, and the estimation of the error distribution. In Section 3 we consider the problem of testing for existence of a transformation in the considered class that leads to a location-scale model. We derive an expansion for the estimator of the joint distribution of covariates and errors. Under the null hypothesis we show weak convergence of the process given by the difference of the estimated joint distribution and the product of the marginals. Consistency of the testing procedures and modifications for the homoscedastic model are discussed. Additionally, we describe bootstrap versions of the hypothesis tests. In Section 4, we present simulations to demonstrate finite sample properties of the profile likelihood estimator for the transformation parameter as well as the hypothesis tests. We illustrate our method on a dataset. All regularity conditions and some of the proofs are collected in Appendices A and B. The other proofs are in a supplementary document.

2. Estimation of the Model

Let $L = \{\Lambda_\vartheta \mid \vartheta \in \Theta\}$ be a parametric class of differentiable and strictly increasing transformations, and let Θ be a nonempty subset of $\mathbb{R}^k$. In this section we assume that there exists some unique $\vartheta_0 \in \Theta$ such that

$$\frac{\Lambda_{\vartheta_0}(Y) - E[\Lambda_{\vartheta_0}(Y)|X]}{(\operatorname{Var}(\Lambda_{\vartheta_0}(Y)|X))^{1/2}} \perp X.$$

Then the covariate and transformed response can be modeled by a nonparametric location-scale model as

$$\Lambda_{\vartheta_0}(Y) = m(X) + \sigma(X)\varepsilon, \quad \varepsilon \perp X, \tag{2.1}$$

where $m(x) = E[\Lambda_{\vartheta_0}(Y)|X = x]$ and $\sigma^2(x) = \operatorname{Var}(\Lambda_{\vartheta_0}(Y)|X = x)$.

2.1. Estimation of the transformation parameter

To estimate the transformation parameter ϑ_0 we use a profile likelihood approach. This type of approach has been proposed by Linton, Sperlich, and Van Keilegom (2008) in the context of homoscedastic transformation models, and has been further used by Heuchenne, Samb, and Van Keilegom (2015) and Colling and Van Keilegom (2015) in the context of the estimation of the error distribution and the development of goodness-of-fit tests for the regression function, respectively. We extend their method to the current setup with heteroscedastic errors.

For $\vartheta \in \Theta$, let $m_\vartheta(x) = E[\Lambda_\vartheta(Y)|X = x]$, $\sigma^2_\vartheta(x) = \mathrm{Var}\,[\Lambda_\vartheta(Y)|X = x]$, and

$$\varepsilon(\vartheta) = \frac{\Lambda_\vartheta(Y) - m_\vartheta(X)}{\sigma_\vartheta(X)}.$$

Let $F_{\varepsilon(\vartheta)}(y) = P(\varepsilon(\vartheta) \le y)$ denote the marginal distribution function of the errors and let $f_{\varepsilon(\vartheta)}(y)$ be the corresponding probability density function. We use the abbreviated notations $\Lambda = \Lambda_{\vartheta_0}$, $\varepsilon = \varepsilon(\vartheta_0)$, $m = m_{\vartheta_0}$, $\sigma^2 = \sigma^2_{\vartheta_0}$, $F_\varepsilon = F_{\varepsilon(\vartheta_0)}$, and $f_\varepsilon = f_{\varepsilon(\vartheta_0)}$.

The conditional distribution $F_{Y|X}(\cdot|x)$ of Y given $X = x$ can then be written as

$$F_{Y|X}(y|x) = F_\varepsilon\Big(\frac{\Lambda(y) - m(x)}{\sigma(x)}\Big),$$

and hence the conditional density $f_{Y|X}(\cdot|x)$ of Y given $X = x$ is

$$f_{Y|X}(y|x) = f_\varepsilon\Big(\frac{\Lambda(y) - m(x)}{\sigma(x)}\Big)\frac{\Lambda'(y)}{\sigma(x)}.$$

Assume we have independent observations (X_i, Y_i), $i = 1, \ldots, n$, distributed as (X, Y), and let $\varepsilon_i = \varepsilon_i(\vartheta_0)$, $i = 1, \ldots, n$. Then, for an arbitrary value $\vartheta \in \Theta$, the log-likelihood can be written as

$$L_\vartheta = \sum_{i=1}^n \Big\{ \log f_{\varepsilon(\vartheta)}\Big(\frac{\Lambda_\vartheta(Y_i) - m_\vartheta(X_i)}{\sigma_\vartheta(X_i)}\Big) + \log \Lambda'_\vartheta(Y_i) - \log \sigma_\vartheta(X_i)\Big\}. \tag{2.2}$$

In order to maximize this log-likelihood with respect to ϑ, we first need to replace the unknown functions $f_{\varepsilon(\vartheta)}$, m_ϑ, and σ_ϑ by suitable estimators. For each $\vartheta \in \Theta$ we estimate $m_\vartheta(x)$ by a local polynomial estimator based on $(X_i, \Lambda_\vartheta(Y_i))$, $i = 1, \ldots, n$. To this end denote the components of X_i by $(X_{i1}, \ldots, X_{id})$ $(i = 1, \ldots, n)$ and let $x = (x_1, \ldots, x_d)$. Let $\hat{m}_\vartheta(x) = \hat{\beta}_0$, where $\hat{\beta}_0$ is the first component of the vector $\hat{\beta}$, that is the solution of the local minimization problem

$$\min_\beta \sum_{i=1}^n \Big\{\Lambda_\vartheta(Y_i) - P_i(\beta, x, p)\Big\}^2 K_h(X_i - x). \tag{2.3}$$

Here, $P_i(\beta, x, p)$ is a polynomial of order p built up with all $0 \leq k \leq p$ products of factors of the form $X_{ij} - x_j$ $(j = 1, \ldots, d)$. The vector β is the vector consisting of all coefficients of this polynomial. Here, for $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$, $K(u) = \prod_{j=1}^d k(u_j)$ is a d-dimensional product kernel, k is a univariate kernel function, $h = (h_1, \ldots, h_d)$ is a d-dimensional bandwidth vector converging to zero when n tends to infinity, and $K_h(u) = \prod_{j=1}^d k(u_j/h_j)/h_j$.

Analogously, for each $\vartheta \in \Theta$ let $\hat{s}_\vartheta$ denote a local polynomial estimator based on $(X_i, \Lambda_\vartheta(Y_i)^2)$, $i = 1, \ldots, n$, and define the variance function estimator as $\hat{\sigma}_\vartheta^2 = \hat{s}_\vartheta - \hat{m}_\vartheta^2$. This estimator has similar properties as a local polynomial estimator based on $(X_i, (\Lambda_\vartheta(Y_i) - \hat{m}_\vartheta(X_i))^2)$, $i = 1, \ldots, n$.

Let $\hat{\varepsilon}_i(\vartheta) = (\Lambda_\vartheta(Y_i) - \hat{m}_\vartheta(X_i))/\hat{\sigma}_\vartheta(X_i)$ and define

$$\hat{f}_{\hat{\varepsilon}(\vartheta)}(y) = \frac{1}{n} \sum_{i=1}^n \ell_g\big(\hat{\varepsilon}_i(\vartheta) - y\big),$$

where ℓ and g are a kernel function and a bandwidth sequence, possibly different from the kernel k and the bandwidth h that were used to estimate the regression and variance function.

We plug the estimators $\hat{m}_\vartheta$, $\hat{\sigma}_\vartheta$ and $\hat{f}_{\hat{\varepsilon}(\vartheta)}$ into the log-likelihood given in (2.2) and obtain a profile likelihood estimator of ϑ as

$$\hat{\vartheta} = \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n \Big\{ \log \hat{f}_{\hat{\varepsilon}(\vartheta)}\Big(\frac{\Lambda_\vartheta(Y_i) - \hat{m}_\vartheta(X_i)}{\hat{\sigma}_\vartheta(X_i)}\Big) + \log \Lambda'_\vartheta(Y_i) - \log \hat{\sigma}_\vartheta(X_i) \Big\}. \quad (2.4)$$

In order to state an asymptotic i.i.d. representation and the asymptotic normality of the estimator $\hat{\vartheta}$, we need some notation. For any function h_ϑ we denote by $\dot{h}_\vartheta = \nabla_\vartheta h_\vartheta$ the vector of partial derivatives of h_ϑ with respect to the components of ϑ. Let

$$G_n(\vartheta) = \frac{1}{n} \sum_{i=1}^n g_\vartheta(X_i, Y_i)$$

be the derivative of the log-likelihood given in (2.2) (divided by n) with respect to ϑ, where

$$g_\vartheta(X_i, Y_i) = \frac{f'_{\varepsilon(\vartheta)}(\varepsilon_i(\vartheta))}{f_{\varepsilon(\vartheta)}(\varepsilon_i(\vartheta))} \Big[\frac{\dot{\Lambda}_\vartheta(Y_i) - \dot{m}_\vartheta(X_i)}{\sigma_\vartheta(X_i)} - \{\Lambda_\vartheta(Y_i) - m_\vartheta(X_i)\} \frac{\dot{\sigma}_\vartheta(X_i)}{\sigma_\vartheta^2(X_i)} \Big] + \frac{\dot{f}_{\varepsilon(\vartheta)}(\varepsilon_i(\vartheta))}{f_{\varepsilon(\vartheta)}(\varepsilon_i(\vartheta))} + \frac{\dot{\Lambda}'_\vartheta(Y_i)}{\Lambda'_\vartheta(Y_i)} - \frac{\dot{\sigma}_\vartheta(X_i)}{\sigma_\vartheta(X_i)}.$$

Then $G_n(\vartheta)$ converges in probability to $G(\vartheta) = E[g_\vartheta(X, Y)]$. We assume that ϑ_0 is the unique zero of G (Assumption (a7) in Appendix A). Our theorem states the asymptotic normality of the estimator $\hat{\vartheta}$. The result shows that the variance of the estimator is the same as in the case where the nonparametric functions $m_\vartheta(x)$, $\sigma_\vartheta(x)$, and $f_{\varepsilon(\vartheta)}(y)$ and their derivatives with respect to ϑ and y are known, which is quite remarkable. The cancellation of all terms derived from estimators of nuisance functions has been observed in other contexts where profile likelihood methods have been used; see e.g., Severini and Wong (1992) among others, where the profile likelihood method internalizes the estimation cost associated with the nonparametric functions. The regularity conditions under which this result is valid are given in Appendix A.

Theorem 1. *Assume* (a1)$-$(a7) *in Appendix A. Then,*

$$\hat{\vartheta} - \vartheta_0 = -\Gamma^{-1}\frac{1}{n}\sum_{i=1}^{n} g_{\vartheta_0}(X_i, Y_i) + o_P(n^{-1/2}),$$
$$n^{1/2}\big(\hat{\vartheta} - \vartheta_0\big) \xrightarrow{d} N\big(0, \Sigma\big),$$

where $\Sigma = \Gamma^{-1}\mathrm{Var}[g_{\vartheta_0}(X,Y)]\Gamma^{-1}$ *and* $\Gamma = \nabla_\vartheta G(\vartheta)^\top|_{\vartheta=\vartheta_0}$.

The proof is in the supplementary document.

2.2. Estimation of regression and variance functions

Once the transformation parameter vector ϑ_0 is estimated, we take

$$\hat{m}(x) = \hat{m}_{\hat{\vartheta}}(x) \qquad \text{and} \qquad \hat{\sigma}^2(x) = \hat{\sigma}^2_{\hat{\vartheta}}(x).$$

Under regularity conditions the estimation of ϑ_0 has no influence on the asymptotic distribution of the centered and scaled estimators $(nh^d)^{1/2}(\hat{m}(x) - E[\hat{m}(x)])$ and $(nh^d)^{1/2}(\hat{\sigma}^2(x) - E[\hat{\sigma}^2(x)])$, since $\hat{\vartheta}$ has a parametric rate of convergence. Therefore, the estimators behave asymptotically as if the true ϑ_0 would be known. The pre-estimation of ϑ_0 does influence the asymptotic distribution of the test statistic because the integrals $\int(\hat{m}_{\vartheta_0} - m)/\sigma\, dF_X$ and $\int(\hat{m}_{\hat{\vartheta}} - \hat{m}_{\vartheta_0})/\sigma\, dF_X$ have the same $n^{1/2}$-rate of convergence (see terms B_n and C_n in the proof of Theorem 2) and a similar statement holds for the variance estimator (see Section 3).

2.3. Estimation of the error distribution

The last unknown component of our heteroscedastic transformation model (2.1) is the distribution F_ε of the error term. We define the residuals as

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i(\hat{\vartheta}) = \frac{\Lambda_{\hat{\vartheta}}(Y_i) - \hat{m}(X_i)}{\hat{\sigma}(X_i)},$$

and take

$$\hat{F}_{\hat{\varepsilon}}(y) = \frac{1}{n}\sum_{i=1}^{n} I\{\hat{\varepsilon}_i \le y\},$$

where I denotes the indicator function. The asymptotic properties of this estimator are studied in the next section, where we study an estimator of the joint distribution of X and ε that includes the estimator $\hat{F}_{\hat{\varepsilon}}(y)$ as a special case.

3. Testing the Validity of the Model

In this section we develop tests for validity of a heteroscedastic semiparametric transformation model. Let $L = \{\Lambda_\vartheta \mid \vartheta \in \Theta\}$ be some parametric class of transformations, Θ some nonempty subset of $\mathbb{R}^k$. Our aim is to test the null hypothesis

$$H_0 : \exists \vartheta \in \Theta \text{ such that } \frac{\Lambda_\vartheta(Y) - E[\Lambda_\vartheta(Y)|X]}{(\mathrm{Var}\,(\Lambda_\vartheta(Y)|X))^{1/2}} \perp X. \tag{3.1}$$

If the null hypothesis is valid then there exists some transformation $\Lambda_{\vartheta_0} \in L$ from which one obtains a nonparametric location-scale model as in (2.1). As we want to test the appropriateness of the parametric family of transformations, our test is a goodness-of-fit test for the chosen family.

3.1. The test statistics and asymptotic distributions under H_0

Let $\hat{\vartheta}$ be some estimator for the true parameter ϑ_0 under H_0 such that a linear expansion

$$\hat{\vartheta} - \vartheta_0 = \frac{1}{n} \sum_{i=1}^{n} g_{\vartheta_0}(X_i, Y_i) + o_P\left(\frac{1}{\sqrt{n}}\right) \tag{3.2}$$

is valid under H_0, where $E[g_{\vartheta_0}(X_i, Y_i)] = 0$, $E[\|g_{\vartheta_0}(X_i, Y_i)\|^2] < \infty$. By Theorem 1, such an expansion is valid for the profile likelihood estimator under some regularity conditions. The joint empirical distribution function of covariates and residuals is

$$\hat{F}_{X,\hat{\varepsilon}}(x, y) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \le x,\ \hat{\varepsilon}_i \le y\},$$

where $\le$ for vectors is meant componentwise. We consider test statistics based on the estimated independence empirical process

$$S_n = \sqrt{n}(\hat{F}_{X,\hat{\varepsilon}} - \hat{F}_X \hat{F}_{\hat{\varepsilon}}), \tag{3.3}$$

where $\hat{F}_X(x) = n^{-1} \sum_{i=1}^{n} I\{X_i \le x\}$ and $\hat{F}_{\hat{\varepsilon}}(y) = n^{-1} \sum_{i=1}^{n} I\{\hat{\varepsilon}_i \le y\}$.

Theorem 2. *Assume* (a1), (a2) *and* (A1)−(A8) *from Appendix A. Then, under* H_0,

$$\begin{aligned}\hat{F}_{X,\hat{\varepsilon}}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \Big(& I\{X_i \le x\}\Big(I\{\varepsilon_i \le y\} + f_\varepsilon(y)(\varepsilon_i + \frac{y}{2}(\varepsilon_i^2 - 1))\Big) \\ & + E\Big[\nabla_\vartheta F_{\varepsilon(\vartheta)|X}(y|X)|_{\vartheta=\vartheta_0} I\{X \le x\}\Big]^\top g_{\vartheta_0}(X_i, Y_i)\Big) + o_P(n^{-1/2})\end{aligned}$$

uniformly with respect to $x \in R_X, y \in \mathbb{R}$.

The proof is given in appendix B. From the theorem one directly obtains the following result for the residual based empirical distribution function defined in Section 2.3.

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

$$\hat{F}_{\hat{\varepsilon}}(y) = \frac{1}{n}\sum_{i=1}^{n}\Big(I\{\varepsilon_i \leq y\} + f_\varepsilon(y)(\varepsilon_i + \frac{y}{2}(\varepsilon_i^2 - 1)) + E\Big[\nabla_\vartheta F_{\varepsilon(\vartheta)|X}(y|X)|_{\vartheta=\vartheta_0}\Big]^\top g_{\vartheta_0}(X_i, Y_i)\Big) + o_P(n^{-1/2})$$

uniformly with respect to $y \in \mathbb{R}$*. The process* $\sqrt{n}(\hat{F}_{\hat{\varepsilon}} - F_\varepsilon)$ *converges weakly in* $\ell^\infty(\mathbb{R})$ *to a centered Gaussian process.*

This corollary generalizes results of Heuchenne, Samb, and Van Keilegom (2015) who consider estimation of the error distribution in a homoscedastic transformation model. The asymptotic expansion directly follows from Theorem 2. The proof of weak convergence is analogous to the proof of Corollary 1 below, and is omitted.

The dominating term in the expansion of $\hat{F}_{\hat{\varepsilon}}(y)$ has expectation $F_\varepsilon(y)$ and, with $\hat{F}_X = F_X + O_p(n^{-1/2})$, one straightforwardly obtains an expansion for the process S_n defined in (3.3):

$$S_n(x, y) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{x,y,\vartheta_0}(X_i, Y_i) + o_P(1) \tag{3.4}$$

uniformly with respect to $x \in R_X, y \in \mathbb{R}$, where

$$\psi_{x,y,\vartheta_0}(X_i, Y_i) = \Big(I\{X_i \leq x\} - F_X(x)\Big)\Big(I\{\varepsilon_i \leq y\} - F_\varepsilon(y) + f_\varepsilon(y)(\varepsilon_i + \frac{y}{2}(\varepsilon_i^2 - 1))\Big) + E\Big[\nabla_\vartheta F_{\varepsilon(\vartheta)|X}(y|X)|_{\vartheta=\vartheta_0}\Big(I\{X \leq x\} - F_X(x)\Big)\Big]^\top g_{\vartheta_0}(X_i, Y_i).$$

Corollary 2. *Under the assumptions of Theorem* 2*, the process* S_n *converges weakly in* $\ell^\infty(R_X \times \mathbb{R})$ *to a centered Gaussian process* S *with covariance* $\mathrm{Cov}\,(S(x, y), S(u, z)) = E[\psi_{x,y,\vartheta_0}(X, Y)\psi_{u,z,\vartheta_0}(X, Y)]$.

The proof is given in appendix B. Let Ψ denote some continuous functional from $\ell^\infty(R_X \times \mathbb{R})$ to $\mathbb{R}$, e.g., $\Psi(s) = \sup_{x,y}|s(x, y)|$ for a Kolmogorov-Smirnov test. Then we reject H_0 with nominal level α if $T_n = \Psi(S_n)$ exceeds a critical value c_α. A bootstrap approximation of c_α is given in Section 3.2. In our simulations, we use the Kolmogorov-Smirnov and Cramér-von Mises test statistics

$$T_{n,KS} = \sqrt{n}\sup_{x,y}|\hat{F}_{X,\hat{\varepsilon}}(x, y) - \hat{F}_X(x)\hat{F}_{\hat{\varepsilon}}(y)|, \tag{3.5}$$

$$T_{n,CM} = n\iint(\hat{F}_{X,\hat{\varepsilon}}(x, y) - \hat{F}_X(x)\hat{F}_{\hat{\varepsilon}}(y))^2 d\hat{F}_X(x) d\hat{F}_{\hat{\varepsilon}}(y). \tag{3.6}$$

3.2. Bootstrap approximation of the critical value

Since the asymptotic distributions of the test statistics depend in a complicated way on unknown quantities, we apply a bootstrap procedure to approximate the critical values. To this end let $\eta_1^*, \dots, \eta_n^*$ be drawn with replacement from standardized residuals $\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_n$, where

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i - n^{-1}\sum_{k=1}^n \hat{\varepsilon}_k}{(n^{-1}\sum_{j=1}^n (\hat{\varepsilon}_j - n^{-1}\sum_{k=1}^n \hat{\varepsilon}_k))^{1/2}}, \quad i = 1, \dots, n. \tag{3.7}$$

Let further $\xi_1, \dots, \xi_n$ be independent standard normals independent of the original sample $\mathcal{Y}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, and let a_n be some positive smoothing parameter. Define bootstrap errors as $\varepsilon_i^* = \eta_i^* + a_n \xi_i$. Methods based on residual empirical processes require smoothing of the bootstrap errors, cf., Neumeyer (2009b), among others. It is easily seen that, conditionally on $\mathcal{Y}_n$, ε_i^* has a smooth distribution function

$$\tilde{F}_{\tilde{\varepsilon}}(y) = \frac{1}{n}\sum_{j=1}^n \Phi(\frac{y - \tilde{\varepsilon}_j}{a_n}),$$

where Φ denotes the standard normal distribution function.

Generate X_i^* from $\hat{F}_X$ and take

$$Y_i^* = \Lambda_{\hat{\vartheta}}^{-1}(Z_i^*), \text{ where } Z_i^* = \hat{m}(X_i^*) + \hat{\sigma}(X_i^*)\varepsilon_i^*, \quad i = 1, \dots, n. \tag{3.8}$$

The bootstrap sample is (X_i^*, Y_i^*), $i = 1, \dots, n$, and fulfills H_0 by construction. To see this let E_n^* and Var_n^* denote the expectation and variance with respect to the conditional distribution $P(\cdot \mid \mathcal{Y}_n)$. Then $E_n^*[\varepsilon_i^* \mid X_i^*] \equiv 0$ and $\mathrm{Var}_n^*(\varepsilon_i^* \mid X_i^*) \equiv 1 + a_n^2$, and thus

$$\frac{\Lambda_{\hat{\vartheta}}(Y_i^*) - E_n^*[\Lambda_{\hat{\vartheta}}(Y_i^*)|X_i^*]}{(\mathrm{Var}_n^*(\Lambda_{\hat{\vartheta}}(Y_i^*)|X_i^*))^{1/2}} = \frac{\varepsilon_i^*}{(1 + a_n^2)^{1/2}} \perp X_i^*$$

(given $\mathcal{Y}_n$). Let T_n denote the test statistic based on the original sample and let T_n^* be the one based on the bootstrap sample. Then H_0 is rejected whenever $T_n > c_{n,\alpha}$, where $P(T_n^* > c_{n,\alpha} \mid \mathcal{Y}_n) = 1 - \alpha$. The critical value $c_{n,\alpha}$ is estimated by the $\lfloor B(1-\alpha) \rfloor$-largest bootstrap test statistic obtained from B replications of the bootstrap data generation.

3.3. Remarks on consistency of the proposed tests

Consider the hypothesis test of Section 3.1 when using the profile likelihood estimator $\hat{\vartheta}$ suggested in Section 2.1. With those notations let

$$p_\vartheta(y|x) = f_{\varepsilon(\vartheta)}\Big(\frac{\Lambda_\vartheta(y) - m_\vartheta(x)}{\sigma_\vartheta(x)}\Big)\frac{\Lambda'_\vartheta(y)}{\sigma_\vartheta(x)}.$$

A consistent estimator (under mild regularity conditions) of the log-likelihood

$$L_\vartheta = \log\Big(\prod_{i=1}^{n} p_\vartheta(Y_i|X_i)\Big)$$

is maximized in order to obtain the profile likelihood estimator of the transformation parameter $\vartheta \in \Theta$ (see (2.2)). Now consider the alternative H_1 that states that there exists no parameter $\vartheta \in \Theta$ such that $p_\vartheta(\cdot|x)$ is the conditional density of Y, given $X = x$. Then L_ϑ/n estimates the expectation

$$E[\log p_\vartheta(Y_i|X_i)] = \int\int (\log p_\vartheta(y|x)) f_{Y|X}(y|x)\, dy dF_X(x)$$

and thus $\hat{\vartheta}$ estimates the value $\vartheta_1 \in \Theta$ which minimizes the expected Kullback-Leibler divergence of the conditional densities $f_{Y|X}$ and p_ϑ,

$$\int\int \Big(\log \frac{f_{Y|X}(y|x)}{p_\vartheta(y|x)}\Big) f_{Y|X}(y|x)\, dy dF_X(x).$$

Thus $\hat{F}_{X,\hat{\varepsilon}}$ in Section 3.1 estimates the joint distribution of X and $\varepsilon(\vartheta_1) = (\Lambda_{\vartheta_1}(Y) - E[\Lambda_{\vartheta_1}(Y)|X])/(\mathrm{Var}\,(\Lambda_{\vartheta_1}(Y)|X))^{1/2}$. Since under H_1 the distribution of $\varepsilon(\vartheta_1)$ depends on X, it follows that, e.g., a Kolmogorov-Smirnov test statistic $T_n = \sup_{x,y} |S_n(x, y)|$ converges to infinity. Thus any test that rejects H_0 whenever T_n exceeds some constant c_α is consistent.

3.4. The homoscedastic transformation model

Let independent copies of (X, Y) be observed and a parametric class of transformations $\{\Lambda_\vartheta \mid \vartheta \in \Theta\}$ be given. Then tests for the null hypothesis

$$H_0 : \exists \vartheta \in \Theta \text{ such that } \Lambda_\vartheta(Y) - E[\Lambda_\vartheta(Y)|X] \perp X \tag{3.9}$$

are also of interest. The validity of the null hypothesis means that a nonparametric location model

$$\Lambda_{\vartheta_0}(Y) = m(X) + \varepsilon, \quad \varepsilon \perp X$$

with $m(x) = E[\Lambda_{\vartheta_0}(Y)|X = x]$ describes the data for some $\vartheta_0 \in \Theta$. Tests for model validity can be derived similarly as in the heteroscedastic case in an obvious manner. An estimator for the transformation parameter analogous to Linton, Sperlich, and Van Keilegom (2008) can be applied where the additive regression estimator is replaced by a purely nonparametric local polynomial estimator. The residuals are then defined as $\hat{\varepsilon} = \Lambda_{\hat{\vartheta}}(Y) - \hat{m}_{\hat{\vartheta}}(X)$. Under slightly weaker assumptions than those stated in Appendix A, similar asymptotic results to those in Section 3.1 can be derived. Additionally, we can use the simplification of the bootstrap in Section 3.2 to implement the test for the validity of (3.9), replacing $\tilde{\varepsilon}_i$ in (3.7) with $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - n^{-1}\sum_{k=1}^n \hat{\varepsilon}_k$, and Z_i^* in (3.8) with $Z_i^* = \hat{m}(X_i^*) + \varepsilon_i^*$.

4. Numerical Simulations

In this section, we report on three simulation studies. All computations were done with R (R Core Team (2015)). We first illustrate the finite sample performance of the estimator $\hat{\vartheta}$ of the transformation parameter in (2.4). We study the performance of the proposed test for checking homoscedasticity under some transformation when it is implemented via the bootstrap described in Section 3.4, and we verify how well the test in Section 3.1 is able to test the assumption of a heteroscedastic transformation structure when the true model gradually deviates from a heteroscedastic transformation model.

In all simulations, we considered the Yeo-Johnson family of transformations:

$$\Lambda_\vartheta(y) = \begin{cases} \frac{(y+1)^\vartheta - 1}{\vartheta}, & y \geq 0, \vartheta \neq 0, \\ \log(y+1), & y \geq 0, \vartheta = 0, \\ -\frac{(-y+1)^{2-\vartheta} - 1}{(2-\vartheta)}, & y < 0, \vartheta \neq 2, \\ -\log(-y+1), & y < 0, \vartheta = 2, \end{cases}$$

proposed by Yeo and Johnson (2000) as a generalization of the Box-Cox family. Concerning the estimation of the transformation parameter, we used the normal kernel whenever a kernel function was needed. To estimate $m(\cdot)$ and $\sigma(\cdot)$, we used the local linear estimator ($p = 1$) using the R package *np* (See Hayfield and Racine (2008)). The bandwidth was chosen by the direct plug-in methodology described by Ruppert, Sheather, and Wand (1995). For estimation of $f_{\varepsilon(\vartheta)}(\cdot)$, we used the bandwidth obtained from the method of Sheather and Jones (1991). All these bandwidth selection methods were implemented in the R package *KernSmooth* (See Wand (2015)).

4.1. Estimation of heteroscedastic transformation parameter

To see how the estimator $\hat{\vartheta}$ in (2.4) works in practice, we generated data from the following heteroscedastic transformation model:

$$\Lambda_{\vartheta_0}(Y_i) = m(X_i) + \sigma(X_i)\varepsilon_i, \; i = 1, \cdots, n, \tag{4.1}$$

where $m(x) = \exp(x) + 1.5$ and $\sigma(x) = 1 + a(x-1)$. Here $X_1, \ldots, X_n$ were independent and uniform on $[0, 1]$, $\varepsilon_1, \ldots, \varepsilon_n$ were independent standard normal, and X_i and ε_i were independent. We let θ_0 be zero here and whenever this model was used further on. For various values of a and n, we calculated $\hat{\vartheta}$ from 200 samples of size $n = 100, 200$, and 400, and computed

Table 1. The bias and mean squared error of the estimator $\hat{\vartheta}$ for $n = 100,\ 200$ and 400.

	$n = 100$		$n = 200$		$n = 400$	
	MEAN	MSE	MEAN	MSE	MEAN	MSE
$a = 0.5$	0.085	0.198	0.035	0.117	0.026	0.062
$a = 0.75$	0.077	0.200	0.048	0.090	0.008	0.053
$a = 1$	0.056	0.228	0.074	0.121	-0.009	0.066

$$\text{MEAN} = \frac{1}{200}\sum_{j=1}^{200}\hat{\vartheta}^{(j)} \text{ and } \text{MSE} = \frac{1}{200}\sum_{j=1}^{200}(\hat{\vartheta}^{(j)} - \vartheta_0)^2,$$

$\hat{\vartheta}^{(j)}$ the estimate of ϑ_0 from the jth sample. The results are given in Table 1. For various values of a, we observe that both the bias and the mean squared error of the estimator decrease as the sample size increases, which suggests the consistency of the estimator.

4.2. Testing for homoscedastic transformation models

To verify the performance of the test proposed in Section 3.4 regarding the assumption of a homoscedastic transformation model, we use model (4.1). The degree of heteroscedasticity decreases as the value of a gets closer to 0 and (4.1) is a homoscedastic transformation model when $a = 0$, which satisfies the null hypothesis (3.9). We investigate how the test behaves as the value of a increases from 0 to 1.

We consider the Kolmogorov-Smirnov and Cramér-von Mises test statistics in (3.5). To find the critical value for the proposed tests, we used 200 bootstrap replications for each sample. For the smooth bootstrap described in Section 3.4, we set a_n to $0.5n^{-1/4}$, as in Neumeyer (2009b). In that reference the validity of a smooth residual bootstrap procedure for the residual-based empirical process was shown rigorously under some conditions on the relationship between the bandwidths h_1 and a_n when $d = 1$ (see assumption A.6 in the reference). They basically assert that

$$\frac{nh_1 a_n^{2+\Delta_1}}{\log(h_1^{-1})} \to \infty, \quad h_1 = o(a_n^{1+\Delta_2}) \text{ for some small positive } \Delta_1, \Delta_2. \tag{4.2}$$

We think that combining that method of proof with ours that could prove the validity of the smooth residual bootstrap for the transformation model under similar conditions. For $a_n \sim n^{-1/4}$ the conditions in (4.2) amount to $nh_1^2\omega_{1,n} \to \infty$, $nh_1^4\omega_{2,n} \to 0$ for some sequences $\omega_{1,n} \to 0$ and $\omega_{2,n} \to \infty$ slowly. Our bandwidth conditions (a2) for $p = d = 1$ say $nh_1^{3+\delta} \to \infty$, $nh_1^4 \to 0$, so they do not contradict the conditions for the bootstrap.

Table 2. The power of the test for verifying the validity of a homoscedastic transformation structure. The power was calculated based on 200 samples. The null hypothesis is satisfied for $a = 0$.

	$n = 100$				$n = 200$			
	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.1$	
	KS	CM	KS	CM	KS	CM	KS	CM
$a = 0$	0.040	0.035	0.075	0.065	0.045	0.045	0.070	0.085
$a = 0.5$	0.085	0.125	0.125	0.160	0.110	0.165	0.180	0.240
$a = 0.75$	0.340	0.460	0.445	0.510	0.545	0.690	0.670	0.775
$a = 1$	0.910	0.980	0.965	0.980	0.995	1.000	1.000	1.000

Table 2 shows the results for the test implemented via the bootstrap described in Section 3.4. The size of the test is somewhat too low, but the power grows to one as the parameter a measuring the degree of heteroscedasticity gets larger. One feature of the results is that the power does not change much until the degree of heteroscedasticity reaches a certain level and then starts to increase rapidly. To explain this peculiar behavior, we show in Figure 1 four plots using data of size $n = 200$ from (4.1). These plots are given for two values of a, and compare the regression function based on the true parameter ϑ_0 with the one based on the estimator $\hat{\vartheta}$.

When $a \neq 0$, the estimator $\hat{\vartheta}$ is not consistent due to the misspecification of the heteroscedastic error structure, and instead targets the pseudo-true parameter $\vartheta^* \neq \vartheta_0$ that maximizes

$$PL(\vartheta) = E(\log f_{\varepsilon_\vartheta}(\Lambda_\vartheta(Y) - m_\vartheta(X)) + \log \Lambda'_\vartheta(Y)), \tag{4.3}$$

where $m_\vartheta(x) = E(\Lambda_\vartheta(Y)|X = x)$ and $\varepsilon_\vartheta = \Lambda_\vartheta(Y) - m_\vartheta(X)$. This value has the interpretation that the corresponding homoscedastic model is the best approximation to the true heteroscedastic transformation model. When the degree of heteroscedasticity is moderate, it is possible that the data look like data coming from a homoscedastic transformation model with transformation parameter $\hat{\vartheta}$ (see the upper right panel of Figure 1. In this case, our test is not able to detect the violation of (3.9) well, and behaves as if the null hypothesis is true. When the degree of heteroscedasticity is severe, the data cannot be considered anymore to come from a homoscedastic transformation model, and it is possible to detect the violation through the dependence between X and $\hat{\varepsilon}$ (see the right lower panel of Figure 1). This feature is different from what was observed in testing for homoscedasticity in regression settings without transformation, such as in Neumeyer (2009a).

Since our testing procedure involves estimation of many parameters and functions, one is interested in how the selection of the smoothing parameters for these

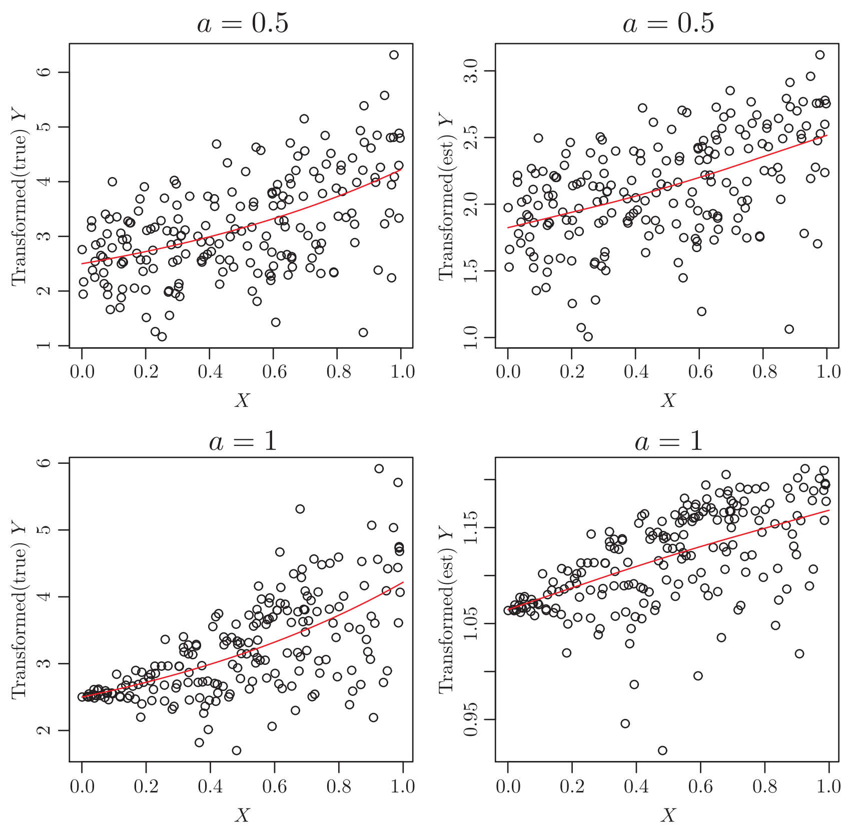

Figure 1. Plot of $\Lambda_{\vartheta=\vartheta_0}(Y_i)$ versus X_i (left panel), and $\Lambda_{\vartheta=\hat{\vartheta}}(Y_i)$ versus X_i (right panel), when $a = 0.5$ (upper panel) and $a = 1$ (lower panel). The solid curves are $m_{\vartheta_0}(\cdot)$ (left) and $m_{\hat{\vartheta}}(\cdot)$ (right).

estimators affects the performance of the proposed tests. We first investigated the impact of the choice of a_n on the test. We reran the simulation for Table 2 but with other choices of a_n: $a_n = 0.25n^{-1/4}, n^{-1/4}$, which produces Table 3. Table 3 suggests that the performance (level or power) of the test is not so sensitive to the choice of a_n. We investigated the impact of bandwidth selection on the behavior of the tests where the test for homoscedastic transformation models requires three bandwidths. To calculate the profile likelihood, we used two bandwidths for $\hat{m}_\vartheta$ and $\hat{f}_{\varepsilon(\vartheta)}$. Once $\hat{\vartheta}$ is obtained, we need another bandwidth to calculate the residual $\hat{\varepsilon}_i = \Lambda_{\hat{\vartheta}}(Y_i) - \hat{m}(X_i)$. In our simulations, these bandwidths were chosen as the optimal bandwidths in terms of MISE using the methods of Ruppert, Sheather, and Wand (1995) and Sheather and Jones (1991). To see the impact of bandwidth selection, we used half of the optimal bandwidth or twice the optimal bandwidth whenever bandwidth selection was necessary and checked the level and power with such bandwidth. The results are summarized in Table

Table 3. The power of the test for verifying the validity of a homoscedastic transformation structure when different bandwidths a_n were used for the smooth bootstrap. The power was calculated based on 200 samples. The null hypothesis is satisfied for $a = 0$.

		$n = 100$				$n = 200$			
		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.1$	
a_n	a	KS	CM	KS	CM	KS	CM	KS	CM
$0.25n^{-1/4}$	$a = 0$	0.025	0.030	0.055	0.055	0.040	0.040	0.065	0.085
	$a = 0.5$	0.080	0.120	0.130	0.155	0.120	0.165	0.180	0.245
	$a = 0.75$	0.300	0.440	0.415	0.510	0.555	0.690	0.645	0.785
	$a = 1$	0.910	0.975	0.955	0.975	0.995	1.000	1.000	1.000
$n^{-1/4}$	$a = 0$	0.035	0.025	0.090	0.070	0.040	0.070	0.080	0.090
	$a = 0.5$	0.095	0.115	0.170	0.170	0.130	0.150	0.170	0.240
	$a = 0.75$	0.335	0.450	0.425	0.520	0.550	0.660	0.640	0.740
	$a = 1$	0.905	0.975	0.955	0.980	0.990	1.000	1.000	1.000

Table 4. The power of the test for verifying the validity of a homoscedastic transformation structure when different bandwidths were used. The power was calculated based on 200 samples. The null hypothesis is satisfied for $a = 0$.

		$n = 100$				$n = 200$			
		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.1$	
	bandwidth	KS	CM	KS	CM	KS	CM	KS	CM
$a = 0$	half	0.030	0.040	0.085	0.065	0.080	0.070	0.010	0.011
	our choice	0.040	0.035	0.075	0.065	0.045	0.045	0.070	0.085
	twice	0.035	0.025	0.060	0.045	0.025	0.040	0.060	0.095
$a = 0.75$	half	0.285	0.360	0.395	0.495	0.590	0.700	0.690	0.780
	our choice	0.340	0.460	0.455	0.510	0.545	0.690	0.670	0.775
	twice	0.325	0.450	0.440	0.545	0.540	0.710	0.650	0.795

4. We see there that the level and power is not so sensitive to the choice of the bandwidths, which makes our procedure applicable in practice.

4.3. Testing for heteroscedastic transformation models

Finally, we illustrate how the test in Section 3.1 works to verify the assumption of a heteroscedastic transformation structure. Here we define two new transformation models. Basically, they are model (4.1), except that the error distribution is as follows.

Model A

$$(\varepsilon|X = x) \sim \begin{cases} N(0, 1^2), & \text{if } 0.5 < x \leq 1; \\ \frac{W - E(W)}{\sqrt{Var(W)}}, \text{ where } W \sim ST(0, 1, \alpha, \nu), & \text{if } 0 \leq x \leq 0.5. \end{cases}$$

Table 5. The power of the test for verifying the validity of a heterocedastic transformation structure from Model A. The power was calculated based on 200 samples. The null hypothesis is satisfied for $\alpha = 0$ and $\nu = \infty$.

	$n = 100$				$n = 200$			
	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.1$	
	KS	CM	KS	CM	KS	CM	KS	CM
$\alpha = 100,\ \nu = 2.1$	0.370	0.445	0.505	0.590	0.710	0.770	0.795	0.850
$\alpha = 0,\ \nu = 2.1$	0.105	0.140	0.170	0.200	0.205	0.270	0.325	0.360
$\alpha = 0,\ \nu = 5$	0.075	0.060	0.105	0.085	0.060	0.060	0.130	0.095
$\alpha = 0,\ \nu = \infty$	0.055	0.060	0.070	0.105	0.080	0.070	0.120	0.135

Model B

$$(\epsilon|X = x) = \begin{cases} N(0, 1^2), & \text{if } 0.5 < x \leq 1; \\ \frac{W-\eta}{\sqrt{2\eta}}, \text{ where } W \sim \chi^2(\eta), & \text{if } 0 \leq x \leq 0.5. \end{cases}$$

Here, $ST(\xi, \Omega, \alpha, \nu)$ is a skew-t distribution with parameters ξ, Ω, α, and ν defined in Azzalini (2005). The parameter α controls the skewness of the distribution and the paramer ν controls kurtosis. Additionally, we set $\sigma(x) = x$ (so $a = 1$). First, note that as $\nu \to \infty$ and $\alpha \to 0$, Model A converges to model (4.1) with $\sigma(x) = x$, which satisfies the assumption of a heteroscedastic transformation structure (the same thing happens as $\eta \to \infty$ in case of Model B). The first and second moments of the conditional error distribution given X coincide with the respective moments under model (4.1). The parameters α, ν, and η determine how much the model violates (3.1). In our simulations, to see how the test performs when the true model gradually deviates from the assumption under the null hypothesis, we investigated the power function as ν changed from ∞ to 2.1 and then as α changed from 0 to 100 for Model A, and as η changed from ∞ to 2 for Model B. Here, ν should be greater than 2 and η should be at least 2 lest one cannot standardize the distribution of W due to variance explosion. For the smooth bootstrap described in Section 3.2, we set a_n to $0.5n^{-1/4}$ and used 200 bootstrap replications.

As seen in the case of homoscedastic transformation models, we observe from Tables 5 and 6 that there is a threshold of difference in two component distributions in the error above which we can detect the violation of the assumption, and the power starts to grow beyond that threshold. Compared to Model A, the power of Model B is somewhat lower; we attribute this to the flexibility of the heteroscedastic transformation model. Since it is a very flexible model, unless the two component distributions in the error are strikingly different from each other, the generated data look like data coming from a heteroscedastic transformation model with appropriately chosen transformation parameter.

Table 6. The power of the test for verifying the validity of a heterocedastic transformation structure from Model B. The power was calculated based on 200 samples. The null hypothesis is satisfied for $\eta = \infty$.

	$n = 100$				$n = 200$			
	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.1$	
	KS	CM	KS	CM	KS	CM	KS	CM
$\eta = 2$	0.215	0.220	0.285	0.310	0.325	0.355	0.455	0.440
$\eta = 3$	0.100	0.165	0.175	0.270	0.155	0.220	0.270	0.295
$\eta = 5$	0.090	0.095	0.140	0.150	0.120	0.125	0.190	0.200
$\eta = 10$	0.050	0.065	0.091	0.125	0.100	0.105	0.140	0.190
$\eta = \infty$	0.065	0.060	0.105	0.115	0.045	0.055	0.100	0.100

Table 7. The calculated P-values for the validity of homoscedastic transformation models concerning the ultrasonic calibration data.

	P-value	
a_n	KS	CM
$0.25n^{-1/4}$	0.883	0.617
$0.5n^{-1/4}$	0.853	0.560
$n^{-1/4}$	0.825	0.490

4.4. Data analysis

To illustrate our method, we analyze the ultrasonic calibration data that can be found in the NIST/SEMATECH e-Handbook of Statistical Methods. The data is available on the website (`http://www.itl.nist.gov/div898/handbook/pmd/section6/pmd631.htm`). The response Y is ultrasonic response and the predictor X is metal distance. Concerning these data, it has been found in the e-book that the data seem to satisfy the assumption of homoscedastic transformation models with the square-root transformation of Y, $\sqrt{Y_i} = m(X_i) + \varepsilon_i,\ i = 1, \ldots, n$. We wanted to test whether our method can rediscover such validity without the information about the appropriate transformation. We considered the Box-Cox transformation family since the square-root transformation is included in the family and all the responses are positive. We calculated the P-values from the proposed test with various choices of a_n. For a more accurate result, the number of bootstrap iterations was set as 400. The estimated transformation parameter $\hat{\vartheta}$ was 0.436, close to 0.5. The P-values in Table 7 suggest that the given data satisfy the assumption of the homoscedastic transformation model, which is consistent with the analysis of the previous study. Additionally, we compare the residual plots of the two regression models, $Y = m(X) + \varepsilon$ and $\Lambda_{\vartheta=0.436}(Y) = m'(X) + \varepsilon'$, where $\{\Lambda_\vartheta(\cdot)\}$ is the family of Box-Cox transformations. The plots (Figure 2) suggest that the the transformation of the response stablizes the variance function in the regression model.

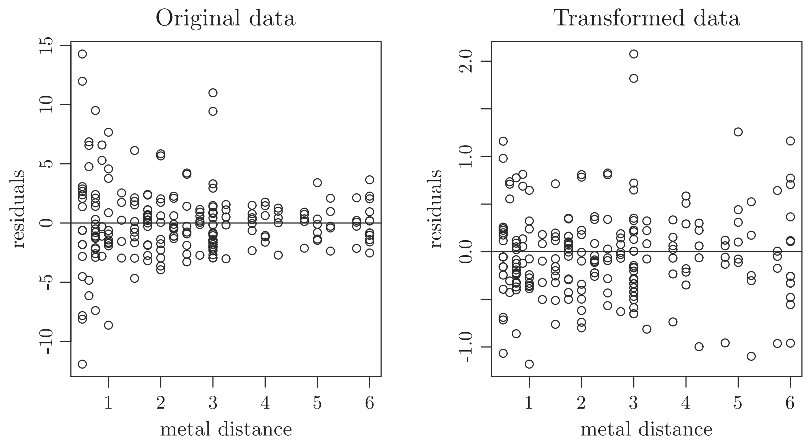

Figure 2. The residual plots of the two regression models, $Y = m(X) + \varepsilon$ and $\Lambda_{\vartheta=0.436}(Y) = m'(X) + \varepsilon'$, where $\{\Lambda_\vartheta(\cdot)\}$ is the family of Box-Cox transformations.

Supplementary Materials

The online supplementary material includes the proof of Theorem 1 and the proof of some auxiliary results used to prove Theorem 2.

Acknowledgements

The authors thank the Co-Editor, an associate editor, and two reviewers for their helpful comments and suggestions, and Oliver Linton for stimulating discussions regarding the profile likelihood method. The first author acknowledges financial support by the DFG (Research Unit FOR 1735 *Structural Inference in Statistics: Adaptation and Effciency*). The research of the second author was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A1A2059875). The research of the third author was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy), and by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

Appendix

Appendix A. Regularity Conditions

For the asymptotic normality of the estimator $\hat{\vartheta}$, we need regularity conditions.

(a1) k is a symmetric probability density function supported on $[-1,1]$, k is $d+1$ times continuously differentiable, and $k^{(j)}(\pm 1)=0$ for $j=0,\ldots,d-1$.

(a2) h_j $(j=1,\ldots,d)$ satisfies $h_j/h_0 \to c_j$ for some $0<c_j<\infty$ and some baseline bandwidth h_0 satisfying $nh_0^{2p+2} \to 0$, for some $p \geq 3$, and $nh_0^{3d+\delta} \to \infty$ for some small $\delta > 0$.

(a3) The kernel ℓ is a symmetric, twice continuously differentiable function supported on $[-1,1]$, $\int u^s \ell(u)du = 0$ for $s=1,\ldots,q-1$, and $\int u^q \ell(u)du \neq 0$ for some $q \geq 4$. The bandwidth g satisfies $ng^6(\log n)^{-2} \to \infty$ and $ng^{2q} \to 0$.

(a4) The support R_X of the covariate X is a compact subset of $\mathbb{R}^d$, the distribution function F_X is $2d+1$-times continuously differentiable, $\inf_{x\in R_X} f_X(x) > 0$, and $\inf_{x\in R_X} \sigma(x) > 0$. The functions $m_\vartheta(x)$, $\dot{m}_\vartheta(x)$, $\sigma_\vartheta(x)$, and $\dot{\sigma}_\vartheta(x)$ are $p+2$ times continuously differentiable with respect to the components of x on $R_X \times \mathcal{N}(\vartheta_0)$, and all derivatives up to order $p+2$ are bounded uniformly in $(x,\vartheta) \in R_X \times \mathcal{N}(\vartheta_0)$, where $\mathcal{N}(\vartheta_0)$ is a neighborhood of ϑ_0.

(a5) The transformation Λ_ϑ satisfies $\sup_{\vartheta\in\Theta, x\in R_X} ||E[\dot{\Lambda}_\vartheta(Y)|X=x]|| < \infty$, $\sup_{x\in R_X} ||E[\dot{\Lambda}^4_{\vartheta_0}(Y)|X=x]|| < \infty$, and the density function of $(\dot{\Lambda}_\vartheta(Y), X)$ exists and is continuous for all $\vartheta \in \Theta$. In addition, $\Lambda_\vartheta(y)$ is three times continuously differentiable with respect to y and ϑ, and there exists a $\delta > 0$ such that

$$E\Big[\sup_{\vartheta':\|\vartheta'-\vartheta\|\leq\delta} \Big|\frac{\partial^{j+r}}{\partial y^j \partial\vartheta_1^{r_1}\ldots\partial\vartheta_k^{r_k}} \Lambda_{\vartheta'}(Y)\Big|\Big] < \infty,$$

for all $\vartheta \in \Theta$ and all $0 \leq j+r \leq 3$, where $r = \sum_{i=1}^k r_i$.

(a6) The error term ε has finite sixth moment and is independent of X. The distribution $F_{\varepsilon(\vartheta)}(y)$ is three times continuously differentiable with respect to y and ϑ,

$$\sup_{y,\vartheta} \Big|\frac{\partial^{j+r}}{\partial y^j \partial\vartheta_1^{r_1}\ldots\partial\vartheta_k^{r_k}} F_{\varepsilon(\vartheta)}(y)\Big| < \infty$$

for all $0 \leq j + \sum_{i=1}^k r_i \leq 2$, $\sup_y |yf'_\varepsilon(y)| < \infty$, $\sup_y |y\dot{f}'_\varepsilon(y)| < \infty$ and $\sup_y |y^2 f''_\varepsilon(y)| < \infty$. The conditional distribution $F_{\varepsilon(\vartheta)|X}(y|x)$ is three times continuously differentiable with respect to y and ϑ,

$$\sup_{y,x,\vartheta} \Big|\frac{\partial^{j+r}}{\partial y^j \partial\vartheta_1^{r_1}\ldots\partial\vartheta_k^{r_k}} F_{\varepsilon(\vartheta)|X}(y|x)\Big| < \infty$$

for all $0 \leq j+\sum_{i=1}^k r_i \leq 2$, $\sup_{y,x} |yf'_{\varepsilon|X}(y|x)| < \infty$, $\sup_{y,x} |y\dot{f}'_{\varepsilon|X}(y|x)| < \infty$ and $\sup_{y,x} |y^2 f''_{\varepsilon|X}(y|x)| < \infty$.

(a7) For all $\eta > 0$, there exists $\epsilon(\eta) > 0$ such that $\inf_{\|\vartheta-\vartheta_0\|>\eta} \|G(\vartheta)\| \geq \epsilon(\eta) > 0$. The matrix Γ defined in Theorem 1 is of full rank.

For the results of Section 3, we need assumptions (a1), (a2) and the following conditions. Let $\|\cdot\|$ denote some vector or matrix norm, as appropriate.

(A1) All partial derivatives of F_X up to order $2d+1$ exist on the interior of its compact support R_X, they are uniformly continuous, and $\inf_{x\in R_X} f_X(x) > 0$.

(A2) All partial derivatives of m and σ up to order $p+2$ exist on the interior of R_X, they are uniformly continuous, and $\inf_{x\in R_X} \sigma(x) > 0$.

(A3) F_ε is twice continuously differentiable, $\sup_y |y f_\varepsilon(y)| < \infty$, $\sup_y |y^2 f_\varepsilon'(y)| < \infty$, and $E(\varepsilon^6) < \infty$.

(A4) $\sup_{y\in\mathbb{R}} E\left[\left\|\nabla_\vartheta F_{\varepsilon(\vartheta)|X}(y|X)|_{\vartheta=\vartheta_0}\right\|\right] < \infty$.

(A5) For the parameter estimator a linear expansion as in (3.2) is valid with $E[g_{\vartheta_0}(X,Y)] = 0$, $E[\|g_{\vartheta_0}(X,Y)\|^2] < \infty$.

(A6) If $F_{Y|X}(\cdot|x)$ and $f_{Y|X}(\cdot|x)$ denote the conditional distribution and density function of Y, given $X = x$, respectively, there exists $\eta > 0$ such that

$$\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\eta} \sup_{z\in\mathbb{R}} \int \Big(|f_{Y|X}'(V_\vartheta(z)|u)|\|\dot{V}_\vartheta(z)\|^2 + f_{Y|X}(V_\vartheta(z)|u)\|\ddot{V}_\vartheta(z)\|\Big) dF_X(x) < \infty,$$

where $V_\vartheta = \Lambda_\vartheta^{-1}$, $\dot{V}_\vartheta = \nabla_\vartheta V_\vartheta$, and $\ddot{V}_\vartheta = (\frac{\partial^2 V_\vartheta}{\partial\vartheta_i\vartheta_j})_{i,j=1,\ldots,k}$. Further,

$$\sup_{y\in\mathbb{R},x\in R_X} \Big\| y \frac{\partial(f_{Y|X}(V_{\vartheta_0}(y)|x)\dot{V}_{\vartheta_0}(y))}{\partial y} \Big\| < \infty.$$

(A7) For some $\eta > 0$, $E[\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\eta} \|\ddot{\Lambda}_\vartheta(Y)\|] < \infty$, $E[\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\eta} \|\dot{\Lambda}_\vartheta(Y)\|^2] < \infty$ and $E[\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\eta} \|\ddot{\Lambda}_\vartheta(Y)\Lambda_\vartheta(Y)\|] < \infty$. Further,

$$E\Big[\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\eta} \|\Lambda_\vartheta(Y)\dot{\Lambda}_\vartheta(Y)\| \,\Big|\, X = x\Big] < \infty,$$
$$E\Big[\sup_{\vartheta:\|\vartheta-\vartheta_0\|\leq\eta} \|\dot{\Lambda}_\vartheta(Y)\| \,\Big|\, X = x\Big] < \infty$$

for almost all $x \in R_X$.

(A8) Assumption (A2) holds with m replaced by $E[\frac{\partial\Lambda_\vartheta(Y)}{\partial\vartheta_i}|_{\vartheta=\vartheta_0}|X = \cdot]$ and σ replaced by $E[\Lambda_{\vartheta_0}(Y)\frac{\partial\Lambda_\vartheta(Y)}{\partial\vartheta_i}|_{\vartheta=\vartheta_0}|X = \cdot]$, for $i = 1,\ldots,k$. Further, $E[\|\dot{\Lambda}_{\vartheta_0}(Y)\|^3] < \infty$ and $E[\|\Lambda_{\vartheta_0}(Y)\dot{\Lambda}_{\vartheta_0}(Y)\|^3] < \infty$.

Remarks on the assumptions. Assumptions (a1)−(a7) are needed for Theorem 1. Here, (a1)−(a3) refers to choices of bandwidths and kernel functions. Assumptions (a4)−(a7) are conditions on the model, analogous to assumptions A.1–A.8 of Linton, Sperlich, and Van Keilegom (2008) with some changes due to heteroscedasticity of our model and application of local polynomial estimators. For Theorem 2, instead of (a3)−(a7), we formulate (A5) about the linear expansion of the estimator $\hat{\vartheta}$. Thus the application of Theorem 2 for different estimators $\hat{\vartheta}$ apart from the profile likelihood estimator is possible. Assumptions (a1), (a2), (A1)−(A3) are typically needed for weak convergence of empirical residual processes, compare to assumptions (C1)−(C5) of Neumeyer and Van Keilegom (2010) in a model without transformation. Assumption (A4) is needed for some Taylor expansion of the, now ϑ-dependent, error distribution with respect to ϑ. In particular they allow interchanging derivatives with integrals that appear in several terms. Assumptions (A6)−(A8) are needed to prove asymptotic expansions of the empirical process, using Taylor expansions with respect to ϑ. For specific classes of transformations, some assumptions can be reformulated or replaced by simpler conditions. For example, (A7) for the Yeo-Johnson family used in the simulations can be deduced from (conditional) moment assumptions on the observations Y.

Appendix B. Proof of the Main Results

B.1. Proof of Theorem 2

Let $\hat{F}_{X,\varepsilon}$ denote the joint empirical distribution function of (X_i, ε_i), $i = 1, \ldots, n$, under H_0. Let further

$$R_n(x,y) = E[I\{X \le x\} I\{\Lambda_{\hat{\vartheta}}(Y) \le y\hat{\sigma}(X) + \hat{m}(X)\} \mid \mathcal{Y}_n] - E[I\{X \le x\} I\{\varepsilon \le y\}],$$

where $\mathcal{Y}_n = \{(X_i, Y_i) \mid i = 1, \ldots, n\}$.

Lemma B.1. Under the assumptions of Theorem 2,

$$\hat{F}_{X,\hat{\varepsilon}}(x,y) = \hat{F}_{X,\varepsilon}(x,y) + R_n(x,y) + o_P(\frac{1}{\sqrt{n}})$$

uniformly with respect to $x \in R_X, y \in \mathbb{R}$.

Proof of Lemma B.1. We need some notation. For $k = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$, let $k. = \sum_{j=1}^d k_j$, $D^k = \partial^{k.}/\partial x_1^{k_1} \ldots \partial x_d^{k_d}$, and

$$\|f\|_{d+\alpha} = \max_{k. \le d} \sup_{x \in R_X} |D^k f(x)| + \max_{k. = d} \sup_{x,x' \in R_X} \frac{|D^k f(x) - D^k f(x')|}{\|x - x'\|^\alpha},$$

where $\| \cdot \|$ is the Euclidean norm in $\mathbb{R}^d$. Let $\mathcal{G}_1 = C_1^{d+\alpha}(R_X)$ be the class of d times differentiable functions f defined on R_X such that $\|f\|_{d+\alpha} \le 1$, and

$\mathcal{G}_2 = \tilde{C}_2^{d+\alpha}(R_X)$ be the class of d times differentiable functions f defined on R_X such that $\|f\|_{d+\alpha} \leq 2$ and $\inf_{x \in R_X} f(x) \geq 1/2$. Let

$$\varphi_{\vartheta,g_1,g_2,y}(X,Y) = I\Big\{\frac{\Lambda_\vartheta(Y) - m(X)}{\sigma(X)} \leq y g_2(X) + g_1(X)\Big\} \\ -I\Big\{\frac{\Lambda_{\vartheta_0}(Y) - m(X)}{\sigma(X)} \leq y\Big\}.$$

With this notation,

$$\sqrt{n}(\hat{F}_{X,\hat{\varepsilon}}(x,y) - \hat{F}_{X,\varepsilon}(x,y) - R_n(x,y)) = G_n\Big(x, \hat{\vartheta}, \frac{\hat{m} - m}{\sigma}, \frac{\hat{\sigma}}{\sigma}, y\Big),$$

where the empirical process

$$G_n(x,\vartheta,g_1,g_2,y) \\ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(I\{X_i \leq x\}\varphi_{\vartheta,g_1,g_2,y}(X_i,Y_i) - E[I\{X \leq x\}\varphi_{\vartheta,g_1,g_2,y}(X,Y)]\Big)$$

(indexed in $x \in R_X, \vartheta \in \Theta, g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2, y \in \mathbb{R}$) converges weakly to a Gaussian process. This follows from Proposition S2.1 in the supplementary document, the Donsker property of $\{I\{X \leq x\} \mid x \in R_X\}$, and because products of uniformly bounded Donsker classes are Donsker (see Example 2.10.8 in van der Vaart and Wellner (1996)). Thus G_n is asymptotically stochastically equicontinuous with respect to

$$\rho\Big((x,\vartheta,g_1,g_2,y),(x',\vartheta',g_1',g_2',y')\Big) \\ = \Big(\mathrm{Var}\Big(I\{X \leq x\}\varphi_{\vartheta,g_1,g_2,y}(X,Y) - I\{X \leq x'\}\varphi_{\vartheta',g_1',g_2',y'}(X,Y)\Big)\Big)^{1/2}$$

(see pages 262 and 263 of van der Vaart (1998)). We have

$$\rho\Big((x,\hat{\vartheta},\frac{\hat{m}-m}{\sigma},\frac{\hat{\sigma}}{\sigma},y),(x,\vartheta_0,0,1,y)\Big) = o_P(\delta_n),$$

where $\delta_n \searrow 0$ by Proposition S2.3 in the supplementary document. Thus, and because $\varphi_{\vartheta_0,0,1,y} \equiv 0$, it follows that

$$P\Big(\sup_{x,y} |\sqrt{n}(\hat{F}_{X,\hat{\varepsilon}}(x,y) - \hat{F}_{X,\varepsilon}(x,y) - R_n(x,y))| > \eta\Big) \\ \leq P\Big(\sup_{\rho((x,\vartheta,g_1,g_2,y),(x',\vartheta',g_1',g_2',y')) \leq \delta_n} |G_n(x,\vartheta,g_1,g_2,y) - G_n(x',\vartheta',g_1',g_2',y')| > \eta\Big)$$

which converges to zero for $n \to \infty$, for all $\eta > 0$. From this the assertion of Lemma B.1 follows.

To finish the proof of Theorem 2 we write $R_n = A_n + B_n + C_n$, where

$$\begin{aligned}
A_n(x,y) = &\quad E[I\{X \le x\} I\{\Lambda_{\hat\vartheta}(Y) \le y\hat\sigma(X) + \hat m(X)\} \mid \mathcal{Y}_n] \\
&- E[I\{X \le x\} I\{\Lambda_{\vartheta_0}(Y) \le y\hat\sigma(X) + \hat m(X)\} \mid \mathcal{Y}_n], \\
B_n(x,y) = &\quad E[I\{X \le x\} I\{\Lambda_{\vartheta_0}(Y) \le y\hat\sigma_{\hat\vartheta}(X) + \hat m_{\hat\vartheta}(X)\} \mid \mathcal{Y}_n] \\
&- E[I\{X \le x\} I\{\Lambda_{\vartheta_0}(Y) \le y\hat\sigma_{\vartheta_0}(X) + \hat m_{\vartheta_0}(X)\} \mid \mathcal{Y}_n], \\
C_n(x,y) = &\quad E[I\{X \le x\} I\{\Lambda_{\vartheta_0}(Y) \le y\hat\sigma_{\vartheta_0}(X) + \hat m_{\vartheta_0}(X)\} \mid \mathcal{Y}_n] \\
&- E[I\{X \le x\} I\{\Lambda_{\vartheta_0}(Y) \le y\sigma_{\vartheta_0}(X) + m_{\vartheta_0}(X)\}].
\end{aligned}$$

For the ease of notation, suppose the parameter ϑ is one-dimensional. We use the notation as in assumption (A6). Then

$$\begin{aligned}
&A_n(x,y) \\
&= \int \Big(F_{Y|X}(V_{\hat\vartheta}(y\hat\sigma(u) + \hat m(u))|u) - F_{Y|X}(V_{\vartheta_0}(y\hat\sigma(u) + \hat m(u))|u) \Big) I\{u \le x\} dF_X(u).
\end{aligned}$$

For the moment fix u and $z = y\hat\sigma(u) + \hat m(u)$, and consider a second order Taylor expansion of the map $\vartheta \mapsto \psi(\vartheta) = F_{Y|X}(V_\vartheta(z)|u)$,

$$\begin{aligned}
\psi(\hat\vartheta) - \psi(\vartheta_0) &= f_{Y|X}(V_{\vartheta_0}(z)|u)\dot V_{\vartheta_0}(z)(\hat\vartheta - \vartheta_0) \\
&\quad + \frac{1}{2}\Big(f'_{Y|X}(V_{\vartheta^*}(z)|u)(\dot V_{\vartheta^*}(z))^2 + f_{Y|X}(V_{\vartheta^*}(z)|u)\ddot V_{\vartheta^*}(z) \Big)(\hat\vartheta - \vartheta_0)^2.
\end{aligned}$$

The value ϑ^* may depend on u and z, but lies between $\hat\vartheta$ and ϑ_0. Because for each $\eta > 0$, $|\hat\vartheta - \vartheta_0| \le \eta$ with probability converging to one, for the proof we may assume $|\vartheta^* - \vartheta_0| \le \eta$ with η, by (A6). A Taylor expansion of ψ motivates the definition of

$$\begin{aligned}
&\tilde A_n(x,y) \\
&= \int f_{Y|X}(V_{\vartheta_0}(y\hat\sigma(u) + \hat m(u))|u)\dot V_{\vartheta_0}(y\hat\sigma(u) + \hat m(u)) I\{u \le x\}\, dF_X(u)(\hat\vartheta - \vartheta_0)
\end{aligned}$$

and yields that

$$\begin{aligned}
&\sup_{x,y} |A_n(x,y) - \tilde A_n(x,y)| \\
&\le (\hat\vartheta - \vartheta_0)^2 \frac{1}{2} \sup_{\vartheta: |\vartheta - \vartheta_0| \le \eta} \sup_{z \in \mathbb{R}} \int \Big(|(f'_{Y|X}(V_\vartheta(z)|u)|(\dot V_\vartheta(z))^2 + f_{Y|X}(V_\vartheta(z)|u)|\ddot V_\vartheta(z)| \Big) \\
&\quad dF_X(x) \\
&= o_P(\frac{1}{\sqrt{n}})
\end{aligned}$$

by (A6). Denote by $\bar A_n$ the same term as $\tilde A_n$, but with the estimators $\hat\sigma$ and $\hat m$ replaced by the true functions σ and m, respectively. From the proof of

Proposition S2.2 in the supplementary document, uniform convergence of $|\hat{\sigma}-\sigma|$ and $|\hat{m}-m|$ to zero in probability follows, and thus by the Mean Value Theorem, the last part of (A6), and $\hat{\vartheta}-\vartheta_0 = O_P(n^{-1/2})$ we obtain $\sup_{x,y}|\tilde{A}_n(x,y) - \bar{A}_n(x,y)| = o_P(n^{-1/2})$. Altogether for A_n, we have uniformly with respect to $x \in R_X$, $y \in \mathbb{R}$,

$$A_n(x,y) = \int f_{Y|X}(V_{\vartheta_0}(y\sigma(u)+m(u))|u)\dot{V}_{\vartheta_0}(y\sigma(u)+m(u))I\{u \le x\}dF_X(u)(\hat{\vartheta}-\vartheta_0) + o_P(\frac{1}{\sqrt{n}}).$$

For C_n we obtain an expansion, uniformly with respect to x, y,

$$\begin{aligned} C_n(x,y) = {} & E\Big[I\{X \le x\}I\Big\{\varepsilon \le y\frac{\hat{\sigma}_{\vartheta_0}(X)}{\sigma(X)} + \frac{\hat{m}_{\vartheta_0}(X)-m(X)}{\sigma(X)}\Big\} \mid \mathcal{Y}_n\Big] \\ & - E[I\{X \le x\}I\{\varepsilon \le y\}] \\ = {} & \int \Big(F_\varepsilon\Big(y\frac{\hat{\sigma}_{\vartheta_0}(u)}{\sigma(u)} + \frac{\hat{m}_{\vartheta_0}(u)-m(u)}{\sigma(u)}\Big) - F_\varepsilon(y)\Big)I\{u \le x\}\,dF_X(u) \\ = {} & f_\varepsilon(y)\Big(y\int \frac{\hat{\sigma}_{\vartheta_0}(u)-\sigma(u)}{\sigma(u)}I\{u \le x\}\,dF_X(u) \\ & + \int \frac{\hat{m}_{\vartheta_0}(u)-m(u)}{\sigma(u)}I\{u \le x\}\,dF_X(u)\Big) + o_P(\frac{1}{\sqrt{n}}) \\ = {} & f_\varepsilon(y)\frac{1}{n}\sum_{i=1}^n(\varepsilon_i + \frac{y}{2}(\varepsilon_i^2-1))\int \frac{1}{h}K^*\Big(\frac{u-X_i}{h}\Big)I\{u \le x\}\,du + o_P(\frac{1}{\sqrt{n}}). \end{aligned}$$

The second but last equality follows by Taylor's expansion, (A3) and the fact that $\int(\hat{m}_{\vartheta_0}-m)^2/\sigma^2\,dF_X = o_P(n^{-1/2})$, $\int(\hat{\sigma}_{\vartheta_0}-\sigma)^2/\sigma^2\,dF_X = o_P(n^{-1/2})$, see the proof of Theorem 2.1 in Neumeyer and Van Keilegom (2010). The last equality follows from (S1.1) and (S1.2) in the supplementary document, a combination of the proof of Lemma A.2 in Neumeyer and Van Keilegom (2010), and the proof of Proposition 2 in Neumeyer and Van Keilegom (2009).

Now let either $Z_i = \varepsilon_i$ or $Z_i = \varepsilon_i^2 - 1$. Then, as in the last part of the proof of Lemma B.1 in the supporting information to Birke and Neumeyer (2013), we have

$$\sup_{x \in R_X}\Big|\frac{1}{n}\sum_{i=1}^n Z_i\Big(\int \frac{1}{h^d}K^*\Big(\frac{u-X_i}{h}\Big)I\{u \le x\}\,du - I\{X_i \le x\}\Big)\Big| = o_P(\frac{1}{\sqrt{n}}).$$

Altogether for C_n we have, uniformly with respect to $x \in R_X$, $y \in \mathbb{R}$,

$$C_n(x,y) = f_\varepsilon(y)\frac{1}{n}\sum_{i=1}^n(\varepsilon_i + \frac{y}{2}(\varepsilon_i^2-1))I\{X_i \le x\} + o_P(\frac{1}{\sqrt{n}}).$$

With B_n we proceed similarly to obtain

$$B_n(x,y) = f_\varepsilon(y)\Big(y \int \frac{\hat\sigma_{\hat\vartheta}(u) - \hat\sigma_{\vartheta_0}(u)}{\sigma(u)} I\{u \le x\}\, dF_X(u) + \int \frac{\hat m_{\hat\vartheta}(u) - \hat m_{\vartheta_0}(u)}{\sigma(u)} I\{u \le x\}\, dF_X(u)\Big) + o_P(\frac{1}{\sqrt{n}})$$

by (A3) and the fact that $\sup_x |\hat m_{\hat\vartheta}(x) - \hat m_{\vartheta_0}(x)| = O_P(n^{-1/2})$, $\sup_x |\hat\sigma_{\hat\vartheta}(x) - \hat\sigma_{\vartheta_0}(x)| = O_P(n^{-1/2})$ (see the proof of Proposition S2.2). Now

$$\begin{aligned} \hat m_{\hat\vartheta}(u) - \hat m_{\vartheta_0}(u) &= \frac{1}{nh^d}\sum_{i=1}^n W_{u,n}\Big(\frac{u-X_i}{h}\Big)(\Lambda_{\hat\vartheta}(Y_i) - \Lambda_{\vartheta_0}(Y_i)) \\ &= \frac{1}{nh^d}\sum_{i=1}^n W_{u,n}\Big(\frac{u-X_i}{h}\Big)\dot\Lambda_{\vartheta_0}(Y_i)(\hat\vartheta - \vartheta_0) + r_n(u), \qquad \text{(B.1)} \end{aligned}$$

where

$$\begin{aligned} &\int \frac{r_n(u)}{\sigma(u)} I\{u \le x\}\, dF_X(u) \\ &\le \frac{1}{2}(\hat\vartheta - \vartheta_0)^2 \int \frac{1}{nh^d}\sum_{i=1}^n \Big|W_{u,n}\Big(\frac{u-X_i}{h}\Big)\Big| \sup_{\vartheta:|\vartheta-\vartheta_0|\le\eta} |\ddot\Lambda_\vartheta(Y_i)| \frac{I\{u\le x\}}{\sigma(u)}\, dF_X(u) \\ &= o_P(n^{-1/2}) \end{aligned}$$

by (A5) and (A7). Proceeding similarly to the expansion of C_n we obtain

$$\begin{aligned} &n \int \frac{\hat m_{\hat\vartheta}(u) - \hat m_{\vartheta_0}(u)}{\sigma(u)} I\{u \le x\}\, dF_X(u) \\ &= (\hat\vartheta - \vartheta_0)\frac{1}{n}\sum_{i=1}^n \dot\Lambda_{\vartheta_0}(Y_i)\int \frac{1}{h^d}K^*\Big(\frac{u-X_i}{h}\Big)\frac{I\{u\le x\}}{\sigma(u)}\, dx + o_P(\frac{1}{\sqrt{n}}) \\ &= (\hat\vartheta - \vartheta_0)E\Big[\dot\Lambda_{\vartheta_0}(Y)\frac{I\{X \le x\}}{\sigma(X)}\Big] + o_P(\frac{1}{\sqrt{n}}). \end{aligned}$$

For the variance we have $\hat\sigma_{\hat\vartheta} - \hat\sigma_{\vartheta_0} = (\hat\sigma^2_{\hat\vartheta} - \hat\sigma^2_{\vartheta_0})/(\hat\sigma_{\hat\vartheta} + \hat\sigma_{\vartheta_0})$, which yields (compare to (B.1))

$$\begin{aligned} &\int \frac{\hat\sigma_{\hat\vartheta}(u) - \hat\sigma_{\vartheta_0}(u)}{\sigma(u)} I\{u \le x\}\, dF_X(u) \\ &= \frac{1}{2}\int \frac{1}{\sigma^2(u)}\frac{1}{nh^d}\sum_{i=1}^n W_{u,n}\Big(\frac{u-X_i}{h}\Big)((\Lambda_{\hat\vartheta}(Y_i))^2 - (\Lambda_{\vartheta_0}(Y_i))^2) I\{u \le x\}\, dF_X(u) \\ &\quad + \frac{1}{2}\int \frac{1}{\sigma^2(u)}(\hat m_{\vartheta_0}(u) - \hat m_{\hat\vartheta}(u))(\hat m_{\vartheta_0}(u) + \hat m_{\hat\vartheta}(u)) I\{u \le x\} dF_X(u) + o_P(\frac{1}{\sqrt{n}}) \end{aligned}$$

$$= (\hat{\vartheta} - \vartheta_0)\Bigg(\frac{1}{2n}\sum_{i=1}^{n}\frac{\partial(\Lambda_\vartheta(Y_i))^2}{\partial\vartheta}\Big|_{\vartheta=\vartheta_0}\int\frac{1}{h^d}K^*\Big(\frac{u-X_i}{h}\Big)\frac{I\{u\le x\}}{\sigma^2(u)}\,du$$

$$-\frac{1}{2n}\sum_{i=1}^{n}\dot{\Lambda}_{\vartheta_0}(Y_i)\int\frac{1}{h^d}K^*\Big(\frac{u-X_i}{h}\Big)\frac{I\{u\le x\}}{\sigma^2(u)}2m(u)\,du\Bigg) + o_P(\frac{1}{\sqrt{n}})$$

$$= (\hat{\vartheta} - \vartheta_0)\frac{1}{n}\sum_{i=1}^{n}\Big(\dot{\Lambda}_{\vartheta_0}(Y_i)\Lambda_{\vartheta_0}(Y_i) - \dot{\Lambda}_{\vartheta_0}(Y_i)m(X_i)\Big)\frac{I\{X_i\le x\}}{\sigma^2(X_i)} + o_P(\frac{1}{\sqrt{n}})$$

$$= (\hat{\vartheta} - \vartheta_0)E\Big[\Big(\dot{\Lambda}_{\vartheta_0}(Y)\Lambda_{\vartheta_0}(Y) - \dot{\Lambda}_{\vartheta_0}(Y)m(X)\Big)\frac{I\{X\le x\}}{\sigma^2(X)}\Big] + o_P(\frac{1}{\sqrt{n}}).$$

Those expansions yield, uniformly with respect to x and y,

$$B_n(x,y) = (\hat{\vartheta} - \vartheta_0)f_\varepsilon(y)E\Big[\dot{\Lambda}_{\vartheta_0}(Y)\Big(\sigma(X) + y\Lambda_{\vartheta_0}(Y) - ym(X)\Big)\frac{I\{X\le x\}}{\sigma^2(X)}\Big] + o_P(\frac{1}{\sqrt{n}}).$$

The expansions derived for A_n, B_n, and C_n now yield

$$R_n(x,y) = (\hat{\vartheta}-\vartheta_0)H_{\vartheta_0}(x,y)+f_\varepsilon(y)\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i+\frac{y}{2}(\varepsilon_i^2-1))I\{X_i\le x\} + o_P(\frac{1}{\sqrt{n}}) \tag{B.2}$$

with

$$H_{\vartheta_0}(x,y) = f_\varepsilon(y)E\Big[\dot{\Lambda}_{\vartheta_0}(Y)\Big(\sigma(X) + y\Lambda_{\vartheta_0}(Y) - ym(X)\Big)\frac{I\{X\le x\}}{\sigma^2(X)}\Big]$$
$$+\int f_{Y|X}(V_{\vartheta_0}(y\sigma(u)+m(u))|u)\dot{V}_{\vartheta_0}(y\sigma(u)+m(u))I\{u\le x\}\,dF_X(u)$$
$$= E\Big[\frac{\partial}{\partial\vartheta}F_{\varepsilon(\vartheta)|X}(y|X)\Big|_{\vartheta=\vartheta_0}I\{X\le x\}\Big].$$

The last equality follows by some tedious but straightforward calculations. The assertion of Theorem 2 follows by Lemma B.1, (B.2), and (A5).

B.2. Proof of Corollary 2

From (3.4) we have

$$S_n(x,y) = G_n\Big(x, y, f_\varepsilon(y), yf_\varepsilon(y), h_{\vartheta_0}(x,y)\Big) + o_P(1)$$

uniformly, where

$$h_{\vartheta_0}(x,y) = E\Big[\nabla_\vartheta F_{\varepsilon(\vartheta)|X}(y|X)\Big|_{\vartheta=\vartheta_0}\Big(I\{X\le x\} - F_X(x)\Big)\Big],$$

and where the process

$$\begin{aligned}&G_n(x,y,z_1,z_2,z_3)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\Big(I\{X_i\le x\}-F_X(x)\Big)\Big(I\{\varepsilon_i\le y\}-F_\varepsilon(y)+z_1\varepsilon_i+\frac{z_2}{2}(\varepsilon_i^2-1)\Big)\\&\quad+z_3g_{\vartheta_0}(X_i,Y_i)\Big),\end{aligned}$$

is indexed in $\mathcal{F}=\{(x,y,z_1,z_2,z_3)\mid x\in R_X,y\in\mathbb{R},z_1,z_2,z_3\in[-K,K]\}$ for some K such that $\sup_y f_\varepsilon(y)\le K,\sup_y|yf_\varepsilon(y)|\le K,\sup_{x,y}|h_{\vartheta_0}(x,y)|\le K$ ((A3) and (A4)). Weak convergence of G_n follows similarly to the proof of Theorem 2 in Neumeyer and Van Keilegom (2009). The key argument is that for the bracketing number $N_{[]}(\eta,\mathcal{F},L_2(P))$, an order $O(\eta^{-7})$ can be derived from the $L_2(P)$-norm

$$\begin{aligned}&\Big(E\Big[\Big(\Big(I\{X_i\le x\}-F_X(x)\Big)\Big(I\{\varepsilon_i\le y\}-F_\varepsilon(y)+z_1\varepsilon_i+\frac{z_2}{2}(\varepsilon_i^2-1)\Big)\\&\quad+z_3g_{\vartheta_0}(X_i,Y_i)-\Big(I\{X_i\le x'\}-F_X(x')\Big)\Big(I\{\varepsilon_i\le y'\}-F_\varepsilon(y')+z_1'\varepsilon_i+\frac{z_2'}{2}(\varepsilon_i^2-1)\Big)\\&\quad-z_3'g_{\vartheta_0}(X_i,Y_i)\Big)^2\Big]\Big)^{1/2}\\&\le C\Big(|F_X(x)-F_X(x')|(1+K^2(1+\mathrm{Var}\,(\varepsilon^2)))+|F_\varepsilon(y)-F_\varepsilon(y')|+(z_1-z_1')^2\\&\quad+(z_2-z_2')^2\mathrm{Var}\,(\varepsilon^2)+(z_3-z_3')^2E[g_{\vartheta_0}^2(X,Y)]\Big)^{1/2}\end{aligned}$$

for some constant C. Weak convergence of S_n follows by consideration of the subclass of $\mathcal{F}$ defined by $z_1=f_\varepsilon(y),z_2=yf_\varepsilon(y),z_3=h_{\vartheta_0}(x,y)$.

References

Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549-567.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scand. J. Statist.* **32**, 159-188.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76**, 296-311.

Birke, M. and Neumeyer, N. (2013). Testing monotonicity of regression funtions — an empirical process approach. *Scand. J. Statist.* **40**, 438-454.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26**, 211-252.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall, New York.

Colling, B. and Van Keilegom, I. (2015). Goodness-of-fit tests in semiparametric transformation models. *TEST*, to appear.

Dette, H., von Lieres und Wilkau, C. and Sperlich, S. (2005). A comparison of different nonparametric methods for inference on additive models. *J. Nonparametr. Stat.* **17**, 57-81.

Einmahl, J. H. J. and Van Keilegom, I. (2008). Specification tests in nonparametric regression. *J. Econom.* **143**, 88-102.

Fan, C. and Fine, J. P. (2013). Linear transformation model with parametric covariate transformations. *J. Amer. Statist. Assoc.* **108**, 701-712.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Gijbels, I., Omelka, M. and Veraverbeke, N. (2015). Estimation of a copula when a covariate affects only marginal distributions. *Scand. J. Statist.*, to appear.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *TEST* **22**, 361-411.

Gu, J., Li, Q. and Yang, J.-C. (2014). Multivariate local polynomial kernel estimators: Leading bias and asymptotic distribution. *Econom. Rev.*, to appear.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926-1947.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *J. Statist. Softw.* **27**, URL http://www.jstatsoft.org/v27/i05/.

Heuchenne, C., Samb, R. and Van Keilegom, I. (2015). Estimating the residuals distribution in semiparametric transformation models. *Electr. J. Statist.*, to appear.

Hlávka, Z., Hušková, M. and Meintanis, S. G. (2011). Test for independence in non-parametric heteroscedastic regression models. *J. Multivariate Anal.* **102**, 816-827.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* **64**, 103-137.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York.

Hušková, M. and Meintanis, S. G. (2010). Tests for the error distribution in nonparametric possibly heteroscedastic regression models. *TEST* **19**, 92-112.

Khan, S., Shin, Y. and Tamer, E. (2011). Heteroskedastic transformation models with covariate dependent censoring. Estimation on a semiparametric transformation model. *J. Bus. Econ. Stat.* **29**, 40-48.

Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation on a semiparametric transformation model. *Ann. Statist.* **36**, 686-718.

Masry, E. (1996a). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-599.

Masry, E. (1996b). Multivariate regression estimation - local polynomial fitting for time series. *Stochastic Process. Appl.* **65**, 81-101.

Mu, Y. and He, X. (2007). Power transformation toward a linear regression quantile. *J. Amer. Statist. Assoc.* **102**, 269-279.

Neumeyer, N. (2009a). Testing independence in nonparametric regression. *J. Multivariate Anal.* **100**, 1551-1566.

Neumeyer, N. (2009b). Smooth residual bootstrap for empirical processes of nonparametric regression residuals. *Scand. J. Statist.* **36**, 204-228.

Neumeyer, N. and Sperlich, S. (2006). Comparison of separable components in different samples. *Scand. J. Statist.* **33**, 477-501.

Neumeyer, N. and Van Keilegom, I. (2009). Change-point tests for the error distribution in nonparametric regression. *Scand. J. Statist.* **36**, 518-541.

Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multivariate Anal.* **101**, 1067-1078.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257-1270.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768-1802.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53**, 683-690.

Stute, W., González-Manteiga, W. and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.* **93**, 141-149.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wand, M. (2015). KernSmooth: Functions for kernel smoothing supporting Wand and Jones (1995). R package version 2.23-15. `http://CRAN.R-project.org/package=KernSmooth`.

Yeo, I-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87**, 954-959.

Zhou, X-H., Lin, H. and Johnson, E. (2009). Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *J. Roy. Statist. Soc. Ser. B* **70**, 1029-1047.

Zhu, L., Fujikoshi, Y. and Naito, K. (2001). Heteroscedasticity checks for regression models. *Sci. China A* **44**, 1236-1252.

Department of Mathematics, University of Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany.

E-mail: neumeyer@math.uni-hamburg.de

Department of Statistics, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul, South Korea 140-742.

E-mail: word5810@gmail.com

Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium.

E-mail: ingrid.vankeilegom@uclouvain.be

(Received November 2014; accepted October 2015)