Statistics and Its Interface Volume 0 (2012) 1–6

Testing linearity in semiparametric regression models

Tapio Nummi*, Jianxin Pan and Nicholas Mesue

One of the fundamental assumptions of a basic multiple linear regression model is that the contribution of each of the model terms is strictly linear. In many cases, this may be an excessive simplification of the complicated relationships. Moreover, it may be difficult or impossible to test the hypothesized model against all possible kinds of relevant alternative models. Therefore, tests that perform well under more general circumstances are also required. This paper considers the semiparametric model, where the contribution of one of the model terms may not be strictly linear, and also proposes an exact F-test for the situation. The method also allows dependent error terms. The performance of the proposed test is illustrated by simulation experiments and in real air pollution and health data.

AMS 2000 subject classifications: Primary 62G08, 60K35; secondary 60K35.
Keywords and phrases: Cubic smoothing splines, Fine particles, F-test, Partial linear model.

1. INTRODUCTION

In many scientific fields, it is often necessary to establish a model that accurately describes real-life relationships. Essentially, the modeling challenge can be driven by the underlying task or data, or both. A traditional statistical approach to modeling is based on linear or non-linear regression model and least squares (or maximum likelihood) methods. Generalized linear models can be considered to be a further extension of these methods. The basic feature in these models is that they are based on a strict assumption of a parametric model. In some cases, these approaches may provide a nice presentation of the data at hand. In many cases, however, the true model form is not known and model choice is restricted to rather arbitrary, parametric competitive models. Over the last two decades, nonparametric methods such as Kernel estimation, local polynomials, regression splines, smoothing splines, and generalized additive models have provided a serious alternative to parametric statistical methods (see e.g. [8, 14, 18] and [19]).

The present paper considers an interesting mixture of parametric and nonparametric models. The model considered here is basically a multiple linear regression model with a possible smooth term. The study was primarily interested in testing whether the smooth term could be exchanged to the linear term so that the full multiple regression model could be used instead of a semiparametric model. Statistical inference of semiparametric regression models have been considered by several authors. However, most of the references concentrate on the methods for independent data (see e.g [1, 2, 5] and [7] for linear models and [10] and [11] for generalized linear models). For dependent data, we may refer to [20] who have considered statistical inference under the semiparametric additive model framework. In summary, most of the references rely on the independence assumption and the exact distribution of the test statistics is either not known or is fairly complicated. Our test relies on the exact F-distribution and also applies under correlated observations. The test is computationally very simple and it is not influenced by the correlation of the explanatory variables and the smooth term.

This article is structured as follows. Section 2 presents the basic cubic smoothing spline model and its estimation. Section 3 extends these results to a semiparamatric model and proposes a new F-test based on a semiparametric model. In Section 4 the performance of the proposed test is illustrated by simulation experiments and in real air pollution and health data from the city of Tampere, Finland. Concluding remarks are provided in Section 5.

2. CUBIC SMOOTHING SPLINES

Suppose that $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is the $n \times 1$ vector of observations and that design points (knots) $x_1, x_2, \ldots, x_n$ are given at a certain interval $[a, b]$ satisfying $a < x_1 < x_2 < \cdots < x_n < b$. A cubic smoothing spline can now be written as follows

$$\mathbf{y} = \mathbf{g} + \epsilon, \tag{1}$$

where $\mathbf{g} = (g(x_1), g(x_2), \ldots, g(x_n))'$, $g(.)$ is a smooth twice differentiable curve and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)' \sim N(\mathbf{0}, \sigma^2\mathbf{R})$, where $\mathbf{R}$ is a covariance matrix with parameters θ.

Let $h_j = x_{j+1} - x_j$, $j = 1, 2, \ldots, (n-1)$. Define the non-zero elements of banded $n \times (n-2)$ and $(n-2) \times (n-2)$ matrices ∇ and Δ, respectively, as

$$\nabla_{k,k} = \frac{1}{h_k}, \quad \nabla_{k+1,k} = -\left(\frac{1}{h_k} + \frac{1}{h_{k+1}}\right), \quad \nabla_{k+2,k} = \frac{1}{h_{k+1}}$$

*Corresponding author.

and

$$\Delta_{k,k+1} = \Delta_{k+1,k} = \frac{h_{k+1}}{6}, \quad \Delta_{k,k} = \frac{h_k + h_{k+1}}{3},$$

where $k = 1, 2, \ldots, (n-2)$. If the roughness matrix is denoted as $\mathbf{K} = \nabla \mathbf{\Delta}^{-1} \nabla'$, the solution (at the design points) to the penalized least squares criterion (PLS)

$$Q_1 = (\mathbf{y} - \mathbf{g})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{g}) + \alpha \mathbf{g}' \mathbf{K} \mathbf{g} \tag{2}$$

for fixed positive α is a cubic smoothing spline

$$\tilde{\mathbf{g}} = (\mathbf{H} + \alpha \mathbf{K})^{-1} \mathbf{H} \mathbf{y} = \mathbf{S}_\alpha \mathbf{y}, \tag{3}$$

where it is denoted $\mathbf{R}^{-1} = \mathbf{H}$ and the so-called *smoother matrix* is $\mathbf{S}_\alpha = (\mathbf{H} + \alpha \mathbf{K})^{-1} \mathbf{H}$. If the covariance matrix $\mathbf{R}$ satisfies the equation $\mathbf{K}\mathbf{R} = \mathbf{K}$ then

$$\mathbf{K} = \mathbf{K}\mathbf{H} \tag{4}$$

and the smoother matrix reduces to the simplified form $\mathbf{S}_\alpha = (\mathbf{I} + \alpha \mathbf{K})^{-1}$, which does not explicitly involve covariance matrix $\mathbf{R}$. The simplified estimator is then

$$\hat{\mathbf{g}} = (\mathbf{I} + \alpha \mathbf{K})^{-1} \mathbf{y}. \tag{5}$$

Note that, in this case, spline estimates are simple linear functions of the observations $y_1, y_2, \ldots, y_n$. A set of covariance matrices satisfying the condition (4) can be generated using the formulas discussed in [12] and [13]. As a special case these structures include, for example, $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$, where $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$, $\mathbf{1} = (1, \ldots, 1)'$ and $\mathbf{x} = (x_1, \ldots, x_n)'$. Note that for simplicity, hereinafter we take $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$, but our results also applies under more general class of covariance structures discussed in [13].

Generally, the smoother matrix $\mathbf{S}_\alpha = (\mathbf{I} + \alpha \mathbf{K})^{-1}$ is not a projection matrix; therefore, for example, the theory of linear models is not directly applicable for statistical inference of smoothing splines. However, it is possible to plausibly approximate $\mathbf{S}_\alpha$ using a projection matrix [13]. Let the eigenvalue decomposition of roughness matrix $\mathbf{K}$ be

$$\mathbf{K} = \mathbf{T} \mathbf{\Lambda} \mathbf{T}', \tag{6}$$

where $\mathbf{T}$ is the matrix of n orthonormal eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{K}$. It can easily be seen that

$$\mathbf{S}_\alpha = \mathbf{T}(\mathbf{I} + \alpha \mathbf{\Lambda})^{-1} \mathbf{T}'. \tag{7}$$

Note that the set of eigenvectors of $\mathbf{S}_\alpha$ and $\mathbf{K}$ is the same and the eigenvalues are connected such that the eigenvalues of $\mathbf{S}_\alpha$ are $\gamma = 1/(1 + \alpha\lambda)$. Hereinafter we assume that eigenvectors are ordered according to eigenvalues of $\mathbf{S}_\alpha$. The first two eigenvalues of $\mathbf{S}_\alpha$ are always one and the first two eigenvectors span the subspace corresponding to the straight line model [9]. In general the sequence of eigenvectors $\mathbf{t}_1, \ldots, \mathbf{t}_n$ appears to increase in complexity like a sequence of orthogonal polynomials defined on $\mathbf{x}$. The spline approximation used here is based on the $c > 2$ first eigenvectors of $\mathbf{T}$. This approximation minimizes the least squares criterion

$$(\mathbf{y} - \mathbf{T}_* \xi)' (\mathbf{y} - \mathbf{T}_* \xi),$$

where $\mathbf{T}_*$ is the matrix of c first eigenvectors of $\mathbf{T}$ and the fit is simply

$$\tilde{\mathbf{y}} = \mathbf{P}\mathbf{y},$$

where $\mathbf{P} = \mathbf{T}_* \mathbf{T}_*'$. It was shown in [13] that the approximation is good for relatively smooth data. It is computationally simple (calculated directly from $\mathbf{K}$) and $\mathbf{P}$ is now a projection matrix, which makes the application of the theory of linear models possible in this framework.

3. SEMIPARAMETRIC MODEL

3.1 Estimation

The semiparametric model to be considered here is an extension of the basic spline model of previous section. The model takes the following form

$$\mathbf{y} = \mathbf{U}\mathbf{b} + \mathbf{g} + \epsilon \tag{8}$$

where the extension is the linear part $\mathbf{U}\mathbf{b}$, where $\mathbf{U}$ is a full rank $n \times k$ model matrix of k explanatory variables (constant term not included) and $\mathbf{b}$ is a k-vector of unknown parameters. Estimation of semiparametric models have been considered in [4, 15] and [17] and applied in [16], for example. Our approach is to minimize the penalized sum of squares (see [4])

$$Q_2 = [\mathbf{y} - (\mathbf{U}\mathbf{b} + \mathbf{g})]' \mathbf{H} [\mathbf{y} - (\mathbf{U}\mathbf{b} + \mathbf{g})] + \alpha \mathbf{g}' \mathbf{K} \mathbf{g}. \tag{9}$$

The minimization with respect to $\mathbf{b}$ and $\mathbf{g}$ yields to estimates

$$\tilde{\mathbf{b}} = [\mathbf{U}'\mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}]^{-1} \mathbf{U}'\mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{y} \tag{10}$$

and

$$\tilde{\mathbf{g}} = \mathbf{S}_\alpha (\mathbf{y} - \mathbf{U}\tilde{\mathbf{b}}) \tag{11}$$

where $\mathbf{S}_\alpha = (\mathbf{H} + \alpha \mathbf{K})^{-1} \mathbf{H}$. If the covariance structure is assumed to be, for example $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$, the condition $\mathbf{K} = \mathbf{K}\mathbf{H}$ holds and it can be shown that

$$\mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha) = \mathbf{I} - \mathbf{S}_\alpha$$

(see Appendix A for details). Therefore, the simplified estimators are

$$\hat{\mathbf{b}} = [\mathbf{U}'(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}]^{-1} \mathbf{U}'(\mathbf{I} - \mathbf{S}_\alpha)\mathbf{y} \tag{12}$$

and $\hat{\mathbf{g}} = \mathbf{S}_\alpha (\mathbf{y} - \mathbf{U}\hat{\mathbf{b}})$, where $\mathbf{S}_\alpha = (\mathbf{I} + \alpha \mathbf{K})^{-1}$. The whole semiparametric curve is then fitted by

$$\hat{\mu} = \mathbf{M}\mathbf{y}, \tag{13}$$

where $\mathbf{M} = \mathbf{S}_\alpha + \tilde{\mathbf{U}}[\tilde{\mathbf{U}}'\mathbf{U}]^{-1}\tilde{\mathbf{U}}'$ and $\tilde{\mathbf{U}} = (\mathbf{I} - \mathbf{S}_\alpha)\mathbf{U}$, respectively. Fixing the dimension of the spline approximation (see the next section for details) gives

$$\hat{\mu} = \mathbf{M}_*\mathbf{y}, \tag{14}$$

where $\mathbf{M}_* = \mathbf{P} + \mathbf{U}_*(\mathbf{U}_*'\mathbf{U}_*)^{-1}\mathbf{U}_*'$ and $\mathbf{U}_* = (\mathbf{I} - \mathbf{P})\mathbf{U}$. Note that $\mathbf{M}_*$ is now a projection matrix and thus lays the grounds for the application of the standard theory of linear models.

Using the approximation (14) makes it easy to investigate the statistical properties of the estimates of the linear part of the model. As discussed in the previous section, parameter estimates for $\mathbf{b}$ are then obtained by

$$\hat{\mathbf{b}} = (\mathbf{U}_*'\mathbf{U}_*)^{-1}\mathbf{U}_*'\mathbf{y} \tag{15}$$

and, if $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$ and $\mathbf{U}_*'\mathbf{X} = \mathbf{0}$, it can easily be seen that

$$Var(\hat{\mathbf{b}}) = \sigma^2(\mathbf{U}_*'\mathbf{U}_*)^{-1}. \tag{16}$$

3.2 Testing the semiparametric model

This study aims to test whether our model is the full linear model. The null hypothesis (*linear model*) is as follows

$$\mathbf{H}_0 : \mu = \mathbf{U}_{k+2}\mathbf{b}_{k+2},$$

where $\mathbf{U}_{k+2} = [\mathbf{U}, \mathbf{X}]$, where $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ and the alternative hypothesis (*semiparametric model*) is

$$\mathbf{H}_a : \mu = \mathbf{U}\mathbf{b} + \mathbf{g},$$

where $\mathbf{g}$ is a smooth term. Testing is now based on the fit obtained by the approximation (14) with $\hat{\mu} = \mathbf{M}_*\mathbf{y}$. If we take $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$, for example, $\mathbf{M}_*(\mathbf{X}\mathbf{D}\mathbf{X}') = \mathbf{X}\mathbf{D}\mathbf{X}'$ and it can easily be seen that

$$(\mathbf{I} - \mathbf{M}_*)(\mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I})(\mathbf{I} - \mathbf{M}_*) = \mathbf{I} - \mathbf{M}_*$$

and therefore the distribution of (see also [3])

$$\sigma^{-2}\mathbf{y}'(\mathbf{I} - \mathbf{M}_*)\mathbf{y} = \sigma^{-2}S_{min} \sim \chi^2_{n-c-k},$$

where we denote $\mathbf{y}'(\mathbf{I} - \mathbf{M}_*)\mathbf{y} = S_{min}$. If now in $\mathbf{M}_*$ we define $\mathbf{P} = \mathbf{T}_*\mathbf{T}_*'$ where $\mathbf{T}_* = [\mathbf{t}_1, \mathbf{t}_2]$ (vectors spanning the linear subspace) and denote the projection matrix as $\mathbf{M}_{OLS}$. Then the distribution of

$$\sigma^{-2}\mathbf{y}'(\mathbf{I} - \mathbf{M}_{OLS})\mathbf{y} = \sigma^{-2}S_{OLS} \sim \chi^2_{n-k-2},$$

where we denote $\mathbf{y}'(\mathbf{I} - \mathbf{M}_{OLS})\mathbf{y} = S_{OLS}$, and further

$$\sigma^{-2}(S_{OLS} - S_{min}) \sim \chi^2_{c-2}.$$

Since $\mathbf{M}_*\mathbf{M}_{OLS} = \mathbf{M}_{OLS}\mathbf{M}_* = \mathbf{M}_{OLS}$ it can also be easily observed that $(\mathbf{I} - \mathbf{M}_*)(\mathbf{M}_* - \mathbf{M}_{OLS}) = \mathbf{0}$ and therefore the sum of squares S_{min} and $S_{OLS} - S_{min}$ are independent. Testing can now be based on

$$F = \frac{(S_{OLS} - S_{min})/(c-2)}{S_{min}/(n-c-k)} \sim F(c-2, n-c-k). \tag{17}$$

Then observing the larger value for F than the quantile $F_{1-\alpha}(c-2, n-c-k)$ yields rejection of the null hypothesis of the full linear model. Note that since $\mathbf{P}$ and $\mathbf{U}_*$ in $\mathbf{M}_*$ are orthogonal F is not influenced by the correlation of the explanatory variables $\mathbf{U}$ and the smooth term $\mathbf{g}$. It should also be emphasized that the exact results holds for fixed c and for certain forms of correlation, e.g. $\mathbf{R} = \mathbf{X}\mathbf{D}\mathbf{X}' + \mathbf{I}$. In the next section, empirical power of the proposed test was studied when c is estimated from the data.

4. NUMERICAL EXAMPLES

4.1 A simulation study

To investigate the empirical power of our methods we conducted a simulation study. We first introduce the mixed model formulation

$$\hat{\mathbf{g}} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\xi}$$

of the spline solution (3). The fitted spline is then obtained as the BLUE (best linear unbiased) and BLUP (best linear unbiased predictor) solutions of the mixed model (see [13])

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\xi + \epsilon$$

where $\mathbf{Z} = \nabla(\nabla'\nabla)^{-1}\mathbf{\Delta}^{-1/2}$, ξ and ϵ are independently distributed as $\xi \sim N(\mathbf{0}, \sigma_\xi^2\mathbf{I})$ and $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The smoothing parameter can be viewed as the ratio $\alpha = \sigma^2\sigma_u^2$. If we define the effective degrees of freedom as $edf = tr(\mathbf{S}_\alpha)$ we can solve for fixed values of edf and σ^2 the corresponding value of σ_ξ^2. In our simulations random vectors were generated from the model

$$\mathbf{y} = \mathbf{U}\mathbf{b} + \mathbf{X}\beta + \mathbf{Z}\xi + \epsilon, \tag{18}$$

where $\mathbf{U}\mathbf{b}$ is the linear part and the smooth term is approximated by $\mathbf{X}\beta + \mathbf{Z}\xi$.

For our study 1000 random vectors were generated from (18) for each value of $edf = 2, 4, 6, 8, 10, 15, 20, 25, 30$ with the corresponding σ_ξ^2. We took $\mathbf{X} = (\mathbf{1}, \mathbf{x})$, $\mathbf{x} = (1, 2, \ldots, 100)'$, $\beta = (500, 1)'$ and $\sigma^2 = 15$. The linear part $\mathbf{U}\mathbf{b}$ of the model was defined using the column vector $\mathbf{u} = \mathbf{x} + \mathbf{e}$ with $\mathbf{b} = 1$, where $\mathbf{e}$ is the vector of 100 realizations of independent standard normal variables. This will create highly correlated smooth and linear parts.

Three different methods for the selection of c was used. The first method is to set c equal to edf (and to corresponding σ_ξ^2) used in simulations. The second method is to minimize the information criterion BIC from $\mathbf{y} = \mathbf{U}\mathbf{b} + \mathbf{T}_*\xi + \epsilon$, where $\mathbf{T}_*$ contains the c first eigenvectors of $\mathbf{S}_\alpha$. In the third method we simply fixed $c = 3$.

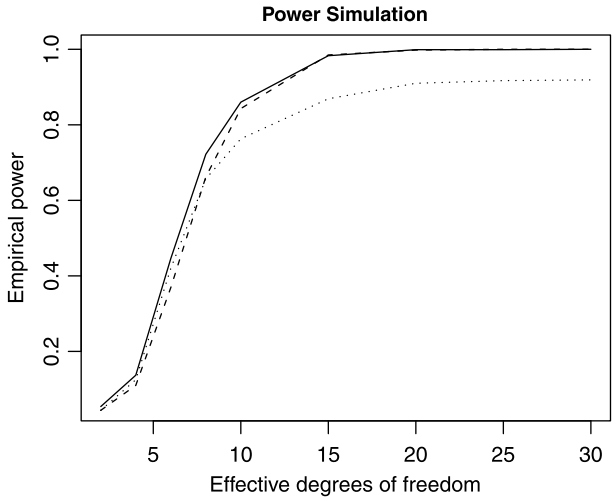

Figure 1. Empirical power as the function of the effective degrees of freedom (dashed line for method 1, solid line for method 2 and dotted line for method 3).

From the Figure 1, it is easily observed that methods 1 and 2 performed practically equally. It seems that basically only minor effects to the empirical power is observed when c is estimated from the data (method 2). These results are also in line with the earlier results by Nummi et al. [13] of the more simple spline model. Also the method 3 performed surprisingly well for low values of $edf < 10$, but for higher values some power is clearly lost.

4.2 Air quality and health in tampere

The example used health and air quality data measured in the medium-sized Finnish city of Tampere daily from January 2006 to August 2008. This data measured air quality variables, weather conditions, traffic density, and health outcomes (such as heart and respiratory diagnosis counts in health centers). This example has concentrated on modeling the fine particle number counts ($< 0,01\,\mu m/cm^3$). Traffic is known to be one of the main sources of air pollution in urban areas. Heating (connected with temperature) is arguably another important source of air pollution, while humidity has also been found to influence fine particle concentrations. Therefore, these variables were included in the model as explanatory variables. Figure 1 plots a logarithm of fine particle counts with temperature. On the basis of Figure 2, it can be argued that the effects of temperature is not linear within the range of observed values, whereas the effects of traffic (mean number of vehicles/h) and humidity (relative %) are assumed to be linear in the present model. The model for the log fine particle counts is as follows

$$\mathbf{y} = \mathbf{U}\mathbf{b} + \mathbf{g} + \epsilon,$$

where the columns of $\mathbf{U}$ consist of measurements for traffic and humidity and $\mathbf{g}$ represents the effects of temperature.

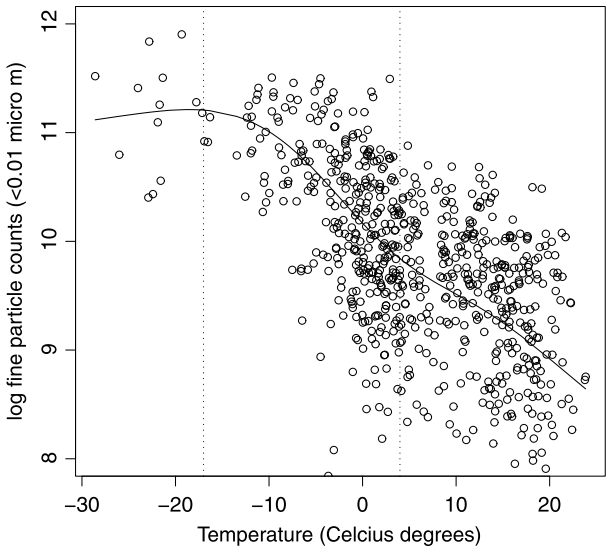

Figure 2. Log fine particle counts as a function of temperature. Note: Some of the original data was excluded due to missing values.

Thus far, we have assumed separate measuring points $x_1, \ldots, x_n$ for the smooth term $\mathbf{g}$ for the model. Unfortunately, the data has multiple measuring times; therefore, instead of the smoother $\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1}$ we used

$$\mathbf{S}_\alpha = \mathbf{N}(\mathbf{N}\mathbf{N}' + \alpha\mathbf{K})^{-1}\mathbf{N}',$$

where $\mathbf{N}$ is an incidence matrix of measuring times. The smoother matrix $\mathbf{S}_\alpha$ was approximated by $\mathbf{T}_*\mathbf{T}_*'$, where the dimension of the approximation was chosen to minimize the BIC criterion. The values of BIC for $c = 1, 2, \ldots, 6$ were: 2518.464, 1205.402, 1211.749, 1207.847, 1205.265 and 1211.364. The minimum is obtained at $c = 5$; therefore, this was chosen as the dimension of the approximation. For this model $S_{min} = 214.3947$ and $S_{OLS} = 220.6037$. This gives

$$F = \frac{6.20904)/3}{214.3947/685} = 6.612715,$$

which is clearly greater than the quantile $F_{0.95}(3, 685) = 2.617906$. Therefore, the null hypothesis of a linear regression model is rejected and the alternative hypothesis about semiparametric regression is accepted.

Figure 2 plots the fitted smooth term of the semiparametric model with linear terms fixed to their mean values. The effects of temperature on the log fine particles counts is clearly not linear (when also other predictors are included). The effects of temperature increase dramatically below approximately 4 Celcius, but no increase in fine particles counts are observed below −18 Celcius. According to the plot of residuals (Figure 3) the model fits the data very well.

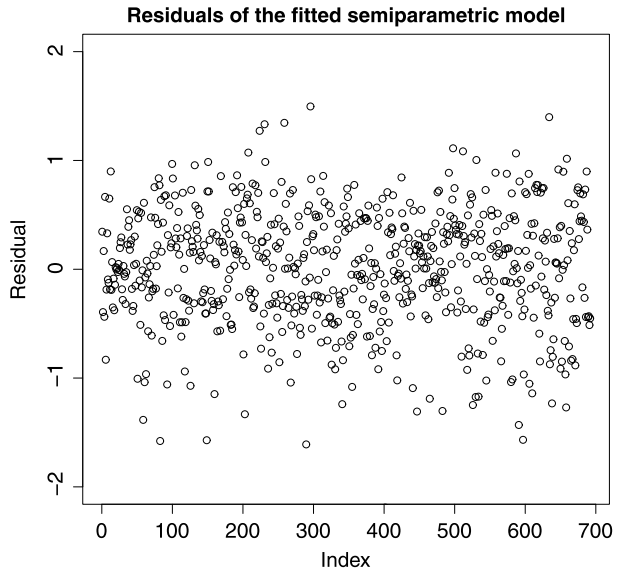

Figure 3. Residuals of the fitted semiparametric model.

5. CONCLUDING REMARKS

This paper has presented a technique that is useful for investigating whether the contribution of a certain covariate in a basic linear model is actually linear or whether it is better to resort to semiparametric modeling. The approach can be useful in situations where the form of the contribution of a covariate is not known (for example, in many medical applications) or if the relation is distorted by outlying observations. The test itself is exact for fixed dimension of the approximation and for certain kind of correlated sequences. However, estimation of the dimension of the approximation seem to have only a minor influence to the empirical power. The test is computationally very simple and it is not influenced by the correlation of the explanatory variables and the smooth term.

APPENDIX A. APPENDIX SECTION

The result is obtained by applying a well known matrix inversion theorem

$$(\mathbf{A}+\mathbf{UBV})^{-1} = \mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{UB}(\mathbf{B}+\mathbf{BVA}^{-1}\mathbf{UB})^{-1}\mathbf{BVA}^{-1}$$

to $\mathbf{H} = \mathbf{R}^{-1} = (\mathbf{XDX}' + \mathbf{I})^{-1}$ and $\mathbf{S}_\alpha = (\mathbf{I} + \alpha\mathbf{K})^{-1} = (\mathbf{I} + \alpha\nabla\boldsymbol{\Delta}^{-1}\nabla')^{-1}$. This yields

$$(\mathbf{I} + \mathbf{XDX}')^{-1} = \mathbf{I} - \mathbf{XD}(\mathbf{D} + \mathbf{DX}'\mathbf{XD})^{-1}\mathbf{DX}'$$

and

$$(\mathbf{I} + \alpha\nabla\boldsymbol{\Delta}^{-1}\nabla')^{-1} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \alpha\Delta^{-1})^{-1}\mathbf{Z}',$$

where $\mathbf{Z} = \nabla(\nabla'\nabla)^{-1}$. Since $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal, we can easily obtain

$$\mathbf{H}(\mathbf{I} - \mathbf{S}_\alpha) = \mathbf{I} - \mathbf{S}_\alpha.$$

ACKNOWLEDGEMENTS

The authors are grateful to environmental inspector Milla Hilli-Lukkarinen from the Environmental Protection Unit at the City of Tampere for providing the data set used in the example. The authors are also grateful to the editors and to the associate editor for the comments that led to the improvements of the paper.

Received 2 December 2011

REFERENCES

[1] Azzalini, A., and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, Series B,* **55**(2) 549–557. MR1224417

[2] Crainiceau, C., Ruppert, D., Claeskens, G., and Wand, M. P. (2005). Exact likelihood ratio tests for penalized splines. *Biometrika*, **92**(1) 91–103. MR2158612

[3] Demidenko, E. (2004). *Mixed Models: Theory and Applications.* Wiley, Hoboken, New Jersey. MR2077875

[4] Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London. MR1270012

[5] Cox, D., and Koh, E. (1989). A smoothing spline based test of model adequacy in polynomial regression. *Annals. of Inst. of Statist. Math.*, **41**(2) 383–400. MR1006497

[6] Eubank, R. L., and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals. of Inst. of Statist. Math.*, **20**(3) 1412–1425. MR1186256

[7] Fan, J., and Huang, L. S. (2001). Goodness-of fit tests for parametric regression models. *Journal of the American Statistical Association*, **96** 640–652. MR1946431

[8] Hastie, T., and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London. MR1082147

[9] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York. MR1851606

[10] Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modelling in generalized linear models. *Journal of the American Statistical Association*, **93** 1461–1474. MR1666641

[11] Liu, A., Meiring, W., and Wang, Y. (2004). Testing generalized linear models using smoothing spline methods. *Statistica Sinica*, **15** 235–256. MR2125730

[12] Nummi, T., and Koskela, L. (2006). Analysis of growth curve data using cubic smoothing splines. *Journal of Applied Statistics*, **35** 681–691. MR2516865

[13] Nummi, T., Pan, J., Siren, T., and Liu, K. (2011). Testing for cubic smoothing splines under dependent data. *Biometrics*, **67**(3) 871–875. DOI: 10.1111/j.1541-0420.2010.01537.x. MR2829261

[14] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, New York, USA. MR1998720

[15] Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, **91** 522–540. MR1814799

[16] Schmalensee, R., and Stoker, T. M. (1999). Household gasoline demand in the United States. *Econometrica*, **67** 645–662.

[17] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50** 413–436. MR0970977

[18] Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, Boca Raton, USA. MR2206355

[19] Wu, H., and Zhang, J. T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley, Hoboken, New Jersey. MR2216899

[20] Zhang, D., and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**(1) 57–74.

Tapio Nummi
School of Health Sciences
FI-33014 University of Tampere
Finland
E-mail address: tapio.nummi@uta.fi

Jianxin Pan
School of Mathematics
The University of Manchester, M13 9PL
UK
E-mail address: jianxin.pan@manchester.ac.uk

Nicholas Mesue
School of Health Sciences
FI-33014 University of Tampere
Finland
E-mail address: mesuenicholas@yahoo.com
url: www.foo.com