View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Bootstrapping for penalized spline regression

b

Göran Kauermann, Gerda Claeskens and Jean D. Opsomer

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0609

Bootstrapping for Penalized Spline Regression[*][†][‡]

Göran Kauermann
Universität Bielefeld

Gerda Claeskens
Katholieke Universiteit Leuven

J. D. Opsomer
Iowa State University

14th February 2006

Abstract

We describe and contrast several different bootstrapping procedures for penalized spline smoothers. The bootstrapping procedures considered are variations on existing methods, developed under two different probabilistic frameworks. Under the first framework, penalized spline regression is considered an estimation technique to find an unknown smooth function. The smooth function is represented in a high dimensional spline basis, with spline coefficients estimated in a penalized form. Under the second framework, the unknown function is treated as a realization of a set of random spline coefficients, which are then predicted in a linear mixed model. We describe how bootstrapping methods can be implemented under both frameworks, and we show in theory and through simulations and examples that bootstrapping provides valid inference in both cases. We compare the inference obtained under both frameworks, and conclude that the latter generally produces better results than the former. The bootstrapping ideas are extended to hypothesis testing, where parametric components in a model are tested against nonparametric alternatives.

[*] Abbreviated title: "Bootstrapping for Penalized Splines."

[†] AMS 1991 subject classifications. Primary-62G08; secondary-62G09.

[‡] Key words: Mixed Model, Nonparametric Regression, Resampling, Nonparametric Hypothesis Testing.

1 Introduction

The objective of nonparametric regression is to model the mean function of a response variable Y by some smooth but otherwise unspecified function $\mu(x)$, with x as continuous covariate. Based on a sample of data pairs (x_i, y_i), $i = 1, \ldots, n$, two important classes of methods for estimating $\mu(x)$ are local approaches (see for instance Fan and Gijbels, 1996) and spline smoothing (see for instance Wahba, 1992 or Eubank, 1999). Both methods can be applied in more complex models like Additive Models (Hastie and Tibshirani, 1990), Varying Coefficient Models (Hastie and Tibshirani, 1993) or in generalized response models (Green and Silverman, 1994 or Bowman and Azzalini, 1997). In recent years, penalized spline regression (often referred to as *P-splines*) has received renewed attention as a powerful alternative smoothing method. Originally suggested by O'Sullivan (1986), the method has been made popular by Eilers and Marx (1996) and more recently through the book by Ruppert, Wand, and Carroll (2003). The main idea of penalized spline regression is to fit the function $\mu(x)$ parametrically with a sufficiently flexible spline basis. Instead of simple parametric estimation, however, a penalty is imposed on the spline coefficients to achieve a smooth fit. One technical benefit of this approach is that it reveals a link to linear mixed models (see Wand, 2003). The resulting affinity to linear mixed models is advantageous and can be exploited in various ways. In particular, the smoothing or penalty parameters are playing the role of a ratio of variances in the mixed model which suggests the application of maximum likelihood theory for estimation (see for instance Kauermann, 2004).

For notational simplicity, we restrict the presentation to the standard smoothing model $Y = \mu(x) + \varepsilon$ with ε as zero mean residuals, even though the examples later in this article mirror more complex models. Estimation of $\mu(x)$ is carried out by penalized spline regression. Under this method, we first replace $\mu(x)$ by the parametric form $X\beta + Zu$, where X is some low dimensional basis, e.g. a line, while Z is high dimensional, e.g. a basis built from truncated line segments. The main assumption is that Z is sufficiently complex and high dimensional, so that the modelling bias $\mu(x) - (X\beta + Zu)$ is of ignorable size compared to the stochastic estimation error. Theoretical results on how large the dimension of the spline basis should be in relation to the sample size are rudimentary, even though Cardot (2002) provides a good starting point. However, it has been found in practice that the actual specification of Z and its dimension has little influence on the fit as long as the dimension of Z is sufficiently large and a penalized fit is pursued. In fact, Ruppert (2002) concludes that "it may be surprising that a default that uses at most 35 or 40 knots [= the dimension of basis Z] could be recommended for effectively all sample sizes and for all smooth regression functions without too many oscillations".

Once a basis is selected, a penalized fit is pursued by imposing a penalty on the spline coefficients u and estimating by least squares regression, which results in a ridge regression estimate. The resulting penalized fit is equivalently achieved by assuming the spline coefficients u to be random, that is formulating an *a priori* distribution on u. This leads to a linear mixed model and the best linear unbiased prediction (BLUP) of u is equivalent to the penalized smooth fit, if the penalty is selected to be equal to the ratio of the variances of ε and u.

Our objective is to develop a bootstrap that takes advantage of the mixed model structure, and to compare it with a bootstrap that treats $\mu(x)$ as fixed and only ε as random. Bootstrapping for such "smoothing models" has a long history, with Härdle and Bowman (1988) and Härdle and Marron (1991) as two important examples. See also Mammen (1993), Härdle, Huet, and Jolivet (1995) or Galindo, Liang, Kauermann, and Carroll (2001) for some extensions. We refer to Shao and Tu (1995) for an overview. A major concern when bootstrapping in smooth models is the bias occurring due to smoothing, which is not accounted for if one applies a naive bootstrap. This requires the use of a pilot estimate with a relatively large smoothing parameter before the actual bootstrapping is pursued (see Härdle and Marron, 1991). Following the discussion in Ruppert, Wand, and Carroll (2003, ch.6), we show here that the bias problem can be circumvented in penalized spline smoothing if a mixed model formulation is used for bootstrapping.

We describe a number of bootstrap versions for both the mixed model and the smoothing model formulations, including simple residual resampling, wild bootstrapping and bootstrapping of correlated spline coefficients. We also show how residuals can be adjusted to compensate for any small sample bias. The adjustment again depends on the model used, that is a smoothing model or a mixed model, respectively. Bootstrapping is employed in our paper for two purposes. First, it serves to mirror estimation variability. That is, we derive bootstrap based confidence bands for our smooth fit. Second, we take advantage of the technique for model validation and model checking. In particular, we use bootstrapping for testing of particular components of the model.

The article is organized as follows. In Section 2, we introduce penalized spline smoothing in the two models considered, i.e. the smoothing model and the linear mixed model. We then suggest two resulting bootstrap procedures. Before providing simulations, we propose some small sample adjustment to improve the performance of the bootstrap routine. The bootstrap is then applied in Section 3 to two data examples making use of additive models. In Section 4 we employ the bootstrap in testing for nonparametric and semiparametric models, which shows the applicability of our suggestions in more complicated regression settings.

2 Penalized Spline Smoothing

2.1 Estimation

We consider the smoothing model

$$y_i|x_i = \mu(x_i) + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ as independent errors. Function $\mu(x)$ is assumed to be smooth but otherwise unspecified. Following the idea of penalized spline smoothing sketched in Section 1, we approximate $\mu(x)$ by $\mu(x_i) = C(x_i)\theta + \delta(x_i)$ where $C(x_i)$ is a high dimensional basis chosen in advance. In this form, $\delta(x)$ denotes the approximation bias of the spline basis in $C(x)$. If $C(x)$ is chosen as a sufficiently flexible basis, $\delta(x)$ does not contain relevant information and will therefore be dropped subsequently. This means we assume the function $\mu(x)$ to be representable by a high dimensional parametric form $C(x)\theta$. It is convenient to decompose $C(x)$ into a low dimensional part X and a high dimensional component Z (see Ruppert, Wand, and Carroll, 2003). For instance $X = (1, x, \dots, x^p)$ can contain a low dimensional polynomial form while Z is a truncated polynomial basis $Z = ((x - \tau_1)_+^p, \dots, (x - \tau_K)_+^p)$, where $(x)_+^p = x^p$ for $x > 0$ and zero otherwise. Following Ruppert (2002), we choose K large but less than the sample size n (or $n - p - 1$). As a practical choice, we suggest $K = \min(n/4, 40)$. Alternatively, one may use the selection routine suggested in Ruppert (2002), but to keep the approach simple we fix K with the above rule of thumb. Once K is chosen, we select the knots τ_k to cover the range of x values using quantiles. This formulation brings us to the parametric model

$$Y|x, u \sim N(X\beta + Zu, \sigma_\varepsilon^2 I) \tag{1}$$

where $Y = (y_1, \ldots, y_n)$ is the vector of response variables and X and Z are the bases vectors built from the observed covariate values $x_1, \ldots, x_n$. We define $\theta = (\beta^T, u^T)$ as parameter for the basis $C = (X, Z)$. The error structure modelled in (1) assumes homogeneity with variance σ_ε^2, even though the bootstrap proposed below will also allow for heterogeneity in the errors.

Simple parametric fitting of θ would lead to unsatisfactory results due to the high dimensionality of C. Instead, θ is estimated in a penalized manner by imposing a penalty on the coefficients in u. This leads to the penalized likelihood criterion

$$l_p(\beta, u, \lambda) = (Y - \mu)^T (Y - \mu) - \lambda u^T \tilde{D} u, \tag{2}$$

with $\mu = C\theta$, λ as penalty parameter steering the amount of smoothness and $\tilde{D}$ as appropriately chosen penalty matrix. For truncated polynomials, it is convenient to chose $\tilde{D}$ as identity matrix (see Ruppert, Wand, and Carroll, 2003), while for a B-spline basis (de Boor, 1978) a difference based penalty is suggested (see Eilers and Marx, 1996). The smooth estimate $\widehat{\mu}$ resulting from (2) then looks like $\widehat{\mu}_\lambda = C\widehat{\theta}_\lambda$, with

$$\widehat{\theta}_\lambda = (C^T C + \lambda D)^{-1} C^T Y \tag{3}$$

where D is a block diagonal matrix built from 0 relating to the unpenalized coefficients β and $\tilde{D}$ relating to u. The coefficient λ acts as a smoothing parameter, which can be chosen by cross validation or using the Akaike criterion, among other methods. For the latter, one minimizes

$$\mathrm{AIC}(\lambda) = \log(Y - \widehat{\mu}_\lambda)^T (Y - \widehat{\mu}_\lambda) + \frac{2\,\mathrm{df}(\lambda)}{n},$$

where $\mathrm{df}(\lambda)$ is the "degrees of freedom" of the fit, commonly chosen as the trace of the smoothing matrix, i.e. $\mathrm{df}(\lambda) = \mathrm{tr}\{(C^T C + \lambda D)^{-1} C^T C\}$.

The penalized estimate in (3) equals a ridge regression estimate with ridging acting on u only. Alternatively, we can motivate the estimator in a different way. Assuming u to be random, one obtains the linear mixed model

$$Y|x, u \sim N(X\beta + Zu, \sigma_\varepsilon^2 I), \quad u \sim N(0, \sigma_u^2 \tilde{D}^-) \tag{4}$$

with $\tilde{D}^-$ as (possibly generalized) inverse of $\tilde{D}$. Under this model, the estimator $\widehat{\mu}_\lambda$ can be interpreted as a posterior Bayes estimator or as best linear unbiased predictor (BLUP) with $\lambda = \sigma_\varepsilon^2/\sigma_u^2$ steering the amount of smoothness. In fact it is easily checked that the BLUP in the linear mixed model (4) is identical to the penalized estimate in the smooth model (1). However, the interpetation of λ is different in the two models. While being a smoothing parameter in the smoothing model, λ is playing the role of a variance ratio in the linear mixed model. In the latter, λ can be estimated together with β using Maximum Likelihood or Restricted Maximum Likelihood (REML, see Harville, 1977) from (4), see e.g. Kauermann (2004).

The objective is now to assess the variability of the estimator $\widehat{\mu}_\lambda$ via bootstrapping. This will be done in two model scenarios. First, we assume that the function $\mu(x) = C\theta$ is unknown and θ is estimated in a penalized form. This corresponds to model (1) and will be subsequently called *smoothing model bootstrap*. Secondly, assuming component u to be random leads to a random function $C\theta$ which is predicted based on data. This is the scenario of model (4) and bootstrapping in this model will be called *mixed model bootstrap*.

2.2 Smoothing Model Bootstrap

We start with smoothing model bootstrapping based on (1). Let λ_p be a smoothing parameter serving as pilot estimate. Then,

$$\widehat{\varepsilon}_p = Y - \widehat{\mu}_p = Y - C(C^TC + \lambda_p D)^{-1}C^TY =: (I - S_\lambda)Y \tag{5}$$

are the resulting residuals. Bootstrapping is now carried out by resampling these residuals with different procedures. One possibility is to employ the estimate $\widehat{\sigma}_\varepsilon^2$ and resample bootstrap errors ε_i^* with replacement from the normal distribution $N(0, \widehat{\sigma}_\varepsilon^2)$. This is usually called *parametric bootstrap* (see e.g. Efron and Tibshirani, 1993). While straightforward to implement, the parametric bootstrap approach is unable to mirror discrepancies from the assumed stochastic model and it is therefore not robust with respect to variance model misspecifications. For this reason, it is commonly recommended to bootstrap errors from the empirical distribution function of the residuals $\widehat{\varepsilon}$. This means we draw the bootstrap errors ε_i^* from $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$ with replacement. In doing so, the distributional assumption of normality is no longer crucial, but homogeneity is assumed since exchangeability of the residuals is requested. We call this bootstrap *residual bootstrap*.

Finally, the homogeneity assumption can be relaxed when working with *wild bootstrap* as introduced in Härdle and Marron (1991). In this case, the ith bootstrap error ε_i^* is drawn from the ith residual $\widehat{\varepsilon}_i$ in the following manner: ε_i^* is drawn from a two point distribution with masses $a_i = \widehat{\varepsilon}_i(1-5^{\frac{1}{2}})/2$ and $b_i = \widehat{\varepsilon}_i(1+5^{\frac{1}{2}})/10$ and sampling probability $P(\varepsilon_i^* = a_i) = (5+5^{\frac{1}{2}})/10$. The rationale of the wild bootstrap is that this method reproduces the first three moments of the original residuals, i.e. $E^*(\varepsilon_i^*) = 0, \quad E^*({\varepsilon_i^*}^2) = \widehat{\varepsilon}_i^{\,2}, \quad E^*({\varepsilon^*}^3) = \widehat{\varepsilon}_i^{\,3}$, where the E^* notation refers to moments taken with respect to the bootstrap distribution. The wild bootstrap is able to better capture local structures like variance heterogeneity, but this flexibility comes at the cost of an increase in variability. In terms of coverage probability, this can lead to undercoverage even if the homoscedastic model is in fact correct. The phenomena is in line with Kauermann and Carroll (2001) and not further explored here.

In this article, we will use both residual and wild bootstrap for penalized spline smoothing. Regardless of the bootstrap used, the corresponding bootstrap observations result from $Y^* = \widehat{\mu} + \varepsilon^*$, and inserting Y^* in (3) leads to bootstrap replicate $\widehat{\mu}_\lambda^*$. One can also choose the smoothing parameter λ according to the bootstrap data Y^*, which makes it possible to take this source of variability into account as well.

We now investigate the bootstrap properties in more depth. Writing $\mu(x) = X\beta + Zu = C\theta$ as before, the deviations between the penalized spline fit and its target can be expressed as

$$\widehat{\mu}_\lambda - \mu = CH_\lambda^{-1}C^T\varepsilon + \text{bias}(\lambda) \tag{6}$$

with $\varepsilon = Y - \mu(x)$ and $H_\lambda = (C^TC + \lambda D)$. The bias term thereby mirrors the traditional smoothing bias resulting as $\text{bias}(\lambda) = -\lambda CH_\lambda^{-1}D\theta$. The corresponding bootstrap version of (6) results by replacing unknown quantities on the right hand side by bootstrap quantities. Hence, let $\widehat{\mu}_p$ be a pilot estimate obtained with bandwidth λ_p (we will say more about the role of λ_p later on). Considering $\widehat{\mu}_p$ as an estimate of μ, we get the bootstrap version of (6) through

$$\widehat{\mu}_\lambda^* - \widehat{\mu}_p = CH_\lambda^{-1}C^T\varepsilon^* + \text{bias}^*(\lambda) \tag{7}$$

where $\text{bias}^*(\lambda)$ simplifies with (6) to

$$\text{bias}^*(\lambda) = -\lambda CH_\lambda^{-1}D\widehat{\theta}_p = \text{bias}(\lambda) + \lambda CH_\lambda^{-1}DH_p^{-1}C^T\varepsilon - \lambda\lambda_p CH_\lambda^{-1}DH_p^{-1}D\theta, \tag{8}$$

with $H_p = (C^TC + \lambda_pD)$.

Since the bootstrap errors ε^* are drawn from the residuals $\widehat{\varepsilon} = Y - \widehat{\mu}_p$, a number of convergence requirements are needed for the bootstrap bias estimator to be valid. As seen from (7), we need ε^* to converge in distribution to ε and $\text{bias}^*(\lambda)$ to converge to $\text{bias}(\lambda)$. For a theoretical investigation we consider the following simple asymptotic scenario. The dimension K of the spline basis is assumed to be large but finite and we assume K to be fixed in advance (see Ruppert, 2002, or Ruppert, Wand, and Carroll, 2003, for a justification of this setting). This scenario allows us to readily derive the asymptotic orders $C^TC = O(n)$ and $(C^TC)^{-1} = O(n^{-1})$, for instance. Moreover, with $(C^TC + \lambda D)^{-1} = O((n+\lambda)^{-1})$ we get from (6) that $\widehat{\mu}_\lambda - \mu = O_p\left(\sqrt{n}/(n+\lambda)\right) + O\left(\lambda/(n+\lambda)\right)$. In particular, $\widehat{\mu}_\lambda$ is $\sqrt{n}$-consistent as long as $\lambda = o(n^{\frac{1}{2}})$, and the bootstrap bias equals

$$\text{bias}^*(\lambda) = \text{bias}(\lambda) + O\left(\frac{\lambda}{(n+\lambda)}\right)\left\{O_p\left(\frac{\sqrt{n}}{n+\lambda_p}\right) + O\left(\frac{\lambda_p}{n+\lambda_p}\right)\right\}. \tag{9}$$

Hence, under the conditions given and with $\lambda_p = o(n^{1/2})$ we ensure the convergence of the bias. It can be shown that the Mean Squared Error based choice of λ has order $O(1)$ (see Kauermann, 2004), so that consistency follows naturally if the smoothing parameter is chosen in a data driven manner, for both pilot and bootstrap versions of λ. In practice and for simplicity we suggest to choose $\lambda = \lambda_p$ which also reduces the numerical effort as the smoothing parameter is selected only once. In principle, however, λ and λ_p can be different.

It remains to investigate convergence of the bootstrap residuals ε^* in (6). This is a standard bootstrap exercise which we solve here by looking at convergence of moments. Note first that $E^*(\varepsilon^*) = 0 = E(\varepsilon)$, where $E^*(\cdot)$ is the expectation with respect to the bootstrap distribution. For the second moment, we obtain for residual bootstrap $E^*(\varepsilon_i^*) = \sum_{j=1}^{n} \widehat{\varepsilon}_j^2/n = \sigma_\varepsilon^2 + O_p(n^{-\frac{1}{2}})$. If wild bootstrapping is pursued, we end up with $E^*(\varepsilon_i^*) = \widehat{\varepsilon}_i^2$ and it is shown later in the article that $E(\widehat{\varepsilon}_i^2) = \sigma_\varepsilon^2 + O(n^{-1})$. Convergence of higher order moments can follow similarly if normality is assumed for ε.

Based on the bootstrap, we can now derive confidence intervals for $\widehat{\mu}_\lambda$ in the conventional way as $[z_l^*, z_u^*] - E^*(\widehat{\mu}_\lambda^* - \widehat{\mu}_\lambda)$ where z_l^* and z_u^* are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap distribution of $\widehat{\mu}^*$, respectively. In practice, one replaces the bootstrap distribution and its expectation E^* by the empirical distribution obtained from repeated simulated bootstrap replicates.

2.3 Mixed Model Bootstrap

The above bootstrap was constructed under the smoothing model, where the unknown function was estimated by penalized least squares regression. Alternatively, we can view the smooth fit as a Posterior Bayes estimate under the mixed model (4). This link is now exploited for the construction of a mixed model bootstrap.

Let $\widehat{\theta}_p = (\widehat{\beta}_p, \widehat{u}_p)$, with $\widehat{u}_p$ as predicted random effects resulting from (3). Accordingly, $\widehat{\varepsilon}$ is the residual as above. In contrast to the smoothing model (1), we now assume the coefficient u to be random, that is we consider the functional form $\mu = X\beta + Zu = X\beta + Z\tilde{D}^{-1/2}v$ with $v \sim N(0, \sigma_\varepsilon^2/\lambda)$ and independent. The stochasticity should be mirrored in the bootstrap and we suggest to draw Y^* via $Y^* = X\widehat{\beta} + Zu^* + \varepsilon^*$ with ε^* and u^* being bootstrapped. As in Section 2.2, there are three different options for bootstrapping both ε^* and u^*. First, a parametric bootstrap can be pursued by drawing u^* from a normal distribution $N(0, \widehat{\sigma}_u^2 \tilde{D}^{-})$ and likewise ε^* from $N(0, \widehat{\sigma}_\varepsilon^2)$. Second, residual bootstrapping can be used by setting $u^* = \tilde{D}^{-1/2}\ v^*$ and drawing v^* from the empirical distribution function of the fitted values $\widehat{v} = \tilde{D}^{1/2}\widehat{u}$. Likewise we draw $\tilde{\varepsilon}^*$ from $\widehat{\varepsilon}$ as above. Finally, we can draw v^* and ε^* using a wild bootstrap from $\widehat{v}$ and $\widehat{\varepsilon}$, respectively. We pursue the latter two options in the following.

Drawing v^* from the fitted values $\widehat{v}$ comes with an additional problem. When resampling errors ε^* one needs the bootstrap mean to be zero, that is $\sum_{i=1}^n \widehat{\varepsilon}_i/n = 0$. This is guaranteed for penalized spline fitting, as can be easily shown. However, a similar property does not hold for the fitted coefficients $\widehat{v}$ so that in order to mirror the mixed model in the bootstrap we have to center the empirical distribution of $\widehat{v}$, that is we draw bootstrap values v^* from $\widehat{v} - \bar{\widehat{v}}$ with $\bar{\widehat{v}}$ as arithmetic mean of $\widehat{v}$. In particular this provides $E^*(v^*) = 0$ for both residual and wild bootstrap.

It should be noted that in the mixed model (4) we are not interested in the random variation of u, but in the prediction of u only, as this builds our predicted fit $\widehat{\mu}$. Our objective is therefore to assess the prediction error

$$\widehat{\mu} - \mu = CH^{-1}C^T\varepsilon - \lambda CH^{-1}D\begin{pmatrix}0\\u\end{pmatrix}$$

with $\mu = X\beta + Zu$ and u considered as random. The corresponding bootstrap version is then

$$\widehat{\mu}_\lambda^* - \mu^* = CH_\lambda^{-1}C^T\varepsilon^* - \lambda CH_\lambda^{-1}D\begin{pmatrix}0\\u^*\end{pmatrix}$$

where $\mu^* = X\widehat{\beta} + Zu^*$ is a random function and $\widehat{\mu}^*_\lambda$ is the resulting fit of Y^*.

Bootstrap convergence is now guaranteed if ε^* and u^* converge in distribution to ε and u, respectively. This convergence could be explored by showing convergence of the moments of ε^* and u^*. Even though this is standard for ε^*, we are faced with a conceptional problem with respect to the asymptotic scenario in the case of the convergence of u^*. The dimension of u is fixed (for fixed K) and replicates are available only by resampling the random function μ. This means that if we consider the function μ as given (but unknown), we treat u as given but unknown and in particular, replicates of u are not available. Hence u^* can not converge in distribution to u in a classical way. It is worth pointing out that this problem holds in the same way in the mixed model (4) when used for smoothing since increasing observations do not provide replicates for u.

Regardless of this conceptual hurdle, we can calculate the bootstrap moments, yielding the first two moments as

$$\begin{aligned} E^*\{\widehat{\mu}^*_\lambda - \mu^*\} &= 0 \\ E^*\left\{(\widehat{\mu}^*_\lambda - \mu^*)^2\right\} &= \lambda^2 \operatorname{diag}\left(C(C^TC + \lambda D)^{-1} D \operatorname{diag}\left(0, \operatorname{Var}^*(u^*)\right) D\, (C^TC + \lambda D)^{-1} C^T\right) \\ &\quad + \operatorname{diag}\left(S_\lambda \operatorname{Var}^*(\varepsilon^*)\, S_\lambda^T\right). \end{aligned} \tag{10}$$

where $(.)^2$ on the left hand side of (10) refers to componentwise squared elements and S_λ as defined in (5). The variances in (10) therefore depend on the bootstrap scheme used. For the residual bootstrap, we find $\operatorname{Var}^*(u^*) = \widehat{\sigma}_u \tilde{D}^-$ and $\operatorname{Var}^*(\varepsilon^*) = \widehat{\sigma}^2_\varepsilon I_n$, where $\widehat{\sigma}^2_u = \left(\widehat{u} - \bar{\widehat{u}}\right)^T \tilde{D} \left(\widehat{u} - \bar{\widehat{u}}\right)/K$ with $\bar{\widehat{u}} = D^{1/2}\, \bar{\widehat{v}}$ and $\widehat{\sigma}^2_\varepsilon = \widehat{\varepsilon}^T\widehat{\varepsilon}/n$. Assuming that $\lambda = \widehat{\sigma}^2_\varepsilon / \widehat{\sigma}^2_u$ and ignoring the centering of v for the moment, we can simplify (10) to $E^*\left\{(\widehat{\mu}^*_\lambda - \mu^*)^2\right\} = \widehat{\sigma}^2_\varepsilon \operatorname{diag}(S_\lambda)$ which mirrors the theoretical findings in Ruppert, Wand, and Carroll (2003, page 190). If in contrast wild bootstrapping is pursued, we find $V^*(u^*) = D^{1/2}\operatorname{diag}\left((\widehat{v} - \bar{\widehat{v}}^2\right) D^{1/2}$ and $\operatorname{Var}^*(\varepsilon^*) = \operatorname{diag}(\widehat{\varepsilon}^2_i)$.

One advantage of the mixed model approach, as also noted in Ruppert, Wand, and Carroll (2003, ch.6), is that the bias due to smoothing in the smoothing model becomes a component of variance by treating u as random. This holds in the same way for the bootstrap. Moreover the variability is increased by the variance of u^*, which takes automatic control of the bias.

2.4 Residual Adjustments for Smoothing Model Bootstrap

In all bootstrap approaches above we draw bootstrap errors ε^* from the residuals $\widehat{\varepsilon}$. Like in other regression contexts, this suffers from a small sample bias since residuals underestimate the true model errors. Therefore, a correction is necessary to provide a reliable performance of the bootstrap.

In the smoothing model (1), we find $\mathrm{E}(\widehat{\varepsilon}_i^2) = \sigma_\varepsilon^2 d_i$ with $d_i = \{(I - S_\lambda)(I - S_\lambda)\}_{ii}$ where subscript ii refers to the ith diagonal element. This suggests replacing $\widehat{\varepsilon}_i$ in the smoothing model bootstrap by

$$\tilde{\varepsilon}_i = \widehat{\varepsilon}_i / \sqrt{d_i} \tag{11}$$

It should be noted that $d_i = 1 + O(n^{-1})$ assuming that $\lambda = o(n^{-1/2})$, so that this adjustment is asymptotically negligible.

Considering now the mixed model bootstrap, where we note that model (4) can be written as $r \sim N(0, \sigma_\varepsilon^2 V_\lambda)$ with $V_\lambda = I + Z\tilde{D}^- Z^T/\lambda$ and $r = Y - X\beta$. The residual for r can be written as

$$\widehat{r} = \left(I - PV_\lambda^{-1}\right) r$$

where $P_\lambda = X(X^T V_\lambda^{-1} X)^{-1} X^T$. This allows to express the fitted spline coefficient $\widehat{u} = (Z^T Z + \lambda\tilde{D})^{-1} Z^T \widehat{r}$ as

$$\widehat{u} = \left(Z^T Z + \lambda\tilde{D}\right)^{-1} Z^T \left(I - P_\lambda V_\lambda^{-1}\right) r.$$

Defining e_j as the jth m dimensional unit vector, we obtain through simple matrix algebra $\mathrm{E}(\widehat{u}_j^2) = \sigma_u^2 c_j$ with

$$c_j = e_j^T \left\{ \tilde{D} Z^T Z \left(Z^T Z + \lambda\tilde{D}\right)^{-1} - \lambda \left(Z^T Z + \lambda\tilde{D}\right)^{-1} Z^T P_\lambda Z \left(Z^T Z + \lambda\tilde{D}\right)^{-1} \right\} e_j$$

for $j = 1, \ldots, m$. Accordingly, the bias in $\widehat{u}_j$ can be corrected by taking $\tilde{u}_j = \widehat{u}_j / \sqrt{c_j}$ for the bootstrap. In the same way, we find $\widehat{\varepsilon} = \widehat{r} - Z\widehat{u}$ as remaining residual. Defining $S_Z = Z(Z^T Z + \lambda\tilde{D})^{-1} Z^T$ and using the fact that $V_\lambda^{-1} = (I - S_Z)$, we find for the second order moment $E(\widehat{\varepsilon}_i^2) = \sigma_\varepsilon^2 q_i$, where

$$q_i = e_i^T \left\{ (I - S_Z) - (I - S_Z) P_\lambda (I - S_Z) \right\} e_i.$$

This in turn suggests to adjust the residuals by $\tilde{\varepsilon}_i = \widehat{\varepsilon}_i/\sqrt{q_i}$ before drawing the mixed model bootstrap.

The above adjustments correct for the small sample size bias in the residuals. In the case of the mixed model bootstrap, an additional source of potential bias comes from the possible model misspecification of the random effect for the coefficients u. We assumed in the mixed model (4) that coefficients u are distributed with correlation matrix $\tilde{D}^{-}$. In case of truncated polynomials, $\tilde{D}^{-}$ is usually set to be the identity matrix for practical convenience, i.e. assuming that the u are i.i.d. However, assuming independence of u in the bootstrap can be inefficient if the true underlying function has a smooth shape.

As an example, we show in Figure 1 (top plot) observations simulated from a sine curve. In order to fit this function by spline regression, we approximate the sine $\mu(x)$ by $X\beta + Zu$ for the right choice of β and u, where $x = (1, x)$ and Z are truncated linear lines $(x - \tau_k)_+$ with $(x)_+ = x$ for $x > 0$ and zero otherwise. The knots are equidistantly distributed over the range of x and the number of knots chosen is 40. The sine shape implies that the coefficient vector u which optimally approximates $\mu(x)$ (in a least squares sense) has adjacent values of u of similar size. This can also be seen from Figure 2 (top plot), where we show the fitted values $\widehat{u}$ as well as the optimal u. The bottom plot in Figure 2 shows the corresponding sample partial autocorrelation function for the elements $(\widehat{u}_1, \widehat{u}_2, \dots, \widehat{u}_{40})$. Autocorrelation is clearly visible.

To mirror this type of correlation within the bootstrap, we can assume that the coefficients in u follow an AR(1) process, which will capture serial correlation between the coefficients. This is achieved by setting $u_l = \rho u_{l-1} + v_l$, with v_l as independent mean zero variables and ρ as autocorrelation. Naturally, more complex correlation structures may also be assumed, but to keep the framework simple we restrict the approach to the AR(1) process here. One can now estimate the autocorrelation parameter ρ from the fitted coefficients. This in turn yields fitted random effect residuals $\widehat{v}_l$, $l = 2, 3, \dots$ obtained from $\widehat{u}_l = \widehat{\rho}\,\widehat{u}_{l-1} + \widehat{v}_l$. Bootstrap errors u^* can now be drawn in the following way. First, a bootstrap sample of u_1^* is drawn, either by wild bootstrapping or setting u_1^* as a random draw from the fitted values $\widehat{u}_l - \bar{\widehat{u}}$ if residual bootstrap is being used. In the next step, we draw v_l^* either with residual or wild bootstrap from $\widehat{v}_l$, $l = 2, 3, \dots$. This in turn leads to replicates u_l^*. Note that dependent on the autocorrelation estimate being used, one might have that $\bar{\widehat{v}}_l \neq 0$ so that some centering might be necessary as well. In practice, due to the construction of an AR(1) process, $\widehat{\widehat{v}}_l$ will be close to zero in particular for a large dimensional basis.

2.5 Simulation

To assess the performance of the proposed routines, we run a small simulation study. We first simulate $n = 200$ observations from the model $\mu(x) + \varepsilon$ with $\mu(x) = 2\sin(\pi x/2)$ and $\varepsilon \sim N(0, 0.25^2)$. One realization from this simulation is shown in Figure 1 (top plot). The covariate x is equidistant on [-2, 2]. For estimation, we use a truncated linear basis with 40 knots equidistantly distributed over the range of x. The smoothing parameter λ_P is chosen using REML, which provides an easy and numerically appealing choice (see also Ruppert, Wand, and Carroll, 2003, p.113). Figure 1 (top plot) shows bootstrap confidence intervals as solid bands for the smoothing model bootstrap and as dotted lines for the mixed model bootstrap for the single realization of the simulated data. Both bootstraps are residual based by resampling from the empirical distribution of $\widehat{\varepsilon}$ and $\widehat{u}$, respectively. In the case of the residual bootstrap, the bias correction discussed in Section 2.2 is also included. The intervals are based on 200 bootstraps, and the bands for both methods are very close to each other.

We now run 200 simulations each with 200 bootstraps to check the coverage probability of the different bootstrap approaches. The mixed model bootstrap of $\widehat{u}$ is carried out in two ways, first by simply resampling u^* from $\widehat{u}$ and secondly by accounting for the correlation structure among $\widehat{u}$ as proposed above. The two lower plots in Figure 1 show the coverage probabilities for bootstrapped confidence bands with nominal coverage level at 95%. It appears that the two versions of the mixed model bootstrap perform slightly better than the smoothing model bootstrap, even though the former exhibits a slight tendency of being too conservative. The smoothing model bootstrap appears to have difficulties at the peaks of the sine curve. In the case of the wild bootstrap, the mixed model bootstrap again performs slightly better. The undercoverage of the smoothing model bootstrap is due to increased variability of the variance estimates and not further explored here (see Kauermann and Carroll, 2001, for an explanation of this phenomenon). For both the residual and the wild bootstrap, incorporating the correlation in the coefficients u does not seem to have a large effect on the coverage probabilities overall.

Next, we explore the effect of the sample size. To do so we simulate data from $\mu(x) = x + \exp(-4x^2)$, as shown in Figure 3 (top plot). The function has locally varying complexity and is therefore challenging for smoothing. The two bottom plots show the coverage probability for the smoothing model bootstrap and the mixed model bootstrap (ignoring in this case any correlation among coefficients u). Increasing the sample size substantially improves the performance of both bootstraps. In particular, for sample size n=400 the coverage of the peak in the middle is clearly better. Overall for both sample sizes, the mixed model bootstrap outperforms the smoothing model bootstrap in this example, by providing simulated coverage probabilities closer to the postulated nominal value.

3 Examples

3.1 Munich Rental Data

To illustrate the bootstrap strategy further, we apply the proposed methods to two real data examples. The first example analyzes data on housing rents (in Euro per squared meter [sqm]) for apartments in the city of Munich, Bavaria, Germany. The data were collected in 2003 as a stratified sample by the city council. The study interviewed 2059 tenants with respect to rent and various other features of their apartments. The data can be downloaded at `www.stat.uni-muenchen.de`. We analyze a subset of the data, namely apartments located in buildings constructed after 1960 and having less than 6 rooms (number of rooms include living room, i.e. 1 room apartment = studio, 2 rooms = 1 bedroom etc.). As further explanatory quantities we consider the continuous covariates x_1: floor space (in sqm), x_2: year of construction and x_3: number of rooms, and the categorical covariates w_1: kitchen ($w_1 = 1$ when apartment is equipped with a kitchen, $w_1 = 0$ otherwise), w_2: location ($w_2 = 1$ if neighborhood is considered "good," $w_2 = 0$ otherwise) and w_3: bath ($w_3 = 1$ if the bathroom is equipped with special features, $w_3 = 0$ otherwise).

With Y denoting the rent per sqm, we consider the additive model $Y = \mu(x, w) + \varepsilon$, where

$$\mu(x, w) = \beta_0 + \mu_1(x_1) + \mu_2(x_2) + \mu_3(x_3) + w_1\beta_1 + w_2\beta_2 + w_3\beta_3.$$

Functions $\mu_l(x_l)$ and $\beta_l, l = 1, 2, 3$, can be estimated by penalized spline regression as follows. By writing $X = (1, X_1, X_2, X_3, w_1, w_2, w_3)$, with X_l as unpenalized part for $\mu_l(x_l)$, and $Z = (Z_1, Z_2, Z_3)$, with Z_l as high dimensional part for fitting $\mu_l(x_l)$, we obtain model (1). For the penalized likelihood (2) we only have to decompose the penalty matrix $\tilde{D}$ to $\mathrm{diag}(\tilde{D}_1, \tilde{D}_2, \tilde{D}_3)$ and attach smoothing parameters $\lambda_1, \lambda_2, \lambda_3$ directly to the corresponding submatrices of $\tilde{D}$. The remaining formulae in the above section are now readily generalized to the additive model fitted here.

The model was fitted using a truncated linear line basis with 10 knots for x_1 and x_2 and 5 knots for x_3, and smoothing parameters were selected by REML; we used wild bootstrapping and bootstrapped uncorrelated spline coefficients for the inference. Figure 4 shows the resulting fits with bootstrap confidence intervals for all estimates, including $\beta_1, \beta_2, \beta_3$. Dotted lines are for mixed model bootstrap, dashed lines are for smoothing model bootstrap; for $\widehat{\beta_l}$ we show mixed model bootstrap only. The confidence intervals for the bootstraps performed under both models are again very close, as was also seen in the simulations above.

Apparently there is a nonlinear effect of floor space with apartments smaller than 50 square meters, say, being increasingly expensive (per sqm). The year of construction has a weak but linear effect with newer houses being more expensive. Moreover, apartments with 2 or 3 rooms are most expensive compared to smaller and larger apartments. The factorial effects β_l all appear to have a positive effect, i.e. for apartments equipped with a kitchen the rent in increased by 70 cents per sqm and likewise for apartments in neighborhood considered as good. The effect of bathrooms with special features is positive but less strong and shows some non-significant behavior based on the mixed model bootstrap. The results look comparable for β_l using the smoothing model bootstrap and are therefore not explicitly shown here.

3.2 USA Phillips Curve Estimation

The Phillips Curve, due to Phillips (1958), is a well established concept in economics. We refer to Chiarella and Flaschel (2000) for a general discussion and motivation. The principal (and simplified) idea is that wage inflation Y depends on the unemployment rate x_1 (after controlling for other quantities). Although Phillips (1958) already discussed a nonlinear relationship between Y and x_1, it has become predominant in economics to work with linear functions only. In this article, we investigate the dependence of wage inflation Y on unemployment rate x_1, inflation x_2 and long term inflation x_3. The latter is also called inflationary climate, the integrated inflation of the last 4 years using a nonparametric approach. See also Flaschel, Kauermann, and Semmler, 2005, for an economic discussion of nonlinearity in this context.

Figure 5, bottom right plot, shows the data for the USA from 1970 onwards. Ignoring the time scale, we embed the data in the Phillips curve context by fitting the model $Y = \mu(x_1, x_2, x_3) + \varepsilon$ with

$$\mu(x_1, x_2, x_3) = \beta_0 + \mu_1(x_1) + \mu_2(x_2) + \mu_3(x_3)$$

and ε as errors. The corresponding fits with residual, wild bootstrap confidence intervals are shown in Figure 5. The fit is obtained with truncated lines with 12 knots and smoothing parameter selected by REML, and inference is based on mixed model bootstrapping with uncorrelated spline coefficients.

There is a slight non-linear shape for unemployment rate, meaning that wages increase less if unemployment is high. Moreover, the effect of the actual inflation is weak and mostly around zero, while long term inflation influences the wage inflation in a sigmoid shape. This means that long term inflation has a roughly linear influence on wage inflation if the long term inflation is in the middle range. For high as well as low long term inflation, the effect on wage inflation is reduced. We will investigate the effects further in a subsequent chapter, where we will test whether the covariate effects are linear or have a non-linear relationship. We again see that both bootstrap approaches give similar confidence intervals.

4 Testing Models using the Bootstrapping

4.1 Testing Parametric versus Nonparametric Models

An important area where bootstrapping can be of practical help is when the focus is on testing different models. In this case, a bootstrap makes it possible to mimic the distribution of a test statistic under the hypothetical model. Assume for instance that we want to test the parametric model $H_{(0)} : \ Y = X\beta + \varepsilon$ against the smooth model $H_{(1)} :$ $Y = X\beta + Zu + \varepsilon$. In the context of smoothing and mixed models this problem has been recently tackled in a series of papers by Crainiceanu and Ruppert (2004), Claeskens (2004) and Crainiceanu, Ruppert, Claeskens, and Wand (2005). The theoretical results derived there rely, among other things, on the assumption of independent homoscedastic errors ε. A test that does not rely on this assumption can be constructed via bootstrapping.

As test statistic for model testing, we take the likelihood ratio with model $H_{(0)}$ defined through $Y|x \sim N(X\beta, \sigma_\varepsilon^2 I)$ and alternative model either the smoothing model (1) or the mixed model (4), respectively. Let $\widehat{\sigma}^2_{(l)}$ denote the error variance estimate under $H_{(l)}$ for $l = 0, 1$. The log likelihood ratio in the smoothing model is then defined through

$$\Lambda_{smooth} = \frac{n}{2} \left\{ -\log(\widehat{\sigma}^2_{(1)}) + \log(\widehat{\sigma}^2_{(0)}) \right\} \tag{12}$$

with $\widehat{\sigma}^2_{(l)} = \widehat{\varepsilon}^T_{(l)} \widehat{\varepsilon}_{(l)}/n$, where $\widehat{\varepsilon}_{(l)}$ are the residuals obtained for model $H_{(l)}$. For the mixed model, we modify the likelihood ratio by employing the Restricted Maximum Likelihood (REML) defined as (see also Harville, 1977)

$$l_{REML}(\sigma_\varepsilon, \lambda) = -\frac{(n-p)}{2} \log(\widehat{\sigma}^2_{\varepsilon,mixed}) - \frac{1}{2} \log |V_\lambda| - \frac{1}{2} \log |X^T V_\lambda^{-1} X|, \tag{13}$$

where $V_\lambda = I + Z\tilde{D}^- Z^T/\lambda$ with $\tilde{D}^-$ as (generalized) inverse of $\tilde{D}$ and $\widehat{\sigma}^2_{\varepsilon,mixed} = (Y - X\widehat{\beta})^T V_\lambda^{-1} (Y - X\widehat{\beta})/(n-p)$. The log RE likelihood ratio is then defined as

$$\Lambda_{REML} = l_{REML}(\widehat{\sigma}^2_{(1),mixed}, \widehat{\lambda}) - l_{REML}(\widehat{\sigma}^2_{(0),mixed}, \lambda = \infty).$$

Note that in the case $\sigma_u^2 = 0$ (or equivalently, $\lambda = \infty$), the mixed model (4) collapses to the simple regression model $H_{(0)}$.

The distribution of Λ_{smooth} and Λ_{REML} under $H_{(0)}$ are difficult to derive analytically in general. We therefore derive it by bootstrapping. To do so, we have to bootstrap data from the $H_{(0)}$ model and refit the models either using the smooth or the mixed model as alternative $H_{(1)}$. Apparently, there is no random effect u in the $H_{(0)}$ model so that we only have to resample residuals, either with residual or wild bootstrapping. Refitting models $H_{(0)}$ and $H_{(1)}$ now provides bootstrap replicates Λ^*_{smooth} and Λ^*_{REML}. If the penalty parameter λ is large, coefficients u are shrunk to zero and there is no evidence for $H_{(1)}$. In fact, as shown in Crainiceanu and Ruppert (2004) ,the probability that the REML estimate $\widehat{\lambda}$ is infinity can exceed 1/2. Therefore in each bootstrap we chose the smoothing parameter λ using a REML estimate to incorporate the variability of estimating λ in bootstrap as well. A test decision can be based on the empirical distribution of Λ^*_{smooth} and Λ^*_{REML}, respectively. In practice, if the bootstrap p-value is at the borderline, one should increase the bootstrap size to guarantee reliable results.

We run a small simulation to show the performance of the routine. First we simulate 200 data points from a simple linear model as shown in Figure 6 (left hand side) for one realization. We fit both alternative and assumed model and assess the significance of the likelihood ratio statistic by smoothing model and mixed model wild bootstrap of size of 200. For 150 simulations, we show in Figure 7 (upper row) the empirical distribution of the resulting bootstrap p-values for both Λ_{smooth} (solid line) and Λ_{REML} (dotted line). Since we simulated from model $H_{(0)}$, this should have a uniform distribution. The large proportion of p-values equal to 1 results from the bootstrap samples where $\widehat{\lambda} \to \infty$ and hence model $H_{(1)}$ collapsed to $H_{(0)}$ (we choose $\widehat{\lambda} > 10^5$ as threshold in the simulation). Focussing on the lower part of the distribution (zoom in plot on the right hand side) we see that the test appears to be consistent since small p-values approximately follow a uniform distribution, with the diagonal line included as reference in the plot.

To assess the power of the test, we simulate data from a quadratic model $y = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon$ with parameter settings $\beta_0 = 0, \beta_1 = 2, \beta_2 = 0.1$. A plot of the data is shown in Figure 6 (right hand side) for one realization. The quadratic shape appears quite weak. The resulting distribution of the p-values is provided in Figure 7 (bottom row).

For comparison, we also include a parametric test of a simple linear model tested against a quadratic model for both simulated cases, i.e. $H_{(0)}: \; y = \beta_0 + x\beta_1 + \varepsilon$ against $H_{(1)}: y = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon$ using a likelihood ration test. Because both models are correctly specified, this test serves as benchmark. The resulting simulated p-values are also included in Figure 6 as dashed line. As can be seen from the plots for data following a linear model and a quadratic model, the bootstrap test behaves soundly by showing a promising power compared to the parametric test. No obvious difference between smooth and mixed model bootstrap is observable for this example. This concurrence has also been observed in other simulations which are not reported here.

4.2 Bootstrapping in Additive and Varying Coefficient Models

The above test situation is somewhat simplistic because, since the hypothetical model is parametric, we did not actually need to use the bootstrap ideas introduced in Section 2 to construct a test. However, both the bootstrapping and the testing ideas we described can be easily extended to more complex models such as Additive or Varying Coefficient Models. As an example of a more complex testing situation, consider the model

$$Y = \beta_0 + \mu_1(x_1) + g\mu_2(x_2) + \varepsilon, \tag{14}$$

where β_0 is the intercept, $\mu_1(.)$ and $\mu_2(.)$ are smooth but unknown functions in x_1 and x_2, and g is a factorial covariate (with binary outcome). If $x_1 \equiv x_2 \equiv x$ then $\mu_2(x)$ describes the multiplicative interaction between g and x, introduced as varying coefficient in Hastie and Tibshirani (1993). If $g \equiv 1$ and x_1 and x_2 are two different covariates, then (14) is better known as Additive Model, extensively discussed in Hastie and Tibshirani (1990). Estimation in (14) can be carried out similarly to our examples above by penalized spline smoothing (see also Marx and Eilers, 1998) by replacing $\mu_l(x)$ by $X_l\beta_l + Z_l u_l$ with X_l as low and Z_l as high dimensional basis. The matrix X_1 does not contain the intercept, since this is explicitly written as β_0 in (14). The same holds for X_2 in the Additive Model. Defining $\theta = (\beta_0, \beta_1, \beta_2, u_1, u_2)$ and $C = (1, X_1, GX_2, Z_1, GZ_2)$ with $G = \text{diag}(g_1, \ldots, g_n)$, we get the penalized fit by $\widehat{\theta}-\theta = H(\lambda)^{-1}C^T\varepsilon - H(\lambda)^{-1}D(\lambda)\theta$, where $H(\lambda) = C^TC + D(\lambda)$ and $D(\lambda) = \text{diag}(0, \lambda_1\tilde{D}_1, \lambda_2\tilde{D}_2)$ for $\lambda = (\lambda_1, \lambda_2)$. Parameter θ is thereby either considered as fixed but unknown, mirroring a model with smooth components (1), or components u_1 and u_2 in θ are treated as random, extending the mixed model (4).

As an example, we present a test on checking an additive model, that is $g = 0$ versus $g = 1$, and x_1 and x_2 as two covariates. For the subsequent simulation we draw x_1 and x_2 independently from a truncated standard normal distribution with support $[-2, 2]$. The shapes of $\mu_1(x_1)$ and $\mu_2(x_2)$ are shown in Figure 8 (top plots) where we show $Y - \mu_2(x_2) = \mu_1(x_1) + \varepsilon$ and $Y - \mu_1(x_1) = \mu_2(x_2) + \varepsilon$, respectively. The error variance is set to 1. We now test hypothesis

$$H_{(0)} : Y = \beta_0 + \mu_1(x_1) \;\; \text{against} \;\; H_{(1)} : Y = \beta_0 + \mu_1(x_1) + \mu_2(x_2).$$

To do so we fit $\mu_l(x_l)$ as $Z_l u_l$ only, i.e. we drop X_l and keep Z_l as truncated linear lines. Hence, when penalized regression is carried out with $\lambda_2 \to \infty$, the fit will correspond to that for the hypothetical model $H_{(0)}$. Let the restricted likelihood function be defined as in (13) with $V_\lambda = I + Z\tilde{\Sigma}_u Z^T$ where $Z = (Z_1, Z_2)$ and $\tilde{\Sigma}_u$ as block diagonal matrix built from $\tilde{D}_1^-/\lambda_1$ and $\tilde{D}_2^-/\lambda_2$. Then setting λ_2 to infinity provides the log REstricted likelihood ratio

$$\Lambda_{REML} = l_{REML}\left(\widehat{\sigma}_{(1)}, (\widehat{\lambda}_1, \widehat{\lambda}_2)\right) - l_{REML}\left(\widehat{\sigma}_{(0)}, (\widehat{\lambda}_1, \widehat{\lambda}_2 = \infty)\right)$$

with $\widehat{\sigma}_{(l)}$ as variance estimates in the resulting model. Correspondingly, the smooth likelihood ratio Λ_{smooth} is defined as in (12). Bootstrapping of the test statistic can now be pursued by either following the mixed model or the smoothing model scenario, respectively. For the former, we sample $Y^* = \widehat{\beta}_0 + Z_1 u_1^* + \varepsilon^*$ where ε^* is drawn from the (adjusted) residuals $\widehat{\varepsilon}_{(1)}$ in the alternative model and u_1^* is drawn from the fitted random effect in the mixed model with u_1 and u_2 as random components. Note that the bootstrap has to be constructed from the fitted values in the $H_{(1)}$ model in order to avoid bias problems occurring due to model misspecifications. Fitting the test statistic Λ_{REML} to the bootstrapped values Y^* provides bootstrap replicates Λ^*_{REML} for the likelihood ratio. To accomplish the variability due to estimation of λ_l we refit $\widehat{\lambda}_l$ for each bootstrap sample selected by its $REML$ estimate.

For the bootstrap based on the smoothing model, we sample $Y^* = W\widehat{\beta} + Z_1\widehat{u}_1 + \varepsilon^*$. Using Y^* to refit the model leads to the bootstrap replicate Λ^*_{Smooth} which is used for validation of the significance of Λ_{smooth}. In this case, ε^* is drawn from the fitted smooth model $H_{(1)}$, while $\widehat{\beta}$ and $\widehat{u}$ are the estimates in the $H_{(0)}$ model. In Figure 8 (bottom plots) we show the distribution of the p-value for simulations under H_0 and under H_1, respectively. The results are based on 100 simulations and $\lambda = (\lambda_1, \lambda_2)$ is estimated using a REML approach. If $\widehat{\lambda}_l > 10^5$, we formally set $\lambda_l \equiv \infty$ and fitted a reduced model with the component excluded. Again, the performance of the bootstrap-based test appears sound and no obvious differences between the smoothing model or mixed model approach can be seen.

5 Examples

5.1 Munich Rental Data

Returning to the Munich Rental data fits in Figure 4, we intend to simplify the model given the linear shape of x_2: year of construction. Moreover, we could test on the significance of w_3: bath, given the bootstrap confidence intervals in the full model contain the zero. We therefore test the following simplified models

$$\begin{array}{lcl} H_0 & : & \mu(x_1) + \mu(x_2) + \mu(x_3) + w_1\beta_1 + w_2\beta_2 + w_3\beta_3 \\ & \text{vs.} & \\ H_{11} & : & \mu(x_1) + \mu(x_2) + x_3\beta_{x_3} + w_1\beta_1 + w_2\beta_2 + w_3\beta_3 \\ H_{12} & : & \mu(x_1) + \mu(x_2) + x_3\beta_{x_3} + w_1\beta_1 + w_2\beta_2. \end{array}$$

The resulting bootstrapped p-values for testing H_{11} against H_0 are 0.22 using the smoothing model bootstrap and 0.19 using the mixed model bootstrap. Testing H_{12} against H_0 we get 0.185 as smoothing model p-value and 0.13 as mixed model p-value, respectively. This suggests that there is no evidence for a non-linear influence of year of construction and special features of the bath do not increase the rent significantly. All other components are significant, with p-values not reported here.

Considering the functional shapes in Figure 5 for the other example data set, we test whether the influence of some of the covariates can be simplified in a linear shape. We therefore pursue a bootstrap test for the following models

$$\begin{aligned} H_0 &: \quad \mu_1(x_1) + \mu_2(x_2) + \mu_3(x_3) \\ &\text{vs.} \\ H_{11} &: \quad x_1\beta_1 + \mu_2(x_2) + \mu_3(x_3) \\ H_{12} &: \quad x_1\beta_1 + x_2\beta_2 + \mu_3(x_3) \\ H_{13} &: \quad x_1\beta_1 + \mu_2(x_2) + x_3\beta_3. \end{aligned}$$

The resulting p-values are shown in Table 1. There is clear evidence for non-linear influence of both long and short term inflation while the unemployment rate has a linear relationship on wage inflation, so that model H_{11} can be used as a final model.

6 Conclusions

We have demonstrated how the link between penalized spline smoothing and linear mixed models can not only be exploited for smoothing but also for bootstrapping. As could be seen in our simulations and examples, the mixed model bootstrap works satisfactory when applied to assess the fit of a smooth function using mixed model confidence bands. For the calculation of confidence intervals, the mixed model formulation provides a better framework for bootstrapping than the traditional smoothing model.

The idea was extended to testing nested models. In this hypothesis testing context, the behavior of the mixed model and smoothing model bootstrap methods appeared to be more similar, with both approaches giving good results. The resulting test appears consistent and powerful at the same time.

References

Bowman, A. W. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. Oxford University Press.

Cardot, H. (2002). Local roughness penalities for regression splines. *Computational Statistics 17*, 89–102.

Chiarella, C. and P. Flaschel (2000). *The Dynamics of Keynesian Monetary Growth: Macro Foundations*. Cambridge, UK: Cambridge University Press.

Claeskens, G. (2004). Restricted likelihood ratio lack of fit tests using mixed spline models. *Journal of the Royal Statistical Society, Series B*, (to appear).

Crainiceanu, C. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B 66*, 165–185.

Crainiceanu, C., D. Ruppert, G. Claeskens, and M. Wand (2005). Exact likelihood ratio tests for penalized splines. *Biometrika 92(1)*, to appear.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrapa*. Chapman & Hall.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11*(2), 89–121.

Eubank, R. (1999). *Nonparametric regression and spline smoothing*. New York: Dekker.

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

Flaschel, P., G. Kauermann, and W. Semmler (2005). Non-parametric testing of wage and price phillips curves for the united states. *Metroeconomica*, (under revision).

Galindo, C. D., H. Liang, G. Kauermann, and R. J. Carroll (2001). Bootstrap confidence intervals for local likelihood, local estimation equations and varying coefficient models. *Statistica Sinica 11*(1), 121–134.

Green, D. J. and B. W. Silverman (1994). *Nonparametric Regression and generalized linear models*. Chapman & Hall.

Härdle, W. and A. W. Bowman (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association. 83*, 102–110.

Härdle, W., S. Huet, and E. Jolivet (1995). Better bootstrap confidence intervals for regression curve estimation. *Statistics 26*(4), 287–306.

Härdle, W. and J. S. Marron (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics 19*(2), 778–796.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association. 72*, 320–338.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T. and R. Tibshirani (1993). Varying–coefficient models. *Journal of the Royal Statistical Society, Series B 55*, 757–796.

Kauermann, G. (2004). A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planing and Inference 127*, 53–69.

Kauermann, G. and R. Carroll (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association 96*, 1387–1396.

Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *The Annals of Statistics 21*(1), 255–285.

Marx, B. D. and P. H. C. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis 28*, 193–209.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science 1*, 502–518.

Phillips, A. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957. *Economica 25*, 283–299.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*, 735–757.

Ruppert, R., M. Wand, and R. Carroll (2003). *Semiparametric Regression.* Cambridge University Press.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. New York: Springer Verlag.

Wahba, G. (1992). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics 18*, 223–249.

	model		
bootstrap	H_{11} versus H_0	H_{12} versus H_0	H_{13} versus H_0
smoothing model	0.31	0.01	< 0.01
mixed model	0.34	0.01	< 0.01

Table 1: Bootstrap p-values for Phillips curve data

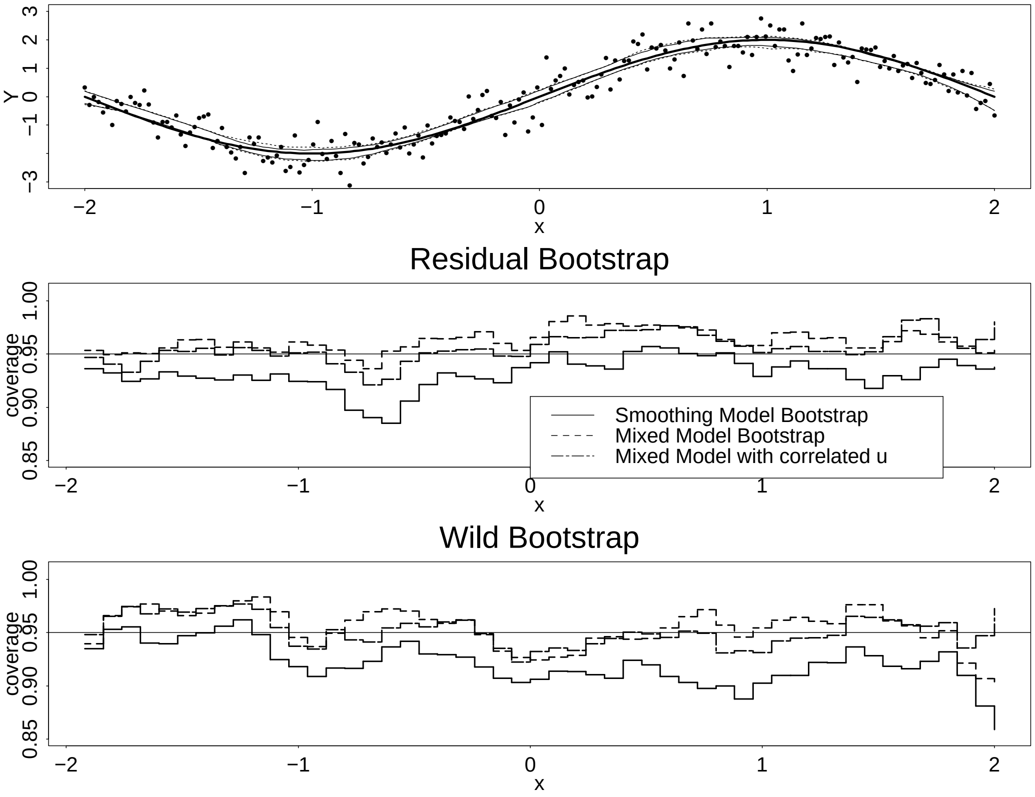

Figure 1: Simulated data from a sine curve (top plot) with bootstrap confidence bands. Bold line shows time curve. Coverage probability based on 200 simulations using residual bootstrapping (middle plot) and wild bootstrapping (bottom plot).

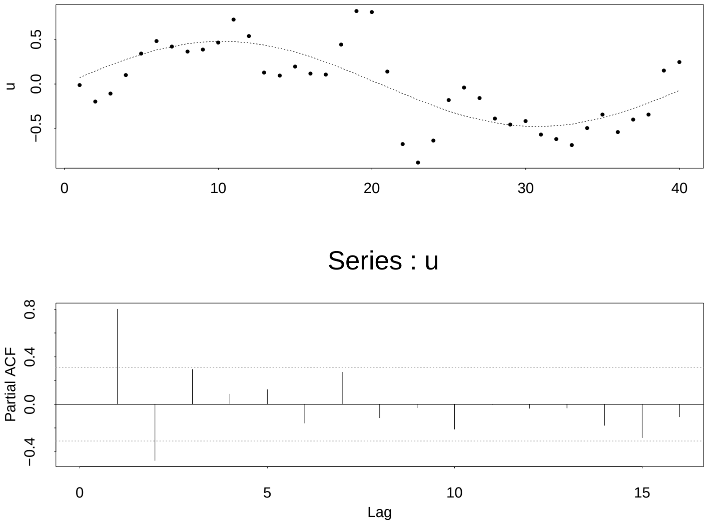

Figure 2: Upper plot shows fitted spline coefficients $\widehat{u}$ with true values u (shown as dotted line). Bottom plot gives the partial autocorrelation function of $\widehat{u}$.

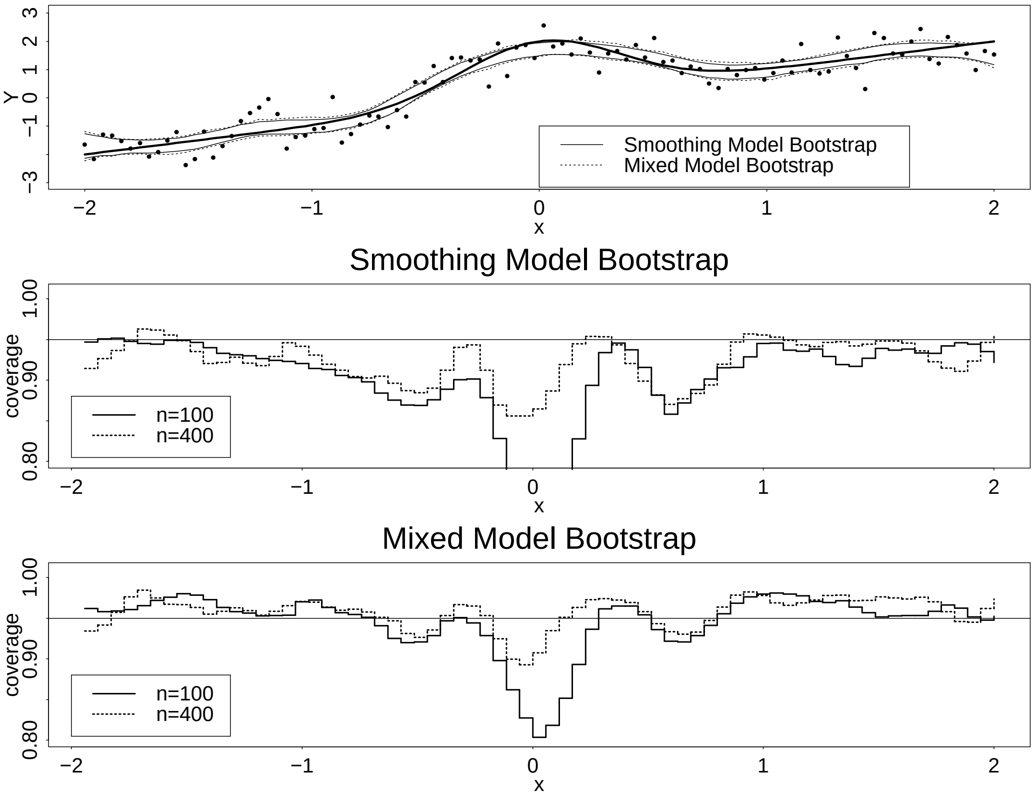

Figure 3: Simulated data with smoothing and mixed model confidence bands (top plot). Bold line shows true curve. Coverage probability based on 200 simulations using a sample size of $n = 100$ and $n = 400$ (two bottom plots).

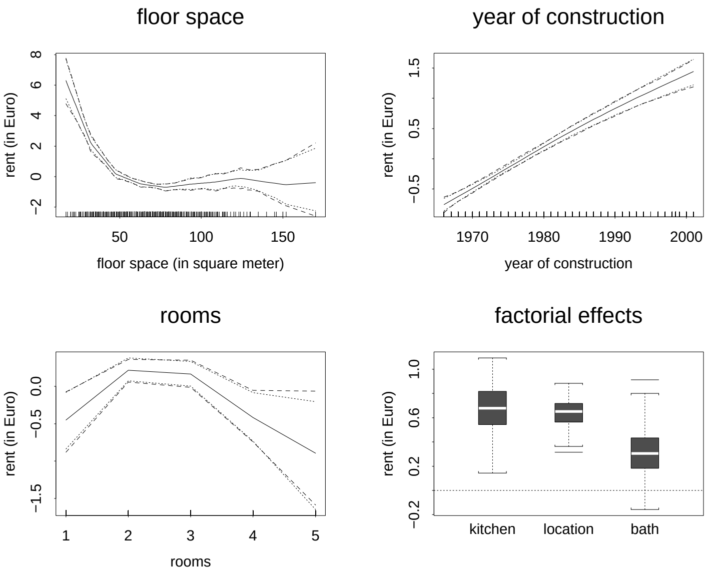

Figure 4: Smooth and parametric effects for Munich rental data. In first three plots, solid lines denote estimated curve, dashed lines are pointwise smoothing model wild bootstrap 95% confidence interval, dotted lines are pointwise mixed model wild bootstrap 95% confidence interval. In fourth plot, boxplots for parametric effects are obtained by mixed model wild bootstrap.

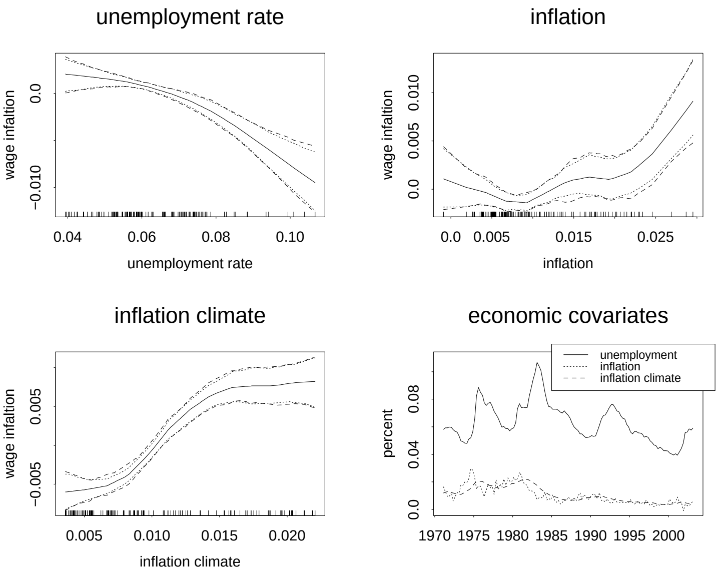

Figure 5: Phillips curves showing the influence on wage inflation for USA (first three plots) and distribution of covariates (bottom right plot). In first three plots, solid lines denote estimated curve, dashed lines are pointwise smoothing model wild bootstrap 95% confidence interval, dotted lines are pointwise mixed model wild bootstrap 95% confidence interval.

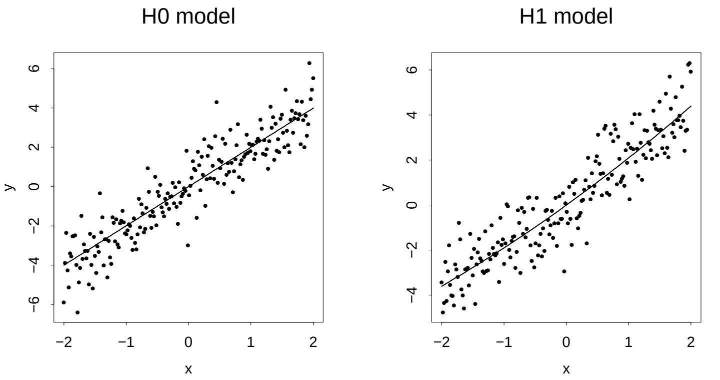

Figure 6: Simulated data from hypothetical (linear) and alternative (quadratic) model.

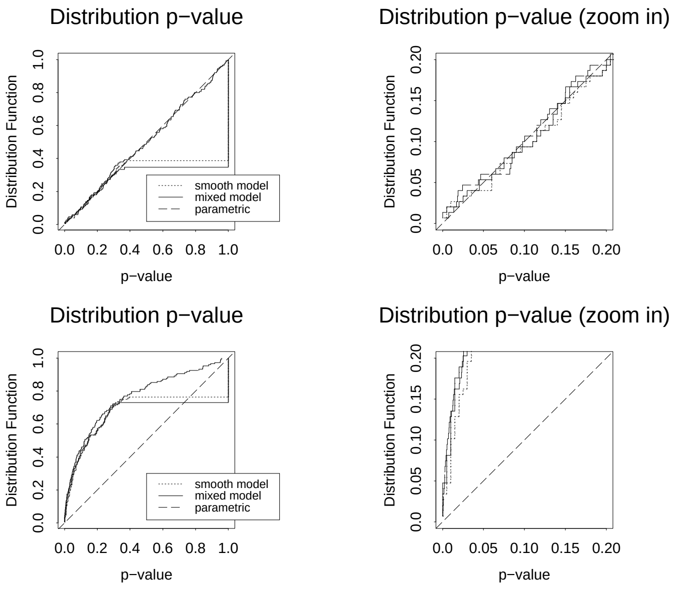

Figure 7: Empirical distribution function of bootstrap p-values under $H_{(0)}$ (upper row) and under the alternative model $H_{(1)}$ (bottom row) from Figure 6.

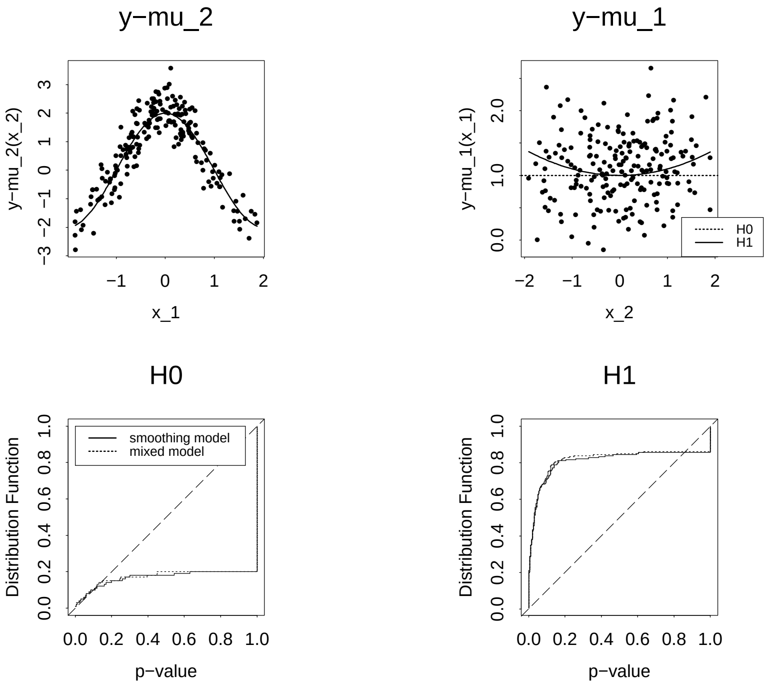

Figure 8: Additive functions in model H_0 and H_1 (top row) and simulated distribution of the p-value.