Computational Statistics and Data Analysis 55 (2011) 1540–1551

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Comparing penalized splines and fractional polynomials for flexible modelling of the effects of continuous predictor variables

Alexander M. Strasak[a], Nikolaus Umlauf[b], Ruth M. Pfeiffer[c], Stefan Lang[b,*]

[a] *Innsbruck Medical University, Schöpfstr. 41, A-6020 Innsbruck, Austria*
[b] *University of Innsbruck, Universitätsstr. 15, A-6020 Innsbruck, Austria*
[c] *National Cancer Institute, 6120 Executive Blvd, Bethesda, MD, 20892-7244, USA*

ARTICLE INFO

Article history:
Received 28 May 2010
Received in revised form 14 October 2010
Accepted 14 October 2010
Available online 21 October 2010

Keywords:
Generalized additive models
Simulation study
Smoothing

ABSTRACT

P(enalized)-splines and fractional polynomials (FPs) have emerged as powerful smoothing techniques with increasing popularity in applied research. Both approaches provide considerable flexibility, but only limited comparative evaluations of the performance and properties of the two methods have been conducted to date. Extensive simulations are performed to compare FPs of degree 2 (FP2) and degree 4 (FP4) and two variants of *P*-splines that used generalized cross validation (GCV) and restricted maximum likelihood (REML) for smoothing parameter selection. The ability of *P*-splines and FPs to recover the "true" functional form of the association between continuous, binary and survival outcomes and exposure for linear, quadratic and more complex, non-linear functions, using different sample sizes and signal to noise ratios is evaluated. For more curved functions FP2, the current default setting in implementations for fitting FPs in R, STATA and SAS, showed considerable bias and consistently higher mean squared error (MSE) compared to spline-based estimators and FP4, that performed equally well in most simulation settings. FPs however, are prone to artefacts due to the specific choice of the origin, while *P*-splines based on GCV reveal sometimes wiggly estimates in particular for small sample sizes. Application to a real dataset illustrates the different features of the two approaches.

1. Introduction

Numerous complex regression techniques are available to flexibly model the functional form of a continuous covariate's effect on outcome. Particularly smoothing approaches, that encompass a broad range of techniques and avoid assumptions of a particular functional form of a relationship between independent variables and outcome have been well-established in the statistical literature, see e.g. Fahrmeir and Tutz (2001), Hastie et al. (2003), Wood (2006a) and Ruppert et al. (2003).

Most smoothing approaches fit into the framework of generalized additive models (GAMs) (Hastie and Tibshirani, 1990) or their extensions (see e.g. Brezger and Lang, 2006). GAMs replace the linear predictor in a generalized linear model (Fahrmeir and Tutz, 2001) by a sum of smooth functions of the individual covariates. Some of the most widely used choices for the smooth functions in GAMs are P(enalized)-splines (e.g. Fahrmeir and Tutz, 2001; Wood, 2006a), and fractional polynomials (Royston and Sauerbrei, 2008).

P-splines approximate an unknown function f by a polynomial spline which can be written as a linear combination of some basis functions. For flexibility, typically a relatively large number of basis functions is used. To prevent overfitting a

[*] Corresponding author. Tel.: +43 512 507 7110; fax: +43 512 507 2851.
E-mail address: stefan.lang@uibk.ac.at (S. Lang).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.10.019

roughness penalty on the regression coefficients is used. Fractional polynomials (FPs) approximate f by the sum of power transformations of the covariates. FPs are more flexible than ordinary polynomials as they allow negative and non-integer powers.

Due to the availability of easy to use software, both, P-splines and FPs have extensively been utilized in various applications. However, despite their popularity only very limited comparisons of the performance and properties of the two methods have been conducted to date. A comparison of P-splines, restricted cubic splines and FPs in Cox proportional hazards models based on a single real dataset found that P-splines and restricted cubic splines were closer to each other than either was to the FPs (Govindarajulu et al., 2007). However, the true functional relationship of exposure and outcome was not known. A simulation study (Royston and Sauerbrei, 2005) and a case study (Royston and Sauerbrei, 2008) compared FPs to pure regression splines with an ad hoc choice of knots, without applying penalties or adaptive knot selection, thus not providing relevant insights.

We therefore compared the performance of P-splines and FPs in extensive simulations and in real data to provide guidance to the practitioner. We focused on assessing the ability of the estimators to recover the functional relationship between independent and dependent variables rather than on prediction. To be practically relevant, the comparison is based on standard implementations of both methods (STATA for FPs, and R and BayesX for P-splines). A particular focus is also on the default settings of the software. In Section 2, we briefly describe GAMs, P-splines and FPs. In Section 3 we compare the methods in simulations for continuous, binary and survival outcomes. In Section 4 we apply both approaches to data on malnutrition in children from the National Family Health Survey from India. Conclusions and recommendations are presented in Section 5.

2. Methods

2.1. Generalized additive models (GAMs)

GAMs assume that the distribution of the response variable y given covariates $x = (x_1, \dots, x_p)'$ belongs to an exponential family. A link function g relates the expected value μ of y to the covariates through

$$g(\mu) = \eta = f_1(x_1) + \cdots + f_p(x_p), \tag{1}$$

where $f_1, \dots, f_p$ are unknown, possibly nonlinear functions. The additive decomposition in (1) allows for good interpretability of the covariate effects and circumvents the curse of dimensionality (Hastie and Tibshirani, 1990). There are two main approaches for modeling the functions $f_1, \dots, f_p$, local polynomial regression and basis functions approaches. Here, we focus on basis functions approaches because both spline-based estimators and FPs are variants of this class.

The basis function approach assumes that an unknown function f in (1) can be approximated by a linear combination of basis functions, $B_1, \dots, B_K$,

$$f(x) = \sum_{k=1}^{K} \beta_k B_k(x), \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)'$ is a vector of unknown regression coefficients. Typically K is a large number to capture the variability of the data. Overfitting is avoided by either a roughness penalty, that is applied to the regression coefficients to ensure smoothness of (2), or by parsimonious selection of basis functions using variable selection methods. P-splines use a roughness penalty approach, while FPs use variable selection methods for adaptive basis function selection.

In the next two subsections we discuss P-splines and FPs in more detail for the simple model $y = f(x) + \varepsilon$.

2.2. P-splines

Spline-based approaches approximate an unknown function f in (1) by a polynomial spline of degree l, defined with respect to a given set of "knots"

$$x_{\min} = \kappa_0 < \kappa_1 < \cdots < \kappa_{m-1} < \kappa_m = x_{\max},$$

which are placed equally or non-equally spaced over the domain of x. A spline has the following two properties:

- In each of the intervals $[\kappa_j, \kappa_{j+1}], j = 0, \dots, m-1$ the spline f is a polynomial of degree l, and
- at the *knots* κ_j the spline is $l-1$ times continuously differentiable.

A spline can be written as a linear combination of $K = m + l$ basis functions (De Boor, 2001) and thus be expressed in the form (2). A widely used basis are local B-splines, which are nonzero only over a domain spanned by $l + 2$ knots.

In a simple regression spline approach, the unknown coefficients $\boldsymbol{\beta}$ in (2) are estimated using standard inference techniques for linear or generalized linear models. The choice of the number and positions of the knots is crucial. A small number of knots may result in a function space which is not flexible enough to capture variability of the data. A large number may lead to overfitting. As a remedy typically a sufficiently large number of knots (between 10 and 40) is defined to ensure enough flexibility. Sufficient smoothness of the fitted curve is achieved through a roughness penalty on the regression coefficients. This leads to a penalized least squares criterion

$$PLS(\lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + P_\lambda(f), \tag{3}$$

to be minimized with respect to the regression coefficients $\boldsymbol{\beta}$ of the spline f. The roughness penalty term $P_\lambda(f)$ depends on a smoothing parameter $\lambda \geq 0$ that governs the trade-off between smoothness and fidelity to the data, see below for the choice of λ.

For non-Gaussian models the penalized least squares criterion is replaced by a penalized log-likelihood criterion of the form

$$PL(\lambda) = l(\boldsymbol{\beta}) - P_\lambda(f), \tag{4}$$

to be maximized with respect to $\boldsymbol{\beta}$.

A widely used penalty proposed by Eilers and Marx (1996) is given by

$$P_\lambda(f) = \lambda \sum_{k=d+1}^{K} (\Delta^d \beta_k)^2, \tag{5}$$

where Δ^d is the difference operator of order d. For $d = 2$ the penalty is a discrete approximation to the integral of squared second order derivatives, a measure for the curvature of the function. A Bayesian interpretation of the penalty can be found in Brezger and Lang (2008). Small values of λ produce a close fit to the data, while large values of λ yield smooth function estimates.

P-splines due to Eilers and Marx are closely related to smoothing splines (Hastie and Tibshirani, 1990). A smoothing spline is derived from the penalized least squares criterion

$$PLS(\lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 \,\mathrm{d}x, \tag{6}$$

where f is assumed to be a smooth function with two continuous derivatives. The function f that minimizes (6) is a natural cubic spline. Smoothing splines are special cases (with $r = 1$) of thin plate regression splines defined for a r-dimensional covariate $\mathbf{x}$ (Wood, 2003). The original smoothing spline is rarely used in practice because in order to minimize (6) a knot has to be placed at every distinct covariate value. In the extreme, there are as many knots (and basis functions) as there are observations. As a remedy, Wood (2003) proposes a low rank (optimal) approximation to smoothing or more generally, thin plate splines.

The choice of the smoothing parameter λ strongly affects the fit of any P-spline. Three main approaches to choose λ are available: first, λ is estimated by minimizing some goodness of fit criterion, such as AIC or GCV (Eilers and Marx, 1996; Currie and Durban, 2002; Wood, 2000, 2004, 2008; Belitz and Lang, 2008). Second, the P-spline is re-expressed as a linear mixed model, and λ is estimated via restricted maximum likelihood (REML; Ruppert et al., 2003; Wand, 2003; Fahrmeir et al., 2004; Kauermann et al., 2009). Finally, a fully Bayesian version of P-splines in combination with Markov Chain Monte Carlo simulation techniques simultaneously estimates the regression coefficients and λ (Lang and Brezger, 2004; Brezger and Lang, 2006; Jullion and Lambert, 2007).

All P-spline approaches can be generalized to a full additive predictor as in (1). The nonlinear functions f_i can be estimated using backfitting (e.g. Hastie and Tibshirani, 1990), an iterative procedure that employs the univariate smoothers described above as building blocks. Direct methods for estimation without resorting to iterative procedures are also available (Marx and Eilers, 1998; Wood, 2000).

For all above mentioned approaches easy to use statistical software is available. A widely used software is the `mgcv` package in R (Wood, 2006c,a). The package allows us to estimate P-splines based on the penalty (5) with corresponding smooth term bs = "ps", see the section on `smooth.terms` in the `mgcv` manual (version 1.6–1.). The default smoother in `mgcv` is the low rank approximation to the smoothing spline mentioned above (smooth term bs = "tp"). In the simulation section we study both variants of P-splines. Smoothing parameter estimation in `mgcv` is based on minimizing GCV or via REML (without resorting to the connection with mixed models). This is done in a very efficient, fast and stable way using the methods described in Wood (2004, 2008). P-splines with the penalty (5) based on REML for smoothing parameter selection can also be estimated within the software package BayesX (Brezger et al., 2005; Belitz et al., 2009). BayesX also implements the full Bayesian approach and supports Cox proportional hazards survival models which are not covered in the `mgcv` package. Cox survival models with P-splines with penalty (5) can also be estimated with the function `coxph` of the R package `survival`.

2.3. Fractional polynomials (FPs)

FPs approximate an unknown function f by a linear combination of M polynomials $x^{p_j}, j = 1, \dots, M$. In ordinary polynomials the powers p_j are restricted to positive integer values, but within the FP modeling framework non-positive and fractional values for p_j are possible. A typical set of admissible powers is given by $p_j \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ where x^0 denotes $ln(x)$. More formally, an *FP* of degree M is defined as

$$FP_M(x) = \sum_{j=1}^{M} \beta_j h_j(x),$$

where $\beta_1, \dots, \beta_M$ are (regression) coefficients and h_j is recursively defined as

$$h_0(x) = 1$$
$$h_j(x) = \begin{cases} x^{p_j} & p_j \neq p_{j-1} \\ h_{j-1}(x)\ \ln(x) & p_j = p_{j-1}. \end{cases} \tag{7}$$

Note that this definition allows repeated powers. For instance, for $M = 2$, $p_1 \neq p_2$ we obtain the fractional polynomial

$$FP_2(x) = \beta_1 x^{p_1} + \beta_2 x^{p_2}$$

and for $M = 2$, $p_2 = p_1$,

$$FP_2(x) = \beta_1 x^{p_1} + \beta_2 x^{p_1} \ln(x).$$

FPs of degree 2, i.e. $M = 2$, is the default setting in all available implementations of FPs. Note that the degree M of FPs denotes the number of basis functions rather than the degree of the polynomials involved.

An obvious limitation of the definition (7) is the requirement $x > 0$ due to $x^0 := \ln(x)$. In implementations, a covariate with negative values is automatically shifted by $x = x + \delta$ to guarantee positivity. However, estimation results are sensitive to the choice of the origin δ, as we show in simulations and the application.

For prespecified order M the regression parameters β_j and the polynomial powers $p_j, j = 1, \dots, M$ are estimated by an algorithm described in Sauerbrei and Royston (1999) and Ambler and Royston (2001). For example, to fit FPs of order 2 one compares the best fitting FP2 to a null model, then a linear model, and finally an FP1 model based on χ^2 tests with 4, 3 and 2 degrees of freedom, respectively. If at any stage the test statistic cannot be rejected, the simpler model is fit to the data, otherwise the FP2 model is selected.

In additive models with multiple covariates the algorithm is combined with a backfitting type algorithm, see Sauerbrei and Royston (1999) for details. There are several criticisms of the above sequential testing approach to model selection. First, the test statistics that are used do not have a χ^2 distribution (Sauerbrei and Royston, 1999). Second, the overall type one error of the procedure may be inflated. To date, investigations of both issues are limited (Ambler and Royston, 2001). Finally, asymptotic properties of FP's have not been studied so far.

Software for fitting additive models based on FPs is available for the statistical computing platforms `STATA` (function `mfp`), `SAS` (macro `mfp8`) and `R` (function `fp` of the package `mfp`), see Sauerbrei et al. (2006). The `R` implementation is restricted to FPs of degree 2, i.e. $M = 2$.

3. Simulation study

3.1. Simulation setup

We compared FPs and spline-based estimators in extensive simulations for continuous, binary and survival outcomes. We applied FPs with degree $M = 2$ (henceforth FP2), the default setting of FP implementations in statistical software packages, and degree $M = 4$ (FP4). We used the function `mfp` in the software package `STATA` to fit the FPs. P-splines were fit to continuous and binary outcomes with the `mgcv` package of `R` (Wood, 2006a) in the following settings: the default smoother, a low rank approximation to the smoothing spline (henceforth LRSS) and cubic P-splines with second order difference penalty (5) (henceforth PS). We applied both spline smoothers with the packages default number of 10 basis functions. We used generalized cross validation (GCV, the default in `mgcv`) and restricted maximum likelihood (REML) to select smoothing parameters. Since survival models are not supported in `mgcv`, we used the `R` package `coxph` and the software `BayesX` (`remlreg` objects) to fit these models. Because of the limitations of both these packages for survival models, we only applied cubic P-splines with second order difference penalty (PS) with 10 basis functions. We do not present detailed numerical results for the default setting for the number of basis functions (17 for `coxph`, 22 for `BayesX`) to allow for a more direct comparison with the Gaussian and binomial simulation results. However, in the text we comment on cases for which the results based on the default settings differed noticeably. Smoothing parameter selection in `coxph` is obtained via AIC, while `BayesX` uses REML.

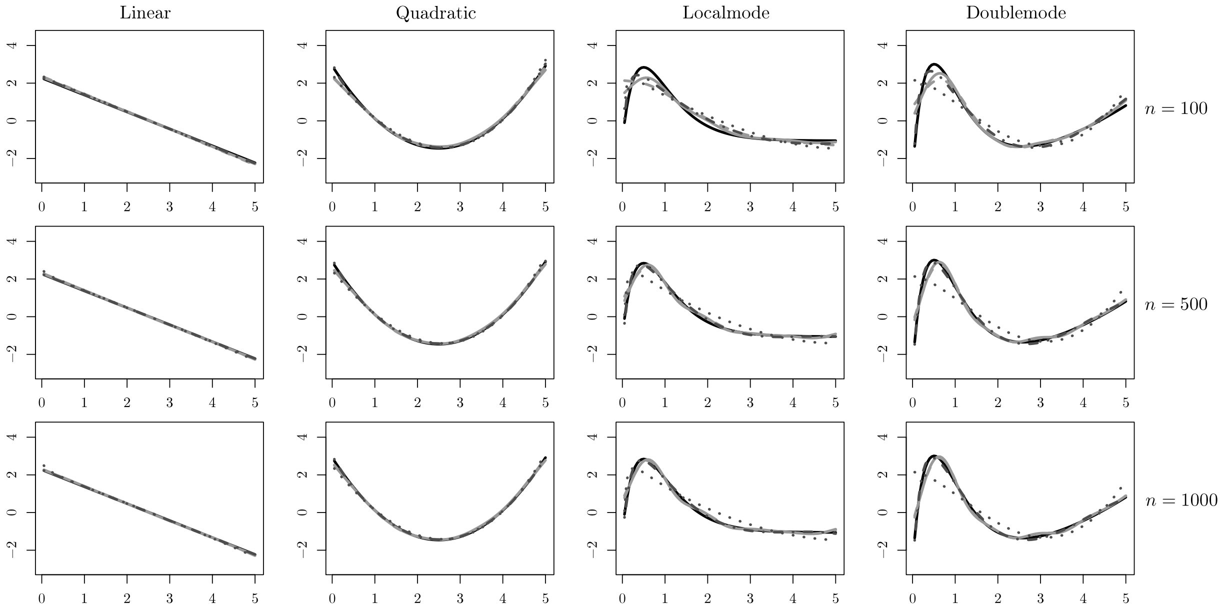

Fig. 1. Gaussian additive model, $\sigma = 0.735$: True curves (black solid lines) and average estimated curves (grey solid lines LRSS/GCV, grey dashed lines LRSS/REML, black dots FP2 and black dashed lines FP4 estimates).

The comparisons were based on data simulated from the following functions of the covariate x (see also Fig. 1):

$$\begin{array}{ll} \text{Linear:} & f_1(x) = -0.9x \\ \text{Quadratic:} & f_2(x) = 0.7 \cdot (x - 2.5)^2 \\ \text{Localmode:} & f_3(x) = 24x \cdot \exp(-2x) \\ \text{Doublemode:} & f_4(x) = 1.3 \cdot (24x \cdot \exp(-2x) + 0.11 \cdot x^2) \end{array} \tag{8}$$

The four functions were scaled such that they all had the same range of 4 units.

For each function f_j in (8), we generated outcome data y from the following four models for one hundred equally spaced design points x between 0.05 to 5:

(i) Gaussian model $y = f_j(x) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. We chose four different values for the error standard deviation: $\sigma = 0.3675$, $\sigma = 0.735$, $\sigma = 1.1025$ and $\sigma = 1.47$ to obtain various magnitudes of signal to noise ratio (SNR).

(ii) Binomial model $y \sim B(1, \pi)$ with

$$\pi = \exp(c \cdot f_j(x)) / \exp(1 + c \cdot f_j(x)),$$

$c = 1, 0.75, 0.5, 0.25$ is a scaling factor chosen to imitate the SNRs of the Gaussian case.

(iii) Survival model (similar to Bender et al., 2005), with hazard rate $\lambda(t) = \lambda_0(t) \exp[0.5 f_j(x)]$ where the baseline hazard $\lambda_0(t)$ is given by

$$\lambda_0(t) = \begin{cases} \cos(x) + 1.2 & x \leq 2\pi \\ 2.2 & x > 2\pi. \end{cases}$$

To obtain censored observations, we generated independent censoring times $C \sim \mathrm{Exp}(0.2)$.

(iv) Gaussian, Binomial and survival models with the additive predictor $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$. To guarantee identifiability of the curves we randomly draw the values of x_j without replacement from equally spaced points between 0.05 and 5. Error variances or scaling factors of functions are identical to those specified in (i)–(iii).

For each of these settings 500 replicated data sets with four different sample sizes $n = 100, 500, 1000, 2000$ were simulated.

The goodness of fit was measured by the empirical mean squared error (MSE),

$$MSE(\hat{f}_j) = 1/S \sum_{s=1}^{S} \left(f_j(x_s) - \hat{f}_j(x_s) \right)^2,$$

where summation is over all design points $x_1, \ldots, x_S$, with $S = 100$.

We also investigated the coverage of pointwise confidence intervals. The *m*gcv package of R and BayesX uses Bayesian rather than frequentist confidence intervals because of their assumed better coverage properties, see e.g. Wood (2006b). The survival package *c*oxph of R is restricted to frequentist confidence intervals.

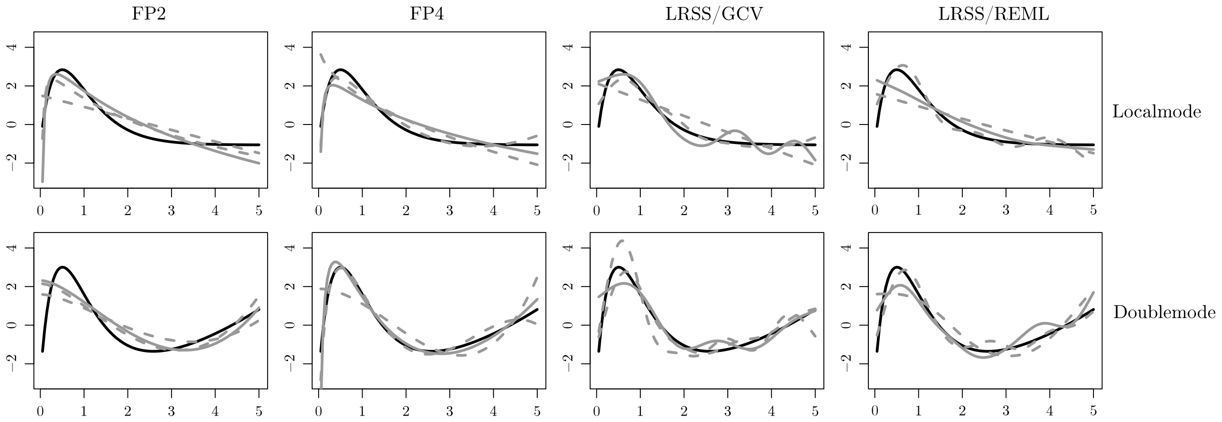

Fig. 2. Gaussian additive model, $\sigma = 0.735$, $n = 100$: Some function estimates of the localmode and doublemode function based on FP2, FP4, LRSS/GCV, LRSS/REML. Shown are the 2.5%, 50% and 97.5% best fits according to the MSE measure. The black solid lines represent the true functions, the grey solid lines the median and grey dashed lines the other quantiles.

3.2. Gaussian responses

Fig. 1 plots average estimated functions, i.e. the mean of $\hat{f}_j$ over all replications, with the true curves for the additive model (iv) for $\sigma = 0.735$ and $n = 100, 500, 1000$. Results for the single predictor models (i)–(iii) and other values of σ and n were similar and are not shown but are available at http://www.uibk.ac.at/statistics/personal/lang/publications/fp_sim_summary.pdf.

Representative for the two spline-based estimators we present results for LRSS. The results for PS are visually indistinguishable from LRSS.

All estimators are unbiased for the linear and quadratic functions in (8) for all choices of sample size, SNR and model type (single or additive predictor). FP4s and the spline estimators also showed very little bias for the localmode and doublemode function, except for $n = 100$, where the estimators are more biased, particularly at the function modes. As expected, the bias decreased for large SNR (figures not shown). An inspection of some individual estimates (Fig. 2) reveals a tendency to underfit for FP4s for small sample sizes ($n = 100$), whereas LRSS based on GCV (to a lesser extent also REML) tended to overfit and produce very unsmooth estimates. The FP2 estimates for the localmode and doublemode functions were considerably biased for all sample sizes and values of σ. The observed patterns are also reflected in the MSE estimates (Table 1). Overall, the estimates based on FP4 resulted in the lowest $\log(\sqrt{MSE})$, followed closely by both spline estimators based on GCV and REML. LRSS and PS produce similar MSE's with somewhat lower values for PS. For the more curved functions f_3 and f_4 PS based on GCV somewhat outperforms FP4. FP2s, however, had a considerably higher $\log(\sqrt{MSE})$ for the localmode and doublemode function.

With the exception of PS based on GCV, average coverage rates of 95% confidence intervals were below the nominal level for the more curved functions doublemode and localmode for all estimators (Table 2). FP4, LRSS based on GCV/REML and PS based on REML were closer to the nominal level (with coverage rates around 85%–90%) than FP2. The coverage decreased as sample size increased for FP4 and LRSS, due to too narrow confidence intervals. The coverage of confidence intervals for PS is not affected by sample size. The undercoverage of FP2 reflects lack of fit. For the quadratic and linear functions, the confidence intervals had nominal levels for both spline estimators whereas FPs produced conservative confidence intervals.

3.3. Binomial responses

Overall, results for binomial responses are similar to the Gaussian case. However, we obtained a considerable number of unreliable results with the FP2 and FP4 estimators and, to a lesser extent, with both spline-based estimators based on GCV, especially for small sample sizes $n = 100$ and $n = 500$. This is illustrated by Fig. 3 panel (a) which shows a particular FP4 estimate for the doublemode function f_4 in (8) with scaling factor $c = 0.75$. Results are improved for $n = 500$ for all function types. The problem appears less frequently for the quadratic and linear function. The reason for this problematic behavior of the FPs is that the support of the design values x is close to zero. The FP basis functions with negative power have an asymptote at zero, and thus yield extremely high values close to zero, which distorts the fitted functions. After shifting all x by adding one unit, the problem disappears (panel (b) in Fig. 3), although the FP based estimates still frequently miss important features of the exposure curves, (Fig. 3) panel (c).

The spline estimators based on GCV also reveals convergence problems showing sometimes extremely rough estimated functions, see Fig. 5, panel (d) (results are shown only for LRSS as PS is very similar). These problems are most pronounced for the additive model (iv) with small sample size, $n = 100$, but occur for all SNRs and all function types. Remarkably, LRSS and PS based on REML do not have convergence problems and almost all estimates produce reasonable results (panel e).

Table 1

Estimated median $\log(\sqrt{MSE})$ of the multivariate models with medium SNR ($\sigma = 0.735$ for Gaussian responses, scaling factor $c = 0.75$ for Binomial outcome, scaling factor $c = 0.5$ for survival models). Numbers in gray cells represent the respective smallest median of the algorithms for each row and distribution.

f	n	Gaussian						Binomial Logit						Survival			
		FP		GCV		REML		FP		GCV		REML		FP		AIC	REML
		FP2	FP4	LRSS	PS	LRSS	PS	FP2	FP4	LRSS	PS	LRSS	PS	FP2	FP4	PS	PS
f_1	100	0.012	0.012	0.021	0.025	0.018	0.018	0.043	0.042	0.146	0.194	0.074	0.076	0.015	0.016	0.427	0.019
	500	0.004	0.002	0.006	0.006	0.004	0.004	0.014	0.011	0.03	0.034	0.028	0.028	0.02	0.007	0.024	0.005
	1000	0.005	0.001	0.003	0.002	0.002	0.002	0.006	0.005	0.013	0.012	0.011	0.01	0.016	0.012	0.012	0.006
	2000	0.008	0.001	0.002	0.001	0.001	0.001	0.004	0.003	0.007	0.006	0.004	0.004	0.022	0.014	0.007	0.036
f_2	100	0.07	0.063	0.072	0.068	0.068	0.061	0.265	0.815	0.338	0.35	0.189	0.174	0.041	0.043	0.423	0.036
	500	0.038	0.011	0.021	0.017	0.022	0.017	0.05	0.048	0.062	0.055	0.057	0.053	0.026	0.013	0.024	0.012
	1000	0.019	0.006	0.013	0.009	0.012	0.009	0.029	0.022	0.034	0.028	0.029	0.025	0.023	0.015	0.012	0.012
	2000	0.02	0.003	0.008	0.005	0.008	0.005	0.015	0.01	0.019	0.017	0.016	0.015	0.021	0.017	0.007	0.046
f_3	100	0.308	0.222	0.21	0.186	0.251	0.268	0.39	0.403	0.409	0.432	0.319	0.318	0.153	0.157	0.445	0.107
	500	0.193	0.04	0.057	0.039	0.059	0.047	0.108	0.107	0.134	0.124	0.156	0.167	0.079	0.029	0.026	0.035
	1000	0.188	0.02	0.039	0.02	0.04	0.029	0.101	0.032	0.076	0.064	0.082	0.082	0.086	0.031	0.013	0.028
	2000	0.186	0.014	0.031	0.01	0.03	0.018	0.09	0.017	0.054	0.039	0.051	0.046	0.092	0.03	0.008	0.074
f_4	100	0.49	0.149	0.225	0.186	0.252	0.258	0.763	0.768	0.56	0.579	0.4	0.404	0.184	0.163	0.426	0.133
	500	0.461	0.037	0.076	0.041	0.079	0.058	0.32	0.085	0.159	0.145	0.189	0.218	0.137	0.029	0.026	0.038
	1000	0.459	0.025	0.058	0.02	0.06	0.039	0.292	0.031	0.1	0.08	0.107	0.095	0.145	0.026	0.013	0.035
	2000	0.458	0.021	0.049	0.01	0.049	0.02	0.289	0.017	0.062	0.038	0.059	0.049	0.144	0.025	0.008	0.014

Table 2

Additive models with medium SNR ($\sigma = 0.735$ for Gaussian responses, scaling factor $c = 0.75$ for Binomial outcome, scaling factor $c = 0.5$ for survival models): Average coverage rates of 95% pointwise confidence intervals of the algorithms for each row and distribution. Cells corresponding to values below a 92.5% level (undercoverage) are marked with dark gray and values larger than a 97.5% level (overcoverage) with light gray.

f	n	Gaussian						Binomial Logit						Survival			
		FP		GCV		REML		FP		GCV		REML		FP		AIC	REML
		FP2	FP4	LRSS	PS	LRSS	PS	FP2	FP4	LRSS	PS	LRSS	PS	FP2	FP4	PS	PS
f_1	100	0.999	0.999	0.939	0.929	0.944	0.939	0.988	0.987	0.94	0.914	0.939	0.937	0.829	0.829	0.85	0.912
	500	0.992	0.999	0.944	0.939	0.947	0.948	0.986	0.998	0.932	0.92	0.925	0.911	0.264	0.663	0.95	0.878
	1000	0.973	0.998	0.92	0.922	0.922	0.932	0.994	0.996	0.93	0.927	0.934	0.931	0.087	0.229	0.951	0.632
	2000	0.938	0.997	0.926	0.931	0.934	0.933	0.993	0.999	0.936	0.933	0.956	0.938	0.001	0.011	0.933	0.103
f_2	100	0.976	0.979	0.947	0.939	0.968	0.962	0.893	0.74	0.945	0.929	0.956	0.948	0.673	0.625	0.845	0.962
	500	0.9	0.98	0.948	0.946	0.964	0.958	0.964	0.973	0.944	0.934	0.953	0.94	0.537	0.644	0.948	0.957
	1000	0.823	0.982	0.939	0.94	0.957	0.953	0.957	0.982	0.946	0.948	0.966	0.961	0.421	0.516	0.949	0.9
	2000	0.705	0.984	0.924	0.941	0.943	0.953	0.958	0.988	0.95	0.945	0.966	0.958	0.296	0.356	0.927	0.455
f_3	100	0.758	0.83	0.82	0.836	0.696	0.641	0.871	0.827	0.837	0.842	0.798	0.787	0.424	0.406	0.838	0.762
	500	0.419	0.918	0.9	0.927	0.893	0.887	0.931	0.948	0.871	0.883	0.803	0.773	0.316	0.59	0.944	0.916
	1000	0.292	0.914	0.858	0.929	0.866	0.871	0.774	0.965	0.92	0.918	0.89	0.854	0.23	0.487	0.945	0.924
	2000	0.2	0.88	0.786	0.941	0.802	0.839	0.54	0.965	0.895	0.919	0.892	0.867	0.158	0.334	0.901	0.754
f_4	100	0.764	0.918	0.9	0.903	0.878	0.849	0.725	0.717	0.874	0.863	0.842	0.82	0.42	0.484	0.839	0.843
	500	0.418	0.929	0.869	0.925	0.871	0.866	0.742	0.942	0.908	0.907	0.865	0.81	0.394	0.64	0.946	0.948
	1000	0.279	0.903	0.818	0.931	0.83	0.834	0.64	0.972	0.908	0.919	0.889	0.872	0.35	0.565	0.94	0.754
	2000	0.199	0.768	0.741	0.944	0.756	0.843	0.498	0.977	0.884	0.928	0.89	0.867	0.322	0.429	0.898	0.56

The median $\log(\sqrt{MSE})$ values show a similar pattern to Gaussian responses (Table 1). After shifting the covariate values away from zero, results were mostly similar for FP4, and LRSS or PS based on GCV and REML. However, for small sample size, $n = 100$, the greater stability of the two spline estimators based on REML results in better estimates. Of note, the MSEs are much larger for FP2 and FP4 on the original scale (data not shown).

The coverage rates of pointwise credible intervals are given in Table 2. By and large, we observe similar patterns as for the Gaussian case.

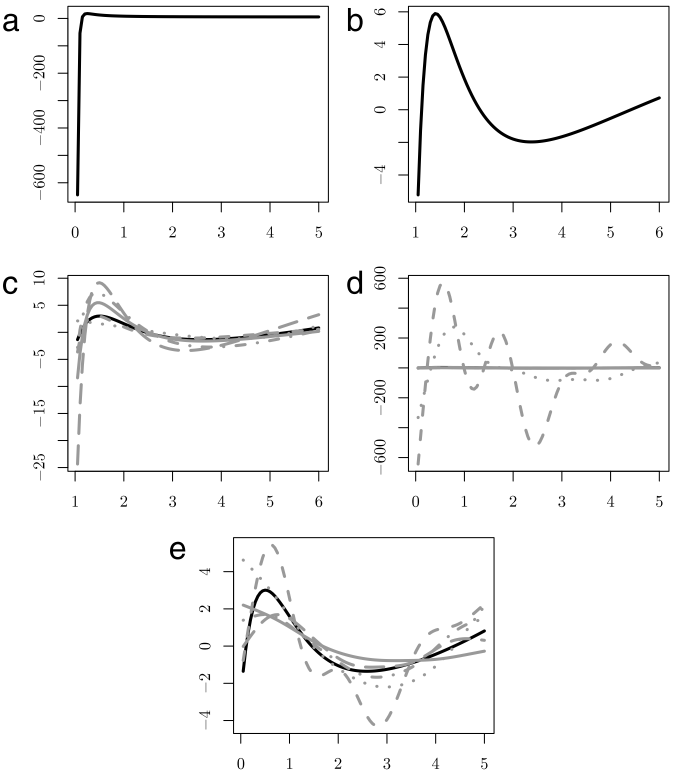

Fig. 3. Binomial additive model, scaling factor $c = 0.75$, $n = 100$: Panel (a): FP4 estimate for a particular replication. Panel (b): FP4 estimate based on a shift of covariate values by one unit. Panels (c)–(e): Some individual function estimates for FP4, LRSS/GCV, LRSS/REML. Shown are the 2.5%, 10%, 50%, 90% and 97.5% best fits according to the MSE measure. The black solid lines represent the true functions, the grey solid lines the median and grey dashed lines the other quantiles.

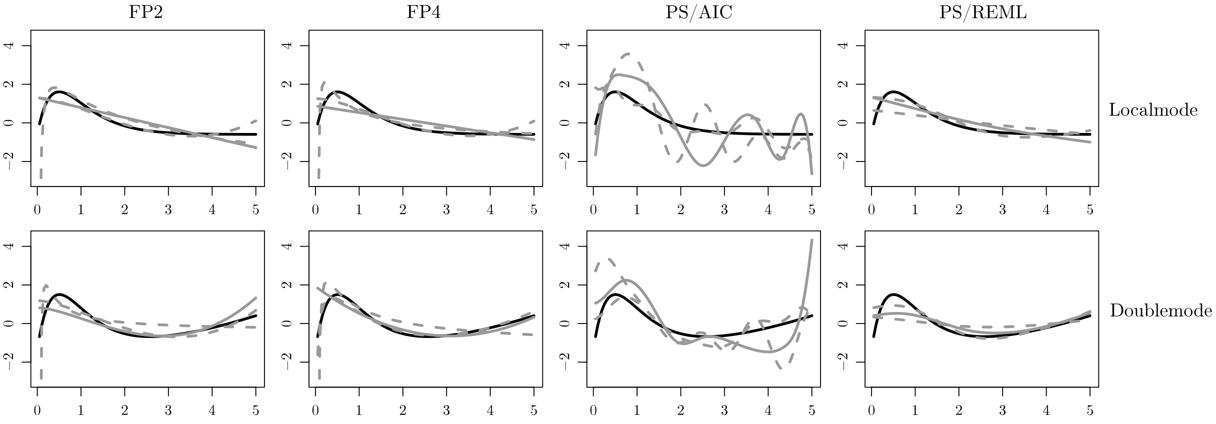

Fig. 4. Survival additive model, $n = 100$: Some function estimates of the localmode and doublemode function based on FP2, FP4, PS/AIC, PS/REML. Shown are the 2.5%, 50% and 97.5% best fits according to the MSE measure. The black solid lines represent the true functions, the grey solid lines the median and grey dashed lines the other quantiles.

3.4. Survival models

Fig. 4 shows for the doublemode and localmode function and $n = 100$ some representative function estimates for the four approaches FP4, FP2, PS/AIC and PS/REML. Table 1 displays estimates of $\log(\sqrt{MSE})$. The results correspond to the Cox-proportional hazards model with the additive predictor (iv) and $n = 100, 500, 1000$ (data for $n = 2000$ are not shown as results were similar to $n = 1000$).

For small sample size, $n = 100$, individual estimates of PS based on AIC are mostly very rough and the results are not reliable (Fig. 4). Notably, the averaged estimated curves (figures not shown) suggest acceptable results. Note also, that the

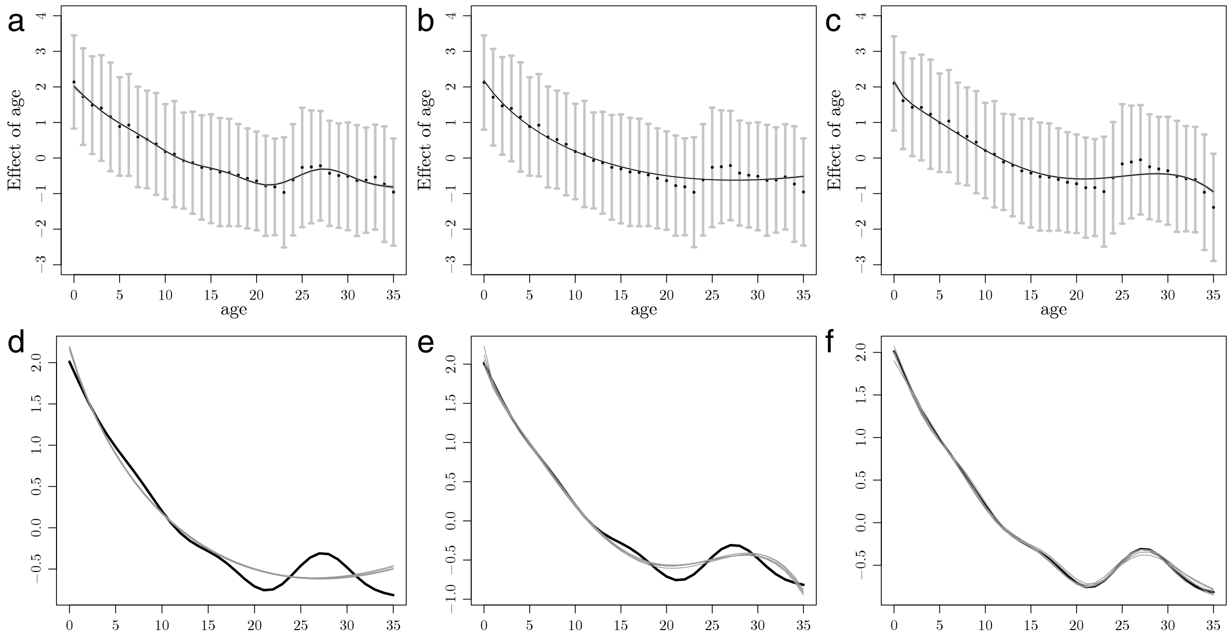

Fig. 5. Malnutrition in India: Estimated effects of the child's age based on LRSS/REML (panel a), FP2 (panel b) and FP4 (panel c). Shown are the estimated functions and 95% confidence intervals and stick plots of means plus/minus standard deviations of the corresponding partial residuals averaged over the distinct covariate values. Panels (d) through (f) show estimation results for simulated data for covariate age. Shown are typical estimates for FP2 in plot (a), FP4 in plot (b) and LRSS/REML in plot (c). The true function $f(age)$, which was generated by the fitted REML model for the India dataset with model equation $y = f(age) + \varepsilon$ and $\varepsilon \sim N(0, 2.17)$, is represented by the solid black lines and estimates with solid grey lines.

wiggliness increases with the number of basis functions. In particular, the default setting in coxph with 17 basis produces completely unreliable results with PS/AIC. The most stable and best estimator for small sample size are PS based on REML. Acceptable results are also obtained with FP4 while FP2 shows strong bias for the doublemode and localmode function.

For sample sizes $n \geq 500$ all estimators are almost unbiased for the quadratic and linear functions (figures not shown). For the localmode and doublemode functions, FP2 estimators again show strong bias while FP4, PS with AIC and REML recover the important features of these functions. However, compared to the Gaussian and binomial outcomes, even with FP4, AIC and REML a noticeable bias can be observed, particularly at the modes of the functions. The best estimator in terms of MSE for sample size $n \geq 500$ is PS based on AIC followed by FP4 and PS based on REML.

Average coverage rates of pointwise credible intervals are typically far beyond the nominal level (Table 2). Only PS with smoothing parameter chosen via AIC for sample sizes $n \geq 500$ produced adequate coverage. The undercoverage of FP4 is a result of confidence intervals that become narrower towards the center of the covariate support and almost collapse to a point, even though the observations are uniformly distributed over the whole range. This phenomenon was not observed for FPs with Gaussian and binomial responses. The undercoverage of FP2 and PS with REML is caused by biased estimates.

4. Data example

We apply *P*-splines and FPs to data from the second National Family Health Survey (NFHS-2) from India, conducted in 1998 and 1999 (see http://www.nfhsindia.org/). We analyse the impact of malnutrition in approximately 30 000 children born in the 3 years preceding the survey. The effect of malnutrition is usually measured by comparing the anthropometric status of children in a given population to a reference population of well nourished children. Here, we focus on stunting or insufficient height for a given age. The outcome variable is defined as

$$z = \frac{H - MH}{\sigma}, \tag{9}$$

where H refers to a child's height at a certain age, and MH and σ refer to the median and the standard deviation of height in the reference population, respectively. We fit the additive model

$$\begin{aligned} z = {} & \beta_0 + f_1(age) + f_2(vacnumb) + f_3(border) + f_4(educm) \\ & + f_5(bmimo) + f_6(biage) + f_7(hhs) + f_8(ai) + \varepsilon, \end{aligned}$$

where $f_1, \ldots, f_8$ are unknown nonlinear functions of the child's age (*age*), the number of the child's vaccinations (*vacnumb*), the birth order (*border*), the mother's years of education (*educm*), body mass index (*bmimo*) and her age at birth (*biage*),

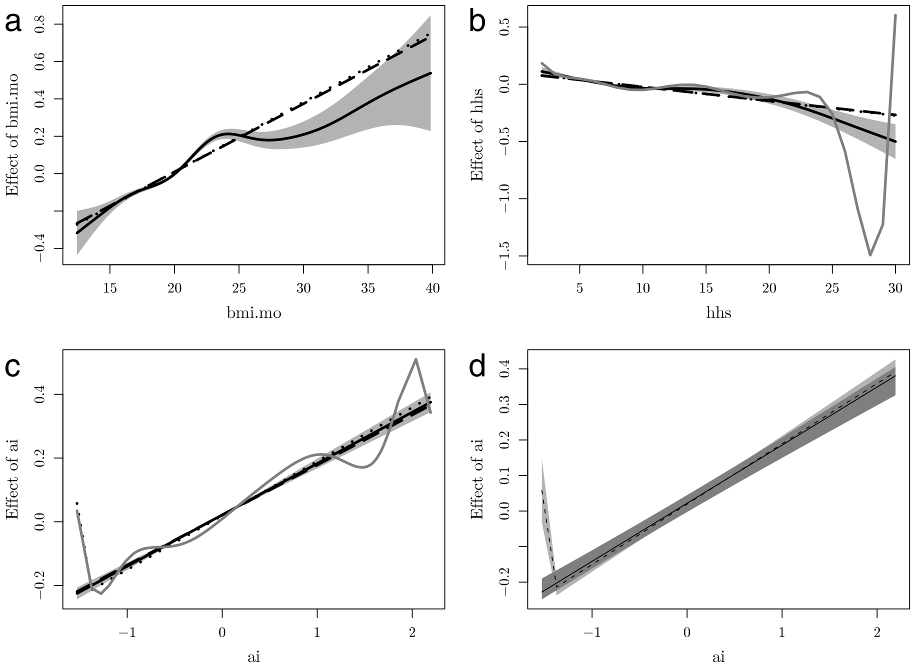

Fig. 6. Malnutrition in India: the black solid lines represent estimates based on LRSS/REML and 95% confidence intervals, dotted lines FP2 and wide dashed lines FP4 estimates. Figures (b) and (c) contain additional estimates based on PS/GCV.

the household size (*hhs*) and an asset index of the household's wealth (*ai*). The errors ε are assumed to be i.i.d. Gaussian variables with common variance σ^2. This model is similar to a model used in Belitz et al. (2010), but with fewer covariates and without considering spatial heterogeneity.

We fit model (4) with the `mgcv` function in `R` using LRSS and PS based on 10 basis functions. The smoothing parameters were estimated by GCV and REML. The spline-based estimates were compared to FP2 and FP4 estimates, obtained from the `mfp` function in `STATA`.

LRSS, PS, FP2 and FP4, produced very similar estimated effects for the three variables *border*, *educmy* and *biage*. The results for these covariates are therefore not shown. The approaches yielded different results for *age*, *bmimo*, *hhs* and *ai*, and we thus present estimated functions for these variables (Fig. 5, top row, Fig. 6) based on the three estimators LRSS/REML (solid line), FP2 (dotted lines) and FP4 (dashed lines). The four spline-based estimators LRSS/REML, LRSS/GCV, PS/REML and PS/GCV produce similar results, with the exception of *hhs* and *ai*, where PS/GCV is very unstable. Therefore Fig. 6(b,c) additionally contains results for PS/GCV. Overall, the estimated functional forms for individual variables agree with Belitz et al. (2010).

The most striking differences between *P*-splines and FP's can be observed for the child's age (Fig. 5). The top row of Fig. 5 shows that the estimated bump around age 25–30 months obtained with the spline estimator LRSS captures a distinct feature in the data. The very narrow confidence bands and the fact that the observations are evenly distributed over the age range indicate that this bump is not caused by outlying observations. Moreover the bump can be explained by a change in the reference standard used in the computation of the outcome variable *z* in Eq. (9). Before the age of 24 months *z* is obtained by comparing the children's height to the heights of middle class US white children. After 24 months *z* was computed based on a cross-section of the overall US population, whose nutritional status is worse than that of white middle class US children, thus causing an apparent improvement in the nutritional status of the Indian children. However, this change in the effect of *age* is missed by the FP2 and FP4 estimators as they are not flexible enough to capture such local phenomena.

To further investigate the behavior of the three methods, we simulated outcome variables from the model $y = f(age) + \varepsilon$ and $\varepsilon \sim N(0, 2.17)$, where $f(age)$ was the *P*-spline based on LRSS/REML fitted model for the India dataset. The bottom row of Fig. 5 further highlights that FP2 and FP4 are not able to detect the underlying structure of the effect of age on outcome.

The methods also differ in the estimated effects for *bmimo* and *hhs* (Fig. 6(a,b)), although the differences are less pronounced. LRSS adapts better to the data than the almost linear fits obtained with FP2 and FP4. However, inspection of the partial residuals shows that all approaches (LRSS/REML, FP4, FP2) give reasonable estimates (plots not shown). PS/GCV produces quite unstable and therefore unreliable results for *hhs*.

Of note is the estimated effect of FP2 for *ai* in panel (c) of Fig. 6. This behavior of FP2 arises since the minimum of *ai* is negative, and the software automatically adds a constant δ to the variable to guarantee positive values. As already mentioned, FPs are not invariant to the choice of origin of a covariate, which causes the behavior of the estimates seen in Fig. 6. Indeed, if we replace *ai* by $ai + 2$, and re-fit the model, this artefact disappears, see panel (d) of Fig. 6. Again, PS/GCV shows unstable behavior (Fig. 6(c)).

Finally, we point out that both approaches could be combined. The estimated spline functions for *vacnumb*, *border*, *educm* and *ai* could be replaced by the more simple and better interpretable FPs. For *border*, *educm* and *ai* FP4 results in a linear fit, while for *vacnumb* a parametric fit with basis functions $vacnumb^{-2}$, $vacnumb^{-2}\log(vacnumb)$ and $vacnumb^{3}$ is obtained.

5. Conclusion

We compared various variants of *P*-splines and fractional polynomials, two widely used smoothing techniques, in extensive simulations and a real data application. The simulations show that the spline-based estimators and fractional polynomials of sufficiently large degree (we used FPs of degree 4) performed similarly in most settings. FPs of degree 2, however, showed considerable bias and consistently higher MSEs compared to all other estimators. Moreover, the real data example revealed that very complex functional forms cannot be detected by FPs of degree 4 (or higher). We also showed that FPs are prone to artefacts because of the dependence of results on the covariate support, while *P*-splines (either LRSS or PS) based on GCV (or AIC in the survival models) reveal sometimes wiggly estimates. The most stable estimators were produced by *P*-splines (either LRSS or PS) based on REML for smoothing parameter selection.

Our findings suggest that *P*-splines are more suited to exploratory data analysis because of their greater flexibility than FPs. We recommend using REML for smoothing parameter selection as it seems the best compromise between goodness of fit and stability of the estimator. The asymptotic behavior of spline-based estimators has been studied (Kauermann et al., 2009, and the references therein), whereas no such results exist for FPs. Spline-based models can be fit using iterative procedures (e.g. backfitting, Markov Chain Monte Carlo) or by direct optimization. In contrast, FPs can only be estimated via an iterative backfitting approach. It is also noteworthy that spline-based curve fitting is more adaptable to specific settings such as cyclic smoothing (e.g. Eilers and Marx, in press), smoothing with shape constraints (e.g. Bollaerts et al., 2006) and locally adaptive smoothing (e.g. Ruppert and Carroll, 2000). While FPs have limitations for exploratory analysis, they are valuable in subsequent steps, to simplify models for better interpretability.

We see several directions for future research. Currently, FPs are estimated in a rather ad hoc procedure that is largely in the spirit of stepwise procedures for linear models. These procedures are not favored by statisticians because of their rather limited theoretical support (Miller, 2002). Hence, there is need for alternative estimation methods. A promising approach is a Bayesian version of FPs suggested by Sabanés Bové and Held (in press). Another problem with FPs, that has been ignored in the literature is the potentially strong dependence of results on the covariate range. A possible remedy could be a mapping of observed covariate values to a "save" interval such that the observed problems are less likely to happen.

While the behavior of spline-based estimators is better understood, the best approach to select smoothing parameters is still not entirely clear. Our simulations suggest that REML may yield more stable results than GCV and AIC, however, theoretical results are needed to support that finding.

Acknowledgements

We thank the two referees for their many valuable comments and suggestions to improve the first version of the paper.

References

Ambler, G., Royston, P., 2001. Fractional polynomial model selection procedures: investigation of type I error rate. Journal of Statistical Computation and Simulation 69, 89–108.

Belitz, C., Brezger, A., Kneib, T., Lang, S., 2009. BayesX Manuals. Technical Report. Department of Statistics, University of Munich. Available at http://www.stat.uni-muenchen.de/~bayesx.

Belitz, C., Hübner, J., Klasen, S., Lang, S., 2010. Determinants of the socioeconomic and spatial pattern of undernutrition by sex in India: a geoadditive semi-parametric regression approach. In: Kneib, T., Tutz, G. (Eds.), Statistical Modelling and Regression Structures. Physika Verlag.

Belitz, C., Lang, S., 2008. Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis 53, 61–81.

Bender, R., Augustin, T., Blettner, M., 2005. Generating survival times to simulate cox proportional hazards models. Statistics in Medicine 24, 1713–1723.

Bollaerts, K., Eilers, P., Van Mechelen, I., 2006. Simple and multiple *P*-spline regression with shape constraints. British Journal of Mathematical and Statistical Psychology 59, 451–469.

Brezger, A., Kneib, T., Lang, S., 2005. Bayesx: analyzing bayesian structured additive regression models. Journal of Statistical Software 14, 1–22.

Brezger, A., Lang, S., 2006. Generalized additive regression based on bayesian *P*-splines. Computational Statistics and Data Analysis 50, 967–991.

Brezger, A., Lang, S., 2008. Simultaneous probability statements for bayesian *P*-splines. Statistical Modelling 8, 141–168.

Currie, D., Durban, M., 2002. Flexible smoothing with *P*-splines: a unified approach. Statistical Modelling 2, 333–349.

De Boor, C., 2001. A Practical Guide to Splines. Springer, New York.

Eilers, P., Marx, B., 2010. Splines, knots, and penalties. Wiley Interdisciplinary Reviews: Computational Statistics, (in press).

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing using *B*-splines and penalized likelihood. Statistical Science 11, 89–121.

Fahrmeir, L., Kneib, T., Lang, S., 2004. Penalized structured additive regression for space–time data: a Bayesian perspective. Statistica Sinica 14, 731–761.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd ed. Springer, New York.

Govindarajulu, U.S., Spiegelman, D., Thurston, S.W., Ganguli, B., Eisen, E.A., 2007. Comparing smoothing techniques in cox models for exposure–response relationships. Statistics in Medicine 26, 3735–3752.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall/CRC, London.

Hastie, T.J., Tibshirani, R.J., Friedman, J., 2003. The Elements of Statistical Learning. Springer, New York.

Jullion, A., Lambert, P., 2007. Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian *P*-splines models. Computational Statistics and Data Analysis 51, 2542–2558.

Kauermann, G., Krivobokova, T., Fahrmeir, L., 2009. Some asymptotic results on generalized penalized spline smoothing. Journal of the Royal Statistical Society 71, 487–503.

Lang, S., Brezger, A., 2004. Bayesian *P*-splines. Journal of Computational and Graphical Statistics 13, 183–212.

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

Miller, A., 2002. Subset Selection in Regression. Chapman & Hall/CRC, Boca Raton, FL.

Royston, P., Sauerbrei, W., 2005. Building multivariable regression models with continuous covariates in clinical epidemiology—with an emphasis on fractional polynomials. Methods of Information in Medicine 44, 561–571.

Royston, P., Sauerbrei, W., 2008. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Wiley.

Ruppert, D., Carroll, R.J., 2000. Spatially adaptive penalties for spline fitting. Australian and New Zealand Journal of Statistics 42, 205–223.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.

Sabanés Bové, D., Held, L., (2010). Bayesian fractional polynomials. Statistics and Computing, (in press).

Sauerbrei, W., Meier-Hirmer, C., Benner, A., Royston, P., 2006. Multivariable regression model building by using fractional polynomials: description of SAS, stata and R programs. Computational Statistics and Data Analysis 50, 3464–3485.

Sauerbrei, W., Royston, P., 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society A 162, 71–94.

Wand, M.P., 2003. Smoothing and mixed models. Computational Statistics 18, 223–249.

Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society 62, 413–428.

Wood, S.N., 2003. Thin-plate regression splines. Journal of the Royal Statistical Society 65, 95–114.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.

Wood, S.N., 2006a. Generalized Additive Models: An Introduction with R. Chapman & Hall.

Wood, S.N., 2006b. On confidence intervals for generalized additive models based on penalized regression splines. Australian and New Zealand Journal of Statistics 48, 445–464.

Wood, S.N., 2006c. R-Manual: The mgcv package, version 1.3-22. Technical Report.

Wood, S.N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society 70, 495–518.