University of Innsbruck

Working Papers in Economics and Statistics

Comparing Penalized Splines and Fractional Polynomials for Flexible Modelling of the Effects of Continuous Predictor Variables

Alexander Strasak, Nikolaus Umlauf,
Ruth Pfeiffer and Stefan Lang

2010-11

Comparing Penalized Splines and Fractional Polynomials for Flexible Modelling of the Effects of Continuous Predictor Variables

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

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

Address for correspondence : Stefan Lang, University of Innsbruck, Department of Statistics, phone: +435125077110, fax: +435125072851, email: stefan.lang@uibk.ac.at

Abstract

P(enalized)-splines and fractional polynomials (FPs) have emerged as powerful smoothing techniques with increasing popularity in several fields of 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. We thus performed extensive simulations to compare FPs of degree 2 (FP2) and degree 4 (FP4) and P-splines that used generalized cross validation (GCV) and restricted maximum likelihood (REML) for smoothing parameter selection. We evaluated 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. We found that for more curved functions FP2, the current default implementation in standard software, showed considerably bias and consistently higher mean squared error (MSE) compared to spline-based estimators (REML, GCV) 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. Finally, we highlight the specific features of the approaches in a real dataset.

Keywords: generalized additive models, GAMs, simulation, smoothing

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 (2006b) and Ruppert et al. (2003).

Most smoothing approaches fit into the framework of generalized additive models (GAMs) (Hastie and Tibshirani 1990) or their extensions (e.g. Fahrmeir et al. 2004). 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 2006b), 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 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 (e.g. Strasak et al. 2009, Eisen et al. 2004, Andre et al. 2004, Stansfeld et al. 2005, Shlipak et al. 2006, Beatty 2009, Ugarte, Goicoa, and Militino 2009, Ellner, Seifu, and Smith 2002, Henley and Peirson 2001, Peterson et al. 2003, Finch et al. 2007). 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 real single dataset (Govindarajulu et al. 2007) found that P-splines and restricted cubic splines were closer to each other than either was to the FPs. However, the true functional relationship of exposures 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 nonlinear functional relationship between independnet 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). In section 2, we briefly describe GAMs, P-splines and fractional polynomials. In section 3 we compare the methods in simulated data 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)

There is a large literature on flexibly modeling and estimating the effect of continuous covariates on outcome (e.g. Hastie, Tibshirani, and Friedman 2003, Fahrmeir and Tutz 2001, Wood 2006b). The vast majority of approaches fits into the framework of generalized additive models (GAMs), see Hastie and Tibshirani (1990). GAMs assume that the distribution of the response variable y given covariates $x = (x_1, \ldots, x_p)'$ belongs to an exponential family. A link function g relates the expected value μ of y to the covariates through

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

where $f_1, \ldots, f_p$ are known, possibly nonlinear functions. The additive decomposition of the covariate effects in (1) allows for good interpretability of the effects and circumvents the curse of dimensionality (Hastie and Tibshirani 1990). There are two main approaches for modeling the functions $f_1, \ldots, f_p$, 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, $B_1, \ldots, B_K$,

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

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ 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 alternatively, 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 functions 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

P-splines as introduced by Eilers and Marx (1996) approximate the unknown function f by a polynomial spline of degree l with equally spaced knots

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

over the domain of x. Because of the equal spacing of knots $\kappa_j = x_{min} + h \cdot j$, $j = 0, \ldots, m$, where $h = (x_{max} - x_{min})/m$. A spline has the following two properties:

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

A spline can be written in terms of a linear combination of $K = m + l$ basis functions (De Boor 2001). The most widely used bases are the truncated power series basis and the B-spline basis. Using a truncated power series basis the function f is

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_l x^l + \sum_{j=1}^{m-1} \beta_{l+j} t_j(x, l), \tag{3}$$

where

$$t_j(x,l) = (x-\kappa_j)_+^l = \begin{cases} (x-\kappa_j)^l & x > \kappa_j \\ 0 & \text{else.} \end{cases}$$

In a simple regression spline approach, the unknown regression coefficients β_k are estimated using standard inference techniques for linear or generalized linear models. The crucial problem with such regression splines is the choice of the number and the position of the knots. A small number of knots may result in a function space which is not flexible enough to capture the variability of the data. A large number may lead to overfitting. As a remedy Eilers and Marx (1996) propose to define a large number of knots (usually between 10 and 40) to ensure enough flexibility. Sufficient smoothness of the fitted curve is achieved through a roughness penalty on the regression coefficients.

Using a truncated power series basis, overfitting is prevented using a quadratic ridge type penalty

$$P(\lambda) = \lambda \sum_{j=1}^{m-1} \beta_{l+j}^2, \tag{4}$$

leading to the penalized least squares criterion

$$PLS(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{j=1}^{m-1} \beta_{l+j}^2 \tag{5}$$

to be minimized with respect to $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{K-1})'$ in (3). Smoothness is controlled by the "smoothing parameter" $\lambda \geq 0$. Small values of λ produce a close fit to the data, while large values of λ yield smooth function estimates.

Despite their simplicity P-splines based on a truncated power series basis in combination with penalty (4) are rarely used in practice, due to the numerical instability of the highly collinear basis functions. In all available P-spline software packages (e.g. `mgcv` of R, BayesX) a local B-splines basis is used instead. There is a close relationship between B-splines and truncated polynomials as B-splines can be computed as differences of truncated powers (Eilers and Marx 2004). For instance B-spline basis functions of degree one are computed as

$$B_j(x,1) = t_{j-2}(x,1) - 2t_{j-1}(x,1) + t_j(x,1) = \Delta^2 t_j(x,1),$$

with t_j defined in (2.2). B-spline basis functions of degree l are given by

$$B_j(x,l) = -1^{l+1}\Delta^{l+1}t_j(x,l)/(h^l l!).$$

For non-equally spaced knots the formulas for computing B-splines are more involved and based on so called divided differences (De Boor (2001)). Extra knots $\kappa_{-l},\ldots,\kappa_{-1}$ left to κ_0 and $\kappa_{m+1},\ldots,\kappa_{m+l}$ right to κ_m are required, so that the truncated polynomials in the above formula are properly defined to compute all basis functions B_j close to the left and right borders. Now the spline f may be written as

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

The local basis also gives rise to alternative penalization. The widely used approach by Eilers and Marx (1996) penalizes the sum of squared d-th order differences

$$P(\lambda) = \lambda \sum_{k=d+1}^{K} \left(\Delta^d \beta_k\right)^2 \tag{6}$$

were Δ^d is the difference operator of order d. The default in most implementations (e.g. `mgcv` in R, BayesX) is $d = 2$, leading to the penalized least squares criterion

$$PLS(\boldsymbol{\beta},\lambda) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \sum_{k=d+1}^{K} \left(\Delta^d \beta_k\right)^2. \tag{7}$$

The penalized least squares criteria (5) and (7) are equivalent, i.e. they produce the same estimates, when $d = l + 1$ and $\lambda_{tr} = (l!h!)\lambda_b$ where λ_{tr} is the smoothing parameter in (5) and λ_b is the smoothing parameter in (7) (Scholz 2004).

A closely related approach by O'Sullivan (1986) replaces the discrete penalty (6) by the integral of squared second order derivatives,

$$P(\lambda) = \int (f''(x))^2\, dx.$$

While P-splines are defined on a somewhat heuristic basis, they work well in practice and are widely used. Recently, researchers have also studied their asymptotic properties, see e.g. Kauermann et al. (2009).

P-splines are closely related to smoothing splines (Reinsch 1967, Green and Silverman 1994, Hastie and Tibshirani 1990). A smoothing spline is derived from the penalized least squares criterion

$$PLS(\lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 \, dx \tag{8}$$

where f is assumed to be a smooth function with two continuous derivatives. The function f that minimizes (8) 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 (8) 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. This low rank approximation is also the default smoother in the `mgcv` package of R (see below for more comments on available software) that we use in our simulation study and the data example.

The choice of the smoothing parameter λ strongly affects the resulting 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 (Wood 2000, Wood 2003, Wood 2004, Wood 2006b, 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, Wand, and Carroll 2003, Wand 2003, Fahrmeir, Kneib, and Lang 2004, Kauermann, Krivobokova, and Fahrmeir 2009). Finally, a fully Bayesian version of P-splines in combination with Markov chain Monte Carlo simulation techniques can be used to simultaneously estimate the regression coefficients and the smoothing parameters (Lang and Brezger 2004, Brezger and Lang 2006, Jullion and Lambert 2007).

For all above mentioned approaches easy to use statistical software is available. Smoothing parameter estimation based on minimizing GCV can be done in a very efficient, fast and stable way using the `mgcv` package of R, see Wood (2006a) and Wood (2006b). Estimation via REML is supported in the current version of `mgcv` (without resorting to the connection with mixed models) or within the software package BayesX (Brezger et al. 2005 and 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 splines can also be estimated using the function coxph of the R package survival.

2.3 Fractional Polynomials (FPs)

FPs approximate the 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

$$\begin{aligned} h_0(x) &= 1 \\ h_j(x) &= \begin{cases} x^{p_j} & p_j \neq p_{j-1} \\ h_{j-1}(x)\, ln(x) & p_j = p_{j-1}. \end{cases} \end{aligned} \tag{9}$$

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. $M = 2$, are the default setting in all available implementations of FPs. 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. $M = 2$.

An obvious limitation of the definition (9) is the requirement $x > 0$ due to $x^0 := ln(x)$. A covariate with negative values is automatically shifted in implementations 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 application.

For prespecified order M the regression parameters β_j and the polynomial powers $p_j, j = 1, \ldots, M$ are estimated by an algorithm described in Sauerbrei and Royston (1999) and Ambler and Royston (2001), and outlined here for FPs of order 2:

- Test the best fitting FP of order 2 against the null model using a χ^2 distributed test statistic with 4 degrees of freedom (dfs). In case of non-significance, the algorithm terminates and the null model is selected.

- Test the best fitting FP of order 2 against a linear fit using a χ^2 distributed test statistic with 3 df. In case of non-significance, a linear fit for x is assumed to be adequate and the algorithm terminates.

- Test the best fitting FP of order 2 against the best fitting FP of order 1 based on a χ^2 distribution with 2 df. In case of significance the order 2 FP, otherwise the order 1 FP is chosen as the best fit.

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).

3 Simulation study

3.1 Simulation setup

We compare FPs and P-splines in extensive simulations for continuous, binary and survival outcomes. We applied FPs with degrees $M = 2$ (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 2006b) using the default smoother, which is a low rank approximation to the smoothing spline, see also section 2.2. We used generalized cross validation (GCV, the default in mgcv) and restricted maximum likelihood (REML) to select smoothing parameters. Survival models are not supported in mgcv we thus used the R package coxph and the software BayesX (remlreg objects) to fit these models. coxph uses AIC for smoothing parameter selection, while BayesX uses REML.

The comparison was based on data simulated from the following functions of the covariate x that are also depicted in Figure 1:

$$\begin{array}{ll} \text{Linear:} & f_1(x) = -0.9x \\ \text{Quadratic:} & f_2(x) = 0.7 \cdot (x - 2.5)^2 \\ \text{Localmode:} & f_3(x) = 24x \cdot \exp(-2x) \\ \text{Doublemode:} & f_4(x) = 1.3 \cdot (24x \cdot \exp(-2x) + 0.11 \cdot x^2) \end{array} \tag{10}$$

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

For each function f_j in (10), 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 choose 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) A survival model (similar to Bender, Augustin, and Blettner 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 $C \sim Exp(0.2)$.

iv) Gaussian, Binomial and survival models with the additive predictor $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$, where each x_j was comprised of one hundred equally spaced points between 0.05 to 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. In summary, we compared the performance of the following approaches: FP2, FP4, P-splines with GCV, REML for continuous and binary outcomes, and FP2, FP4, P-splines based on AIC, REML for survival data.

The goodness of the 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 $S = 100$.

3.2 Gaussian responses

Figure 2 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 here but are available at

http://www.uibk.ac.at/statistics/personal/lang/publications/fp_ sim_ summary.pdf.

All estimators are unbiased for the linear and quadratic functions in (10) for all choices of sample size, SNR and model type (single or additive predictor). FP4s and the P-spline estimators also showed very little bias for the localmode and doublemode function. An exception is the case $n = 100$, here these estimators are more biased, particularly at the modes of the functions. As expected, the bias decreased for large SNR (figures not shown). A inspection of some individual estimates (figure 3) reveals a tendency to underfitting for FP4s for small sample sizes ($n = 100$), whereas P-splines 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 function were considerably biased for all sample sizes and values of σ. The observed patterns are also reflected in the MSE estimates (table 1). The estimates based on FP4 resulted in a lowest $log(\sqrt{MSE})$, followed closely by P-splines based on GCV and REML. FP2s, however, had a considerably higher $log(\sqrt{MSE})$ for the localmode and doublemode function.

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 and P-splines based on GCV and REML were closer to the nominal level (with coverage rates around 85-90%) than FP2. The coverage decreased with sample size for FP4 and the P-splines estimators, due to too narrow confidence intervals (figure 4). The undercoverage of FP2 reflects lack of fit. For the quadratic and linear function the nominal level was kept by the P-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 the P-splines based on GCV (figure 5), especially for small sample sizes $n = 100$ and $n = 500$. This is illustrated by figure 5 a) which shows a particular FP4 estimate for the doublemode function f_4 in (10) and scaling factor $c = 0.75$. Results are somewhat 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 figure 5), although the FP based estimates still miss important features of the exposure curves in many cases, see panel c).

The P-spline estimator based on GCV also reveals convergence problems showing sometimes extremely rough estimated functions, see figure 5, panel d). 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, P-splines based on REML do not have convergence problems and almost all estimates produce reasonable results (panel e).

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 P-splines based on GCV and REML. For small sample size, $n = 100$, the greater stability of the P-splines based on REML however 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 and illustrated in figure 4. Again, similar to the Gaussian case, FP4 has better coverage than the other approaches even for the more curved functions doublemode and localmode.

3.4 Survival models

Figure 6 and table 1 show average estimated functions and estimates of $log(\sqrt{MSE})$ for the Cox-proportional hazards model with the additive predictor iv) and $n = 100, 500, 1000$ (the case $n = 2000$ is not shown as results were similar to the setting with $n = 1000$).

For small sample size, $n = 100$, P-splines based on AIC are very rough and the results are not reliable. The most stable and best estimator for small sample size are P-splines based on REML. Acceptable results are also obtained with FP4 while FP2 shows strong bias for the doublemode and localmode function $-f_3$ in (10).

For sample sizes $n \geq 500$ all estimators are almost unbiased for the quadratic and linear functions, f_2 and f_1 respectively, in 10. For the localmode and doublemode functions, FP2 estimators again show strong bias while FP4, P-splines 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 for sample size $n \geq 500$ is the P-spline based on AIC followed by FP4 and the P-spline based on REML.

Average coverage rates of pointwise credible intervals are typically far beyond the nominal level (table 2). Only P-splines with smoothing parameter chosen via AIC for the quadratic and linear function produced adequate coverage. The reason for the undercoverage of FP4 is exemplified in figure 4. We observe that undercoverage is a result of confidence intervals that become narrower towards the center of the covariate support although the observations are uniformly distributed over the whole range. In the center the confidence interval almost collapses to a point. This phenomenon was not observed for FPs with Gaussian and binomial responses. The undercoverage of FP2 and P-splines with REML is caused by the biased estimates.

3.5 Simulation summary

We briefly summarize our findings regarding the performance of the methods.

- *Quality of fit:* The performance of the P-spline based estimators and fractional polynomials of degree 4, FP4, is similar, with FP4 resulting in slightly lower MSE. Fractional polynomials of degree 2, FP2s, do not adequately capture relationships that are more complex than quadratic.

- *Coverage rate of confidence intervals:* Coverage rates of FP4 are close to or above the nominal level for Gaussian and binomial outcomes, but not for survival models. Coverage rates for the P-spline estimators are often below the nominal level for the more complex functions for all types of outcomes. In survival models P-splines based on REML shows undercoverage even for the less curved functions. Coverage rates of FP2 are often below the nominal level, due to the large bias of the estimator.

- *Stability of estimators:* The most stable estimator are P-splines based on REML. Particularly for small sample sizes ($n = 100$) and more complex functions, the FP estimators strongly depend on the covariate support. P-splines with GCV and (for survival models) AIC are also prone to unstable behavior, i.e. bumpy estimates, for small sample sizes. Note that the similar behavior of GCV and AIC is not surprising as both goodness of fit criteria are equivalent in large samples.

- *Computing time:* The `mgcv` function of R used in the simulations of Gaussian and binomial responses is extremely fast, producing results in (milli)seconds. The FP estimators are sometimes up to 200 times slower (table 3). For survival models computing times of all estimation procedures are similar (table 4). However, computing time is a function of both the estimation algorithm as well as the implementation. In particular, for `coxph` of R and BayesX there is room for improvement, as `coxph` uses a simple grid search to find the AIC best model and BayesX uses a standard Newton algorithm for optimization. The limitation for fractional polynomials seems to be the computer intensive stepwise selection type estimation algorithm.

4 Data example

In this section 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/). Our analysis focuses on the impact of malnutrition in approximately 30000 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

$$z = \frac{H - MH}{\sigma}, \tag{11}$$

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 following additive model to the data

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

where $f_1, \ldots, f_8$ are unknown nonlinear functions of the child's age (*age*), the number of vaccinations after the child's birth (*vacnumb*), the birth order (*border*), the mother's years of education (*educm*), the mother's body mass index (*bmimo*), the mother's age at birth (*biage*), 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 with common variance σ^2 across subjects. 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. The smoothing parameters were estimated by GCV and REML. Since GCV and REML produced similar results we only present those based on REML. The spline based estimates were compared to FP2 and FP4 estimates, obtained from the `mfp` function in STATA.

Figure 7 presents the estimated functions based on the three estimators REML (solid line), FP2 (dotted lines) and FP4 (dashed lines). For REML pointwise 95% confidence intervals are included. Overall the estimated functional forms for individual variables agreed with the literature (e.g. Belitz et al. 2010). P-splines, FP2 and FP4, produced very similar estimates of the effects of three variables, *border*, *educmy* and *biage*. However, for *age*, *bmimo* and *hhs* we observed pronounced differences between the methods (figure 8). The top row of figure 8 shows that the estimated bump around age 25-30 months obtained with the spline estimator captures a distinct feature in the data, and is not an artefact of the method. 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 equation (11). 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 REML fitted model for the India dataset. Figure 9 further highlights that FP2 and FP4 are not able to detect the underlying structure of the effect of age on outcome.

The three methods also differ in the estimated effects for *bmimo* and *hhs*, although the differences are less pronounced. The spline based estimator adapts better to the data revealing monotonic decreasing respectively increasing effects of *bmimo* and *hhs* rather than the almost linear fits obtained with FP2 and FP4. However, the partial residuals show that all approaches give reasonable estimates (middle and bottom panel of figure 8).

Of note is the estimated effect of FP2 for *ai*, in the right bottom panel of figure 8. 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 the figure. Indeed, if we replace *ai* by $ai + 2$, and re-fit the model, this artefact disappears, see figure 10.

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 simpler 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 P-splines and fractional polynomials (FPs), two widely used smoothing techniques in empirical science, 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 can not be detected by fractional polynomials of any degree. We also showed that FPs are prone to artefacts because of the dependence of results on the covariate support, while P-splines based on GCV (or AIC in the survival models) reveal sometimes wiggly estimates. The most stable estimators were produced by P-splines 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. The latter may be of great value in subsequent analysis 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, see for instance Miller (2002). Hence there is need for alternative estimation methods. A promising approach is a Bayesian version of FPs that has been published recently by Sabanés Bové and Held (2010). Another problem with FPs, that has been ignored in the literature is the sometimes strong dependence of results on the covariate range. A possible remedy could be a mapping of observed covariate values in a "save" interval such that the observed problems are less likely to happen.

Although, the behavior of splines based estimators is better understood, the best criterion or approach for smoothing parameter selection is still not entirely clear. Our findings suggest that selection of the smoothing parameter based on REML is more stable than GCV and AIC, however, to date no theoretical results exist to support that finding.

References

Ambler, G. and P. Royston (2001). Fractional polynomial model selection procedures: Investigation of type i error rate. *Journal of Statistical Computation and Simulation 69*, 89–108.

Andre, C., A. P. de Barros, L. A. A. Pereira, and P. H. N. Saldiva (2004). The distributed lag between air pollution and intrauterine mortality. *Epidemiology 15*, 51.

Beatty, T. (2009). Semiparametric quantile engel curves and expenditure elasticities: a penalized quantile regression spline approach. *Applied Economics 12*, 1533–1542.

Belitz, C., A. Brezger, T. Kneib, and S. Lang (2009). Bayesx manuals. Technical report, Department of Statistics, University of Munich. Available at http://www.stat.uni–muenchen.de/˜bayesx.

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

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

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

Brezger, A., T. Kneib, and S. Lang (2005). Bayesx: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software 14*, 1–22.

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

De Boor, C. (2001). *A practical Guide to Splines.* Springer, New York.

Eilers, P. and B. Marx (2004). plines, knots, and penalties. Technical report, Department of Experimential Statistics LSU. Available at http://www.stat.lsu.edu/faculty/marx/.

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

Eisen, E. A., I. Agalliu, B. A. Coull, S. W. Thurston, and H. Checkoway (2004). Smoothing in occupational cohort studies: an illustration based on penalized splines. *Occupational and Environmental Medicine 61*, 854–860.

Ellner, S. P., Y. Seifu, and R. H. Smith (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology 83*, 2256–2270.

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

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2. Auflage).* Springer, New York.

Finch, B. K., N. Lim, W. Perez, and D. Phuong Do (2007). Toward a population health model of segmented assimilation: The case of low birth weight in los angeles. *Sociological Perspectives 50*, 445–468.

Govindarajulu, U. S., D. Spiegelman, S. W. Thurston, and E. A. Ganguli, B. amd Eisen (2007). Comparing smoothing techniques in cox models for exposure–response relationships. *Stat Med 26*(20), 3735–52.

Green, P. and P. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

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

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

Henley, A. and J. Peirson (2001). Non-linearities in electricity demand and temperature: Parametric versus non-parametric methods. *Oxford Bulletin of Economics and Statistics 59*, 149–162.

Jullion, A. and P. Lambert (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., T. Krivobokova, and L. Fahrmeir (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society B 71*, 487–503.

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

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

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science 1*, 505–527.

Peterson, D. A., L. J. Grossback, J. A. Stimson, and A. Gangl (2003). Congressional response to mandate elections. *American Journal of Political Science 47*, 411–426.

Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik 10*, 177–183.

Royston, P. and W. Sauerbrei (2005). Building multivariable regression models with continuous covariates in clinical epidemiology – with an emphasis on fractional polynomials. *Methods Inf Med. 44*, 561–71.

Royston, P. and W. Sauerbrei (2008). *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley.

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

Sabanés Bové, D. and L. Held (2010). Bayesian fractional polynomials. *Statistics and Computing, to appear*.

Sauerbrei, W., C. Meier-Hirmer, A. Benner, and P. Royston (2006). Multivariable regression model building by using fractional polynomials: Description of sas, stata and r programs. *Computational Statistics and Data Analysis 50*, 3464–85.

Sauerbrei, W. and P. Royston (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.

Scholz, T. (2004). *Flexible Modellierung kategorialer Regressionsvariablen*. Dr. Hut Verlag.

Shlipak, M. G., R. Katz, M. J. Sarnak, L. F. Fried, A. B. Newman, C. Stehman-Breen, S. L. Seliger, B. Kestenbaum, B. Psaty, R. P. Tracy, and D. S. Siscovick (2006). Cystatin c and prognosis for cardiovascular and kidney outcomes in elderly persons without chronic kidney disease. *Ann Intern Med 145*(4), 237–46.

Stansfeld, S. A., B. Berglund, C. Clark, I. Lopez-Barrio, P. Fischer, E. Ohrström, M. M. Haines, J. Head, S. Hygge, I. van Kamp, and B. F. Berry (2005). Ranch study team. aircraft and road traffic noise and children's cognition and health: a cross-national study. *Lancet 365*(9475), 1942–9.

Strasak, A. M., S. Lang, T. Kneib, L. J. Brant, J. Klenk, W. Hilbe, W. Oberaigner, E. Ruttmann, L. Kaltenbach, H. Concin, G. Diem, K. P. Pfeiffer, H. Ulmer, and V. S. Group. (2009). Use of penalized splines in extended cox–type additive hazard regression to flexibly estimate the effect of time–varying serum uric acid on risk of cancer incidence: a prospective, population–based study in 78,850 men. *Annals of Epidemiolgy 19*, 15–24.

Ugarte, M. D., T. Goicoa, and A. F. Militino (2009). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics*, 1533–1542.

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

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

Wood, S. (2006a). R - manual: The mgcv package, version 1.3 - 22. Technical report.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society B 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. (2006b). *Generalized Additive Models: An Introduction with R.* Chapman & Hall.

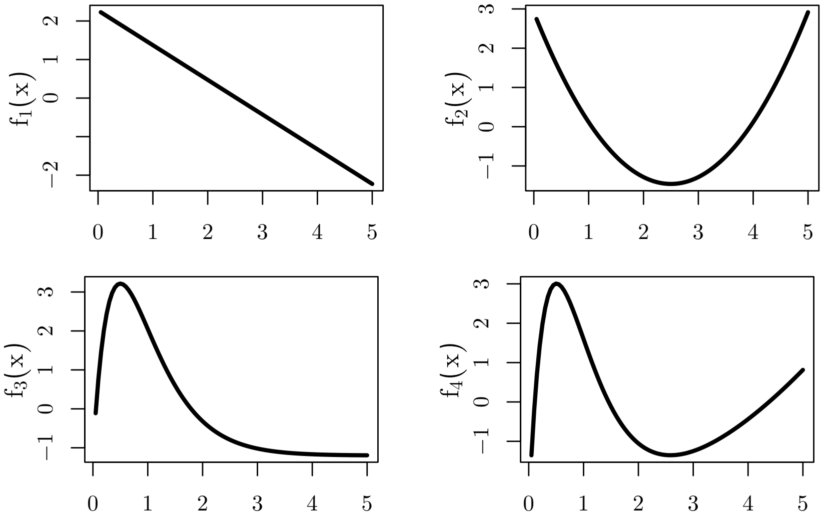

Figure 1: Functions used for simulations.

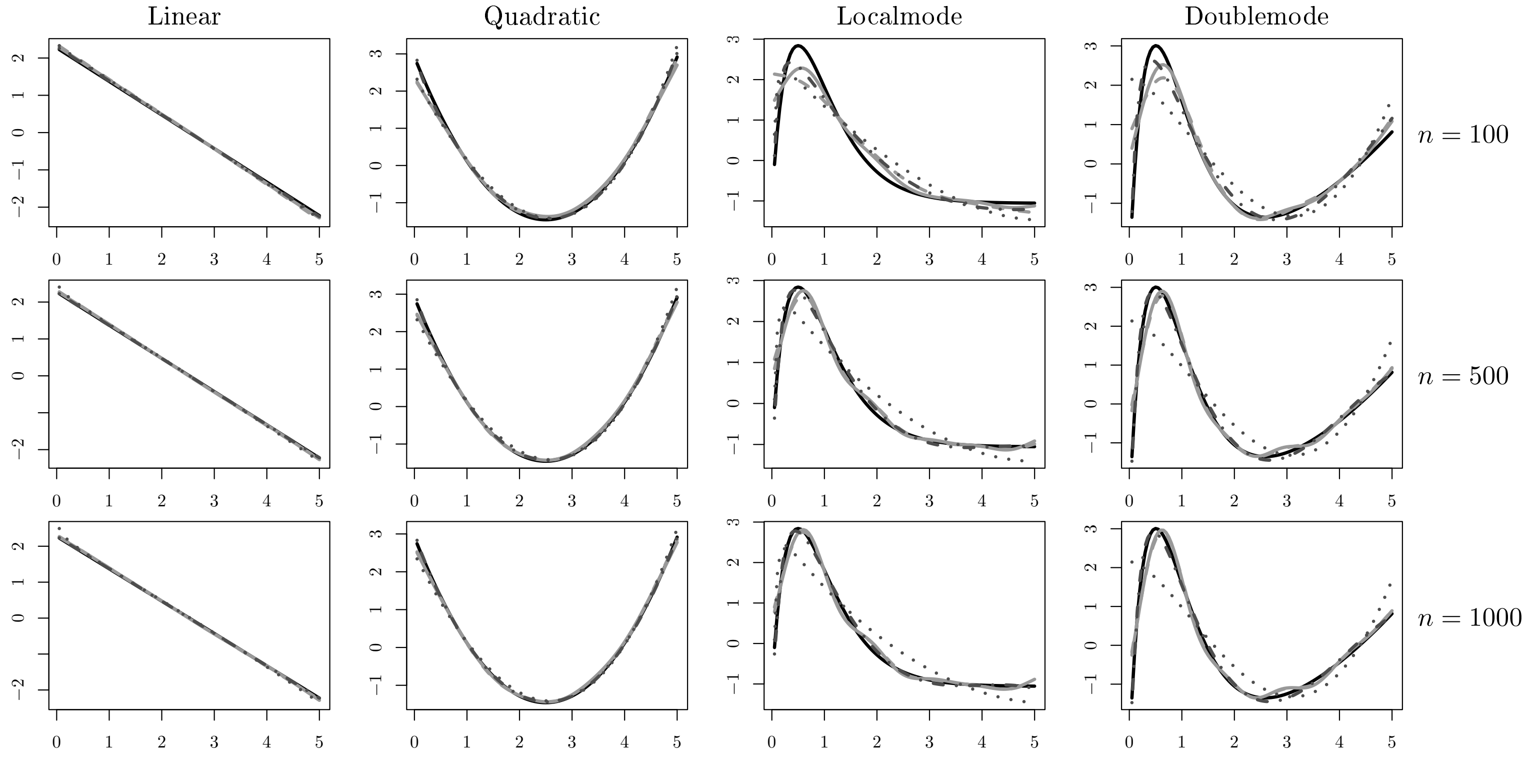

Figure 2: Gaussian additive model, $\sigma = 0.735$: True curves (black solid lines) and average estimated curves (grey solid lines GCV, grey dashed lines REML, black dots FP2 and black dashed lines FP4 estimates).

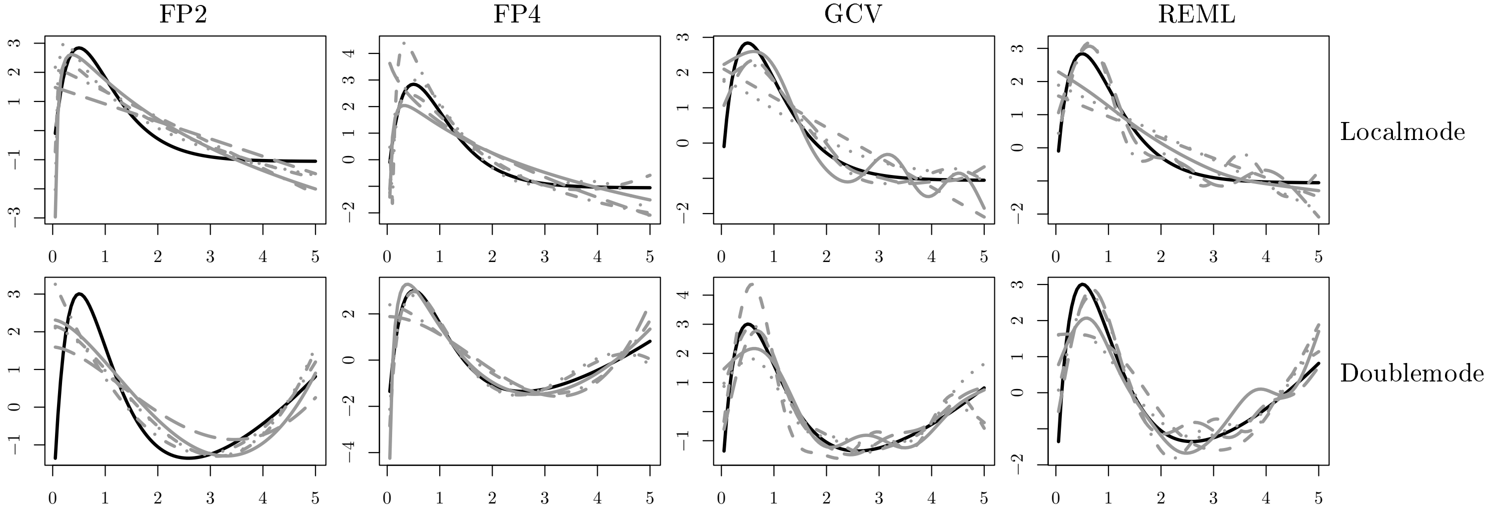

Figure 3: Gaussian additive model, $\sigma = 0.735$, $n = 100$*: Some function estimates for the localmode and doublemode function. Shown are the 2.5,10,50,90 and 97.5 percent best fits according to the MSE measure. The black solid lines represent the true functions, the grey solid lines the median and grey dashed lines the other quantiles.*

f_j	n	Gaussian				Binomial Logit				Survival			
		FP2	FP4	GCV	REML	FP2	FP4	GCV	REML	FP2	FP4	AIC	REML
f_1	100	**0.012**	**0.012**	0.021	0.018	0.043	**0.042**	0.146	0.074	**0.015**	**0.015**	4.479	0.029
	500	0.004	**0.002**	0.006	0.004	0.014	**0.011**	0.03	0.028	**0.02**	**0.02**	0.05	0.028
	1000	0.005	**0.001**	0.003	0.002	0.006	**0.005**	0.013	0.011	**0.016**	**0.016**	0.021	0.034
	2000	0.008	**0.001**	0.002	**0.001**	0.004	**0.003**	0.007	0.004	0.022	0.022	**0.01**	0.065
f_2	100	0.07	**0.063**	0.072	0.068	0.265	0.815	0.338	**0.189**	**0.041**	**0.041**	4.488	0.047
	500	0.038	**0.011**	0.021	0.022	0.05	**0.048**	0.062	0.057	**0.026**	**0.026**	0.049	0.039
	1000	0.019	**0.006**	0.013	0.012	0.029	**0.022**	0.034	0.029	0.023	0.023	**0.02**	0.045
	2000	0.02	**0.003**	0.008	0.008	0.015	**0.01**	0.019	0.016	0.021	0.021	**0.011**	0.066
f_3	100	0.308	0.222	**0.21**	0.251	0.39	0.403	0.409	**0.319**	0.153	0.153	5.115	**0.12**
	500	0.193	**0.04**	0.057	0.059	0.108	**0.107**	0.134	0.156	0.079	0.079	**0.051**	0.09
	1000	0.188	**0.02**	0.039	0.04	0.101	**0.032**	0.076	0.082	0.086	0.086	**0.02**	0.082
	2000	0.186	**0.014**	0.031	0.03	0.09	**0.017**	0.054	0.051	0.092	0.092	**0.011**	0.18
f_4	100	0.49	**0.149**	0.225	0.252	0.763	0.768	0.56	**0.4**	0.184	0.184	4.587	**0.138**
	500	0.461	**0.037**	0.076	0.079	0.32	**0.085**	0.159	0.189	0.137	0.137	**0.05**	0.074
	1000	0.459	**0.025**	0.058	0.06	0.292	**0.031**	0.1	0.107	0.145	0.145	**0.021**	0.066
	2000	0.458	**0.021**	0.049	0.049	0.289	**0.017**	0.062	0.059	0.144	0.144	**0.012**	0.187

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 boldface represent the respective smallest median of four algorithms for each row and distribution.*

f	n	Gaussian				Binomial Logit				Survival			
		FP2	FP4	GCV	REML	FP2	FP4	GCV	REML	FP2	FP4	AIC	REML
f_1	100	0.999	0.999	0.939	0.944	0.988	0.987	0.94	0.939	0.829	0.829	0.94	0.875
	500	0.992	0.999	0.944	0.947	0.986	0.998	0.932	0.925	0.264	0.663	0.932	0.494
	1000	0.973	0.998	0.92	0.922	0.994	0.996	0.93	0.934	0.087	0.229	0.93	0.319
	2000	0.938	0.997	0.926	0.934	0.993	0.999	0.936	0.956	0.001	0.011	0.936	0.192
f_2	100	0.976	0.979	0.947	0.968	0.893	0.74	0.945	0.956	0.673	0.625	0.945	0.935
	500	0.9	0.98	0.948	0.964	0.964	0.973	0.944	0.953	0.537	0.644	0.944	0.681
	1000	0.823	0.982	0.939	0.957	0.957	0.982	0.946	0.966	0.421	0.516	0.946	0.442
	2000	0.705	0.984	0.924	0.943	0.958	0.988	0.95	0.966	0.296	0.356	0.95	0.24
f_3	100	0.758	0.83	0.82	0.696	0.871	0.827	0.837	0.798	0.424	0.406	0.837	0.685
	500	0.419	0.918	0.9	0.893	0.931	0.948	0.871	0.803	0.316	0.59	0.871	0.638
	1000	0.292	0.914	0.858	0.866	0.774	0.965	0.92	0.89	0.23	0.487	0.92	0.484
	2000	0.2	0.88	0.786	0.802	0.54	0.965	0.895	0.892	0.158	0.334	0.895	0.216
f_4	100	0.764	0.918	0.9	0.878	0.725	0.717	0.874	0.842	0.42	0.484	0.874	0.767
	500	0.418	0.929	0.869	0.871	0.742	0.942	0.908	0.865	0.394	0.64	0.908	0.77
	1000	0.279	0.903	0.818	0.83	0.64	0.972	0.908	0.889	0.35	0.565	0.908	0.623
	2000	0.199	0.768	0.741	0.756	0.498	0.977	0.884	0.89	0.322	0.429	0.884	0.252

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. Cells corresponding to values below a 92.5% level (undercoverage) are marked with dark grey and values larger than a 97.5% level (overcoverage) with light grey.

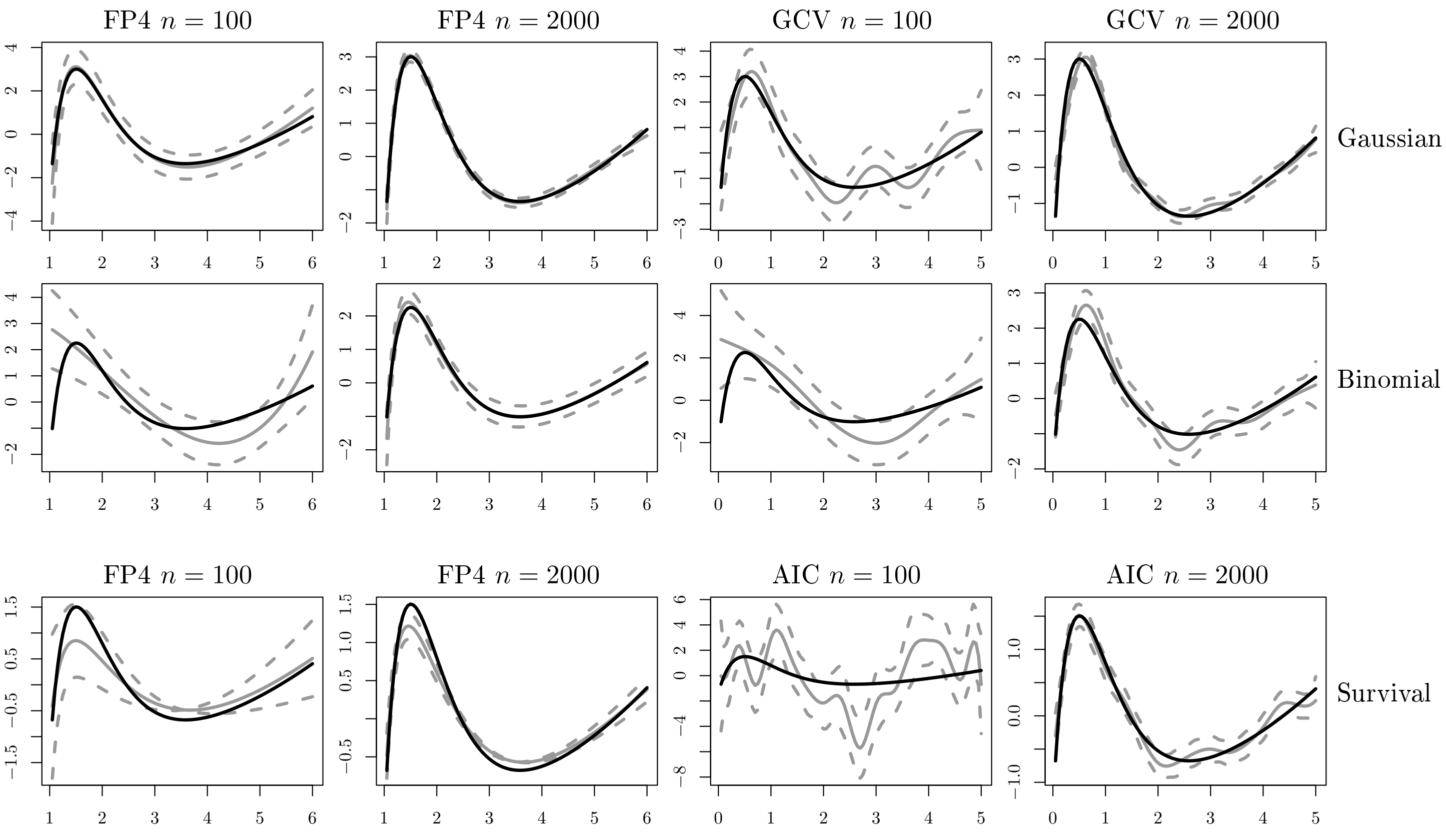

Figure 4: Additive models with medium SNR ($\sigma = 0.735$ for Gaussian responses, scaling factor $c = 0.75$ for Binomial outcome, scaling factor $c = 0.5$ for survival models): Typical 95% pointwise confidence intervals for the doublemode function.

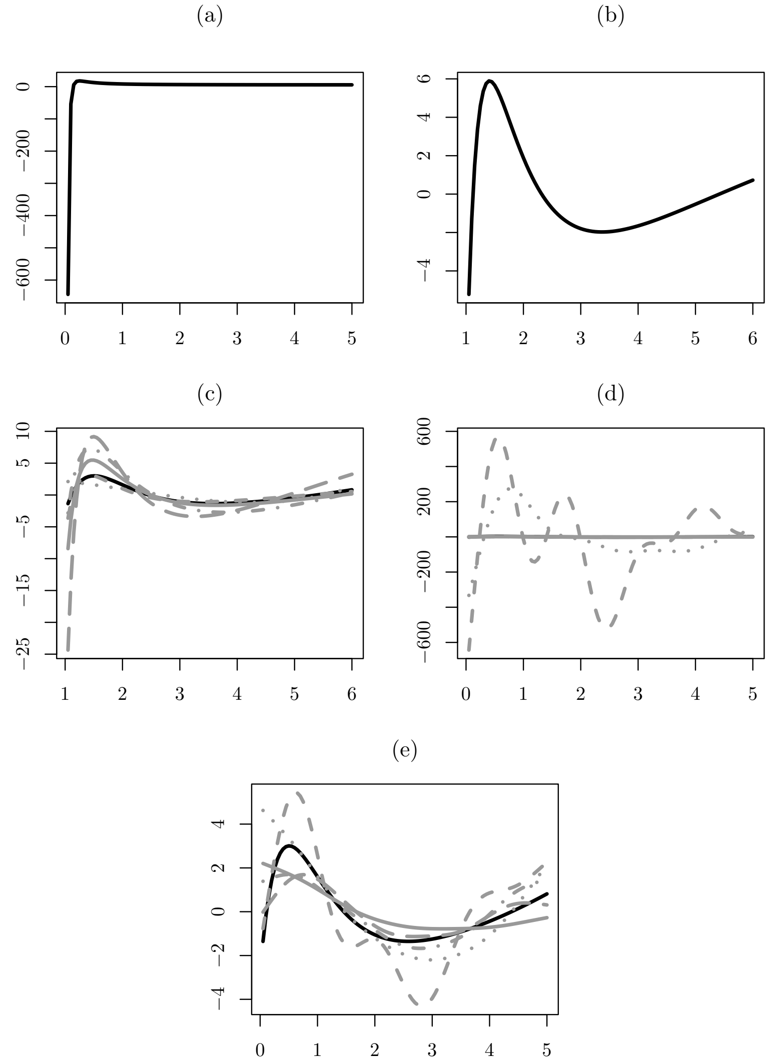

Figure 5: Binomial additive model, scaling factor c = 0.75*: Panel a): FP4 estimate for a particular replication. Panel b): FP4 estimate based on a shift of covariate values by one unit. Panels c)-e): Some individual function estimates for FP4, P-splines based on GCV, P-splines based on REML. Shown are the 2.5,10,50,90 and 97.5 percent 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.*

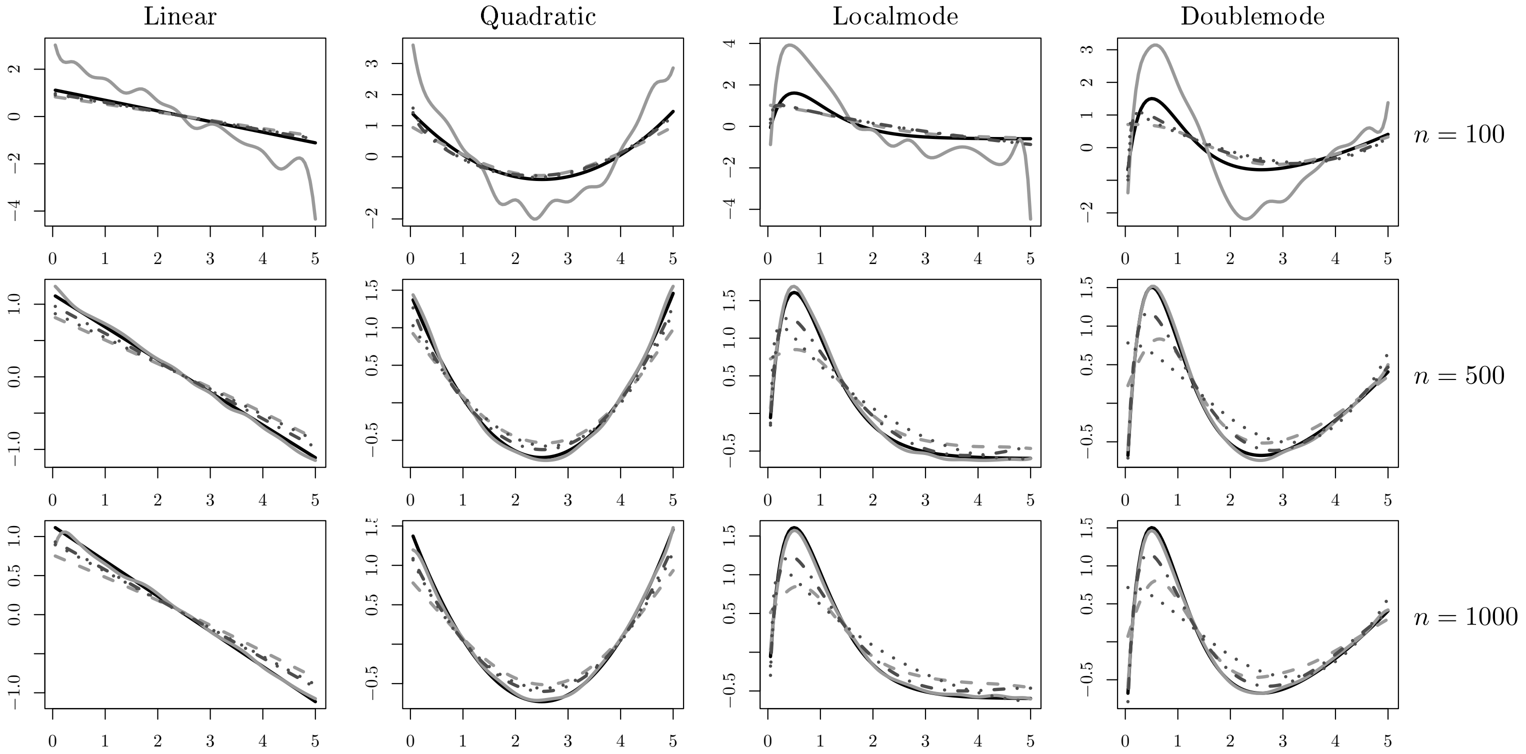

Figure 6: Survival additive model: True functions (black solid lines) and average estimated functions (grey solid lines P-splines based on AIC, grey dashed lines P-splines based on REML, black dots FP2 and black dashed lines FP4 estimates).

REML	GCV	FP2	FP4	Scale	n
1.61	21.3	10.065	148.064	$c = 1$	100
1.75	5.27	7.361	84.763	$c = 0.75$	
2.02	2.52	6.699	70.632	$c = 0.5$	
1.97	2.3	4.927	45.266	$c = 0.25$	
3.42	4.42	28.972	475.539	$c = 1$	500
3.07	4.24	25.176	461.989	$c = 0.75$	
2.66	3.71	23.714	370.146	$c = 0.5$	
2.94	3.88	15.613	145.761	$c = 0.25$	
5.71	6.85	47.018	1060.325	$c = 1$	1000
4.86	7.47	45.826	684.641	$c = 0.75$	
5.51	6.97	42.552	661.518	$c = 0.5$	
4.81	5.89	33.548	413.669	$c = 0.25$	
10.61	11.37	156.296	2380.226	$c = 1$	2000
8.9	11.19	70.122	1668.426	$c = 0.75$	
9.02	12.19	65.174	1317.968	$c = 0.5$	
8.23	11.83	52.296	885.402	$c = 0.25$	

Table 3: Estimation times in seconds for the logit models with additive logit-mean structure based on 10 replications. The results are obtained on a Intel Core2 Duo CPU E6550 processor with 2.33GHz and 3.5GB RAM storage on a Windows XP operating system.

REML	AIC	FP2	FP4	n
180.18	4.28	37.203	458.634	100
154.57	13.48	69.08	1007.509	500
217.68	24.66	110.1	1882.827	1000
330.4	39.25	173.641	2832.071	2000

Table 4: Estimation times in seconds for Cox regression models with $\eta = \sum f_i(x_i)$ based on 10 replications. The results are obtained on a Intel Core2 Duo CPU E6550 processor with 2.33GHz and 3.5GB RAM storage on a Windows XP operating system.

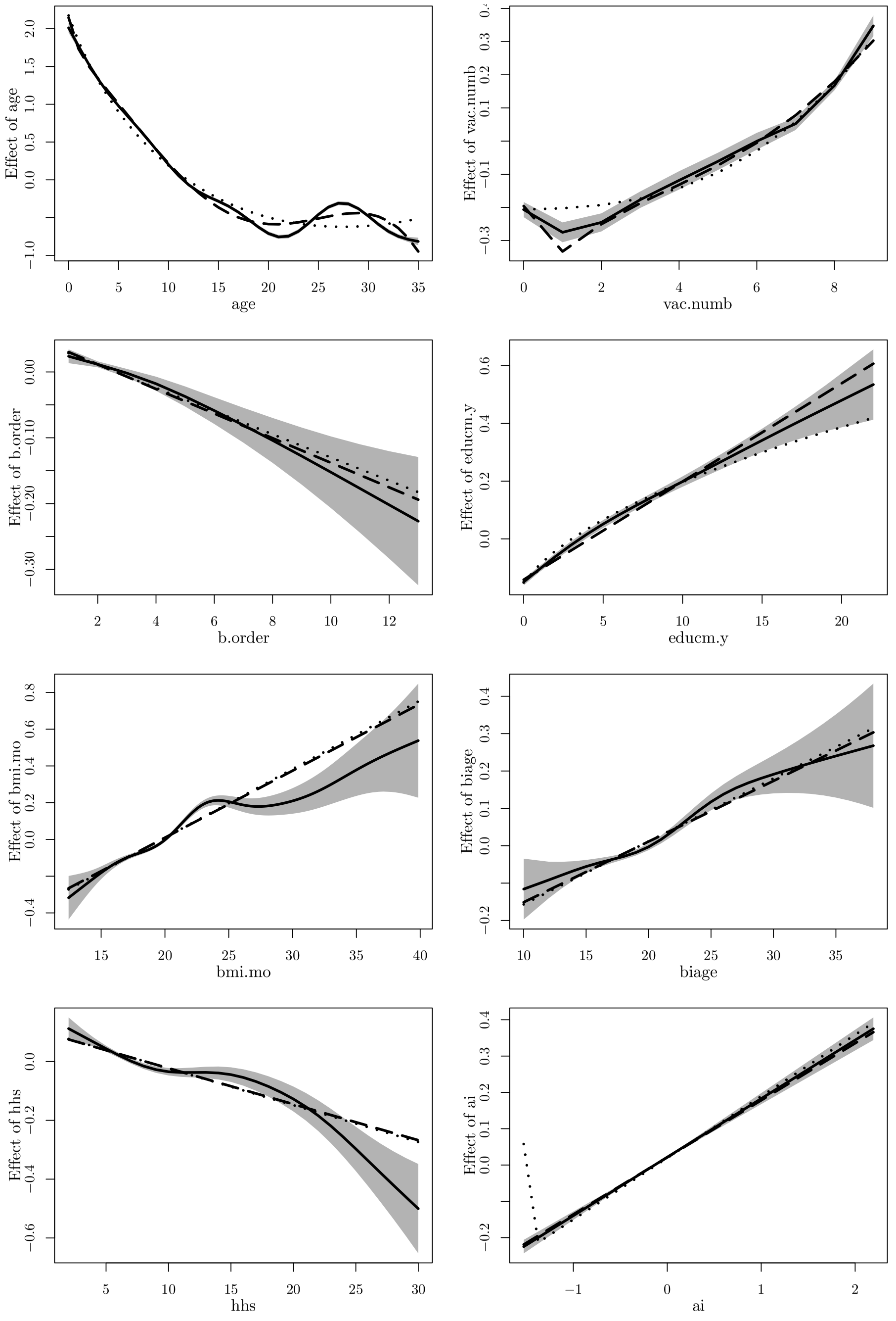

Figure 7: Malnutrition in India: the solid lines represent estimates for P-splines based on REML and 95% confidence intervals, dotted lines FP2 and wide dashed lines FP4 estimates.

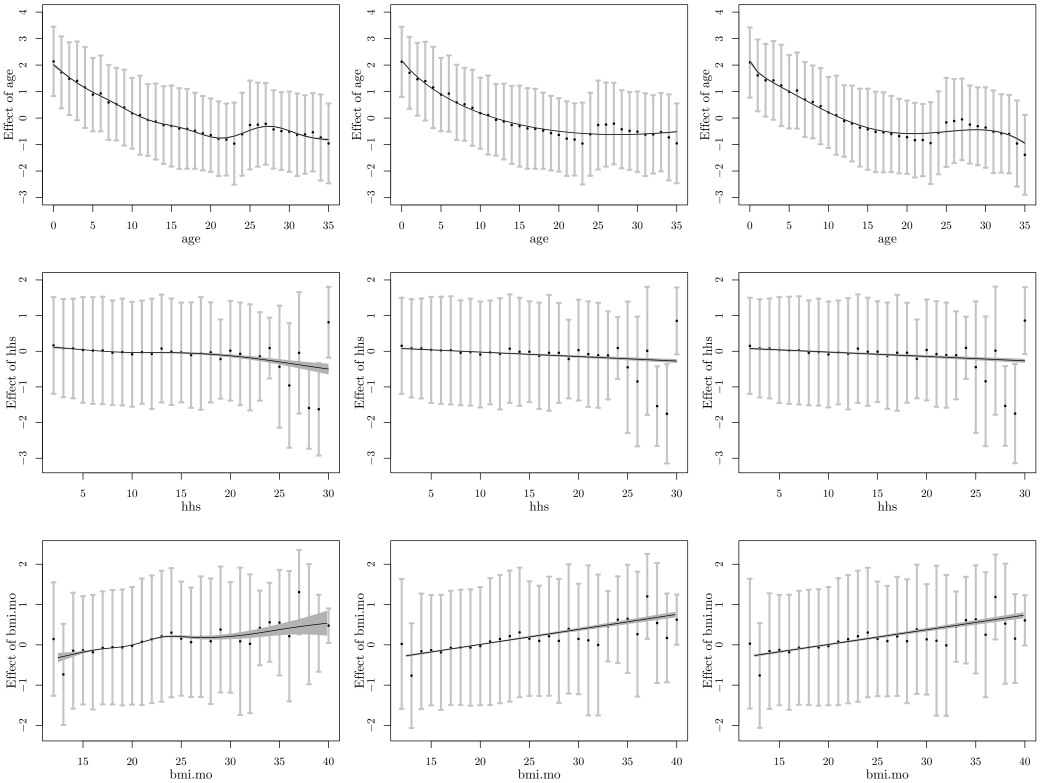

Figure 8: Malnutrition in India: Estimated effects for the age of the child (top row), the household size (middle row) and the mothers body mass index (bottom row). From left to right the estimates correspond to P-splines based on REML, FP2 and FP4. Shown are the estimated functions and 95% confidence intervals and stick plots of means plus/minus standard deviations of the corresponding partial residuals averaged over the distinct (or rounded) covariate values.

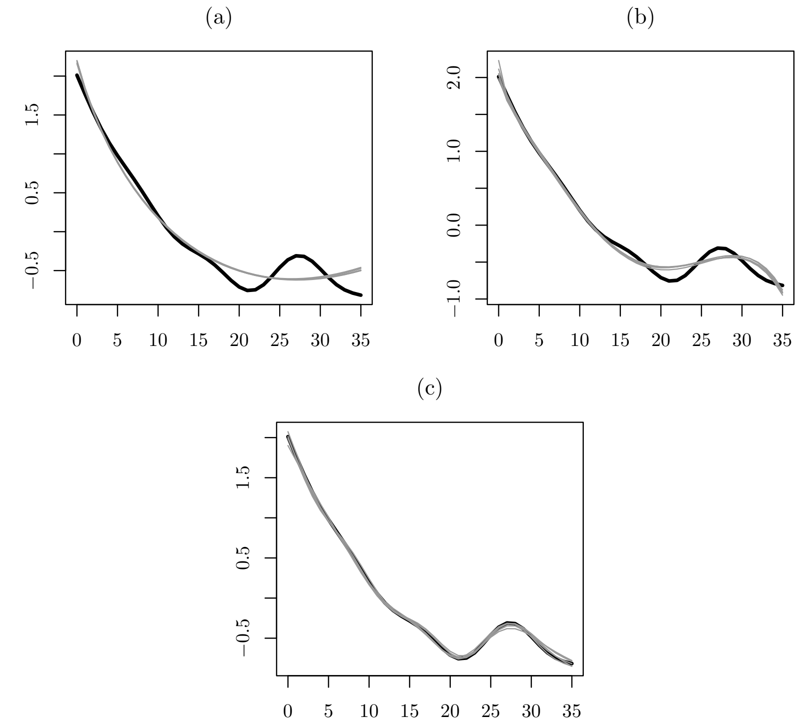

Figure 9: Estimation results for simulated data for covariate age. Shown are typical estimates for FP2 in plot (a), FP4 plot (b) and REML in plot(c). The true function $f(age)$, which was generated by the fitted REML model for the India dataset with model equation $y = f(age) + \varepsilon$ and $\varepsilon \sim N(0, 2.17)$, is represented by the solid black lines and estimates with solid grey lines.

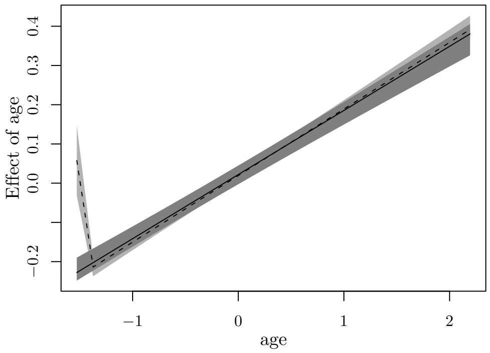

Figure 10: Malnutrition in India: Estimated effects for age based on FP2 with 95% pointwise confidence intervals. The dark shaded area corresponds to the estimate obtained with the original covariate values. The light shaded area is the estimate obtained for the transformed covariate, i.e. the constant 2 is added to the values of ai.

University of Innsbruck – Working Papers in Economics and Statistics
Recent papers

2010-11 **Alexander Strasak, Nikolaus Umlauf, Ruth Pfeiffer and Stefan Lang:** Comparing Penalized Splines and Fractional Polynomials for Flexible Modelling of the Effects of Continuous Predictor Variables

2010-10 **Wolfgang A. Brunauer, Sebastian Keiler and Stefan Lang:** Trading strategies and trading profits in experimental asset markets with cumulative information

2010-09 **Thomas Stöckl and Michael Kirchler:** Trading strategies and trading profits in experimental asset markets with cumulative information

2010-08 **Martin G. Kocher, Marc V. Lenz and Matthias Sutter:** Psychological pressure in competitive environments: Evidence from a randomized natural experiment: Comment

2010-07 **Michael Hanke and Michael Kirchler:** Football Championships and Jersey Sponsors' Stock Prices: An Empirical Investigation

2010-06 **Adrian Beck, Rudolf Kerschbamer, Jianying Qiu and Matthias Sutter:** Guilt from Promise-Breaking and Trust in Markets for Expert Services - Theory and Experiment

2010-05 **Martin Gächter, David A. Savage and Benno Torgler:** Retaining the Thin Blue Line: What Shapes Workers' Intentions not to Quit the Current Work Environment

2010-04 **Martin Gächter, David A. Savage and Benno Torgler:** The relationship between Stress, Strain and Social Capital

2010-03 **Paul A. Raschky, Reimund Schwarze, Manijeh Schwindt and Ferdinand Zahn:** Uncertainty of Governmental Relief and the Crowding out of Insurance

2010-02 **Matthias Sutter, Simon Czermak and Francesco Feri:** Strategic sophistication of individuals and teams in experimental normal-form games

2010-01 **Stefan Lang and Nikolaus Umlauf:** Applications of Multilevel Structured Additive Regression Models to Insurance Data

---

2009-29 **Loukas Balafoutas:** How much income redistribution? An explanation based on vote-buying and corruption. *Revised version forthcoming in Public Choice.*

2009-28 **Rudolf Kerschbamer, Matthias Sutter and Uwe Dulleck:** The Impact of Distributional Preferences on (Experimental) Markets for Expert Services

2009-27 **Adrian Beck, Rudolf Kerschbamer, Jianying Qiu and Matthias Sutter:** Car Mechanics in the Lab - Investigating the Behavior of Real Experts on Experimental Markets for Credence Goods

2009-26 **Michael Kirchler, Jürgen Huber and Thomas Stöckl:** Bubble or no Bubble - The Impact of Market Model on the Formation of Price Bubbles in Experimental Asset Markets

2009-25 **Rupert Sausgruber and Jean-Robert Tyran:** Tax Salience, Voting, and Deliberation

2009-24 **Gerald J. Pruckner and Rupert Sausgruber:** Honesty on the Streets - A Natural Field Experiment on Newspaper Purchasing

2009-23 **Gerlinde Fellner, Rupert Sausgruber and Christian Traxler:** Testing Enforcement Strategies in the Field: Legal Threat, Moral Appeal and Social Information

2009-22 **Ralph-C. Bayer, Elke Renner and Rupert Sausgruber:** Confusion and Reinforcement Learning in Experimental Public Goods Games

2009-21 **Sven P. Jost:** Transfer Pricing Risk Awareness of Multinational Corporations - Evidence from a Global Survey

2009-20 **Andrea M. Leiter and Engelbert Theurl:** The Convergence of Health Care Financing Structures: Empirical Evidence from OECD-Countries

2009-19 **Francesco Feri and Miguel A. Meléndez-Jiménez:** Coordination in Evolving Networks with Endogenous Decay

2009-18 **Harald Oberhofer:** Firm growth, European industry dynamics and domestic business cycles

2009-17 **Jesus Crespo Cuaresma and Martin Feldkircher:** Spatial Filtering, Model Uncertainty and the Speed of Income Convergence in Europe

2009-16 **Paul A. Raschky and Manijeh Schwindt:** On the Channel and Type of International Disaster Aid

2009-15 **Jianying Qiu:** Loss aversion and mental accounting: The favorite-longshot bias in parimutuel betting

2009-14 **Siegfried Berninghaus, Werner Güth, M. Vittoria Levati and Jianying Qiu:** Satisficing in sales competition: experimental evidence

2009-13 **Tobias Bruenner, Rene Levinský and Jianying Qiu:** Skewness preferences and asset selection: An experimental study

2009-12 **Jianying Qiu and Prashanth Mahagaonkar:** Testing the Modigliani-Miller theorem directly in the lab: a general equilibrium approach

2009-11 **Jianying Qiu and Eva-Maria Steiger:** Understanding Risk Attitudes in two Dimensions: An Experimental Analysis

2009-10 **Erwann Michel-Kerjan, Paul A. Raschky and Howard C. Kunreuther:** Corporate Demand for Insurance: An Empirical Analysis of the U.S. Market for Catastrophe and Non-Catastrophe Risks

2009-09 **Fredrik Carlsson, Peter Martinsson, Ping Qin and Matthias Sutter:** Household decision making and the influence of spouses' income, education, and communist party membership: A field experiment in rural China

2009-08 **Matthias Sutter, Peter Lindner and Daniela Platsch:** Social norms, third-party observation and third-party reward

2009-07 **Michael Pfaffermayr:** Spatial Convergence of Regions Revisited: A Spatial Maximum Likelihood Systems Approach

2009-06 **Reimund Schwarze and Gert G. Wagner:** Natural Hazards Insurance in Europe – Tailored Responses to Climate Change Needed

2009-05 **Robert Jiro Netzer and Matthias Sutter:** Intercultural trust. An experiment in Austria and Japan

2009-04 **Andrea M. Leiter, Arno Parolini and Hannes Winner:** Environmental Regulation and Investment: Evidence from European Industries

2009-03 **Uwe Dulleck, Rudolf Kerschbamer and Matthias Sutter:** The Economics of Credence Goods: On the Role of Liability, Verifiability, Reputation and Competition. *Revised version forthcoming in American Economic Review.*

2009-02 **Harald Oberhofer and Michael Pfaffermayr:** Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996)

2009-01 **Loukas Balafoutas:** How do third parties matter? Theory and evidence in a dynamic psychological game.

2008-27 **Matthias Sutter, Ronald Bosman, Martin Kocher and Frans van Winden:** Gender pairing and bargaining – Beware the same sex! *Revised version published in Experimental Economics, Vol. 12 (2009): 318-331.*

2008-26 **Jesus Crespo Cuaresma, Gernot Doppelhofer and Martin Feldkircher:** The Determinants of Economic Growth in European Regions.

2008-25 **Maria Fernanda Rivas and Matthias Sutter:** The dos and don'ts of leadership in sequential public goods experiments.

2008-24 **Jesus Crespo Cuaresma, Harald Oberhofer and Paul Raschky:** Oil and the duration of dictatorships.

2008-23 **Matthias Sutter:** Individual behavior and group membership: Comment. *Revised Version published in American Economic Review, Vol.99 (2009): 2247-2257.*

2008-22 **Francesco Feri, Bernd Irlenbusch and Matthias Sutter:** Efficiency Gains from Team-Based Coordination – Large-Scale Experimental Evidence. *Revised and extended version forthcoming in American Economic Review.*

2008-21 **Francesco Feri, Miguel A. Meléndez-Jiménez, Giovanni Ponti and Fernando Vega Redondo:** Error Cascades in Observational Learning: An Experiment on the Chinos Game.

2008-20 **Matthias Sutter, Jürgen Huber and Michael Kirchler:** Bubbles and information: An experiment.

2008-19 **Michael Kirchler:** Curse of Mediocrity - On the Value of Asymmetric Fundamental Information in Asset Markets.

2008-18 **Jürgen Huber and Michael Kirchler:** Corporate Campaign Contributions as a Predictor for Abnormal Stock Returns after Presidential Elections.

2008-17 **Wolfgang Brunauer, Stefan Lang, Peter Wechselberger and Sven Bienert:** Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna.

2008-16 **Harald Oberhofer, Tassilo Philippovich:** Distance Matters! Evidence from Professional Team Sports. *Extended and revised version forthcoming in Journal of Economic Psychology.*

2008-15 **Maria Fernanda Rivas and Matthias Sutter:** Wage dispersion and workers' effort.

2008-14 **Stefan Borsky and Paul A. Raschky:** Estimating the Option Value of Exercising Risk-taking Behavior with the Hedonic Market Approach. *Revised version forthcoming in Kyklos.*

2008-13 **Sergio Currarini and Francesco Feri:** Information Sharing Networks in Oligopoly.

2008-12 **Andrea M. Leiter:** Age effects in monetary valuation of mortality risks - The relevance of individual risk exposure.

2008-11 **Andrea M. Leiter and Gerald J. Pruckner:** Dying in an Avalanche: Current Risks and their Valuation.

2008-10 **Harald Oberhofer and Michael Pfaffermayr:** Firm Growth in Multinational Corporate Groups.

2008-09 **Michael Pfaffermayr, Matthias Stöckl and Hannes Winner:** Capital Structure, Corporate Taxation and Firm Age.

2008-08 **Jesus Crespo Cuaresma and Andreas Breitenfellner:** Crude Oil Prices and the Euro-Dollar Exchange Rate: A Forecasting Exercise.

2008-07 **Matthias Sutter, Stefan Haigner and Martin Kocher:** Choosing the carrot or the stick? – Endogenous institutional choice in social dilemma situations. Revised version forthcoming in *Review of Economic Studies*.

2008-06 **Paul A. Raschky and Manijeh Schwindt:** Aid, Catastrophes and the Samaritan's Dilemma.

2008-05 **Marcela Ibanez, Simon Czermak and Matthias Sutter:** Searching for a better deal – On the influence of group decision making, time pressure and gender in a search experiment. *Revised version published in Journal of Economic Psychology, Vol. 30 (2009): 1-10.*

2008-04 **Martin G. Kocher, Ganna Pogrebna and Matthias Sutter:** The Determinants of Managerial Decisions Under Risk.

2008-03 **Jesus Crespo Cuaresma and Tomas Slacik:** On the determinants of currency crises: The role of model uncertainty. *Revised version accepted for publication in Journal of Macroeconomics.*

2008-02 **Francesco Feri:** Information, Social Mobility and the Demand for Redistribution.

2008-01 **Gerlinde Fellner and Matthias Sutter:** Causes, consequences, and cures of myopic loss aversion - An experimental investigation. *Revised version published in The Economic Journal, Vol. 119 (2009), 900-916.*

2007-31 **Andreas Exenberger and Simon Hartmann:** The Dark Side of Globalization. The Vicious Cycle of Exploitation from World Market Integration: Lesson from the Congo.

2007-30 **Andrea M. Leiter and Gerald J. Pruckner:** Proportionality of willingness to pay to small changes in risk - The impact of attitudinal factors in scope tests. *Revised version forthcoming in Environmental and Resource Economics.*

2007-29 **Paul Raschky and Hannelore Weck-Hannemann:** Who is going to save us now? Bureaucrats, Politicians and Risky Tasks.

2007-28 **Harald Oberhofer and Michael Pfaffermayr:** FDI versus Exports. Substitutes or Complements? A Three Nation Model and Empirical Evidence.

2007-27 **Peter Wechselberger, Stefan Lang and Winfried J. Steiner:** Additive models with random scaling factors: applications to modeling price response functions.

2007-26 **Matthias Sutter:** Deception through telling the truth?! Experimental evidence from individuals and teams*. Revised version published in The Economic Journal, Vol. 119 (2009), 47-60.*

2007-25 **Andrea M. Leiter, Harald Oberhofer and Paul A. Raschky:** Productive disasters? Evidence from European firm level data. *Revised version forthcoming in Environmental and Resource Economics.*

2007-24 **Jesus Crespo Cuaresma:** Forecasting euro exchange rates: How much does model averaging help?

2007-23 **Matthias Sutter, Martin Kocher and Sabine Strauß:** Individuals and teams in UMTS-license auctions. *Revised version with new title "Individuals and teams in auctions" published in Oxford Economic Papers, Vol. 61 (2009): 380-394****).***

2007-22 **Jesus Crespo Cuaresma, Adusei Jumah and Sohbet Karbuz:** Modelling and Forecasting Oil Prices: The Role of Asymmetric Cycles. *Revised version accepted for publication in The Energy Journal.*

2007-21 **Uwe Dulleck and Rudolf Kerschbamer:** Experts vs. discounters: Consumer free riding and experts withholding advice in markets for credence goods. *Revised version published in International Journal of Industrial Organization*, *Vol. 27, Issue 1 (2009): 15-23.*

2007-20 **Christiane Schwieren and Matthias Sutter:** Trust in cooperation or ability? An experimental study on gender differences*. Revised version published in Economics Letters, Vol. 99 (2008): 494-497.*

2007-19 **Matthias Sutter and Christina Strassmair:** Communication, cooperation and collusion in team tournaments – An experimental study. *Revised version published in: Games and Economic Behavior, Vol.66 (2009), 506-525.*

2007-18 **Michael Hanke, Jürgen Huber, Michael Kirchler and Matthias Sutter:** The economic consequences of a Tobin-tax – An experimental analysis. *Revised version forthcoming in Journal of Economic Behavior and Organization.*

2007-17 **Michael Pfaffermayr:** Conditional beta- and sigma-convergence in space: A maximum likelihood approach. *Revised version forthcoming in Regional Science and Urban Economics.*

2007-16 **Anita Gantner:** Bargaining, search, and outside options. *Published in: Games and Economic Behavior, Vol. 62 (2008), pp. 417-435.*

2007-15 **Sergio Currarini and Francesco Feri:** Bilateral information sharing in oligopoly.

2007-14 **Francesco Feri:** Network formation with endogenous decay.

2007-13 **James B. Davies, Martin Kocher and Matthias Sutter:** Economics research in Canada: A long-run assessment of journal publications. *Revised version published in: Canadian Journal of Economics, Vol. 41 (2008), 22-45.*

2007-12 **Wolfgang Luhan, Martin Kocher and Matthias Sutter:** Group polarization in the team dictator game reconsidered. *Revised version published in: Experimental Economics, Vol. 12 (2009), 26-41.*

2007-11 **Onno Hoffmeister and Reimund Schwarze:** The winding road to industrial safety. Evidence on the effects of environmental liability on accident prevention in Germany.

2007-10 **Jesus Crespo Cuaresma and Tomas Slacik:** An “almost-too-late” warning mechanism for currency crises. *(Revised version accepted for publication in Economics of Transition***)**

2007-09 **Jesus Crespo Cuaresma, Neil Foster and Johann Scharler:** Barriers to technology adoption, international R&D spillovers and growth.

2007-08 **Andreas Brezger and Stefan Lang:** Simultaneous probability statements for Bayesian P-splines.

2007-07 **Georg Meran and Reimund Schwarze:** Can minimum prices assure the quality of professional services? (*Accepted for publication in European Journal of Law and Economics*)

2007-06 **Michal Brzoza-Brzezina and Jesus Crespo Cuaresma:** Mr. Wicksell and the global economy: What drives real interest rates?.

2007-05 **Paul Raschky:** Estimating the effects of risk transfer mechanisms against floods in Europe and U.S.A.: A dynamic panel approach.

2007-04 **Paul Raschky and Hannelore Weck-Hannemann:** Charity hazard - A real hazard to natural disaster insurance. *Revised version forthcoming in: Environmental Hazards.*

2007-03 **Paul Raschky:** The overprotective parent - Bureaucratic agencies and natural hazard management.

2007-02 **Martin Kocher, Todd Cherry, Stephan Kroll, Robert J. Netzer and Matthias Sutter:** Conditional cooperation on three continents. *Revised version published in: Economics Letters, Vol. 101 (2008): 175-178.*

2007-01 **Martin Kocher, Matthias Sutter and Florian Wakolbinger:** The impact of naïve advice and observational learning in beauty-contest games.

University of Innsbruck

Working Papers in Economics and Statistics

2010-11

Alexander Strasak, Nikolaus Umlauf, Ruth Pfeiffer and Stefan Lang

Comparing Penalized Splines and Fractional Polynomials for Flexible Modelling of the Effects of Continuous Predictor Variables

Abstract

P(enalized)-splines and fractional polynomials (FPs) have emerged as powerful smoothing techniques with increasing popularity in several fields of 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. We thus performed extensive simulations to compare FPs of degree 2 (FP2) and degree 4 (FP4) and P-splines that used generalized cross validation (GCV) and restricted maximum likelihood (REML) for smoothing parameter selection. We evaluated 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. We found that for more curved functions FP2, the current default implementation in standard software, showed considerably bias and consistently higher mean squared error (MSE) compared to spline-based estimators (REML, GCV) 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. Finally, we highlight the specific features of the approaches in a real dataset.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)