Open access • Journal Article • DOI:10.1007/S11222-010-9170-7

Bayesian fractional polynomials — Source link

Daniel Sabanés Bové, Leonhard Held

Institutions: University of Zurich

Published on: 01 Jul 2011 - Statistics and Computing (Springer US)

Topics: Linear model, Bayesian probability and Feature selection

Bayesian fractional polynomials

Sabanés Bové, D; Held, L

Bayesian fractional polynomials

Abstract

This paper sets out to implement the Bayesian paradigm for fractional polynomial models under the assumption of normally distributed error terms. Fractional polynomials widen the class of ordinary polynomials and offer an additive and transportable modelling approach. The methodology is based on a Bayesian linear model with a quasi-default hyper-g prior and combines variable selection with parametric modelling of additive effects. A Markov chain Monte Carlo algorithm for the exploration of the model space is presented. This theoretically well-founded stochastic search constitutes a substantial improvement to ad hoc stepwise procedures for the fitting of fractional polynomial models. The method is applied to a data set on the relationship between ozone levels and meteorological parameters, previously analysed in the literature.

other $k-1$ covariates are fixed. Formally, this is

$$\eta(\boldsymbol{x}) := \mathbb{E}(y \mid \boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i, \tag{1.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_k)^T$. Of course, such a formulation can lead to incorrect inference if the true relationship is far from linear for certain x_i. An immediate generalization that retains additive effects is to substitute $\beta_i x_i$ with $f_i(x_i)$ in (1.1), i. e.

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} f_i(x_i). \tag{1.2}$$

Nonparametric smoothers are very flexible methods for estimating the unknown functions f_i, see Ruppert, Wand, and Carroll (2003) for a recent review. They emerged in the last two decades and had their breakthrough with the definition of the generalized additive model (Hastie and Tibshirani 1990). However, the resulting models are difficult to summarize in closed form, as each function f_i is in itself a linear combination of complicated basis functions, e. g. B-spline basis functions. Moreover, the local behaviour of scatterplot smoothers can lead to artifacts in the resulting function and prohibits any extrapolation outside the observed data range. Finally, estimation of the associated smoothing parameters may become difficult, if k is large.

On the other hand, *ad hoc* approaches, such as equating f_i with a polynomial of low degree and comparing the model fit to that of a linear function, are common in applied statistics. Lying within the framework of traditional parametric models, these global models are easy to understand and communicate, but have severe disadvantages: their form is quite limited and resorting to higher degrees may lead to unplausible features, in particular near the minimum and maximum of x_i. Therefore, Box and Tidwell (1962) restricted themselves to polynomials of degree one or two before estimating the best powers among all real numbers iteratively. They introduced the transformation now known as the Box-Tidwell transformation,

$$x^{(a)} = \begin{cases} x^a & \text{if } a \neq 0, \\ \log(x) & \text{if } a = 0, \end{cases} \tag{1.3}$$

where a is a real number. Few other attempts to develop methodology for systematic parametric covariate transformation had been made until Royston and Altman (1994)

extended the classical polynomials to a class which they called fractional polynomials (FPs). This contribution is one of the most cited papers in *Applied Statistics* with more than 400 citations at the time of writing, which illustrates that this method has been well-received by applied researchers. Royston and Altman (1997) show that FPs "are particularly good at providing concise and accurate formulae" for representing smooth relationships between y and the x_i. From a simulation study on the Cox model, Govindarajulu, Malloy, Ganguli, Spiegelman, and Eisen (2009) conclude that FPs are among the least biased smoothing methods for fitting non-linear exposure effects. So although Ambler and Royston (2001) acknowledge that finding very complex non-linear relationships may require more complex non-parametric regression methods, the FP approach has clearly established a prominent role in the non-linear parametric methodology.

An FP of degree m with powers $p_1 \leq \cdots \leq p_m$ and respective coefficients $\alpha_1, \ldots, \alpha_m$ is

$$\begin{aligned} f^m(x; \boldsymbol{\alpha}, \boldsymbol{p}) &= \sum_{j=1}^{m} \alpha_j h_j(x), \quad \text{where} \\ h_0(x) &= 1 \quad \text{and} \\ h_j(x) &= \begin{cases} x^{(p_j)} & \text{if } p_j \neq p_{j-1}, \\ h_{j-1}(x)\log(x) & \text{if } p_j = p_{j-1} \end{cases} \quad \text{for} \quad j = 1, \ldots, m. \end{aligned} \tag{1.4}$$

Note that the definition of $h_j(x)$ allows repeated powers. The brackets around the exponent denote the Box-Tidwell transformation (1.3). For $m \leq 3$, Royston and Altman (1994) constrained the set of possible powers p_j to the set

$$\mathcal{S} = \left\{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3\right\}, \tag{1.5}$$

which encompasses the classic polynomial powers $1, 2, 3$ but also offers square roots and reciprocals. Royston and Sauerbrei (2008, section 4.6) argue that this set is sufficient to approximate all powers in the interval $[-2, 3]$. However, sometimes there are reasons to extend this set, see e. g. Shkedy, Aerts, Molenberghs, Beutels, and van Damme (2006). A problematic aspect of the logarithm inclusion is that $x > 0$ is required, which may require a prior transformation of the original variable z. Often used is a shift $x = z + \xi$ with a natural point of origin ξ. Royston and Sauerbrei (2008, section 5.4) discuss

sensitivity of the results depending on the choice of origin. Data-driven estimation of ξ is also possible, but generally not recommended (Royston and Altman 1994).

For example, an FP with $m = 3$ powers in its power vector $\boldsymbol{p} = (p_1, p_2, p_3) = (-\frac{1}{2}, 2, 2)$ would be

$$f^3(x; \boldsymbol{\alpha}, \boldsymbol{p}) = \alpha_1 x^{-\frac{1}{2}} + \alpha_2 x^2 + \alpha_3 x^2 \log(x),$$

where the last term reflects the repeated power 2. Note that, given the degree and powers, the function is linear in the unknown coefficients. Indeed, when using FPs as model functions f_i in (1.2), this gives the same form as in (1.1):

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} f_i^{m_i}(x_i; \boldsymbol{\alpha}_i, \boldsymbol{p}_i) = \beta_0 + \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij} h_{ij}(x_i). \tag{1.6}$$

So besides having more summands the linear predictor $\eta(\boldsymbol{x})$ is unchanged and established estimation procedures apply. We call a model with structual assumption (1.6) a multiple FP model.

It is worthwhile to gauge the complexity of the model space that has just been described. Suppose we continue examining k continuous covariates $x_1, \ldots, x_k$ and content ourselves with a maximum degree of $m_{max} \leq 3$ for each $f_i^{m_i}$, i. e. $0 \leq m_i \leq m_{max}$ for $i = 1, \ldots, k$, where $m_i = 0$ denotes the omission of x_i from the model. From the power set $\mathcal{S}$, m powers are chosen, which need not be different because of the inclusion of logarithmic terms for repeated powers, cf. (1.4). Therefore, for only a single covariate x, the number of possible fractional polynomials with degree $m = 0, 1, 2, 3$ is $d(m) = 1, 8, 36$ and 120, respectively. The model space complexity grows exponentially as a function of the number k of covariates. For example, already for a moderate degree $m_{max} = 2$ and $k = 5$ covariates $(1 + 8 + 36)^5 = 184\,528\,125$ different models exist, which illustrates that the search for the best model is expensive.

Royston and Altman (1994) conduct inference about the best degrees $\{m_i\}$ and powers $\{\boldsymbol{p}_i\}$ (where $p_{ij} \in \mathcal{S}, j = 1, \ldots, m_i$) for the corresponding $f_i(x_i) = f_i^{m_i}(x_i; \boldsymbol{\alpha}_i, \boldsymbol{p}_i)$ in (1.2) by implementing maximum likelihood in an iterative backfitting-like routine. Of course, this algorithm may miss the best model in the restricted range of degrees as not every combination of fractional polynomials is given a chance. This type of stepwise backward elimination was slightly modified by Sauerbrei and Royston (1999), in order to reduce the increase in the type I error rate inherent to the multiple testing setting, cf. Ambler and Royston (2001).

In this paper we implement the Bayesian paradigm for fitting and selecting a multiple FP model under the assumption of normally distributed error terms. We use a hyper-g prior for the regression coefficients as recently proposed in Liang, Paulo, Molina, Clyde, and Berger (2008). Section 2 defines the models to be considered, which can be viewed as a collection of special Bayesian linear models. An algorithm for posterior sampling from the model space is presented and model selection and averaging are discussed in Section 3. The approach is applied to real data in Section 4. Section 5 discusses the paper findings and possible extensions.

2 Model definition

2.1 The multiple fractional polynomial model as a linear model

Consider the linear model with intercept,

$$\boldsymbol{y} = \beta_0 + \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where the $(n \times p)$-design matrix $\boldsymbol{B} = \big(B_i(\boldsymbol{x}_j)\big)_{ji}$ with row indices $j = 1, \ldots, n$ and column indices $i = 1, \ldots, p$ is a function of explanatory variables $\boldsymbol{x}_j$ of the jth observation ($j = 1, \ldots, n$). The responses $\boldsymbol{y}$, the errors $\boldsymbol{\varepsilon}$ and the regression coefficients $\boldsymbol{\beta}$ are appropriate column vectors of length n, n and p, respectively. The assumption of independent homoscedastic normally distributed error terms ε_j results in $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{I}_n$ denotes the identity matrix of dimension n. Hence, $\boldsymbol{y}$ also follows a multivariate normal distribution with the same covariance matrix and mean vector $\boldsymbol{\mu} = \boldsymbol{1}_n\beta_0 + \boldsymbol{B}\boldsymbol{\beta}$, which determines the likelihood $f(\mathcal{D} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2)$, where $\mathcal{D} = \{y_j, \boldsymbol{x}_j\}_{j=1}^n$ denotes the observed data.

A special way of defining the design matrix $\boldsymbol{B}$ is through the use of FPs. In this case, the basis functions B_i are chosen as the transformations h_{ij} in (1.6), and with the appropriate parameter vector $\boldsymbol{\beta} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)^T$, where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{im_i})$, the FP approach has been embedded into the linear model framework. The transformations h_{ij} are determined by the power vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k$ through their definition (1.4), so that each multiple FP model can be represented by a vector $\boldsymbol{\theta}$ of ordered tuples:

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k) \quad \text{with} \\ \boldsymbol{p}_i &= (p_{i1} \leq p_{12} \leq \cdots \leq p_{im_i}). \end{aligned}$$

The model parameter space $\boldsymbol{\Theta}$ contains all such $\boldsymbol{\theta}$ which fulfill the restriction of the power set $\mathcal{S}$ given in (1.5). Note that $\boldsymbol{\theta}$ is of varying dimension $p_{\boldsymbol{\theta}} := \sum_{i=1}^{k} m_i$. For the null model, $p_{\boldsymbol{\theta}} = 0$, because the ith tuple $\boldsymbol{p}_i$ is empty if the covariate x_i is not included in the model ($m_i = 0$). Quantities that depend on the model are henceforth subscripted with $\boldsymbol{\theta}$. The columns of the covariates' design matrix $\boldsymbol{B}_{\boldsymbol{\theta}}$ are centered such that

$$\mathbf{1}_n^T \boldsymbol{B}_{\boldsymbol{\theta}} = \mathbf{0}_{p_{\boldsymbol{\theta}}}^T$$

to ensure that the intercept β_0 is a common parameter with identical interpretation in all models.

Note that we could reparametrize the inclusion of x_i with a binary variable inclusion indicator $\gamma_i = \mathbb{I}(m_i > 0)$. However, the reparametrization of a non-empty power vector $\boldsymbol{p}_i$ by an additional lower-level set of binary indicators would not be straightforward nor natural, because the recursive FP definition (1.4) would need to be obscured. By contrast, our parametrization retains the FP form, and of course also allows probability statements about variable inclusion, cf. section 3.2.

2.2 Prior specification

We use the hyper-g prior of Liang et al. (2008), which is constructed as follows. Jeffreys' prior is used for the regression variance σ^2. Conditional on $\sigma^2, g > 0$, an improper flat prior on the intercept β_0 and a mean-zero normal prior with covariance matrix $\sigma^2 g \cdot (\boldsymbol{B}_{\boldsymbol{\theta}}^T \boldsymbol{B}_{\boldsymbol{\theta}})^{-1}$ on the remaining model-specific coefficients in $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ are used:

$$\begin{aligned} f(\sigma^2) &\propto (\sigma^2)^{-1}, \\ f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\theta}} \,|\, \sigma^2, g) &\propto (\sigma^2 g)^{-\frac{p_{\boldsymbol{\theta}}}{2}} \left| \boldsymbol{B}_{\boldsymbol{\theta}}^T \boldsymbol{B}_{\boldsymbol{\theta}} \right|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2 g} \left\| \boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}} \right\|^2 \right\}. \end{aligned}$$

An advantage of this so-called g-prior (Zellner 1986) is that it accounts for multicollinearity, because *a priori* coefficients of almost collinear columns are highly correlated and have a large variance, which reflects that they should have the same magnitude and are hard to estimate. The covariance factor $g > 0$ is assumed to be independent of σ^2 with prior density

$$f(g) = \frac{a-2}{2}(1+g)^{-\frac{a}{2}},$$

where $a \in (3, 4]$ ensures that the posterior mean $\mathbb{E}(g \,|\, \boldsymbol{\theta}, \mathcal{D})$ is finite in any given model $\boldsymbol{\theta}$. Moreover, the implied prior on the factor $t = g/(g+1)$, which shrinks the mean vector

$\boldsymbol{\mu}$ towards the intercept β_0, does not favor small values of t (heavy shrinkage) more than the uniform distribution obtained from $a = 4$, see Liang et al. (2008) on p. 415.

This prior has several desirable asymptotic properties (Liang et al. 2008, section 4). For $n \geq p_{\boldsymbol{\theta}} + 3$, the information paradox of the fixed-g prior is resolved: the Bayes factor of a model with $R^2_{\boldsymbol{\theta}} \to 1$ versus the null model can grow in parallel without restraint, where $R^2_{\boldsymbol{\theta}}$ is the coefficient of determination for the OLS estimate with components $\hat{\beta}_0^{OLS} = \bar{y}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{OLS} = (\boldsymbol{B}_{\boldsymbol{\theta}}^T \boldsymbol{B}_{\boldsymbol{\theta}})^{-1} \boldsymbol{B}_{\boldsymbol{\theta}}^T \boldsymbol{y}$. Moreover, whenever the true model is not the null model, the *maximum a posteriori* (MAP) model is consistent for the true model when $n \to \infty$. The hyper-g prior also produces Bayesian model average (BMA) estimates which are consistent under prediction of new responses. Thus, although it might be a strong assumption that the prior variance of the regression parameters depends on the error variance σ^2, the utilized prior remedies the deficiencies of the ordinary conjugate normal-gamma and g-priors while still being computationally tractable.

Turning to the prior on the models, prior independence of the FP transformations can be specified by assuming $f(\boldsymbol{\theta}) = f(\boldsymbol{p}_1) \times \cdots \times f(\boldsymbol{p}_k)$. For a single covariate x_i one noninformative prior is based on the idea that each degree $0 \leq m_i \leq m_{max}$ has the same prior probability, and that, given the degree m_i, each combination of powers $p_{i1} \in \mathcal{S}, \ldots, p_{im_i} \in \mathcal{S}$ is equally probable *a priori*. The number of degrees is $m_{max} + 1$ and the number of different FPs for degree m_i was denoted as $d(m_i)$. Thus, this model prior can be formulated as

$$f(\boldsymbol{p}_i) = f(p_{i1}, \ldots, p_{im_i} \,|\, m_i) f(m_i) = d(m_i)^{-1} (m_{max} + 1)^{-1} \tag{2.2}$$

In this case, the null model has the highest prior probability $(m_{max} + 1)^{-k}$. This prior directly penalizes non-parsimonious models, which helps to concentrate the posterior model probability in a small part of the model space and thus eases the model inference in section 3.

If non-identifiable models exist in the original description of the model space, the definition of the prior of a specific power vector $\boldsymbol{p}_i$ in (2.2) is to be understood as a definition up to a multiplicative constant, the kth power of which normalizes the model prior $f(\boldsymbol{\theta})$ to a valid prior distribution. This is necessary, as we intend to assign such models a zero prior probability.

2.3 Posterior distribution of parameters

The posterior density of the parameters $\beta_0, \boldsymbol{\beta}_{\boldsymbol{\theta}}, \sigma^2$ for a specific model $\boldsymbol{\theta}$ and covariance factor g is

$$f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\theta}}, \sigma^2 \,|\, \mathcal{D}, g) \propto (\sigma^2)^{-(\frac{n+p_{\boldsymbol{\theta}}}{2}+1)} g^{-\frac{p_{\boldsymbol{\theta}}}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\|\boldsymbol{y}-\boldsymbol{\mu}\|^2 + \frac{1}{g}\|\boldsymbol{B}_{\boldsymbol{\theta}}\boldsymbol{\beta}_{\boldsymbol{\theta}}\|^2\right]\right\}.$$

This kernel can be shown to belong to the normal inverse-gamma distribution (Denison, Holmes, Mallick, and Smith 2002, p. 16)

$$\beta_0, \boldsymbol{\beta}_{\boldsymbol{\theta}}, \sigma^2 \,|\, \mathcal{D}, g \sim \mathrm{N}_{p_{\boldsymbol{\theta}}+1}\,\mathrm{IG}\left(\boldsymbol{m}_{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}}, \frac{n-1}{2}, c_{\boldsymbol{\theta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} n^{-1} & \mathbf{0}_{p_{\boldsymbol{\theta}}}^T \\ \mathbf{0}_{p_{\boldsymbol{\theta}}} & \frac{g}{g+1}(\boldsymbol{B}_{\boldsymbol{\theta}}^T\boldsymbol{B}_{\boldsymbol{\theta}})^{-1} \end{pmatrix}, \tag{2.3}$$

$$\boldsymbol{Z}_{\boldsymbol{\theta}} = (\mathbf{1}_n, \boldsymbol{B}_{\boldsymbol{\theta}}),$$

$$\boldsymbol{m}_{\boldsymbol{\theta}} = \boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{Z}_{\boldsymbol{\theta}}^T\boldsymbol{y} = \begin{pmatrix} \bar{y} \\ \frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{OLS} \end{pmatrix}, \tag{2.4}$$

$$c_{\boldsymbol{\theta}} = \boldsymbol{y}^T\left[\boldsymbol{I}_n - \boldsymbol{Z}_{\boldsymbol{\theta}}\boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{Z}_{\boldsymbol{\theta}}^T\right]\boldsymbol{y}/2. \tag{2.5}$$

The marginal posterior density of the shrinkage factor $t = g/(g+1)$ is

$$f(t \,|\, \mathcal{D}) \propto (1-t)^{(p_{\boldsymbol{\theta}}+a-2)/2-1}(1-R_{\boldsymbol{\theta}}^2 t)^{-(n-1)/2}.$$

Liang et al. (2008) have derived a closed form expression for the posterior mean of t. However, we want to incorporate the posterior uncertainty with respect to t in our analysis. To achieve this, we need to be able to sample from $f(t \,|\, \mathcal{D})$. This can be done by inversion, since the unnormalized cumulative distribution function (cdf) can be

obtained by a change of variable to $u = (1 - R^2_{\boldsymbol{\theta}})/(1 - R^2_{\boldsymbol{\theta}} t)$:

$$\begin{aligned}
\tilde{F}_{\boldsymbol{\theta}}(q) &\propto \int_0^q f(t \,|\, \mathcal{D})\, dt \\
&\propto \int_{1-R^2_{\boldsymbol{\theta}}}^{\frac{1-R^2_{\boldsymbol{\theta}}}{1-R^2_{\boldsymbol{\theta}} q}} u^{[(n-1)/2-(p_{\boldsymbol{\theta}}+a-2)/2]-1}(1-u)^{(p_{\boldsymbol{\theta}}+a-2)/2-1}\, du \\
&\propto B_{\boldsymbol{\theta}}\left(\frac{1-R^2_{\boldsymbol{\theta}}}{1-R^2_{\boldsymbol{\theta}} q}\right) - B_{\boldsymbol{\theta}}(1-R^2_{\boldsymbol{\theta}}),
\end{aligned}$$

where $B_{\boldsymbol{\theta}}$ is the cdf of the Beta distribution with shape parameters $(n - p_{\boldsymbol{\theta}} - a + 1)/2$ and $(p_{\boldsymbol{\theta}} + a - 2)/2$. The normalization constant of the shrinkage factor cdf is

$$\tilde{F}_{\boldsymbol{\theta}}(1) = 1 - B_{\boldsymbol{\theta}}(1 - R^2_{\boldsymbol{\theta}}),$$

yielding the posterior cdf $F_{\boldsymbol{\theta}}(q) = \tilde{F}_{\boldsymbol{\theta}}(q)/\tilde{F}_{\boldsymbol{\theta}}(1)$. The inverse cdf can be derived from that as

$$F^{-1}_{\boldsymbol{\theta}}(p) = \left(1 - \frac{1-R^2_{\boldsymbol{\theta}}}{B^{-1}_{\boldsymbol{\theta}}\left(p + (1-p)B_{\boldsymbol{\theta}}(1-R^2_{\boldsymbol{\theta}})\right)}\right) \Bigg/ R^2_{\boldsymbol{\theta}}. \tag{2.6}$$

This allows effective inverse sampling from the model-specific posterior distribution of the shrinkage factor t, and hence the covariance factor $g = t/(1-t)$.

3 Model inference

Inference on the space $\boldsymbol{\Theta}$ of all possible models $\boldsymbol{\theta}$ grounds on the posterior model probabilities

$$f(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{f(\mathcal{D} \,|\, \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathcal{D})}. \tag{3.1}$$

The hyper-g prior is convenient because it allows a closed form for the marginal likelihood $f(\mathcal{D} \mid \boldsymbol{\theta})$ of a model $\boldsymbol{\theta}$. From Liang et al. (2008) we have

$$\begin{aligned} f(\mathcal{D} \mid \boldsymbol{\theta}) &= \frac{f(\mathcal{D} \mid g, \boldsymbol{\theta}) f(g)}{f(g \mid \mathcal{D}, \boldsymbol{\theta})} \\ &= \frac{\Gamma(\frac{n-1}{2}) \left\| \boldsymbol{y} - \mathbf{1}_n \bar{y} \right\|^{-(n-1)} (a-2) {}_2F_1\left(\frac{n-1}{2}; 1; \frac{p_{\boldsymbol{\theta}}+a}{2}; R^2_{\boldsymbol{\theta}}\right)}{\sqrt{\pi}^{(n-1)} \sqrt{n} (p_{\boldsymbol{\theta}} + a - 2)} \\ &\propto \frac{{}_2F_1\left(\frac{n-1}{2}; 1; \frac{p_{\boldsymbol{\theta}}+a}{2}; R^2_{\boldsymbol{\theta}}\right)}{p_{\boldsymbol{\theta}} + a - 2}, \end{aligned} \tag{3.2}$$

where factors which are not model-specific have been omitted in the last step and the Gaussian hypergeometric function has the integral representation

$${}_2F_1(a; b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a} \, dt,$$

cf. Abramowitz and Stegun (1964, section 15.3). See Appendix A for the numerical calculation of the marginal likelihood and related quantities.

Posterior inference is conducted in two steps. First, posterior model probabilities are estimated. This requires sampling from the model space (Section 3.1), when an exhaustive computation of all marginal likelihoods is infeasible. Second, the posterior distribution of FP curves in the most probable model or in a model average is estimated by Monte Carlo (Section 3.2).

3.1 Posterior model sampling

As shown in Section 1, the model space may get very large due to its exponential growth in the number of covariates k. This often renders an exhaustive computation of all posterior model probabilities $f(\boldsymbol{\theta} \mid \mathcal{D})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ via (3.2), (2.2) and (3.1) infeasible. Instead of utilizing *ad hoc* search strategies such as stepwise procedures, we are going to sample from the posterior distribution $f(\boldsymbol{\theta} \mid \mathcal{D})$ via a Markov Chain Monte Carlo (MCMC) sampler, which is an adaption of the Metropolis-Hastings sampler by Denison et al. (2002, pp. 53 ff. and p. 97). The approach is similar to the MCMC model composition by Madigan and York (1995).

The proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$ is formed by four different move types, which define how to jump from the current model $\boldsymbol{\theta}$ to the new model $\boldsymbol{\theta}'$:

BIRTH Randomly select one of the covariates with FP degree $m_i < m_{max}$. Add a power to its $\boldsymbol{p}_i$ after randomly drawing it from $\mathcal{S}$.

DEATH Randomly select one of the covariates with FP degree $m_i > 0$. Remove a randomly chosen power from its $\boldsymbol{p}_i$.

MOVE Randomly select one of the covariates with FP degree $m_i > 0$. Remove a randomly chosen power from its $\boldsymbol{p}_i$, then randomly draw a power from $\mathcal{S}$ and add it to $\boldsymbol{p}_i$.

SWITCH Randomly select one of the covariates with non-empty power vector $\boldsymbol{p}_i$. Randomly select one of the other covariates with power vector $\boldsymbol{p}_j$. Switch the power vectors $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$.

Note that the *SWITCH* move is only sensible for $k > 1$ covariates, but for $k = 1$ all models could easily be evaluated without any model sampling. The *SWITCH* move is designed to be able to efficiently trace models with high posterior probability even in situations where covariates are almost collinear. Each proposal begins with the probabilistic choice of one of the move types, with the four probabilities $b_{p_{\boldsymbol{\theta}}}$, $d_{p_{\boldsymbol{\theta}}}$, $m_{p_{\boldsymbol{\theta}}}$ and $s_{p_{\boldsymbol{\theta}}}$ depending on the current dimension $p_{\boldsymbol{\theta}}$ of the whole parameter vector $\boldsymbol{\theta}$:

$$\begin{aligned}
&b_{p_{\boldsymbol{\theta}}} = 1, \quad d_{p_{\boldsymbol{\theta}}} = m_{p_{\boldsymbol{\theta}}} = s_{p_{\boldsymbol{\theta}}} = 0 && \text{if } p_{\boldsymbol{\theta}} = 0,\\
&b_{p_{\boldsymbol{\theta}}} = d_{p_{\boldsymbol{\theta}}} = m_{p_{\boldsymbol{\theta}}} = s_{p_{\boldsymbol{\theta}}} = \frac{1}{4} && \text{if } 0 < p_{\boldsymbol{\theta}} < p_{max},\\
&b_{p_{\boldsymbol{\theta}}} = 0, \quad d_{p_{\boldsymbol{\theta}}} = m_{p_{\boldsymbol{\theta}}} = s_{p_{\boldsymbol{\theta}}} = \frac{1}{3} && \text{if } p_{\boldsymbol{\theta}} = p_{max},
\end{aligned}$$

where the value $p_{max} := \min\{n - 3 - a, k \times m_{max}\}$ takes into account that more than $n - 3 - a$ powers would render the posterior distributions in the model improper (Liang et al. 2008, p. 420).

The proposed new model $\boldsymbol{\theta}'$ is accepted with probability

$$\alpha(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}) = \min\left\{1, \frac{f(\mathcal{D} \,|\, \boldsymbol{\theta}')}{f(\mathcal{D} \,|\, \boldsymbol{\theta})} \frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}')}{q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})}\right\},$$

which is the usual Metropolis-Hastings acceptance probability; in case of rejection the chain stays at the previous model $\boldsymbol{\theta}$. The only parts of $\alpha(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})$ which still need to be computed are the prior odds $f(\boldsymbol{\theta}')/f(\boldsymbol{\theta})$ and the proposal ratio $q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}')/q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})$, because

the Bayes factor $f(\mathcal{D} \,|\, \boldsymbol{\theta}')/f(\mathcal{D} \,|\, \boldsymbol{\theta})$ is known from (3.2). Both prior odds and proposal ratio depend on the proposed move type.

For example, suppose a *BIRTH* proposed to add a power p to the ith FP which formerly had the degree m_i. Using the prior independence of the power vectors and (2.2), the prior odds amount to

$$\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} = \frac{f(\boldsymbol{p}_i')}{f(\boldsymbol{p}_i)} = \frac{d(m_i+1)^{-1}(m_{max}+1)^{-1}}{d(m_i)^{-1}(m_{max}+1)^{-1}} = \frac{d(m_i)}{d(m_i+1)} = \frac{m_i+1}{|\mathcal{S}|+m_i}.$$

The proposal probability $q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})$ of this specific *BIRTH* move is

$$q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}) = b_{p_{\boldsymbol{\theta}}} \times \frac{1}{|\mathcal{F}|} \times \frac{1}{|\mathcal{S}|},$$

where $\mathcal{F} = \{j : |\boldsymbol{p}_j| < m_{max}\}$ collects the indices of the covariates in model $\boldsymbol{\theta}$ that could receive an additional power. The reverse probability of reaching the old model $\boldsymbol{\theta}$ from the proposed model $\boldsymbol{\theta}'$ by a converse *DEATH* move is

$$q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}') = d_{p_{\boldsymbol{\theta}}+1} \times \frac{1}{|\mathcal{P}'|} \times \frac{\mathbf{1}_{\boldsymbol{p}_i'}(p)}{m_i+1},$$

where $\mathcal{P}' = \{j : |\boldsymbol{p}_j'| > 0\}$ abbreviates the index set of present covariates in the proposed model $\boldsymbol{\theta}'$. The multiplicity of the newly chosen power p in $\boldsymbol{p}_i'$ is denoted by $\mathbf{1}_{\boldsymbol{p}_i'}(p)$. Altogether we obtain

$$\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}')}{q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})} = \frac{d_{p_{\boldsymbol{\theta}}+1}}{b_{p_{\boldsymbol{\theta}}}} \frac{|\mathcal{F}|}{|\mathcal{P}'|} \frac{\mathbf{1}_{\boldsymbol{p}_i'}(p) \cdot |\mathcal{S}|}{|\mathcal{S}|+m_i}.$$

The prior odds and proposal ratios for the *DEATH*, *MOVE* and *SWITCH* proposals are computed analogously, see Appendix B. The sampling algorithm can be modified without much effort to enable the selection of categorical covariates using "fixed form covariates groups": for each non-reference category of a categorical covariate, a binary design variable is included in the corresponding covariate group, which is then included as a whole in each FP model or not. By contrast to the continuous FP terms, the form of the design variables is naturally fixed here. While we already have implemented this extension of particular practical relevance, we omit the details here because the selection of fixed form covariates groups is not an original feature of the FP approach.

We have been able to analytically marginalize the likelihood over the parameters β_0, $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ and σ^2 and have arrived at the compact formula (3.2) for the marginal likelihood. So

when the algorithm jumps to a new model $\boldsymbol{\theta}'$, we can immediately compute the posterior model probability (3.1) up to the unknown multiplicative constant $f(\mathcal{D})^{-1}$. Let the models that have been visited by the algorithm be collected in $\hat{\boldsymbol{\Theta}}$. The normalizing constant can be approximated by the sum over its elements,

$$f(\mathcal{D}) \approx \sum_{\boldsymbol{\theta} \in \hat{\boldsymbol{\Theta}}} f(\mathcal{D} \,|\, \boldsymbol{\theta}) f(\boldsymbol{\theta}), \tag{3.3}$$

and the values $f(\mathcal{D} \,|\, \boldsymbol{\theta}) f(\boldsymbol{\theta})$ of the visited models $\boldsymbol{\theta} \in \hat{\boldsymbol{\Theta}}$ can be normalized with this sum, to obtain estimates $\hat{f}(\boldsymbol{\theta} \,|\, \mathcal{D})$. Of course, these estimates will be too high, because the sum for the normalization constant is not taken over the whole model space $\boldsymbol{\Theta}$. In a similar context George and McCulloch (1997) propose a more elaborated estimator, which requires a preliminary run of the MCMC sampler. Here, the visited part $\hat{\boldsymbol{\Theta}}$ is effectively interpreted as an estimate of the whole model space $\boldsymbol{\Theta}$.

The sampling algorithm has strong connections to the simulated annealing approach, which has also been utilized for frequentist model selection procedures, e. g. by Brooks, Friel, and King (2003), as we need not base inference on the model frequencies in the Markov chain. However, the MCMC construction ensures that for sufficiently long chains the best models will be visited finally, as the chain converges to the true posterior distribution $f(\boldsymbol{\theta} \,|\, \mathcal{D})$. From this perspective, the sampling algorithm appears as a seemingly simple search algorithm for the best models. The search is local in a sense, because in the algorithm the current model is slightly modified to propose a model from the current model's neighborhood, and if the proposed model's posterior probability is higher, then it is essentially accepted (modulo the proposal ratio). If the proposed model's posterior probability is lower, then the algorithm might still accept the new model, so that our approach is superior to stepwise or backfitting approaches, which get easily stuck in local maxima.

3.2 Model selection and averaging

An intuitive approach is the selection of the model $\boldsymbol{\theta}_{MAP}$ with the highest posterior probability, which can be estimated by the algorithm described in Section 3.1. The alternative is to take into account the uncertainty in model selection by marginalising over the set of possible models. The resulting hypermodel is a BMA with weights given by the posterior model probabilities. In general it will not be part of the original model

space $\boldsymbol{\Theta}$, but in our application the BMA mean curve is again an FP, typically with a higher degree than m_{max}.

The hypermodel can be estimated by drawing samples from the posterior in three hierarchical steps:

1. Draw a model from the estimated posterior model distribution $\hat{f}(\boldsymbol{\theta} \,|\, \mathcal{D})$.

2. Sample a shrinkage factor $t = g/(1+g)$ from $f(t \,|\, \mathcal{D}, \boldsymbol{\theta})$, using the quantile function (2.6) for inverse sampling.

3. Sample the intercept β_0 and the coefficient vector $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ from the Student distributions (see Denison et al. (2002, p. 238) for the parametrization used)

$$\beta_0 \,|\, \mathcal{D}, \boldsymbol{\theta}, g \sim \mathrm{t}\left(\bar{y}, \frac{2c_{\boldsymbol{\theta}}}{n(n-1)}, n-1\right)$$
$$\text{and} \quad \boldsymbol{\beta}_{\boldsymbol{\theta}} \,|\, \mathcal{D}, \boldsymbol{\theta}, g \sim \mathrm{t}_{p_{\boldsymbol{\theta}}}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{OLS}, \frac{2c_{\boldsymbol{\theta}} g}{(n-1)(g+1)}(\boldsymbol{B}_{\boldsymbol{\theta}}^T \boldsymbol{B}_{\boldsymbol{\theta}})^{-1}, n-1\right).$$

Samples from linear combinations of $\boldsymbol{\beta}_{\boldsymbol{\theta}}$, especially FP curve points $f_i(x_i)$, are easily obtained during the last step (see Appendix C for details on the computation of posterior summaries). Samples from the regression variance can be drawn from the inverse-gamma distribution $\sigma^2 \,|\, \mathcal{D}, \boldsymbol{\theta}, g \sim \mathrm{IG}((n-1)/2, c_{\boldsymbol{\theta}})$, if needed. Note that the above simulation is necessary, because we have marginalized analytically over the model parameters β_0, $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ and σ^2 before exploring the model space.

As the model sampling algorithm will typically visit hundreds of thousands of FP models, it is impractical to include all of them in step 1 above. Thus we will adapt the "Occam's Window" strategy of Raftery, Madigan, and Hoeting (1997) and only save a fixed number of best models for the BMA, collected in the set $\hat{\boldsymbol{\Theta}}^{loc} \subset \hat{\boldsymbol{\Theta}}$. The whole $\hat{\boldsymbol{\Theta}}$ will only be used to calculate variable inclusion probabilities $\hat{\pi}_i$ in addition to the "local" counterparts $\hat{\pi}_i^{loc}$, via

$$\hat{\pi}_i^{(loc)} := \sum_{\boldsymbol{\theta} \in \hat{\boldsymbol{\Theta}}^{(loc)} : \boldsymbol{p}_i \neq \emptyset} \hat{f}(\boldsymbol{\theta} \,|\, \mathcal{D}), \quad i = 1, \ldots, k. \tag{3.4}$$

If posterior inference given a single (best) model is desired, one simply omits step 1 of the above algorithm and always uses the same $\boldsymbol{\theta}$. Similarly, one can define other subsets of $\hat{\boldsymbol{\Theta}}^{(loc)}$ and average over their elements. For example, Barbieri and Berger (2004) propose

the median probability model that selects all variables with $\hat{\pi}_i \geq 1/2$. As we also consider transformations of the continuous covariates in addition to their selection, our median probability model could be a BMA of those models which do not contain powers for those covariates with $\hat{\pi}_i \leq 1/2$. We may also rerun the model sampling algorithm on the model subset or choose a lower threshold for the inclusion probabilities, following the approach of Fouskakis, Ntzoufras, and Draper (2009).

4 Application

We will apply the Bayesian FP approach to the ozone data that was first analyzed by Breiman and Friedman (1985) (with the alternating conditional expectations (ACE) algorithm). Nine variables with the same maximum FP degree $m_{max} = 2$ had been considered in the model selection procedure. They had been preliminarily transformed to ensure positivity and to avoid numerical issues with large numbers. To assess the predictive performance of the Bayesian FP models, we randomly select 30 observations that shall form a test set. The training set which is used to fit the models comprises the remaining 300 records. More details on the data set can be found in Appendix D. The hyperparameter is set as $a = 4$.

In order to explore the vast model space of cardinality $756 \cdot 10^{12}$, we have run the search algorithm for 1 000 000 iterations. This task required only 11 minutes (on an Intel T2500 with 2 GHz running Ubuntu 9.10), because we have used a fast `C++` implementation of the model search algorithm. The `R`-package with a comfortable `R`-interface and corresponding binaries for Windows and Mac operating systems are available from `R`-Forge (`http://r-forge.r-project.org/projects/bfp`).

Two computational problems had to be solved before it was possible to implement the sampler successively. First, most unnormalized posterior probabilities had been smaller than 10^{-308} and it had been impossible to display them in double precision. Fortunately, modern `C++` compilers offer an extended precision floating-point data type (long double) and compatible exponential and log functions, which solved this problem in a straightforward manner. Second, a naive implementation of the summation (3.3) of these values had turned out to be insufficient, because large cancellations between summands of different magnitudes had occurred. A sophisticated 'distillation algorithm' for floating-

point summation (Anderson 1999) had already been implemented by Kenneth Wilder[i]. Though it consumes more memory and computing time, it has delivered sensible results, which appear to be correct. It is important to mention that no probability estimates are necessary for mere model ranking, as the (log) unnormalized posterior probabilities can be used for an equivalent comparison.

Note that we also ran the sampler with three other hyperparameter choices $a \in \{3.1, 3.4, 3.7\}$ for this data set, which barely changed the results. Furthermore, three additional runs of the algorithm using $a = 4$ with different random number generator seeds yielded very similar results. While emphasizing that the method is not sensitive to the hyperparameter value, this also suggests that the chain length is sufficient to explore the set of models with high posterior probability.

The transformation parameters and posterior inclusion probabilities are shown in Table 1. Only z_0 and $z_6, \ldots, z_{10}$ have probabilities greater than 0.7, as z_4 is borderline significant with the local inclusion probability dropping below 0.5. These results roughly correspond with those of Breiman and Friedman (1985), whose ACE algorithm selected z_0 and $z_7, \ldots, z_{10}$. One reason for this good correspondence may be that only a very mild transformation of y is proposed by the ACE procedure, so the considered dependent variable is almost the same. The local inclusion probabilities that are based on the saved 3 000 models with the highest posterior probabilities are quite similar to the global inclusion probabilities. This indicates that at least in this respect $\hat{\boldsymbol{\Theta}}^{loc}$ constitutes a sensibly reduced model set. The `mfp` algorithm (Sauerbrei, Meier-Hirmer, Benner, and Royston 2006) yields the model

$$\eta(\boldsymbol{x}) = x_0 + x_0 \log x_0 + x_4 + x_5 + x_6 + x_7^{\frac{1}{2}} + x_7^2 + x_8 + x_9^3 + x_{10}^{-\frac{1}{2}} + x_{10}^{-\frac{1}{2}} \log x_{10}, \quad (4.1)$$

which includes all covariates at least linearly, and is not among the saved best 3 000 models with posterior probability $2 \cdot 10^{-7}$.

The top ten models are summarized in Table 2. While the first column contains the product of the marginal likelihood and the prior model probability, normalized within all visited models by (3.3), the second column refers to the frequencies of the models in the model sampling path. The two estimates differ considerably because the MCMC

[i] See `http://sites.google.com/site/jivsoft/Home/accurately-sum-the-elements-of-a-c---vector` for the original source.

algorithm has not yet converged to the posterior model distribution, which is not relevant here because we simply use it as a model search tool. Surprisingly, the top ten models agree on which variables to include and only vary in the powers contained in the respective power vectors. The MAP model configuration is

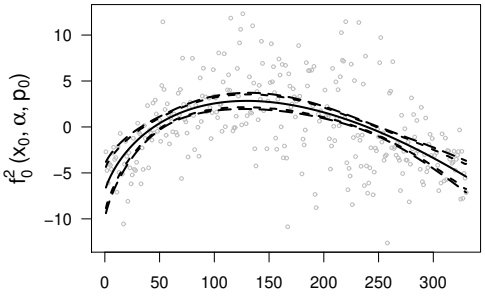
$$\begin{aligned}\mathbb{E}(y \,|\, \boldsymbol{x}, \boldsymbol{\theta}_{MAP}) = \beta_0 &+ \alpha_{01} x_0 + \alpha_{02} x_0 \log(x_0) + \alpha_{61} x_6 \\ &+ \alpha_{71} x_7^3 + \alpha_{81} x_8^2 + \alpha_{91} x_9^3 + \alpha_{10,1} x_{10}^{-\frac{1}{2}} + \alpha_{10,2} \log(x_{10})\end{aligned} \tag{4.2}$$

and can be obtained from the FP powers in the first row of Table 2 via the FP definition (1.4): for example, the MAP model contains the repeated power 1 for the first covariate x_0, which results in the FP part $\alpha_{01} x_0 + \alpha_{02} x_0 \log(x_0)$. The estimated FP parts are graphed on the original scales in Figure 1. The plotted curves result from Monte Carlo estimation using 20 000 samples from the posterior distribution of the coefficients in the MAP model, see Appendix C.1 for details. Note that the estimated mean curve matches the true mean curve obtained by using the posterior expected coefficients (up to Monte Carlo error). Yet, just plugging in the posterior expected shrinkage value 0.9897 into the covariance matrix of the posterior normal inverse-gamma distribution would lead to underestimation of the uncertainty, that means the credible intervals would be too small.

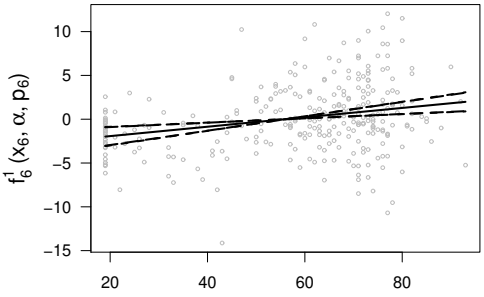
A comparison of the FPs for z_0 (day of the year), z_7 (temperature at Sandberg) and z_{10} (visibility) with their counterparts in Breiman and Friedman's (1985) Figure 5 reveals similarities. On the other hand, the functions for z_8 (inversion base height) and z_9 (pressure gradient) are quadratic and cubic power transformations with peaks at the negative of their shifts 0 and 70, respectively. This differs from Breiman and Friedman's (1985) transformations, which have their peaks at 1 000 and 0.

The FP mixtures of the BMA over the saved 3 000 models have been estimated by drawing 30 000 samples from their posterior distributions. The results are shown in Figure 2, see Appendix C.2 for details on the computations. Note that the estimates for f_6 and f_8 are based on less than 30 000 samples due to local inclusion probabilities $\hat{\pi}_i^{loc}$ smaller than unity, see Table 1. The function shapes are in general similar to those in the MAP model, but the uncertainty is larger of course. The mean estimate for z_8 exhibits a peak around 1 000 and approaches Breiman and Friedman's (1985) ACE transformation. Note that the centering of the design matrix columns is essential in order to obtain sensible results here, because correlations between the intercept and the

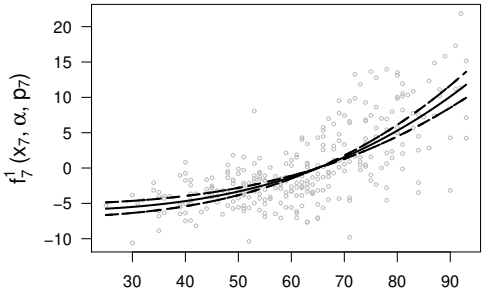
Figure 1 – *FPs estimates (means, solid lines) for the MAP model for the ozone data. The functions are plotted on the original covariate scales. Pointwise (short dashed lines) as well as simultaneous (long dashed lines) 95%-HPD intervals are given. The points are partial residuals to the FP mean curves.*



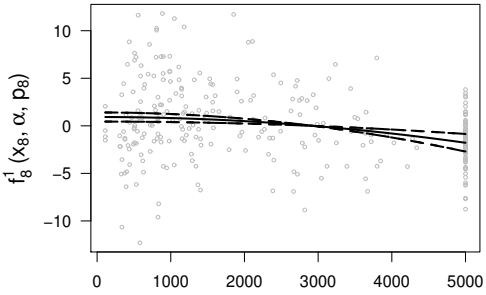
(a) Covariate z_0 (day of the year), cf. Breiman and Friedman's (1985) Figure 5f



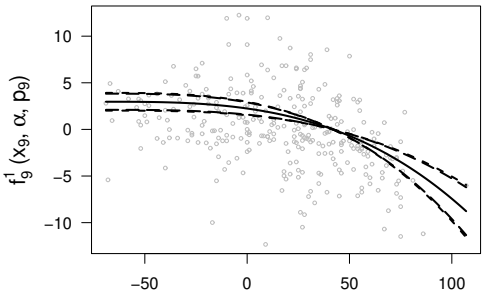
(b) Covariate z_6 (relative humidity [%])



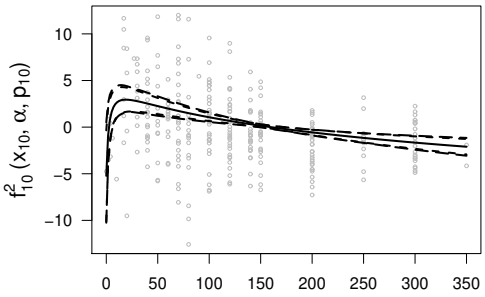
(c) Covariate z_7 (temperature at Sandberg [°F]), cf. Breiman and Friedman's (1985) Figure 5b



(d) Covariate z_8 (inversion base height [feet]), cf. Breiman and Friedman's (1985) Figure 5c



(e) Covariate z_9 (pressure gradient [mm Hg]), cf. Breiman and Friedman's (1985) Figure 5d



(f) Covariate z_{10} (visibility [miles]), cf. Breiman and Friedman's (1985) Figure 5e

Table 1 – *Preliminary transformation parameters and posterior inclusion probabilities for the nine covariates considered in the sampling process, which discovered* 907 986 *models constituting* $\hat{\boldsymbol{\Theta}}$. *Inclusion probabilities were estimated from* $\hat{\boldsymbol{\Theta}}$ *or from the best found* 3 000 *models in* $\hat{\boldsymbol{\Theta}}^{loc}$ *via* (3.4). *Note that the shifts* ξ_i *and scales* ζ_i *for the transformation* $x_i = (z_i + \xi_i)/\zeta_i$ *were chosen as in the* `mfp` *algorithm, see Appendix D for details.*

	ξ_i	ζ_i	$\hat{\pi}_i$	$\hat{\pi}_i^{loc}$
z_0	0	100	1.0000	1.0000
z_4	0	10 000	0.5758	0.3812
z_5	1	10	0.2629	0.1692
z_6	0	100	0.8447	0.8767
z_7	0	100	0.9994	1.0000
z_8	0	1 000	0.7039	0.7567
z_9	70	100	1.0000	1.0000
z_{10}	2	100	0.9991	1.0000
z_{11}	0	100	0.0886	0.0595

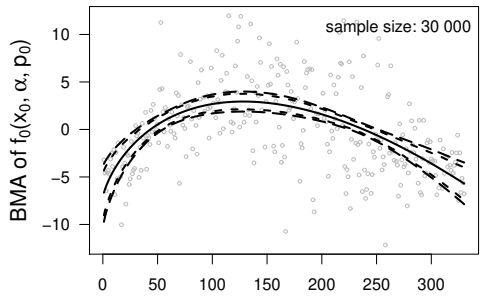
FPs would result in much larger and non-interpretable credible bands.

Three different models were compared by computing their predictions $\{\hat{y}_i\}$ for the test set data and quantifying the distance of these predictions to the actual values $\{y_i\}$ by means of the root mean squared prediction error (RMSPE). The `mfp` model (4.1) results in RMSPE = 3.579. The MAP model (4.2), which had been found by sampling from the posterior model distribution, is more successful with 3.512. Its RMSPE is even better than the result 3.571 of the BMA, whose predictions have been obtained by averaging over all 3 000 model-specific predictions.

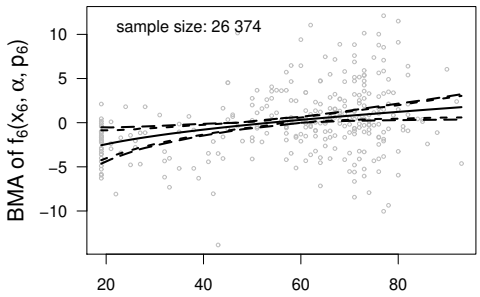
5 Discussion

This paper has implemented the multiple FP modelling approach, which combines variable selection and "parsimonious parametric modelling" (Royston and Altman 1994) of the covariate effects, within a Bayesian framework for normal linear regression. The

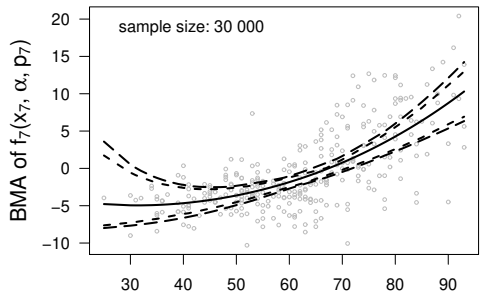
Figure 2 – *FPs mixture estimates (solid lines) for the BMA over the best 3 000 FP models for the ozone data. The functions are plotted on the original covariate scales. Pointwise (short dashed lines) as well as simultaneous (long dashed lines) 95%-HPD intervals are given. The points are partial residuals to the FP mean curves: the sample sizes underlying each function estimate are printed in the top corners.*



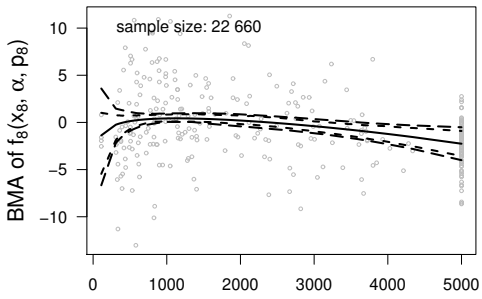
(a) Covariate z_0 (day of the year)



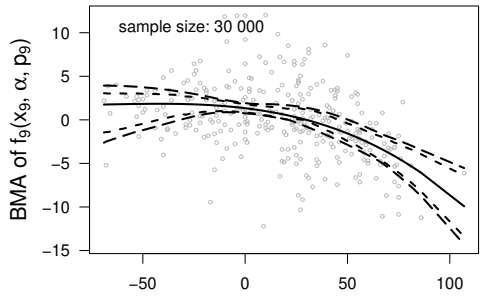
(b) Covariate z_6 (relative humidity [%])



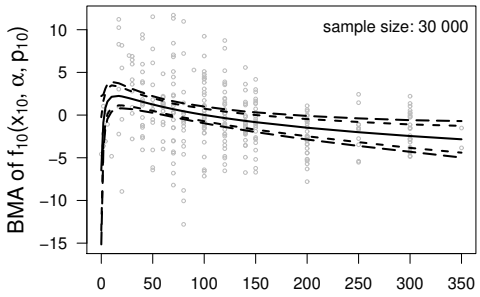
(c) Covariate z_7 (temperature at Sandberg [°F])



(d) Covariate z_8 (inversion base height [feet])



(e) Covariate z_9 (pressure gradient [mm Hg])



(f) Covariate z_{10} (visibility [miles])

Table 2 – *Summary of the top ten models in terms of posterior probability. The power vectors* $\boldsymbol{p}_4, \boldsymbol{p}_5$ *and* $\boldsymbol{p}_{11}$ *have always been empty and hence are omitted from the table.*

i	$\hat{f}(\boldsymbol{\theta}_i \mid \mathcal{D})^{\text{a}}$ $\times 10^4$	$\hat{f}_{freq}(\boldsymbol{\theta}_i \mid \mathcal{D})^{\text{b}}$ $\times 10^4$	$\log f(\mathcal{D} \mid \boldsymbol{\theta}_i)$	$\mathbb{E}(t \mid \mathcal{D}, \boldsymbol{\theta}_i)$	$\boldsymbol{p}_0$	$\boldsymbol{p}_6$	$\boldsymbol{p}_7$	$\boldsymbol{p}_8$	$\boldsymbol{p}_9$	$\boldsymbol{p}_{10}$
1	22.74	0.75	-40515.99	0.989687	1, 1	1	3	2	3	-0.5, 0
2	21.68	0.33	-40516.03	0.989683	1, 1	0.5	3	2	3	-0.5, 0
3	18.53	0.87	-40516.19	0.989668	1, 1	1	3	2	3	-0.5, -0.5
4	18.10	0.69	-40516.21	0.989666	1, 1	1	3	3	3	-0.5, 0
5	17.52	0.15	-40516.25	0.989663	1, 1	0.5	3	3	3	-0.5, 0
6	16.19	0.13	-40516.33	0.989655	1, 1	1	3	2	3	-1, 0
7	15.74	0.16	-40516.35	0.989653	1, 1	0.5	3	2	3	-1, 0
8	15.44	0.43	-40516.37	0.989651	1, 1	1	3	3	3	-0.5, -0.5
9	15.40	0.08	-40516.38	0.989651	1, 1	0.5	3	3	3	-0.5, -0.5
10	15.31	0.10	-40516.38	0.989650	1, 1	0	3	2	3	-0.5, 0

[a] The posterior probabilities are proportional to the exponential transformation of the sum of $\log f(\mathcal{D} \mid \boldsymbol{\theta})$ and $\log f(\boldsymbol{\theta}) = -\sum_{j=1}^{k} \log\big[d(m_j)(m_{max}+1)\big]$, where here $m_{max} = 2$. We obtain $\log f(\boldsymbol{\theta}_i) = -4\log(8) - 2\log(36) - 9\log(3) = -25.372315$ for all models $i = 1, \ldots, 10$.

[b] Model frequencies in the Markov chain of the model sampling algorithm.

Bayesian perspective allows coherent inference for models, covariate inclusion and FPs. Model selection is the main issue and has been addressed by a stochastic search algorithm that is a form of an MCMC algorithm. This path is computationally more demanding than simple stepwise search procedures, but it is theoretically well-founded. Model averaging is a valuable alternative, which directly accounts for model uncertainty, and the used hyper-g prior ensures that the resulting predictions are consistent for the true FP model's predictions.

Simultaneous covariate and transformation selection in the linear model has been done by Hoeting and Ibrahim (1998) and Hoeting, Raftery, and Madigan (2002), who give examples from the Box-Cox family of transformations and change-point transformations, respectively. Gottardo and Raftery (2009) use Box-Cox transformations also for the response variable. However, this complicates the MCMC algorithm considerably.

The proposed prior distributions express noninformativeness both about the models and the model parameters in order to do justice to the situations in which the modelling approach will usually be applied. We have used a quasi-default prior where only the

hyperparameter $a \in (3, 4]$ has to be chosen by the user. We have conducted a sensitivity study which has shown that the results are not sensitive to the hyperparameter choice in this range, and only abnormal choices $a \gg 4$ lead to a stronger shrinkage of the fit towards the mean. So our approach avoids potentially dangerous manual tuning of the smoothing parameter g and at the same time allows the computation of the marginal likelihood for each model in question. The 'Shotgun Stochastic Search' algorithm by Hans, Dobra, and West (2007) could therefore in principle be applied here, and we plan to test its implementation for the Bayesian FPs in the future. This search algorithm would be advantageous to efficiently use the full parallel computing power of clusters of multiple computers or future many-core workstations. Moreover, we do not need to implement complex reversible jump MCMC algorithms as that proposed by Jasra, Stephens, and Holmes (2007) to effectively traverse the model space.

The computational costs of the method are moderate, which is at least partly due to the use of a compiled language for the algorithm implementation. Besides allowing all maximum degrees the user wishes, the approach can take account of model uncertainty via Bayesian model averaging. This possibility should be used for checking the conclusions drawn from single models.

Immediate extensions of the implemented FPs could include other transformations. For instance, other powers in the set $\mathcal{S}$ or the exponential function would provide a bigger model class. Even trigonometric functions could be useful for the description of, e. g., seasonal data or blood measurements. Another improvement of the current procedure would be to provide the opportunity to contain hierarchical interactions in the linear predictor. At the moment, only manual input of products of covariate vectors is possible, and this does not prevent the algorithm from proposing non-hierarchical and thus non-interpretable models. The implementation of hierarchical interactions (with non-hierarchical models having prior probability zero) would necessitate adaption of the move types and hence adaption of the acceptance probability formulas.

Furthermore, the sampling approach used can readily be extended to distributions for which auxiliary variable methods that complement the linear regression model exist. For example, Holmes and Held (2006) extend the Albert and Chib (1993) method for probit regression to binary and multinomial logistic regression models. An integration of their findings into the multiple FP approach could be fruitful, as logistic regression is probably the second most important regression model. Similarly, for Poisson regres-

sion models auxiliary mixture sampling has been proposed by Frühwirth-Schnatter and Wagner (2006) and Frühwirth-Schnatter, Frühwirth, Held, and Rue (2009).

Acknowledgments

This paper has benefited from detailed comments by the Associate Editor and two referees. We are also grateful to Willi Sauerbrei for helpful comments in the initial phase of this project.

References

Abramowitz, M. and I. A. Stegun (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (ninth Dover printing, tenth GPO printing ed.). New York: Dover.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Ambler, G. and P. Royston (2001). Fractional polynomial model selection procedures: Investigation of type I error rate. *Journal of Statistical Computation and Simulation 69*(1), 89–108.

Anderson, I. J. (1999). A distillation algorithm for floating-point summation. *SIAM Journal on Scientific Computing 20*(5), 1797–1806.

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *Annals of Statistics 32*(3), 870–897.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science 10*(1), 3–66.

Box, G. E. P. and P. W. Tidwell (1962). Transformation of the independent variables. *Technometrics 4*(4), 531–550.

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association 80*(391), 580–598.

Brooks, S. P., N. Friel, and R. King (2003). Classical model selection via simulated annealing. *Journal of the Royal Statistical Society. Series B (Methodological) 65*(2), 503–520.

Denison, D. G. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Chichester: Wiley.

Fouskakis, D., I. Ntzoufras, and D. Draper (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics 3*(2), 663–690.

Frühwirth-Schnatter, S. and H. Wagner (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika 93*(4), 827–841.

Frühwirth-Schnatter, S., R. Frühwirth, L. Held, and H. Rue (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing 19*(4), 479–492.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*(2), 339–373.

Gottardo, R. and A. Raftery (2009). Bayesian robust transformation and variable selection: a unified approach. *The Canadian Journal of Statistics 37*(3), 361–380.

Govindarajulu, U. S., E. J. Malloy, B. Ganguli, D. Spiegelman, and E. A. Eisen (2009). The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *International Journal of Biostatistics 5*(1), 1–19.

Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association 102*(478), 507–516.

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

Hoeting, J. A. and J. G. Ibrahim (1998). Bayesian predictive simultaneous variable and transformation selection in the linear model. *Journal of Computational Statistics and Data Analysis 28*(1), 87–103.

Hoeting, J. A., A. E. Raftery, and D. Madigan (2002). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics 11*(3), 485–507.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*(1), 145–168.

Jasra, A., D. A. Stephens, and C. C. Holmes (2007). Population-based reversible jump Markov chain Monte Carlo. *Biometrika 94*(4), 787–807.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association 103*(481), 410–423.

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review 63*(2), 215–232.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

Royston, P. and D. Altman (1997). Approximating statistical functions by using fractional polynomial regression. *Journal of the Royal Statistical Society. Series D (The Statistician) 46*(3), 411–422.

Royston, P. and D. G. Altman (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 43*(3), 429–467.

Royston, P. and W. Sauerbrei (2008). *Multivariable Model-building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continous Variables*. Wiley Series in Probability and Statistics. Chichester: Wiley.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

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. *Journal of Computational Statistics and Data Analysis 50*(12), 3464–3485.

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. Series A (Statistics in Society) 162*(1), 71–94.

Shkedy, Z., M. Aerts, G. Molenberghs, P. Beutels, and P. van Damme (2006). Modelling force of infection from prevalence data using fractional polynomials. *Statistics in Medicine 25*(9), 1577–1591.

Sutradhar, B. C. (1986). On the characteristic function of multivariate Student t-distribution. *The Canadian Journal of Statistics 14*(4), 329–337.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Volume 6 of *Studies in Bayesian Econometrics and Statistics*, Chapter 5, pp. 233–243. Amsterdam: North-Holland.

A Numerical calculation of hyper-g quantities

In order to calculate the Bayes factor (Liang et al. 2008, formula (17)) and the posterior expectation of g or the shrinkage factor $g/(1+g)$ given model $\boldsymbol{\theta}$ (Liang et al. 2008, formulas (18) and (19)), integrals of the common form

$$\psi_{\boldsymbol{\theta}}(b,c) := \int_0^\infty g^{b-1}(1+g)^{(n-1-p_{\boldsymbol{\theta}}-c)/2}\left[1+(1-R^2_{\boldsymbol{\theta}})g\right]^{-(n-1)/2}\,dg$$

need to be computed. This results from

$$\begin{aligned} f(g\,|\,\mathcal{D},\boldsymbol{\theta}) &\propto f(\mathcal{D}\,|\,g,\boldsymbol{\theta})f(g\,|\,\boldsymbol{\theta}) \\ &\propto \frac{(1+g)^{(n-1-p_{\boldsymbol{\theta}})/2}}{\left[1+(1-R^2_{\boldsymbol{\theta}})g\right]^{(n-1)/2}}\frac{a-2}{2}(1+g)^{-a/2} \\ &= \frac{a-2}{2}(1+g)^{(n-1-p_{\boldsymbol{\theta}}-a)/2}\left[1+(1-R^2_{\boldsymbol{\theta}})g\right]^{-(n-1)/2} \end{aligned}$$

for non-null models $\boldsymbol{\theta}$. The normalizing constant of this posterior density is the Bayes factor of model $\boldsymbol{\theta}$ versus the null-model $\mathcal{M}_N$,

$$BF(\boldsymbol{\theta} : \mathcal{M}_N) = \frac{a-2}{2} \int_0^\infty (1+g)^{(n-1-p_{\boldsymbol{\theta}}-a)/2} \left[1 + (1 - R^2_{\boldsymbol{\theta}})g\right]^{-(n-1)/2} dg = \frac{a-2}{2} \psi_{\boldsymbol{\theta}}(1, a).$$

The *a posteriori* expected value of g in model $\boldsymbol{\theta}$ is thus

$$\mathbb{E}(g \,|\, \mathcal{D}, \boldsymbol{\theta}) = \frac{\int_0^\infty g \frac{a-2}{2} (1+g)^{(n-1-p_{\boldsymbol{\theta}}-a)/2} \left[1 + (1 - R^2_{\boldsymbol{\theta}})g\right]^{-(n-1)/2} dg}{\frac{a-2}{2} \psi_{\boldsymbol{\theta}}(1, a)} = \frac{\psi_{\boldsymbol{\theta}}(2, a)}{\psi_{\boldsymbol{\theta}}(1, a)}.$$

Similarly, the posterior expected value of $t = g/(g+1)$ given the model $\boldsymbol{\theta}$ is

$$\mathbb{E}(t \,|\, \mathcal{D}, \boldsymbol{\theta}) = \frac{\psi_{\boldsymbol{\theta}}(2, a+2)}{\psi_{\boldsymbol{\theta}}(1, a)}. \tag{A.1}$$

The first way of computing the $\psi_{\boldsymbol{\theta}}$ function can be derived by employing the change of integration variable g to $t := g/(g+1)$. The integration range is mapped onto the unit interval and by the integral representation of the Gaussian hypergeometric function (Liang et al. 2008, formula (20)) we obtain

$$\begin{aligned} \psi_{\boldsymbol{\theta}}(b, c) &= \int_0^1 t^{b-1} (1-t)^{(p_{\boldsymbol{\theta}}+c)/2-b-1} (1 - R^2_{\boldsymbol{\theta}} t)^{-(n-1)/2} \, dt \\ &= Beta\left(b, \frac{p_{\boldsymbol{\theta}}+c}{2} - b\right) \cdot {}_2F_1\left(\frac{n-1}{2}; b; \frac{p_{\boldsymbol{\theta}}+c}{2}; R^2_{\boldsymbol{\theta}}\right). \end{aligned}$$

Liang et al. (2008) have reported occasional numerical difficulties with the Gaussian hypergeometric function in the Cephes library (available from netlib). We have implemented their alternative Laplace approximation, but its use was not necessary in our applications.

B Model sampling acceptance probabilities

Suppose a *DEATH* happened and removed a power p from the ith FP. The prior odds then are calculated analogously and equal

$$\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} = \frac{d(m_i)}{d(m_i - 1)} = \frac{|\mathcal{S}| - 1 + m_i}{m_i}.$$

Similarly, the proposal probabilities are computed in the same manner as for a *BIRTH*. The results are

$$q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}) = d_{p_{\boldsymbol{\theta}}} \times \frac{1}{|\mathcal{P}|} \times \frac{\mathbf{1}_{\boldsymbol{p}_i}(p)}{m_i} \quad \text{and} \quad q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}') = b_{p_{\boldsymbol{\theta}}-1} \times \frac{1}{|\mathcal{F}'|} \times \frac{1}{|\mathcal{S}|}.$$

Thus, for a *DEATH* move,

$$\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}')}{q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})} = \frac{b_{p_{\boldsymbol{\theta}}-1}}{d_{p_{\boldsymbol{\theta}}}} \frac{|\mathcal{P}|}{|\mathcal{F}'|} \frac{|\mathcal{S}| - 1 + m_i}{\mathbf{1}_{\boldsymbol{p}_i}(p) \cdot |\mathcal{S}|}.$$

Second, consider a *MOVE* which substituted the power q for the power p in the ith FP. Obviously, the prior odds are one, because the decisive degrees $\{m_j\}_{j=1}^{k}$ have not changed. The proposal probabilities are

$$q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}) = m_{p_{\boldsymbol{\theta}}} \times \frac{1}{|\mathcal{P}|} \times \frac{\mathbf{1}_{\boldsymbol{p}_i}(p)}{m_i} \times \frac{1}{|\mathcal{S}|} \quad \text{and} \quad q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}') = m_{p_{\boldsymbol{\theta}}} \times \frac{1}{|\mathcal{P}|} \times \frac{\mathbf{1}_{\boldsymbol{p}'_i}(q)}{m_i} \times \frac{1}{|\mathcal{S}|},$$

differing only in one number because the degrees and, consequently, the dimension w and the number of present covariates have not been altered. The proposal ratio hence reduces to the ratio of the number of powers q in $\boldsymbol{p}'_i$ of the new model $\boldsymbol{\theta}'$ to the number of powers p in $\boldsymbol{p}_i$ of the current model $\boldsymbol{\theta}$, i. e.

$$\frac{q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}')}{q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta})} = \frac{\mathbf{1}_{\boldsymbol{p}'_i}(q)}{\mathbf{1}_{\boldsymbol{p}_i}(p)}.$$

Lastly, suppose a *SWITCH* exchanged the power vectors of the ith and the jth FP. The prior odds are one, because

$$\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} = \frac{f(\boldsymbol{p}'_i) f(\boldsymbol{p}'_j)}{f(\boldsymbol{p}_i) f(\boldsymbol{p}_j)}$$

and $\boldsymbol{p}'_i = \boldsymbol{p}_j$, $\boldsymbol{p}'_j = \boldsymbol{p}_i$. If by chance $\boldsymbol{p}_i = \boldsymbol{p}_j$, then obviously the proposal ratio equals one, and we do not need to think about probabilities contributed by *MOVE*s which also result in the same model vector, and vice versa for the *MOVE* acceptance probabilities. If $\boldsymbol{p}_i \neq \boldsymbol{p}_j$, the proposal probabilities are

$$\begin{aligned} q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}) &= s_{p_{\boldsymbol{\theta}}} \times \frac{1}{|\mathcal{P}|} \times \frac{1}{k-1} \times \big[\mathbb{I}(m_i > 0) + \mathbb{I}(m_j > 0)\big], \\ q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}') &= s_{p_{\boldsymbol{\theta}'}} \times \frac{1}{|\mathcal{P}'|} \times \frac{1}{k-1} \times \big[\mathbb{I}(m'_i > 0) + \mathbb{I}(m'_j > 0)\big], \end{aligned}$$

and they are equal because the dimension $p_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}'}$ and the number of present covariates $|\mathcal{P}| = |\mathcal{P}'|$ have not changed. So also the proposal ratio of the *SWITCH* moves equals one.

C Posterior summaries

Having explored the posterior model space that was defined in Section 2, one is interested in at least two things: first, one wants to get a general idea of the posterior model distribution. For instance, one would like to know how probable the inclusion of certain covariates is or what models are most plausible after taking account of the observed data. Another kind of posterior summary is BMA, which can serve us as a benchmark for single models. Second, if one selects a single model which is the 'best' in terms of posterior probability or interpretability, point estimates and credible intervals for its coefficients or crediblity regions for the FP functions are of particular interest. The necessary methods are developed in this Section and were applied in the context of an elaborate example in Section 4.

C.1 Describing a single FP model

In this section we will introduce techniques for summarizing a single multiple FP model.

Estimation of coefficients and regression variance Having decided on a certain model, the intercept β_0 and the various regression coefficients $\{\alpha_{ij}\}$, which had been collected into the large coefficient vector $\boldsymbol{\beta}$, can be treated equally and are thus denoted as $\beta_0, \ldots, \beta_p$. As was shown in Section 3, *a posteriori* the whole vector $\boldsymbol{\beta}$ follows a p-variate Student distribution, conditional on the covariance factor g. Because subvectors of a vector with multivariate Student distribution are themselves t-distributed with their respective parts of the mean vector and diagonal block of the original scale matrix as parameters (Sutradhar 1986), the ith coefficient follows a univariate t-distribution:

$$\beta_i \,|\, \mathcal{D}, g \sim \mathrm{t}(m_i, 2b/(n-1)V_{ii}, n-1), \tag{C.1}$$

where $\boldsymbol{m} = (m_0, \ldots, m_p)^T$ and $\boldsymbol{V} = (V_{ij})_{0 \le i,j \le p}$ are assumed. Standardization leads to a central t-distribution with unit scale and the same degrees of freedom, i. e. with $s_i = 2b/(n-1)V_{ii}$ we can write

$$\frac{\beta_i - m_i}{\sqrt{s_i}} \,|\, \mathcal{D}, g \sim \mathrm{t}(n-1).$$

If the uncertainty from g should be taken into account for the model-specific part $\boldsymbol{\beta}$, the law of iterated expectations yields

$$\begin{aligned}\hat{\boldsymbol{\beta}} := \mathbb{E}(\boldsymbol{\beta} \,|\, \mathcal{D}) &= \mathbb{E}\left[\mathbb{E}\left(\boldsymbol{\beta} \,|\, \mathcal{D}, g\right) \,|\, \mathcal{D}\right] \\ &= \mathbb{E}\left[\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{(OLS)} \,|\, \mathcal{D}\right] \\ &= \mathbb{E}\left(t \,|\, \mathcal{D}\right) \cdot \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{(OLS)} \end{aligned} \tag{C.2}$$

that is the OLS estimate scaled by (A.1).

A conditional posterior equal-tailed $(1-\alpha)$-credible interval for β_i that is centered around the posterior mean m_i can be calculated numerically only if g is held fixed. Otherwise, equal-tailed or highest posterior density (HPD) credible intervals can easily be Monte Carlo estimated via N samples, say, obtained from the sampling algorithm in Section 3.1. Equal-tailed credible intervals are bounded by the empirical $(1-\alpha)/2$- and $(1+\alpha)/2$-quantiles of the samples. The HPD intervals may be calculated in the following manner. Let the number of samples to be included in the empirical HPD interval be $l = [N(1-\alpha)]$. After ordering the samples, the width of all $N-l$ possible contiguous sets comprising l elements is calculated. The set with minimal width is then the empirical $(1-\alpha)$-HPD interval.

Likewise, one can proceed to estimate the marginal posterior distribution of the regression variance σ^2.

Estimation of FP curves If a FP function is part of the linear predictor, experiencing the estimated relationship and uncertainty about it visually will be more helpful to the user than reading credible interval bounds of the associated coefficients $\{\alpha_j\}$. Since the approach is a form of additive modelling, the illustration of the effect fortunately boils down to making a graph of a univariate function—namely the FP estimate $f^{\hat{m}}(x; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{p}})$. The estimates of the power vector $\boldsymbol{p}$ and the degree m are assumed fixed here, as they are part of the model definition, and only uncertainty about the coefficient vector $\boldsymbol{\alpha}$ remains to be considered.

Evaluation of the function estimate can be implemented by building a fine grid of x-values in the observed range and interpolating the function values at these abscissae. Each ordinate is computed by transforming x into the design vector $\boldsymbol{h}(x) =$

$(h_1(x), \ldots, h_m(x))$ and multiplying it with the point estimate $\hat{\boldsymbol{\alpha}}$ which is the appropriate subvector of the grand posterior mode (C.2). Recall that the model parameters determine the transformations $\{h_j\}$ via (1.4). This point estimate,

$$f^{\hat{m}}(x; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{p}}) = \boldsymbol{h}(x)\hat{\boldsymbol{\alpha}}, \tag{C.3}$$

is the posterior expected function ordinate in the given model.

Pointwise credible intervals are well suited for expressing the range of plausible *function values* at a certain predictor value. Yet, they are not qualified for illustrating the range of plausible FP *functions*. In particular, the credible level $1 - \alpha$ cannot be interpreted as the posterior probability for a curve with coefficients drawn from the posterior to be embedded in the region circumscribed by the connected lower and upper pointwise bounds. The issue is more urgent here than, for example, in spline regression, where the approximating functions are 'local' by nature. The FPs belong to the family of parametric models, that is why they are 'global', meaning that a change of the function in one point affects the whole curve.

A simulation-based approach to constructing a simultaneous $(1 - \alpha)$-credible region could proceed as follows. One starts with drawing N samples $\boldsymbol{\alpha}^{(i)}$, $i = 1, \ldots, N$, from the posterior distribution $\boldsymbol{\alpha} \,|\, \mathcal{D}$. This again works like the algorithm sketched in Section 3.1, that is one samples models covariance factors $g^{(i)}$ using the inverse sampling scheme and samples t-distributed vectors $\boldsymbol{\alpha}^{(i)}$ using the formulas in step 2 with location vector and scale matrix determined by the respective $g^{(i)}$ via (2.4) and (2.3). Afterwards one computes the respective function estimates $f^{\hat{m}}(x; \boldsymbol{\alpha}^{(i)}, \hat{\boldsymbol{p}})$ at a grid of k abscissae. An algorithmically advantageous formulation of a simultaneous $(1 - \alpha)$-credible region for the function which always includes the mean curve can then be derived from the nonparametric approach that was developed by Besag, Green, Higdon, and Mengersen (1995), particularly from their one-sided upper simultaneous credible band (SCB). It is based on the $(N \times k)$-matrix of the function values $\big(v_j^{(i)} := f^{\hat{m}}(x_j; \boldsymbol{\alpha}^{(i)}, \hat{\boldsymbol{p}})\big)_{ij}$, where each function estimate is allocated in one row, and the different function estimates at a certain x-value are allocated in one column each. Let the absolute distances between the function values and the mean curve values be collected in a matrix

$$\big(d_j^{(i)} := |v_j^{(i)} - \boldsymbol{h}(x_j)\hat{\boldsymbol{\alpha}}|\big)_{ij}$$

of the same dimension. Now each column of $(d_j^{(i)})_{ij}$ is ordered seperately to obtain the

ranks $\{r_j^{(i)}\}, j = 1, \ldots, k$, of the absolute distances at each of the k grid points. Denote the number of functions which shall be included in the credible set by $l := [N(1-\alpha)]$ and the lth order statistic from the set of rowwise maximum ranks $\{\max_{j=1,\ldots,k} r_j^{(i)}\}_{i=1}^N$ by r^*. The upper bound on the ranks, r^*, determines the SCB that consists of the k elementwise ranges

$$\min, \max \left\{ v_j^{(i)} \,|\, r_j^{(i)} \leq r^*, i = 1, \ldots, N \right\}, \quad j = 1, \ldots, k.$$

Unlike the equal-tailed SCB of Besag et al. (1995), this credible region will in general not be invariant to strictly monotone transformations of the values $\{v_j^{(i)}\}$. In this respect the proposed SCB resembles the single HPD interval. However, this fact should not concern us unduly, as we usually will not want to consider transformations of the FP function values after having calculated the credible band.

The distance between the model fit and the data can be gauged by adding partial residuals to each function plot. For the ith FP they are defined as

$$\hat{\varepsilon}_j^{(i)} = f^{\hat{m}_i}(x_{ij}; \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{p}}_i) + \hat{\varepsilon}_j, \quad j = 1, \ldots, n,$$

where the jth raw residual $\hat{\varepsilon}_j$ is the difference between the response y_j and the model fit $\hat{y}_j$. The fit $\hat{y}_j$ of the multiple FP model is the posterior mean of the linear predictor η in (1.6) evaluated at $\boldsymbol{x}_j$. The rationale behind this definition of fit is based on the fact that in linear regression with identity link the linear predictor models the mean $\mathbb{E}(y_j \,|\, \boldsymbol{x}_j)$ of the response y_j directly. The posterior expectation of the modelled mean $\beta_0 + \boldsymbol{b}_j^T \boldsymbol{\beta}$ simply is the linear combination $\bar{y} + \boldsymbol{b}_j^T \hat{\boldsymbol{\beta}}$. The design vector $\boldsymbol{b}_j$ depends on the covariate values $\boldsymbol{x}_j$ via (1.6). HPD intervals for the modelled mean may again be Monte Carlo estimated by applying the sampling scheme in Section 3.1.

Via the raw residuals, the partial residuals take into account all other variables. A satisfying fit of the ith FP is indicated if the function estimate reflects the plotted relationship between the covariate x_i and the partial residuals $\{\hat{\varepsilon}_j^{(i)}\}$ quite well.

C.2 Describing the posterior model distribution

Having explored the whole or a part of the posterior model distribution by an exhaustive search or a posterior sampling procedure, respectively, one is not only interested in a single model, but also in the model distribution. Besides analyzing a table of the most probable models, the BMA approach can be insightful.

If a FP covariate is included in an FP model, the conditional distribution of the FP curve is of interest. This distribution can be estimated by the algorithm described in Section 3.1. Every sample of uncertain covariates is conditional on the inclusion of the covariate, so the inclusion probabilities (3.4) must always be examined in parallel.

Based on a sample of size N that is well above the size of $\hat{\boldsymbol{\Theta}}^{(loc)}$ pointwise estimates and credible intervals for the averages of FP functions are available after linear transformation of the associated coefficients via the appropriate design vectors. Note that Bayesian model averages of the single FP coefficients $\{\alpha_{ij}\}$ are not very meaningful, as one is not interested in the coefficient of e. g. x^{-1} given that it is included in the design vector. Only the average FP function as a whole is informative. Simultaneous credible bands for the partial predictor functions can be estimated by applying the procedure, that was described in the former section, on the function samples. The only difference is the sample space—while in the previous section we sampled from a single model, we now sample from a model average.

The simplest BMA fit $\hat{y}_j$ for the jth response value y_j is the marginal posterior mean of the modelled linear predictor at the independent values $\boldsymbol{x}_j$. It arises from the model-specific fits through posterior model probability weighted averaging by applying the law of iterated expectations:

$$\begin{aligned}\hat{y}_j &= \mathbb{E}_{\boldsymbol{\beta}\,|\,\mathcal{D}}\big(\eta(\boldsymbol{x}_j)\,|\,\mathcal{D}\big) = \mathbb{E}_{\boldsymbol{\theta}\,|\,\mathcal{D}}\left\{\mathbb{E}_{\boldsymbol{\beta}\,|\,\boldsymbol{\theta},\mathcal{D}}\big(\beta_0 + \boldsymbol{b}_{\boldsymbol{\theta},j}^T\boldsymbol{\beta}_{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta},\mathcal{D}\big)\right\} \\ &= \bar{y} + \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \boldsymbol{b}_{\boldsymbol{\theta},j}^T\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}\,|\,\mathcal{D}) \\ &\approx \bar{y} + \sum_{\boldsymbol{\theta}\in\hat{\boldsymbol{\Theta}}} \boldsymbol{b}_{\boldsymbol{\theta},j}^T\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}\,|\,\mathcal{D}).\end{aligned} \tag{C.4}$$

This allows again heuristic goodness-of-fit checks for the residuals $y_j - \hat{y}_j$, for instance plotting the partial residuals as desribed in the previous subsection.

D Ozone data description

The ozone data presented in Breiman and Friedman (1985) detail the relationship between atmospheric ozone concentration and meteorology in the Los Angeles basin. The data is available by FTP from Leo Breiman's website. Breiman and Friedman (1985) wanted to predict the maximum one-hour average ozone concentration of the next day from nine meteorological variables. All variables are listed in Table 3.

The authors used their alternating conditional expectations (ACE) algorithm to estimate the nonparametric transformations of both the response and the independent variables that maximize the fraction of variance explained by the multiple linear regression. There is a link to our Bayesian approach, but we aim to maximize the posterior model probability within a model space that only contains parametric transformations of the independent variables.

Table 3 – *Description of the variables in the ozone data set, which spans all 366 days of the leap year 1976.*

Variable	Description	Measurement location	Missing
y	Maximum 1-hour average ozone level [ppm]	Upland, CA	5
z_1	Month		
z_2	Day of month		
z_3	Day of week		
z_4	500 millibar pressure height [m]	Vandenberg AFB	12
z_5	Wind speed [mph]	LAX	
z_6	Relative humidity [%]	LAX	15
z_7	Temperature [°F]	Sandberg, CA	2
z_8	Inversion base height [feet]	LAX	15
z_9	Pressure gradient [mm Hg] from LAX to Daggett, CA		1
z_{10}	Visibility [miles]	LAX	
z_{11}	Inversion base temperature [°F]	LAX	14
z_{12}	Temperature [°F]	El Monte, CA	139

Since the temperature values at El Monte are missing for 139 days, which is more than a third of the total 366 records, this variable (z_{12}) is not included in the analysis. The covariates z_7 and z_{11}, which are temperature readings at Sandberg and at Los Angeles International Airport, respectively, may serve as partial surrogate variables, because of their high linear correlations (0.91 and 0.93) with z_{12} in the data set. After omitting the incomplete cases, we arrive at 330 observations, which is the sample size reported

by Breiman and Friedman (1985). Borrowing from them, we form an additional time variable z_0 that contains the day of the year in order to capture extra seasonal variation. The month and day variables z_1, z_2 and z_3 are not used.

The transformation method is adopted from the `mfp` algorithm (Sauerbrei, Meier-Hirmer, Benner, and Royston 2006). Each original covariate z_i is shifted and rescaled as $x_i = (z_i + \xi_i)/\zeta_i$, where the shift ξ_i and the scale ζ_i are computed as follows. If the smallest observed value $\min_j z_{ij}$ is positive, no shift is made. Otherwise, the shift parameter is equated with the smallest positive increment in successive ordered values minus the minimum value and rounded up to the next first decimal place:

$$\xi_i = \begin{cases} 0 & \text{if } z_{i(1)} > 0, \\ \left\lceil \min_{z_{i(j)} \neq z_{i(j+1)}} \left\{ \left| z_{i(j+1)} - z_{i(j)} \right| - z_{i(1)} \right\} \cdot 10 \right\rceil /10 & \text{if } z_{i(1)} \leq 0 \end{cases}$$

The decimal log mean $r = \log_{10} \left\{ \frac{1}{n} \sum_{j=1}^{n} (z_{ij} + \xi_i) \right\}$ of the shifted values defines the scale parameter ζ_i via $\zeta_i = 10^{\text{sign}(r)} \cdot [|r|]$. Thus, small values are scaled up and big values are scaled down by powers of 10.

The dates of the test set observations are given in Table 4.

Table 4 – *Dates of the test set records in DD/MM/1976 format.*

17/1	11/2	10/3	26/3	31/3	7/4	10/4	27/4	2/5	15/5
17/5	20/5	6/6	7/6	10/6	6/7	10/7	20/7	27/7	2/8
3/8	9/8	11/9	28/9	13/10	15/11	30/11	4/12	7/12	14/12