Bayesian Fractional Polynomials

Daniel Sabanés Bové   Leonhard Held

Biostatistics Unit, Institute of Social and Preventive Medicine
University of Zurich

Consulting seminar

Zurich, 04/03/2010

Outline

1. Introduction
2. Bayesian Inference for Linear FP Models
   - Model Definition
   - Model inference
3. Application: Ozone data
4. Outlook and Discussion

University of Zurich

Additive linear regression modelling

Consider linear regression model with additive effects structure:

$$\mathbb{E}(y \mid \boldsymbol{x}) = \eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} f_i(x_i)$$

$\Rightarrow$ How to choose the transformations f_i ?

flexible $\leftarrow$	Nonparametric smoothers	(Automatic) parametric nonlinear modelling	(Manual) ad-hoc transformations	Assumption of linearity	$\rightarrow$ simple
	e. g. splines	e. g. Fractional Polynomials	e. g. classic polynomials	just $f_i(x_i) = \beta_i x_i$	

University of Zurich

Fractional Polynomials (Royston and Altman 1994)

$$f_i(x_i) = f(x_i; \boldsymbol{\alpha}_i, \boldsymbol{p}_i)$$

- Degree m_i, is the number of
- powers in the power vector $\boldsymbol{p}_i = (p_{i1} \leq \cdots \leq p_{im_i})$ from
- the set $\mathcal{S} = \left\{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, \ldots, \max\{3, m_i\}\right\}$.
- Coefficients are in the vector $\boldsymbol{\alpha}_i \in \mathbb{R}^{m_i}$.

$\Rightarrow$ Encompasses classic polynomial powers, e. g.
$\boldsymbol{p}_i = (1, 2) \Rightarrow f_i(x_i) = \alpha_{i1} x_i + \alpha_{i2} x_i^2$

$\Rightarrow$ Offers reciprocals, square roots and logarithm, e. g.
$\boldsymbol{p}_i = (-1, -\frac{1}{2}, 0) \Rightarrow f_i(x_i) = \alpha_{i1}/x_i + \alpha_{i2}/\sqrt{x_i} + \alpha_{i3} \log(x_i)$

$\Rightarrow$ Allows "repeated powers" with logarithm products, e. g.
$\boldsymbol{p}_i = (-2, -2) \Rightarrow f_i(x_i) = \alpha_{i1}/x_i^2 + \alpha_{i2} \log(x_i)/x_i^2$

University of Zurich

The special linear model for one observation

Rewrite as a special linear model:

$$\begin{aligned}\eta(\boldsymbol{x}) &= \beta_0 + \sum_{i=1}^{k} f(x_i; \boldsymbol{\alpha}_i, \boldsymbol{p}_i) \\ &= \beta_0 + \sum_{i=1}^{k} \boldsymbol{h}_i(x_i)^T \boldsymbol{\alpha}_i \\ &= \beta_0 + \boldsymbol{b}_{\boldsymbol{\theta}}^T \boldsymbol{\beta}_{\boldsymbol{\theta}}.\end{aligned}$$

- $\boldsymbol{h}_i(x_i) \in \mathbb{R}^{m_i}$ is the design vector determined by $\boldsymbol{p}_i$.
- $\boldsymbol{\theta} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_k)$ indexes the models,
- $p_{\boldsymbol{\theta}} = \dim(\boldsymbol{\theta}) = \sum_{i=1}^{k} m_i$ is the model dimension.
- $\boldsymbol{b}_{\boldsymbol{\theta}} = (\boldsymbol{h}_1(x_1)^T, \dots, \boldsymbol{h}_k(x_k)^T)^T \in \mathbb{R}^{p_{\boldsymbol{\theta}}}$,
- $\boldsymbol{\beta}_{\boldsymbol{\theta}} = (\boldsymbol{\alpha}_1{}^T, \dots, \boldsymbol{\alpha}_k{}^T)^T \in \mathbb{R}^{p_{\boldsymbol{\theta}}}$.

University of Zurich

The special linear model for n observations

For n observations collected in $\mathcal{D} = \{y_j, \boldsymbol{x}_j\}_{j=1}^{n}$
we can write

$$\boldsymbol{y} = \mathbf{1}_n \beta_0 + \boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ is the vector of iid errors.
- $\boldsymbol{B}_{\boldsymbol{\theta}} \in \mathbb{R}^{n \times p_{\boldsymbol{\theta}}}$ is the design matrix,
  a column-centered version of $(\boldsymbol{b}_{1\boldsymbol{\theta}}, \dots, \boldsymbol{b}_{n\boldsymbol{\theta}})^T$
  $\Rightarrow \beta_0$ is a common parameter of all models $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

University of Zurich

Bayesian inference plan

1. Choose prior for parameters $\boldsymbol{\xi_\theta} = (\beta_0, \boldsymbol{\beta_\theta}, \sigma^2, g)$ in model $\boldsymbol{\theta}$:

$$f(\boldsymbol{\xi_\theta} \,|\, \boldsymbol{\theta}) = f(\beta_0) f(\boldsymbol{\beta_\theta} \,|\, \sigma^2, g) f(\sigma^2) f(g).$$

2. Likelihood for the parameters is normal:

$$f(\mathcal{D} \,|\, \boldsymbol{\xi_\theta}, \boldsymbol{\theta}) = \mathrm{N}_n(\boldsymbol{y} \,|\, \mathbf{1}_n \beta_0 + \boldsymbol{B_\theta} \boldsymbol{\beta_\theta}, \sigma^2 \boldsymbol{I}_n).$$

3. Compute "marginal likelihood" of the model $\boldsymbol{\theta}$:

$$f(\mathcal{D} \,|\, \boldsymbol{\theta}) = \int_{\boldsymbol{\Xi_\theta}} f(\mathcal{D} \,|\, \boldsymbol{\xi_\theta}, \boldsymbol{\theta}) f(\boldsymbol{\xi_\theta} \,|\, \boldsymbol{\theta}) \, d\boldsymbol{\xi_\theta}.$$

4. Choose prior model probabilities $f(\boldsymbol{\theta})$.
5. Posterior model probabilities are then

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

Hyper-g prior $f(\boldsymbol{\xi}_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta})$ for model parameters

- Jeffreys' prior on $\sigma^2 > 0$:

$$f(\sigma^2) \propto 1/\sigma^2$$

- Flat prior on $\beta_0 \in \mathbb{R}$:

$$f(\beta_0) \propto 1$$

- g-prior on $\boldsymbol{\beta}_{\boldsymbol{\theta}} \in \mathbb{R}^{p_{\boldsymbol{\theta}}}$ (Zellner 1986):

$$\boldsymbol{\beta}_{\boldsymbol{\theta}} \,|\, \sigma^2, g \sim \mathrm{N}_{p_{\boldsymbol{\theta}}}\left(\mathbf{0}, \sigma^2 g \cdot (\boldsymbol{B}_{\boldsymbol{\theta}}^T \boldsymbol{B}_{\boldsymbol{\theta}})^{-1}\right)$$

- (Hyper-)Prior on covariance factor $g > 0$ (Liang et al. 2008):

$$\frac{g}{g+1} \sim \mathrm{Be}(1, \tfrac{a}{2} - 1), \quad a \in (3, 4]$$

$\Rightarrow$ Avoids manual choice of shrinkage parameter g

$\Rightarrow$ Several desirable asymptotic properties

$\Rightarrow$ Closed form for marginal likelihood $f(\mathcal{D} \,|\, \boldsymbol{\theta})$

University of Zurich

Shrinkage prior $f(\boldsymbol{\theta})$ on the model space

Assume for the prior model probabilities $f(\boldsymbol{\theta})$:

1. FP powers of different covariates are independent:

$$f(\boldsymbol{\theta}) = f(\boldsymbol{p}_1) \times \cdots \times f(\boldsymbol{p}_k)$$

2. For the ith covariate, each degree $0 \leq m_i \leq m_{max}$ is equally probable
3. Given the degree m_i, each combination of powers $\boldsymbol{p}_i \in \mathcal{S}^{m_i}$ is equally probable $\Rightarrow f(\boldsymbol{p}_i)$

$\Rightarrow$ Null model $\boldsymbol{\theta}_{\text{null}}$ with $p_{\boldsymbol{\theta}_{\text{null}}} = 0$ has highest prior probability:

$$f(\boldsymbol{\theta}_{\text{null}}) = 1/(m_{max} + 1)^k.$$

University of Zurich

Model inference

Problem:

- Model space $\boldsymbol{\Theta}$ gets very large for $k \geq 5$
- e. g. $m_{max} = 2, k = 5 \Rightarrow |\boldsymbol{\Theta}| = (1 + 8 + 36)^5 = 184\,528\,125$

$\Rightarrow$ Exhaustive exploration of $f(\boldsymbol{\theta} \,|\, \mathcal{D})$ not feasible!

Solution:

1. Sample from the distribution $f(\boldsymbol{\theta} \,|\, \mathcal{D})$ using a special Markov chain Monte Carlo (MCMC) algorithm.
2. Interpret the visited models as an estimate $\hat{\boldsymbol{\Theta}}$.
3. Take the *maximum a posteriori* (MAP) FP model $\hat{\boldsymbol{\theta}}_{\text{MAP}}$,
4. and/or do Bayesian Model Averaging (BMA) with weights

$$\hat{f}(\boldsymbol{\theta} \,|\, \mathcal{D}) \propto f(\mathcal{D} \,|\, \boldsymbol{\theta}) f(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \hat{\boldsymbol{\Theta}}.$$

Monte Carlo parameter estimation

- After the model sampling, estimate the parameters (across models) by Monte Carlo.
- For $l = 1, \ldots, L$ do:
  1. Draw $\boldsymbol{\theta}^{[l]}$ from $\hat{f}(\boldsymbol{\theta} \,|\, \mathcal{D})$.
  2. Sample the model parameters $\boldsymbol{\xi}_{\boldsymbol{\theta}^{[l]}}$:
     1. Draw $g^{[l]} \,|\, \boldsymbol{\theta}^{[l]}, \mathcal{D}$ using inverse sampling
     2. Draw $\beta_0^{[l]}, \boldsymbol{\beta}^{[l]} \,|\, g^{[l]}, \boldsymbol{\theta}^{[l]}, \mathcal{D}$ from Student t distributions
     3. Draw $\sigma^{2[l]} \,|\, g^{[l]}, \boldsymbol{\theta}^{[l]}, \mathcal{D}$ from inverse gamma distribution
- Often impractical to include whole $\hat{\boldsymbol{\Theta}}$ in the BMA $\Rightarrow$ use smaller subset $\hat{\boldsymbol{\Theta}}^{\text{local}} \subset \hat{\boldsymbol{\Theta}}$
- Posterior inference in the MAP model by setting $\boldsymbol{\theta}^{[l]} \equiv \hat{\boldsymbol{\theta}}_{\text{MAP}}$.

University of Zurich

Automatic variable and transformation selection

- If $\boldsymbol{p}_i = \emptyset \Rightarrow x_i$ is not included in the model

$\Rightarrow$ Estimate posterior variable inclusion probabilities

$$\hat{\pi}_i = \hat{\mathbb{P}}(x_i \text{ included} \mid \mathcal{D}) = \sum_{\boldsymbol{\theta} \in \hat{\Theta} : \boldsymbol{p}_i \neq \emptyset} \hat{f}(\boldsymbol{\theta} \mid \mathcal{D})$$

- Given the inclusion of x_i, look at FP curve uncertainty (possibly averaged across models):
  1. At a grid of x_i-values and for all samples $l = 1, \dots, L$, calculate the ordinate

  $$f_i(x_i) = f(x_i; \boldsymbol{\alpha}_i^{[l]}, \boldsymbol{p}_i^{[l]}) = \boldsymbol{h}_i^{[l]}(x_i)^T \boldsymbol{\alpha}_i^{[l]}$$

  2. For each grid point x_i, we can compute mean/quantiles or (pointwise) highest posterior density intervals for $f_i(x_i)$.
  3. We can also compute simultaneous credible bands for the transformation f_i.

Application: Ozone data (Breiman and Friedman 1985)

Variable	Description	Missing	Shift γ_i	Scale ζ_i
y	Maximum 1-hour average ozone level [ppm]	5	-	-
z_0	Day of the year	-	0	100
z_4	500 millibar pressure height [m]	12	0	10 000
z_5	Wind speed [mph]	-	1	10
z_6	Relative humidity [%]	15	0	100
z_7	Temperature [°F]	2	0	100
z_8	Inversion base height [feet]	15	0	1 000
z_9	Pressure gradient [mm Hg] (LA to Daggett)	1	70	100
z_{10}	Visibility [miles]	-	2	100
z_{11}	Inversion base temperature [°F]	14	0	100

- Data from the 366 days of 1976 in the Los Angeles basin
- Preprocessing of covariates: $x_i := (z_i + \gamma_i)/\zeta_i$
- Exclude 30 observations as a test set $\Rightarrow n = 300$
- Set $m_{max} = 2$ and $a = 4$ for hyper-g prior

University of Zurich

Model and variable uncertainty results

- $1\,000\,000$ iterations, $|\hat{\boldsymbol{\Theta}}| = 907\,986$, top $3\,000$ saved $\Rightarrow \hat{\boldsymbol{\Theta}}^{\text{local}}$
- 11 minutes on a 2 GHz laptop using the R-package bfp

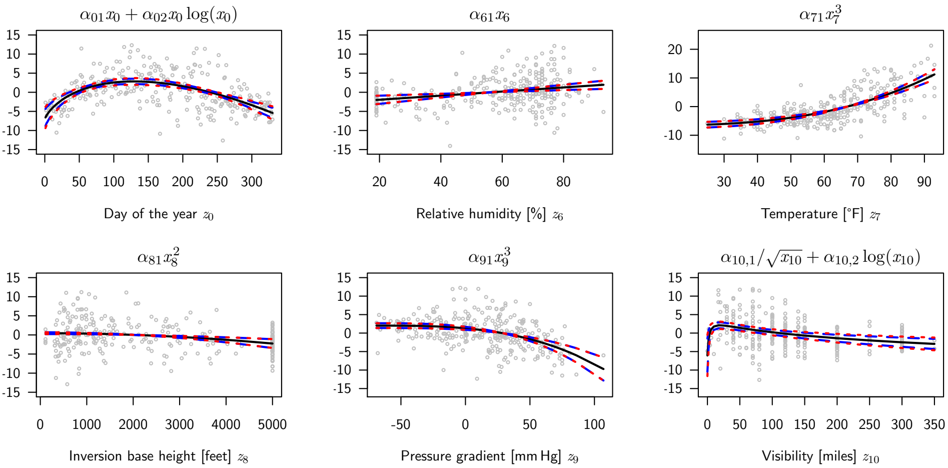
$\Rightarrow$ Posterior variable inclusion probabilities (global and local):

Variable:	x_0	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
$\hat{\pi}$:	1.0000	0.5758	0.2629	0.8447	0.9994	0.7039	1.0000	0.9991	0.0886
$\hat{\pi}^{\text{local}}$:	1.0000	0.3812	0.1692	0.8767	1.0000	0.7567	1.0000	1.0000	0.0595

$\Rightarrow$ Top ten models do not include x_4, x_5 and x_{11}:

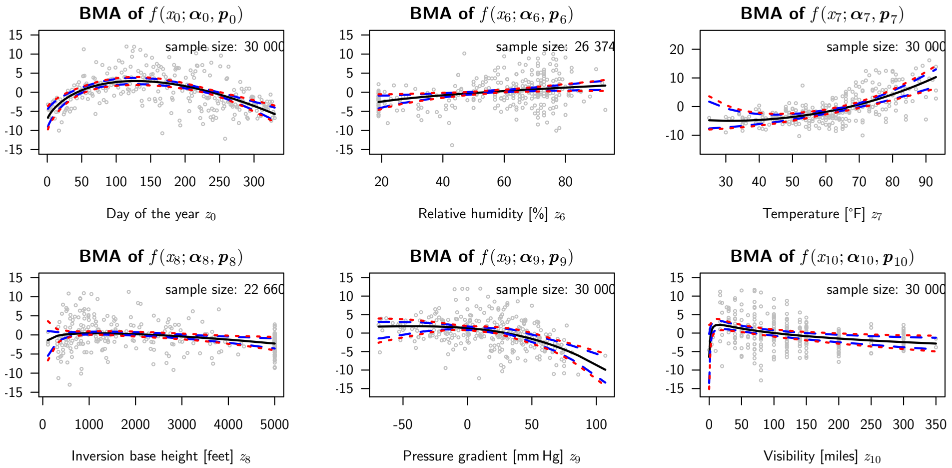
No.	$\hat{f}(\boldsymbol{\theta} \mid \mathcal{D})$ $\times 10^4$	$\hat{f}_{freq}(\boldsymbol{\theta} \mid \mathcal{D})$ $\times 10^4$	$\boldsymbol{p}_0$	$\boldsymbol{p}_6$	$\boldsymbol{p}_7$	$\boldsymbol{p}_8$	$\boldsymbol{p}_9$	$\boldsymbol{p}_{10}$
1	22.74	0.75	1, 1	1	3	2	3	-0.5, 0
2	21.68	0.33	1, 1	0.5	3	2	3	-0.5, 0
3	18.53	0.87	1, 1	1	3	2	3	-0.5, -0.5
4	18.10	0.69	1, 1	1	3	3	3	-0.5, 0
5	17.52	0.15	1, 1	0.5	3	3	3	-0.5, 0
6	16.19	0.13	1, 1	1	3	2	3	-1, 0
7	15.74	0.16	1, 1	0.5	3	2	3	-1, 0
8	15.44	0.43	1, 1	1	3	3	3	-0.5, -0.5
9	15.40	0.08	1, 1	0.5	3	3	3	-0.5, -0.5
10	15.31	0.10	1, 1	0	3	2	3	-0.5, 0

FP curves in the MAP model $\hat{\boldsymbol{\theta}}_{\text{MAP}}$



University of Zurich

Compare: FP curves in the BMA over $\hat{\Theta}^{\text{local}}$



University of Zurich

Outlook: Extension to Generalized Linear Models (GLMs)

- Consider a (non-normal) GLM with fixed dispersion ϕ.
- The linear predictor $\eta(\boldsymbol{x})$ is modelled by FPs.
- We stick to the g-prior on $\boldsymbol{\beta_\theta}$,
  and allow any hyper-prior $f(g)$.
- Problem: No $f(g)$ available with closed form for $f(\mathcal{D} \mid \boldsymbol{\theta})$!

$\Rightarrow$ Solution: Use (numerical) integrated Laplace approximation to get (fast) estimates $\hat{f}(\mathcal{D} \mid \boldsymbol{\theta})$ (Rue, Martino, and Chopin 2009)

- Tuning-free MCMC algorithm for parameter estimation after the model search

$\Rightarrow$ Estimates $\hat{f}_{\mathrm{MCMC}}(\mathcal{D} \mid \boldsymbol{\theta})$ (Chib and Jeliazkov 2001) as byproducts for checks of $\hat{f}(\mathcal{D} \mid \boldsymbol{\theta})$

University of Zurich

Discussion

- Details are in the paper (Sabanés Bové and Held 2010).
- Installing the R-package from R-Forge is easy:

```
install.packages("bfp", repos="http://R-Forge.R-project.org")
```

- Root mean squared prediction errors on the test were:
  MAP (3.512) $<$ BMA (3.571) $<$ `mfp` (3.579),
  but how good are the predictions in general?
- No sensitivity of results to choice of a
  in the recommended range $a \in (3, 4]$.
- $p_{\boldsymbol{\theta}} \leq n - 3 - a$ is required for proper posteriors.
- Rerun model search to check if number of iterations suffices.
- Other stochastic search algorithms may be useful.

University of Zurich

Thank you for your attention

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.

Chib, S. and I. Jeliazkov (2001).
Marginal likelihood from the Metropolis-Hastings output.
Journal of the American Statistical Association 96(453), 270–281.

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.

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.

Rue, H., S. Martino, and N. Chopin (2009).
Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.
Journal of the Royal Statistical Society. Series B (Methodological) 71(2), 319–392.

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

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.