Comparison of Multivariable Fractional Polynomials with Splines and Penalised Splines

Aris Perperoglou[1], Daniela Dunkler[2], Matthias Schmid[3] and Willi Sauerbrei[4] for STRATOS TG2

[1]Newcastle University, UK
[2]Medical University of Vienna, Austria
[3]Institute for Medicine Biometry, Informatics and Epidemiology, Bonn, Germany
[4]University of Freiburg, Germany

Outline

- Observational studies
- Variable selection with Fractional polynomials, Spline approaches and Penalised methods
- Application to PIMA & PBC data
- Simulations
- Discussion

TG2 Focus: Observational Studies – Regression models

- **Typical situation:** Several variables, mix of continuous and (ordered) categorical variables

- **Aim** of a study has strong influence on the analysis strategy
- Three conceptual modelling approaches:
  - Explanatory, descriptive, predictive
- Interest here: **descriptive model** (aims to capture the data structure parsimoniously)

- **Main issues:** (similar in different types of regression models)
- **Which variables to include? Which functional forms for continuous variables?**
- **Use subject-matter knowledge for modelling... but for some variables, data-driven choice inevitable**

Variable selection & choice of functional forms

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei, Aris Perperoglou, Matthias Schmid, Michal Abrahamowicz, Heiko Becher, Harald Binder, Daniela Dunkler, Frank E. Harrell Jr, Patrick Royston & Georg Heinze for TG2 of the STRATOS initiative

Diagnostic and Prognostic Research **4**, Article number: 3 (2020) | Cite this article

- Variable selection in the presence of non-linear relationships of covariates is an even more complicated exercise. In fact, decisions regarding the inclusion/exclusion of specific variables and modelling of the functional forms of both these variables and potential confounders may depend on each other in a complex way.

Do we need variable selection?

- *…guided by principles such as the need for interpretability, reproducibility and transportability, we prefer a simple model unless the data indicate the need for greater complexity. (Royston & Sauerbrei, 2008)*

- *(variable selection)… from a pragmatic point of view, aims at determining which covariates have the strongest effects on the response of interest, whereas from a statistical perspective it represents a means to achieve balance between goodness of fit and parsimony. By effectively identifying a subset of important covariates we can both enhance model interpretability and improve prediction accuracy. (Marra & Wood, 2012)*

Fractional polynomial models

- Describe for one covariate, X
- Fractional polynomial of degree m for X with powers $p_1, \ldots, p_m$ is given by

$$FPm(X) = \beta_1 X^{p1} + \ldots + \beta_m X^{pm}$$

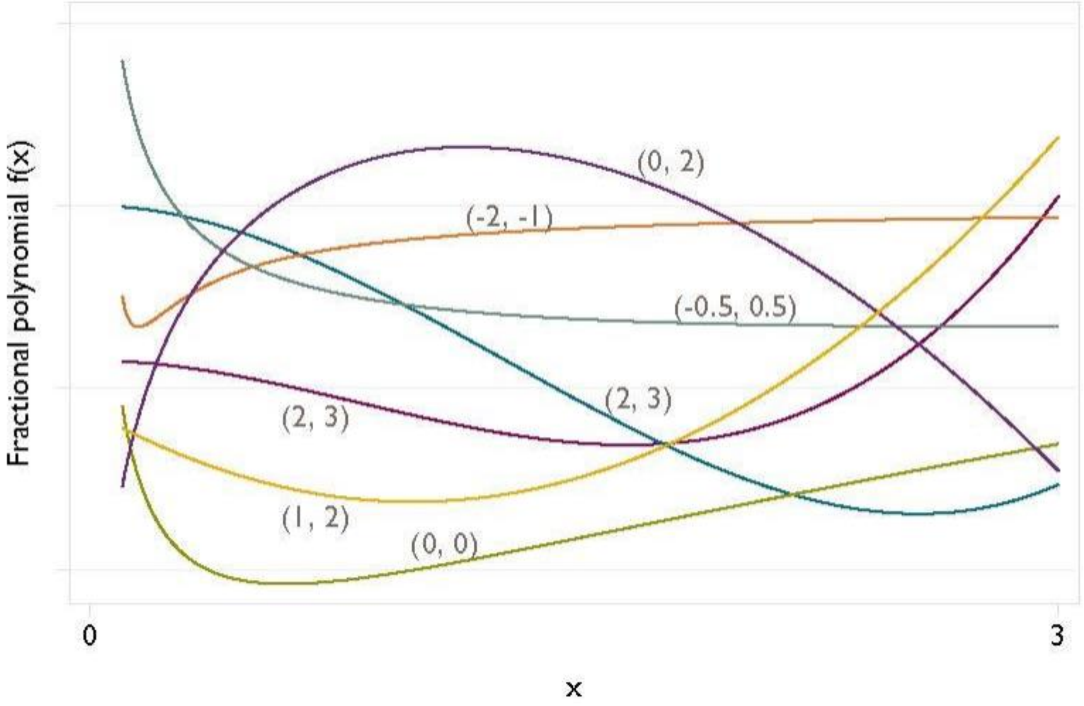
- Powers $p_1, \ldots, p_m$ are taken from a special set

$$\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$

- Usually $m = 1$ or $m = 2$ is sufficient for a good fit
- Repeated powers ($p_1 = p_2$)

$$\beta_1 X^{p1} + \beta_2 X^{p1} \log X$$

- 8 FP1, 36 FP2 models



Function Selection Procedure and Multivariable FP

FSP

- Define most complex function allowed, common choice FP2; deviance difference as the criteria; determine significance level α_1

	df	p-value
Any effect?		
Best FP2 versus null	4	
Linear function suitable?		
Best FP2 versus linear	3	
FP1 sufficient?		
Best FP2 vs. best FP1	2	

- Combine backward elimination of weak variables with search for best FP functions
- Determine fitting order from full linear model
- Apply FSP selection procedure to each X in turn, fixing functions (but not βs) for other X's
- Cycle until FP functions (i.e. powers) and variables selected do not change
- Significance level may be different for the two parts – selection of variables (α_2) and selection of variable forms (α_1)

Splines are also simple polynomials

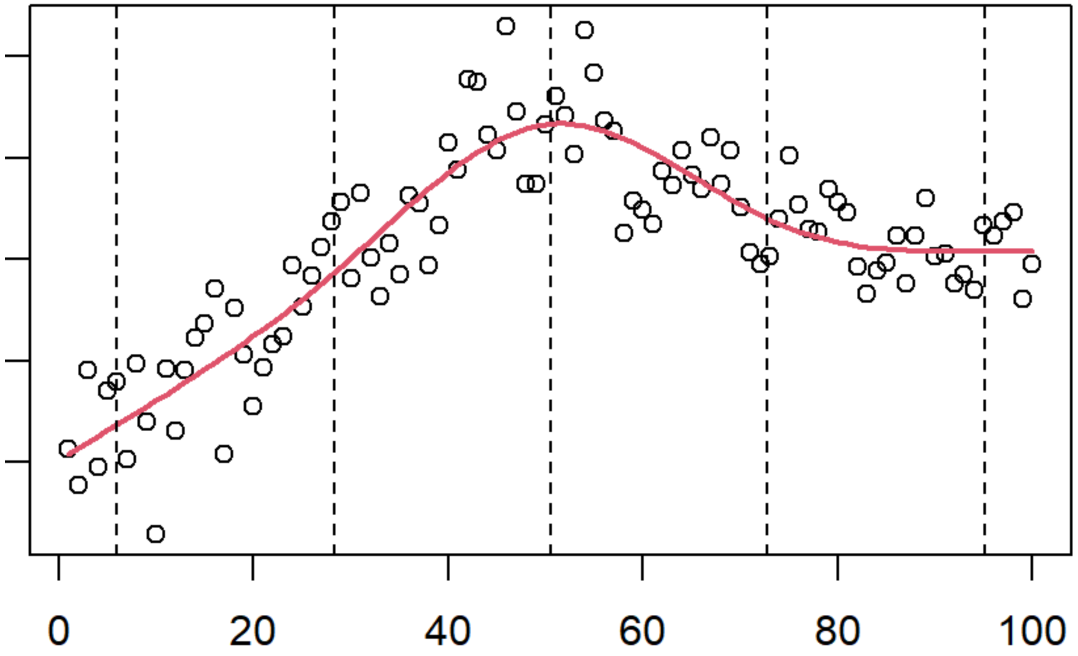
- Set of piecewise polynomials, each of degree d
- Joined together at a set of knots $\tau_1, \ldots, \tau_K$
- Continuous in value and sufficiently smooth at the knots

A restricted cubic regression spline is defined by:

being a cubic function between the set of fixed knots $\tau_1, \ldots, \tau_K$

being a linear function for $x < \tau_1$ and $x > \tau_K$

being continuous with continuous first and second derivative



Natural Splines are restricted cubic splines with cubic b-splines as functions between knots

A review of spline function procedures in R

Aris Perperoglou, Willi Sauerbrei, Michal Abrahamowicz & Matthias Schmid

BMC Medical Research Methodology **19**, Article number: 46 (2019) | Cite this article

Restricted Cubic Splines and Multivariable Regression Splines (MVRS)

SSP

- Determine the most complex model in terms of knots "df(m)"; m often depends on sample size; knots are chosen at predetermined percentiles of distribution of x; deviance difference as criteria; determine significance level α_1

	Df	p-value
Any effect? Best df(m) versus null	m+1	
Linear function suitable? Best df(m) versus linear	m	
df(m) needed? Best df(m) vs. df(1)	m-1	
….	…	

- Predictors are considered in decreasing order of significance in a full linear model
- The algorithm cycles over the predictors, updating the model
- Procedure terminates when no further variables included in the model and df for splines are chosen for continuous variables
- Royston, Sauerbrei suggested df(m=4,8)
- Procedure can be easily adapted to other spline bases, eg b-splines, natural splines
- MVSS also suggested for cubic smoothing splines (based on edf)

Generalised additive models and mgcv

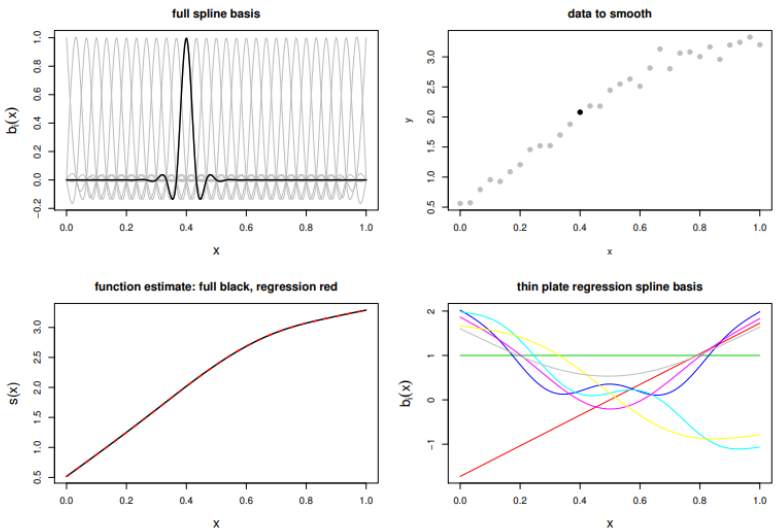
A generalised additive model GAM (Hastie and Tibshirani 1990) connects a response Y_i to linear components and smooth functions:

$$g\{E(Y_i)\} = X_i\theta + \sum_j f_j(x_{\{ij\}})$$

Where g(.) a prespecified link functions, X_i a linear component of the model and f_j some smooth functions.

Example: eigen based spline "tp"

- The "tp", *thin plate regression spline* basis is an eigen approximation to a thin plate spline (including cubic spline in 1 dimension).



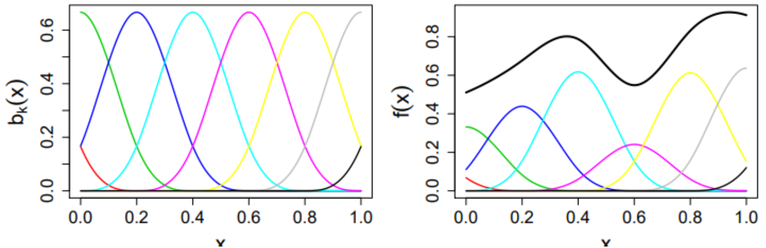
Example: P-splines "ps"

- Eilers and Marx have popularized the use of B-spline bases with discrete penalties.
  - If $b_k(x)$ is a B-spline and β_k an unknown coefficient, then

$$f(x) = \sum_k^K \beta_k b_k(x).$$

  - Wiggliness can be penalized by e.g.

$$\mathcal{P} = \sum_{k=2}^{K-1} (\beta_{j-1} - 2\beta_j + \beta_{j+1})^2 = \beta^{\mathsf{T}}\mathbf{S}\beta.$$



Practical Variable Selection for GAMs

Penalised maximum likelihood estimation can be used to control overfit.
In practice a GAM is fitted by iterative minimisation of:

$$\left\|\sqrt{W^{[k]}}\left(z^{[k]} - X\beta\right)\right\|^2 + \sum_j \lambda_j \beta^T S_j \beta, wrt\ \beta$$

Large values of λ_j will control smooth term but will not force it out of the model.

(Marra & Wood, Comp Stat & Data Analysis 2011)

Double Penalty

$$\lambda_j \beta^T S_j \beta + \lambda_j^* \beta^T S_j^* \beta$$

Any spline type smoother can be decomposed into two component functions: a component in the **range space** of the penalty (**λ**) and a component in the **null space** of penalty (**λ^***).

As an example, when using a cubic spline penalty large λ values would force spline towards a linear form and λ^* would penalise straight line components to zero.

Shrinkage approach

Replace smoothing penalty matrix S_j with

$$\tilde{S}_j = U_j \widetilde{\Lambda_j} U_j^T$$

where U_j is an eigenvector matrix associated with j smooth function and $\widetilde{\Lambda_j}$ a corresponding diagonal eigenvalue matrix except for the zero eigenvalues replaced by ε, a small proportion of the smallest strictly positive eigenvalues of S.

This forces eigenvalues of $\tilde{S}_j$ associated with the penalty null space to be different from zero.

Datasets

Prediction of diabetes onset

- Dataset from an investigation of potential predictors for the onset of diabetes in a cohort of 768 female Pima Indians, of whom 268 developed diabetes.
- **Response:** binary outcome diabetes (0/1)
- **Continuous Predictors:** number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin, diabetes pedigree function, bmi and age
- Substantial missing values imputed once by ice in STATA

Set available in http://biom131.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/#datasets

Survival of PBC patients

- Mayo Clinic trial in PBC conducted between 1974 and 1984. A total of 312 PBC patients randomized in a placebo controlled trial of the drug D-penicillamine.
- **Response**: Survival time, 125 deaths
- **Continuous Predictors:** age, serum albumin, serum bilirunbin, serum cholesterol, urine copper, triglycerides
- **Categorical/Ordinal:** presence of ascites, spiders (malformations of the skin), edema (no, untreated or treated) histological stage of disease

Set available in R

Models

	MFP	MVRS	NS	TS1	TS2	PS
function	Fractional polynomials	Natural splines	Natural splines	Thin plate regression splines	Thin plate regression splines	P-splines
maximum df	4 df (2FPs)	5	9	9	9	9
variable selection	BE + FSP	BE + SSP	shrinkage	shrinkage	double penalty	double penalty
R library	mfp	script	mgcv	mgcv	mgcv	mgcv

Results extract (PIMA data)

mfp

```
mfp(formula = Outcome ~ fp(Pregnancies, df
= 4) + …+ fp(Age, df = 4), family =
"binomial",     select = 0.01)
```

	df.init	slct	alpha	df.final	pw1	pw2
Glucose	4	0.01	0.05	1	1	.
BMI	4	0.01	0.05	2	-2	.
Pregn	4	0.01	0.05	0	.	.
Diab	4	0.01	0.05	1	1	.
Age	4	0.01	0.05	4	0	3
Blood	4	0.01	0.05	0	.	.
Skin	4	0.01	0.05	0	.	.
Insuln	4	0.01	0.05	0	.	.

mgcv

```
gam(Outcome ~ s(Pregnancies,bs = 'tp') +
s(Age,bs = 'tp'), family = "binomial",
select= TRUE, method="REML")
```

Approximate significance of smooth terms:

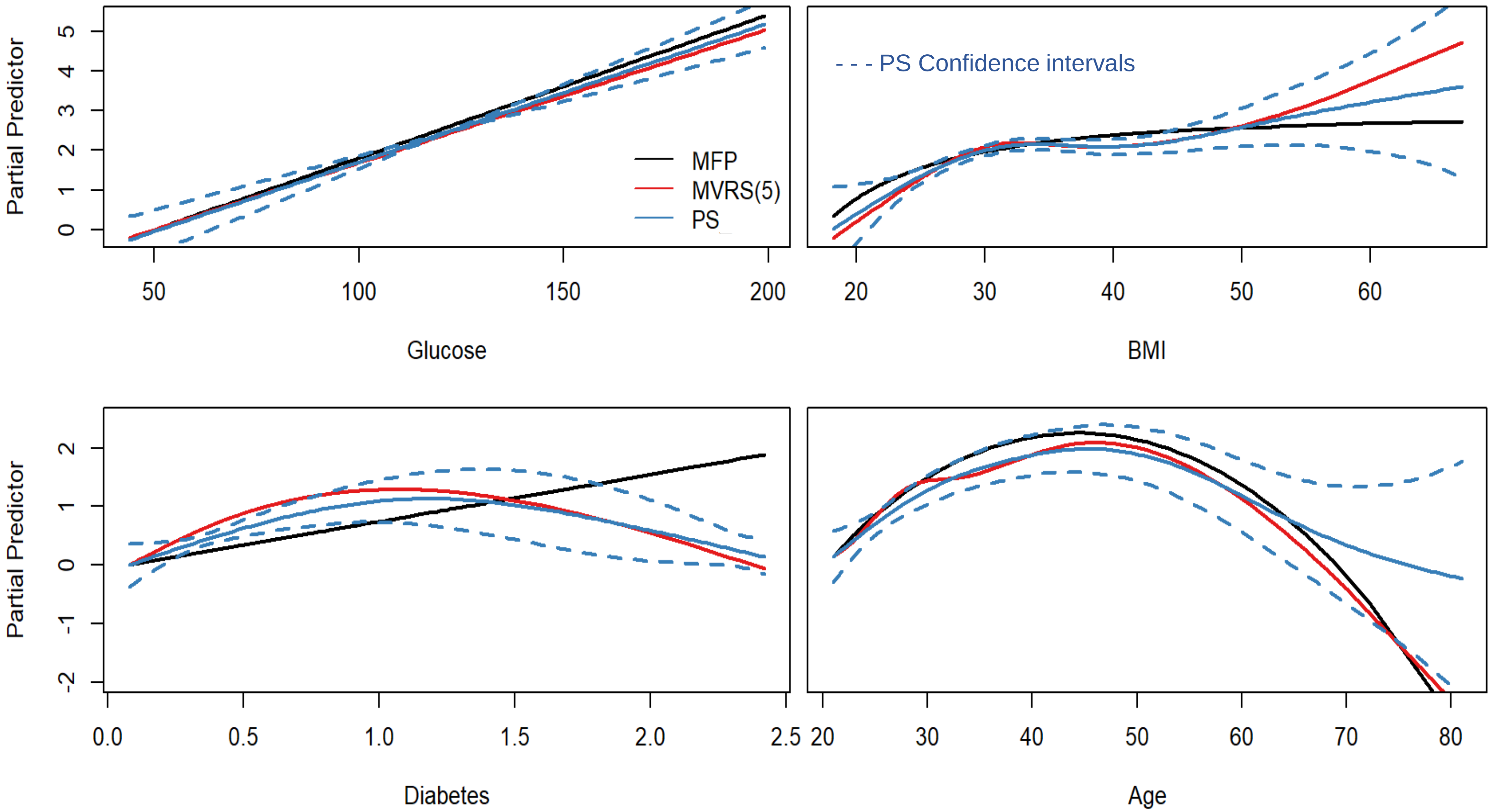
	edf	Ref.df	Chi.sq	p-value
s(Glucose)	0.989	9	89.347	< 2e-16
s(BMI)	3.665	9	30.695	< 2e-16
s(Pregnancies)	1.106	9	2.903	0.06618
s(DiabetesPed)	1.677	9	9.814	0.00183
s(Age)	3.098	9	28.168	< 2e-16
s(BloodPressure)	0.000	9	0.000	0.42658
s(SkinThickness)	0.000	9	0.000	0.99424
s(Insulin)	0.099	9	0.108	0.30852

Variables included

All approaches seem to agree on variable inclusion bar MVRS that also included pregnancies .

Variables	MFP(2)	MVRS(5)	TS_1	TS_2	PS_2	NS
	power	df	edf	edf	edf	edf
Glucose	lin	1	1.3	1.0	1.0	2.1
BMI	-2	5	3.7	3.9	3.7	3.7
Pregnancies	-	1	0.6	0.6	0.5	0.6
Diabetes	lin	2	0.9	1.8	1.4	1.6
Age	-2	5	3.0	2.9	2.7	3.0
Systolic	-	-	0.0	0.1	0.1	0.1
Biceps	-	-	0.0	0.0	0.0	0.0
Insulin	-	-	0.0	0.0	0.5	0.0

Functional Forms

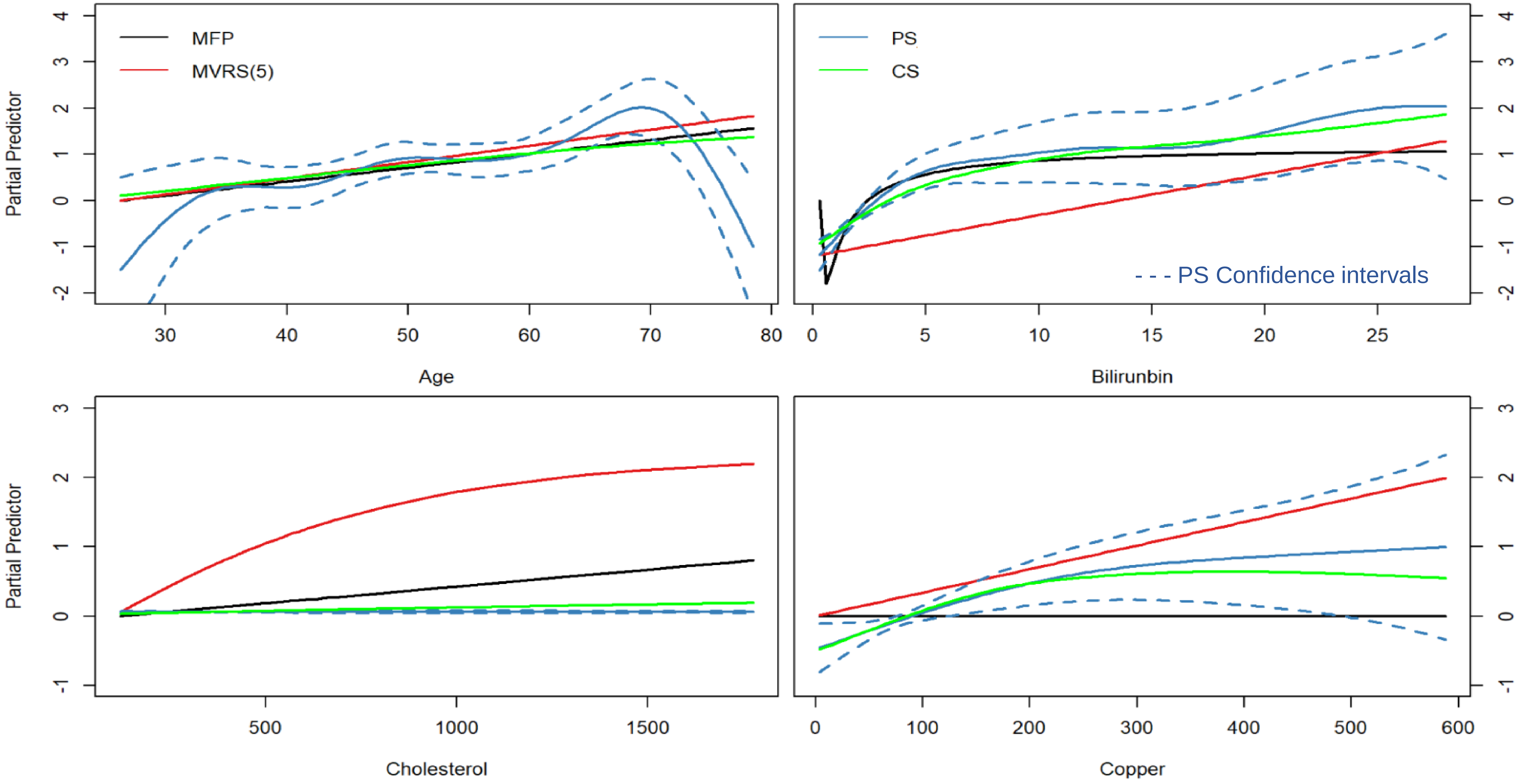


PBC data

Variables	MFP(2)	MVRS(5)	TS_1	TS_2	PS_2	NS
	power	df	edf	edf	edf	edf
age	lin	1	5.8	5.7	4.9	1.1
bili	-2, -1	1	3.9	4.6	3.8	2.7
chol	1	2	0.0	0.0	0.0	0.2
albumin	-	-	0.9	0.9	0.8	1.4
copper	-	1	0.9	1.4	1.6	1.7
trig	-	1	0.8	0.8	0.8	0.6
asc	in	in	in	in	in	-
spiders	-	in	-	-	-	-
edema	in	in	-	-	-	in
stage	in	in	in	in	in	in

Methods disagree on inclusion

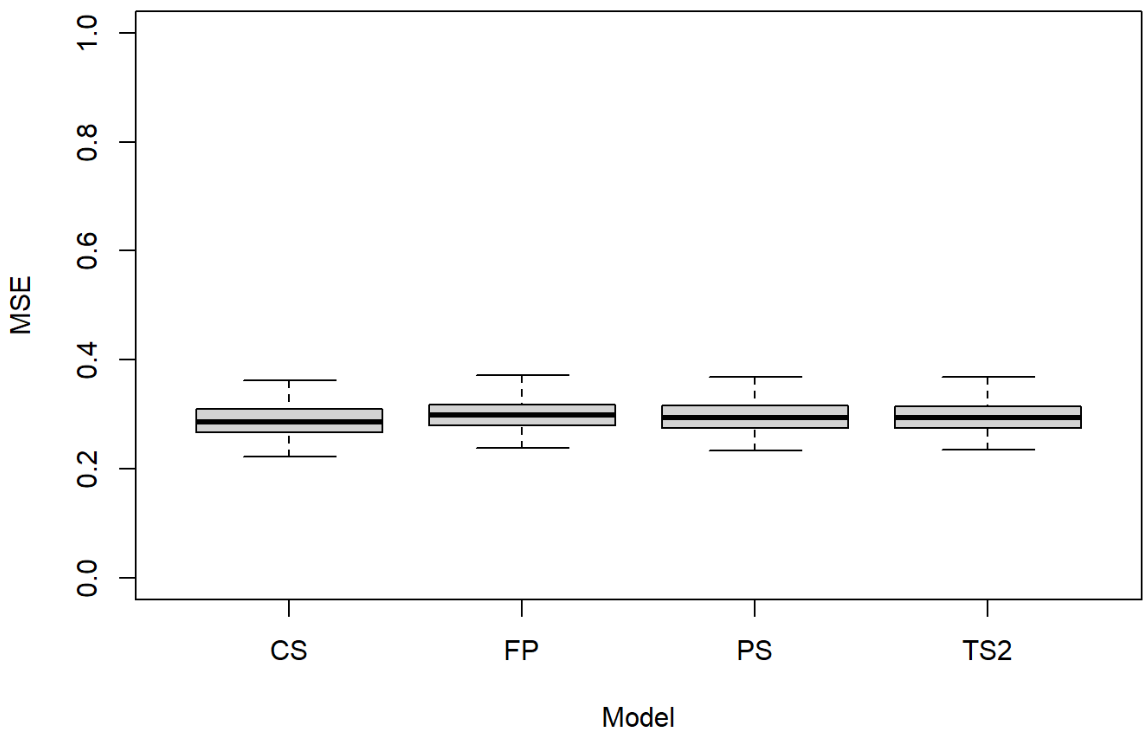
Functional forms



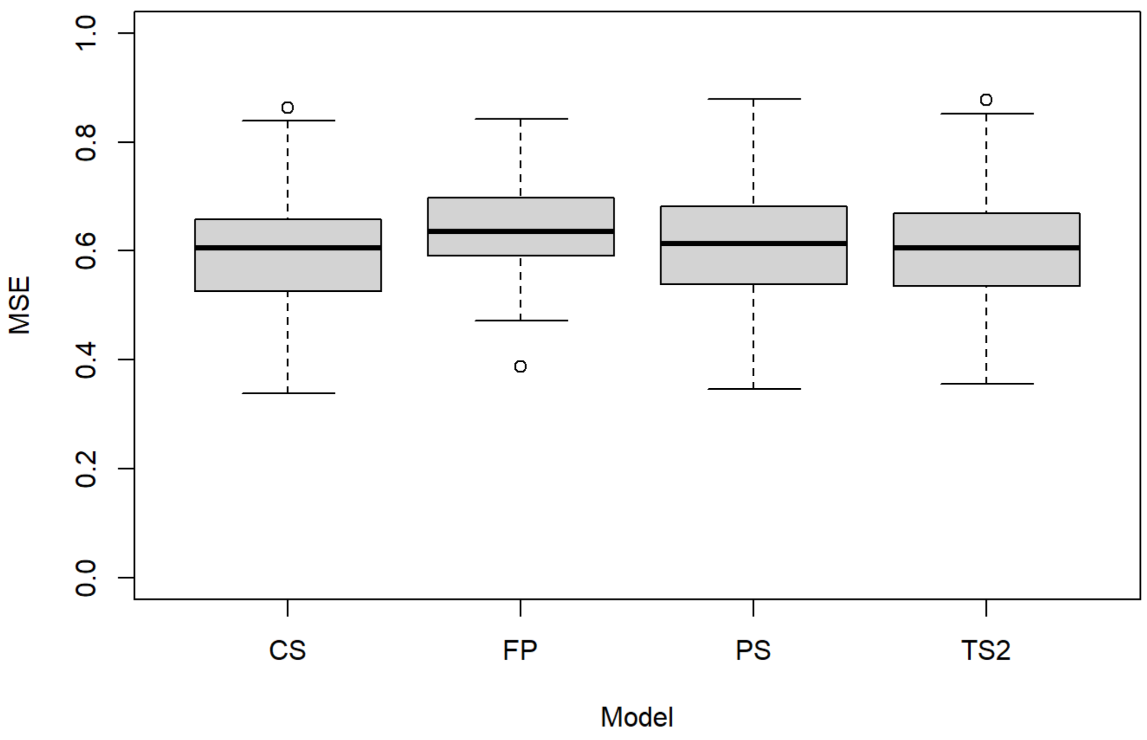
Prediction error

- 100 bootstrap samples for each dataset, leave 10% out for each sample.

PIMA data



PBC data



Simulation

200 iterations of n normal responses

- n = 400, n=1200

8 continuous covariates

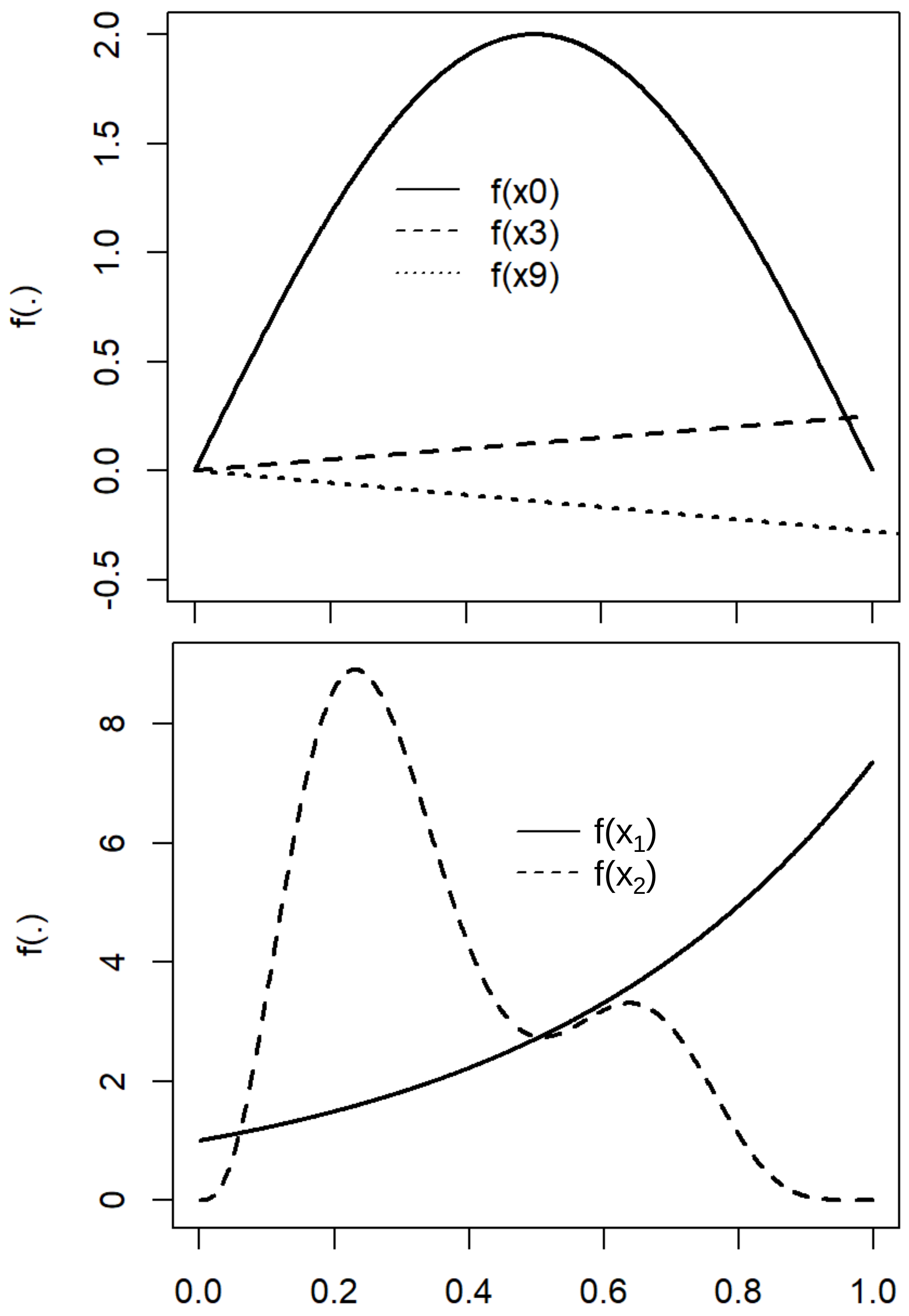
- 5 known functions (right) and 3 spurious (x_4-x_6)

2 binary covariates

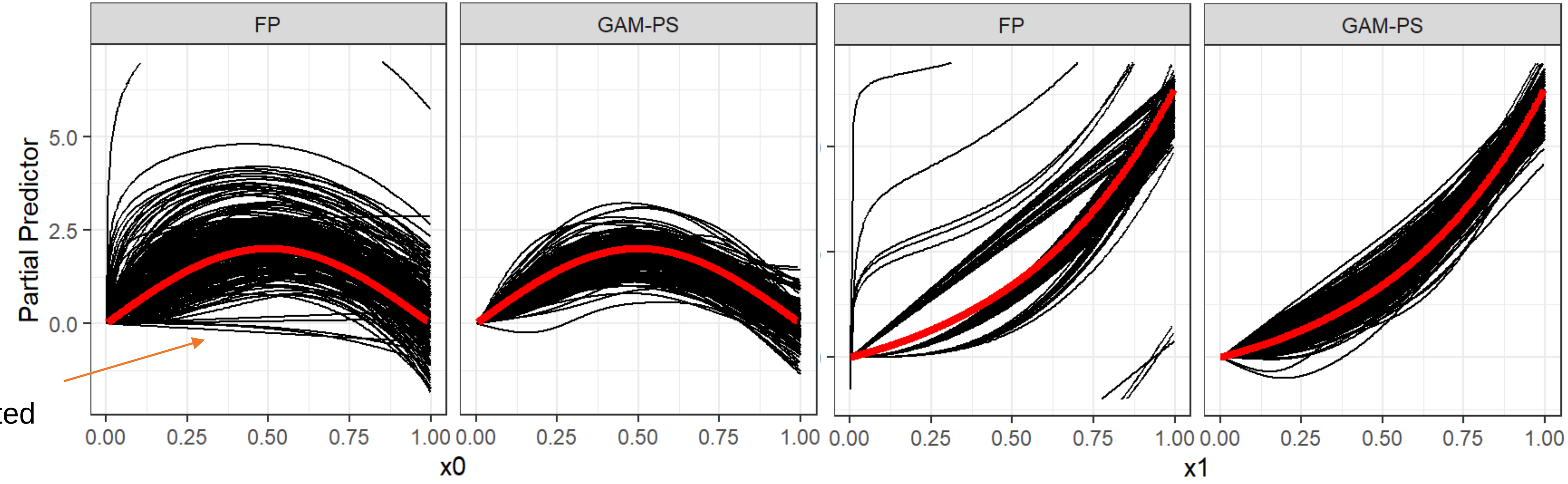
- 1 spurious (x_7), 1 related to outcome (0.72*x8)

$$y= f(x_0) + f(x_1) + f(x_2) + f(x_3) + 0.72* x8 + f(x_9) + \varepsilon$$

Very limited setting similar to Gu and Wahba
(four univariate term example, from function gamSim in mgcv)
More interesting simulations to follow, with correlated variables, and more features.

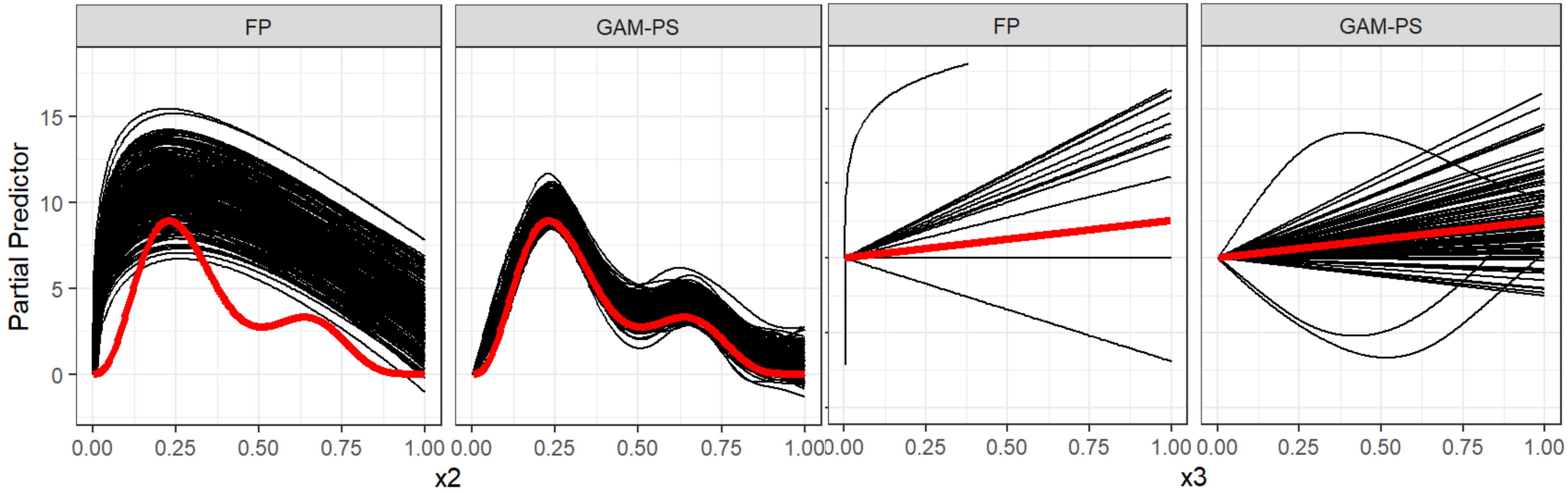


n = 400



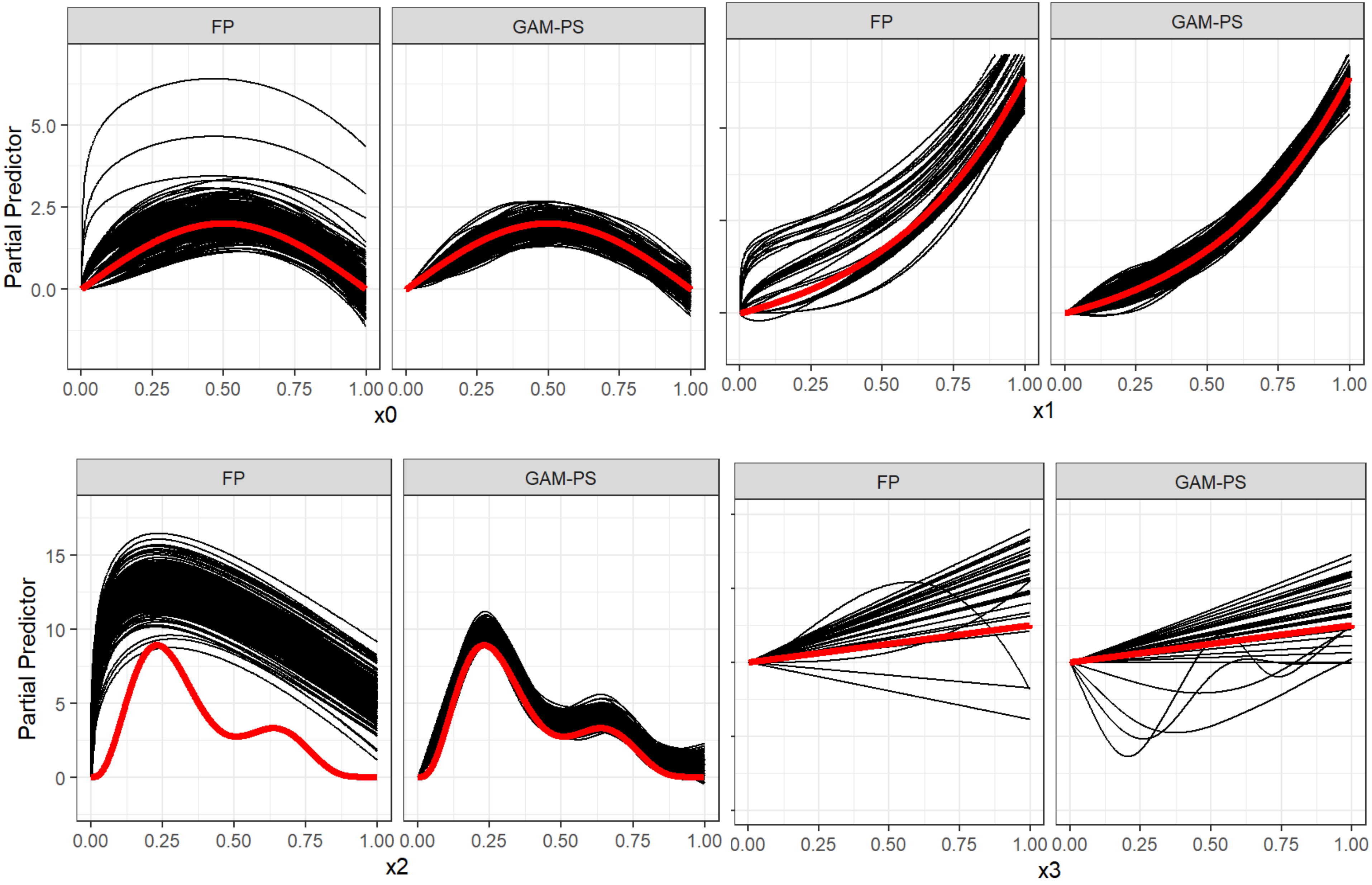
Low power and BIC:
linear function selected

X	FP	PS	TS	TP	CS
x0	200	200	200	200	200
x1	200	200	200	200	200
x2	200	200	200	200	200
x3	18	30	7	20	16
x4	11	25	3	17	16
x5	18	34	4	23	12
x6	19	27	3	16	8
x7	29	33	33	32	33
x8	178	193	192	192	192
x9	140	139	130	143	132



n=1200

	FP	PS	TS1	TS2	CS
x0	200	200	200	200	200
x1	200	200	200	200	200
x2	200	200	200	200	200
x3	42	47	19	40	47
x4	27	29	2	15	12
x5	23	29	6	14	12
x6	13	25	3	11	6
x7	32	31	33	34	33
x8	200	200	200	200	200
x9	199	199	199	199	199



Discussion

- **Choice of parameters can alter effects (significance levels, AIC/BIC for MFP, maximum df for splines, choice of penalty, knots, etc). All results here produced at software default.**
- In agreement with Royston & Sauerbrei (2008), MFP and spline approaches provide roughly comparable models.
- Approaches where closer in logistic regression setting with a fair sample size of 768 observations. Differences were more obvious in smaller sample size (survival model).
- MSE from all models showed little difference between approaches. However, main interest here is in models for description.
- In simulated data, where more flexibility is required, FP(2) may not be enough. Equally, penalised splines will not always correctly identify a linear relationship.
- Penalised approaches (double penalty) can be computationally expensive but can still handle moderate sample sizes.
- **Limitation: simple simulation setting, small number of non-correlated variables.**

References

- Sauerbrei, Willi, et al. "**State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues.**" *Diagnostic and prognostic research* 4.1 (2020): 1-18.
- Royston, Patrick, and Willi Sauerbrei. ***Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*****.** John Wiley & Sons, 2008.
- Marra, Giampiero, and Simon N. Wood. **"Practical variable selection for generalized additive models."** *Computational Statistics & Data Analysis* 55.7 (2011): 2372-2387.
- Perperoglou, Aris, et al. **"A review of spline function procedures in R."** *BMC medical research methodology* 19.1 (2019): 1-16.
- Gu, Chong, and Grace Wahba. "**Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method**." SIAM Journal on Scientific and Statistical Computing (1991): 383-398.