A comparison of spline methods in R for building explanatory models

Aris Perperoglou on behalf of TG2
STRATOS Initiative, University of Essex

ISCB2017

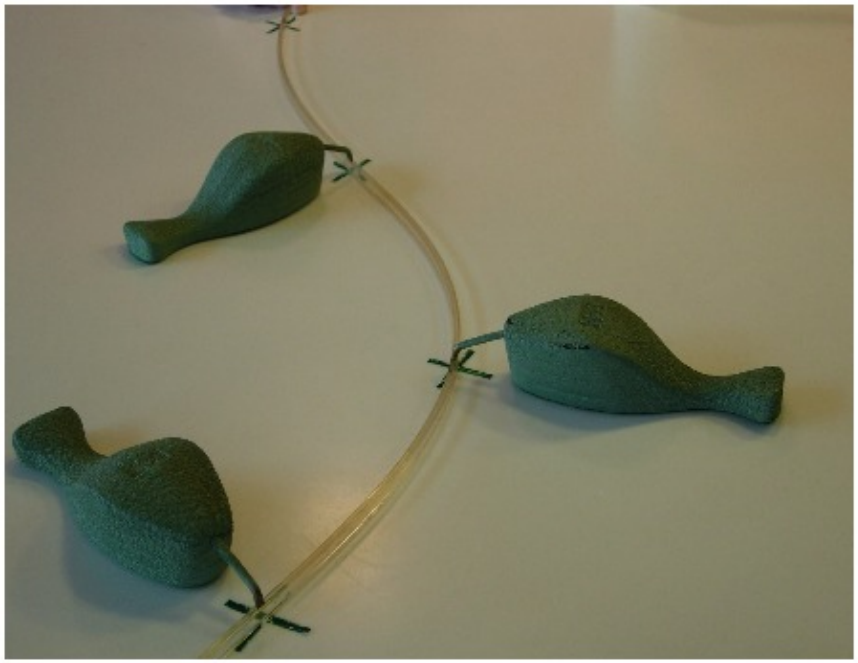
Introduction

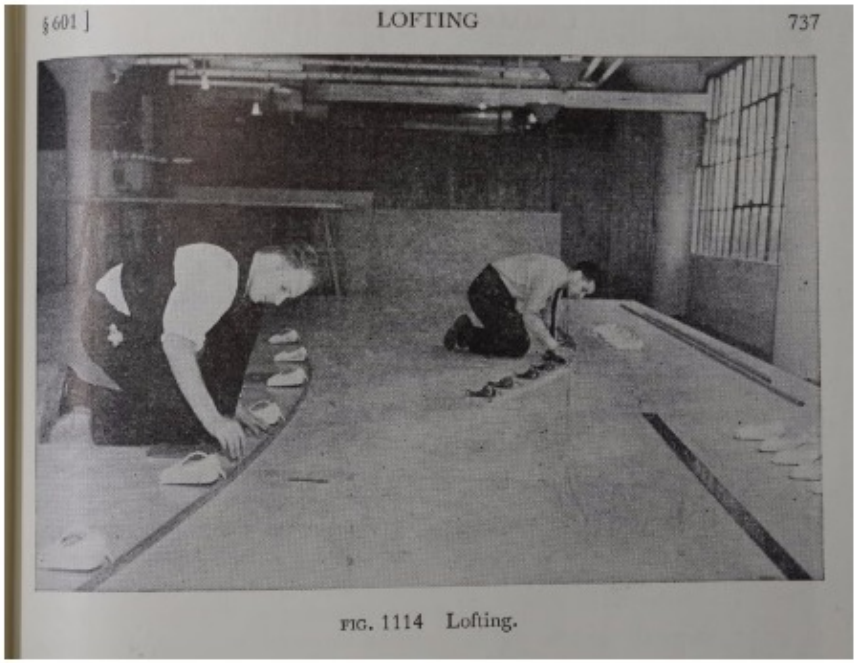
Topic Group 2: Selection of variables and functional forms in multivariable analysis

. . . main focus of TG2 is to identify influential variables and gain insight into their individual and joint relationship with the outcome. Two main challenges are selection of variables for inclusion in a multivariable explanatory model, and choice of the functional forms for continuous variables (Harrell 2001, Sauerbrei et al. 2007).

- identify and assess methods currently used in practice
- compile and publish guidance on selection of variables and their functional forms

Splines





§601] LOFTING 737

FIG. 1114 Lofting.

Splines

- Splines are piecewise polynomial functions
- Given the range of a continuous variables, define points on this interval (knots)
- Fit a simple polynomial between these knots.
- Depending on the selection of knots and the type of polynomial the type of spline is named as:
  - polynomial, natural, restricted regression spline (**de Boor 1978, Harrel 2013**)
  - truncated power basis, b-spline (**de Boor 1978**), p-spline (**Marx and Eilers 1996**)
  - smoothing splines (see Generalized Additive Models (**Hastie and Tibshirani 1990**)), penalized regression spline (**Wood 2006**)

The need

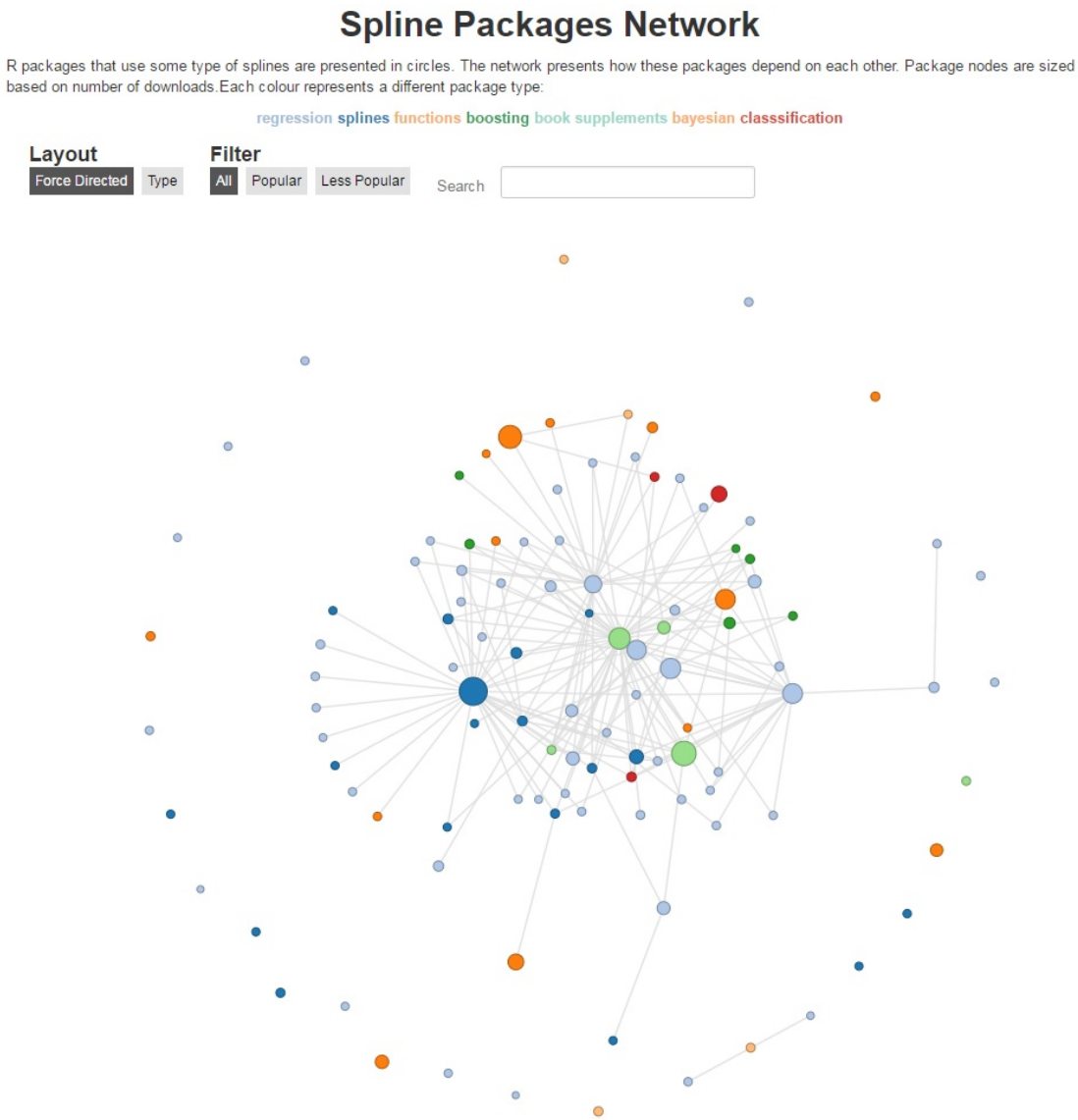
- Although mathematical properties are well understood the use of splines can be challenging for applied statisticians.
- Lack of statistical education. Most researchers are not *taught* how to use splines.
- Lack of thorough comparisons between different approaches.
- We aim to develop systematic guidance for using splines in applications focusing on level 2 expertise:
  - . . . we point to **methodology** which is perhaps slightly below state of the art, but **doable by every experienced analyst**. We should refer to **advantages and disadvantages of competing approaches**, point to the **importance and implications of underlying assumptions**, and stress the necessity of sensitivity analyses. . . question. **Sufficient guidance about software plays a key role that this approach is also used in practise**.

The splines project:

- Identify R packages that provide the ability to use splines as a functional transformation in univariable and/or multivariable analysis.
- Collect information on packages.
- Compile documentation for practical use.
- Evaluate quality of available packages.
- Compare approaches, further guidance for use in practice.

Challenges

- ISCB2015: More than 6200 packages available on CRAN.
- Banf2016: CRAN package repository features 8670 available packages.
- ISCB2017: More than 10980 packages available on CRAN.


Spline Packages Network
R packages that use some type of splines are presented in circles. The network presents how these packages depend on each other. Package nodes are sized based on number of downloads.Each colour represents a different package type:
regression splines functions boosting book supplements bayesian classsification
Layout
Force Directed
Type
Filter
All
Popular
Less Popular
Search

Types of splines in R

The splines Package

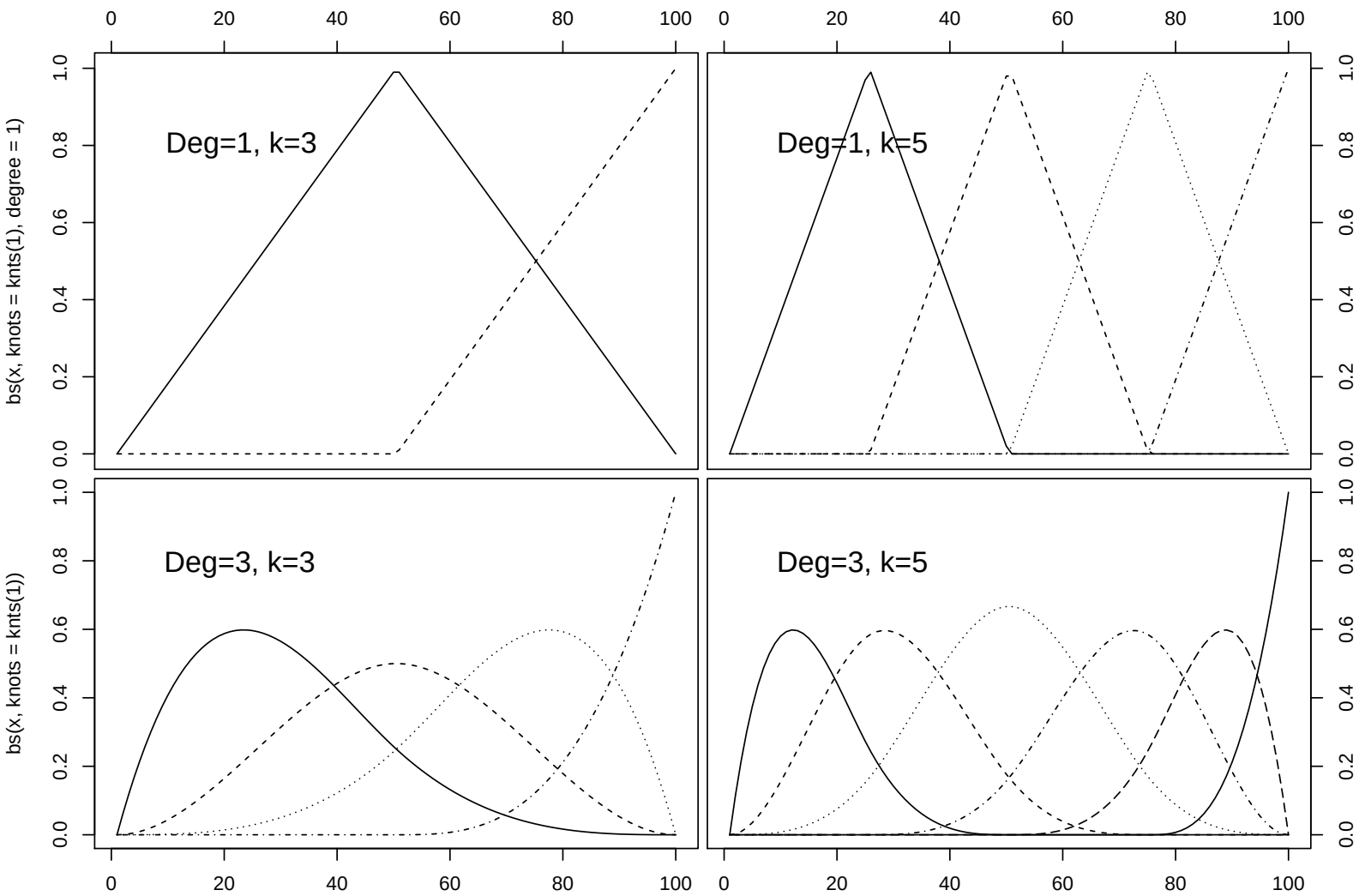
- bs: B-Spline Basis for Polynomial Splines

```
bs(x,df=NULL,knots=NULL,degree=3,Boundary.knots=range(x))
```

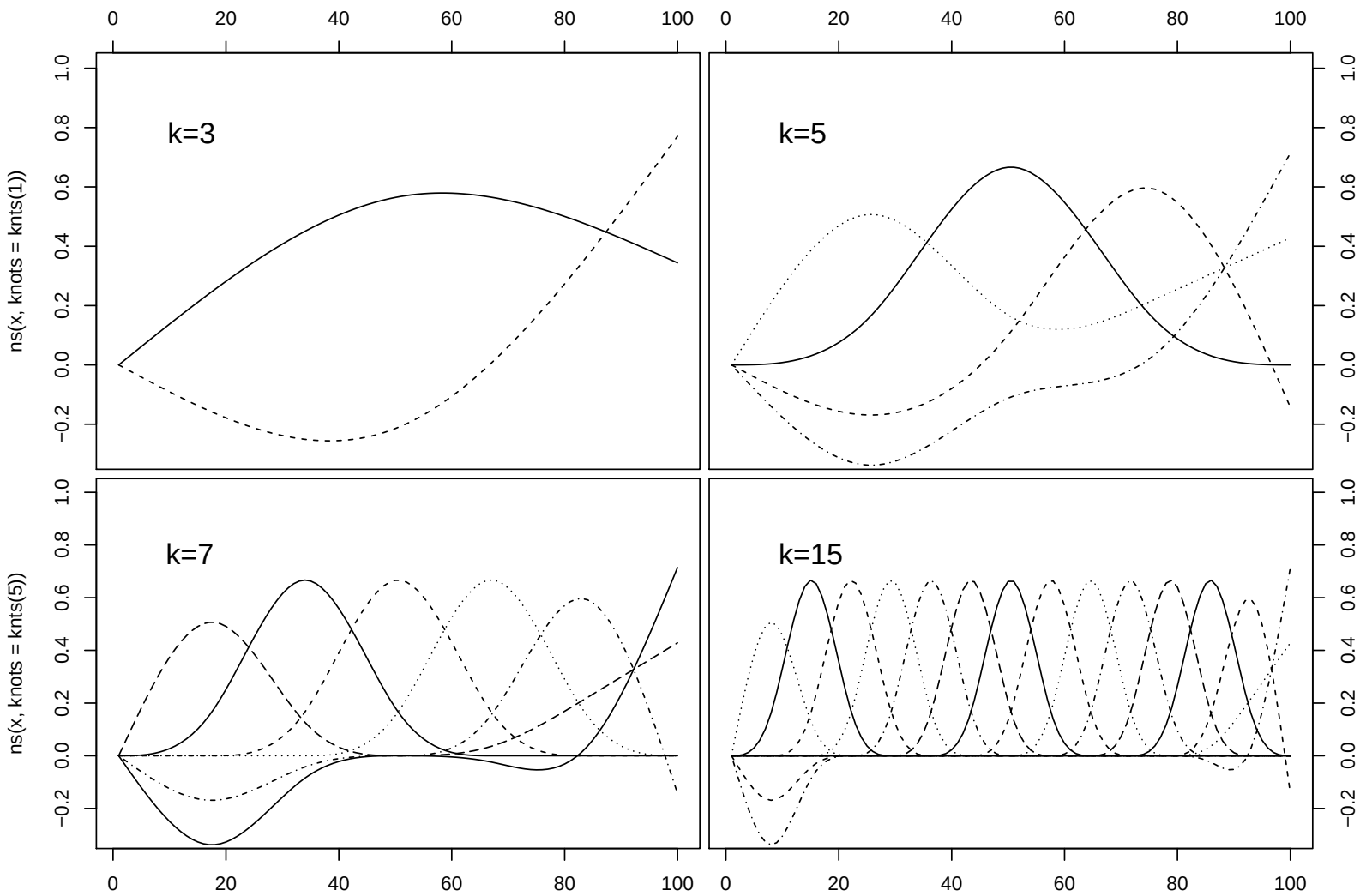
- ns: Generate a Basis Matrix for Natural Cubic Splines

```
ns(x,df=NULL,knots=NULL,Boundary.knots=range(x))
```

B-splines basis



Natural splines

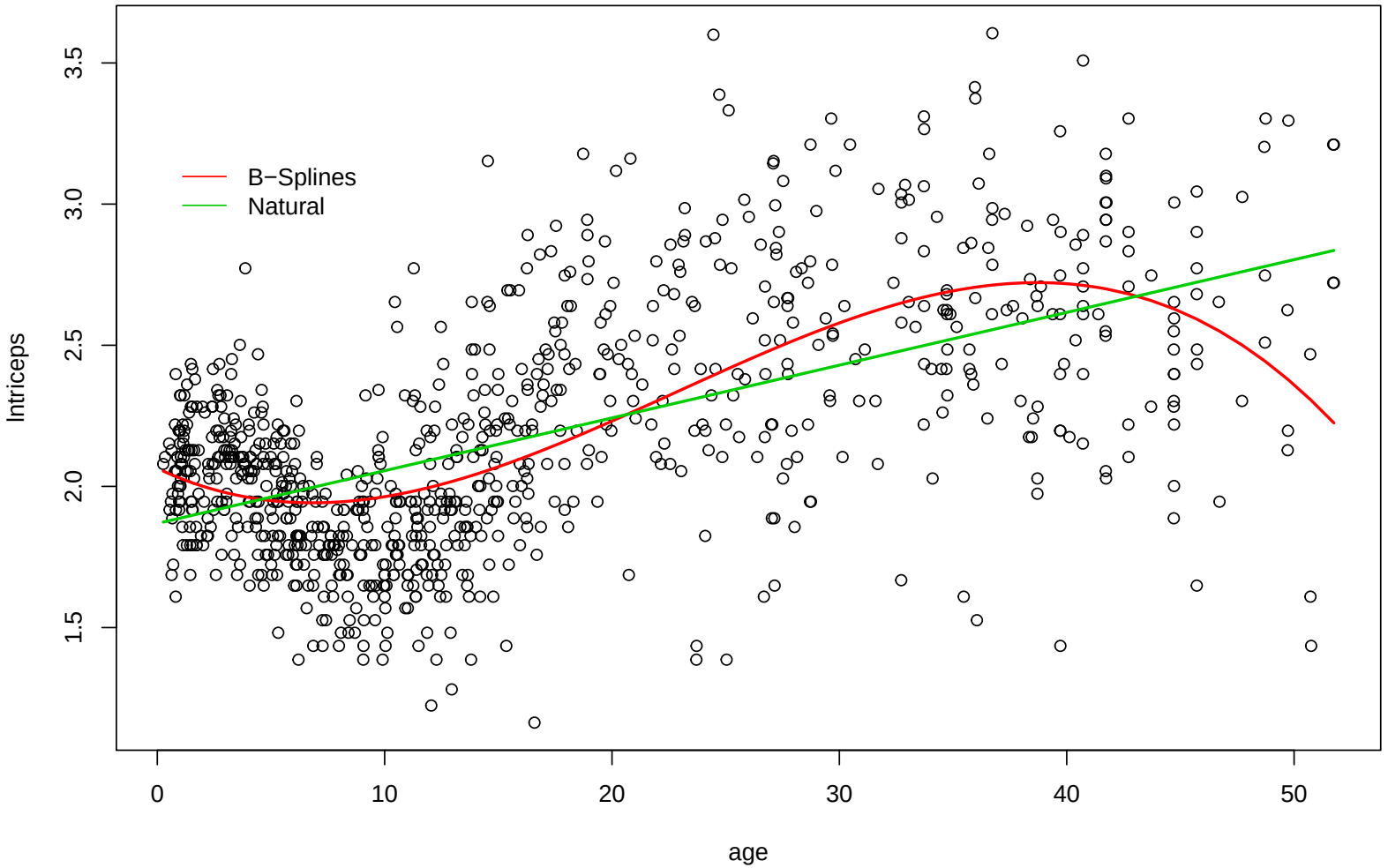


Scatterplot smoothing: Tricepts skinfold thikness by age

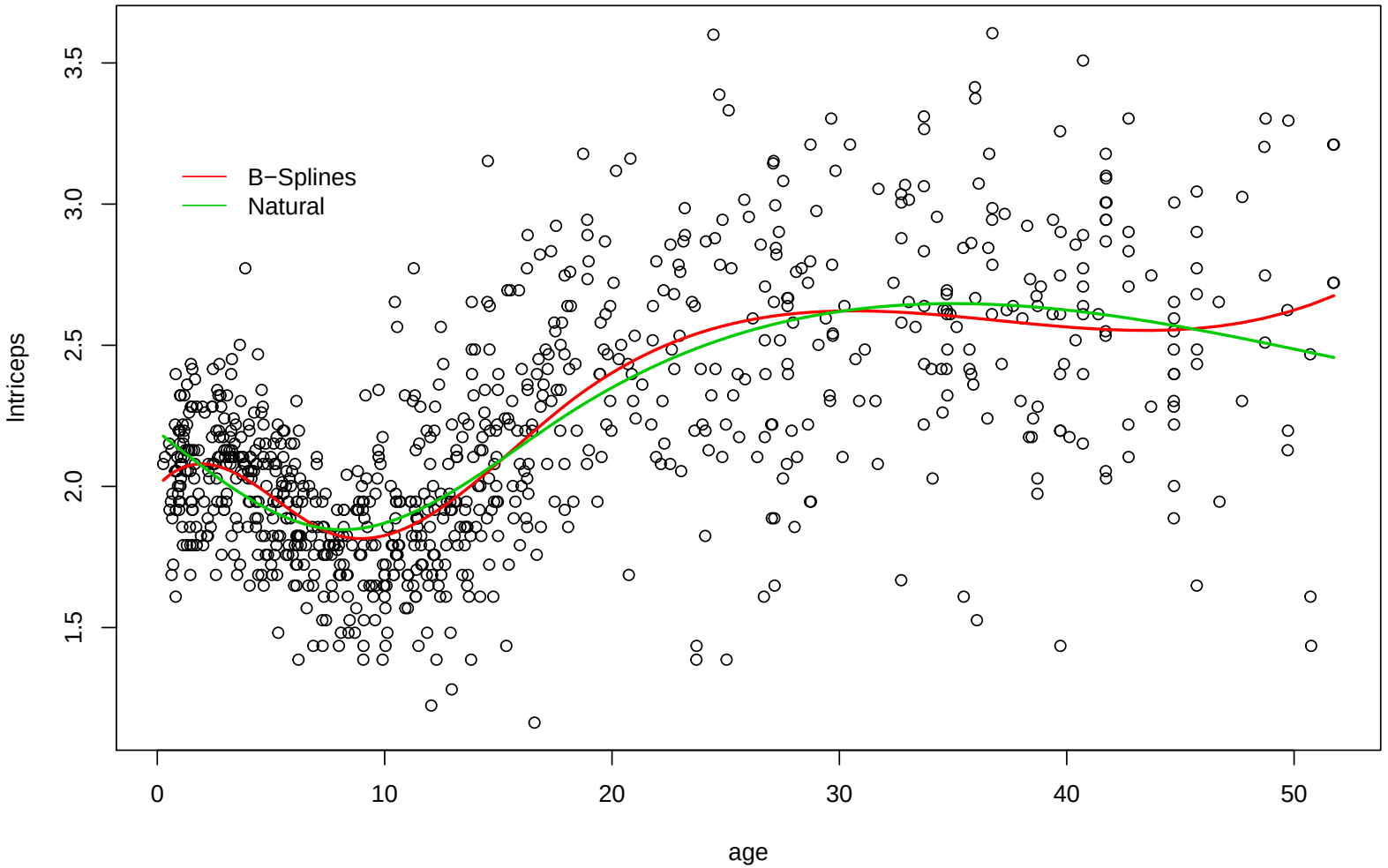
- 892 females under 50 years in three villages in West Africa
- Figure 1 shows the relationship between age and triceps skinfold thickness measured in log scale.
- For more information about the data see (Cole and Green 1992) and (Royston and Sauerbrei 2008).

```
plot(x,y,ylab="lntriceps",xlab="age")
fit.bs = lm(y ~ bs(x))
fit.ns = lm(y~ns(x))
lines(x, predict(fit.bs, data.frame(x=x)), col=2,lwd=2)
lines(x, predict(fit.ns, data.frame(x=x)), col=3,lwd=2)
```

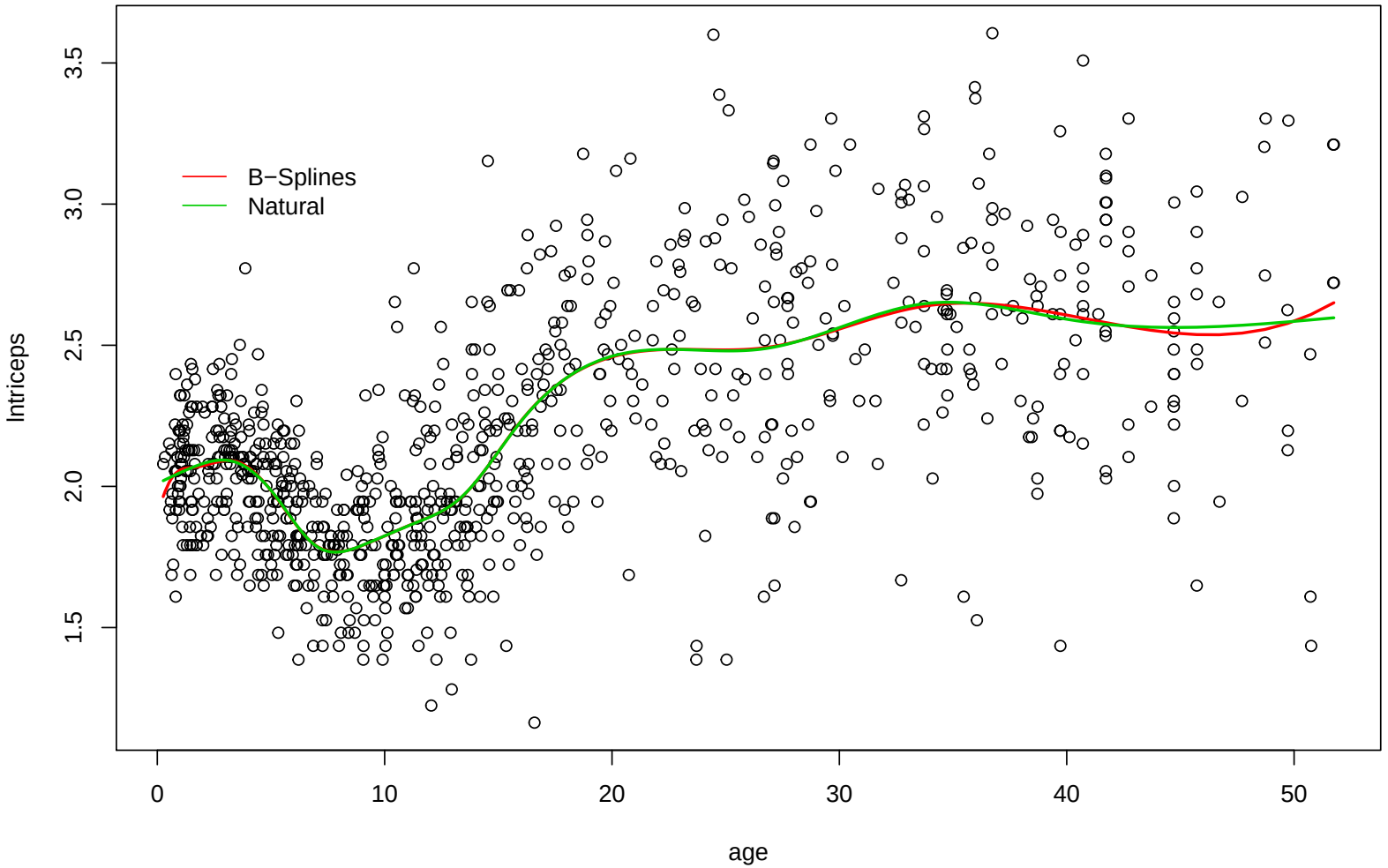
Scatterplot smoothing (default values)



Scatterplot smoothing (k=4)

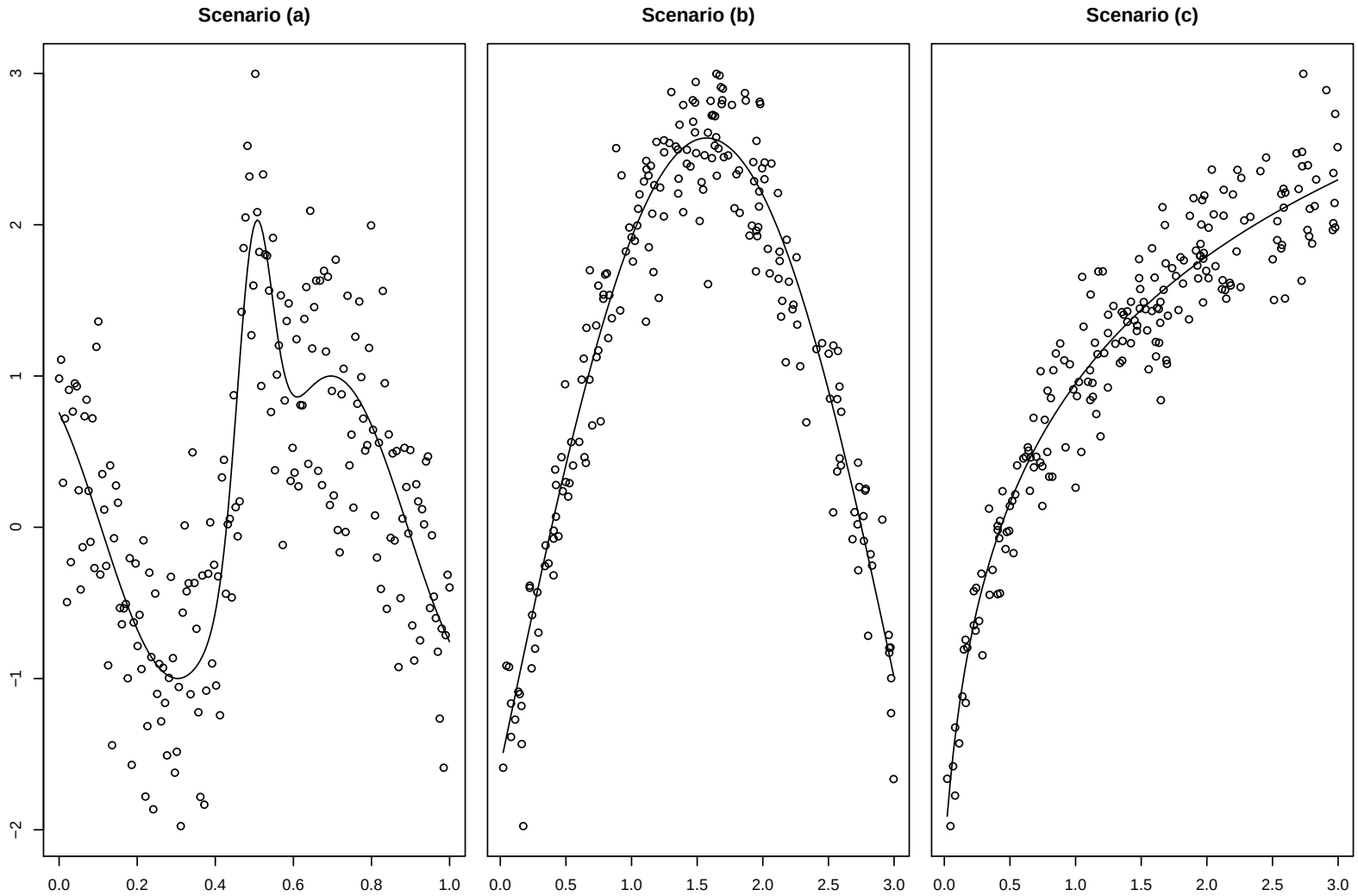


Scatterplot smoothing (k=12)

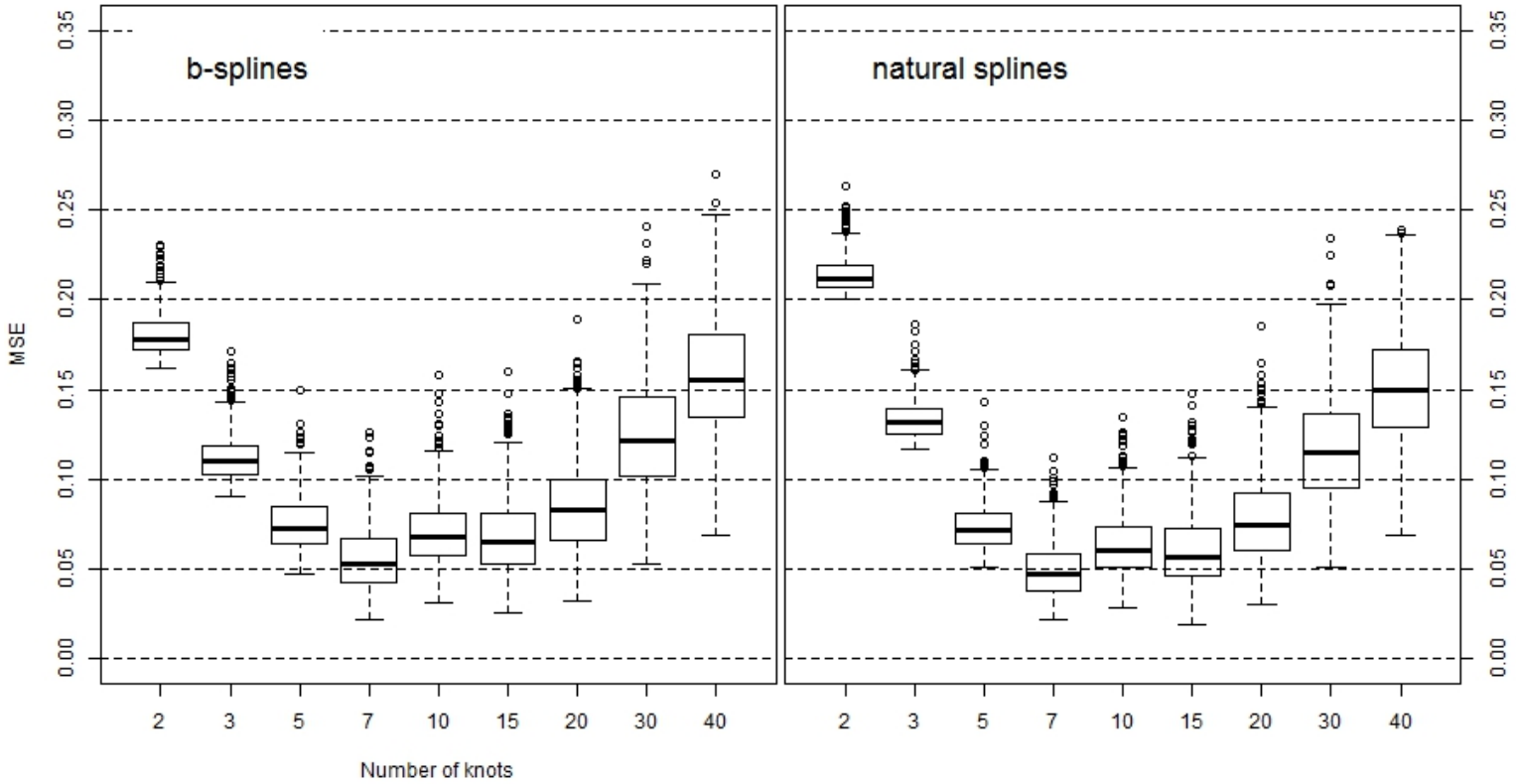


Simulation

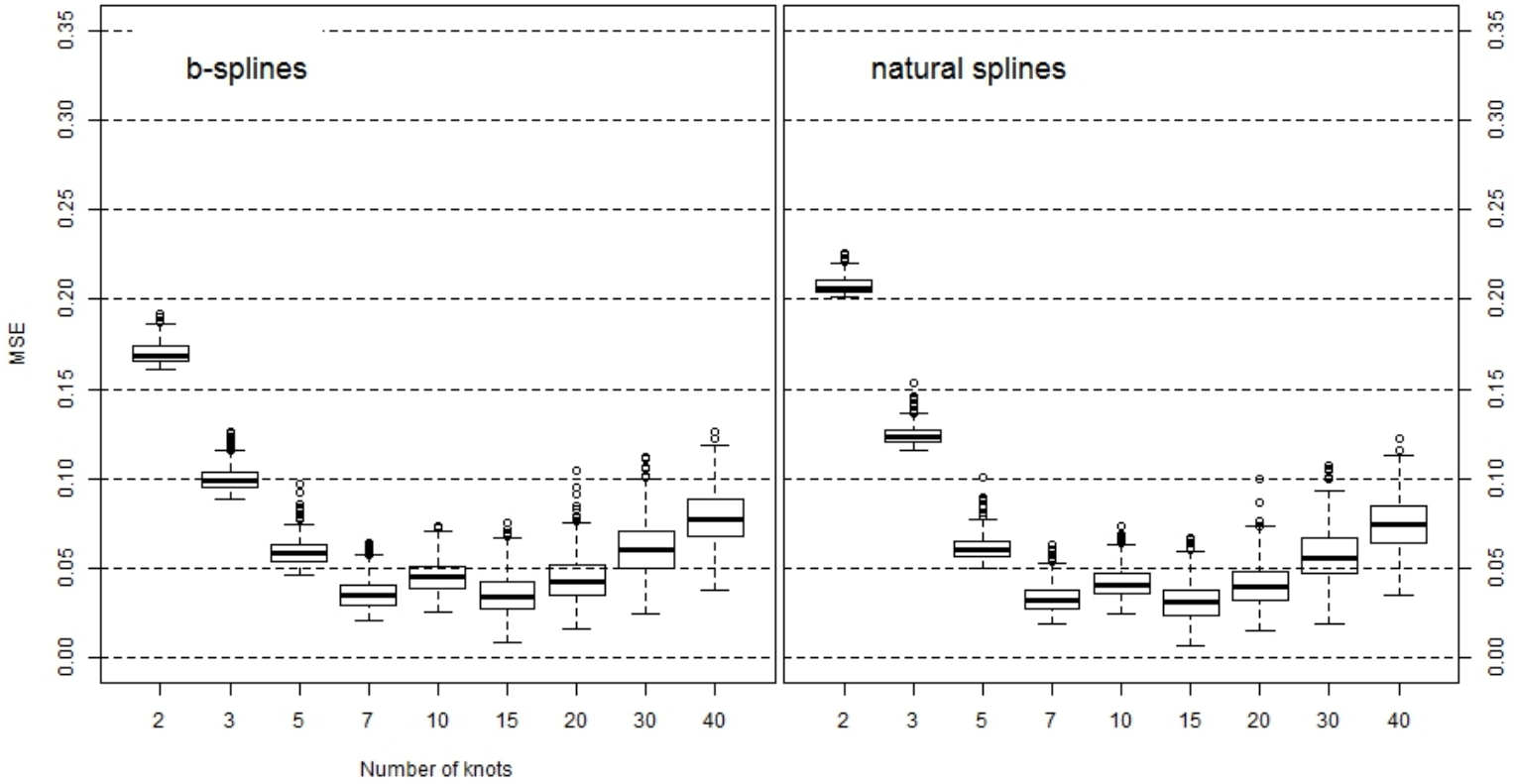
- Mean Square Error depending on the number of knots
- Three different scenarios:



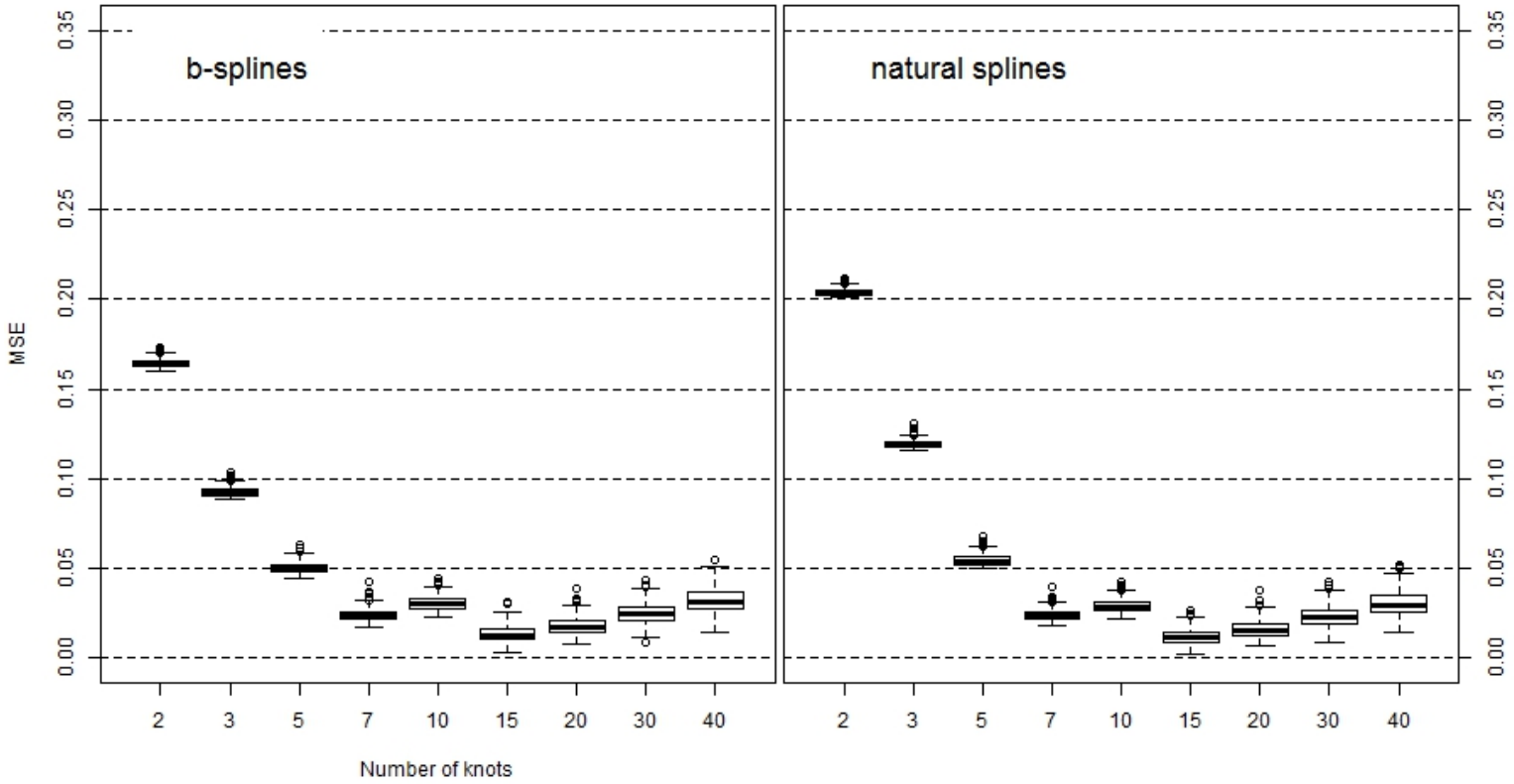
Sim (a): n=100


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
3
5
7
10
15
20
30
40
Number of knots

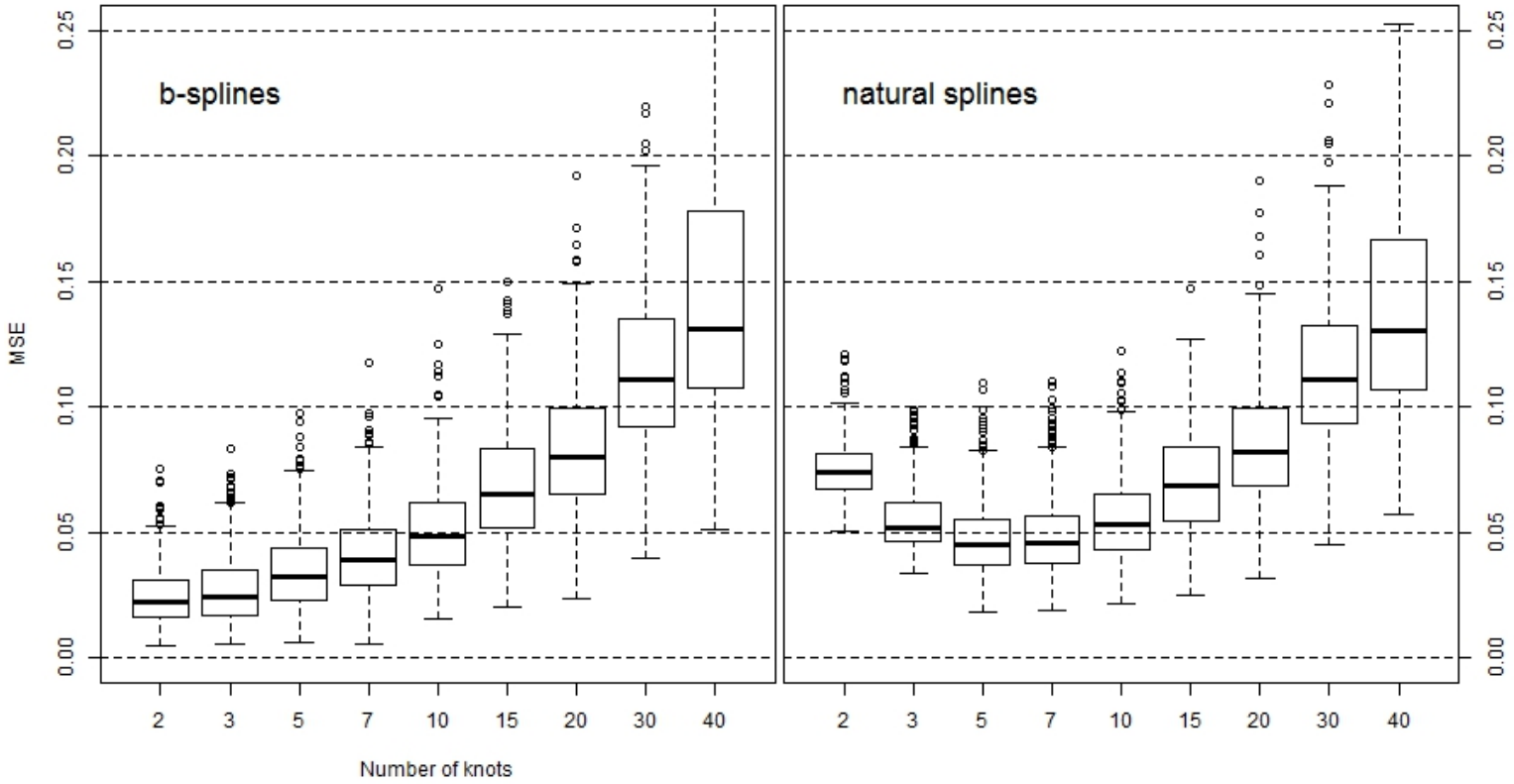
Sim (a): n=200


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
3
5
7
10
15
20
30
40
Number of knots

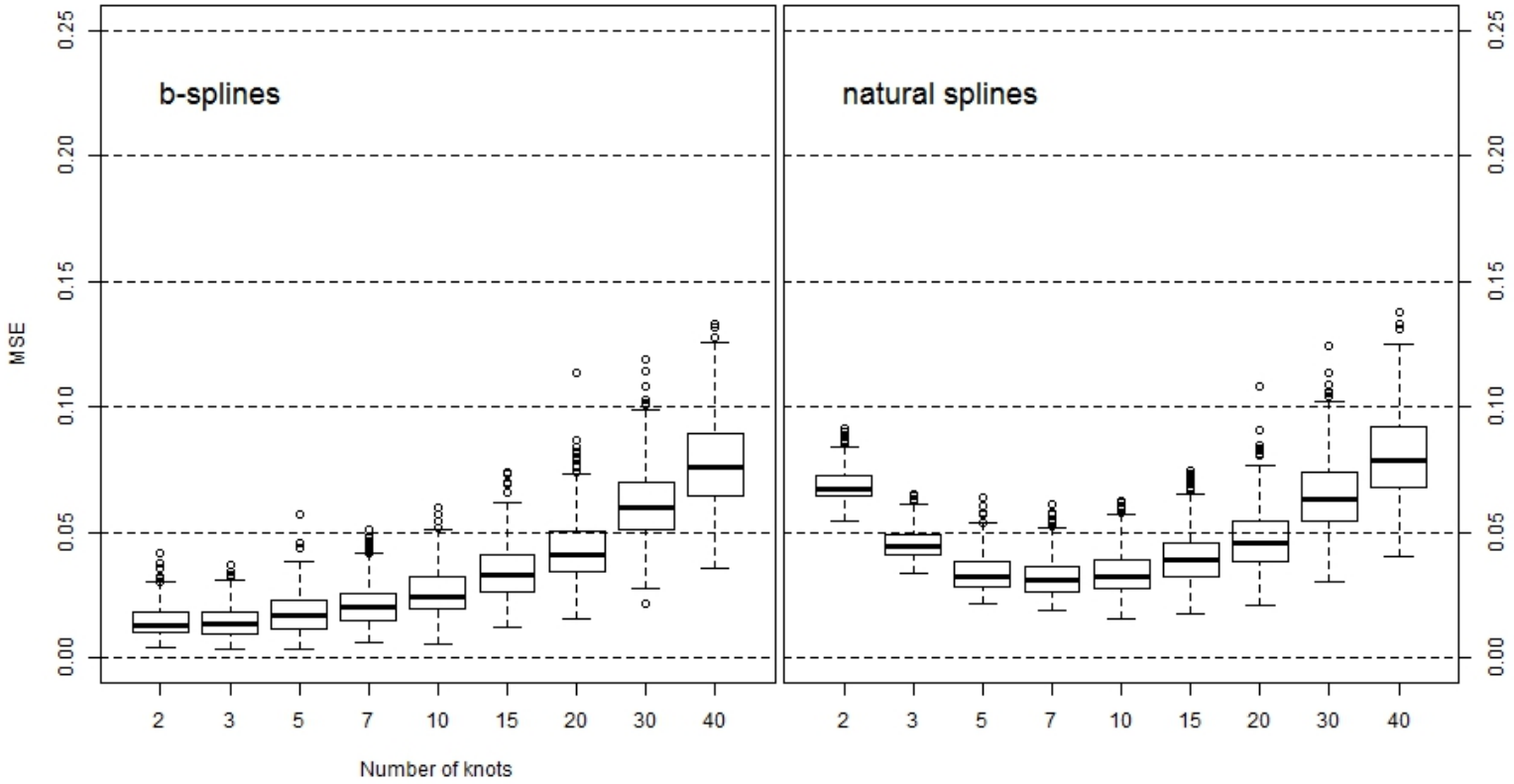
Sim (a): n=500


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
3
5
7
10
15
20
30
40
Number of knots

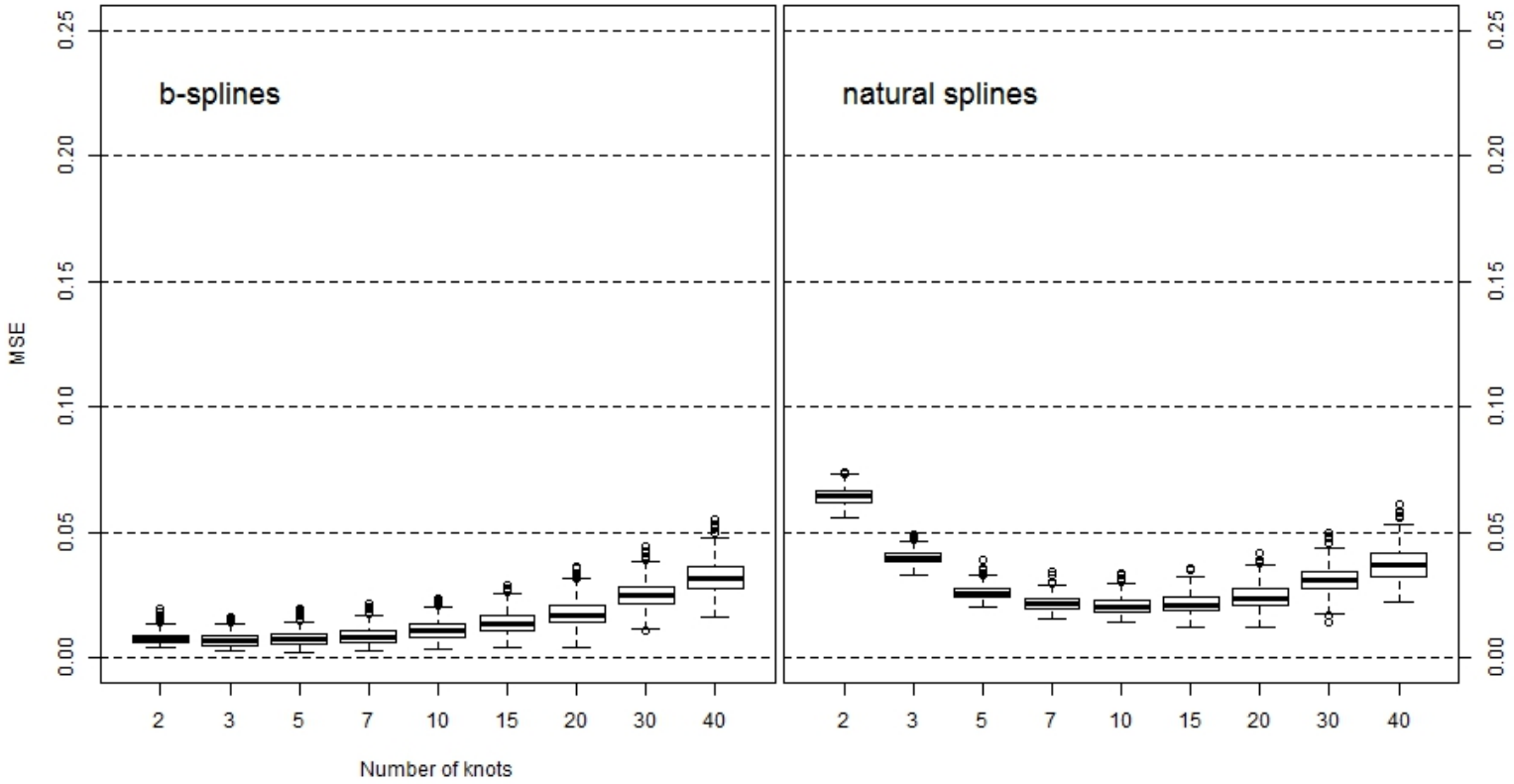
Sim (b): n=100


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
2
3
5
7
10
15
20
30
40
Number of knots

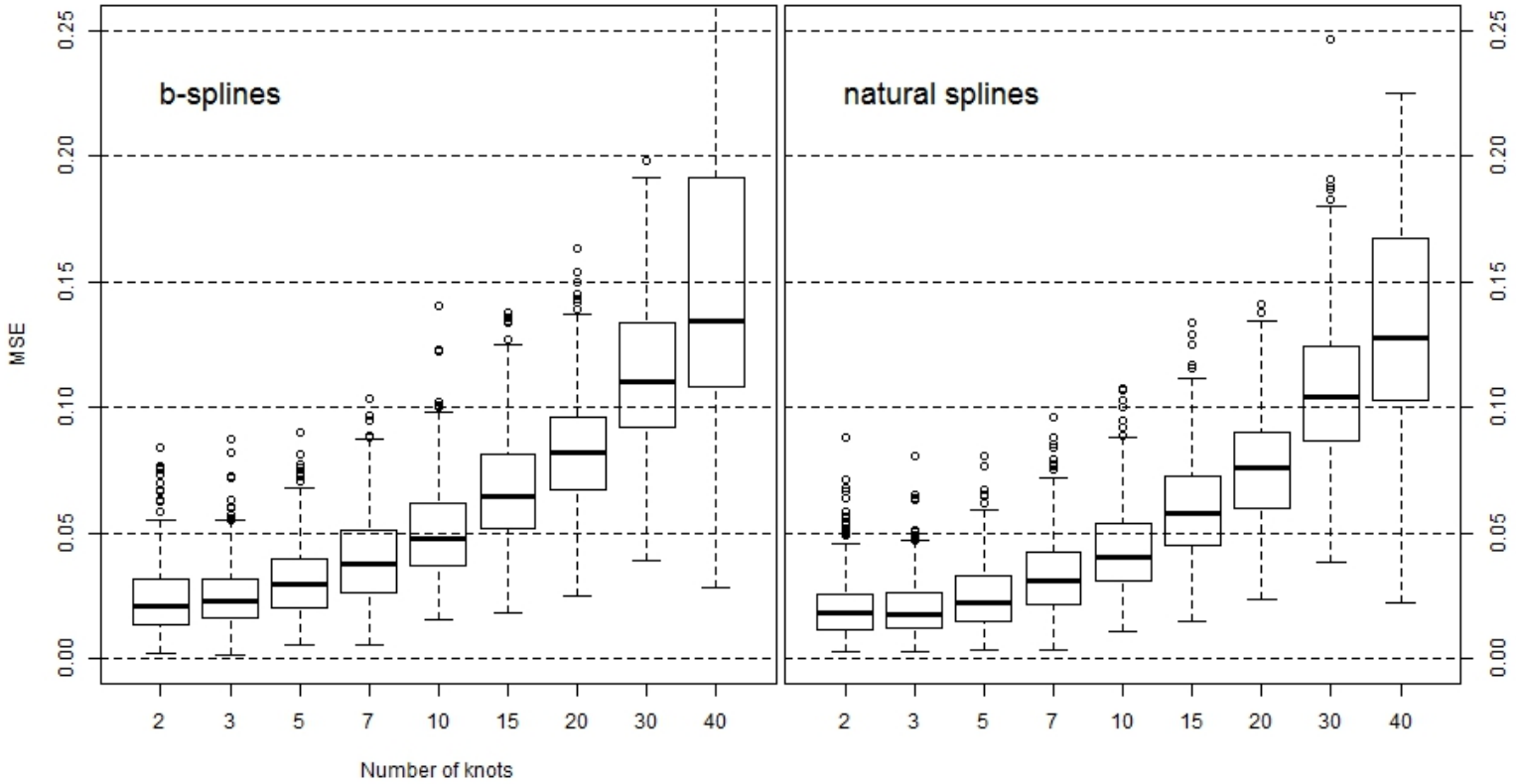
Sim (b): n=200


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
2
3
5
7
10
15
20
30
40
Number of knots

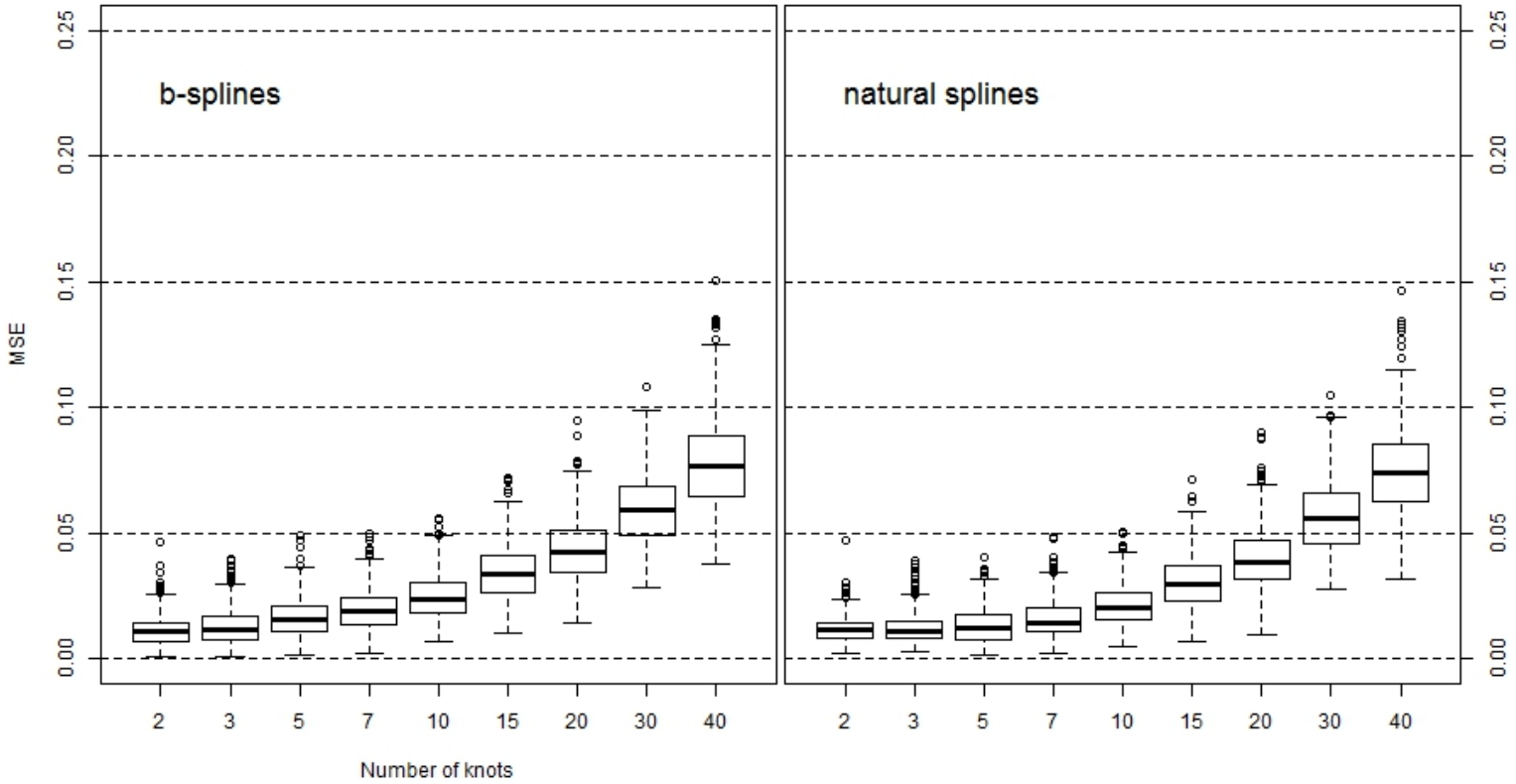
Sim (b): n=500


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
2
3
5
7
10
15
20
30
40
Number of knots

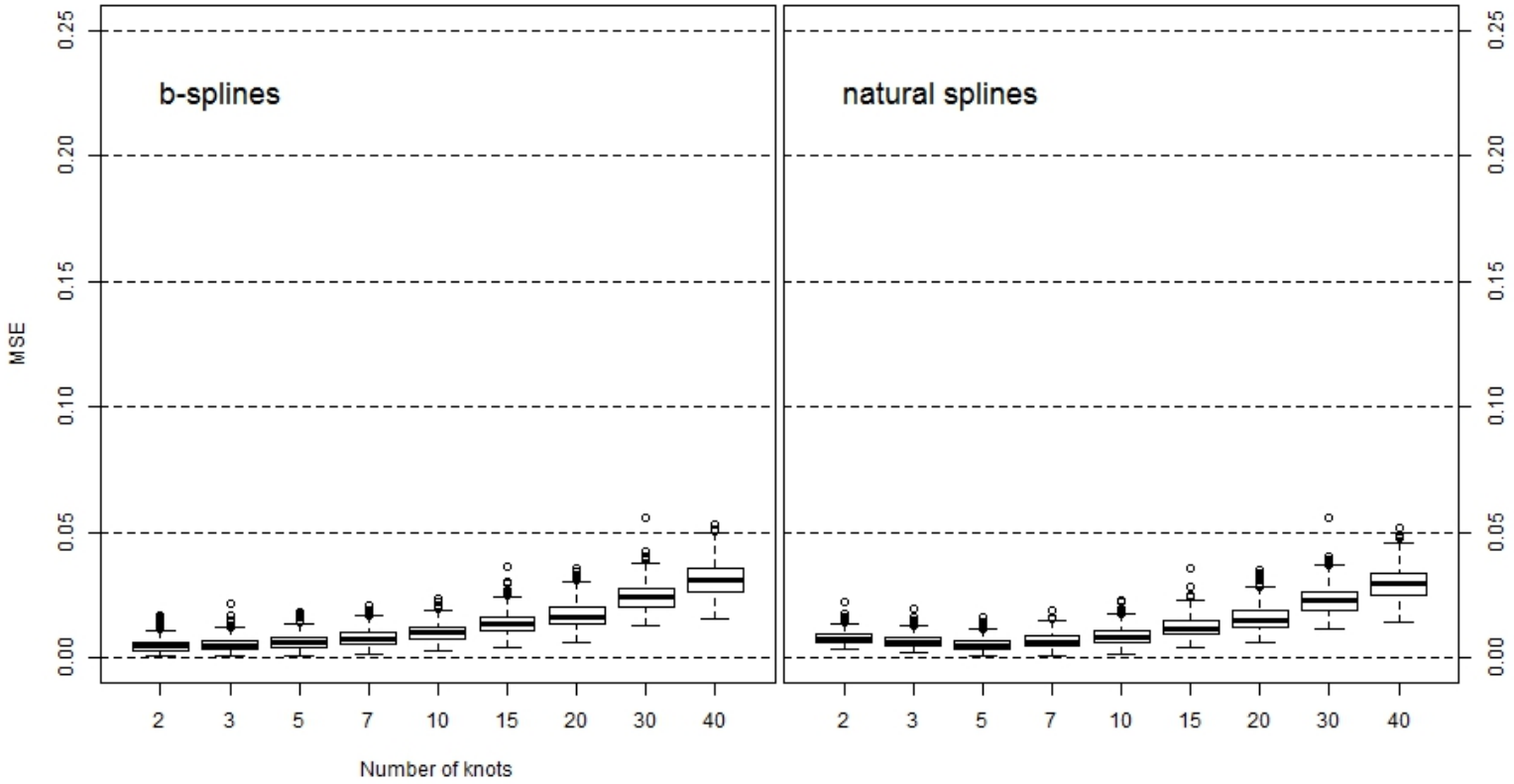
Sim (c): n=100


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
2
3
5
7
10
15
20
30
40
Number of knots

Sim (c): n=200


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
2
3
5
7
10
15
20
30
40
Number of knots

Sim (c): n=500


b-splines
natural splines
MSE
0.00
0.05
0.10
0.15
0.20
0.25
2
3
5
7
10
15
20
30
40
Number of knots

Some ideas

- There is no such thing as an "optimal" number of knots
- MSE varies with number of knots, sample size, type of spline and type of data
- B-splines tend to behave better with a smaller number of parameters. They should outperform natural splines when the fit needed is quite flexible
- The choice of basis becomes less relevant as number of knots increases

Other packages

Package	Description	Authors
gss	General Smoothing Splines	C Gu
polspline	Polynomial spline routines	C Kooperberg
pspline	Penalized Smoothing Splines	B Ripley
cobs	Constrained B-Splines	PT Ng et al
crs	Categorical Regression Splines	JS Racine et al
bigsplines	Smoothing Splines for Large Samples	NE Helwig
bezier	Bezier Curve and Spline Toolkit	A Olsen
freeknotsplines	Free-Knot Splines	S Spiriti et al
Orthogonal splinebasis	Orthogonal B-Spline Functions	A Redd
pbs	Periodic B Splines	S Wang
logspline	Logspline density estimation routines	C Kooperberg
episplineDensity	Density Estimation Exponential	S Buttrey et al
Hmisc, rms	restricted cubic splines, plots	F Harrel

Regression Packages

A brief overview of packages

Package	Downloads	Vignette	Book	Website	Datasets
quantreg	2001231	X	X		7
mgcv	1438166	X	X		2
survival	1229305	X	X		33
VGAM	297308	X	X	X	50
gbm	271362			X	3
gam	168143		X	X	1
gamlss	78295	X	X	X	29

A brief overview of packages

Response	mgcv	quantreg	VGAM	gbm	gam	gamlss
Linear	X	X	X	X	X	X
Categorical	X		X	X	X	X
Count	X	X	X	X	X	X
Survival	X		X	X	X	X
Quantile Reg		X	X	X	X	X
Multivariate	X				X	X
Nonlinear			X		X	X
Reduced Rank			X		X	
Other	X	X	X	X	X	X

Splines in packages

package	bs	ns	s	p-splines
gam	x	x	x	
mgcv	x	x	x	x
VGAM	x	x	x	x
gamlss	x	x	cs	x

Criteria for significance of non-linear effect and variable selection methods

package	non-linear	var selection
mgcv	df	double penalty
quantreg		lasso
survival	residuals	
VGAM		
gbm		boosting
gam		stepGAM
gamlss		stepGAIC

Common bases in mgcv

- tp: Thin plate regression spline, ts as tp but with a modification to the smoothing penalty
- ds: Duchon splines (generalisation of thin plate splines)
- ps: p-splines
- cr: Cubic regression splines, cs specifies a shrinkage version of cr, cc specifies a cyclic cubic regression splines i.e. a penalized cubic regression splines whose ends match, up to second derivative. re: parametric terms penalized by a ridge penalty

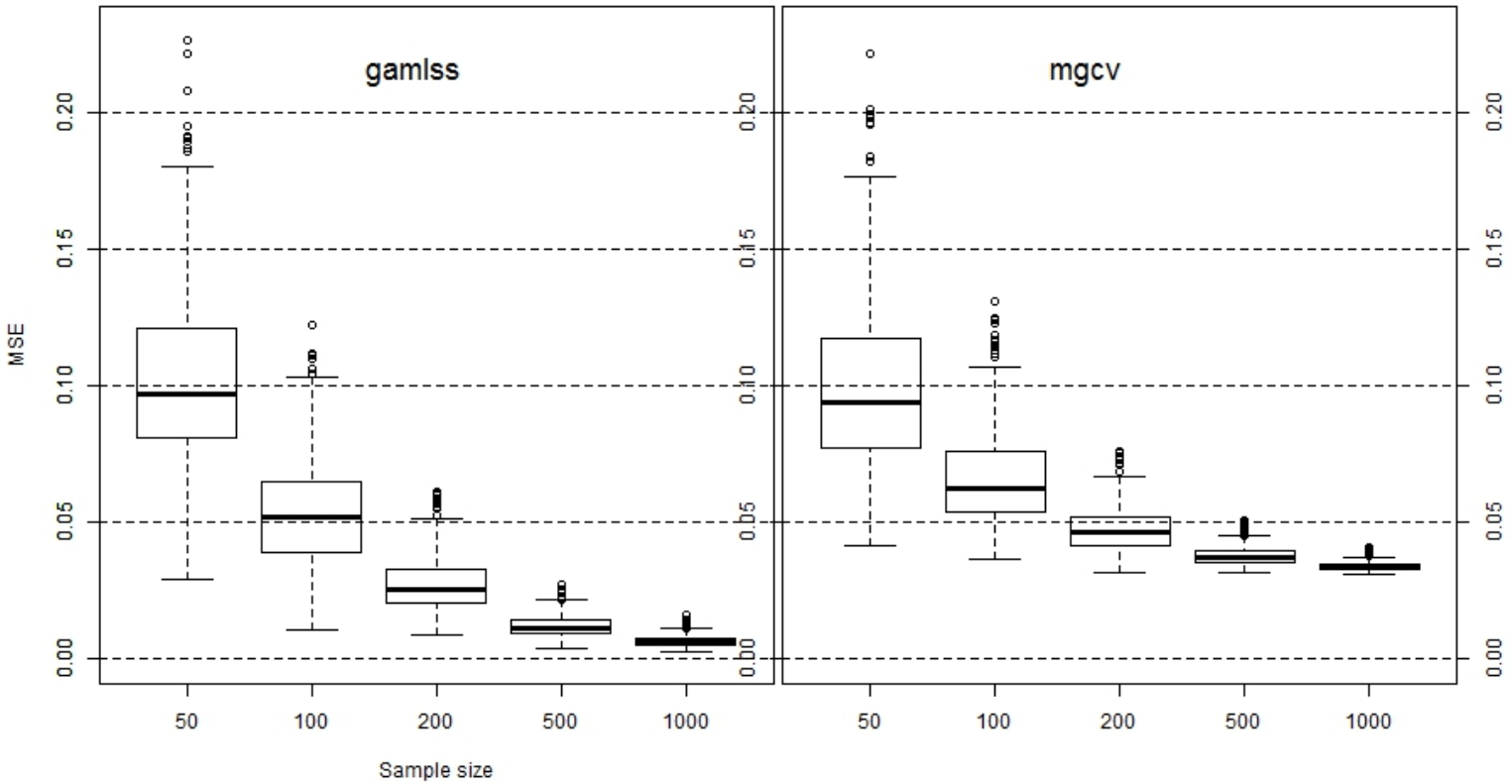
Common basis in Gamlss

Table 5.1: Implemented **gamlss** additive functions

Additive terms	R Name	Section
Cubic splines	cs()	5.1
Varying coefficient	vc()	5.2
Penalized splines	ps()	5.3
`loess`	lo()	5.4
Fractional polynomials	fp()	5.5
Random effects	random()	5.6.1
Random effects	ra()	5.6.2
Random coefficient	rc()	5.6.3

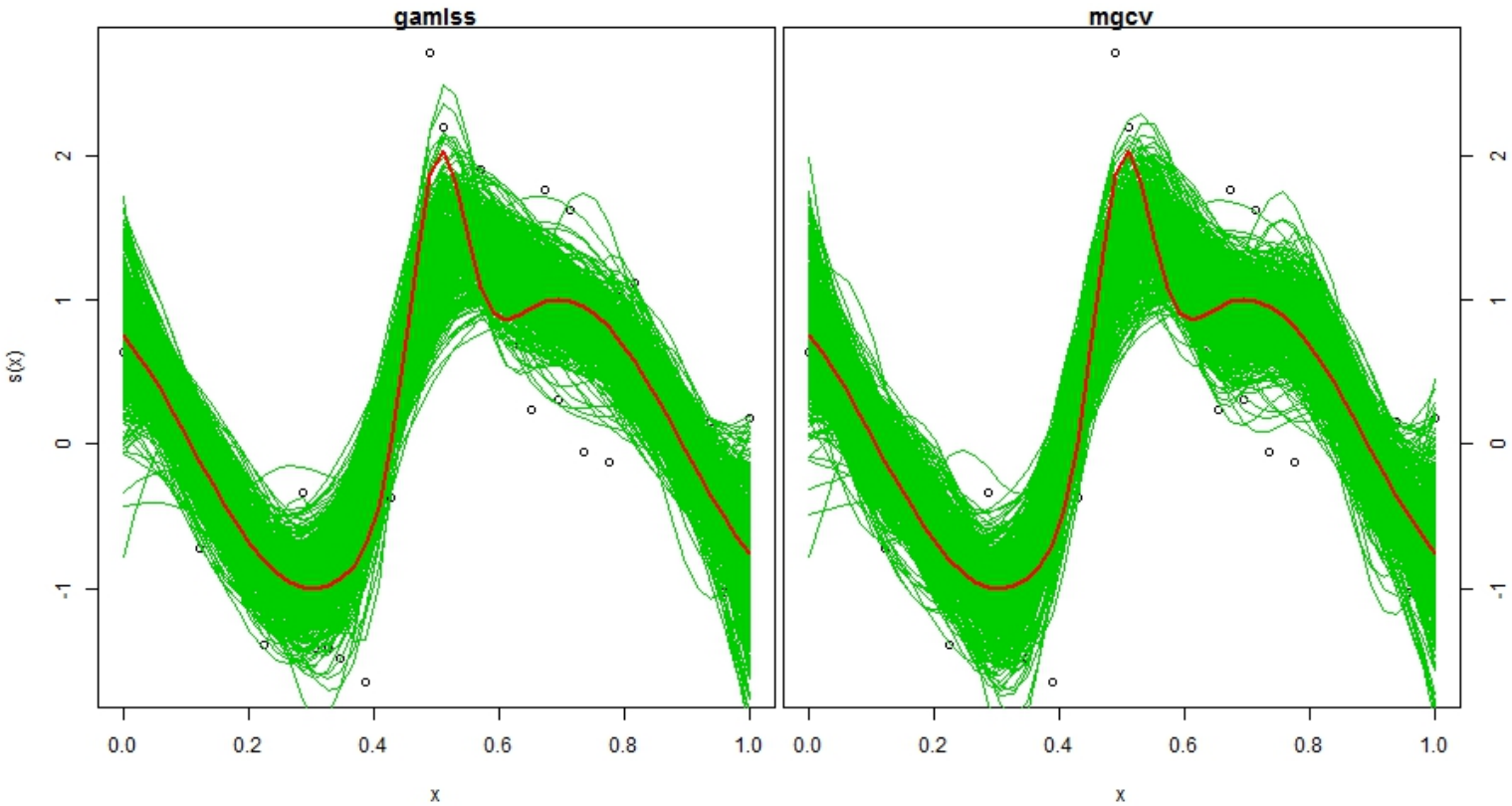
Comparison

- Testing at default values thin plate regression splines (Wood 06) vs p-splines (Eilers & Marx 96)



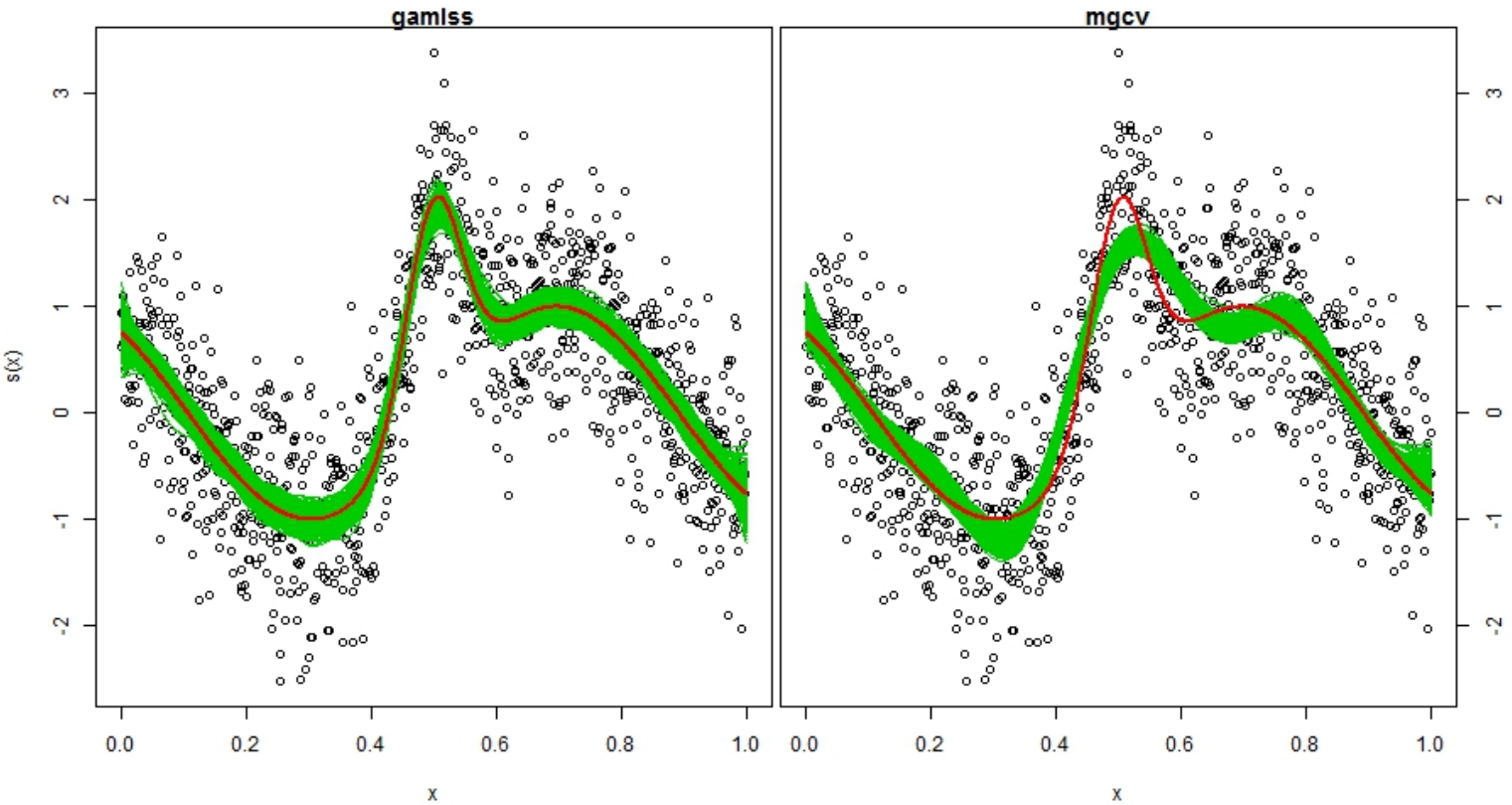
P-splines vs Thin Plate Regression Splines

- Sample size n=50



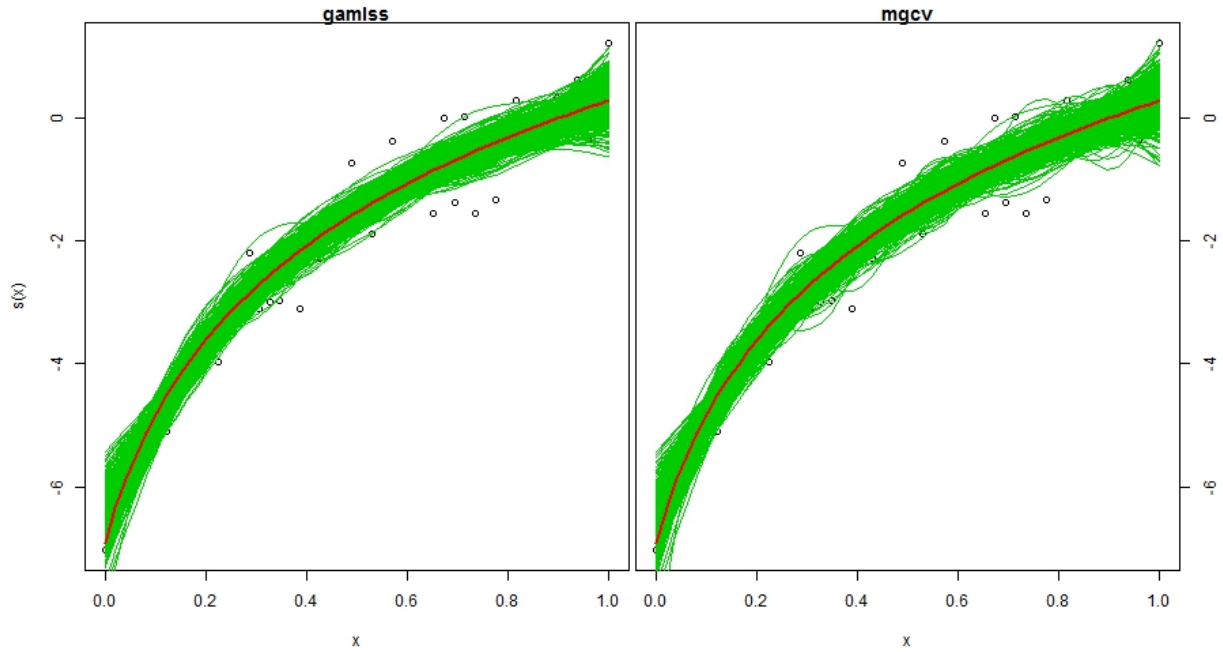
P-splines vs Thin Plate Regression Splines

- Sample size n=1000



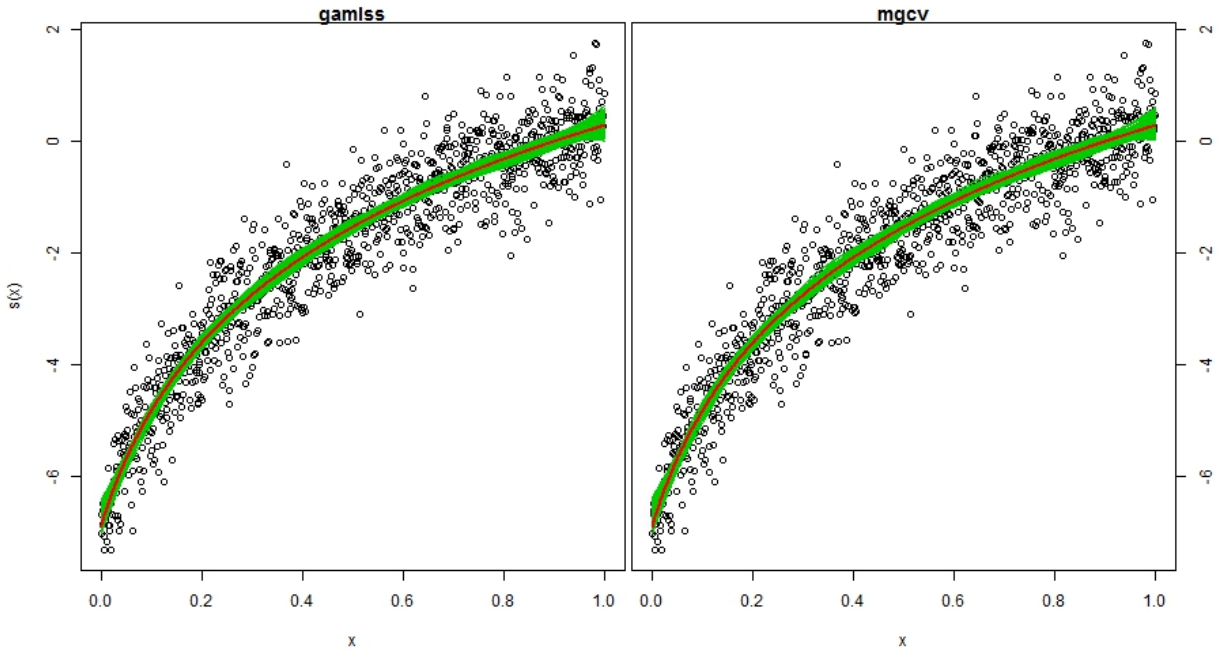
P-splines vs Thin Plate Regression Splines

- Sample size n=50



P-splines vs Thin Plate Regression Splines

- Sample size n=1000



MSE p-splines vs thin plate splines vs smoothing splines

scenario (a)					
n	50	100	200	500	1000
mgcv tp	.102	.067	.049	.038	.034
gamlss pb	.106	.055	.029	.012	.006
gam s	.160	.146	.137	.133	.131
scenario (b)					
mgcv tp	.039	.021	.010	.000	.000
gamlss pb	.039	.023	.012	.000	.000
gam s	.035	.019	.013	.000	.034
scenario (c)					
n	50	100	200	500	1000
mgcv tp	.046	.026	.013	.000	.000
gamlss pb	.042	.023	.012	.000	.000
gam s	.039	.023	.015	.010	.009

Discussion

Some points

- Anything is possible for the experts, not so straight forward for applied analysts.
- Not an easy task to suggest an overall strategy.
- Paper on review of spline methods in R almost completed.
- Paper on evaluation of spline methods in R in progress.