Review and comparison of spline procedures in multivariable regression.

Aris Perperoglou TG2 STRATOS

RSS 2018

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

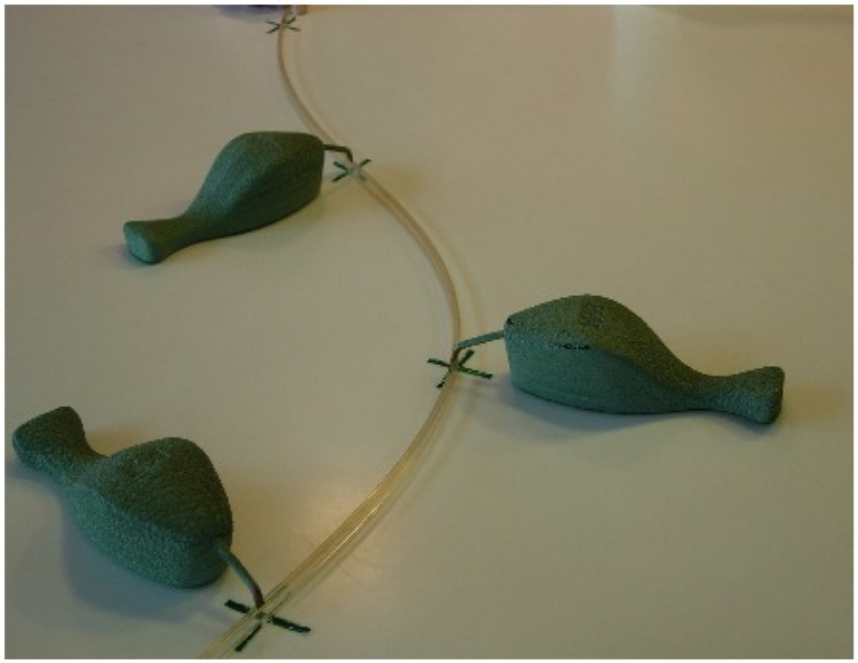
. . . main focus of TG2 is to identify influential variables and gain insight into their individual and joint relationship with the outcome. Two main challenges are selection of variables for inclusion in a multivariable explanatory model, and choice of the functional forms for continuous variables (Harrell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

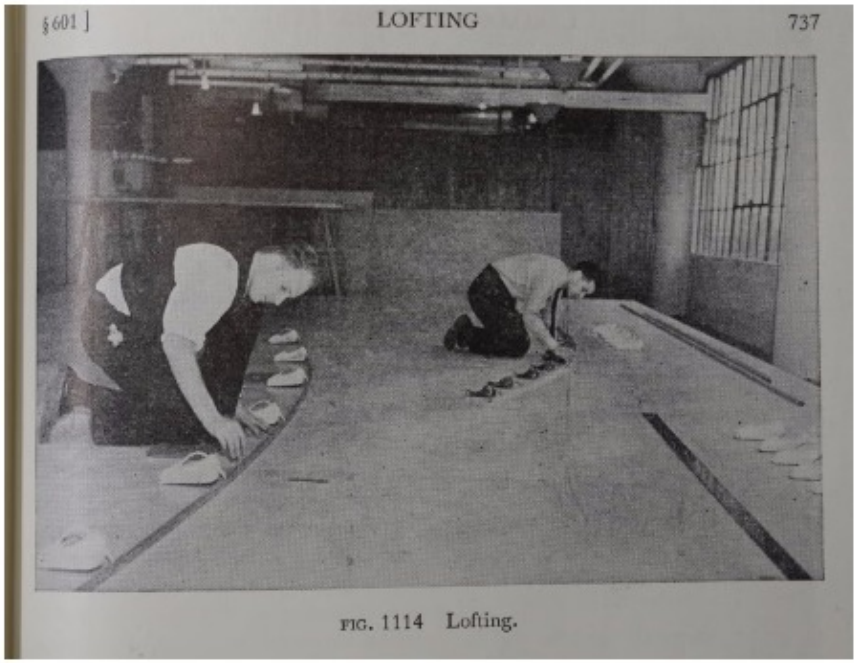
- identify and assess methods currently used in practice
- compile and publish guidance on selection of variables and their functional forms

The Subject

- Fit a statistical model of the form $g(Y|X) = \beta_0 + f(X)$
  - p explanatory variables $X = (X_1, ..., X_p)$
  - f unknown, allowed to be nonlinear but should be interpretable
- Common specification: $f(X_1, ..., X_p) = f_1(X_1) + ... + f_p(X_p)$
  - Generalized additive models (GAMs)
- Splines are a popular method to estimate $f_1, ..., f_p$
  - GAM books by Hastie/Tibshirani and Wood are hugely popular ($>$14000 and $>$6000 citations, respectively)

Splines

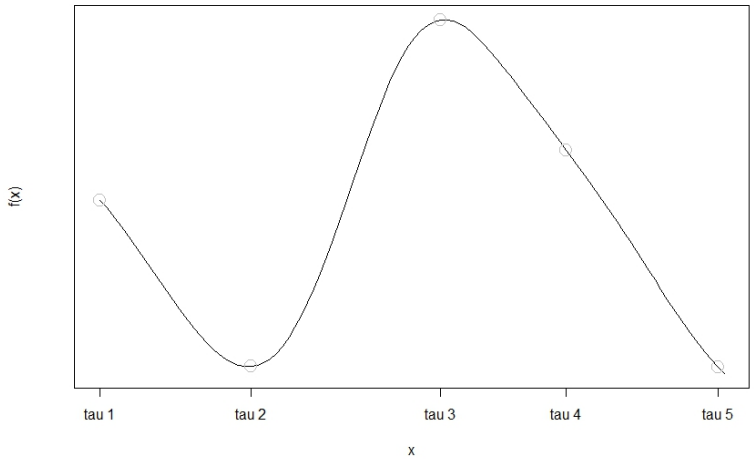




§601] LOFTING 737

FIG. 1114 Lofting.

Definition of Splines



- Set of piecewise polynomials, each of degree d
- Joined together at a set of knots $\tau_1, ..., \tau_K$
- Continuous in value, sufficiently smooth at the knots
- Depending on knots and type of polynomial spline is named as:
  - polynomial, natural, restricted regression spline, truncated power basis, b-spline (**de Boor 1978**)
  - smoothing splines: p-spline (**Marx and Eilers 1996**), penalized regression spline (**Wood 2006**)

Spline Estimation

- Splines are typically represented by a set of (non-unique) basis functions $B_1, ..., B_{K+d+1}$
  - $f(X) = \sum_{k=1}^{K+d+1} \beta_k B_k(X)$
  - Linearization of the estimation problem
- Many types & subtypes, e.g.,
  - Natural splines: Required to be linear in τ_1 and τ_K
  - Penalized splines: Minimization of a critierion of the form

$$\text{Sum of Residuals} + \lambda J_m$$

with smoothing parameter λ and "wiggliness" parameter m

The problem

- Spline modeling involves the selection of a comparatively large number of parameters
- Number & placement of knots
- Choice of basis & restrictions
- Choice of smoothing parameter / optimization procedure
- Choice of penalty order
- Mathematical properties are well understood, but . . .

The Problem (2)

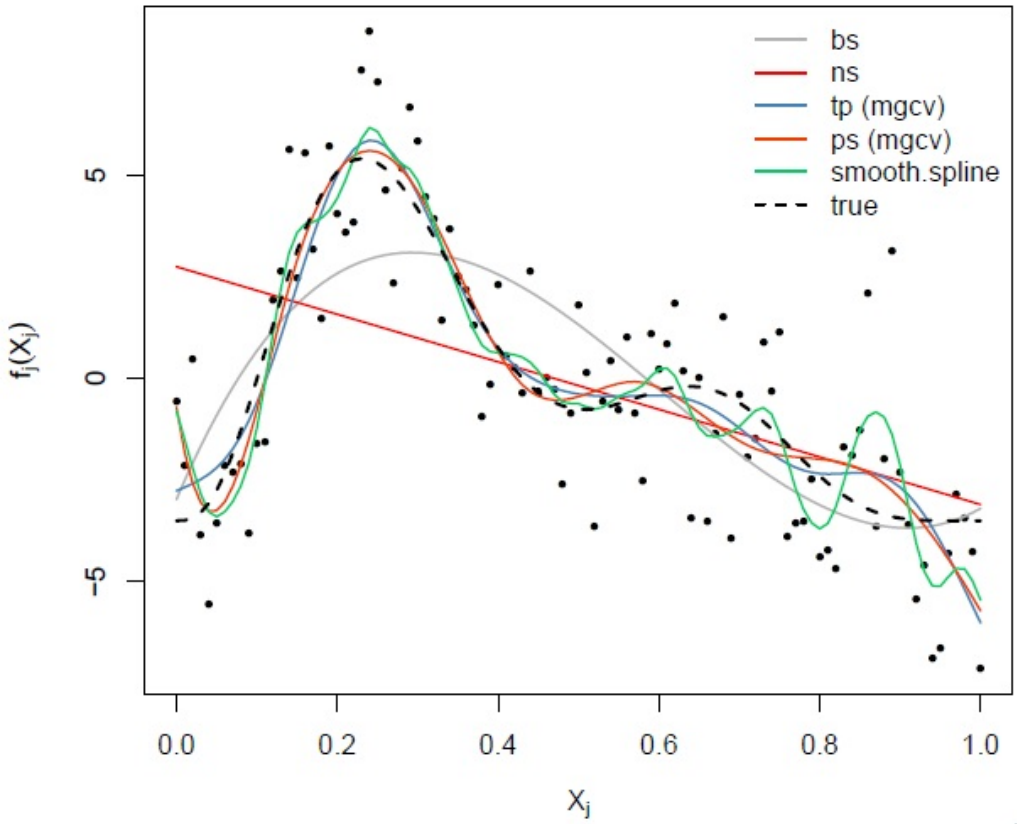
- Little guidance for applied statisticians
- Even for level 2 users it is hard to choose between competing approaches

Broadly speaking the default penalized thin plate regression splines tend to give the best MSE performance, but they are slower to set up than the other bases. The knot based penalized cubic regression splines (with derivative based penalties) usually come next in MSE performance, with the P-splines doing just a little worse. However the P-splines are useful in non-standard situations.

(Excerpt from smooth.terms help file in R package mgcv)

The problem (3)

- Little guidance (so can we rely on software?)
- Default results of some spline procedures in R:



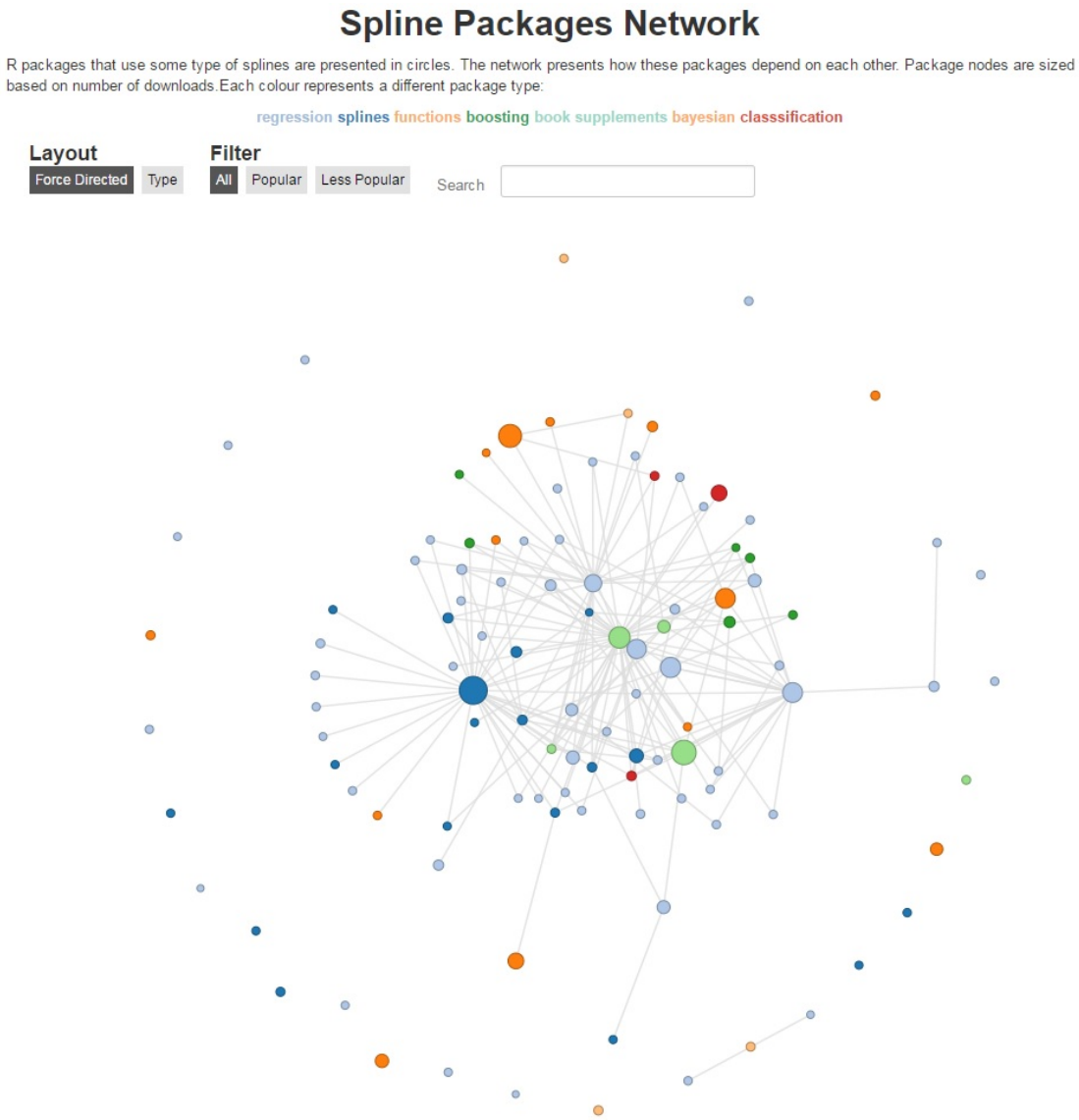
R Cran packages

- May 2010, there existed 2,445 packages
- May 2015, more than 6200 packages
- May 2016, more than 8500
- August 2018: 12,879
  - 7,377 package maintainers
  - 243 updates last week
  - 9,728,940 downloads last week

Splines packages

- more than 300 splines packages:
  - Spline basis: splines, splines2, gss, polspline, pspline, cobs, crs, bigsplines, bezier, freeknotsplines, Orthogonal splinebasis, pbs, logspline, epispline...
  - Regression: gam, mgcv, VGAM, gbm,survival, gamlss...
  - Miscalenea: Hmisc, MASS...

Network


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

The splines Package

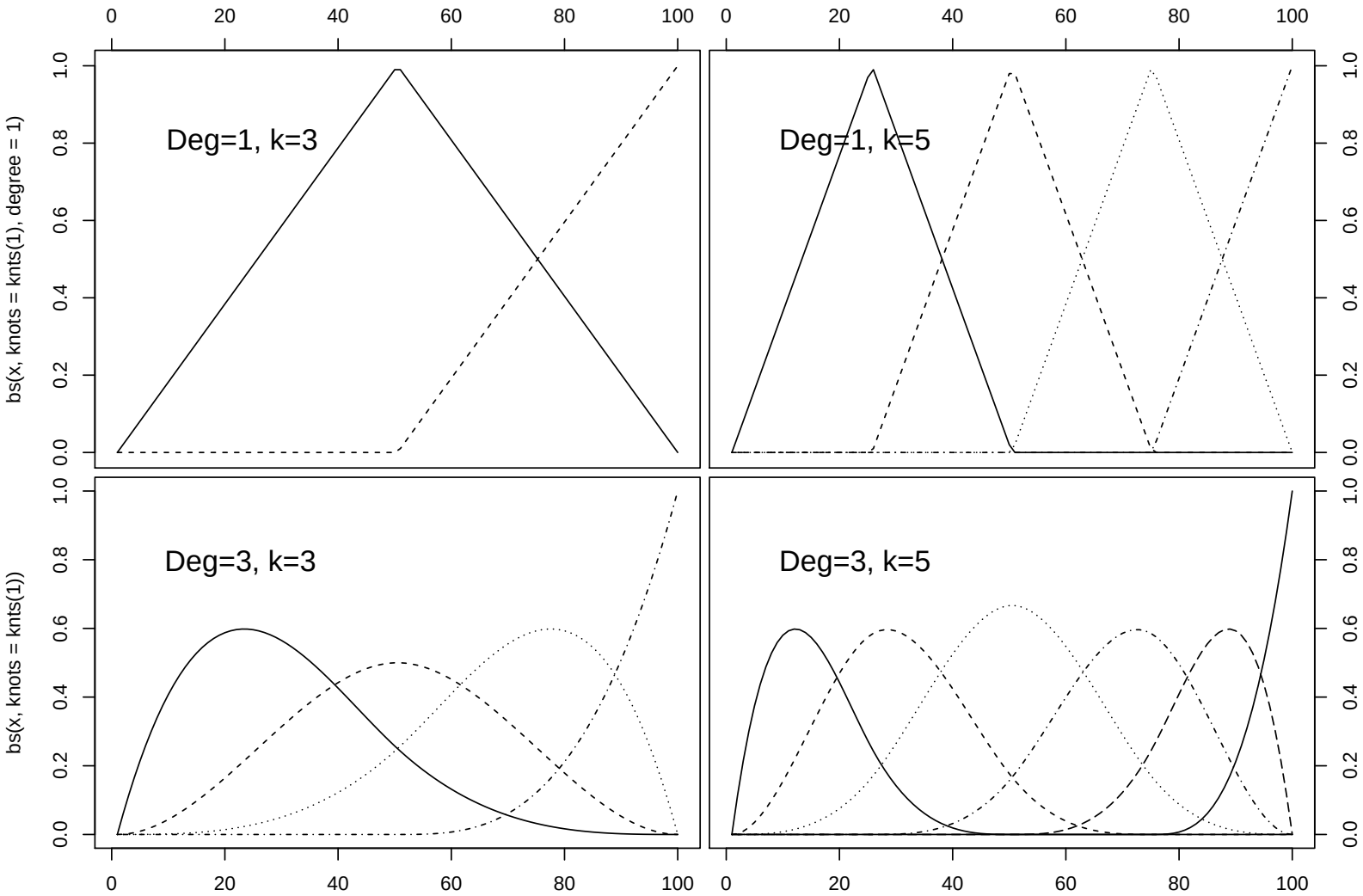
- bs: B-Spline Basis for Polynomial Splines

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

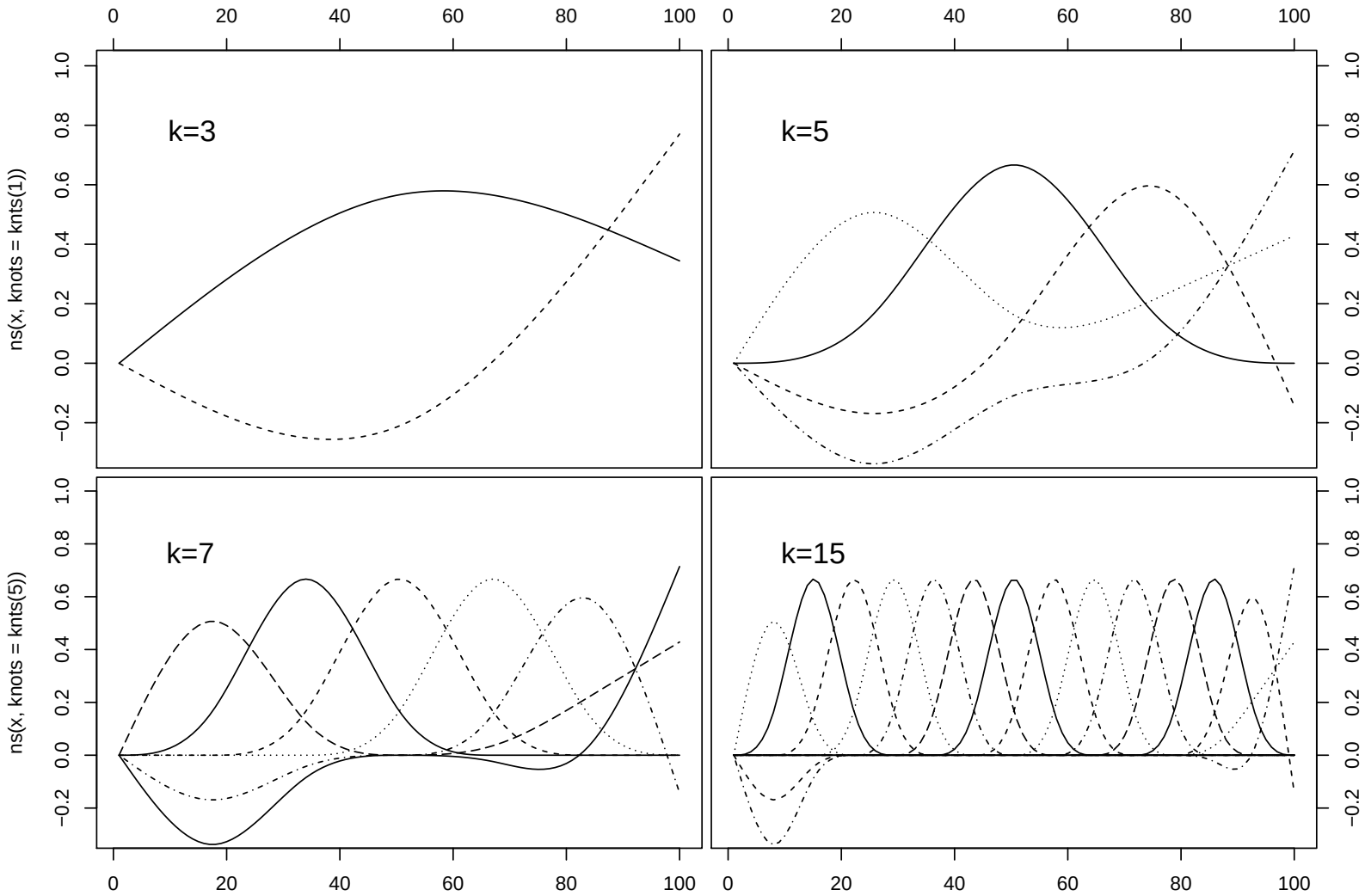
- ns: Generate a Basis Matrix for Natural Cubic Splines

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

B-splines basis



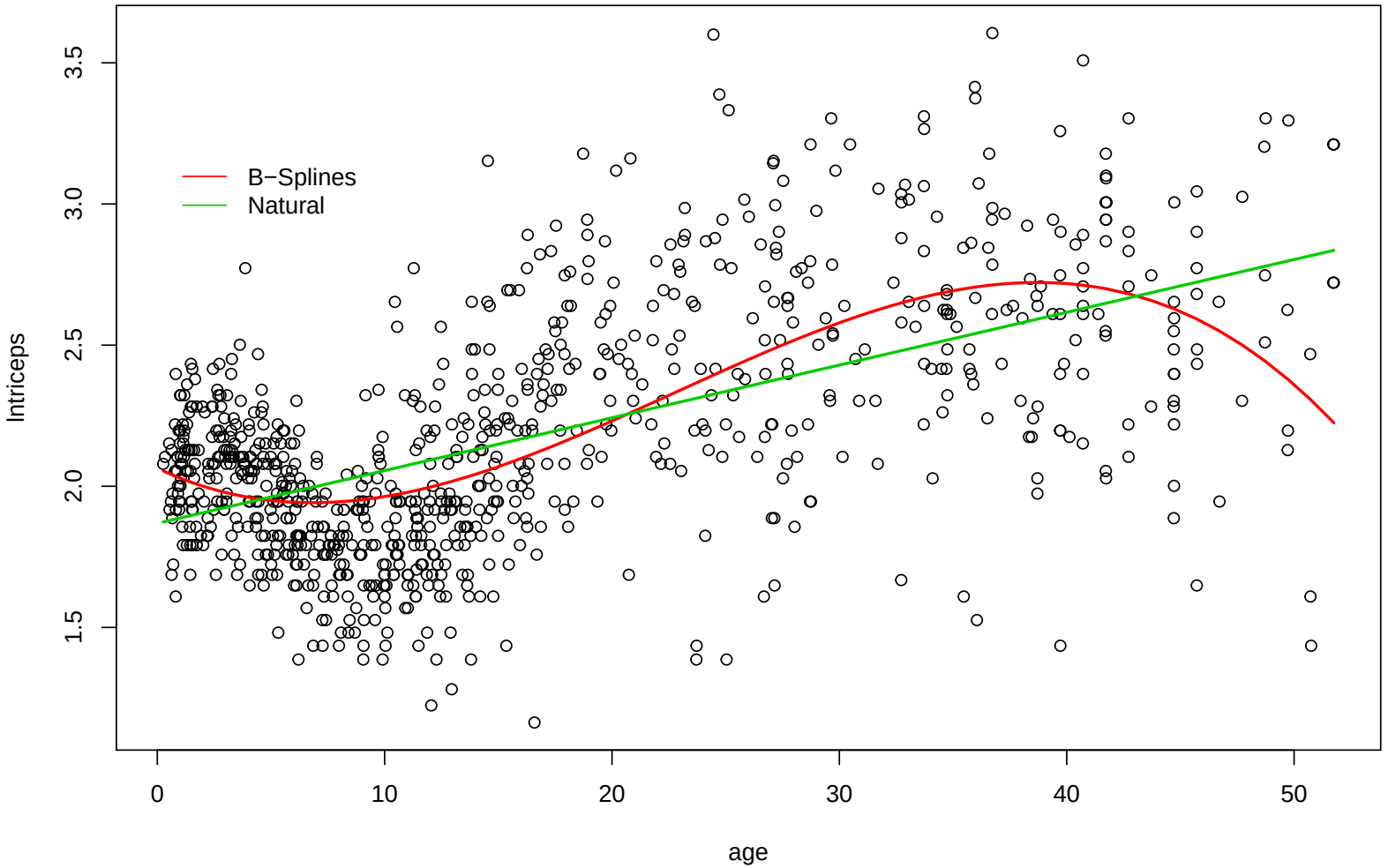
Natural splines



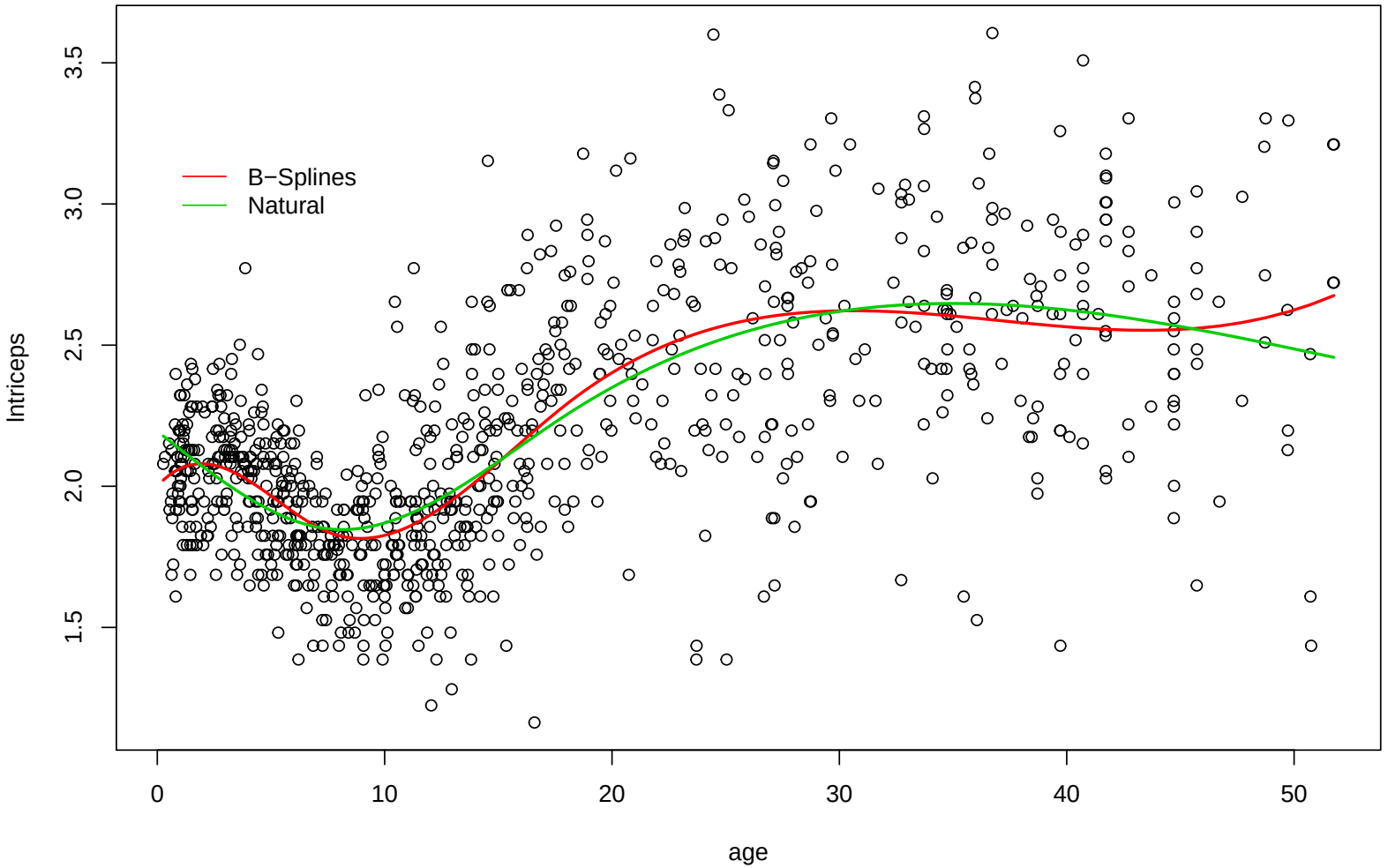
Scatterplot smoothing: Tricepts skinfold thikness by age

```
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)
```

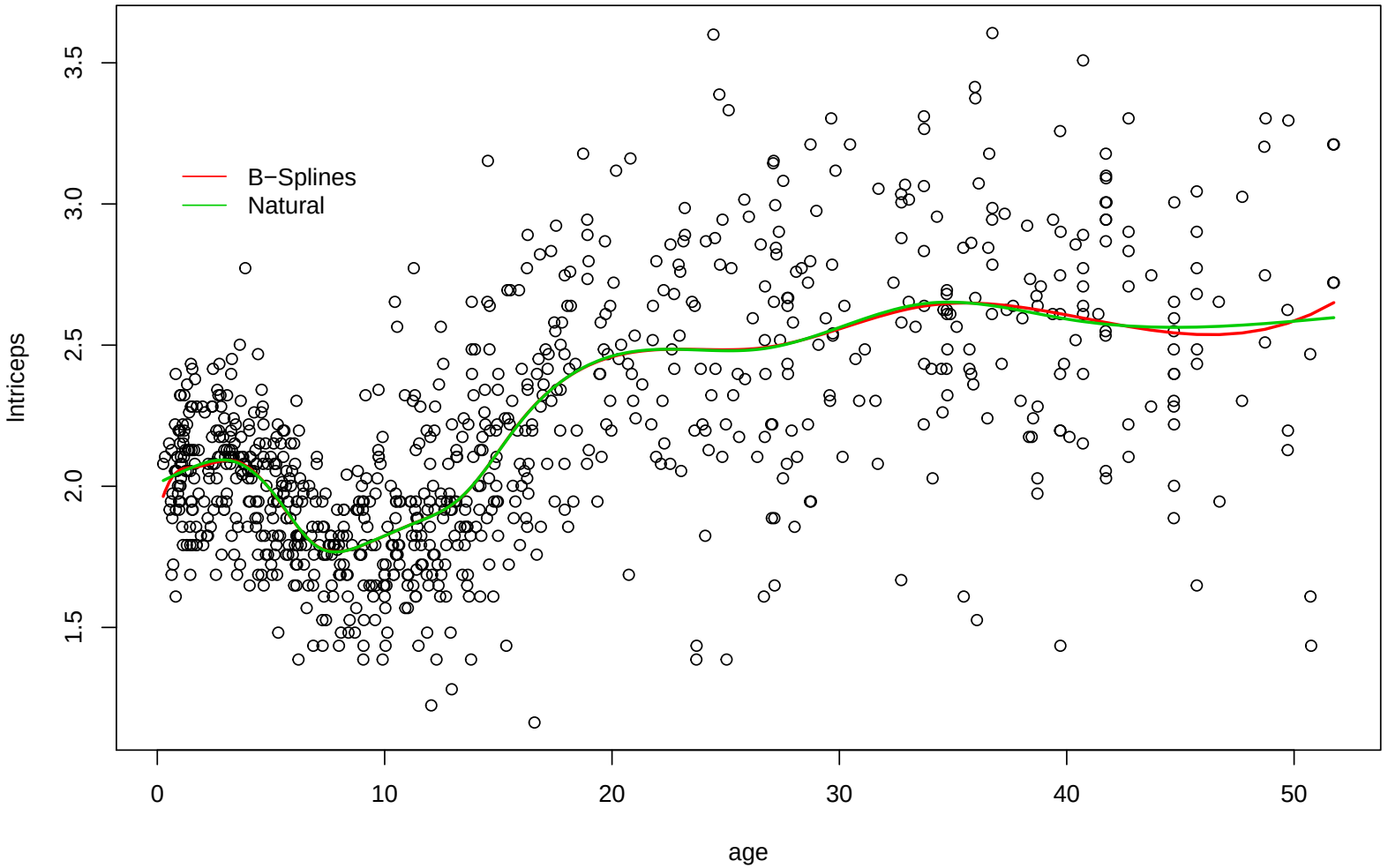
Scatterplot smoothing (default values)



Scatterplot smoothing (k=4)



Scatterplot smoothing (k=12)



Regression packages

A brief overview of packages

Package	Downloads	Vignette	Book	Website	Datasets
mgcv	2974575	x	x		2
survival		x	x		33
VGAM	599436	x	x	x	50
gbm	598233			x	3
gam			x	x	1
gamlss	185273	x	x	x	29

A brief overview of packages (2)

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

mgcv vs gamlss

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

Thin Plate Regression Splines

- Low-rank approximation of thin plate splines
- Knot positions = data locations (with sub-sampling of data locations if n is large)
- Defaults in mgcv:
  - Degree 3
  - Estimation with integrated second-order derivative penalty
  - 9 coefficients per smooth term (null space dimension (= 2) plus 8 minus intercept)
  - Optimization of smoothing parameter via GCV

Penalized Cubic Regression Splines

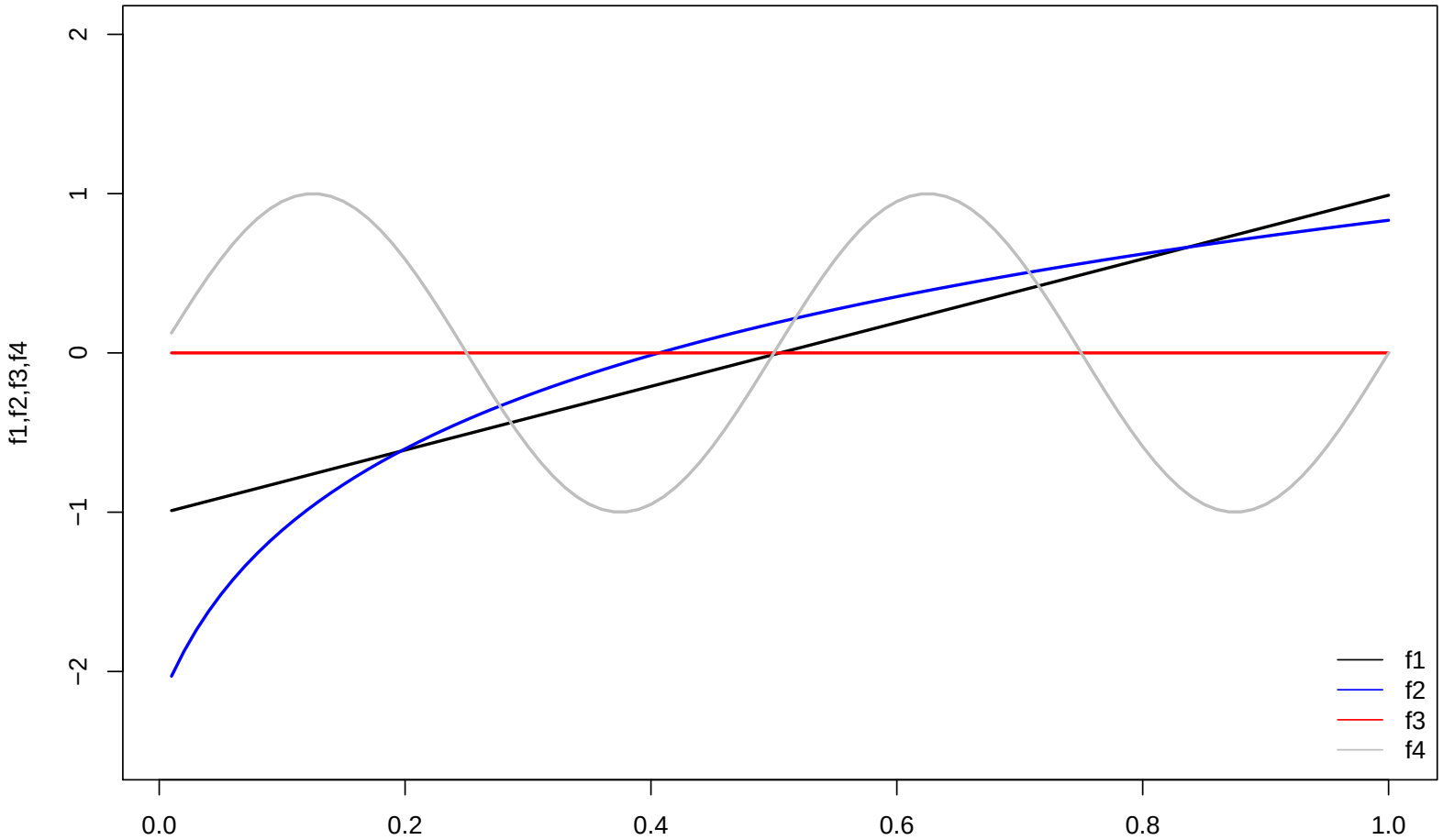
- Natural cubic splines with k knots, integrated second-order derivative penalty
- Based on cardinal spline basis (constructed such that j-th basis function is 1 at the j-th knot and 0 at the other knots, $1 \leq j \leq k$)
- Knots are placed evenly throughout the ordered covariate values
- Defaults in mgcv:
  - 10 knots per smooth term (9 coefficients: # knots minus intercept)
  - Optimization of smoothing parameter via GCV

P-Splines

- Polynomial splines, based on B-spline basis
- Integrated squared derivative penalty is approximated by an m-th order difference penalty
- Knots are placed evenly throughout the ordered covariate values
- Defaults in mgcv:
  - Cubic splines (degree 3) with second-order difference penalty
  - 6 inner knots and 2 boundary knots per smooth term (9 coefficients: # inner knots + degree 3 + 1 minus intercept)
  - Optimization of smoothing parameter via GCV

Simulation Design

- Model: $Y = f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + \epsilon$
- $f_1(X_1) = X1, f_2(X_2) = \log(X2 + 0.05), f_3(X_3) = 0, f_4(X_4) = \sin(4\pi X_4)$



Previous work by Matthias Schmid (mgcv: tp, cr, ps)

Simulation Design (2)

- 100 simulation runs with sample sizes $n = 100, 300, 500$
- Data values of X_1, X_2, X_3, X_4: independent permutations of $1/n, 2/n, \ldots, n/n$
- Use standardized values of $f_j(X_j)$, $j = 1, 2, 3, 4$
- $\epsilon \sim \mathcal{N}(\sigma^2)$
- σ^2 adjusted such that $R^2 = 0.75$
- For $n = 300$: Additionally investigate $R^2 = 0.25, 0.5$
- Run `gam` with tp, cr and ps implementations (using default procedures)
- Defaults in **mgcv** ensure that all spline bases have the same dimensionality
- Evaluation: covariate-wise mean squared error, $\int_{x_j} (f_j - \hat{f}_j)^2 dP_{x_j}$

Summary of the Simulation Study

- Regression setting with reasonably large sample sizes
- Setting refers to "typical" predictor-response relationships, not too wiggly
- Uncorrelated predictors, no outliers in X

$\Rightarrow$ In this setting, **mgcv** defaults worked well

$\Rightarrow$ Differences between tp, cr and ps appear to be negligible

Common bases in gamlss

Table 5.1: Implemented **gamlss** additive functions

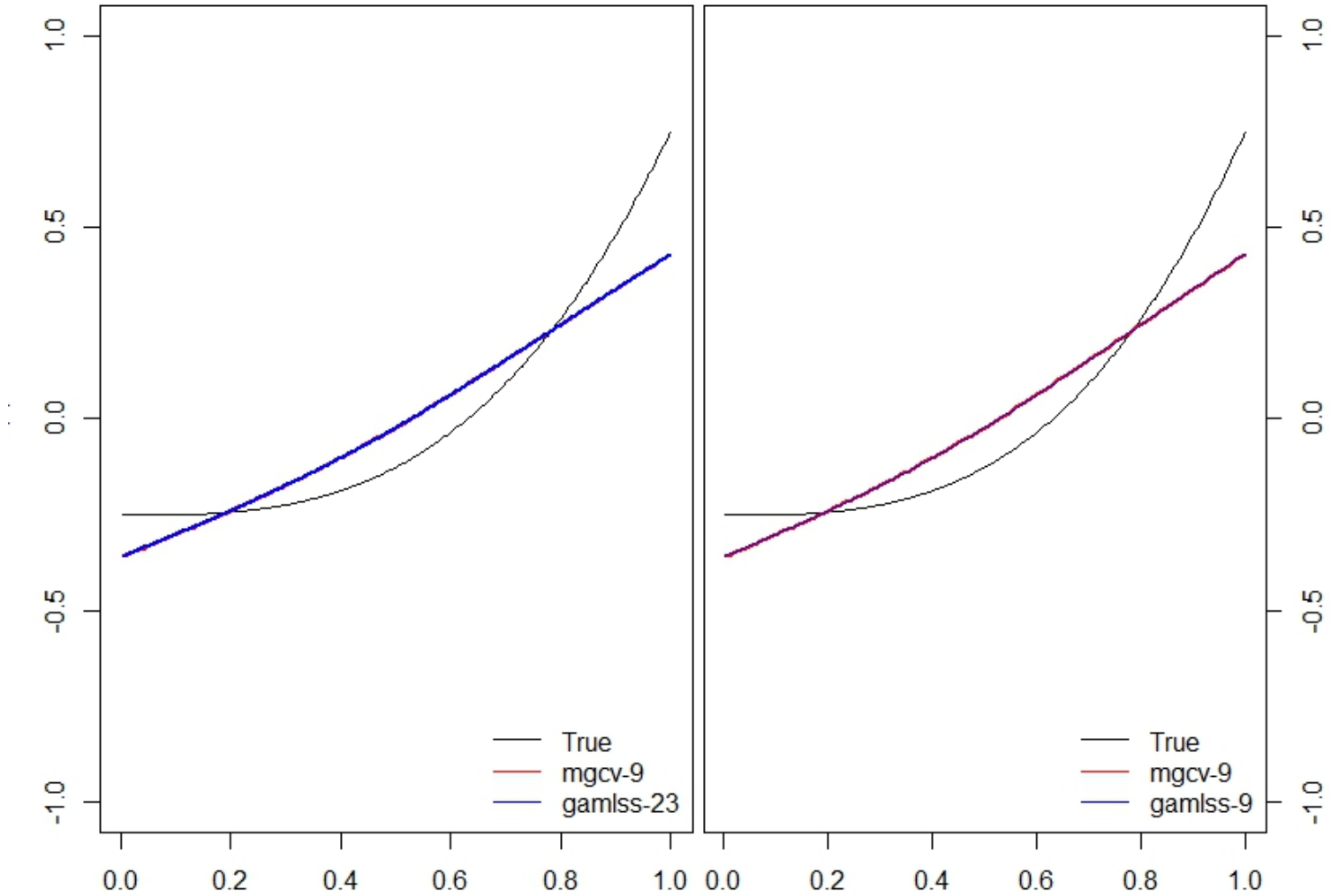
Additive terms	R Name	Section
Cubic splines	cs()	5.1
Varying coefficient	vc()	5.2
Penalized splines	ps()	5.3
`loess`	lo()	5.4
Fractional polynomials	fp()	5.5
Random effects	random()	5.6.1
Random effects	ra()	5.6.2
Random coefficient	rc()	5.6.3

P-Splines in gamlss

- Defaults in gamlss:
  - b-splines (degree 3) with second-order difference penalty
  - 20 inner knots and 2 boundary knots per smooth term (23 coefficients: # inner knots + degree 3 + 1 minus intercept)
  - Optimization of smoothing parameter via ML

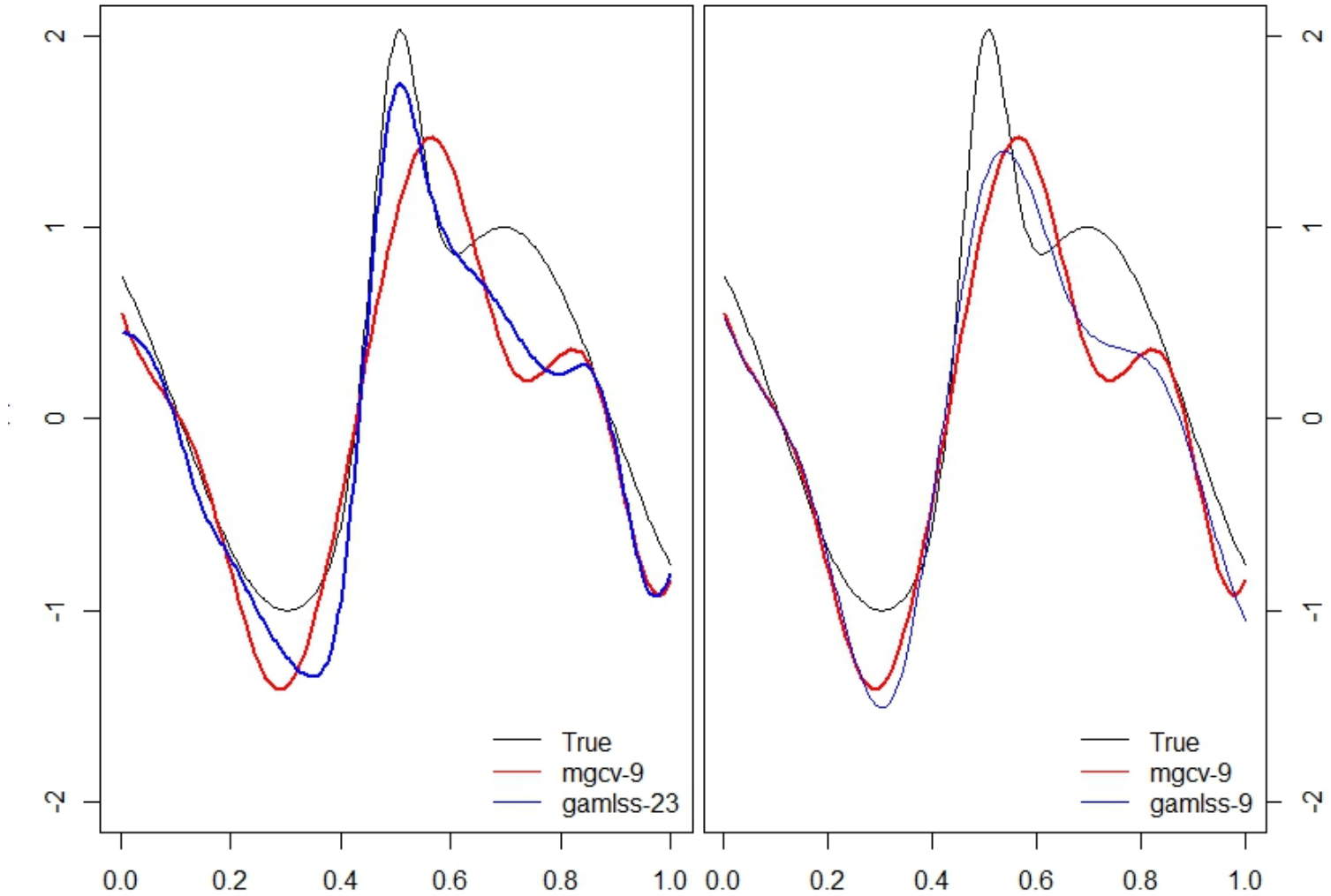
Simple comparison

- $Y = x^3 + \epsilon$
- left: default values, right: same # of knots



Simple comparison

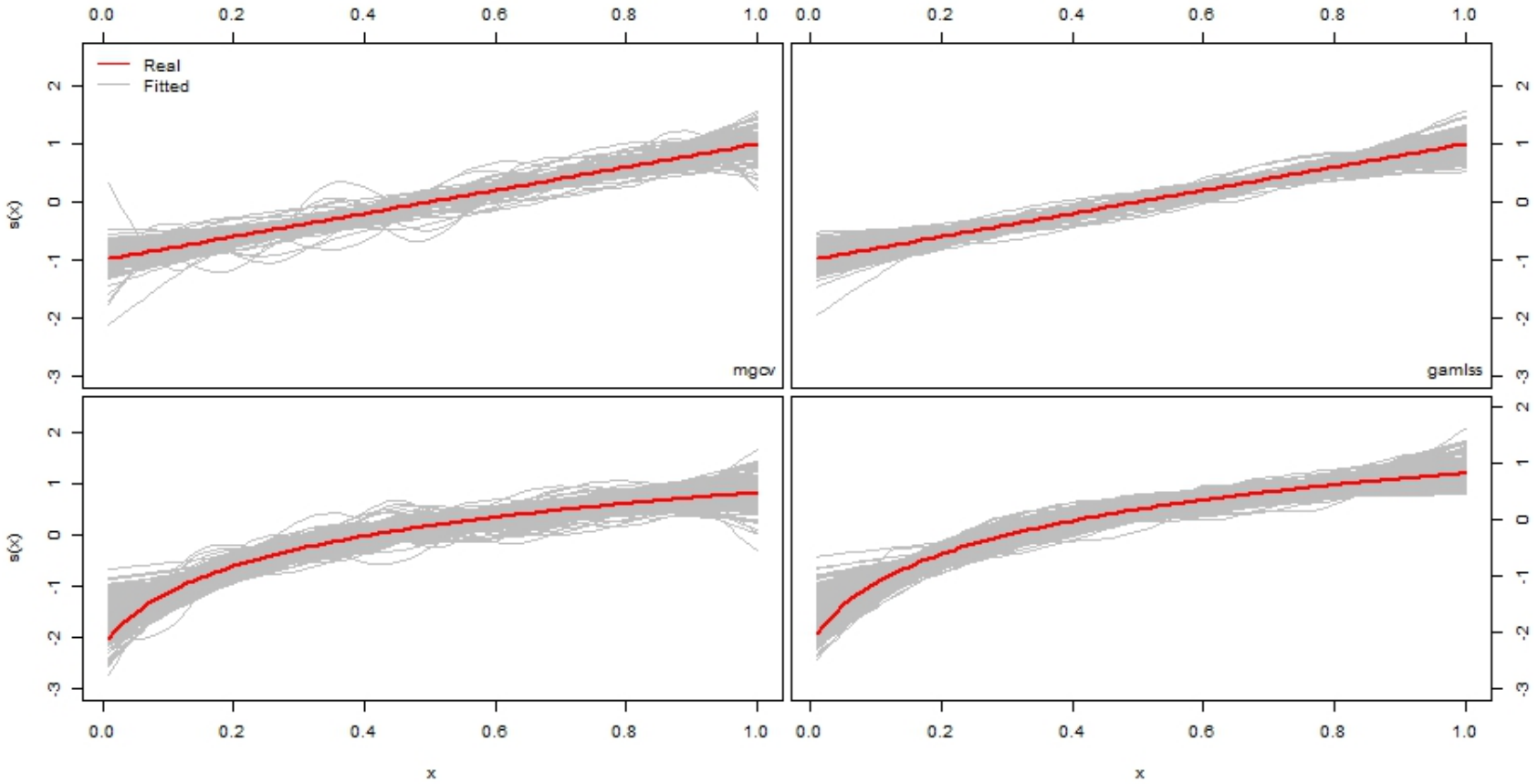
- $Y = \sin(4x - 2) + 2\exp(-\frac{16^2}{(x-.5)^2}) + \epsilon$
- left: default values, right: same # of knots



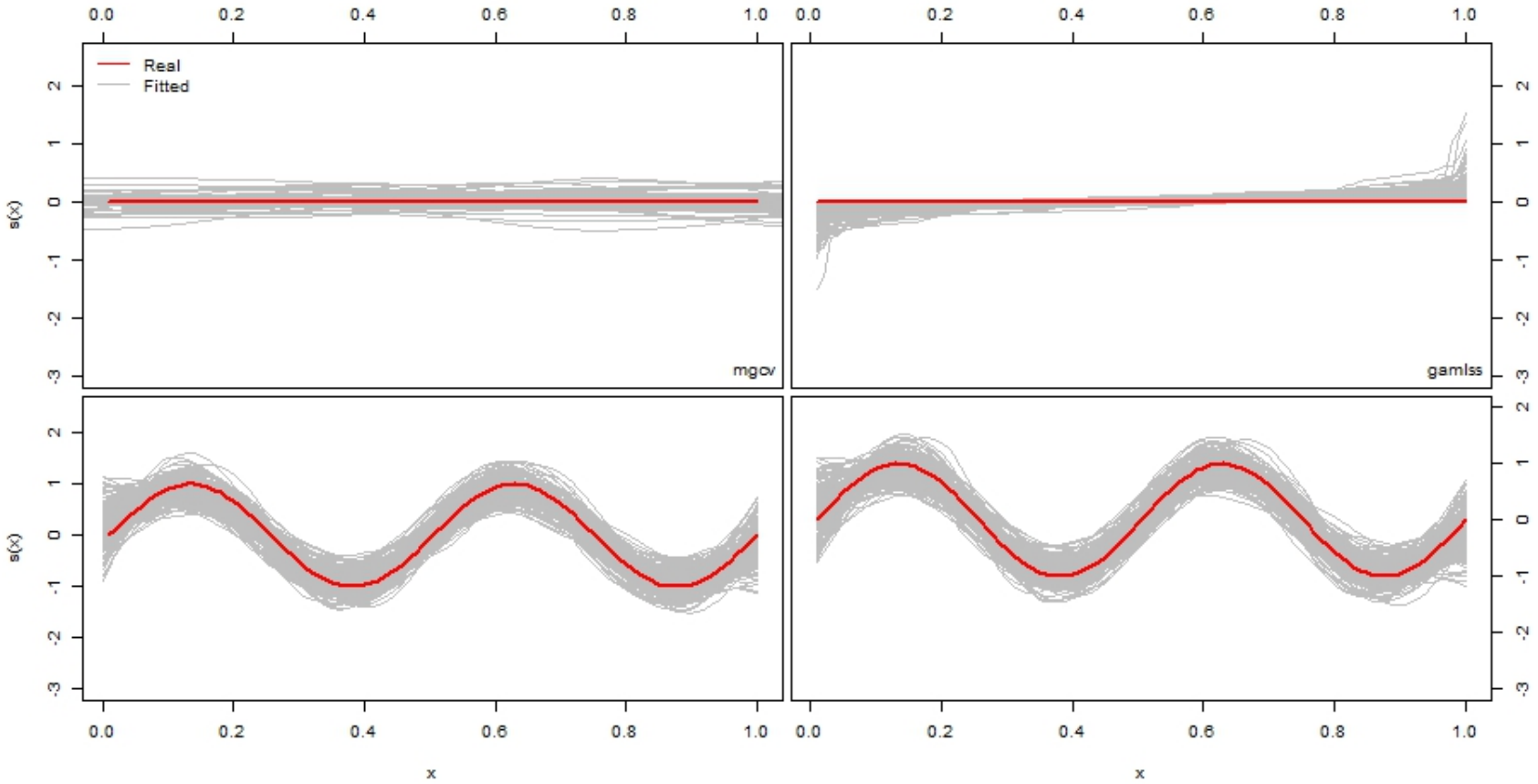
mgcv Vs gamlss

- Compare fits when using default values of mgcv vs gamlss
- mgcv:
  - Low-rank approximation of thin plate splines
  - 9 coefficients per smooth term
  - optimised with GCV
- gamlss:
  - b-splines (degree 3) with second-order difference penalty
  - 23 coefficients per smooth term
  - optimised with ML

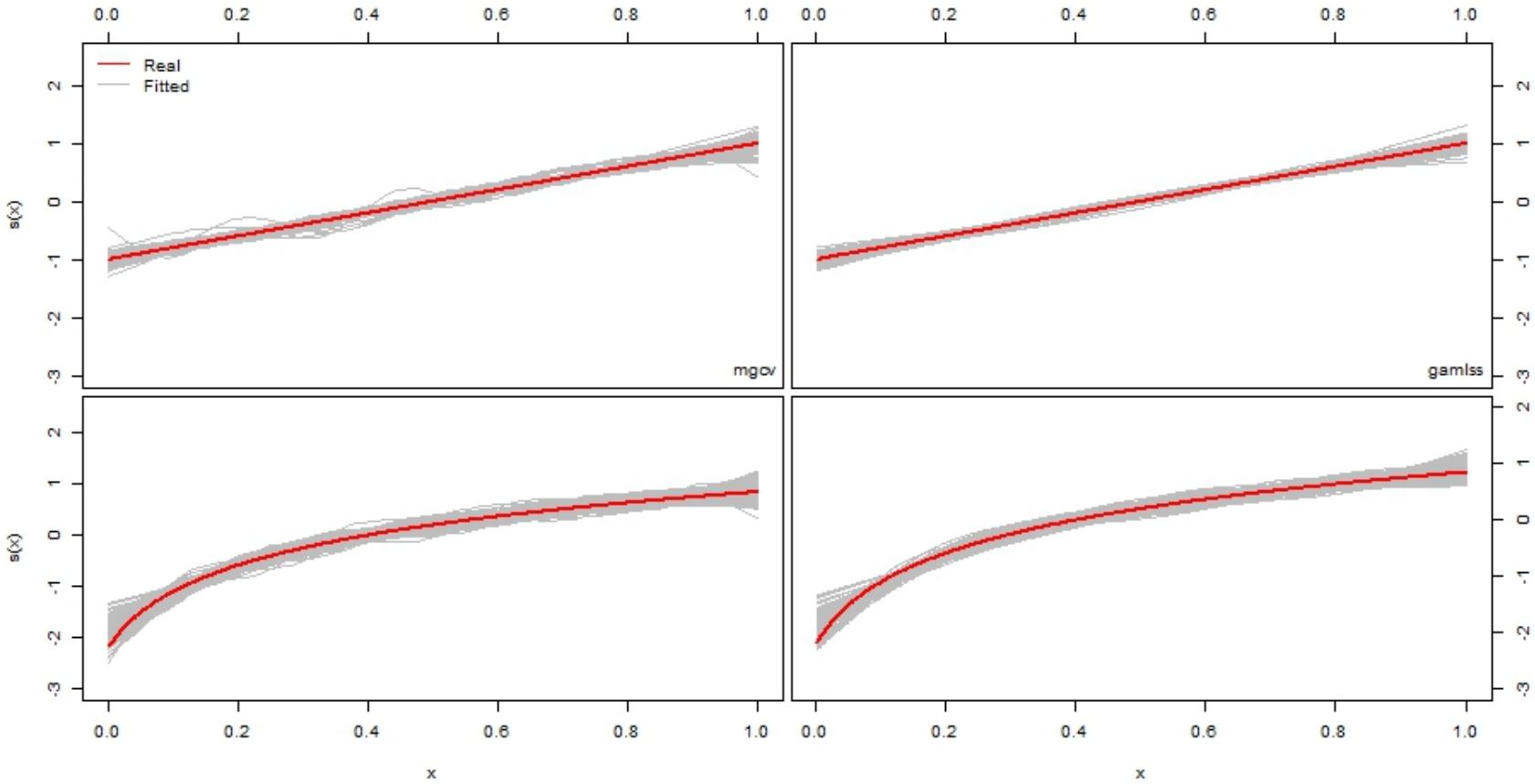
n=100



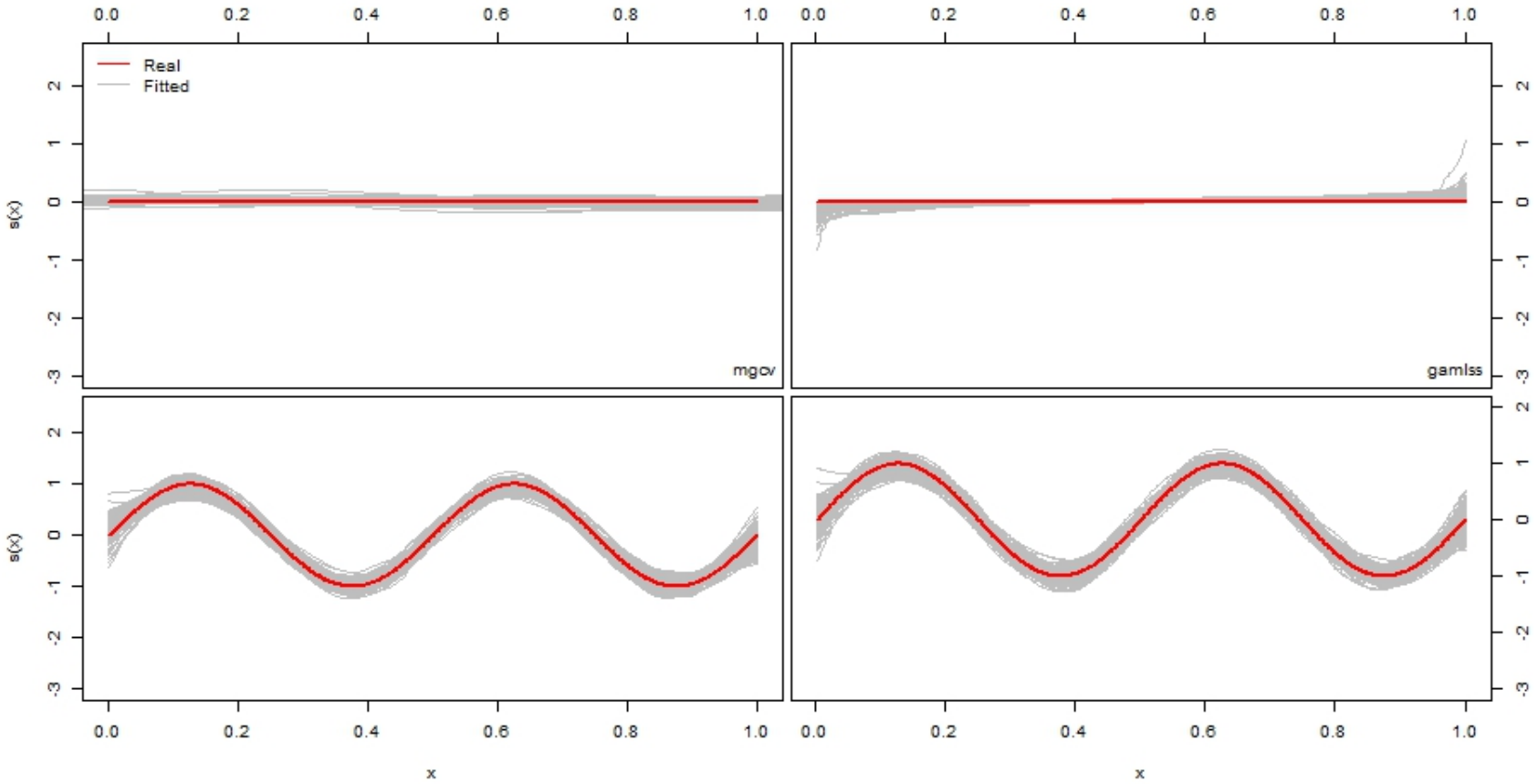
n=100



n=500



n=500



Discussion

- both pagkages give similar results
- however: p-splines seem more stable

Discussion

- Paper on R spline methods submitted for publication
- Paper on thorough comparison currently on the way
- Next steps: Correlated predictors, more noise variables, less smooth variable transformations