Multivariable Regression Modelling

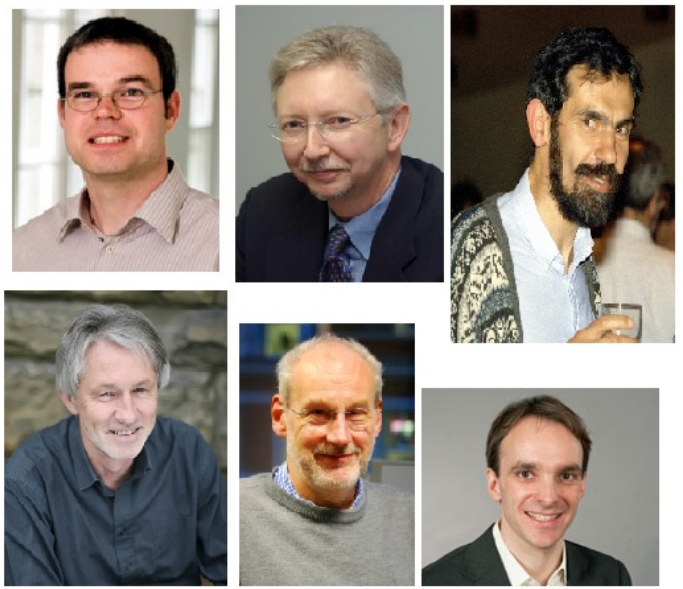
A review of available spline packages in R.

Aris Perperoglou for TG2

ISCB 2015

TG2 Members

- Michal Abrahamowic, Quebec Canada
- Willi Sauerbrei, Freiburg Germany
- Harald Binder, Mainz Germany
- Frank Harrell, Nashville USA
- Patrick Royston, London UK
- Matthias Schmid, Bonn Germany
- Aris Perperoglou, Colchester UK



Outline

- Introduction: Topic Group 2 and the R splines project
- Splines: from bases to software
- R packages
- Some results
- Discussion

Introduction: TG2

TG2: Selection of variables and functional forms in multivariable analysis

- The main focus of TG2 is to **identify** influential variables and gain insight into their individual and joint relationship with an outcome of interest.
- Which variables are related with the outcome and how to:
  - **choose** relevant predictors
  - determine **functional form** of continuous predictors.

General regression model

- Consider an outcome of interest y and let $x_1, x_2, ..., x_p$ some information on p covariates

$$g(y) = f(x_1, \beta_1) + f(x_2, \beta_2) + ... + f(x_p, \beta_p)$$

- In practice:
  - Most often a linear relationship of x with $g(y)$ is assumed
  - Still people tend to categorise covariates. See **Greenland 1995, Royston, Altman & Sauerbrei 2006** for reasons against it.
  - Sometimes a simple transformation might be considered, such as $\log(x), x^2$...
  - Fractional Polynomials **Royston & Altman 1994**
  - Splines

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**)
  - b-spline (**de Boor 1978**), p-spline (**Marx and Eilers 1996**), penalized regression spline (**Wood 2006**)
  - smoothing splines (see Generalized Additive Models (**Hastie and Tibshirani 1990**))...

Decisions

- Although splines are a powerful tool, in practice the number of decisions that the user has to make complicate the modelling problem.
  - Type of spline
  - Number of knots
  - Position of knots
  - Order of the spline (complexity)

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**.

R: Identify and assess methods used in practice

1. Available software (focus on R): Identify *all* packages within R that provide the ability to use splines as a functional transformation in univariable and/or multivariable analysis.
2. Collect information on packages.
3. Compile documentation for practical use.
4. Evaluate quality of available packages.
5. Compare approaches, further guidance for use in practice.

Why R

- R is the most widely spread programme amongst research statisticians, e.g. people that do research in statistics. Also dominates the field of *data science*.
- It is free and comes with an open software licence, which is use at your own cost.
- Although CRAN requires the basic documentation of all functions contained in submitted packages, help files are often not detailed enough to fully understand how the implemented methods work.

Spline basics

Preliminary idea

- Consider the straight line regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

- The corresponding **basis** for the model is:

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

- A polynomial basis can also be:

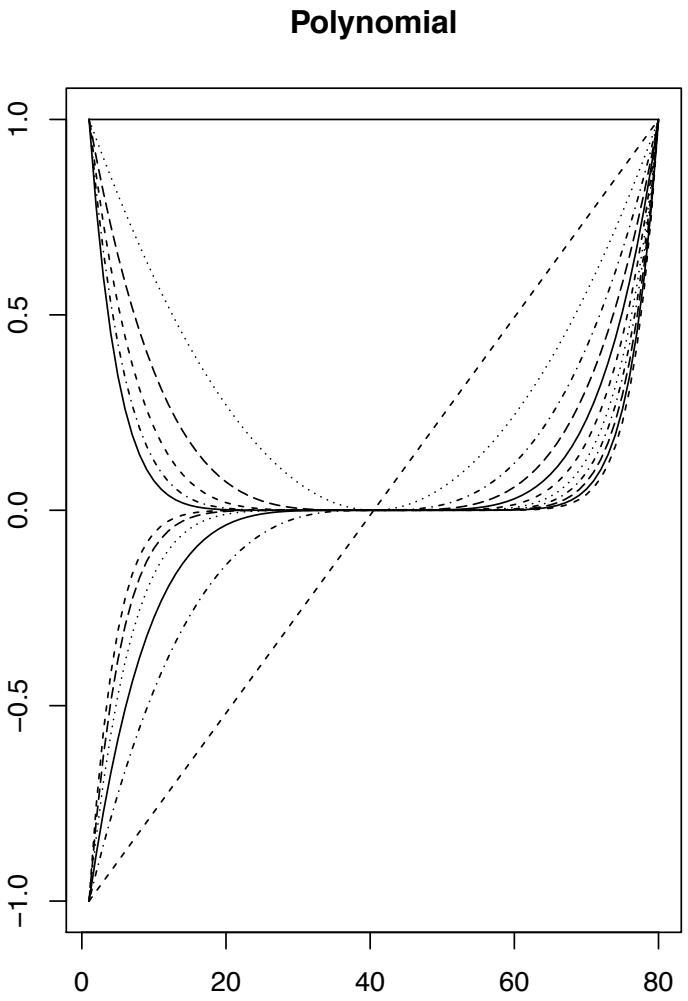
$$U = \begin{bmatrix} 1 & u_1 & u_1^2 & u_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & u_n & u_n^2 & u_n^3 \end{bmatrix}$$

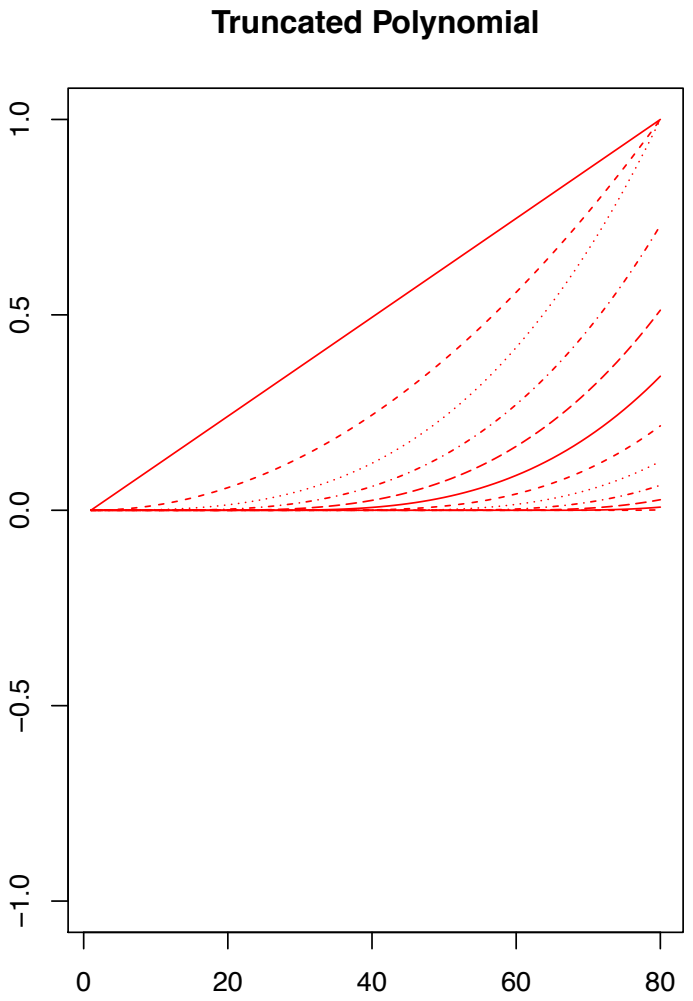
- where $u = \frac{x - \min(x)}{\max(x) - min(x)}$

Spline basics: From bases to software

```
pbase <- function(x, p) {
  u <- (x - min(x)) / (max(x) - min(x))
  u <- 2 * (u - 0.5);
  P <- outer(u, seq(0, p, by = 1), "^")
  P }
```

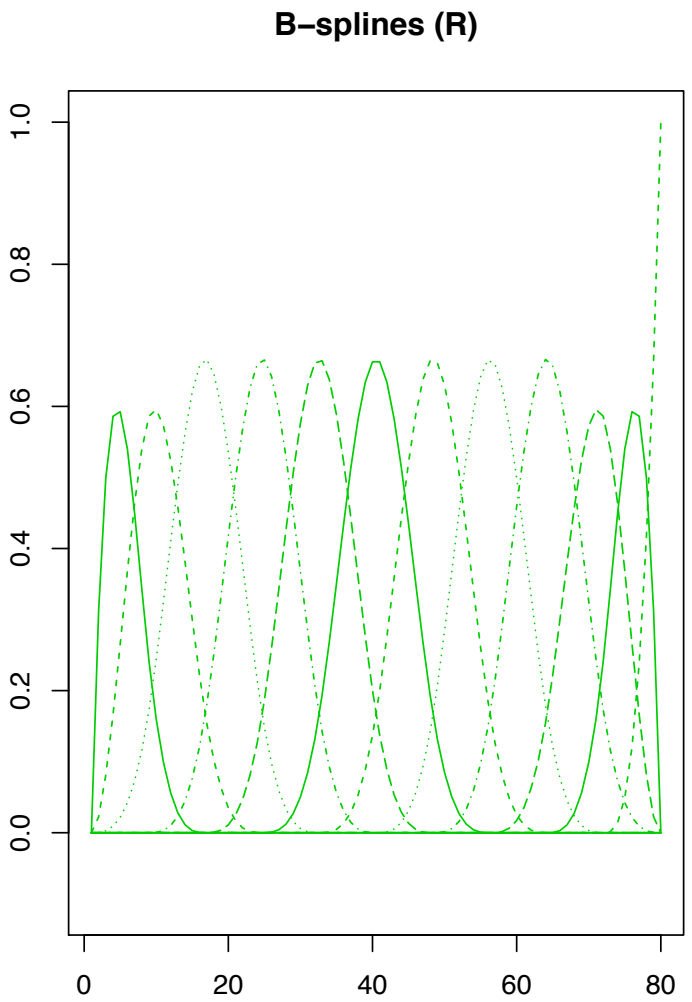
Spline basics: From bases to software

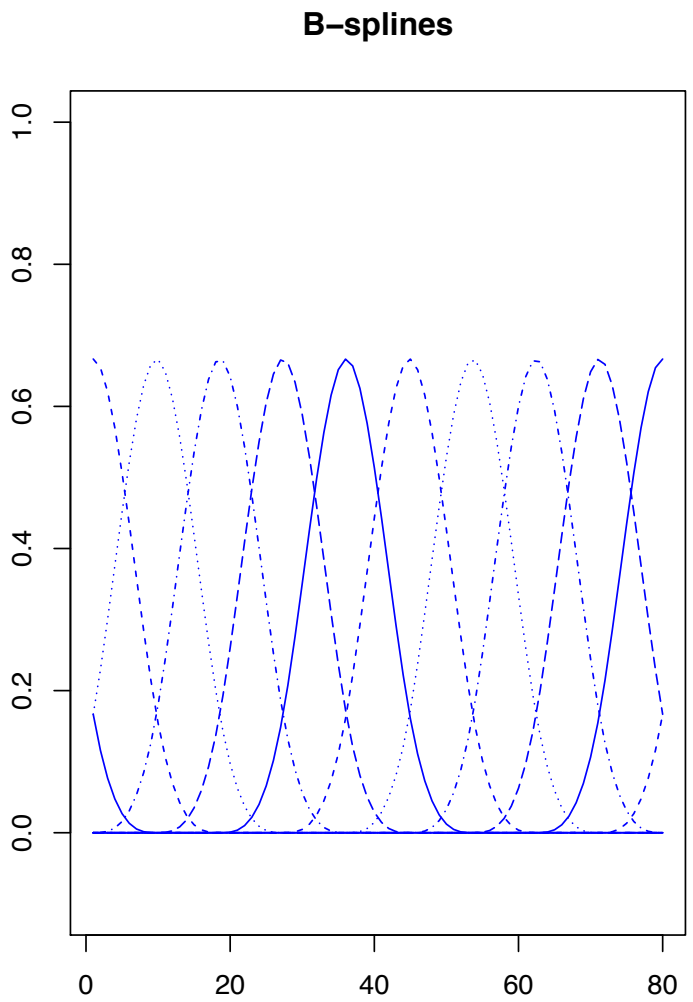




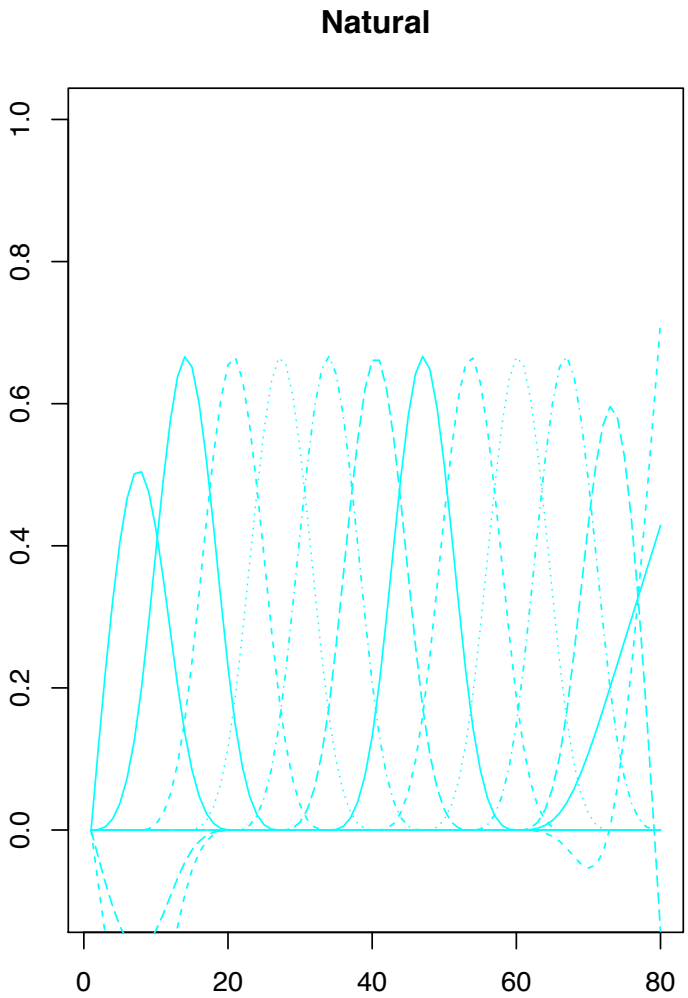
Spline basics: cubic B-splines with 12 knots

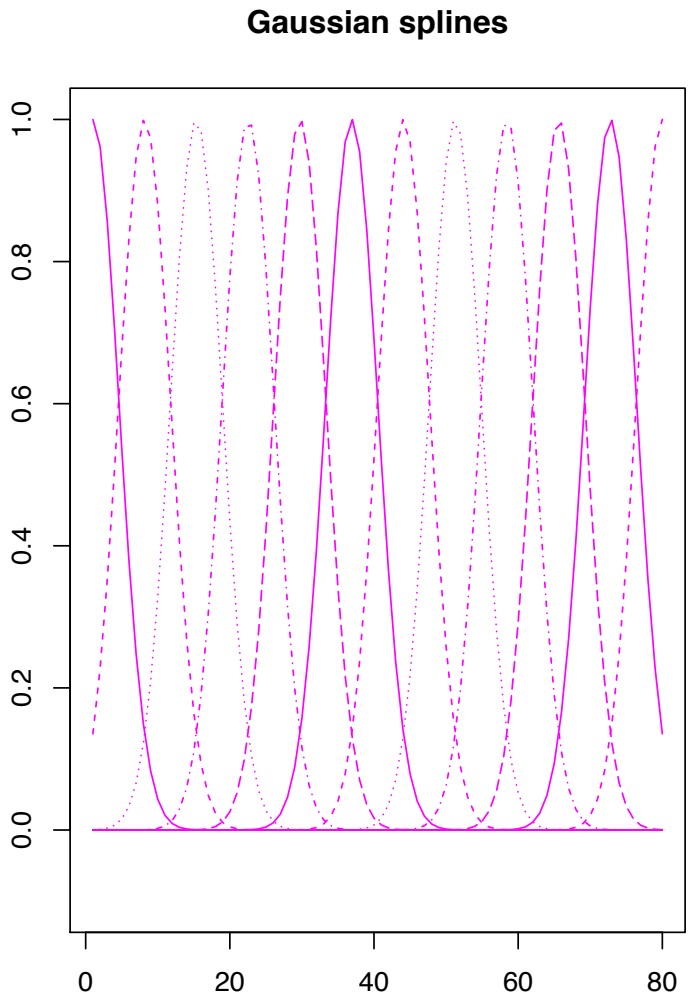
- B-splines are a common choice





Spline basics: Natural and Gaussian Splines (12 knots)





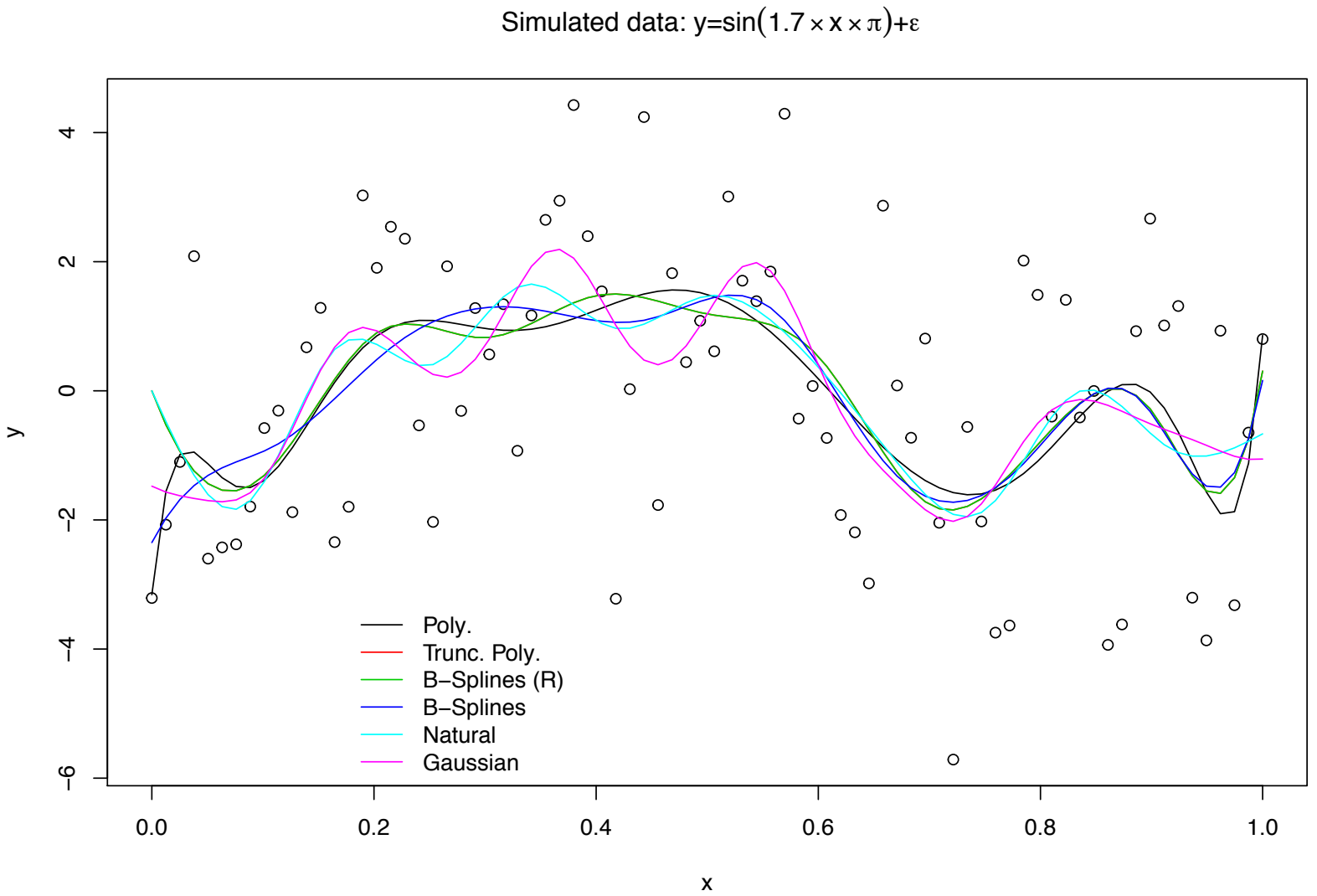
Least square fit

- A general spline model with k knots forms an X matrix basis.
- The ordinary least squares fit can be written as:

$$\hat{y} = X\hat{\beta} \quad \text{where} \quad \widehat{\beta} \quad \text{minimizes} \quad ||\mathrm{y} - \mathrm{X}\beta||^2$$

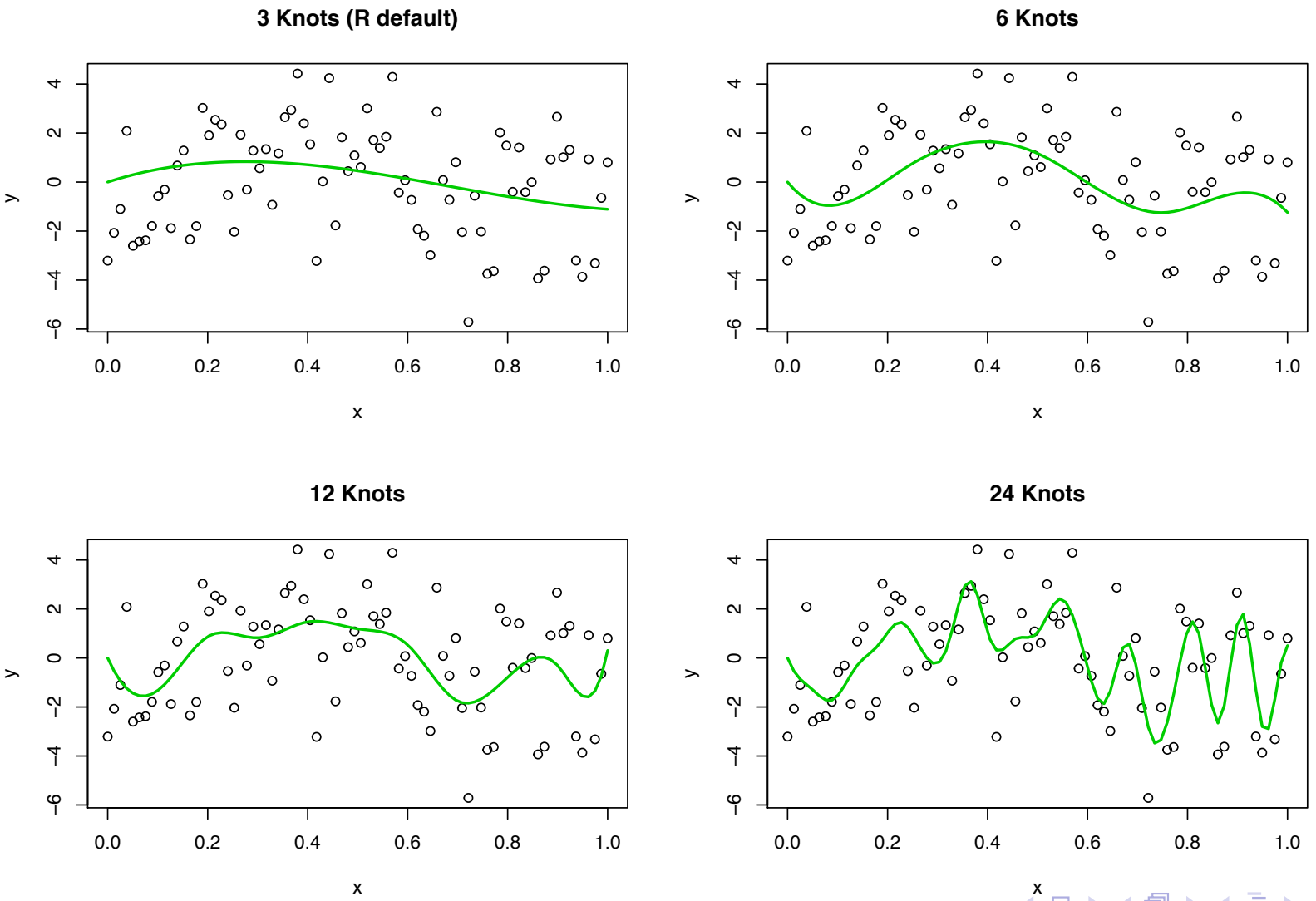
- See **Ruppert, Wand and Carroll 2003** for a good overview of univariable analysis.

Choice of splines: Differences when fitting data

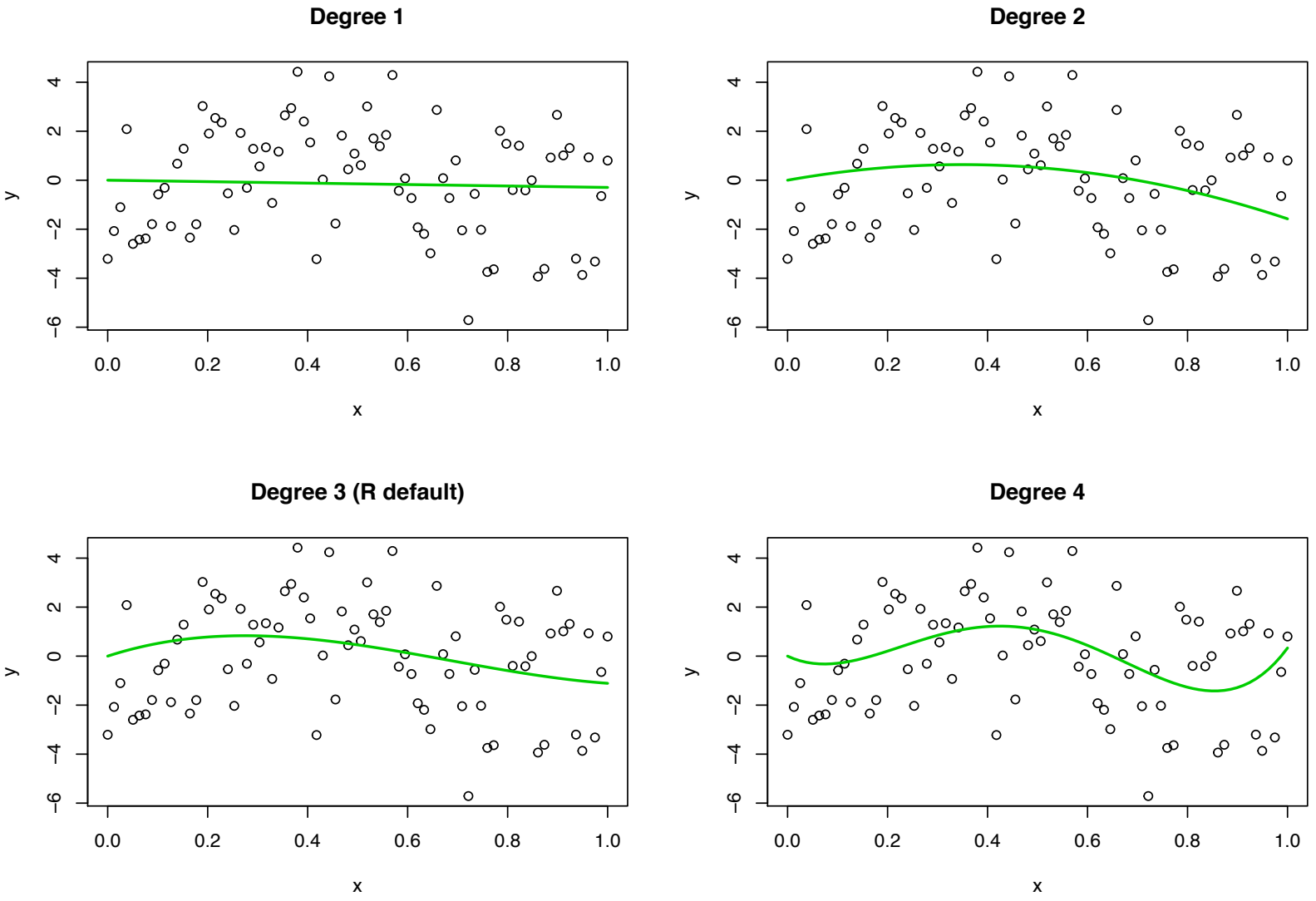


Choice of knots: cubic B-splines fit

- Number of knots will lead to different fit:



Choice of complexity: b-splines with 12 knots

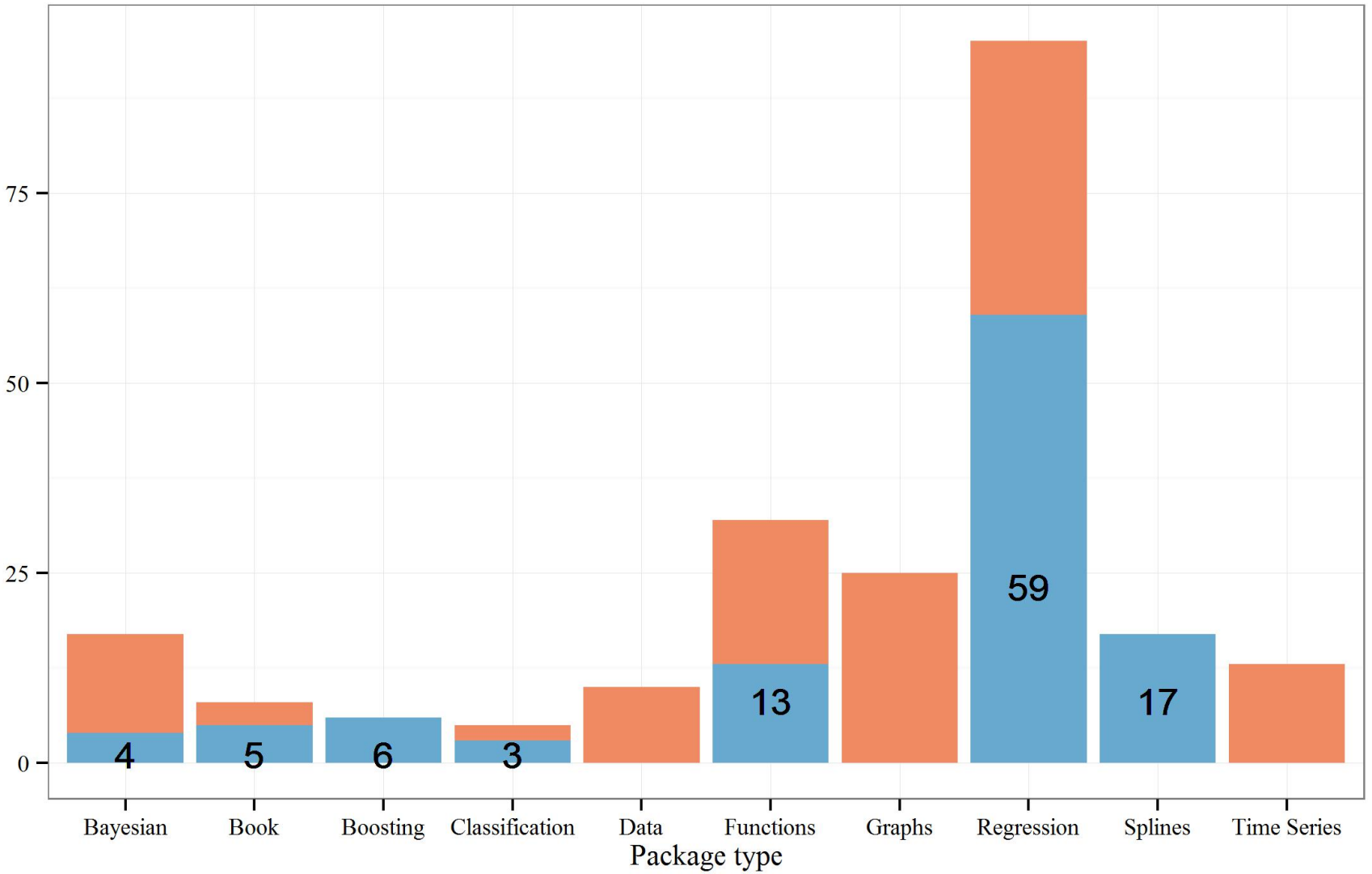


The search for R Packages

R packages

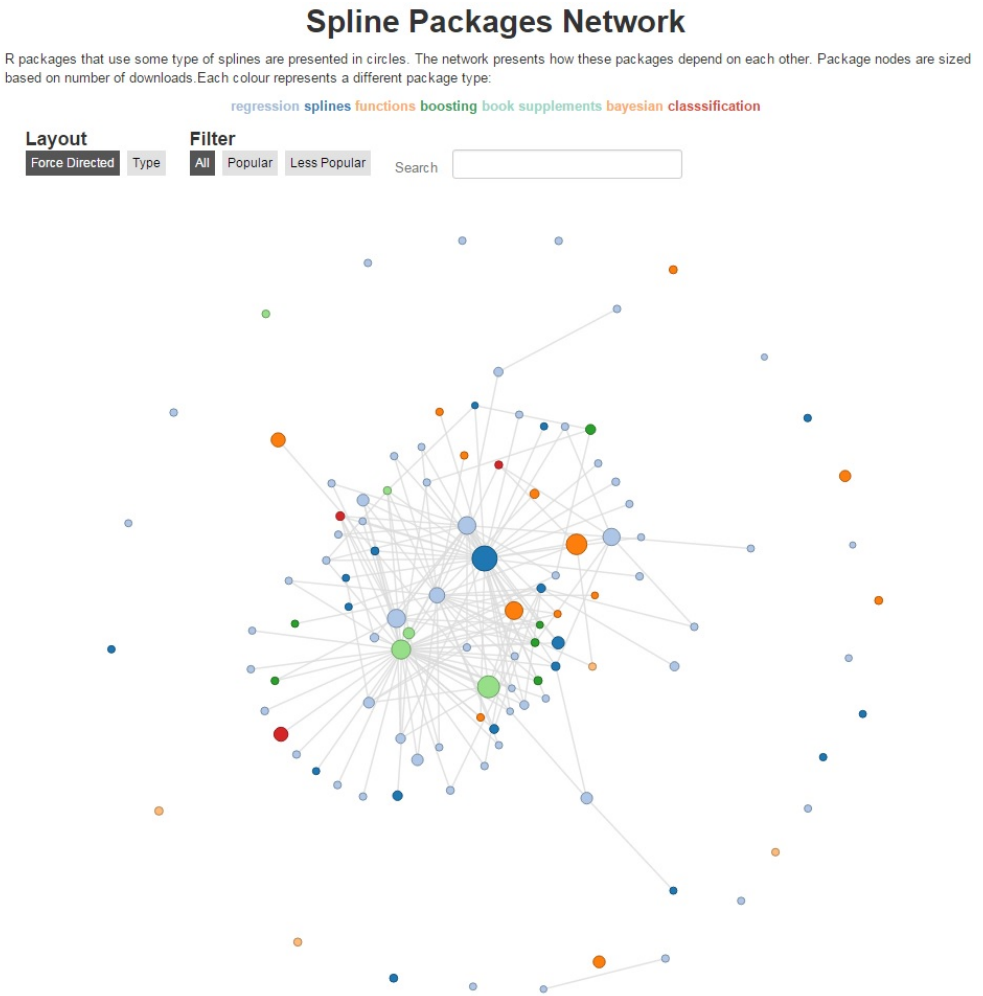
- More than 6200 packages available on CRAN.
- Even more available on GITHUB, RGforce etc that we do not look into.
- Used a java programme to scan CRAN and identify all packages that have a vague relation to splines.
- A total of 519 packages were identified (May 2015) out of which 229 were flagged as relevant

Subset of 109 packages were selected for further analysis



Interdependencies

http://drperpo.github.io/docs/rpackages/splines.html



What makes a good package?

- Packages with vignettes
- Packages with a web site companion/book/documentation
- Packages with real life data
- Packages with clear examples of methods
- Packages with clear references

Out of selected: packages with vignettes

- Bayesian: 1
- Boosting: 2
- Classification: 1
- Functions: 4
- Regression: 14
- Splines: 1

Results (1): Spline creators

Packages, downloads and reverse dependencies (RD)

package	down	RD	Description
splines		124	Regression spline functions and classes
polsplines	35419	0	Polynomial spline routines
logspline	19513	3	Logspline density estimation routines
cobs	11596	2	Constrained B-Splines
pencopula	4892	0	Copula Estimation with P-Splines
orthogonalsplinebasis	3959	0	Orthogonal B-Spline Functions
pbs	3790	0	Periodic B Splines
bezier	3546	1	Bezier Curve and Spline Toolkit

- Downloads does not mean unique users.

Packages, downloads and reverse dependencies (RD)

package	down	RD	Description
gss	109259	5	General Smoothing Splines
pspline	23313	6	Penalized Smoothing Splines
MBA	15689	6	Multilevel B-spline Approximation
crs	11288	2	Categorical Regression Splines
cobs99	5859	0	Constrained B-splines – outdated version
Kpart	4747	0	Spline fitting
freeknotsplines	4228	0	Free-Knot Splines
bigsplines	4184	1	Smoothing Splines for Large Samples
sspline	3955	0	Smoothing Splines on the Sphere
episplineDensity	1042	0	Density Estimation Exponential Epi-splines

Results (2): Regression packages

General features of popular regression packages

package	downloaded	vignette	book	website	datasets
mgcv	380232	x	x		2
quantreg	347203	x	x		7
survival	243212	x	x		33
VGAM	92091	x	x	x	50
gbm	86729			x	3
gam	62916		x	x	1
gamlss	27084	x	x	x	29

- VGAM and gamlss have separate data packages.

Regression models per package

response	mgcv	quantreg	survival	VGAM	gbm	gam	gamlss
Linear	x	x		x	x	x	x
Categorical	x			x	x	x	x
Count Reg	x	x		x	x	x	x
Survival	x	x	x	x	x	x	x
Quantile Reg				x	x	x	x
Multivariate	x			x		x	x
Nonlinear				x		x	x
Reduced Rank		x		x		x	
Other	x	x		x	x	x	x

Can the package fit penalized splines and random effects

package	p- splines	RE
mgcv	x	x
quantreg	x	x
survival	x	x
VGAM	x	x
gbm	x	
gam	x	
gamlss	x	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

Variable selection in other packages

- Some packages include functions to perform variable selection: `leaps`, `rms`, `gamboostLSS`, `subselect`, `glmulti`, `glmulti`, `spikeSlabGAM`,`spikelab`
- The list is not exhaustive.

Post fit functions, extract coefficients, predictors, plot terms and other useful functions

package	coeff/fitted	plot terms	conf.int
mgcv	x	x	x
quantreg	x	x	
survival	x	x	x
VGAM	x	x	x
gbm	x	x	
gam	x	x	x
gamlss	x	x	x

Some more questions:

- Are examples based on real or simulated datasets?
- Are results for the examples nicely and understandable presented?
- Give details about the most important default values. Any reason for the default values given?

Outcomes

- A paper with preliminary results within a year
- Further research into the quality of packages, advantages/disadvantages, worked out examples etc to follow

References: Functional forms and Splines

- Greenland, S. (1995). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology, 6(4), 450-454.
- Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med, 25(1), 127-141.
- Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Applied Statistics, 429-467.
- De Boor, C. (1978). A practical guide to splines. Mathematics of Computation.
- Harrell, F. E. (2013). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer Science & Business Media.
- Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. Statistical science, 89-102.
- Wood, S. (2006). Generalized additive models: an introduction with R. CRC press.
- Hastie, T. J., & Tibshirani, R. J. (1990). Generalized additive models (Vol. 43). CRC Press.
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric regression (No. 12). Cambridge university press.

References: Modelling and R packages

- Wood, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. R news, 1(2), 20-25.
- Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: B (Statistical Methodology), 65(1), 95-114.
- Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society: B (Statistical Methodology), 70(3), 495-518.
- Yee, Thomas W., and C. J. Wild. "Vector generalized additive models." Journal of the Royal Statistical Society. B (Methodological) (1996): 481-493.
- Yee, T. W. (2010). The VGAM package for categorical data analysis. Journal of Statistical Software, 32(10), 1-34.
- Therneau, T. M., & Grambsch, P. M. (2000). Modeling survival data: extending the Cox model. Springer Science & Business Media.
- Koenker, R. (2005). Quantile regression (No. 38). Cambridge university press.
- Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society:C (Applied Statistics), 54(3), 507-554.
- Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, 23(7), 1-46.