The GeDS R package: Geometrically Designed Variable-Knot Splines in the context of GLM(GNM) modelling

Andrea Lattuada[1]

DEAMS, University of Trieste

June 8, 2017

[1] Joint work with Dimitrina S. Dimitrova, Vladimir K. Kaishev and Richard J. Verrall

Summary

The problem

- In several situations, data $\{(y_i, \boldsymbol{z}_i)\}_{i=1}^n$ should be modelled according to

$$E(y|\boldsymbol{z}_i) = f(\boldsymbol{z}_i)$$

or, more generally, to

$$E(y|\boldsymbol{z}_i) = \mu_i \text{ and } g(\mu_i) = f(\boldsymbol{z}_i)$$

- We will concentrate here on the univariate case, hence we will use in the remainder z instead of $\boldsymbol{z}$
- If the 'shape' of f is known, it can be estimated within the parametric framework.
- The characteristics of f are identified by a finite number of parameters and the estimation procedure takes place in a finite dimensional space

Splines

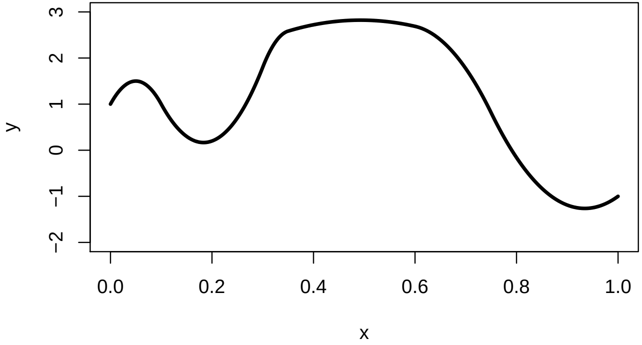
If f is unknown, it can be estimated in an infinite dimensional space in the non-parametric framework or in a finite (but high) dimensional space

By means of Taylor expansion any smooth function can be locally approximated by a Polynomial curve

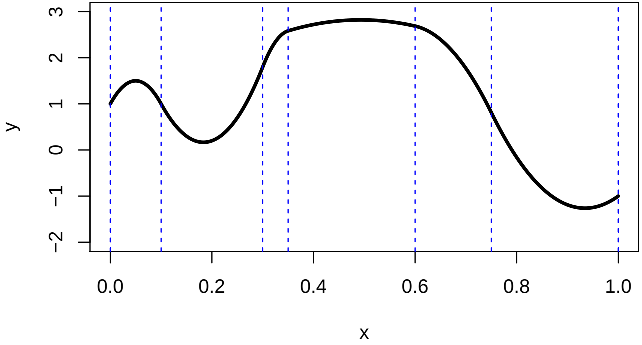
Definition: a function $f : [a, b] \to \mathbb{R}$ is an lth order polynomial spline defined on the knots $\{t_j\}_{j=1}^{m}$ such that $a = t_1 \leq \cdots \leq t_m = b$ if:

- it belongs to $\mathcal{C}^{(l-2)}$
- it is a polynomial of degree $l-1$ in $[t_j, t_{j+1}]$

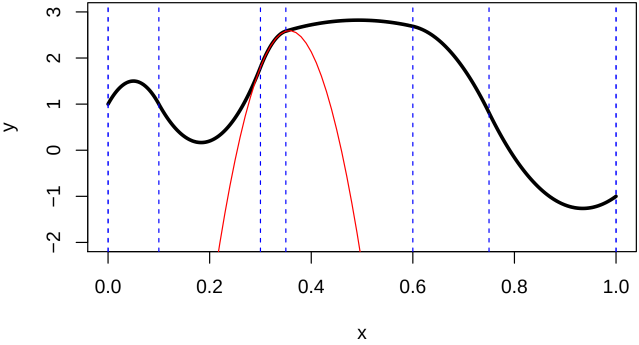
Some basic properies



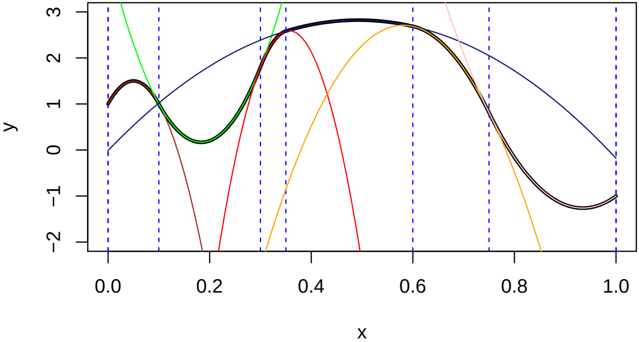
Some basic properies



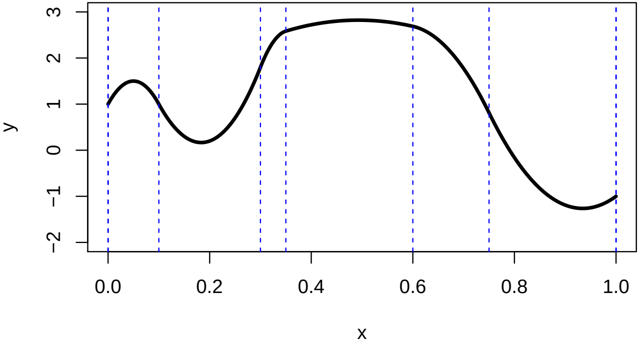
Some basic properies



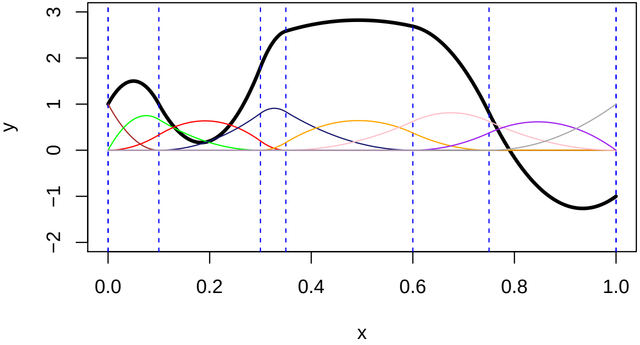
Some basic properies



Some basic properies



Some basic properies



Basis Splines

Representing a spline as a linear combination of basis functions $B_{j,l}$, we have

$$f(z) \approx f^*(z) = \sum_j \beta_j B_{j,l}(z)$$

and one can estimate $\boldsymbol{\beta}$ via standard regression tools such as

- Least Squares under the assumptions of the classical Linear Model
- Maximum Likelihood if we are in the GLM framework

Basic B-spline properties

The set of all the lth order splines defined on the knots $\boldsymbol{t} = \{t_j\}_{j=1}^{m}$ is the space $S_{\boldsymbol{t},l}$. A basis of this space is

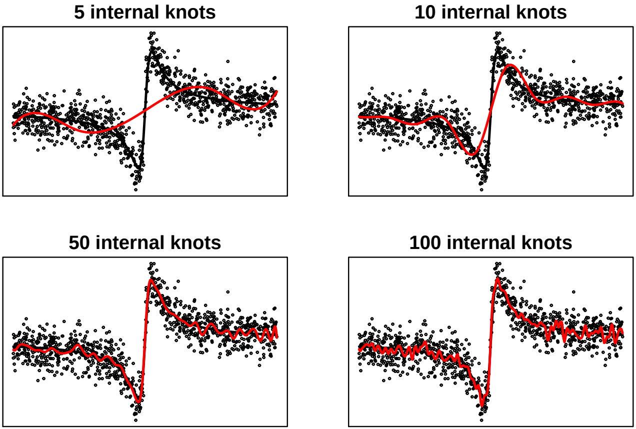
$$N_{j,l}(z) := (t_{j+l} - t_j)[t_j, \ldots, t_{j+l}](\cdot - z)_+^{l-1} \tag{1}$$

$\{N_{j,l}\}_{j=1}^{m-l}$ are called B-splines

- Non-negative: $N_{j,l}(z) \geq 0$
- Local support: $N_{j,l}(z) = 0$ iff $z \notin [t_j, \ldots, t_{j+l}]$
- Partition of the unity: $\sum_{j=1}^{m-l} N_{j,l}(z) = 1$

Semiparametric Regression

If $\{t_j\}_{j=1}^m$ is set ex ante, the estimation procedure takes place in a finite dimensional parameter space and the problem becomes a problem of Semi-Parametric regression



Penalized regression

In general, the higher the number of bases, the wigglier will be the estimated curve

The issue can be addressed introducing a penalization proportional to a measure of wiggliness in the estimating procedure

However there are some open questions

- How to measure the wiggliness of a function
- How to choose the penalization
- Is the choice of the same penalization on the whole domain a limitation?

The penalization can be chosen via

→ visual inspection of the results

→ ML-REML approaches, taking advantage of the mixed model representation

→ model selection criteria, such as GCV, AIC and UBRE

→ again in the mixed model representation, but in the Bayesian framework

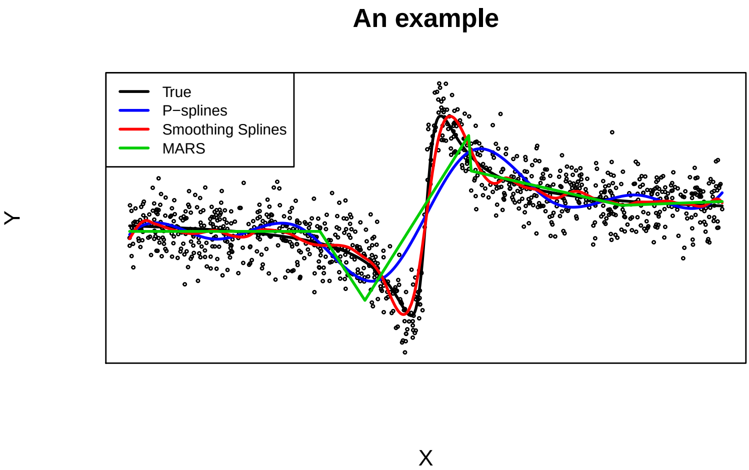
Non-Parametrc framework I

Several procedures help in producing estimates in the fully Non-Parametric framework

- Local regression procedures, such as Loess, Kernel Smoothing, Nearest Neighbour
- Adaptive procedures
- Smoothing splines (e.g. Gu, 2014)

In general all the methods in which knots are not set ex-ante and the number of parameters of the resulting fit is not known in advance

Example



Control Polygon I

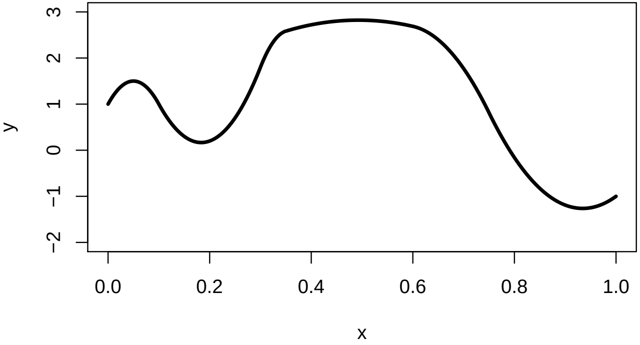
The shape of a spline is controlled by its Control Polygon, i.e. the polygon whose vertices are $\{(\xi_j, \beta_j)\}_{j=1}^{p}$ with

$$\xi_j = \frac{t_{j+1} + \cdots + t_{j+l-1}}{l-1}$$

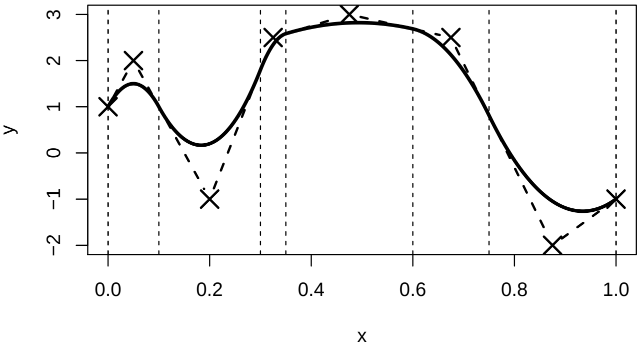
Properties:

- Convex hull property
- The spline follows the shape of the polygon
- The spline is a variation diminishing approximation to its polygon

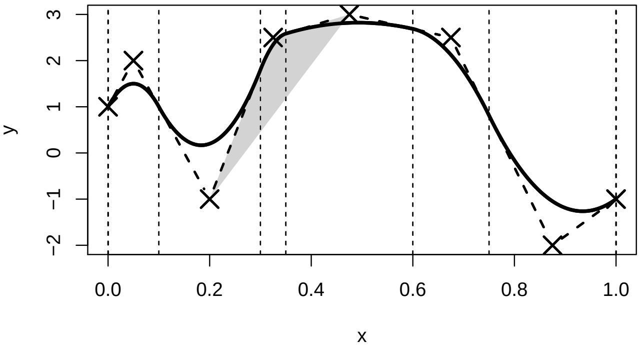
Control Polygon II



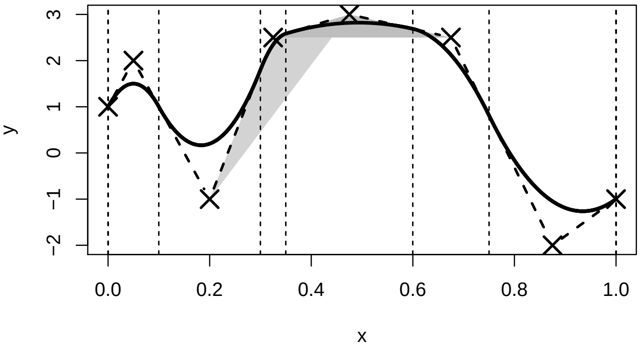
Control Polygon II



Control Polygon II



Control Polygon II



GeDS

Geometrically Designed Spline Regression is a methodology that allows to perform spline regression in an adaptive way.
Parameters estimated by the method are:

- the knot locations $\{t_j\}_{j=1}^{m}$
- the coefficients $\boldsymbol{\beta}$
- the order of the spline l

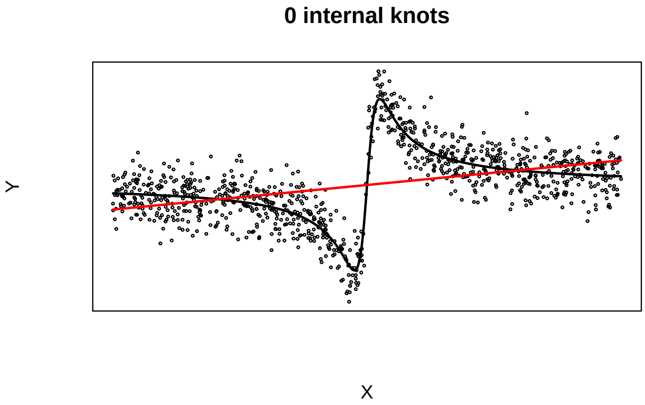
The algorithm is composed of two stages:

Stage A where f is estimated via a second order spline

Stage B where from the second order spline, higher order splines are computed

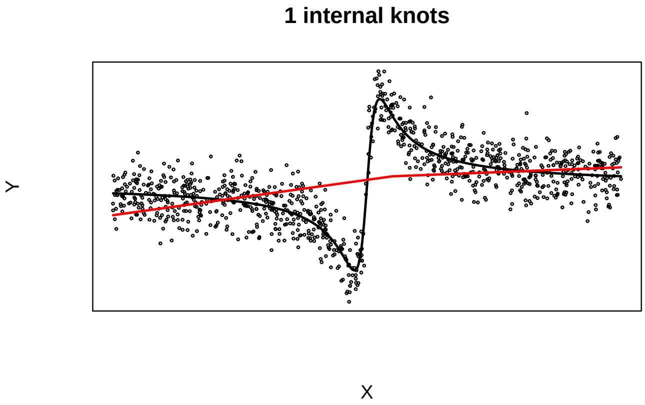
Stage A

Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



Stage A

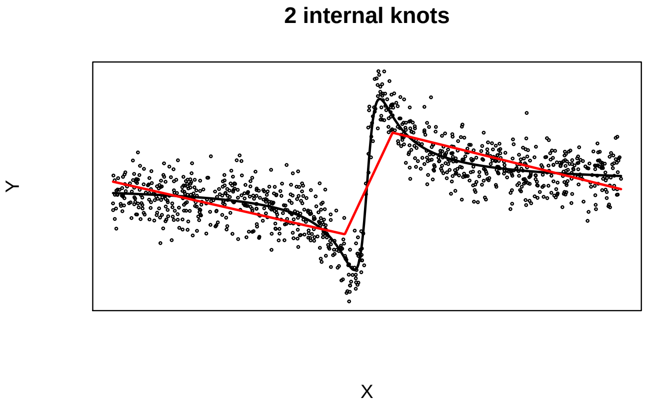
Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



Stage A

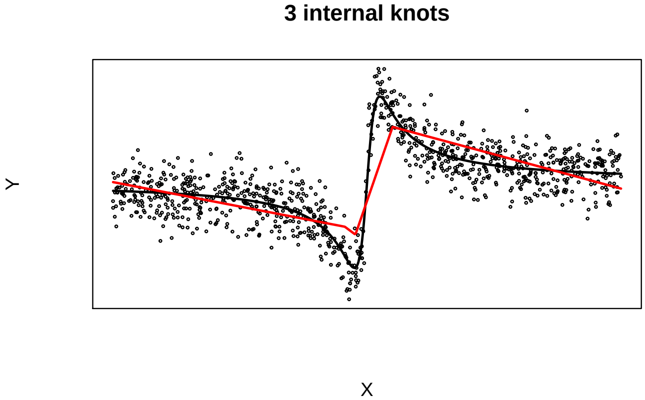
Stage A embeds a knot addition scheme

Starting from 2 couples of boundary knots, new knots are sequentially added



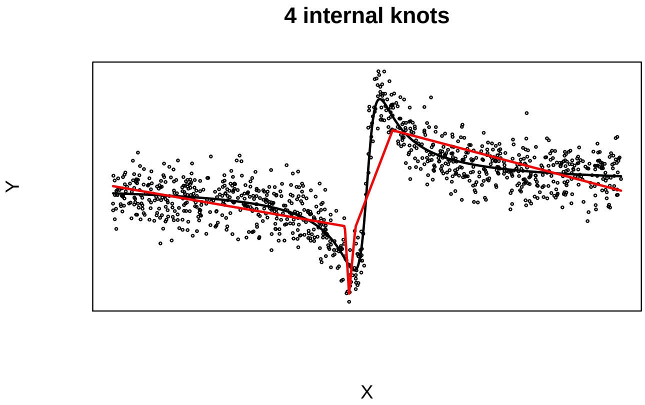
Stage A

Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



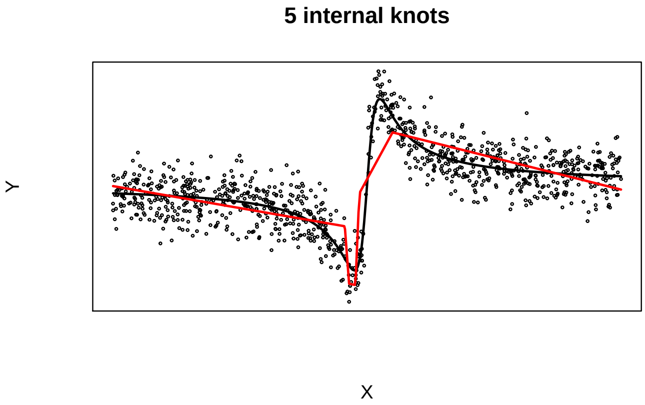
Stage A

Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



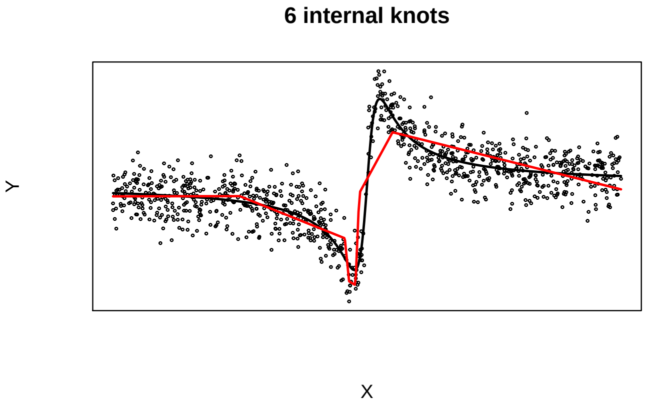
Stage A

Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



Stage A

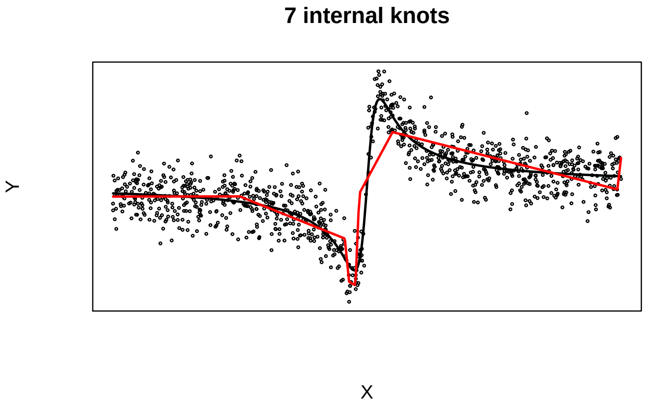
Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



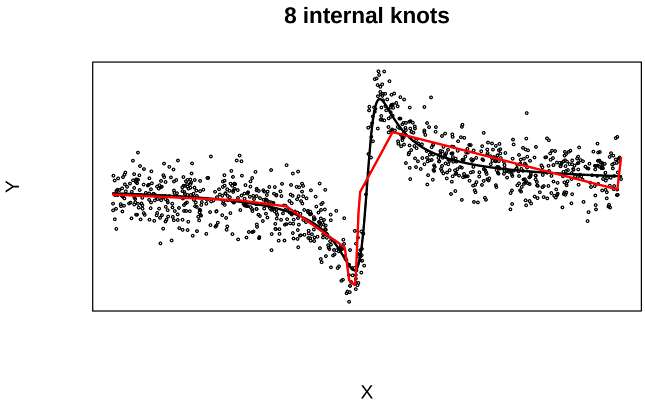
Stage A

Stage A embeds a knot addition scheme

Starting from 2 couples of boundary knots, new knots are sequentially added



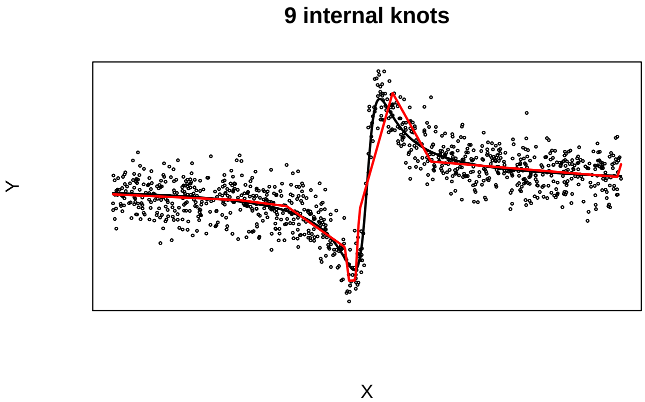
Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



Stage A

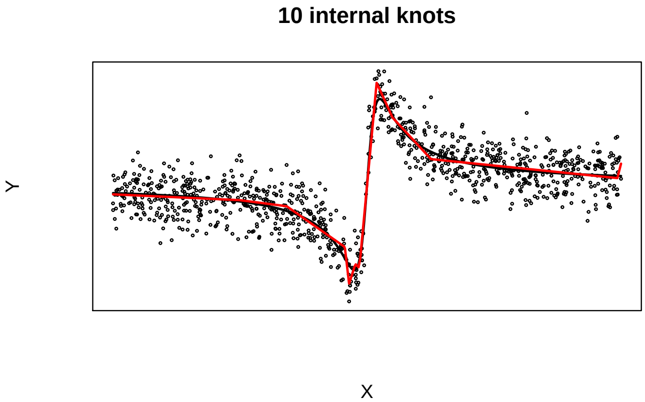
Stage A embeds a knot addition scheme

Starting from 2 couples of boundary knots, new knots are sequentially added



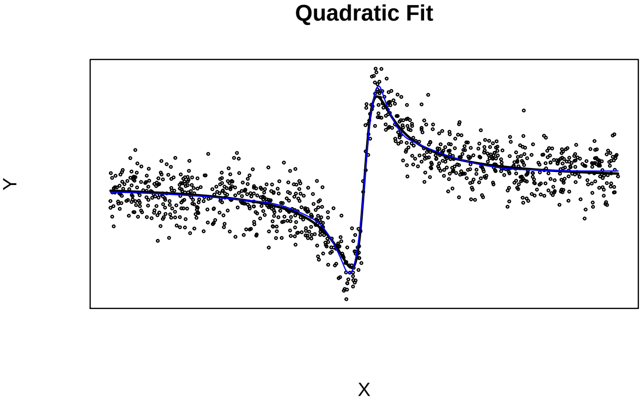
Stage A

Stage A embeds a knot addition scheme
Starting from 2 couples of boundary knots, new knots are sequentially added



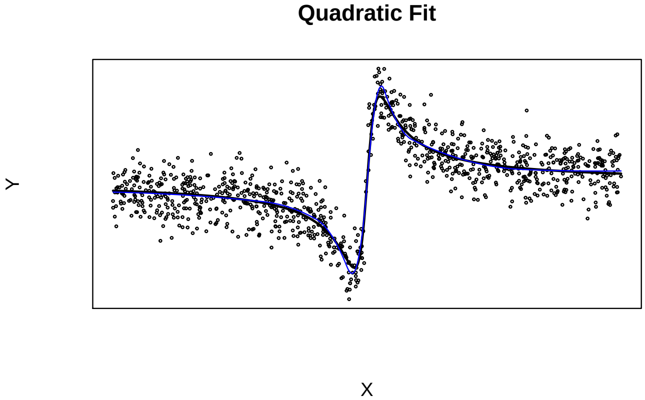
Stage B

In Stage B, starting from the result from stage A the knot locations for the higher order spline are computed



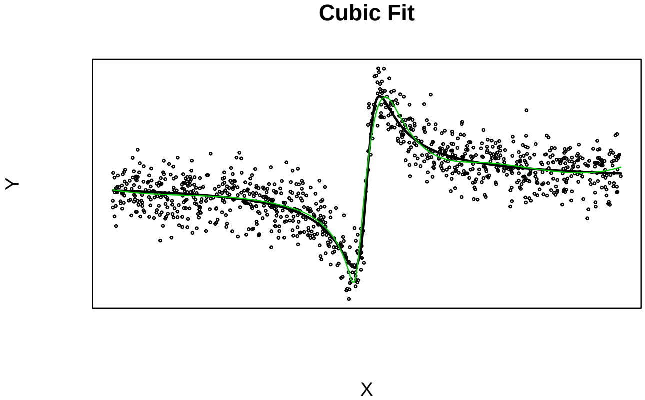
Stage B

In Stage B, starting from the result from stage A the knot locations for the higher order spline are computed



Stage B

In Stage B, starting from the result from stage A the knot locations for the higher order spline are computed



Comparison with existing methods I

GeDS Algorithm has been implemented in the R package `GeDS`

Main functions of the package are `NGeDS` and `GGeDS` that allow to perform the regression, both in the linear and in the generalized linear frameworks

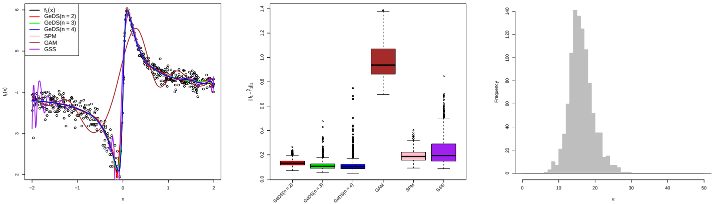
Similar R packages that work in the same context are

- `mgcv`, implementing the Generalized Additive Models (Wood, 2006)
- `ssanova`, implementing smoothing splines (Gu, 2014)
- `SemiPar`, implementing semi-parametric models (Wand, 2014)

Comparison with existing methods II

We performed a simulation study on some test functions in order to check whether GeDS regression performances are good
In particular:

- we simulate 2000 samples of 500 Poisson distributed data,
- we run GeDS regression on them and we store the number of knots
- we run regressions according to other R packages
- we check the goodness of fit of the estimates according to the L^1 norm



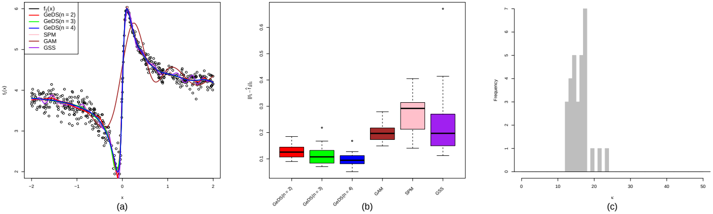
(a) (b) (c)

Comparison with existing methods III

The results seem to be quite good, however:

- The goodness of fit of GeDS regression shows that there are some outliers
- The number of knots selected is quite variable
- All the algorithms have some input parameters that should be properly tuned

We simulate 30 samples with the same characteristics as before and we tuned the parameters by hand for each of them



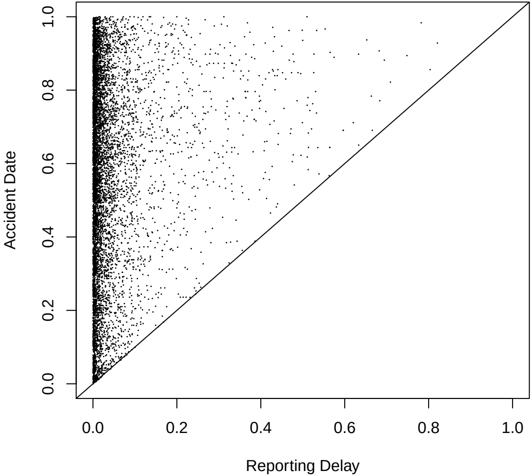
An application in Claims reserving

We use a dataset containing couples $(x_{i1}, x_{i2})_{i=1}^{N}$, $N = 8122$, where x_{i1} is the accident date and x_{i2} the reporting delay of the ith claim.

Unfortunately, we have:

- no information about the sizes
- no information about the exposures

But still we can use them in order to study IBNR claims

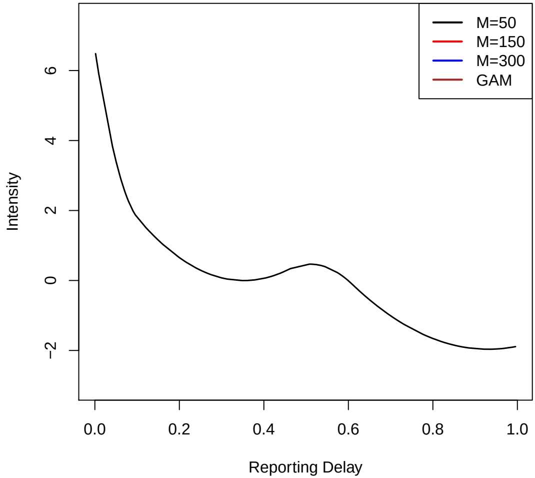

1.0
0.8
0.6
0.4
0.2
0.0
Accident Date
0.0
0.2
0.4
0.6
0.8
1.0
Reporting Delay

Setting the problem

We partition the support in M^2 squares (or rectangles) R_j, $j = 1, \ldots, M^2$
Let y_j be the counts of points falling in the rectangle R_j, then

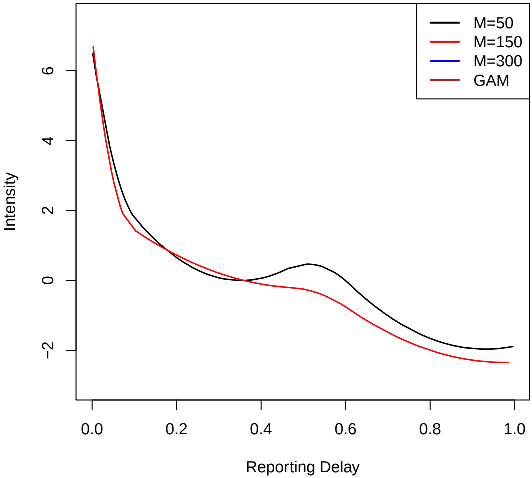
$$y_j \dot{\sim} Poi(\mu_j)$$

In order to assess the number of IBNR claims the actuary is interested in fitting the function $\mu_{\cdot} = \mu(x_1, x_2)$ and in particular to the predictions on the lower triangle, where $x_2 > x_1$.
If one assumes also $\log \mu(x_1, x_2) = \alpha + f_1(x_1) + f_2(x_2)$, see e.g. England and Verrall (2002), GeDS regression can be successfully applied.

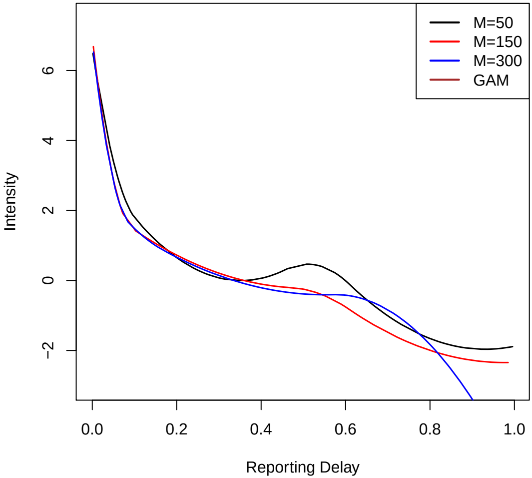


Intensity
Reporting Delay
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
M=50
M=150
M=300
GAM

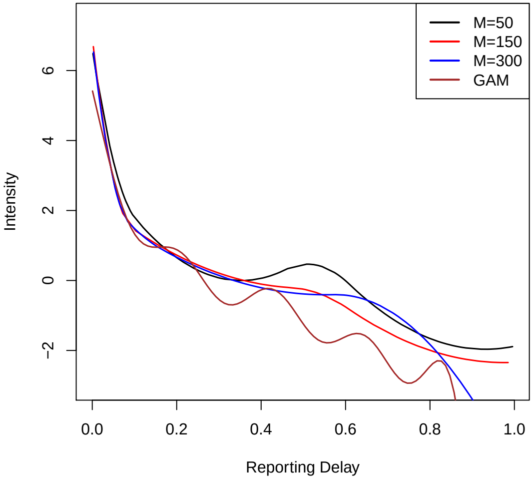
Results


M=50
M=150
M=300
GAM
Intensity
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8
1.0
Reporting Delay

Results


M=50
M=150
M=300
GAM
Intensity
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8
1.0
Reporting Delay

Results



M=50
M=150
M=300
GAM
Intensity
Reporting Delay
−2
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0

Conclusions

- We presented a novel approach to perform non-parametric regression
- The method can be very efficient in some cases, in particular when the smoothness of the objective function is not homogeneous over the whole domain
- in general it needs to be ‘driven’
- the methodology is implemented in an R package, that soon will be submitted to CRAN
- we presented an application in actuarial practice of claims reserving

Main References I

De Boor, C. (2001). A Practical Giude to Splines. Revised Edition. *Springer–Verlag*.

Dimitrova, D.S., Kaishev, V.K., Lattuada, A. and Verrall, R.J. (2017). Geometrically Designed Variable Knot Splines in Generalized (Non-)Linear Models. Submitted

England, P.D., and Verrall, R.J. (2002). Stochastic claims reserving in general insurance. *BAJ*, **8**(3), 443–518.

Gu, C. (2014). Smoothing Spline ANOVA Models: R Package **gss**. *Journal of Statistical Software*, **58**(5), 1–25.

Main References II

- Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. *Chapman & Hall, London*
- Wand, M.P. (2014). **SemiPar**: Semiparametic Regression. *R package version 1.0-4.1*. URL http://CRAN.R-project.org/package=SemiPar.
- Wood, S.N. (2006). Generalized Additive Models: An Introduction with R. *Chapman & Hall/CRC Press*.