Flexible regression modelling and P-splines approximations

Irène Gijbels

KU Leuven
Department of Mathematics and Leuven Statistics Research Centre
Belgium

Prague, August 31–September 4, 2015

European Young Statisticians Meetings

1st EYSM, 1978, Wiltshire, UK	
⋮	
5th EYSM, 1987, Aarhus	
6th EYSM, 1989, Prague	
⋮	
8th EYSM, 1993, Vilnius	**1st NANRC, 1993, Berkeley**
⋮	⋮
19th EYSM, 2015, Prague	17th NANRC, 2015, Seattle

North American New Researchers Conference (NANRC); organized by the Institute of Mathematical Statistics

young .... **new** ...... 55+ ......

young in spirit .... passion for research and scientific curiosity ...

Outline

multiple linear regression model: $\boxed{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d + \varepsilon}$

• mean regression function

$$E\left(\varepsilon | X_1, \ldots, X_d\right) = 0 \implies \boxed{E(Y|X_1, \ldots, X_d) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j}$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\; E\left(Y - \beta_0 - \sum_{j=1}^{d} \beta_j X_j\right)^2$$

• quantile regression function

denote (for $0 \leq \tau \leq 1$) : $F^{-1}_{\varepsilon|X_1,\ldots,X_d}(\tau) = \inf_z \left\{z : F_{\varepsilon|X_1,\ldots,X_d}(z) \geq \tau\right\}$

the τth conditional quantile of ε

$\Longrightarrow$ the τth conditional quantile of Y given $X_1, \ldots, X_d$

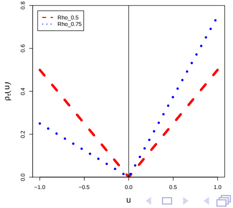
$$q_\tau(Y|X_1,\ldots,X_d) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + F^{-1}_{\varepsilon|X_1,\ldots,X_d}(\tau)$$

if $F^{-1}_{\varepsilon|X_1,\ldots,X_d}(\tau) = F^{-1}_{\varepsilon}(\tau)$, then

$$q_\tau(Y|X_1,\ldots,X_d) = \underbrace{\beta_0 + F^{-1}_{\varepsilon}(\tau)}_{=\beta_0^\tau} + \sum_{j=1}^{d} \beta_j X_j$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T = \text{argmin}_{\,\boldsymbol{\beta}}\, E\,\rho_\tau\left(Y - \beta_0 - \sum_{j=1}^{d} \beta_j X_j\right)$$

$$\rho_\tau(z) = \begin{cases} \tau\, z & \text{if } z > 0 \\ -(1-\tau)\, z & \text{otherwise} \end{cases}$$



Outline

mean regression function

observations:
$(Y_1, X_{11}, \ldots, X_{1d}), \ldots, (Y_n, X_{n1}, \ldots, X_{nd})$ i.i.d. from $(Y, X_1, \ldots, X_d)$

estimation of the mean regression coëfficients

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\; E\left(Y - \beta_0 - \sum_{j=1}^{d} \beta_j X_j\right)^2$$

Ordinary Least-Squares method:

$$\min_{\beta_0, \beta_1, \ldots, \beta_d} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \beta_j X_{ij}\right)^2 \quad \Longrightarrow \quad \widehat{\beta}_j^{\text{OLS}}\ , j = 0, 1, \ldots, d$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{pmatrix}$$

$n \times (d+1)$ design matrix

least-squares minimization problem: $\boxed{\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}$

$$\Longrightarrow \widehat{\boldsymbol{\beta}}^{\text{OLS}} = (\widehat{\beta}_0^{\text{OLS}}, \cdots, \widehat{\beta}_d^{\text{OLS}})^T$$

provided the inverse of the matrix $\mathbf{X}^T\mathbf{X}$ exists, the solution is

$$\boxed{\widehat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}$$

inference about $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$ follows rather easily from this expression

some assumptions are needed of course ...

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_d X_{id} + \varepsilon_i \qquad i = 1, \ldots, n$$

if the ε_i's are independent and identically distributed

with $E(\varepsilon_i | X_{i1}, \ldots, X_{id}) = 0$ and $\mathrm{Var}(\varepsilon_i | X_{i1}, \ldots, X_{id}) = \sigma^2$ then denoting

$$\mathcal{X} = \{(X_{11}, \ldots, X_{1d}), \ldots, (X_{n1}, \ldots, X_{nd})\}$$

- the least-squares estimator is a (conditionally) unbiased estimator: $\mathrm{E}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} | \mathcal{X}\right) = \boldsymbol{\beta}$
- the conditional variance-covariance matrix of $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is: $\mathbf{V}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} | \mathcal{X}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$
- the OLS estimator is an unbiased estimator
  it has the lowest variance of all unbiased estimators
- **but** with increasing correlation between the explanatory variables, the covariances between the corresponding estimated coefficients increase
  in other words: a strong correlation between the explanatory variables can be problematic
  or, the quantity $(\mathbf{X}^T\mathbf{X})^{-1}$ can be large ...

consider estimators that may have a small bias but have a lower variance ....

Ridge regression (Hoerl & Kennard (1970), ...) :

$$\min_{\beta_0,\beta_1,\ldots,\beta_d} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=0}^{d} \beta_j^2 \right\} \qquad \lambda > 0$$

or, in matrix notation (for simplicity without intercept)

$$\boxed{\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}}$$

$$\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{d} \beta_j^2 \qquad \text{the } L_2\text{-norm of the vector } \boldsymbol{\beta}$$

provided the inverse of the matrix $\mathbf{X}^T\mathbf{X}$ exists, the solution is

$$\boxed{\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_d \right)^{-1} \mathbf{X}^T \mathbf{Y}}$$

with $\mathbf{I}_d$ the identity matrix of dimension $d \times d$

$$\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \mathbf{X}^T \mathbf{Y} \quad = \quad (\underbrace{\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_d}_{\boldsymbol{S}_\lambda})^{-1} \mathbf{X}^T \mathbf{X}\, \widehat{\boldsymbol{\beta}}^{\text{OLS}}$$

(conditional) bias and variance-covariance matrix of the Ridge regression estimator:

$$E\left(\widehat{\boldsymbol{\beta}}^{\text{Ridge}} | \mathcal{X}\right) = \boldsymbol{\beta} - \lambda \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \boldsymbol{\beta} = \boldsymbol{\beta} - \lambda \boldsymbol{S}_\lambda^{-\mathbf{1}} \boldsymbol{\beta}$$

(conditional) variance-covariance matrix of $\left(\widehat{\boldsymbol{\beta}}^{\text{Ridge}}\right)$

$$\mathbf{V}\left(\widehat{\boldsymbol{\beta}}^{\text{Ridge}} | \mathcal{X}\right) = \sigma^2 \boldsymbol{S}_\lambda^{-\mathbf{1}} \mathbf{X}^T \mathbf{X} \boldsymbol{S}_\lambda^{-\mathbf{1}} \leq \lambda^{-2} \mathbf{V}\left(\widehat{\boldsymbol{\beta}}^{\text{OLS}} | \mathcal{X}\right)$$

- the bias depends on λ, the design matrix and the true $\boldsymbol{\beta}$
- the variance is smaller than that of the OLS estimator

in case of an orthogonal design matrix $\mathbf{X}$, i.e. when $\mathbf{X}^T\mathbf{X} = \mathbf{I}_d$, we have that $\boldsymbol{S}_\lambda = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_d = (1+\lambda)\mathbf{I}_d$, and

$$\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = \frac{1}{1+\lambda}\,\widehat{\boldsymbol{\beta}}^{\text{OLS}}$$

the Ridge parameter results in a **shrinkage** of the least-squares regression coefficients, **but** none of the coefficients will be put to zero (**no selection**)

the Ridge regression minimization problem is equivalent to the minimization problem

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \qquad \text{subject to } \|\boldsymbol{\beta}\|_2^2 \leq s$$

with $s > 0$ a shrinkage/regularization parameter

Outline

- **Ordinary least-squares**: $\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right\}$
- **Ridge regression**:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \qquad \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{d} \beta_j^2 \ \ L_2\text{-norm}$$

- **Least Absolute Shrinkage and Selection Operator (LASSO)** :

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\} \qquad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{d} |\beta_j| \ \ L_1\text{-norm}$$

(Tibshirani (1996, 2014), Lockhart *et al.* (2014), ...

- **Bridge regression** ($0 < \gamma < 1$) :

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_\gamma^\gamma \right\} \qquad \|\boldsymbol{\beta}\|_\gamma^\gamma = \sum_{j=1}^{d} |\beta_j|^\gamma \quad L_\gamma\text{-norm}$$

(Frank & Frieman (1993), Fu (1998), Knight & Fu (2000), ...)

- **Elastic net**:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}$$

(Zou & Hastie (2005), Wu (2012), Slawski (2012), Zhou (2013), ...)

- **Adaptive LASSO**:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} w_j \, |\beta_j| \right\}$$

with $w_j > 0$ weights (depending on the data), e.g. $w_j = \dfrac{1}{\left|\widehat{\beta}_j^{\mathsf{OLS}}\right|}$

(Zou (2006), Potscher & Schneider (2009), ...)

- **...**

the optimization problem of the LASSO method can be re-expressed via the equivalent optimization problem

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \qquad \text{subject to } \|\boldsymbol{\beta}\|_1 \leq s$$

with $s > 0$ a shrinkage/regularization parameter

in case of an orthogonal design matrix $\mathbf{X}$, there is an explicit relationship between the ordinary least-squares estimator $\widehat{\boldsymbol{\beta}}^{\mathsf{OLS}}$ and the LASSO regression estimator

$$\widehat{\boldsymbol{\beta}}_j^{\mathsf{LASSO}} = \mathrm{sign}\left(\widehat{\boldsymbol{\beta}}_j^{\mathsf{OLS}}\right) \max\left(0, \left|\widehat{\boldsymbol{\beta}}_j^{\mathsf{OLS}}\right| - \lambda\right) \qquad j = 1, \ldots, d$$

this clearly shows the **shrinking** and **selection** effect of the LASSO method

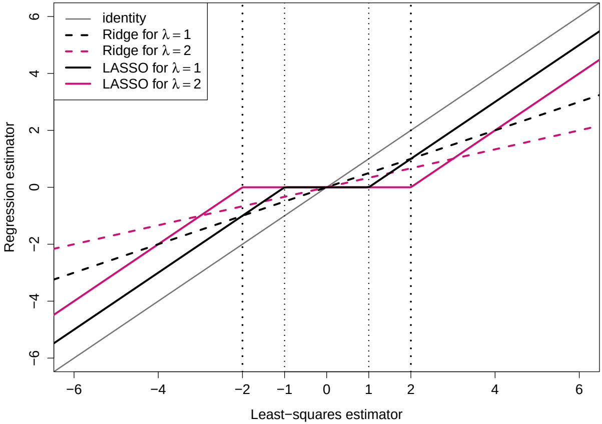


Figure: *OLS estimates and some other estimates.*

the nonnegative garrote method (Breiman (1995))

basic idea:

find shrinkage factors $(c_1, \ldots, c_d)$ *that shrink the least-squares regression coefficients: instead of an estimated coefficient* $\widehat{\beta}_j^{OLS}$ *one considers* $c_j\, \widehat{\beta}_j^{OLS}$

a shrinkage should

- not alter the sign of a covariate's influence in the linear model
- be globally a real shrinkage of the original regression coefficients:

$$c_j \geq 0\,, \text{for } j = 1, \ldots, d, \qquad \text{and} \qquad \sum_{j=1}^{d} c_j \leq s \qquad \text{with } s \leq d$$

the nonnegative garrote shrinkage factors $\widehat{c}_j$ are found by solving the optimization problem

$$\begin{cases} \min\limits_{c_1,\ldots,c_d} \sum\limits_{i=1}^{n} \left(Y_i - \sum\limits_{j=1}^{d} c_j \widehat{\beta}_j^{\mathsf{OLS}} X_{ij} \right)^2 \\ \text{subject to } \ 0 \leq c_j \text{ , for } j = 1, \ldots, d, \quad \text{and} \quad \sum\limits_{j=1}^{d} c_j \leq s \end{cases}$$

for given s, also equivalent to the optimization problem

$$\begin{cases} \min\limits_{c_1,\ldots,c_d} \left\{ \sum\limits_{i=1}^{n} \left(Y_i - \sum\limits_{j=1}^{d} c_j \widehat{\beta}_j^{\mathsf{OLS}} X_{ij} \right)^2 + \lambda \sum\limits_{j=1}^{d} c_j \right\} \\ \text{subject to } \ 0 \leq c_j \text{ , for } j = 1, \ldots, d \text{ ,} \end{cases}$$

for given $\lambda > 0$

the nonnegative garrote (NNG) estimator of the regression coefficient β_j (model without intercept term) is

$$\widehat{\beta}_j^{\mathsf{NNG}} = \widehat{c}_j \widehat{\beta}_j^{\mathsf{OLS}} \qquad j = 1, \ldots, d$$

and in the special case of an orthogonal design matrix :

$$\widehat{c}_j = \max\left(0, 1 - \frac{\lambda}{\left(\widehat{\beta}_j^{\mathsf{OLS}}\right)^2}\right)$$

$\Longrightarrow$ **shrinking** and **selection** effect (if $\left(\widehat{\beta}_j^{\mathsf{OLS}}\right)^2 < \lambda$)

example: Boston Housing data

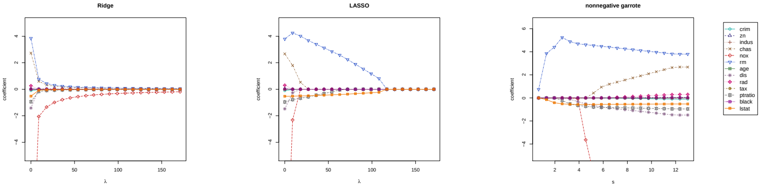


Figure: *Boston housing data: Estimated coefficients in function of the regularization parameter for Ridge,* LASSO *and* NNG *methods.*

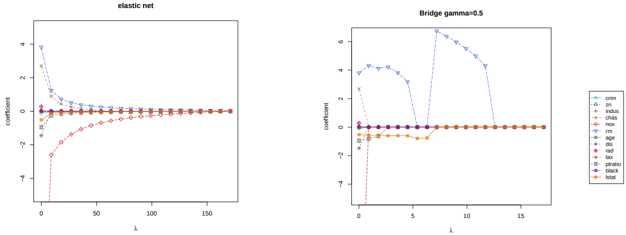


Figure: *Boston housing data: Estimated coefficients in function of the regularization parameter, for elastic net and Bridge.*

Least Absolute Shrinkage and Selection Operator (LASSO) :

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\} \qquad \|\boldsymbol{\beta}\|_1 = \sum_{j=0}^{d} |\beta_j| \ L_1\text{-norm}$$

the added term does not need to be an L_p-type of norm nor a combination of norms of $\boldsymbol{\beta}$

it can be any positive-valued function that regularizes the regression coefficients in general, one can consider the optimization problem

$$\boxed{\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}) \right\}}$$

$J(\cdot)$ a given penalty function, that penalizes the resulting estimator in case the function-value $J(\boldsymbol{\beta})$ is too large

in the literature (statistics, but also numerical analysis, engineering, ...) there are a wealth of regularization techniques that result from including a penalty term

in general, the penalty term $J(\boldsymbol{\beta})$ is of a form $\boxed{J(\boldsymbol{\beta}) = \sum_{j=1}^{d} \gamma_j \psi(\beta_j)}$

$\gamma_j > 0$ weights; $\psi(\cdot) \geq 0$ a function satisfying some conditions

important properties distinguishing between the various $\psi(\cdot)$ functions:

- the smoothness (mainly differentiability) of the function at zero;
- the convexity or nonconvexity of the function

Smoothed Clipped Absolute Deviation (SCAD) penalty (Fan (1997), Antoniadis & Fan (2001), ...) :

$$\psi'(|\beta|) = \lambda \left\{ I\{|\beta| \le \lambda\} + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I\{|\beta| > \lambda\} \right\} \qquad a > 2$$

the integral of this leads to the penalty

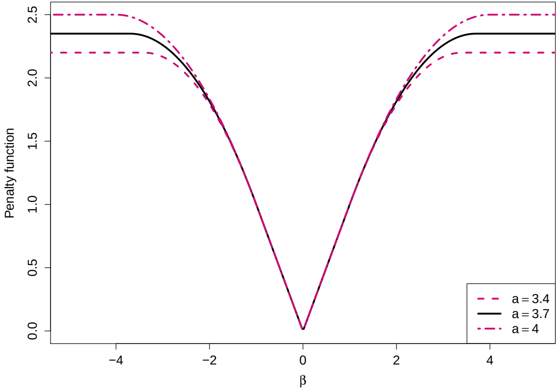


Figure: *SCAD penalty: non-differentiable at zero and nonconvex penalty, for three values of a.*

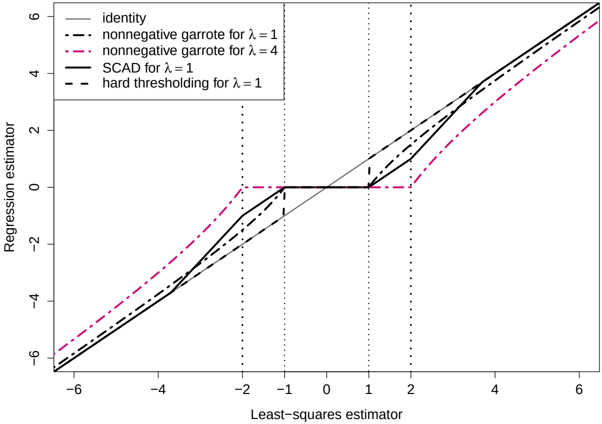


Figure: *OLS estimates and some other estimates.*

frequently-used hyperbolic type of penalty function : $\psi(\beta) = \sqrt{\gamma + \beta^2}$

for various values of γ, including $\gamma = 0$ when the function reduces to the absolute value function $\psi(\beta) = |\beta|$ (the L_1-penalty)

functions are convex and either differentiable at zero (for $\gamma > 0$) or non-differentiable at zero (for $\gamma = 0$)

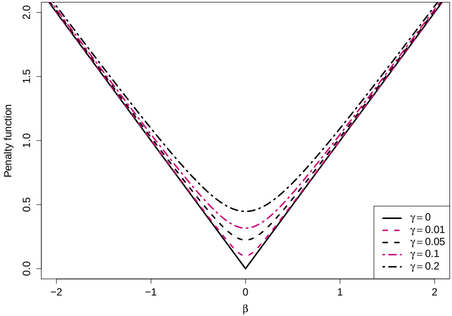


Figure: *Examples of differentiable and non-differentiable convex penalties* ($\psi(\beta) = \sqrt{\gamma + \beta^2}$).

another example: $\psi(\beta) = 1 - \exp(-\gamma\beta^2)$

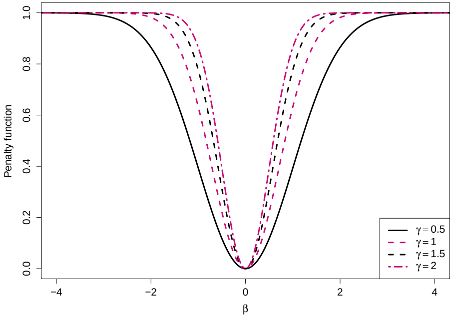


Figure: *Examples of differentiable nonconvex penalties ($\psi(\beta) = 1 - \exp(-\gamma\beta^2)$).*

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{d} \gamma_j \psi(\beta_j) \qquad \text{more generally} \qquad J(\boldsymbol{\beta}) = \sum_{j=1}^{d} \gamma_j \psi(d_j^T \boldsymbol{\beta})$$

with $\gamma_j > 0$ weights and where d_j are given linear operators

if ψ is a convex function, then J forces the regularized solution $\widehat{\boldsymbol{\beta}}$ of the considered optimization problem to be such that $|d_j^T \widehat{\boldsymbol{\beta}}|$ is small

special class of penalty functions : d_j finite difference operators

- difference operator of order 1: $\Delta^1 \beta_j = \beta_j - \beta_{j-1}$
- difference operator of order 2: $\Delta^2 \beta_j = \beta_j - 2\beta_{j-1} + \beta_{j-2}$
- difference operator of order k (with $k \in I\!N$), denoted by $d_j = \Delta^k$:

$$\Delta^k \beta_j = \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} \beta_{j-\ell}$$

using a finite order difference operator Δ^k encourages solutions $\widehat{\boldsymbol{\beta}}$ with neighboring coefficients having similar values

Outline

flexible univariate regression model and P-splines approximations

$$\boxed{Y = \mu(X) + \varepsilon}$$

$\mu(x)$ **unknown** univariate function

without loss of generality: X takes values in $[0, 1]$

$$E(\varepsilon|X = x) = 0 \implies \mu(x) = E(Y|X = x)$$

assume: μ can be approximated by a set of basis functions $B_1(\cdot), \ldots, B_m(\cdot)$:

$$\boxed{\mu(x) \approx \sum_{j=1}^{m} \alpha_j B_j(x)}$$

AIM: **estimate the coefficients** $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$

- examples of basis functions: wavelets, polynomial splines, ...
- crucial choice: number m of basis functions

popular choice of basis functions: **B-splines basis functions** $\{B_1(\cdot;q),\ldots,B_m(\cdot;q)\}$

- functions $B_j(x;q)$, are piecewise polynomial functions of degree q;
- $(q-1)$-st derivative is a continuous function on $[0,1]$, but not differentiable in the points $t_0,t_1,\ldots,t_K$ in the interval $[0,1]$, called the knot points;
- often one works with normalized B-splines, i.e. satisfying $\sum_{j=1}^{m} B_j(x;q)=1$, and equidistant knot points $t_0=0, t_1=1/K,\ldots,t_{K-1}=(K-1)/K, t_K=1$ in the interval $[0,1]$
- with $K+1$ equidistant knot points and q the degree of the polynomial pieces, there are $m=K+q$ basis functions that span the space of functions on $[0,1]$ that are splines of degree q

given knots $0 < t_1 < \cdots < t_K < 1$; B-splines are polynomial pieces of degree q joined together at each knot t_k

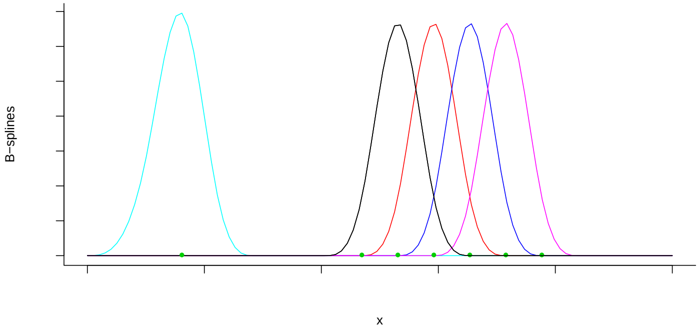


Figure: *Some functions of a B-splines basis.*

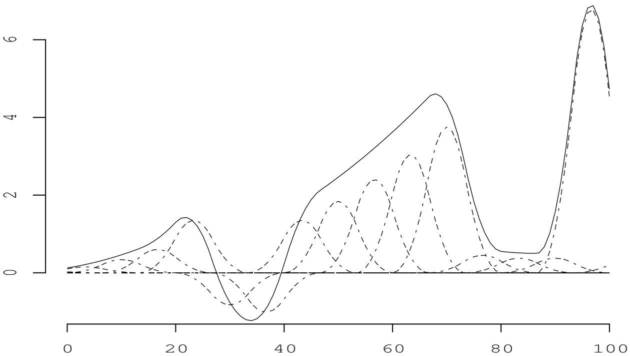


Figure: *Illustration of B-spline constructed smooth curve.*

dashed curves: scaled basis functions; heights are the coefficients

solid curve: resulting smooth curve as sum of scaled B-splines

$$\boxed{\mu(x) \approx \sum_{j=1}^{m} \alpha_j B_j(x;q) \qquad\qquad m = K + q}$$

- if $\mu(\cdot)$ belongs to this space of functions, then $\mu(x) = \sum_{j=1}^{m} \alpha_j B_j(x;q)$;
- if $\mu(\cdot)$ does not belong to this space, then one needs to deal with a **modeling bias**:
  - take a large number of knot points K (increasing as such the flexibility of the model)
  - control the risk of overfitting (too many parameters) by introducing a penalty to the least-squares approximation method

with $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. observations from (X, Y)

resulting optimization problem

$$\boxed{\min_{\alpha_1,\ldots,\alpha_m} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(X_i;q) \right)^2 + \lambda J(\boldsymbol{\alpha}) \right\}}$$

$\lambda > 0$ smoothing parameter

commonly-used penalty function: $J(\boldsymbol{\beta}) = \sum_{j=1}^{d} \gamma_j \psi(d_j^T \boldsymbol{\beta})$ with $\psi(\beta) = \beta^2$, $d_j = \Delta^k$

P-splines optimization problem

$$\boxed{\min_{\alpha_1,\ldots,\alpha_m} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(X_i; q) \right)^2 + \lambda \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2 \right\}}$$

$\Longrightarrow$ results in a **sparse representation** for curves that are smooth on a large part of the domain (since for smooth curves neighbouring coefficients of B-splines will be close)

Eilers & Marx (1996), ...

in matrix notation ...

notations:

$$(X_1, X_2, \cdots, X_n) \qquad \mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^T \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$$

$$\mathbf{B} = \begin{pmatrix} B_1(X_1) & B_2(X_1) & \cdots & B_m(X_1) \\ B_1(X_2) & B_2(X_2) & \cdots & B_m(X_2) \\ \vdots & \vdots & & \vdots \\ B_1(X_i) & B_2(X_i) & \cdots & B_m(X_i) \\ \vdots & \vdots & & \vdots \\ B_1(X_n) & B_2(X_n) & \cdots & B_m(X_n) \end{pmatrix} \quad \text{matrix of dim } n \times m$$

$$\mathbf{B}(X_i) = (B_1(X_i), B_2(X_i), \cdots, B_m(X_i)) \quad \text{vector of dim } 1 \times m$$

objective function to be minimized, with respect to $\boldsymbol{\alpha}$:

$$\boxed{\sum_{i=1}^{n} (Y_i - \mathbf{B}(X_i)\boldsymbol{\alpha})^2 + \lambda J(\boldsymbol{\alpha})}$$

minimize $\left\{ \sum_{i=1}^{n} (Y_i - \mathbf{B}(X_i)\boldsymbol{\alpha})^2 + \lambda J(\boldsymbol{\alpha}) \right\}$ with respect to $\boldsymbol{\alpha}$

$$\boxed{\min_{\boldsymbol{\alpha}} \left\{ (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \mathbf{D}_k^T \mathbf{D}_k \boldsymbol{\alpha} \right\}}$$

matrix $\mathbf{D}_k$ for the k-th order difference operator :

$\sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2 = \boldsymbol{\alpha}^T \mathbf{D}_k^T \mathbf{D}_k \boldsymbol{\alpha}$

matrix $\mathbf{D}_k$ = a matrix of dimension $(m-k) \times m$

example: for a B-spline basis of degree 2, and 5 knots (i.e. $K = 4$), $m = 6$, and the matrix $\mathbf{D}_2$ is a matrix of dimension 4×6

$$\mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}$$

$$\min_{\boldsymbol{\alpha}} \left\{ (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \mathbf{D}_k^T \mathbf{D}_k \boldsymbol{\alpha} \right\}$$

solution to the optimization problem:

penalized regression estimator:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_m)^T$$

estimator of the function μ:

$$\boxed{\widehat{\mu}(x) = \sum_{j=1}^{m} \widehat{\alpha}_j B_j(x; q)}$$

extension to a **generalized linear model**

Y: response variable $\qquad$ X: covariate (univariate)

cond. distrib. of Y given $X = x$ is from an **exponential family** distr.

$$\boxed{f_{Y|X}(y|x) = \exp\left(\frac{y\theta(x) - b(\theta(x))}{\phi} + c(y_i, \phi)\right)}$$

$b(\cdot)$ and $c(\cdot)$ known functions; $\qquad$ ϕ : known scale parameter

$\theta(\cdot)$ unknown function

$E(Y|X = x) = b'(\theta(x)) = \mu(x)$ $\qquad$ $\text{Var}(Y|X = x) = \phi\, b''(\theta(x))$

$g(\mu(x)) = \eta(x)$ $\qquad$ g the **link function**

$\eta(\cdot)$ the **predictor function**, to be estimated

generalized <u>linear</u> models: $\eta(x) =$ a linear function of x

Examples

- **Normal regression** with additive errors: $f_{Y|X}(y|x) \sim \mathsf{N}\left(\mu(x); \sigma^2\right)$

  link function: $g(t) = t$ (identity) predictor fct $\eta(x) = \mu(x)$

- **Logistic regression**: $f_{Y|X}(y|x) \sim \text{Bernoulli}\,(1; \mu(x))$

  0-1 response type of variable Y $\mu(x)=$ conditional probab.

  link fct: $g(t) = \log \dfrac{t}{1-t}$ (logit) predictor fct $\eta(x) = \log \frac{\mu(x)}{1-\mu(x)}$

- **Poisson regression**: $f_{Y|X}(y|x) \sim \text{Poisson}\,(\mu(x))$

  counts type of r.v. Y $\mu(x)=$ Poisson intensity function

  link function: $g(t) = \log(t)$ predictor fct $\eta(x) = \log\left(\mu(x)\right)$

regression analysis:

from observation $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

estimate the predictor function $\eta(\cdot)$

- standard parametric model: $\eta(x) = \eta(x; \boldsymbol{\alpha})$

  ex.: generalized linear models; $\eta(x; \boldsymbol{\alpha})$ a function linear in $\boldsymbol{\alpha}$

- nonparametric estimation: several techniques, ..., e.g. penalization techniques

$$\boxed{\eta(x) \approx \sum_{k=1}^{m} \alpha_k B_k(x)}$$

objective function to be maximized:

$$\boxed{\textit{maximize}_{\eta\in\textit{function space}} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i,\eta(X_i)) - \lambda J(\eta) \right\}}$$

ℓ =**log-likelihood** $\qquad J(\cdot)$ is a roughness functional (penalty)

1st term: discourages the lack of fit of η to the data

2nd term: penalizes the roughness of η

$\lambda > 0$: smoothing parameter controlling trade-off between 2 terms

nonparametric setting: $\eta(x) \approx \sum_{k=1}^{m} \alpha_k B_k(x)$, with m large enough

$$\eta(X_i) \approx \mathbf{B}(X_i)\boldsymbol{\alpha}$$

penalized log-likelihood estimator:

$$\boxed{\text{maximize}_{\boldsymbol{\alpha}} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i,\mathbf{B}(X_i)\boldsymbol{\alpha}) - \lambda J(\alpha) \right\}}$$

$$\text{maximize}_{\boldsymbol{\alpha}} \quad \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \mathbf{B}(X_i)\boldsymbol{\alpha}) - \lambda J(\alpha) \right\}$$

- how to do the optimization of the penalized log-likelihood ?
- algorithm for carrying out the optimization ?
- statistical properties and asymptotic analysis of the penalized maximum likelihood estimators of $\boldsymbol{\alpha}$, of $\eta(\cdot)$ and of $\mu(\cdot)$, ... ?
- bias, variance of the estimators, consistency + rate of convergence, asymptotic distributional results, ...
- finite-sample performance ?

Antoniadis, G. & Nikolova (2011), Li *et al.* (2012), ...

flexible multiple regression models and P-splines approximations

Y response variable $\qquad X_1, \ldots X_d$ covariates

a multiple linear regression model, $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d + \varepsilon$, assumes a linear influence of each of the covariates on the response variable

- **additive regression model**

influence of X_j on Y is modeled via an **unknown** univariate function f_j:

$$\boxed{Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \varepsilon\ , \qquad \text{with} \quad E\left(f_j(X_j)\right) = 0}$$

$(Y_1, X_{11}, \ldots, X_{1d}), \ldots, (Y_n, X_{n1}, \ldots, X_{nd})$ i.i.d. observations from $(Y, X_1, \ldots, X_d)$ satisfying the additive model

how to obtain estimators for the d unknown functions ?

$\boldsymbol{f}_j = (f_j(X_{1j}), \ldots, f_j(X_{nj}))^T$ the column vector of all f_j function values (evaluated at the observed values of X_j)

P-splines estimation of the functions f_j can be done as follows

Step 1: Initialization step: put $\widehat{f}_0 = n^{-1} \sum_{i=1}^{n} Y_i$, and $\widehat{\boldsymbol{f}}_j = \mathbf{0}$, for $j = 1, \ldots, d$;

Step 2: for $j = 1, \ldots, d$, calculate the residuals $\boldsymbol{e}_j = \mathbf{Y} - \sum_{\ell \neq j} \widehat{\boldsymbol{f}}_\ell$, and use univariate P-splines regression applied to $\boldsymbol{e}_j$, to estimate $\boldsymbol{f}_j$;

Step 3: Repeat *Step 2* until convergence.

$\Longrightarrow$ consistent estimation of $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d$

Eilers & Marx (2002), Antoniadis, G. & Verhasselt (2012b)

- **varying coefficient regression model**

multiple linear regression model: $Y = \beta_0 + \beta_1 X^{(1)} + \ldots + \beta_d X^{(d)} + \varepsilon$

complex data

flexible modelling $\longrightarrow$ **varying coefficient regression model**:

$$\boxed{Y(\mathbf{t}) = \beta_0(\mathbf{t}) + \beta_1(\mathbf{t}) X^{(1)}(\mathbf{t}) + \ldots + \beta_d(\mathbf{t}) X^{(d)}(\mathbf{t}) + \varepsilon(\mathbf{t})}$$

$(t \in \mathcal{T} = [0, T])$

$\varepsilon(t)$ independent of $(X^{(1)}(t), \ldots, X^{(d)}(t), t)$

Hastie & Tibshirani (1993), Hoover *et al.* (1998), Fan & Zhang (2008), Lu *et al.* (2008), Wang *et al.* (2008), ...,
Antoniadis, G. & Verhasselt (2012a), Andriyana (2014), ...

$$
\begin{aligned}
Y(t) &= \beta_0(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_d(t)X^{(d)}(t) + \varepsilon(t) \\
&= \mathbf{X}(t)^T\boldsymbol{\beta}(t) + \varepsilon(t)
\end{aligned}
$$

where $\mathbf{X}(t) = \left(X^{(0)}(t), X^{(1)}(t), \ldots, X^{(d)}(t)\right)^T$ covariate vector at time t with $X^{(0)}(t) \equiv 1$

$\boldsymbol{\beta}(t) = (\beta_0(t), \beta_1(t), \ldots, \beta_d(t))^T$

vector of $(d+1)$ unknown **univariate** regression coefficients at time t

$\beta_0(t)$ is the baseline effect

assume that $\varepsilon(t)$ is a mean zero stochastic process at time t

first aim: estimate the **mean regression function**

$$
E(Y(t)|\mathbf{X}(t), t) = \beta_{\mathbf{0}}(\mathbf{t}) + \beta_{\mathbf{1}}(\mathbf{t})X^{(1)}(t) + \ldots + \beta_{\mathbf{d}}(\mathbf{t})X^{(d)}(t)
$$

observational setting: longitudinal data setup

n independent subjects/individuals

for each individual i: measurements repeated over a time period

measurements at time points $t_{i1}, \ldots, t_{iN_i}$

N_i different measurements for response and all explanatory variables:

$Y(t_{ij}) = Y_{ij}$

$X^{(k)}(t_{ij}) = X_{ij}^{(k)} \quad k = 1, \ldots, d \Longrightarrow \mathbf{X}(t_{ij}) \overset{\text{not.}}{=} \mathbf{X}_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(d)})^T$

total number of observations over all individuals:

$$N = \sum_{i=1}^{n} N_i$$

example: CD4 data example

the data are a subset from the Multicenter AIDS Cohort Study (Kaslow *et al.* (1987))

contain repeated measurements of physical examinations, laboratory results, CD4 cell counts and CD4 percentages of 283 homosexual men who became HIV-positive between 1984 and 1991

unequal numbers of repeated measurements and different measurement times for each individual

the number of repeated measurements ranged from 1 to 14, with a median of 6 and mean of 6.57

the number of distinct time points was 59

response variable :

$Y(t)$ = CD4 percentage at time t after infection

covariates:

- $X_i^{(1)}$ the smoking status of the i-th individual (1 or 0 if the individual ever or never smoked cigarettes)
- $X_i^{(2)}$ the centered age at HIV infection for the i-th individual
- $X_i^{(3)}$ the centered pre-infection CD4 percentage

aim: try to evaluate the mean effects of cigarette smoking, pre-HIV infection CD4 cell percentage and age at HIV infection on the CD4 percentage after infection response:

the conditional mean function

$$E(Y(t)|\mathbf{X}(t),t) = \beta_{\mathbf{0}}(\mathbf{t}) + \beta_{\mathbf{1}}(\mathbf{t})X^{(1)}(t) + \ldots + \beta_{\mathbf{d}}(\mathbf{t})X^{(d)}(t)$$

longitudinal data: $\left(t_{ij}, Y_{ij}, X_{ij}^{(1)}, \ldots, X_{ij}^{(d)}\right)$

$$i = 1, \ldots, n, \quad j = 1, \ldots, N_i \qquad N = \sum_{i=1}^{n} N_i$$

estimation of the $(d+1)$ unknown **univariate** regression functions $\beta_k(t)$, $k = 0, \ldots, d$

P-spline estimator for the regression coefficient function $\beta_k(\cdot)$

Lu, Zhang & Zhu (2008), Wang & Huang (2008), ...

$$E(Y(t)|\mathbf{X}(t),t) = \beta_\mathbf{0}(\mathbf{t}) + \beta_\mathbf{1}(\mathbf{t})X^{(1)}(t) + \ldots + \beta_\mathbf{d}(\mathbf{t})X^{(d)}(t)$$

suppose: each unknown function $\beta_k(t)$, $k = 0, \ldots, d$, can be approximated by a B-spline basis expansion

$$\boxed{\beta_k(t) \approx \alpha_{k1}B_{k1}(t;\nu_k) + \ldots + \alpha_{km_k}B_{km_k}(t;\nu_k) = \sum_{\ell=1}^{m_k} \alpha_{k\ell}B_{k\ell}(t;\nu_k)}$$
$$= \boldsymbol{\alpha}_k^T \mathbf{B}_k(t;\nu_k)$$

$$\boldsymbol{\alpha}_\mathbf{k} = (\alpha_\mathbf{k1}, \ldots, \alpha_{\mathbf{km_k}})^\mathbf{T} \qquad \mathbf{B}_k(t;\nu_k) = (B_{k1}(t;\nu_k), \ldots, B_{km_k}(t;\nu_k))^T$$

$m_k = u_k + \nu_k$ $\qquad$ $u_k + 1 =$ number of knot points

where $\{B_{k\ell}(\cdot;\nu_k) : \ell = 1, \ldots, u_k + \nu_k = m_k\}$ is the ν_k-th degree B-spline basis with $u_k + 1$ equidistant knots for the k-th component

$$\boxed{\beta_k(t_{ij}) \approx \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t_{ij};\nu_k)}$$

the P-spline estimates of the regression coefficients $\alpha_{k\ell}$ are obtained by minimizing $S(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \ldots, \boldsymbol{\alpha}_d^T)^T \in I\!R^{m_{\text{tot}} \times 1}$, where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{km_k})^T$ and $m_{\text{tot}} = \sum_{k=0}^{d} m_k$:

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(Y_{ij} - \sum_{k=0}^{d} \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t_{ij};\nu_k) X_{ij}^{(k)} \right)^2 + \sum_{k=0}^{d} \lambda_k \boldsymbol{\alpha}_k^T \mathbf{D}_{d_k}^T \mathbf{D}_{d_k} \boldsymbol{\alpha}_k$$

d_k is the differencing order for the k-th component

$\lambda_k > 0$ are the $(d+1)$ smoothing parameters

$$
\begin{aligned}
S(\boldsymbol{\alpha}) &= \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(Y_{ij} - \sum_{k=0}^{d} \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t_{ij}; \nu_k) X_{ij}^{(k)} \right)^2 + \sum_{k=0}^{d} \lambda_k \boldsymbol{\alpha}_k^T \mathbf{D}_{d_p}^T \mathbf{D}_{d_k} \boldsymbol{\alpha}_k \\
&= \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\alpha})^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\alpha}) + \boldsymbol{\alpha} \mathbf{Q}_\lambda \boldsymbol{\alpha}
\end{aligned}
$$

$$
\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^T
$$

$$
\mathbf{B}(t) = \begin{pmatrix} B_{01}(t; q_0) & \ldots & B_{0m_0}(t; q_0) & 0 \ldots 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \ddots & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 \ldots 0 & B_{d1}(t; q_d) & \ldots & B_{dm_d}(t, q_d) \end{pmatrix}
$$

$$
\mathbf{U}_{ij}^T = \mathbf{X}_{ij}^T \mathbf{B}(t_{ij}) \in I\!R^{1 \times m_{\text{tot}}} \qquad \mathbf{X}_{ij} = \left(1, X^{(1)}(t_{ij}), \ldots, X^{(d)}(t_{ij})\right)^T
$$

$$
\mathbf{U}_i = (\mathbf{U}_{i1}^T, \ldots, \mathbf{U}_{iN_i}^T)^T \in I\!R^{N_i \times m_{\text{tot}}}
$$

$\mathbf{W}_i = \text{diag}\left(N_i^{-1}, \ldots, N_i^{-1}\right) \in I\!R^{N_i \times N_i}$ (a diagonal matrix with N_i times N_i^{-1} on the diagonal)

$\mathbf{Q}_\lambda = \text{diag}\left(\lambda_0 \mathbf{D}_{d_0}^T \mathbf{D}_{d_0}, \ldots, \lambda_d \mathbf{D}_{d_d}^T \mathbf{D}_{d_d}\right) \in I\!R^{m_{\text{tot}} \times m_{\text{tot}}}$ (a block diagonal matrix with the matrices $\lambda_k \mathbf{D}_{d_k}^T \mathbf{D}_{d_k}$ on the diagonal)

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\alpha})^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\alpha}) + \boldsymbol{\alpha} \mathbf{Q}_\lambda \boldsymbol{\alpha}$$

if $\sum_{i=1}^{n} \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i + \mathbf{Q}_\lambda$ is invertible then $S(\boldsymbol{\alpha})$ has a unique minimizer

$$\boxed{\widehat{\boldsymbol{\alpha}} = \Big(\sum_{i=1}^{n} \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i + \mathbf{Q}_\lambda \Big)^{-1} \sum_{i=1}^{n} \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i}$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_0^T, \ldots, \widehat{\boldsymbol{\alpha}}_d^T)^T$ and $\widehat{\boldsymbol{\alpha}}_k = (\widehat{\alpha}_{k1}, \ldots, \widehat{\alpha}_{km_k})^T$ for $k = 0, \ldots, d$

the P-spline estimate of $\boldsymbol{\beta}(t)$ is then

$$\widehat{\boldsymbol{\beta}}(t) = \mathbf{B}(t) \widehat{\boldsymbol{\alpha}} = (\widehat{\beta}_0(t), \ldots, \widehat{\beta}_d(t))^T \quad \text{with} \quad \widehat{\beta}_k(t) = \sum_{\ell=1}^{m_k} \widehat{\alpha}_{k\ell} B_{k\ell}(t; \nu_k)$$

theoretical results are established for the case that the number of knots $u_k + 1$ (and thus $m_k = u_k + \nu_k$) grows with n

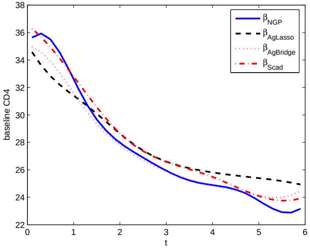
$\beta_k(\cdot)$ is not a spline function itself, but can be **approximated** by a spline function

theoretical results

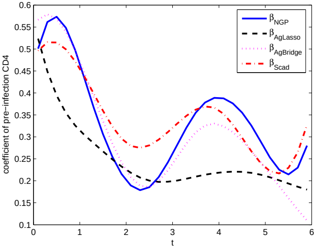
- consistency result (+ rate)

$$\|\widehat{\beta}_k - \beta_k\|_{L_2} = \left\{ \int_{\mathcal{T}} \left(\widehat{\beta}_k(t) - \beta_k(t) \right)^2 dt \right\}^{1/2} = O_P\left(\left(\tfrac{1}{n^2} \sum_{i=1}^{n} \frac{1}{N_i} \right)^{q/(2q+1)} \right)$$

- asymptotic normality



(a)



(b)

Figure: Aids data. Fitted (a) baseline effect; (b) coefficient of pre-infection CD4.

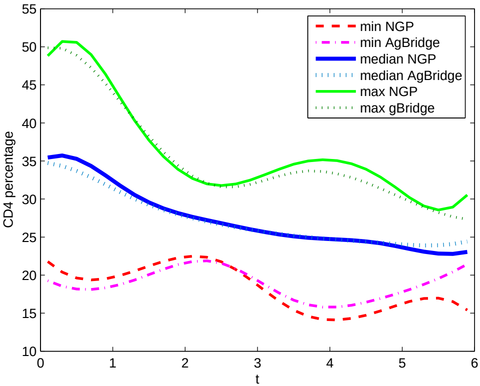


Figure: Aids data. Fitted CD4 percentage for person with minimum (-27.6841), median (-0.3841) and maximum (26.3159) centered pre-infection CD4.

Outline

if d is large, we need to **select** also which variables have an important influence $\Longrightarrow$ **variable selection**

simultaneous estimation and variable selection

- **estimation consistency**:

$$\widehat{\beta}_k - \beta_k \to 0, \text{ as } n \to \infty \text{ (e.g. in } L_2 \text{ sense)} \qquad + \text{ rate}$$

- **variable selection consistency**:

  suppose that the true $\beta_k = 0$; then we want

$$P\left\{\widehat{\beta}_k \neq 0\right\} \to 0, \text{ as } n \to \infty$$

we discuss briefly a variable selection method for additive models and for varying coefficient models

• additional regression models

$\widehat{f}_j^{\text{init}}(X_j)$ an initial estimator of $f_j(X_j)$

nonnegative garrote variable selection method then consists of finding the nonnegative garrote shrinkage factors c_j via the minimization problem:

$$\begin{cases} \min_{c_1,\ldots,c_d} \left\{ \sum_{i=1}^{n} \left(Y_i - \widehat{f}_0^{\text{init}} - \sum_{j=1}^{d} c_j \widehat{f}_j^{\text{init}}(X_{ij}) \right)^2 + \lambda \sum_{j=1}^{d} c_j \right\} \\ \text{subject to } 0 \leq c_j \text{, for } j = 1, \ldots, d \end{cases}$$

denote by $(\widehat{c}_1, \ldots, \widehat{c}_d)$, the solution to this minimization problem

the associated nonnegative garrote estimator for the function f_j is given by

$$\widehat{f}_j^{\text{NNG}}(\cdot) = \widehat{c}_j \widehat{f}_j^{\text{init}}(\cdot)$$

Yuan (2007), Cantoni *et al.* (2011) and Antoniadis *et al.* (2012b), Huang *et al.* (2010) and Marra and Wood (2011)....

• **varying coefficient models** variable selection for the varying coefficient model, based on longitudinal data

obtain nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_1, \ldots, \widehat{c}_d)$ from the optimization problem

$$\left\{ \begin{array}{l} \displaystyle\min_{c_1,\ldots,c_d} \left\{ \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(Y_{ij} - \widehat{\beta}_0^{\text{init}}(t_{ij}) - \sum_{p=1}^{d} X_{ij}^{(p)} c_p \widehat{\beta}_p^{\text{init}}(t_{ij}) \right)^2 + \lambda \sum_{p=1}^{d} c_p \right\} \\ \text{subject to } \; 0 \leqslant c_p \, , \text{for}\, p = 1, \ldots, d \end{array} \right.$$

$\widehat{\beta}_p^{\text{init}}(\cdot)$ is an initial estimator for the regression coefficient function $\beta_p(\cdot)$

Antoniadis *et al.* (2012a) and Verhasselt (2014)

Wang *et al.* (2008) and Xue and Qu (2012), ...

$\Longrightarrow$ **grouped regularization techniques**

Outline

varying coefficient models

$$\boxed{Y(\mathbf{t}) = \beta_0(\mathbf{t}) + \beta_1(\mathbf{t})X^{(1)}(\mathbf{t}) + \ldots + (\mathbf{t})\beta_d(\mathbf{t})X^{(d)} + \varepsilon(\mathbf{t})}$$

$q_\tau\left(\varepsilon(t)|X^{(1)}(t), \ldots, X^{(d)}(t)\right) = 0$

$\varepsilon(t)$ independent of $(X^{(1)}(t), \ldots, X^{(d)}(t), t)$

second aim: estimate **τth conditional quantile function** $(0 < \tau < 1)$

$$q_\tau(Y(t)|\mathbf{X}(t), t) = \beta_{\mathbf{0}}(\mathbf{t}) + \beta_{\mathbf{1}}(\mathbf{t})X^{(1)}(t) + \ldots + \beta_{\mathbf{d}}(\mathbf{t})X^{(d)}(t)$$

the conditional quantile

$$q_\tau(Y(t)|\mathbf{X}(t), t) = \beta_{\mathbf{0}}(\mathbf{t}) + \beta_{\mathbf{1}}(\mathbf{t})X^{(1)}(t) + \ldots + \beta_{\mathbf{d}}(\mathbf{t})X^{(d)}(t)$$

can be approximated via normalized B-splines

unknown regression coefficient functions $\beta_k(\cdot)$: can be of different degree of smoothness; B-splines of degree ν_k to approximate the coefficient function $\beta_k(t)$, for $k = 0, \ldots, d$:

$$\begin{aligned} \beta_k(t) \approx \alpha_{k1} B_{k1}(t;\nu_k) + \ldots + \alpha_{km_k} B_{km_k}(t;\nu_k) &= \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t;\nu_k) \\ &= \boldsymbol{\alpha}_k^T \mathbf{B}_k(t;\nu_k) \end{aligned}$$

$$\boldsymbol{\alpha}_{\mathbf{k}} = (\alpha_{\mathbf{k1}}, \ldots, \alpha_{\mathbf{km_k}})^{\mathbf{T}} \qquad \mathbf{B}_k(t;\nu_k) = (B_{k1}(t;\nu_k), \ldots, B_{km_k}(t;\nu_k))^T$$

$m_k = u_k + \nu_k$ $\qquad$ $u_k + 1 =$ number of knot points

estimation of global vector of **all unknown coefficients**
$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^{\mathbf{T}}, \ldots, \boldsymbol{\alpha}_{\mathbf{p}}^{\mathbf{T}})^{\mathbf{T}}$

quality of the fit measured via the goodness-of-fit quantity

$$\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \rho_\tau \left(Y_{ij} - \sum_{k=0}^{p} \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t_{ij}; \nu_k) X_{ij}^{(k)} \right)$$

reducing the modelling bias: use a large number of basis functions

but this leads to overfitting

... prevent this to happen by adding a penalty term

... adding a penalty term: minimize

$$\boxed{\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \rho_\tau \left(Y_{ij} - \sum_{k=0}^{d} \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t_{ij}; \nu_k) X_{ij}^{(k)} \right) + \sum_{k=0}^{d} \sum_{\ell=d_k+1}^{m_k} \lambda_k \left| \Delta^{d_k} \alpha_{k\ell} \right|^{\gamma}}$$

where $\gamma > 0$

$\lambda_k > 0$, $k = 0, \ldots, d$: smoothing parameters

Δ^{d_k} = the d_kth order differencing operator of the kth variable, with $d_k \in \mathbb{N}$

denote by $\widehat{\boldsymbol{\alpha}}_k$ the resulting P-splines estimator for the vector $\boldsymbol{\alpha}_k$, $k = 0, \ldots, d$

estimator for the τth conditional quantile function?

$$
\begin{aligned}
q_\tau(Y(t)|\mathbf{X}(t), t) &= \boldsymbol{\beta_0(t)} + \boldsymbol{\beta_1(t)} X^{(1)}(t) + \ldots + \boldsymbol{\beta_d(t)} X^{(d)}(t) \\
&\approx \sum_{k=0}^{d} \sum_{\ell=1}^{m_k} \alpha_{k\ell} B_{k\ell}(t_{ij}; \nu_k) X_{ij}^{(k)}
\end{aligned}
$$

P-splines estimator of the conditional regression quantile :

$$
\widehat{q}_\tau(Y_{ij}|\mathbf{X}_{ij}, t_{ij}) = \sum_{k=0}^{d} \sum_{\ell=1}^{m_k} \widehat{\alpha}_{k\ell} B_{k\ell}(t_{ij}; \nu_k) X_{ij}^{(k)}
$$

important issues :

- choices of γ, λ_k's, ...
- how to solve the optimization problem (algorithms, ...)
- can we show consistency, asymptotic distributional results ?

Andriyana *et al.* (2014, 2015), ...

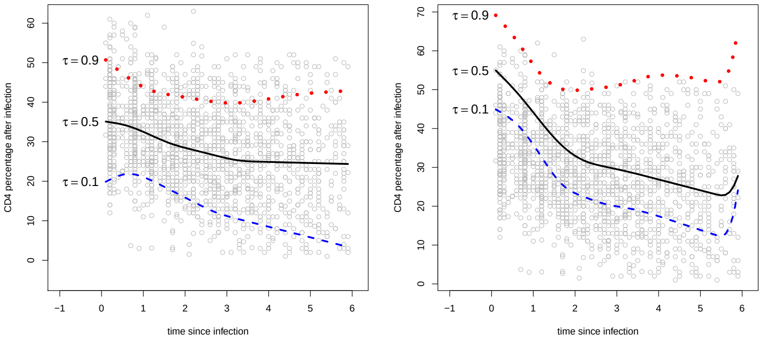


Figure: *Estimated quantile curves: $\tau = 0.1$ (dashed curves), $\tau = 0.5$ (solid curves) and $\tau = 0.9$ (dotted curves) for (left) median and (right) maximum of covariate values.*

median covariate case: nonsmoking, 32.6 years old patient, with pre-infection CD4 of 42.3%

$\tau = 0.5$: estimated to have a CD4 percentage of 24.37% after 6 years many issues not touched upon ...

- what if $d \gg n$?
- what if the variance/dispersion of the error term cannot assumed to be constant (heteroscedasticity)?

  can we estimate this heteroscedasticity in a flexible manner ?
- what about robust methods for variable selection ?
- how to prevent estimated quantile curves of different orders to cross ?
- what if data are not i.i.d. ?
- how to deal with functional data?