The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-23

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Generalized linear models and penalized likelihood regression

Irène Gijbels

Katholieke Universiteit Leuven
Department of Mathematics
B-3001 Leuven, Belgium

Generalized linear models and penalized likelihood regression

Irène Gijbels

Department of Mathematics & Leuven Statistics Research Centre

Katholieke Universiteit Leuven, Belgium

Joint work with Anestis Antoniadis and Mila Nikolova

Outline

- Introduction: models and basic elements
- Penalties and regularization
- Optimization of the penalized likelihood
- Statistical properties and Asymptotic analysis
- Choice of regularization parameters
- Simulations and example

Introduction: models and basic elements

Generalized models

Y: response variable $\qquad\qquad$ 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 linear models: $\eta(x) =$ a linear function of x

Examples

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

  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 $\eta(x) = \log\,(\mu(x))$

McCullagh & Nelder (1989)

regression analysis:

from observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

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

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

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

- nonparametric estimation: several techniques

penalized log-likelihood:

$$\boxed{\text{maximize} \quad Z_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \eta(x_i)) - \lambda J(\eta)}$$

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 controling trade-off between 2 terms flexible estimation approach:

represent $\eta(\cdot)$ as a linear combination of known basis functions $h_1(x), h_2(x), \cdots, h_p(x)$

$$\boxed{\eta(x) = \sum_{k=1}^{p} \beta_k h_k(x) \qquad k = 1, \ldots, p}$$

AIM: estimate the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$

examples of basis functions: wavelets, polynomial splines, ...

crucial choice: number p of basis functions

- small p: may not be flexible enough to capture variability of data
- large p: may lead to overfitting

regularization: use a highly parametrized model and impose a penalty on large fluctuations of fitted curve notations:

$$\mathbf{x} = (x_1, x_2, \cdots, x_n) \qquad \mathbf{y} = (y_1, y_2, \cdots, y_n) \qquad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$$

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_p(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_p(x_2) \\ \vdots & \vdots & & \vdots \\ h_1(x_i) & h_2(x_i) & \cdots & h_p(x_i) \\ \vdots & \vdots & & \vdots \\ h_1(x_n) & h_2(x_n) & \cdots & h_p(x_n) \end{pmatrix}$$ matrix of dim $n \times p$

$$\mathbf{h}(x_i) = (h_1(x_i), h_2(x_i), \cdots, h_p(x_i))$$ vector of dim $1 \times p$

objective function to be maximized in some function space

$$\boxed{Z_n(\boldsymbol{\beta}) = \tfrac{1}{n} \sum_{i=1}^{n} \ell(y_i, \mathbf{h}(x_i)\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}) \equiv \frac{1}{n} L_{\mathbf{y}}(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta})}$$

for given basisfunctions $h_1(\cdot), \cdots, h_p(\cdot)$, penalty function $J(\cdot)$ and smoothing parameter λ

$$\boxed{\text{maximize} \quad Z_n(\boldsymbol{\beta}) = \tfrac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{h}(x_i)\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}) \equiv \frac{1}{n} L_{\mathbf{y}}(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta})}$$

$$\eta(x_i) = \mathbf{h}(x_i)\boldsymbol{\beta} \qquad g(\mu(x_i)) = \eta(x_i) = \mathbf{h}(x_i)\boldsymbol{\beta} \qquad i = 1, \cdots, n$$

allow p to be large, and control the risk of overfitting the data by using an adequate penalty J on the coefficients

Eilers & Marx (1996), Ruppert & Carroll (2000), ...

what choice of basisfunctions?

Truncated power basis

knot points $t_1 < t_2 < \cdots < t_K$

d integer, $d \geq 1$

truncated power basis for polynomial of degree d regression splines with knots $t_1 < t_2 < \cdots < t_K$

$$\boxed{\{1, x, \ldots, x^d, (x - t_1)_+^d, \ldots, (x - t_K)_+^d\}} \qquad z_+ = \max(z, 0)$$

continuous up to $(d-1)$st derivative

representation of a univariate function f in terms of these $(d+1+K)$ basis functions

$$f(x) = \sum_{k=0}^{d} \beta_k x^k + \sum_{j=1}^{K} \beta_{p+j}(x - t_j)_+^d$$

each coefficient β_{d+j} is identified as a jump in the d-th derivative of f at the corresponding knot ($\longrightarrow$ easy interpretation)

sometimes not desirable because computationally less stable

de Boor (1978) and Dierckx (1993)

- normalized B-splines basis of order q with knots $0 < t_1 < \cdots < t_K < 1$: set of degree $(q-1)$ splines

$$\{B^q_{Kj}, j = 1, \ldots, q + K\}$$

- functions B^q_{Kj} are positive and have local support: are non-zero only on an interval which covers no more than q + 1 knots
- equivalently: at any point x there are no more than q B-splines that are non-zero
- recursive relationship to describe B-splines; provides a very stable numerical computation algorithm
- moderately large number of knots (usually between 20 and 40) to ensure enough flexibility
- quadratic penalty based on differences of adjacent B-spline coefficients to guarantee sufficient smoothness of fitted curves

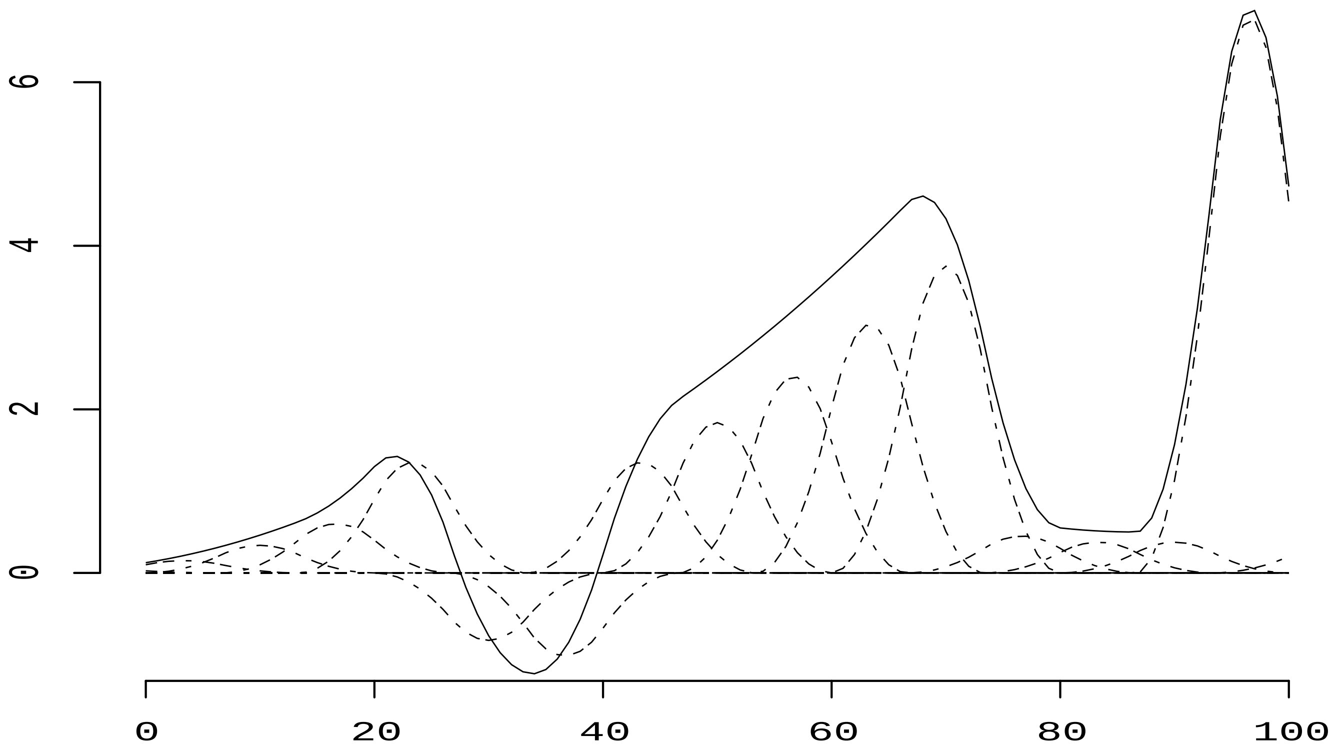

Figure 1: Illustration of B-spline constructed smooth curve.

dashed curves: scaled basis functions; heights are the coefficients

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

- quadratic regularization: $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$
- in the setting of Bayesian MAP estimation and Markov random fields (Geman & Clure (184, 1987), Besag (1974, 1989), ...):

$$\boxed{J(\boldsymbol{\beta}) = \sum_{k=1}^{r} \gamma_k \psi(d_k^T \boldsymbol{\beta})}$$

$\gamma_k > 0$ weights d_k linear operators

- for $\psi(\cdot)$ convex: J pushes solution $\widehat{\boldsymbol{\beta}}$ to be s.t. $|d_k^T \widehat{\boldsymbol{\beta}}|$ is small
- in particular: if d_k are finite difference operators, neighboring coefficients of $\widehat{\boldsymbol{\beta}}$ are encouraged to have similar values ($\widehat{\boldsymbol{\beta}}$ involves homogeneous zones)
- if $d_k = e_k$, then J encourages the components $\widehat{\boldsymbol{\beta}}_k$ to have small magnitude

- choice of $J(\boldsymbol{\beta})$ depends strongly on the basis functions used
- for a truncated power basis functions of degree d; coefficients of basis functions at the knots involve jumps of d-th derivative (large coeff. are associated with singularities in the fct):

$$\boxed{J(\boldsymbol{\beta}) = \sum_k \gamma_k \psi(\beta_k) \qquad \gamma_k > 0}$$

  no reason that neighboring coefficients of $\boldsymbol{\beta}$ have close values
- example: $\psi(\cdot) = |\cdot|$

  Mammen & Van de Geer (1997), Ruppert & Carroll (1997), Yu & Ruppert (2001), Antoniadis & Fan (2001), ...

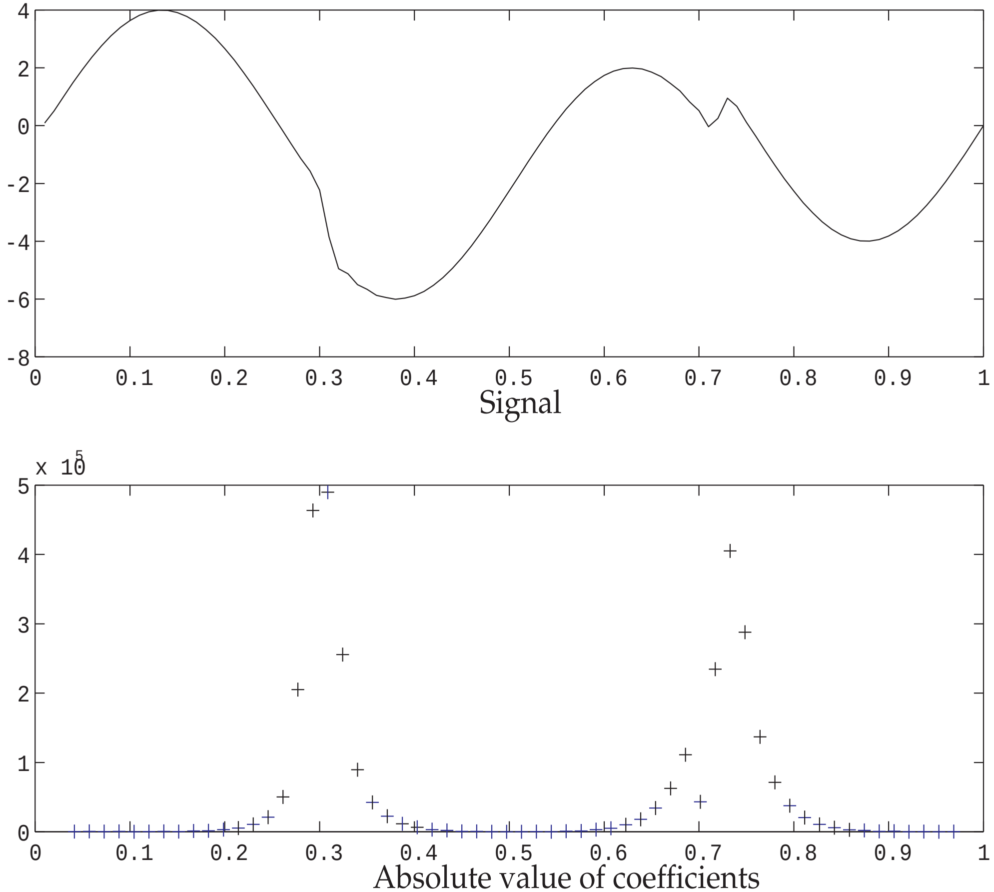

Figure 2: Behavior of coeff. of function in a truncated power basis.

- for B-splines basis: penalties on neighbor B-spline coeff. ensure that neighboring coeff. do not differ too much from each other when η is smooth

- absolute values of first order or second order differences are maximum at singularity points of curve

- penalties such as $J(\boldsymbol{\beta}) = \sum^{r} \gamma_k \psi(d_k^T \boldsymbol{\beta})$ are more adequate

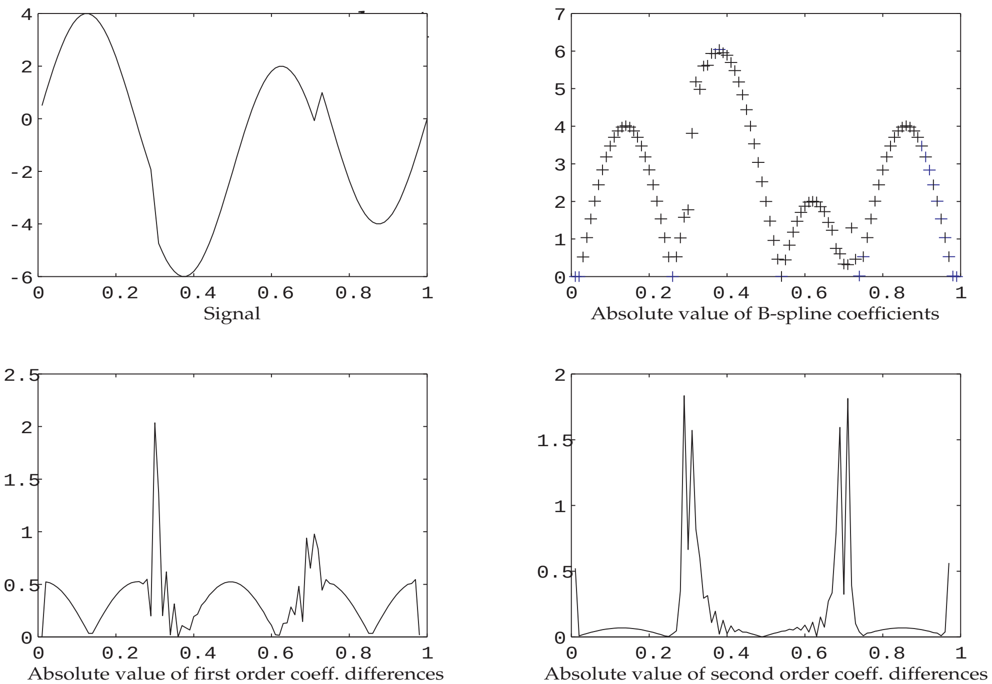

Figure 3: Behavior of coefficients of function in a B-splines basis.

$$J(\boldsymbol{\beta}) = \sum_k \gamma_k \psi(\beta_k) \qquad\qquad J(\boldsymbol{\beta}) = \sum_{k=1}^{r} \gamma_k \psi(d_k^T \boldsymbol{\beta})$$

general type of penalty functions $\psi(\cdot)$

- ◇ L_2 or quadratic penalty $\psi(\beta) = |\beta|^2$ ridge type regression
- ◇ L_1 penalty $\psi(\beta) = |\beta|$ LASSO type regression

  Donoho & Johnstone (1994), Tibshirani (1996), Klinger (2000) ...

- ◇ L_q $(0 \le q \le 1)$ penalty $\psi(\beta) = |\beta|^q$ bridge regression

  Frank & Friedman (1993), Ruppert & Carroll (1997), Fu (1998), Knight & Fu (2000), Yu & Ruppert (2001), ...

- usually: ψ symmetric around 0 and increasing on $[0, +\infty)$
- ψ can be convex or non-convex, smooth or non-smooth

what is a good penalty function? Antoniadis & Fan (2001)

- ○ gives an estimator that avoids excessive bias (unbiasedness)
- ○ forces sparse solutions to reduce model complexity (sparsity)
- ○ avoids unnecessary variation (stability)
- ○ from computational viewpoint: resulting optimization problem should be (easily) solvable

AIM: summarize and unify mean features of $\psi(\cdot)$ that determine essential properties of maximizer $\widehat{\boldsymbol{\beta}}$ of $Z_n(\boldsymbol{\beta})$

Convex	
Smooth at zero	Singular at zero
1. $\psi(\beta) = \|\beta\|^\alpha,\ \alpha > 1$	6. $\psi(\beta) = \|\beta\| \quad \psi'(0^+) = 1$
2. $\psi(\beta) = \sqrt{\alpha + \beta^2}$	7. $\psi(\beta) = \alpha^2 - (\|\beta\| - \alpha)^2 I\{\|\beta\| < \alpha\}$
3. $\psi(\beta) = \log(\cosh(\alpha\beta))$	$\psi'(0^+) = 2\alpha$
4. $\psi(\beta) = \beta^2 - (\|\beta\| - \alpha)^2 I\{\|\beta\| > \alpha\}.$	
5. $\psi(\beta) = 1 + \|\beta\|/\alpha - \log\left(1 + \|\beta\|/\alpha\right)$	

Nonconvex	
Smooth at zero	Singular at zero
8. $\psi(\beta) = \alpha\beta^2/(1 + \alpha\beta^2)$	12. $\psi(\beta) = \|\beta\|^\alpha, \alpha \in (0,1) \quad \psi'(0^+) = \infty$
9. $\psi(\beta) = \min\{\alpha\beta^2, 1\}$	13. $\psi(\beta) = \alpha\|\beta\|/(1 + \alpha\|\beta\|) \quad \psi'(0^+) = \alpha$
10. $\psi(\beta) = 1 - \exp\left(-\alpha\beta^2\right)$	14. $\psi(0) = 0,\ \psi(\beta) = 1, \forall\beta \neq 0$ discont.
11. $\psi(\beta) = -\log\left(\exp(-\alpha\beta^2) + 1\right)$	15. $\psi(\beta) = \log(\alpha\|\beta\| + 1) \quad \psi'(0^+) = \alpha$
	16. $\int_0^\beta \psi'(u)du \quad \psi'(\|\beta\|)$ $= \alpha\{I\{\|\beta\| \leq \alpha\} + \frac{(a\alpha - \|\beta\|)_+}{(a-1)\alpha}\{\|\beta\| > \alpha\}\}$ $a > 2$

Penalties and regularization

Smooth regularization: $\boxed{J(\boldsymbol{\beta}) = \sum_{k=1}^{r} \gamma_k \psi(d_k^T \boldsymbol{\beta})}$

- Convex penalties: typically consider $\boxed{J(\boldsymbol{\beta}) = \boldsymbol{\beta}^T D(\boldsymbol{\gamma}) \boldsymbol{\beta}}$

  $D(\boldsymbol{\gamma})$ positive definite matrix; examples:

  - $D(\boldsymbol{\gamma})$ diagonal matrix with elements γ_k

$$J(\boldsymbol{\beta}) = \sum_k \gamma_k \beta_k^2 \qquad \psi(\beta) = \beta^2 \qquad d_k = e_k$$

  - $D(\boldsymbol{\gamma})$ a banded matrix corresponding to a quadratic form of finite differences of components of $\boldsymbol{\beta}$

how to solve the optimization problem?

- for fixed λ and γ: estimator of β is obtained recursively by an iterated re-weighted least squares algorithm (cfr generalized linear models)
- with quadratic regularization: more or less like classical maximum penalized likelihood; may not be acceptable when the function to recover is less regular

- in the later case use non-quadratic convex penalties
  - EXAMPLE: hyperbolic potential $\psi(t) = \sqrt{\alpha + t^2}$ is very frequently used
    is a smooth approximation to $|t|$, since $\psi(t) \to |t|$ as $\alpha \searrow 0$

  - main characteristics of these functions (cfr 1—5 in Table 1): $\psi(\cdot)$ has a strict minimum at zero and $\psi'(\cdot)$ is almost constant (but > 0) except in a nhd of the origin
  - when $L_{\mathbf{y}}$ is strictly concave and ψ is convex, or $L_{\mathbf{y}}$ is concave and ψ is strictly convex, the penalized log-likelihood $Z_n(\boldsymbol{\beta})$ is guaranteed to have a unique maximizer

- Non-convex penalties
  - typically $\psi(t)$ is (nearly) constant for large values of $|t|$ (cfr 8—11 in Table 2)
  - main difficulty: the penalized log-likelihood $Z_n(\boldsymbol{\beta})$ is non-concave and may exhibit a large number of local maxima
  - no way to guarantee the finding of a global maximizer
  - computational cost is generally high

Non-smooth regularization: $\boxed{J(\boldsymbol{\beta}) = \sum_{k=1}^{r} \gamma_k \psi(d_k^T \boldsymbol{\beta})}$

to estimate less regular fct's: use penalties that are singular at zero

- L_1 LASSO penalty: $\psi(\beta) = |\beta|$ non-smooth at zero, but convex ($\longrightarrow$ sparse solutions, asympt. optimal minimax estimators, ...)
- hyperbolic potential $\psi(\beta) = \sqrt{\alpha + \beta^2}$ is a smooth version of the LASSO penalty, also convex
- Smoothed Clipped Absolute Deviation (SCAD) penalty (cfr nr 16)

  non-smooth, non-convex

solving the optimization problem?

non-convex penalties: difficult (or even impossible) task
convex non-smooth at the origin penalties: feasible task (see later)

Optimization of the penalized likelihood

general: some elements from optimization theory

- the function $\boldsymbol{\beta} \to -Z_n(\boldsymbol{\beta})$ is said to be coercive if

$$\lim_{\|\boldsymbol{\beta}\| \to +\infty} -Z_n(\boldsymbol{\beta}) = +\infty$$

- since $J(\boldsymbol{\beta})$ is nonnegative, function $\boldsymbol{\beta} \to J(\boldsymbol{\beta})$ is bounded by below

  if in addition $\boldsymbol{\beta} \to L_{\mathbf{y}}(\boldsymbol{\beta})$ is bounded above, then $-Z_n$ is coercive if at least one of the two terms J or $-L_{\mathbf{y}}$ is coercive

- for Gaussian and Poisson nonp. GLM models, $-Z_n(\boldsymbol{\beta})$ is coercive

- for Bernoulli nonparametric GLM model, $-Z_n(\boldsymbol{\beta})$ is not coercive

  the addition of a suitable penalty term (e.g. a quadratic term) to $J(\boldsymbol{\beta})$ makes $-Z_n(\boldsymbol{\beta})$ coercive (see e.g. Park & Hastie (2006)) in general: existence and uniqueness of solutions

- if $-Z_n$ is coercive, for every $c \in I\!R$, the set $\{\boldsymbol{\beta} : -Z_n(\boldsymbol{\beta}) \leq c\}$ is bounded
- if Z_n is continuous the value $\sup_{\boldsymbol{\beta}} Z_n$ is finite and the set of the optimal solutions $\{\widehat{\boldsymbol{\beta}} \in I\!R^p : Z_n(\widehat{\boldsymbol{\beta}}) = \sup_{\boldsymbol{\beta} \in I\!R^p} Z_n\}$ is nonempty and compact
- in general, beyond its global maxima, Z_n may exhibit local maxima
- if in addition Z_n is strictly concave, then for every $\mathbf{y} \in I\!R^n$, there is a unique maximizer
- analyzing the maximizers of a non-concave Z_n is much more difficult
- in the Gaussian case with $\mathbf{H}^T\mathbf{H}$ invertible and J non-convex, the regularity of local and global maximizers of Z_n has been studied by Durand & Nikolova (2005) and Nikolova (2005)

assume: penalties are symmetric and nonnegative

consider 2 situations in our nonparametric GLM models:

Geman's class of penalties	and	δ-class of penalties

$\diamond$ Geman's class of penalties: functions ψ satisfying

- ψ is in $\mathcal{C}^2$ and convex on $[0, +\infty[$
- $t \rightarrow \psi(\sqrt{t})$ is concave $[0, +\infty[$
- $\psi'(t)/t \rightarrow M < \infty$ as $t \rightarrow \infty$
- $\lim\limits_{t \nearrow 0} \psi'(t)/t$ exists

we have shown the existence of a unique solution and discuss a computational algorithm to find it (via half-quadratic optimization)

examples of such penalties: numbers 2, 3, 4 and 5 in Table 1

◇ δ-class of penalties: penalties with properties

- ψ is monotone increasing on $[0, +\infty[$
- ψ is in $\mathcal{C}^1$ on $\mathbb{R}\backslash\{0\}$ and continuous in 0
- $\lim_{t\to 0} \psi'(t)t = 0$

named δ-class: since it essentially consists of penalties that are non smooth at the origin but can be approximated by a quadratic function in a δ-nhd of the origin

for this class we will find an approximate solution to the optimization problem and provide bias and variance expressions

Optimalization with penalties in the δ-class

how to deal with nondifferentiabilty of such penalties?

approximate penalized log-likelihood $Z_n(\boldsymbol{\beta})$ by $Z_\delta(\boldsymbol{\beta})$ by replacing penalty $J(\boldsymbol{\beta}) = \sum_k \gamma_k \psi(\beta_k)$ by $J_\delta(\boldsymbol{\beta}) = \sum_k \gamma_k \psi_\delta(\beta_k)$

ψ_δ : fct equal to ψ away from 0 (at a distance $\delta > 0$) and a "smooth quadratic" version of ψ in a δ-nhd of zero (e.g. Tishler & Zang (1982))

- define smooth version of ψ:

$$\boxed{\psi_\delta(s) = \begin{cases} \psi(s) & \text{if} \quad s > \delta \\ \frac{\psi'(\delta)}{2\delta} s^2 + [\psi(\delta) - \psi'(\delta)\delta/2] & \text{if} \quad 0 \le s \le \delta \end{cases}}$$

- then

$$\psi''_\delta(s) = \begin{cases} \psi''(s) & \text{if} \quad s > \delta \\ \frac{\psi'(\delta)}{\delta} & \text{if} \quad 0 \le s \le \delta \end{cases}$$

  and for all $s \ge 0$ $\quad \lim_{\delta \downarrow 0} \psi_\delta(s) = 0$

- score function for the approximate penalized log-likelihood $Z_\delta(\boldsymbol{\beta})$

$$\boxed{u_\delta(\boldsymbol{\beta}) = s(\mathbf{y}, \boldsymbol{\beta}) + \lambda D(\boldsymbol{\gamma})\mathbf{g}_\delta(\boldsymbol{\beta})}$$

$s(\mathbf{y}, \boldsymbol{\beta}) = (\partial L_{\mathbf{y}}(\boldsymbol{\beta})/\partial \beta_j)_{j=1,\dots,p}$

$\mathbf{g}_\delta(\boldsymbol{\beta})$ = $(p \times 1)$ vector with corresponding j-th component $g_\delta(|\beta_j|)$

$$g_\delta(|\beta_j|) = \begin{cases} -\psi'_\delta(|\beta_j|) & \text{if} \quad \beta_j \geq 0 \\ +\psi'_\delta(|\beta_j|) & \text{if} \quad \beta_j < 0 \end{cases}$$

- for any $\boldsymbol{\beta}$ fixed:

$$\lim_{\delta \downarrow 0} \mathbf{g}_\delta(\boldsymbol{\beta}) = \mathbf{g}(\boldsymbol{\beta})$$

$\mathbf{g}(\boldsymbol{\beta}) = (g(|\beta_1|), \dots, g(|\beta_p|))^T$ with $g(|\beta_p|) = \psi'(\beta_j|)I\{\beta_j \neq 0\}$

- score function $u_\delta(\boldsymbol{\beta})$ converges to $u(\boldsymbol{\beta})$ as $\delta \downarrow 0$, where

$$\boxed{u(\boldsymbol{\beta}) = s(\mathbf{y}, \boldsymbol{\beta}) + \lambda D(\boldsymbol{\gamma})\mathbf{g}(\boldsymbol{\beta})}$$

- $\widehat{\boldsymbol{\beta}}(\delta)$, a root of approximate penalized score equations (i.e. $u_\delta(\widehat{\boldsymbol{\beta}}(\delta)) = 0$)
- since penalty function ψ_δ is strictly convex, such an estimator exists and is unique even in situations where the maximum likelihood principle diverges
- fast computation of the estimator can be done by standard Fisher scoring procedure

Statistical properties & Asymptotic analysis

Bias and variance $p < n$ for δ-class penalties

- sample bias and variance properties
- for fixed diagonal matrix $D(\boldsymbol{\gamma})$ of weights and fixed penalization parameter λ: let $\boldsymbol{\beta}^*$ be a maximizer of the expected penalized log-likelihood
- in case of uniqueness: equivalent to root of the expected penalized score equation, i. e. $\mathbb{E}(u(\boldsymbol{\beta}^*)) = 0$
- what is the estimation error induced by our regularized procedure?

linear Taylor expansion

$$0 = u_\delta(\widehat{\boldsymbol{\beta}}(\delta)) \approx u_\delta(\boldsymbol{\beta}^*) + \{\mathbf{H}_L(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma}) G(\boldsymbol{\beta}^*;\delta)\} (\widehat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}^*)$$

$G(\boldsymbol{\beta}^*;\delta)$ = diag. matrix with entries $\partial g_\delta(|\beta_j|)/\partial\beta_j = \psi''_\delta(|\beta_j|)$

we get : $\boxed{\widehat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}^* \approx \{\mathbf{H}_L(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma}) G(\boldsymbol{\beta}^*;\delta)\}^{-1} u_\delta(\boldsymbol{\beta}^*)}$

- since β^* is a root of $\mathbb{E}(u(\boldsymbol{\beta}))$, we have $\mathbb{E}(u_\delta(\boldsymbol{\beta}^*)) = \lambda D(\boldsymbol{\gamma}) \mathbf{g}_\delta(\boldsymbol{\beta}^*)$ and therefore
- $\widehat{\boldsymbol{\beta}}(\delta)$ has bias $\boxed{\{\mathbf{H}_L(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma}) G(\boldsymbol{\beta}^*;\delta)\}^{-1} \mathbb{E}(u_\delta(\boldsymbol{\beta}^*))}$
- $$\boxed{\text{var}(\widehat{\boldsymbol{\beta}}(\delta)) = \{\mathbf{H}_L(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma}) G(\boldsymbol{\beta}^*;\delta)\}^{-1} \text{var}(s(\mathbf{y},\boldsymbol{\beta}^*)) \{\mathbf{H}_L(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma}) G(\boldsymbol{\beta}^*;\delta)\}^{-1}}$$
- bias and variance depend on the behavior of the eigenvalues of $\{\mathbf{H}_L(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma}) G(\boldsymbol{\beta}^*;\delta)\}^{-1}$ and their limits as $\delta \downarrow 0$ with $\lambda > 0$ fixed ($\longrightarrow$ detailed study)

General Asymptotic Analysis

AIM: obtain asymptotic results of estimators $\widehat{\boldsymbol{\beta}}_n$ minimizing

$$-Z_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{h}(x_i)\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}) \equiv \frac{1}{n} L_{\mathbf{y}}(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta})$$

2 cases: p fixed and finite and $p = p_n$ and $p_n \to \infty$

case p fixed and finite

under regularity conditions (on the log-likelihood; cfr conditions that guarantee normality of ordinary MLE)

$$a_n = \lambda_n \max\{\gamma_j \psi'(|\beta_{0j}|); \beta_{0j} \neq 0\} < \infty$$

THEOREM: Let the probability density of our model satisfy the regularity conditions. Assume $\lambda_n \to 0$ as $n \to \infty$. If $b_n := \lambda_n \max\{\gamma_j |\psi''(|\beta_{0j}|)|; \beta_{0j} \neq 0\} \to 0$, then there exists a local minimizer $\widehat{\boldsymbol{\beta}}_n$ of the penalized likelihood such that $\|\widehat{\boldsymbol{\beta}} - \beta_0\| = O_P(n^{-1/2} + a_n)$

case $p = p_n$ and $p_n \to \infty$

for some non-concave penalized likelihood function; see e.g. Fan & Peng (2004)

Regularity conditions (on penalty and on growth rate of dim. p_n)

(a) $\liminf_{\beta \to 0^+} \psi'(\beta) > 0$

(b) $a_n = O(n^{-1/2})$

(c) $a_n = o((np_n)^{-1/2})$

(d) $b_n = \max_{1 \leq j \leq p_n} \{\gamma_j |\psi''(|\beta_j|)|; \beta_j \neq 0\} \to 0$

(e) $b_n = o_P(p_n^{-1/2})$

(f) exists C and D such that when x_1 and $x_2 > C\lambda_n$, $\lambda_n |\psi''(x_1) - \psi''(x_2)| \leq D|x_1 - x_2|$

under such conditions previous theorem extends to case $p_n \to \infty$

Choice of the regularization parameters

- L-curve approach adapted to Generalized linear model context

  Belge, Kilmer & Miller (2002)

- Alternative approach

  estimated predictor depends on scaling of basisfct's

  overcoming drawback by standardizing basisfct's in advance

$$\overline{h}_j = \frac{1}{n}\sum_{i=1}^{n} h_j(x_i) \qquad\qquad \widetilde{s}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left[h_j(x_i) - \overline{h}_j\right]^2$$

  adjust threshold parameters γ_k appropriately: $\gamma_k = \sqrt{\widetilde{s}_k^2}$

  with this choice, any scaled version $\kappa[\mathbf{H}(\mathbf{x})]_j$ would yield the threshold $\widetilde{\gamma}_k = |\kappa|\gamma_k$

  data-driven choices: $\gamma_k = \sqrt{\widetilde{s}_k^2}$, select λ by Generalized Cross Validation

Simulations and example

test functions: with jumps or with discontinuities in derivatives

Quadratic loss

Gaussian noise

2 test functions: heavisine function and corner function

100 simulations in each experiment (same design points each time; from uniform U$(0, 1)$)

signal-to-noise ratio is 4 (= $\sqrt{\text{Var}(f(X))/\sigma^2}$) $\quad n = 200$

4 procedures (all based on regression splines):

- Ridge regression (quadratic loss and L_2 penalty on coeff.)
- LASSO regression (quadratic loss and L_1 penalty on coeff.)
- SARS Spatially Adaptive Regression Splines (Zhou & Shen (2001))
- Half-Quadratic regularization procedure (quadratic loss and hyperbolic potential $\psi(\beta) = \sqrt{\alpha + \beta}$; convex and smooth)

truncated power basis of degree 3, with 40 equispaced knots;

threshold parameters selected adjusting to stdev of each basis function; smoothing parameter λ selected by 10-fold GCV

for SARS procedure: default values of hyperparameters

measure of quality: $\boxed{\text{MASE}\,(\widehat{\eta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}(x_i) - \eta(x_i)\right)^2}$

Poisson regression

$Y_i \sim$ Poisson$(\mu(x_i))$ $\qquad$ $\mu(\cdot)$ = exponential (heavisine function)

SARS not designed for treating Poisson distributed data

3 procedures:

- Ridge regression
- Half-Quadratic regularization procedure
- SPIC procedure by Imoto & Konishi (2003); B-splines procedure based on an information criterion

truncated power basis of degree 3, with 40 equispaced knots;

threshold and smoothing parameters: as before

for SPIC procedure: B-splines with 30 knots; smoothing parameter selected by SPIC procedure

Analysis of AIDS data

- AIDS data (Stasinopoulos & Rigby (1992))
- concerns the quarter yearly frequency count of reported AIDS cases in the UK from January 1983 to September 1990
- after deseasonalising this time series, one suspects a break in the relationship between the number of AIDS cases and the time measured in quarter years
- model Y (deseasonalised frequency of AIDS cases) by a Poisson distribution with mean a polynomial spline function of x, the time measured in quarter years
- use half quadratic procedure (HQ) with spline basis based on 12 knots

seemingly a break point at about July 1987 as also suggested by Stasinopoulos & Rigby (1992)

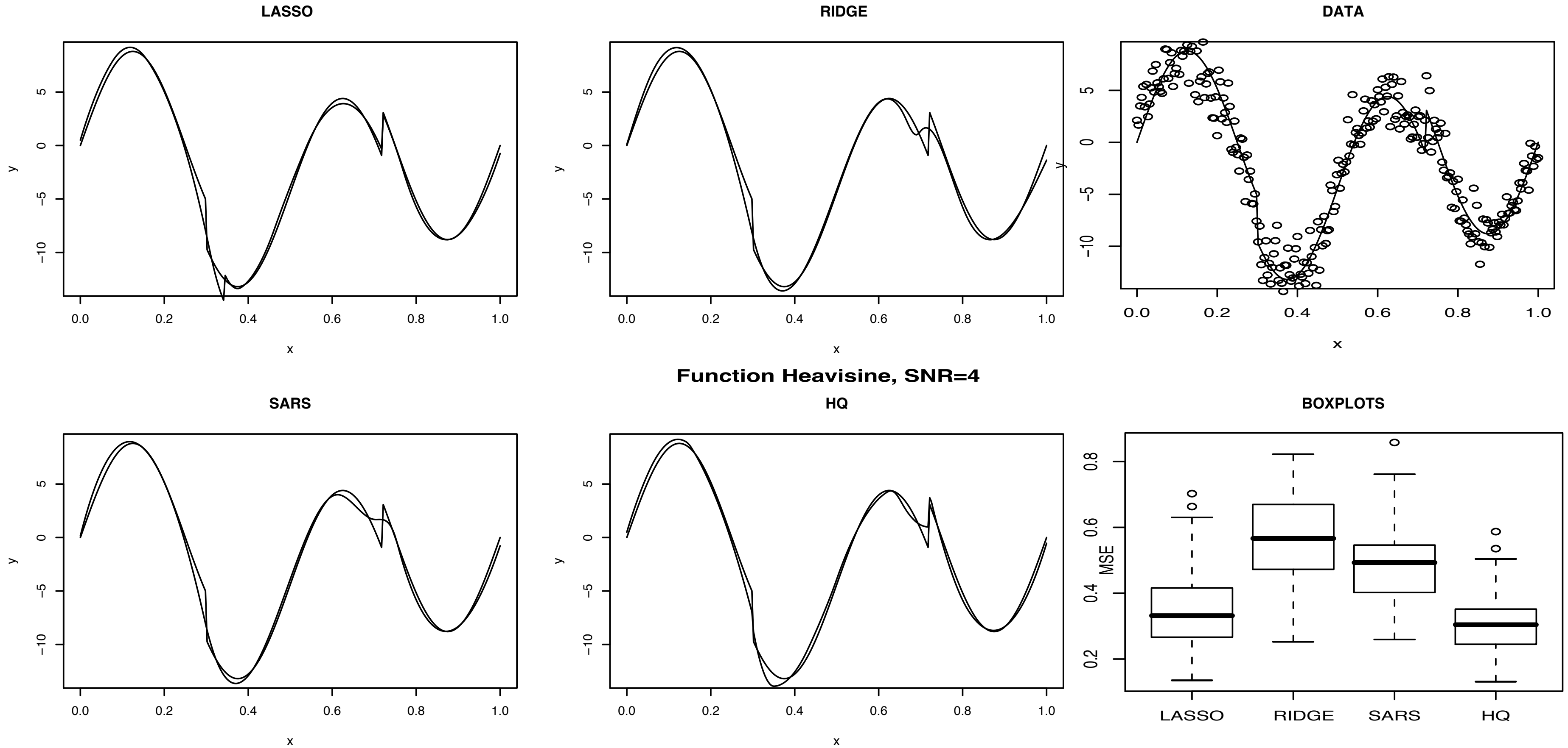

Figure 4: Simulated example: Gaussian noise; `heavisine` function.

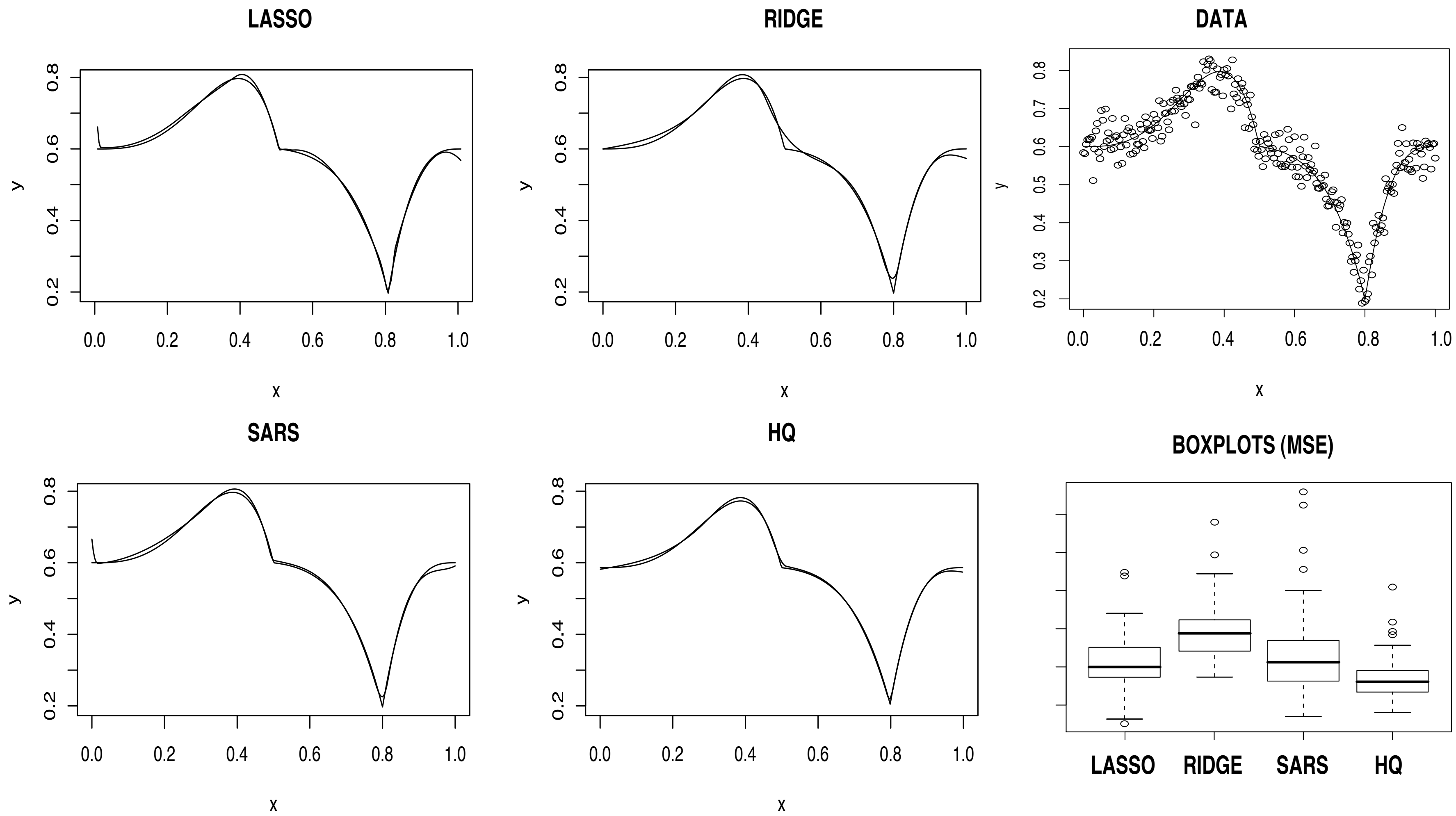

Figure 5: Simulated example: Gaussian noise; `corner` function.

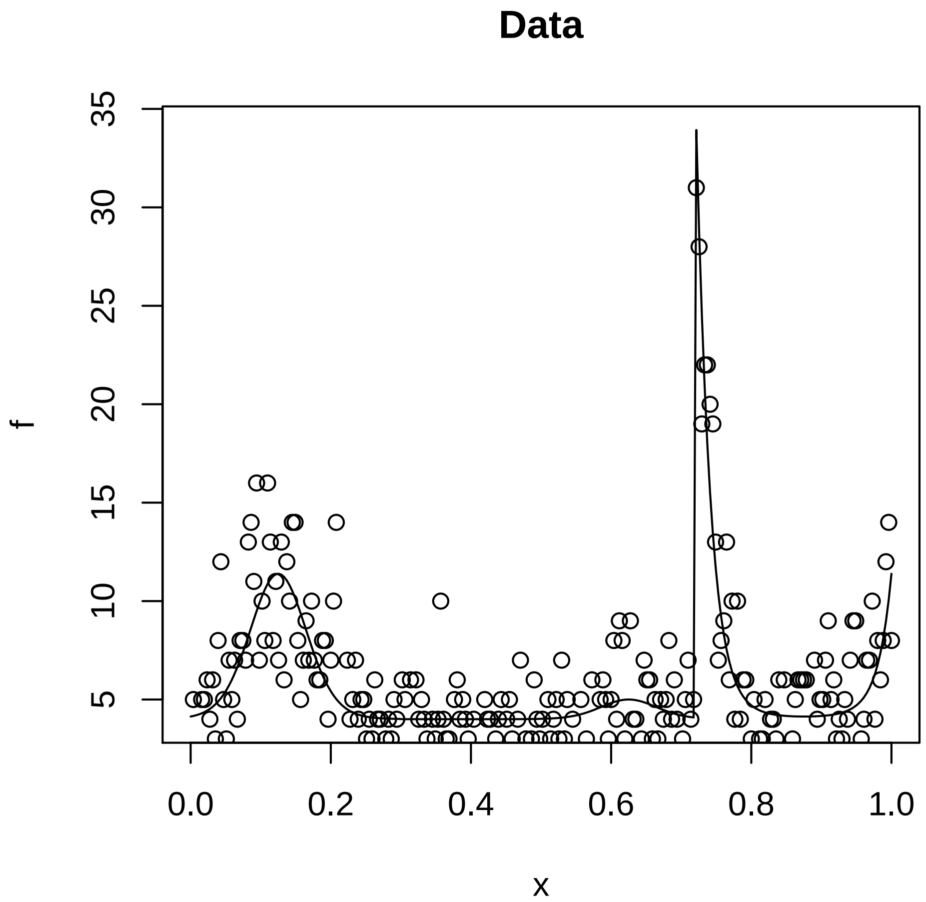

Figure 6: Simulated data: Poisson regression; `exp(heavisine)` function.

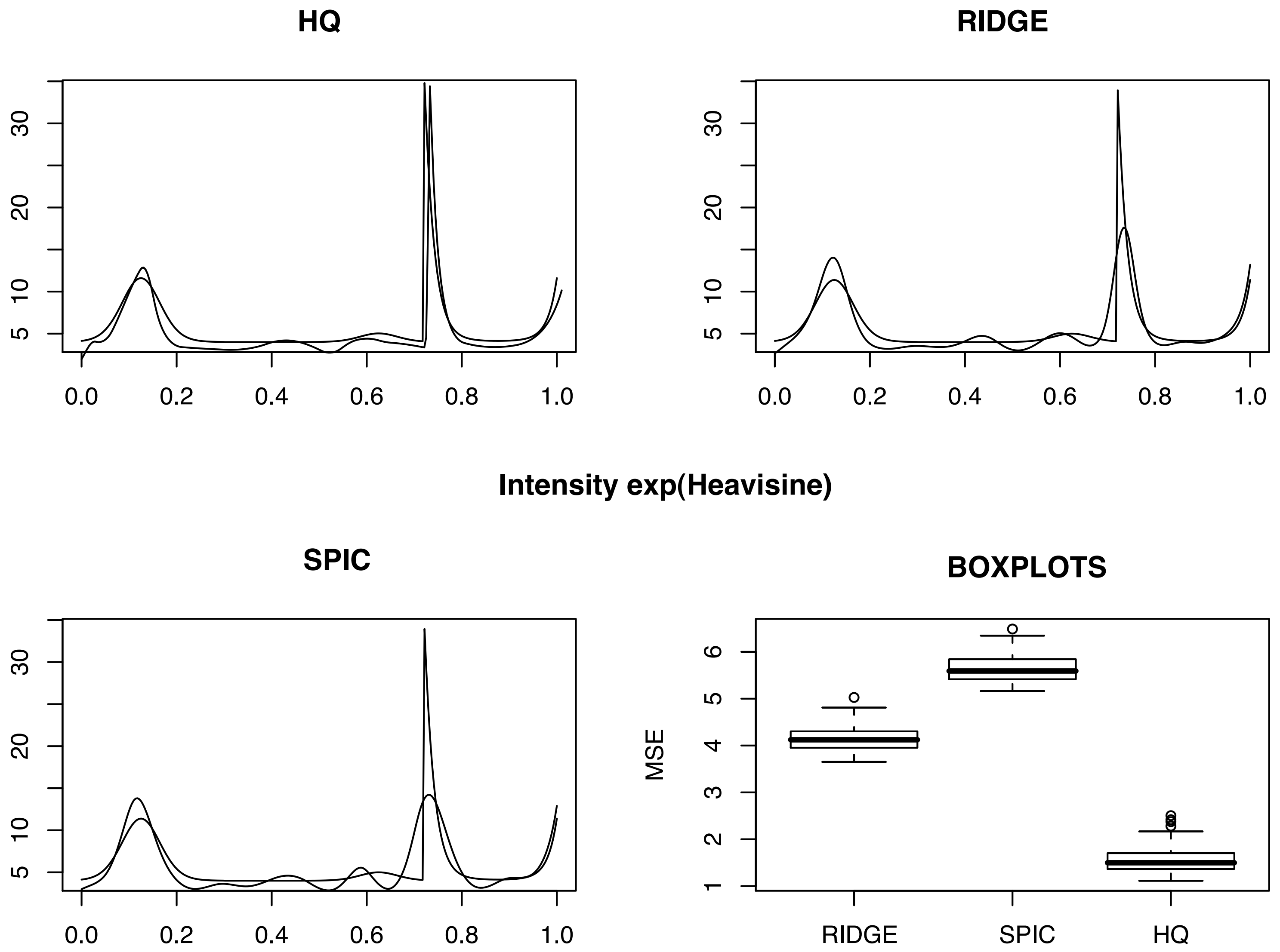

Figure 7: Simulated data: Poisson regression; `exp(heavisine)` function.

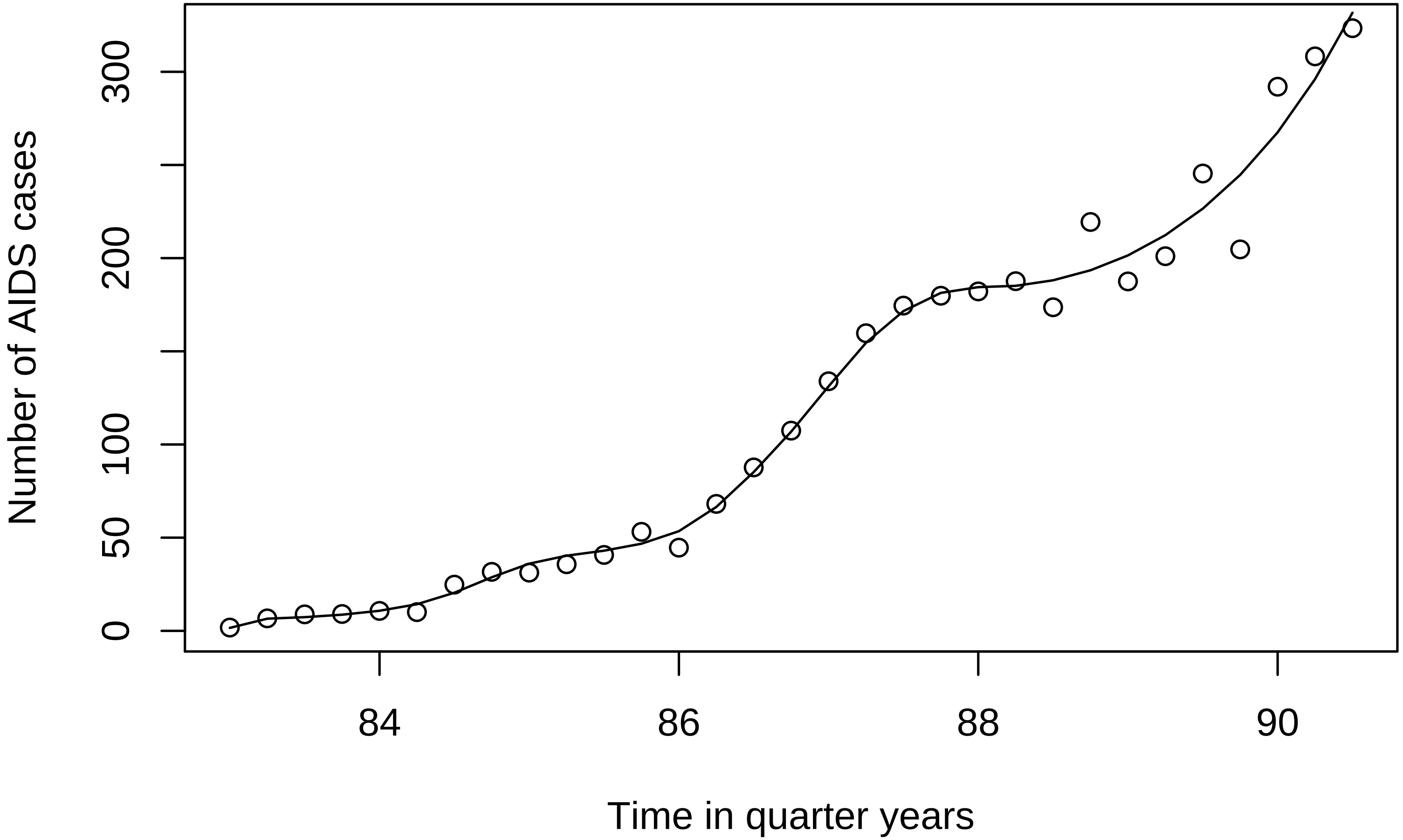

Figure 8: Half Quadratic penalized fit to the deseasonalized AIDS data.