Smoothing and variable selection using P-splines

Irène Gijbels

Department of Mathematics & Leuven Statistics Research Center

Katholieke Universiteit Leuven, Belgium

Joint work with Anestis Antoniadis, Anneleen Verhasselt, Sophie Lambert-Lacroix

Anestis Antoniadis

- XIVth International Biometrics Conference in Namur, Belgium, in July 1988
- first reading:
  "Nonparametric **penalized** maximum likelihood estimation of the intensity of a counting process", *AISM*, 1989
- joint work on
  - **model selection** using **wavelet** decomposition (with G. Grégoire)
  - **change point detection**
  - change point detection in hazard function (with B. MacGibbon)
  - **unfolding sphere size distributions** using **wavelets** (with J. Fan)

- penalized **wavelet monotone regression** (with J. Bigot)
- smoothing non equispaced **heavy noisy data** with **wavelets** (with Jean-Michel Poggi)
- **penalized likelihood regression** for generalized linear models with nonquadratic penalties (with Mila Nikolova)
- **variable selection** in additive models and in varying coefficient models using **P-splines** (with Anneleen Verhasselt and Sophie Lambert-Lacroix)

Seminal contributions to many areas

- wavelets and their applications
- intensity function estimation
- survival analysis and point processes
- inverse problems (in particular Poisson inverse problems)
- constraint estimation
- analysis of functional data
- development of statistical methods for microarray data
- $\cdots$

excellent and very dynamic researcher

extremely broad knowledge

Some characteristics

- many services to the profession
  - associate editor of several international journals
  - serving in scientific evaluation boards for many years
  - ...
- very supporting to young (moderate and old) researchers
  - always helping with scientific advise
  - a remarkable honesty and modesty
  - ...

an admirable personality; an example for many ...

THANK YOU!

Additive models: introduction

Y: response variable

$(X_1, \ldots, X_d)$ vector of d explanatory variables

additive model $$Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \varepsilon \qquad E\left(f_j(X_j)\right) = 0$$

ε random noise term; mean 0 and variance σ^2

f_j unknown univariate functions

often only a few components f_j are different from 0

aim: to *select* and *estimate* the non-zero f_j components

Nonnegative garrote method

Original nonnegative garrote method

proposed by Breiman (1995) in a multiple linear regression model

data $(Y_i, X_{i1}, \ldots, X_{id})$ from

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ij} + \varepsilon_i \qquad i = 1, \ldots, n$$

$\widehat{\beta}_j^{\text{OLS}}$ ordinary least squares estimator for β_j

basic idea: the nng method shrinks the least squares estimators $\widehat{\beta}_j^{\text{OLS}}$

shrinkage done via: $c_j\,\widehat{\beta}_j^{\text{OLS}}$ with $c_j \geq 0$ and a bound on $\sum_{j=1}^{d} c_j$

task : how to find the shrinkage factors c_j?

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

$$\begin{cases} (\widehat{c}_1, \ldots, \widehat{c}_d) = \text{argmin}_{c_1,\ldots,c_d} \dfrac{1}{2} \displaystyle\sum_{i=1}^{n} \Big(Y_i - \widehat{\beta}_0^{\text{OLS}} - \sum_{j=1}^{d} c_j \widehat{\beta}_j^{\text{OLS}} X_{ij} \Big)^2 \\ \text{s.t. } 0 \leq c_j \ (j = 1, \ldots, d), \quad \displaystyle\sum_{j=1}^{d} c_j \leq s \end{cases}$$

for given s, or equivalently

$$\begin{cases} (\widehat{c}_1, \ldots, \widehat{c}_d) = \text{argmin}_{c_1,\ldots,c_d} \left\{ \frac{1}{2} \displaystyle\sum_{i=1}^{n} \Big(Y_i - \widehat{\beta}_0^{\text{OLS}} - \sum_{j=1}^{d} c_j \widehat{\beta}_j^{\text{OLS}} X_{ij} \Big)^2 + \theta \sum_{j=1}^{d} c_j \right\} \\ \text{s.t. } 0 \leq c_j \ (j = 1, \ldots, d) \end{cases}$$

for given $\theta > 0$

$s > 0$ and $\theta > 0$; regularization parameters (see e.g. Xiong (2010)) the nonnegative garrote estimator of the regression coefficient β_j is

$$\widehat{\beta}_j^{\text{NNG}} = \widehat{c}_j \widehat{\beta}_j^{\text{OLS}}$$

special case: orthogonal design, i.e. $\mathbf{X}'\mathbf{X} = \mathbf{I}_n$

$$\widehat{c}_j = \left(1 - \frac{\theta}{(\widehat{\beta}_j^{\text{OLS}})^2}\right)_+ \qquad z_+ = \max(z, 0)$$

the larger θ, the stronger the shrinkage effect

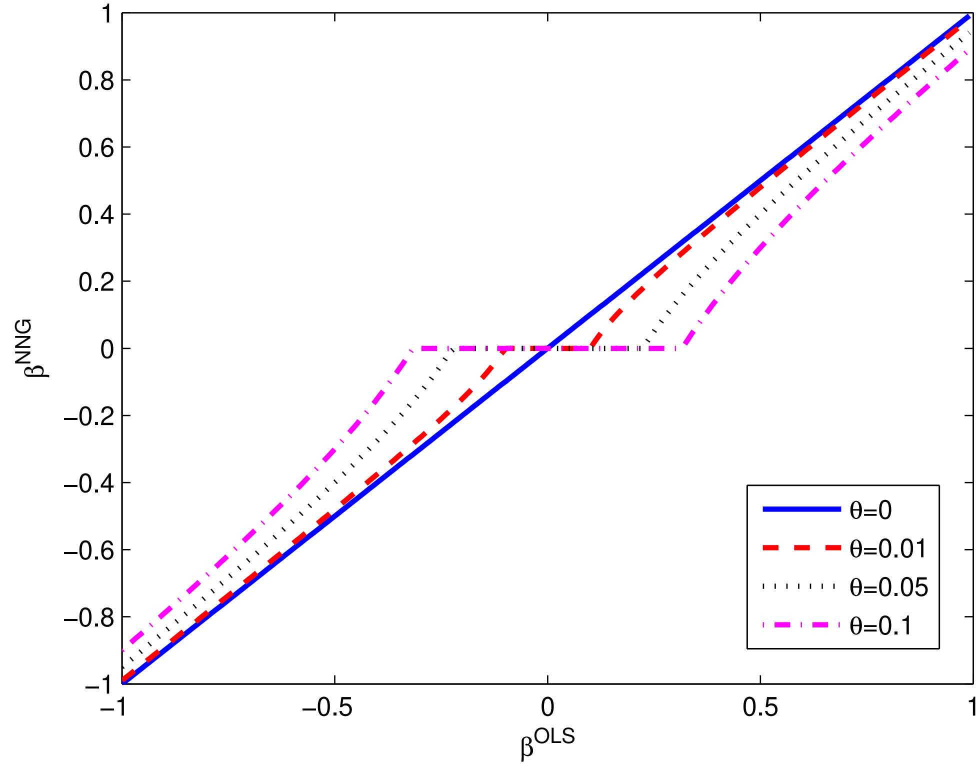


Figure 1: Shrinkage effect of the nonnegative garrote for different θ's

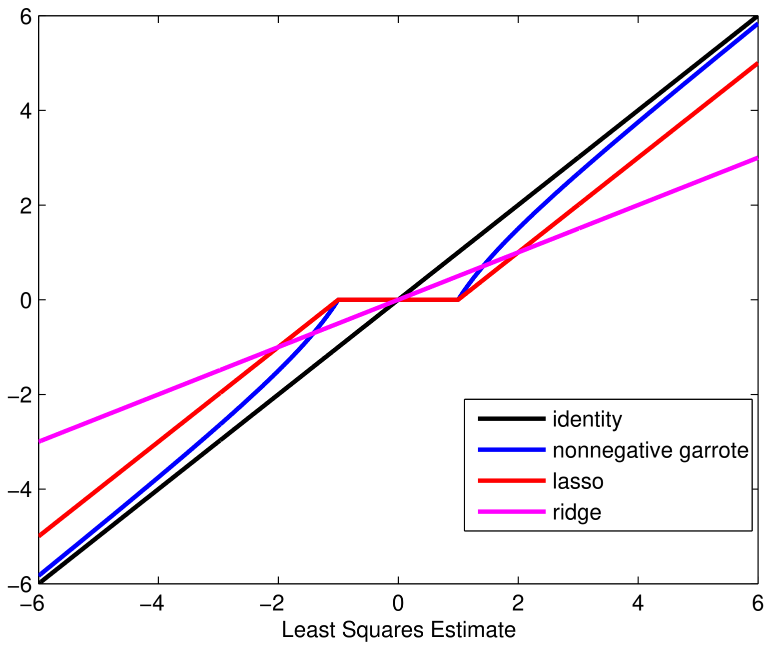
Relation with other estimation methods

LASSO

$$\begin{cases} (\widehat{\beta}_1^{\text{Lasso}}, \ldots, \widehat{\beta}_d^{\text{Lasso}}) = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \beta_j X_{ij} \right)^2 \\ \text{s.t. } \sum_{j=1}^{d} |\beta_j| \leq s \end{cases}$$

Ridge :

$$\begin{cases} (\widehat{\beta}_1^{\text{Ridge}}, \ldots, \widehat{\beta}_d^{\text{Ridge}}) = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \beta_j X_{ij} \right)^2 \\ \text{s.t. } \sum_{j=1}^{d} \beta_j^2 \leq s \end{cases}$$



More relations with (other) thresholding rules

(see, e.g., literature on wavelet methods, work by Anestis Antoniadis)

hard-thresholding rule
$$\delta_\lambda^H\left(\widehat{\beta}_j\right) = \begin{cases} 0 & \text{if} \quad |\widehat{\beta}_j| \le \lambda \\ \widehat{\beta}_j & \text{if} \quad |\widehat{\beta}_j| > \lambda \end{cases}$$

soft-thresholding rule
$$\delta_\lambda^S\left(\widehat{\beta}_j\right) = \begin{cases} 0 & \text{if} \quad |\widehat{\beta}_j| \le \lambda \\ \widehat{\beta}_j - \lambda & \text{if} \quad \widehat{\beta}_j > \lambda \\ \widehat{\beta}_j + \lambda & \text{if} \quad \widehat{\beta}_j < -\lambda \end{cases}$$

- ◇ hard-thresholding (**a discontinous function**): **'keep'** or **'kill'** rule
- ◇ soft-thresholding (**a continuous function**): **'shrink'** or **'kill'** rule

Bruce & Gao (1996) and Marron, Adak, Johnstone, Newmann & Patil (1998), ...,
Gao (2008), ...

another thresholding rule

Antoniadis & Fan (2001) suggested the SCAD (Smoothed Clipped Absolute Deviation) thresholding rule

$$\delta_\lambda^{\text{SCAD}}\left(\widehat{\beta}_j\right) = \begin{cases} \text{sign}\left(\widehat{\beta}_j\right)\max\left(0, |\widehat{\beta}_j|\lambda\right) & \text{if} \quad |\widehat{\beta}_j| \leq 2\lambda \\ \dfrac{(a-1)\widehat{\beta}_j - a\lambda\text{sign}\left(\widehat{\beta}_j\right)}{a-2} & \text{if} \quad 2\lambda < |\widehat{\beta}_j| \leq a\lambda \\ \widehat{\beta}_j & \text{if} \quad |\widehat{\beta}_j| > a\lambda \end{cases}$$

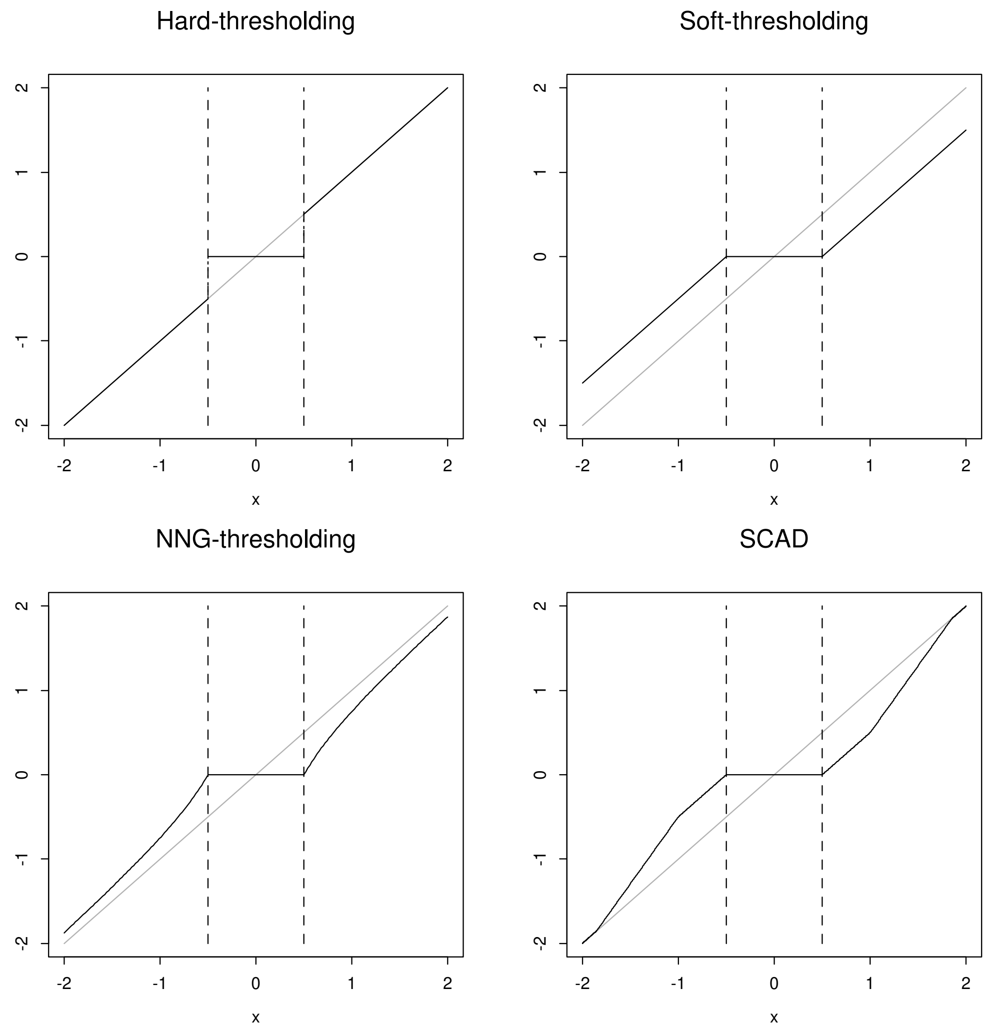
where $a > 2$

this is also a **'shrink'** or **'kill'** rule (**a piecewise linear function**)

this rule does not over-penalize large values of $|\widehat{\beta}_j|$ and hence does not create excessive bias when the regression coefficients are large

Antoniadis & Fan (2001), based on a Bayesian argument, have recommended to use the value $a = 3.7$

thresholding functions δ_λ



Functional nonnegative garrote method

extension of the nng to additive models: see Cantoni, Fleming & Ronchetti (2011) and Yuan (2007)

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

the nonnegative garrote shrinkage factors are then found via

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

the nonnegative garrote estimate of f_j is $\boxed{\widehat{f}_j^{\text{NNG}}(\cdot) = \widehat{c}_j \widehat{f}_j^{\text{init}}(\cdot)}$

Cantoni, Fleming & Ronchetti (2011): use smoothing splines for the initial estimator

using P-splines ...

univariate P-spline estimation

P-splines, introduced by Eilers & Marx (1996), in the univariate nonparametric smoothing context

$$Y_i = f(X_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n$$

P-splines are an extension of regression splines with a penalty on the coefficients of adjacent B-splines

(X_i, Y_i), for $i = 1, \ldots, n$, with $X_i \in [0, 1] \subset I\!R$

regression spline model: approximate $f(x)$ with $\boxed{\sum_{j=1}^{m} \alpha_j B_j(x; q)}$

where $\{B_j(\cdot; q) : j = 1, \ldots, K + q = m\}$ is the q-th **degree** B-spline basis, using normalized B-splines such that $\sum_j B_j(x; q) = 1$, with $K + 1$ equidistant **knot points** $t_0 = 0, t_1 = \frac{1}{K}, \ldots, t_K = 1$ in $[0, 1]$

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$: unknown column vector of regression coefficients penalized least squares estimator $\widehat{\alpha}$ is the minimizer of

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{m} \alpha_j B_j(X_i; q) \Big)^2 + \lambda \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2$$

$\lambda > 0$: smoothing parameter

Δ the differencing operator: $\Delta^k \alpha_j = \sum_{t=0}^{k} (-1)^t \binom{k}{t} \alpha_{j-t}$, with $k \in \mathbb{N}$

examples: $k = 1 : \Delta^1 \alpha_j = \alpha_j - \alpha_{j-1}$

$k = 2 : \Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$

rewriting in matrix-notation:

$$S(\boldsymbol{\alpha}) = (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})'(\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' \mathbf{D}_k' \mathbf{D}_k \boldsymbol{\alpha}$$

the elements B_{ij} of $\mathbf{B}$ ($\in \mathbb{R}^{n \times m}$) are $B_j(X_i; q)$

$\mathbf{D}_k$ ($\in \mathbb{R}^{(m-k) \times m}$) : matrix representation of the kth order differencing operator Δ^k

Additive model: Nonnegative garrote with P-splines initial estimator

additive model $$Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \varepsilon \qquad E\left(f_j(X_j)\right) = 0$$

key ingredients to prove the consistency of the nonnegative garrote with P-splines

- (i) consistency result for (univariate) P-splines (Claeskens, Krivobokova & Opsomer (2009))
- (ii) extension of a univariate smoothing estimator to additive models via backfitting (rely on results from Horowitz, Klemelä & Mammen (2006))
- (iii) on a consistency result for the functional nonnegative garrote (e.g. Yuan (2007))

Consistency of nonnegative garrote with P-splines

notation : $\mathbf{f}_j = (f_j(X_{1j}), \cdots, f_j(X_{nj}))'$

Theorem:

Under some assumptions, and if $\frac{\theta}{n}$ tends to 0 such that $\kappa_n = n^{\frac{-(q+1)}{2q+3}} = o(\frac{\theta}{n})$, then (given $X_{ij} = x_{ij}$)

(1). $P(\widehat{\mathbf{f}}_j^{\text{NNG}} = \mathbf{0}) \to 1$ for any j such that $f_j = 0$

(2). $\sup_j \frac{1}{n}||\mathbf{f}_j - \widehat{\mathbf{f}}_j^{\text{NNG}}||_2^2 = O_P\big((\frac{\theta}{n})^2\big)$

in other words: the nonnegative garrote method with P-splines is

- variable selection consistent (1)
- estimation consistent (2)

Other selection methods

- COSSO (Component Selection and Smoothing Operator, Lin & Zhang (2006))
- ACOSSO (Adaptive Component Selection and Smoothing Operator, Storlie, Bondell, Reich & Zhang (2010))
- APSO (Adaptive P-splines Selection Operator, Antoniadis *et al.* (2011))

Varying coefficient models: introduction

(Fan and Zhang (2000), ...)

$$Y(t) = \beta_0(t) + \sum_{p=1}^{d} X^{(p)}(t)\beta_p(t) + \varepsilon(t) = \mathbf{X}(t)'\boldsymbol{\beta}(t) + \varepsilon(t)$$

$Y(t)$ is the response at time t $(t \in \mathcal{T} = [0, T])$

$\mathbf{X}(t) = (X^{(0)}(t), \ldots, X^{(d)}(t))'$ covariate vector at time t with $X^{(0)}(t) \equiv 1$

$\boldsymbol{\beta}(t) = (\beta_0(t), \ldots, \beta_d(t))'$ the vector of coefficients at time t

$\beta_0(t)$ is the baseline effect

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

longitudinal data, i.e. samples with n independent subjects or individuals each measured repeatedly over a time period

the j-th measurement for subject i of $(t, Y(t), \mathbf{X}(t))$ is denoted by $(t_{ij}, Y_{ij}, \mathbf{X}_{ij})$

$1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant N_i$

N_i is the number of repeated measurements of subject i

t_{ij} is the measurement time, Y_{ij} is the observed response at time t_{ij} and $\mathbf{X}_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(d)})'$

$N = \sum_{i=1}^{n} N_i$ is the total number of observations

P-spline estimation in varying coefficient models

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

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

$$\beta_p(t) = \sum_{l=1}^{m_p} B_{pl}(t; q_p)\alpha_{pl}$$

where $\{B_{pl}(\cdot; q_p) : l = 1, \ldots, K_p + q_p = m_p\}$ is the q_p-th degree B-spline basis with $K_p + 1$ equidistant knots for the p-th component the P-spline estimates of the regression coefficients α_{pl} are obtained by minimizing $S(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_0, \ldots, \boldsymbol{\alpha}'_d)' \in \mathbb{R}^{\text{dim} \times 1}$, where $\boldsymbol{\alpha}_p = (\alpha_{p1}, \ldots, \alpha_{pm_p})'$ and $\text{dim} = \sum_p m_p$:

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(Y_{ij} - \sum_{p=0}^{d} \sum_{l=1}^{m_p} X_{ij}^{(p)} B_{pl}(t_{ij}; q_p) \alpha_{pl} \right)^2 + \sum_{p=0}^{d} \lambda_p \boldsymbol{\alpha}'_p \mathbf{D}'_{k_p} \mathbf{D}_{k_p} \boldsymbol{\alpha}_p$$

k_p is the differencing order for the p-th component

λ_p are the smoothing parameters

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \Big(Y_{ij} - \sum_{p=0}^{d} \sum_{l=1}^{m_p} X_{ij}^{(p)} B_{pl}(t_{ij}; q_p) \alpha_{pl} \Big)^2 + \sum_{p=0}^{d} \lambda_p \boldsymbol{\alpha}_p' \mathbf{D}_{k_p}' \mathbf{D}_{k_p} \boldsymbol{\alpha}_p$$

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

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

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

$$\mathbf{U}_{ij}' = \mathbf{X}_{ij}' \mathbf{B}(t_{ij}) \in \mathbb{R}^{1 \times \text{dim}}$$

$$\mathbf{U}_i = (\mathbf{U}_{i1}', \ldots, \mathbf{U}_{iN_i}')' \in \mathbb{R}^{N_i \times \text{dim}}$$

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

$\mathbf{Q}_\lambda = \mathrm{diag}(\lambda_0 \mathbf{D}_{k_0}' \mathbf{D}_{k_0}, \ldots, \lambda_d \mathbf{D}_{k_d}' \mathbf{D}_{k_d}) \in \mathbb{R}^{\text{dim} \times \text{dim}}$ (a block diagonal matrix with the matrices $\lambda_p \mathbf{D}_{k_p}' \mathbf{D}_{k_p}$ on the diagonal)

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

introducing further matrix notations:

$$||\widetilde{\mathbf{Y}} - \widetilde{\mathbf{U}} \boldsymbol{\alpha}||_2^2 + \boldsymbol{\alpha} \mathbf{Q}_\lambda \boldsymbol{\alpha}$$

$$\begin{aligned}
\mathbf{Y} &= (\mathbf{Y}_1', \ldots \mathbf{Y}_n')' \in \mathbb{R}^{N \times 1} \\
\mathbf{W} &= \text{diag}(\mathbf{W}_i)_{i=1,\ldots,n} \in \mathbb{R}^{N \times N} \\
\widetilde{\mathbf{Y}} &= \mathbf{W}^{1/2} \mathbf{Y} \\
\mathbf{U} &= [\mathbf{U}_0, \ldots, \mathbf{U}_d] \\
\widetilde{\mathbf{U}} &= \mathbf{W}^{1/2} \mathbf{U}
\end{aligned}$$

consistency results are proved for the case that the number of knots $K_p + 1$ (and thus $m_p = K_p + m_p$) grows with n

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

theoretical results

- consistency result
  $\|\widehat{\beta}_p(t) - \beta_p(t)\|_{L_2} = O_P\left(\left(\frac{1}{n^2}\sum_{i=1}^n \frac{1}{N_i}\right)^{q/(2q+1)}\right)$
- asymptotic normality

Nonnegative garrote selection method

the nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_0, \ldots, \widehat{c}_d)'$ are obtained from the optimization problem

$$\begin{cases} \min_{c_0,\ldots,c_d} \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{d} \left(Y_{ij} - \sum_{p=0}^{d} X_{ij}^{(p)} c_p \widehat{\beta}_p^{\text{init}}(t_{ij}) \right)^2 + \theta \sum_{p=0}^{d} c_p \\ \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d) \end{cases}$$

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

$\theta > 0$ is a regularization parameter

we use the P-spline estimator as an initial estimator some more matrix notations:

$$\begin{cases} \min_{\mathbf{c}} ||\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\mathbf{c}||_2^2 + \theta \sum_{p=0}^{d} c_p \\ \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d) \end{cases}$$

where

$$\begin{aligned}
\mathbf{Y} &= (\mathbf{Y}_1', \ldots \mathbf{Y}_n')' \in I\!R^{N\times 1} \\
\mathbf{W} &= \text{diag}(\mathbf{W}_i)_{i=1,\ldots,n} \in I\!R^{N\times N} \\
\widetilde{\mathbf{Y}} &= \mathbf{W}^{1/2}\mathbf{Y} \\
z_i^{(p)} &= (X_{i1}^{(p)}, \ldots, X_{iN_i}^{(p)}) \ \text{diag}(\widehat{\beta}_p^{\text{init}}(t_{ij}))_{j=1,\ldots,N_i} \in I\!R^{1\times N_i} \\
\mathbf{Z}_p &= (z_1^{(p)}, \ldots, z_n^{(p)})' \in I\!R^{N\times 1} \\
\mathbf{Z} &= [\mathbf{Z}_0, \ldots, \mathbf{Z}_d] \\
\widetilde{\mathbf{Z}} &= \mathbf{W}^{1/2}\mathbf{Z} \\
\mathbf{c} &= (c_0, \ldots, c_d)'
\end{aligned}$$

the P-spline estimator

$$\widehat{f}_p(t) = X^{(p)}(t)\widehat{\beta}_p(t)$$

for the p-th component $f_p(t) = X^{(p)}(t)\beta_p(t)$ is consistent

it can be shown that the nonnegative garrote estimator with the P-spline estimator as initial estimator for $\beta_p(t)$

$$\widehat{f}_p^{\mathrm{NNG}}(t) = \widehat{c}_p \widehat{f}_p(t)$$

is
- · estimation consistent
- · variable selection consistent

other selection methods: Adaptive P-spline Selection Operator (APSO) ...

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

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

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

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

varying coefficient model for Y_{ij}
$Y_{ij} = \beta_0(t_{ij}) + \sum_{p=1}^{3} X_i^{(p)} \beta_p(t_{ij}) + \varepsilon_{ij}$

$\beta_0(t)$ is the baseline CD4 percentage, represents the mean CD4 percentage t years after the HIV infection for a nonsmoker with average pre-infection CD4 percentage and average age at infection

Table 1: Aids data. Summary parameters.

method	NS	RSS
ngp	2	110.7756
APSO	2	113.1924

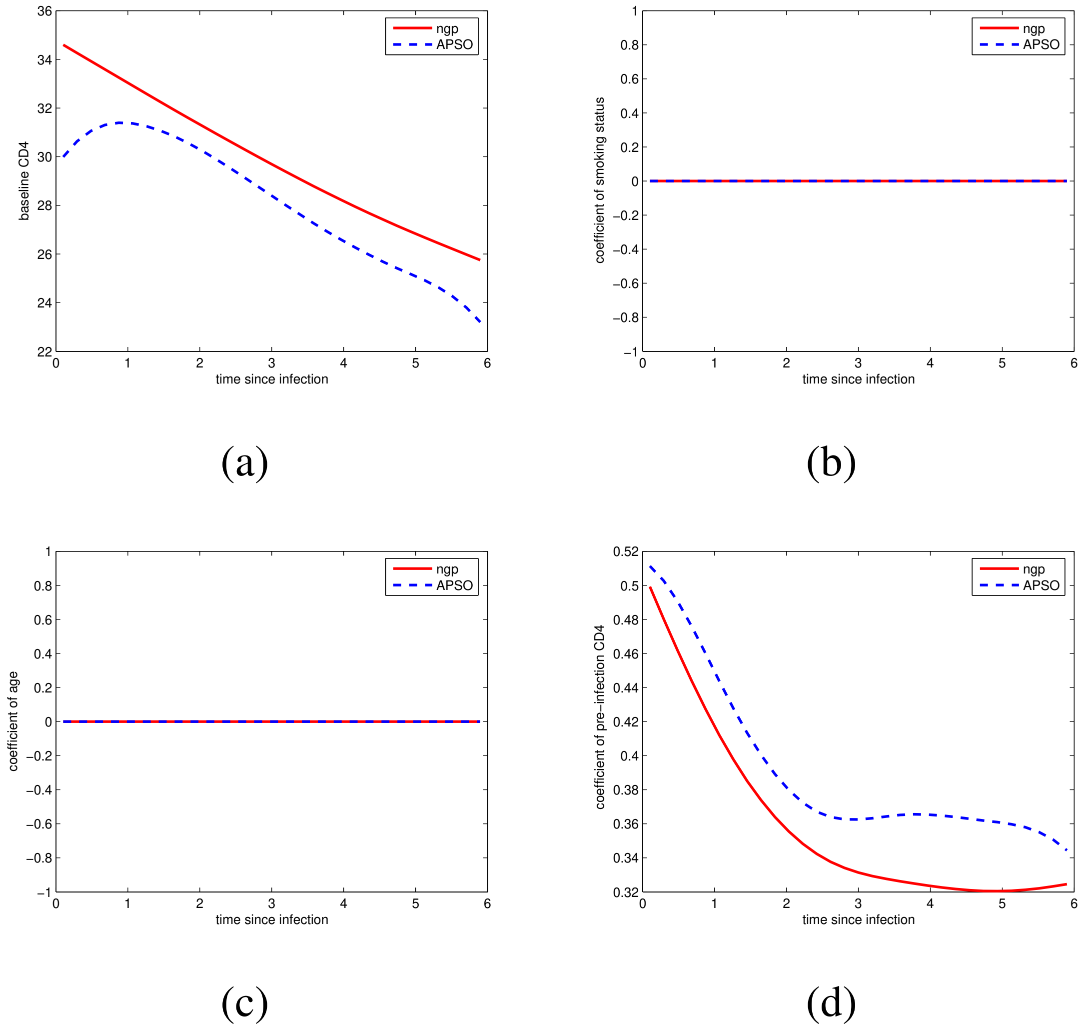


Figure 2: Aids data. Fitted (a) baseline effect; (b) coefficient of smoking status; (c) coefficient of age at HIV infection; (d) coefficient of pre-infection CD4.

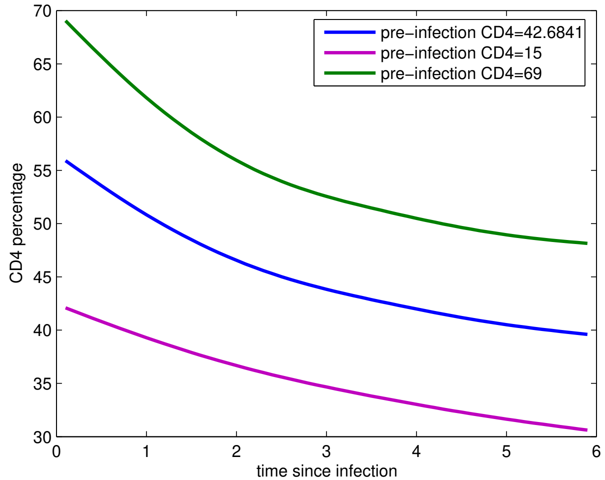


Figure 3: Aids data. Fitted CD4 percentage for 3 pre-infection CD4 percentages.

Grouped regularization methods

$$Y(t) = \beta_0(t) + \sum_{p=1}^{d} X^{(p)}(t)\beta_p(t) + \varepsilon(t) = \mathbf{X}(t)'\boldsymbol{\beta}(t) + \varepsilon(t)$$

approximation in terms of a basis of smooth functions

$$\beta_p(t) \approx \sum_{\ell=1}^{L_p} B_\ell^{(p)}(t)\gamma_{p\ell} \qquad (L_p \text{ large})$$

as before, introducing the appropriate matrix notations

$$\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(Y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{\ell=1}^{L_k} \gamma_{k,\ell} X_i^{(k)}(t_{ij}) B_\ell^{(k)}(t_{ij}) \right)^2 \equiv \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2$$

$\widetilde{\mathbf{Z}}$: dimension $N \times (\sum_{p=0}^{d} L_p)$ $\qquad$ $\boldsymbol{\gamma}$: dimension $(\sum_{p=0}^{d} L_p) \times 1$

grouped Lasso regularization

minimize

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \lambda \sum_{p=0}^{d} w_p \|\boldsymbol{\gamma}_p\|_2 \qquad w_p = \sqrt{L_p}$$

with respect to the vector of parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0', \ldots, \boldsymbol{\gamma}_d')'$

defining $\boxed{p_\lambda(v) = \lambda\, v}$, for $v \geq 0$, this can be written as

$$\text{minimize} \qquad \frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{p=0}^{d} p_\lambda \left(w_p \|\boldsymbol{\gamma}_p\|_2 \right) \qquad w_p = \sqrt{L_p}$$

Liu and Zhang (2008)

grouped SCAD regularization

the SCAD penalty $p_\lambda(v)$, for $v \geq 0$, is

$$p_\lambda(v) = \begin{cases} \lambda v & \text{if} \quad 0 \leq v \leq \lambda \\ -\dfrac{v^2 - 2a\lambda v + \lambda^2}{2(a-1)} & \text{if} \quad \lambda < v < a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{if} \quad v \geq a\lambda \end{cases}$$

grouped SCAD procedure: minimize

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\gamma\|_2^2 + \sum_{p=0}^{d} p_\lambda(\omega_p\|\boldsymbol{\gamma}_p\|_2) \qquad w_p = \sqrt{L_p}$$

with respect to the vector of parameters γ

possible approach

a Taylor expansion of $p_\lambda(v)$ for v around v_0 gives

$$p_\lambda(v) \approx p_\lambda(v_0) + \frac{1}{2}\frac{p'_\lambda(v_0)}{v_0}(v^2 - v_0^2)$$

(see Fan and Li (2001))

this leads to solving a Ridge-regression type of problem
(restricted to $d_n < n$)

first grouped SCAD regularization procedure

minimize

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{p=0}^{d} p_\lambda(\omega_p \|\boldsymbol{\gamma}_p\|_2) \qquad w_p = \sqrt{L_p}$$

with respect to the vector of parameters γ

second grouped SCAD regularization procedure

minimize with respect to γ

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{p=0}^{d} p_\lambda(\|\boldsymbol{\gamma}_p\|_1)$$

algorithm for solving this optimization problem: in Breheny & Huang (2009)

grouped Bridge regularization

penalty function, for $v > 0$,

$$\boxed{p_\lambda(v) = \lambda|v|^q} \qquad \text{with } 0 < q < 1$$

grouped Bridge approach: minimizing the objective function

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{p=0}^{d} p_\lambda(\omega_p\|\boldsymbol{\gamma}_p\|_2)$$

an algorithm for solving this optimization problem: from Breheny & Huang (2009)

Simulation studies

comparison of 5 methods of grouped regularization methods

- `gbridge`: grouped Bridge ($q = 1/2$); local coordinate descent algorithm from Breheny & Huang (2009)
- `gscad`: grouped SCAD; idem
- `gMCP`: grouped MCP; idem
- `glasso 1`: grouped Lasso; idem
- `glasso 2`: grouped Lasso; implemented by Meier, van de Geer & Bühlman (2008)

tuning parameter λ chosen by a BIC-type of criterion:

$$\log\left(\frac{\mathsf{RRS}_\lambda}{\sum_{i=1}^{n} N_i}\right) + \frac{\log(\sum_{i=1}^{n} N_i)}{\sum_{i=1}^{n} N_i}\mathsf{df}_\lambda$$

where RSS_λ is the residual sum of squares

df_λ is the number of nonzero coefficients of $\widehat{\gamma}$

simulation model

(from Huang, Wu & Zhou (2002) and Wang, Li & Huang (2008))

$$Y_i(t_{ij}) = \beta_0(t_{ij}) + \sum_{p=1}^{23} \beta_p(t_{ij}) X_i^{(p)}(t_{ij}) + \varepsilon_i(t_{ij}), \quad i = 1, \ldots, n \quad j = 1, \ldots, \widetilde{N}$$

intercept term and the three true relevant variables:

$$\beta_0(t) = 15 + 20 \sin\left(\frac{\pi t}{60}\right) \qquad \beta_1(t) = 2 - 3\cos\left(\frac{\pi(t-25)}{15}\right)$$

$$\beta_2(t) = 6 - 0.2t \qquad \beta_3(t) = -4 + \frac{(20-t)^3}{2000} \qquad t \in [1, 30]$$

remaining coefficients: $\beta_p(t) = 0$, $p = 4, \ldots, 23$

time points t_{ij} given by $1, 2, \ldots, 30$ ($\widetilde{N} = 30$) and $n = 100$

simulation of three relevant variables $X_i^{(k)}(t)$, $k = 1, \ldots, 3$:

- at any point t, the variable $X_i^{(1)}(t)$ is sampled uniformly from $[t/10, 2 + t/10]$
- conditioning on $X_i^{(1)}(t)$, the variable $X_i^{(2)}(t)$ is centered Gaussian with variance given by $(1 + X_i^{(1)}(t))/(2 + X_i^{(1)}(t))$
- the variable $X_i^{(3)}(t)$ is independent of $X_i^{(1)}$ and $X_i^{(2)}$ and is a Bernoulli random variable with success rate equal to 0.6

the irrelevant variables $X_i^{(k)}$, $k = 4, \ldots, 23$ are paths of centered Gaussian process with covariance function Cov$(X_i^{(k)}(t), X_i^{(k)}(s)) = 4\exp(-|t - s|)$

the irrelevant variables are independent between them and of the others three first variables three levels of noise for the random error: $\sigma = 1,$ 1.25 and 2

corresponds to signal-to-noise ratio (SNR) : 6.39, 5.11 and 3.19

SNR is defined by $\gamma^T \mathbf{Z}^T \mathbf{Z} \boldsymbol{\gamma} / N$

for each simulated data set: use cubic splines with five equidistant internal knots

number of simulations: 500

reported criteria:

- mean value of the tuning parameter λ
- the average number of variables selected
- the average number of truly zero variables that where selected (false positives)
- the average number of truly nonzero variables that where not selected (false negatives)
- the mean and standard deviation of the model error: $(\widehat{\gamma} - \gamma)^T \mathbf{Z}^T \mathbf{Z} (\widehat{\gamma} - \gamma)/N$

Table 2: Selection model ability

	λ	S	FP	FN	ME
$\sigma = 1$					
gbridge	0.006	4.038	0.038	0	0.0121 (0.0033)
gscad	0.199	8.088	4.088	0	0.0324 (0.0142)
gMCP	0.206	7.360	3.360	0	0.0311 (0.0124)
glasso 1	0.114	4.000	0.000	0	0.0141 (0.0066)
glasso 2	0.061	4.102	0.102	0	3.1458 (0.0497)
$\sigma = 2$					
gbridge	0.0146	4.052	0.052	0	0.0482 (0.0137)
gscad	0.294	6.350	2.350	0	0.1222 (0.0316)
gMCP	0.299	6.050	2.050	0	0.1103 (0.0317)
glasso 1	0.145	4.000	0.000	0	0.0587 (0.0407)
glasso 2	0.110	4.586	0.586	0	3.5621 (0.1209)

remarks from this simulation study:

- for all signal-to-noise ratios, `gbridge` and `glasso 1` lead to the best result for the selection ability and for the model error compared to the other methods
- the `gbridge` method is better in model error while `glasso 1` method is better in selection ability
- the `gscad` and `gMCP` procedures are not very good in selection ability: the number of false positives is rather high
- the implementation of group lasso (`glasso 2`) gives relatively correct result in selection model but leads to very bad result in term of model error

typical performance of the estimators of the four first coefficients for a signal-to-noise ratio = 1.25

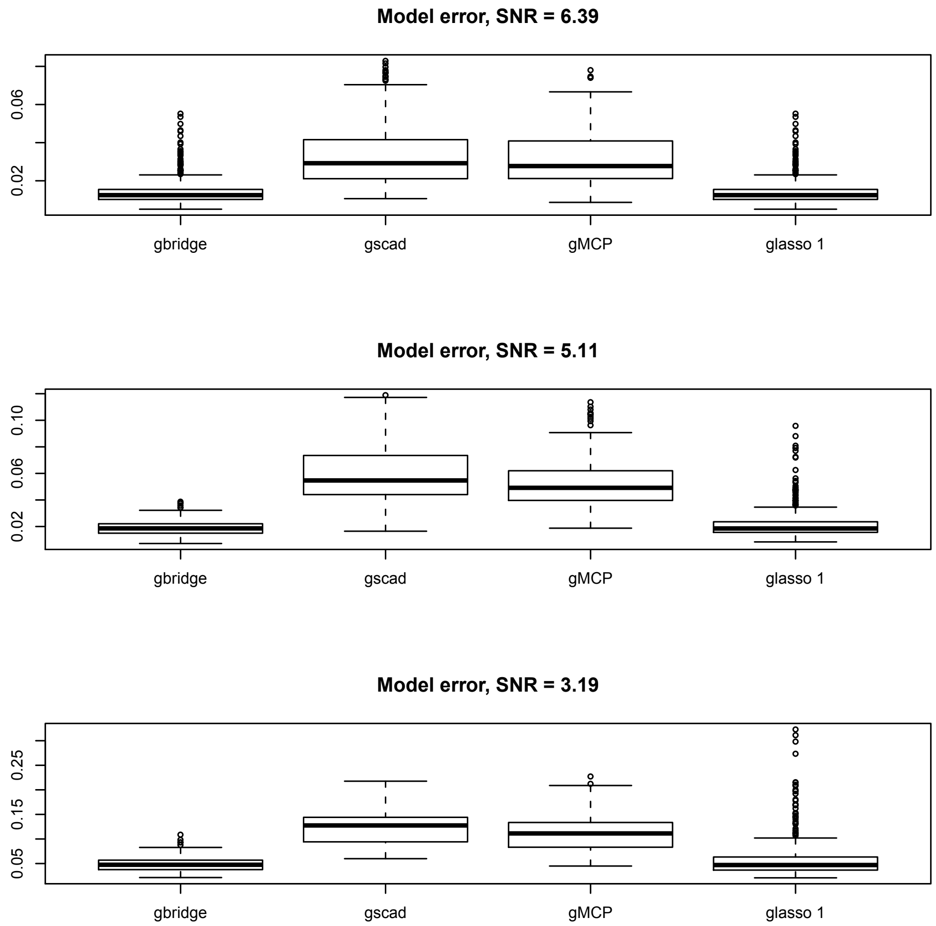


Figure 4:

THANK YOU!