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
- ...

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

\[
Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \varepsilon \quad E(f_j(X_j)) = 0
\]

\( \varepsilon \) random noise term; mean 0 and variance \( \sigma^2 \)

\( f_j \) 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 garroote method

Original nonnegative garroote method

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

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

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

\( \hat{\beta}_j^{\text{OLS}} \)
ordinary least squares estimator for \( \beta_j \)
**basic idea**: the nng method shrinks the least squares estimators $\hat{\beta}_j^{\text{OLS}}$

shrinkage done via: $c_j \hat{\beta}_j^{\text{OLS}}$ 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 $\hat{c}_j$ are found by solving

\[
(\hat{c}_1, \ldots, \hat{c}_d) = \arg\min_{c_1, \ldots, c_d} \frac{1}{2} \sum_{i=1}^{n} \left( Y_i - \hat{\beta}_0^{\text{OLS}} - \sum_{j=1}^{d} c_j \hat{\beta}_j^{\text{OLS}} X_{ij} \right)^2
\]

\[\text{s.t. } 0 \leq c_j (j = 1, \ldots, d), \quad \sum_{j=1}^{d} c_j \leq s\]

for given $s$, or equivalently

\[
(\hat{c}_1, \ldots, \hat{c}_d) = \arg\min_{c_1, \ldots, c_d} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left( Y_i - \hat{\beta}_0^{\text{OLS}} - \sum_{j=1}^{d} c_j \hat{\beta}_j^{\text{OLS}} X_{ij} \right)^2 + \theta \sum_{j=1}^{d} c_j \right\}
\]

\[\text{s.t. } 0 \leq c_j (j = 1, \ldots, d)\]

for given $\theta > 0$

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

$$\hat{\beta}_j^{NNG} = \hat{c}_j \hat{\beta}_j^{OLS}$$

special case: orthogonal design, i.e. $X'X = I_n$

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

the larger $\theta$, the stronger the shrinkage effect
Figure 1: Shrinkage effect of the nonnegative garrote for different $\theta$'s
Relation with other estimation methods

\textbf{LASSO} \begin{align*}
&\left\{ (\hat{\beta}_1^{\text{Lasso}}, \ldots, \hat{\beta}_d^{\text{Lasso}}) = \arg\min_{\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{align*}

\textbf{Ridge:} \begin{align*}
&\left\{ (\hat{\beta}_1^{\text{Ridge}}, \ldots, \hat{\beta}_d^{\text{Ridge}}) = \arg\min_{\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{align*}
More relations with (other) thresholding rules

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

hard-thresholding rule

\[ \delta^H_\lambda (\hat{\beta}_j) = \begin{cases} 
0 & \text{if } |\hat{\beta}_j| \leq \lambda \\
\hat{\beta}_j & \text{if } |\hat{\beta}_j| > \lambda 
\end{cases} \]

soft-thresholding rule

\[ \delta^S_\lambda (\hat{\beta}_j) = \begin{cases} 
0 & \text{if } |\hat{\beta}_j| \leq \lambda \\
\hat{\beta}_j - \lambda & \text{if } \hat{\beta}_j > \lambda \\
\hat{\beta}_j + \lambda & \text{if } \hat{\beta}_j < -\lambda 
\end{cases} \]

diamond Hard-thresholding (a discontinuous function): ‘keep’ or ‘kill’ rule

diamond 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

\[
\delta^\text{SCAD}_\lambda (\hat{\beta}_j) = \begin{cases} 
\text{sign}(\hat{\beta}_j) \max\left(0, |\hat{\beta}_j| \lambda\right) & \text{if } |\hat{\beta}_j| \leq 2\lambda \\
(a - 1)\hat{\beta}_j - a\lambda \text{sign}(\hat{\beta}_j) & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda \\
\hat{\beta}_j & \text{if } |\hat{\beta}_j| > a\lambda
\end{cases}
\]

where \(a > 2\)

this is also a ‘shrink’ or ‘kill’ rule (a piecewise linear function)

this rule does not over-penalize large values of \(|\hat{\beta}_j|\) 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 $\delta_\lambda$

![Graphs of thresholding functions](image.png)
Functional nonnegative garrote method

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

start with an initial estimator \( \hat{f}_{j}^{\text{init}}(X_j) \) of \( f_j(X_j) \)

the nonnegative garrote shrinkage factors are then found via

\[
\begin{align*}
\min_{c_1, \ldots, c_d} & \left\{ \frac{1}{2} \sum_{i=1}^{n} \left( Y_i - \hat{f}_{0}^{\text{init}} - \sum_{j=1}^{d} c_j \hat{f}_{j}^{\text{init}}(X_{ij}) \right)^2 + \theta \sum_{j=1}^{d} c_j \right\} \\
\text{s.t.} & \quad 0 \leq c_j (j = 1, \ldots, d)
\end{align*}
\]

the nonnegative garrote estimate of \( f_j \) is \( \hat{f}_{j}^{\text{NNG}}(\cdot) = \tilde{c}_j \hat{f}_{j}^{\text{init}}(\cdot) \)

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

\[(X_i, Y_i), \text{ for } i = 1, \ldots, n, \text{ with } X_i \in [0, 1] \subset \mathbb{R} \]

regression spline model: approximate \( f(x) \) with

\[ \sum_{j=1}^{m} \alpha_j B_j(x; q) \]

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 \([0, 1]\)

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

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

$\lambda > 0$: smoothing parameter

$\Delta$ 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:

$$S(\alpha) = (Y - B\alpha)'(Y - B\alpha) + \lambda \alpha' D_k' D_k \alpha$$

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

$D_k$ ($\in \mathbb{R}^{(m-k) \times m}$): matrix representation of the $k$th order differencing operator $\Delta^k$
Additive model: Nonnegative garroote with P-splines initial estimator

\[
Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \varepsilon \quad E(f_j(X_j)) = 0
\]

key ingredients to prove the consistency of the nonnegative garroote 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 garroote (e.g. Yuan (2007))
Consistency of nonnegative garrote with P-splines

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

Theorem:

Under some assumptions, and if \( \frac{n}{\theta} \) tends to 0 such that
\[
k_n = n \frac{(q+1)}{2q+3} = o\left( \frac{\theta}{n} \right), \text{ then (given } X_{ij} = x_{ij} \text{)}
\]

(1). \( P(\hat{f}_{j}^{NNG} = 0) \to 1 \) for any \( j \) such that \( f_j = 0 \)

(2). \( \sup_j \frac{1}{n} \| f_j - \hat{f}_{j}^{NNG} \|_2^2 = O_P \left( \left( \frac{\theta}{n} \right)^2 \right) \)

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), ...)

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

\( Y(t) \) 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 \( X^{(0)}(t) \equiv 1 \)

\( \beta(t) = (\beta_0(t), \ldots, \beta_d(t))' \) 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 \((t, Y(t), X(t))\) is denoted by 
\[(t_{ij}, Y_{ij}, X_{ij})\]

\[1 \leq i \leq n, \ 1 \leq j \leq N_i\]

\( N_i \) 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 \( X_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(d)})' \)

\[ N = \sum_{i=1}^{n} N_i \] 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 $\alpha_{pl}$ are obtained by minimizing $S(\alpha)$ with respect to $\alpha = (\alpha'_0, \ldots, \alpha'_d)' \in \mathbb{R}^{\text{dim} \times 1}$, where $\alpha_p = (\alpha_{p1}, \ldots, \alpha_{pm_p})'$ and $\text{dim} = \sum_p m_p$:

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

$k_p$ is the differencing order for the $p$-th component

$\lambda_p$ are the smoothing parameters
\[
S(\alpha) = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left( Y_{ij} - \sum_{p=0}^{d} \sum_{l=1}^{m_p} X_{ij}^{(p)} B_{pl}(t_{ij}; q_p) \alpha_{pl} \right)^2 + \sum_{p=0}^{d} \lambda_p \alpha'_p D'_k D_k \alpha_p \\
= \sum_{i=1}^{n} (Y_i - U_i \alpha)' W_i (Y_i - U_i \alpha) + \alpha Q_\lambda \alpha
\]

\[
Y_i = (Y_{i1}, \ldots, Y_{iN_i})'
\]

\[
B(t) =
\begin{pmatrix}
B_{01}(t; q_0) & \cdots & B_{0m_0}(t; q_0) & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & \cdots & 0 & \ddots & 0 & \cdots & 0 & \cdots & 0 \\
0 & \cdots & 0 & 0 & \cdots & 0 & B_{d1}(t; q_d) & \cdots & B_{dm_d}(t, q_d)
\end{pmatrix}
\]

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

\[
U_i = (U'_{i1}, \ldots, U'_{iN_i})' \in \mathbb{R}^{N_i \times \text{dim}}
\]

\[
W_i = \text{diag} \left( N_i^{-1}, \ldots, N_i^{-1} \right) \in \mathbb{R}^{N_i \times N_i} \quad \text{(a diagonal matrix with } N_i \text{ times} \ N_i^{-1} \text{ on the diagonal)}
\]

\[
Q_\lambda = \text{diag}(\lambda_0 D'_{k_0} D_{k_0}, \ldots, \lambda_d D'_{k_d} D_{k_d}) \in \mathbb{R}^{\text{dim} \times \text{dim}} \quad \text{(a block diagonal matrix with the matrices } \lambda_p D'_{k_p} D_{k_p} \text{ on the diagonal)}
\]
\[ S(\alpha) = \sum_{i=1}^{n} (Y_i - U_i \alpha)' W_i (Y_i - U_i \alpha) + \alpha Q_\lambda \alpha \]

introducing further matrix notations:

\[
\| \tilde{Y} - \tilde{U} \alpha \|_2^2 + \alpha Q_\lambda \alpha
\]

\[
Y = (Y'_1, \ldots, Y'_n)' \in \mathbb{R}^{N \times 1}
\]

\[
W = \text{diag}(W_i)_{i=1,\ldots,n} \in \mathbb{R}^{N \times N}
\]

\[
\tilde{Y} = W^{1/2}Y
\]

\[
U = [U_0, \ldots, U_d]
\]

\[
\tilde{U} = W^{1/2}U
\]
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
  \[
  \| \hat{\beta}_p(t) - \beta_p(t) \|_{L^2} = O_P \left( \left( \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{N_i} \right)^{q/(2q+1)} \right)
  \]

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

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

$\widehat{\beta}_p^{\text{init}}(\cdot)$: 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:

\[
\begin{aligned}
\min_c & \|\tilde{Y} - \tilde{Z}c\|_2^2 + \theta \sum_{p=0}^{d} c_p \\
\text{s.t.} & \quad 0 \leq c_p \ (p = 0, \ldots, d)
\end{aligned}
\]

where

\[
\begin{align*}
Y &= (Y'_1, \ldots, Y'_n)' \in \mathbb{R}^{N \times 1} \\
W &= \text{diag}(W_i)_{i=1,\ldots,n} \in \mathbb{R}^{N \times N} \\
\tilde{Y} &= W^{1/2}Y \\
\tilde{z}^{(p)}_i &= (X^{(p)}_{i1}, \ldots, X^{(p)}_{iN_i}) \text{diag}((\hat{\beta}^{\text{init}}_p(t_{ij}))_{j=1,\ldots,N_i} \in \mathbb{R}^{1 \times N_i} \\
Z_p &= (\tilde{z}^{(p)}_1, \ldots, \tilde{z}^{(p)}_n)' \in \mathbb{R}^{N \times 1} \\
Z &= [Z_0, \ldots, Z_d] \\
\tilde{Z} &= W^{1/2}Z \\
c &= (c_0, \ldots, c_d)'
\end{align*}
\]
the P-spline estimator

\[ \hat{f}_p(t) = X^{(p)}(t) \hat{\beta}_p(t) \]

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 \( \beta_p(t) \)

\[ \hat{f}^\text{NNG}_p(t) = \hat{c}_p \hat{f}_p(t) \]

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>
<thead>
<tr>
<th>method</th>
<th>NS</th>
<th>RSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngp</td>
<td>2</td>
<td>110.7756</td>
</tr>
<tr>
<td>APSO</td>
<td>2</td>
<td>113.1924</td>
</tr>
</tbody>
</table>
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

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

approximation in terms of a basis of smooth functions

\[ \beta_p(t) \approx \sum_{\ell=1}^{L_p} B^{(p)}(t)\gamma_{p\ell} \quad (L_p \text{ large}) \]

as before, introducing the appropriate matrix notations

\[
\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left( Y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{\ell=1}^{L_k} \gamma_{k,\ell} X^{(k)}_i(t_{ij}) B^{(k)}(t_{ij}) \right)^2 \equiv \| \tilde{Y} - \tilde{Z}\gamma \|_2^2
\]

\( \tilde{Z} : \text{dimension } N \times (\sum_{p=0}^{d} L_p) \quad \gamma : \text{dimension } (\sum_{p=0}^{d} L_p) \times 1 \)
grouped Lasso regularization

minimize

$$\frac{1}{2n} \| \tilde{Y} - \tilde{Z} \gamma \|_2^2 + \lambda \sum_{p=0}^{d} w_p \| \gamma_p \|_2 \quad w_p = \sqrt{L_p}$$

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

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

minimize

$$\frac{1}{2n} \| \tilde{Y} - \tilde{Z} \gamma \|_2^2 + \sum_{p=0}^{d} p_{\lambda}(w_p \| \gamma_p \|_2) \quad w_p = \sqrt{L_p}$$

Liu and Zhang (2008)
grouped SCAD regularization

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

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

grouped SCAD procedure: minimize

$$\frac{1}{2n} \| \tilde{Y} - \tilde{Z}\gamma \|_2^2 + \sum_{p=0}^{d} p_\lambda(\omega_p \| \gamma_p \|_2) \quad w_p = \sqrt{L_p}$$

with respect to the vector of parameters $\gamma$
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

\[
\frac{1}{2n} \| \tilde{Y} - \tilde{Z} \gamma \|_2^2 + \sum_{p=0}^{d} p \lambda (\omega_p \| \gamma_p \|_2) \quad w_p = \sqrt{L_p}
\]

with respect to the vector of parameters \( \gamma \)

second grouped SCAD regularization procedure

minimize with respect to \( \gamma \)

\[
\frac{1}{2n} \| \tilde{Y} - \tilde{Z} \gamma \|_2^2 + \sum_{p=0}^{d} p \lambda (\| \gamma_p \|_1)
\]

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

penalty function, for \( v > 0 \),

\[
p_{\lambda}(v) = \lambda|v|^q
\]

with \( 0 < q < 1 \)

grouped Bridge approach: minimizing the objective function

\[
\frac{1}{2n} \|\tilde{Y} - \tilde{Z}\gamma\|_2^2 + \sum_{p=0}^{d} p_{\lambda}(\omega_p \|\gamma_p\|_2)
\]

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

comparison of 5 methods of grouped regularization methods

- \textit{gbridge}: grouped Bridge \((q = 1/2)\); local coordinate descent algorithm from Breheny & Huang (2009)

- \textit{gscad}: grouped SCAD; idem

- \textit{gMCP}: grouped MCP; idem

- \textit{glasso 1}: grouped Lasso; idem

- \textit{glasso 2}: grouped Lasso; implemented by Meier, van de Geer & Bühlman (2008)
tuning parameter $\lambda$ chosen by a BIC-type of criterion:

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

where $\text{RSS}_\lambda$ is the residual sum of squares

$\text{df}_\lambda$ is the number of nonzero coefficients of $\hat{\gamma}$
simulation model

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

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

intercept term and the three true relevant variables:

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

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

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

time points \( t_{ij} \) given by 1, 2, \ldots, 30 (\( \tilde{N} = 30 \)) 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

$\text{Cov}(X_i^{(k)}(t), X_i^{(k)}(s)) = 4 \exp(-|t - s|)$

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 $\gamma^T Z^T Z \gamma / N$

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 $\lambda$
- the average number of variables selected
- the average number of truly zero variables that were selected (false positives)
- the average number of truly nonzero variables that were not selected (false negatives)
- the mean and standard deviation of the model error:
  $$\frac{(\hat{\gamma} - \gamma)^T Z^T Z (\hat{\gamma} - \gamma)}{N}$$
Table 2: Selection model ability

<table>
<thead>
<tr>
<th></th>
<th>λ</th>
<th>S</th>
<th>FP</th>
<th>FN</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>σ = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gbridge</td>
<td>0.006</td>
<td>4.038</td>
<td>0.038</td>
<td>0</td>
<td>0.0121 (0.0033)</td>
</tr>
<tr>
<td>gscad</td>
<td>0.199</td>
<td>8.088</td>
<td>4.088</td>
<td>0</td>
<td>0.0324 (0.0142)</td>
</tr>
<tr>
<td>gMCP</td>
<td>0.206</td>
<td>7.360</td>
<td>3.360</td>
<td>0</td>
<td>0.0311 (0.0124)</td>
</tr>
<tr>
<td>glasso 1</td>
<td>0.114</td>
<td>4.000</td>
<td>0.000</td>
<td>0</td>
<td>0.0141 (0.0066)</td>
</tr>
<tr>
<td>glasso 2</td>
<td>0.061</td>
<td>4.102</td>
<td>0.102</td>
<td>0</td>
<td>3.1458 (0.0497)</td>
</tr>
</tbody>
</table>

| σ = 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!