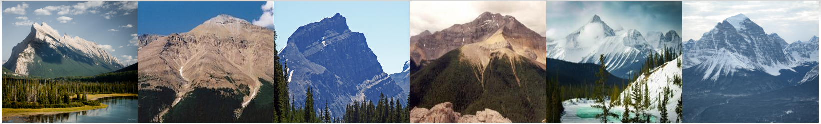
Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

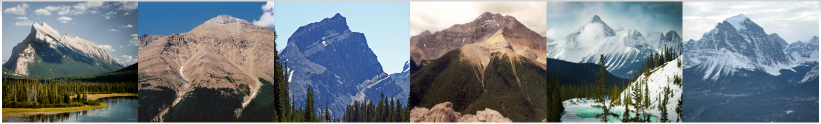
Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada



Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada

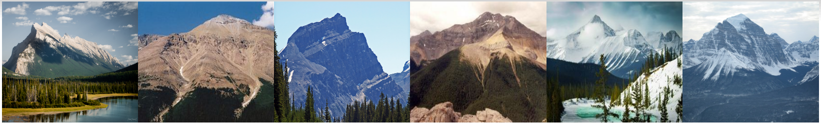


Rundle

Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada

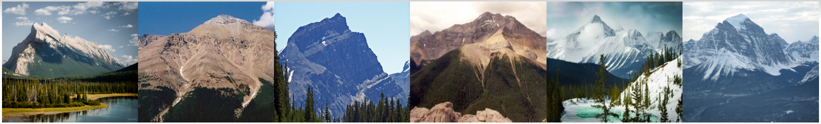


Rundle **O**bservation Pk.

Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada

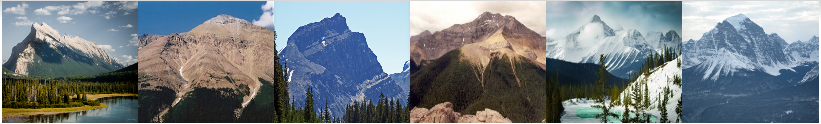


Rundle **O**bservation Pk. **B**ow Pk.

Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada

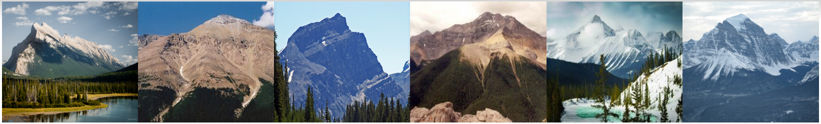


Rundle **O**bservation Pk. **B**ow Pk. **U**topia Mt.

Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada



Rundle **O**bservation Pk. **B**ow Pk. **U**topia Mt. **S**arbach Mt.

Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada

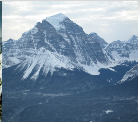

Rundle


Observation Pk.


Bow Pk.


Utopia Mt.


Sarbach Mt.


Temple

Change-point Estimation and Inference in Nonparametric Regression Using Different Regularization Concepts

Jetřichovice, January 24th, 2014

Matúš Maciak
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB, Canada

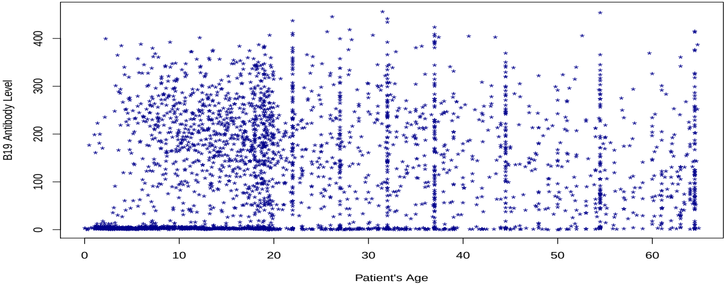
Joint work with Ivan Mizera

Mt. Temple (3.543 meters) July 26, 2013

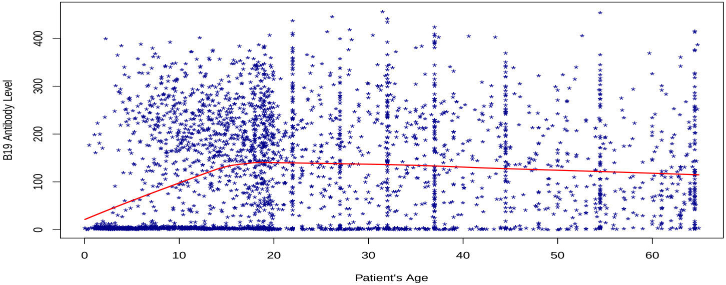
Motivation: Parvovirus B19 Data

- University of Hasselt, Belgium (2008)

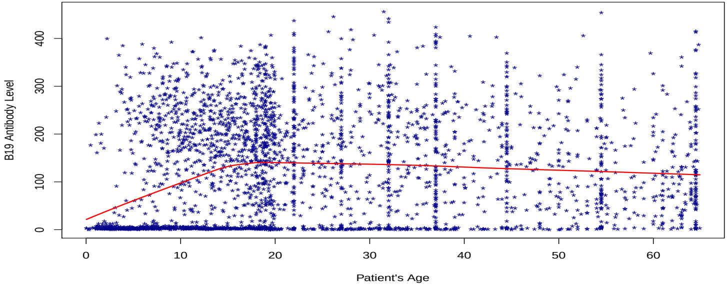
Parvovirus B19 Data



Parvovirus B19 Data - LOWESS



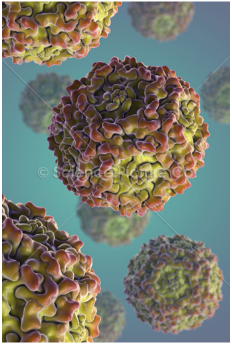
Parvovirus B19 Data - LOWESS



- data analyzed by many authors using various modeling approaches; *(Hens et al. (2010), Maciak (2008), Nardone et al.(2007), etc.)*
- mostly, authors expect some change-points to be present; *(different theoretical and practical limitations)*

Parvovirus B19 Data

- **B19 virus:** mostly known for causing a disease in a pediatric population;
- **Transmission:** respiratory droplets, mostly children at the age of 6 to 10;
- **Infectivity:** individuals after infection generally assumed to be immune;
- **Epidemiology:** increase in the number of cases is seen every three to four years;
- **Data:** over 3000 patients collected in Belgium (November 2001 – March 2003);



Change-points in Regression

- **One-sided estimates** ⇒ **segmented estimation;**
  Antoch et al. (2006); Csörgo and Horváth (1997);
- **Jump detection algorithms** ⇒ **segmented estimation;**
  Horváth and Kokoszka (2002); Qui and Yandell (1998);
- **Permutation tests** ⇒ **segmented estimation;**
  Kim at al. (2009, 2000);
- **Bayesian approach** ⇒ **segmented estimation;**
  Martinez-Beneito et al. (2011); Carlin et al. (1992);

Change-points in Regression

- **One-sided estimates ⇒ segmented estimation;**
  Antoch et al. (2006); Csörgo and Horváth (1997);
- **Jump detection algorithms ⇒ segmented estimation;**
  Horváth and Kokoszka (2002); Qui and Yandell (1998);
- **Permutation tests ⇒ segmented estimation;**
  Kim at al. (2009, 2000);
- **Bayesian approach ⇒ segmented estimation;**
  Martinez-Beneito et al. (2011); Carlin et al. (1992);

- **Total Variation Penalty ⇒ automatic selection using sparsity;**
  Harchaoui and Lévy-Leduc (2010);

The Underlying Model

- random sample $\{(X_i, Y_i);\ i = 1, \ldots, n \in \mathbb{N}\}$, true population $F_{(X,Y)}$;
- the alertdependence structure of Y given X is assumed to take a form

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

- where function m can be additively decomposed as:

$$m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x),$$

The Underlying Model

- random sample $\{(X_i, Y_i);\ i = 1, \dots, n \in \mathbb{N}\}$, true population $F_{(X,Y)}$;
- the alertdependence structure of Y given X is assumed to take a form

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \dots, n,$$

- where function m can be additively decomposed as:

$$m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x),$$

- $\hookrightarrow$ different smoothing assumptions posed on $m_0, s_0, \dots, s_{p-1}$;
  (smooth function m_0 with some background shock processes $s_0, \dots, s_{p-1}$)

The Underlying Model

- random sample $\{(X_i, Y_i);\ i = 1, \ldots, n \in \mathbb{N}\}$, true population $F_{(X,Y)}$;
- the alertdependence structure of Y given X is assumed to take a form

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

- where function m can be additively decomposed as:

$$m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x),$$

- $\hookrightarrow$ different smoothing assumptions posed on $m_0, s_0, \ldots, s_{p-1}$;
  (smooth function m_0 with some background shock processes $s_0, \ldots, s_{p-1}$)
- $\hookrightarrow$ for identifiability reasons we also assume that

$$\sum_{i=1}^{n} s_j^{(\ell)}(X_i) = 0, \qquad \forall j = 0, \ldots, p-1, \text{ and } \ell = 0, \ldots, j,$$


TIME
AIDS
Debates and
Protests
The
New
Mr.
Dollar
Fed Nominee
Alan Greenspan

Model Estimation Using Splines

- available data: $\{(X_i, Y_i);\ i = 1, \ldots, n\}$
- function to estimate: $m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x)$

Model Estimation Using Splines

- available data: $\{(X_i, Y_i);\ i = 1, \ldots, n\}$
- function to estimate: $m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x)$

- **Smoothing Splines approach (with change-points):**
  - X_i's observations $\rightarrow$ knots $\{\xi_i;\ i = 1, \ldots, n\}$ $\Rightarrow$ basis functions $\psi_i(x)$;
    $\hookrightarrow$ basis coefficients $\beta_S \in \mathbb{R}^K$, where $m_0(x) = \sum_{i=1}^{K} \beta_S^{(i)} \psi_i(x)$;
  - jump function $s_0(x)$ $\rightarrow$ grid of (hypothetical) jump-locations $\xi_{01}, \ldots, \xi_{0k_0}$;
    $\hookrightarrow$ jump generating basis: zero order truncated basis $\psi_{0j}(x) = (x - \xi_{0,j})_+^0$;
  - ... ...
  - $(p-1)$-order jump function $s_{p-1}(x)$ $\rightarrow$ grid points $\xi_{(p-1)1}, \ldots, \xi_{(p-1)k_{p-1}}$;
    $\hookrightarrow$ $(p-1)$-order jump generating basis: $\psi_{(p-1)j}(x) = (x - \xi_{(p-1),j})_+^{p-1}$;

Model Estimation Using Splines

- available data: $\{(X_i, Y_i);\ i = 1, \ldots, n\}$
- function to estimate: $m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x)$

- **Smoothing Splines approach (with change-points):**
  - X_i's observations $\rightarrow$ knots $\{\xi_i;\ i = 1, \ldots, n\}$ $\Rightarrow$ basis functions $\psi_i(x)$;
    $\hookrightarrow$ basis coefficients $\boldsymbol{\beta}_S \in \mathbb{R}^K$, where $m_0(x) = \sum_{i=1}^{K} \beta_S^{(i)} \psi_i(x)$;
  - jump function $s_0(x)$ $\rightarrow$ grid of (hypothetical) jump-locations $\xi_{01}, \ldots, \xi_{0k_0}$;
    $\hookrightarrow$ jump generating basis: zero order truncated basis $\psi_{0j}(x) = (x - \xi_{0,j})_+^0$;
  - ... ...
  - $(p-1)$-order jump function $s_{p-1}(x)$ $\rightarrow$ grid points $\xi_{(p-1)1}, \ldots, \xi_{(p-1)k_{p-1}}$;
    $\hookrightarrow$ $(p-1)$-order jump generating basis: $\psi_{(p-1)j}(x) = (x - \xi_{(p-1),j})_+^{p-1}$;

- ideally, we have $k_0 = \cdots = k_{p-1} \equiv k$ and $\xi_{0,j} = \cdots = \xi_{(p-1)j}$ for all $j = 1, \ldots, k$;
- Smoothing spline coefficients $\boldsymbol{\beta}_S$ with a corresponding design matrix $\mathbb{X}_S$ and jump generating (sparse) coefficients $\boldsymbol{\beta}_J$ with a corresponding design matrix $\mathbb{X}_J$;

Minimization formulation

- finite dimensional minimization problem

$$\underset{\beta_S,\, \beta_J}{\textit{Minimize}} \quad \left\| \boldsymbol{Y} - (\mathbb{X}_S \mathbb{X}_J) \begin{pmatrix} \beta_S \\ \beta_J \end{pmatrix} \right\|_2^2 + \lambda_1 \left\| \mathbb{W} \begin{pmatrix} \beta_S \\ \beta_J \end{pmatrix} \right\|_2^2 + \lambda_2 \|\beta_J\|_1$$

- for some $\lambda_1, \lambda_2 > 0$ and $\mathbb{W} = \mathbb{V}^\top \mathbb{V}$, where $\mathbb{V} = (V_{\ell_1 \ell_2})_{\ell_1, \ell_2}$, such that

$$V_{\ell_1 \ell_2} = \int \psi''_{\ell_1}(x) \psi''_{\ell_2}(x) \mathrm{d}x$$

Minimization formulation via LASSO

- for any given $\lambda_1 > 0$ one can apply simple algebra to express the original minimization as

$$\underset{\beta_S,\,\beta_J}{\textit{Minimize}} \quad \left\| \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} \mathbb{X}_S & \mathbb{X}_J \\ \sqrt{\lambda_1}\mathbb{W}_1 & \sqrt{\lambda_1}\mathbb{W}_2 \end{pmatrix} \begin{pmatrix} \beta_S \\ \beta_J \end{pmatrix} \right\|_2^2 + \lambda_2 \|\beta_J\|_1$$

where $(\mathbb{W}_1, \mathbb{W}_2) = \mathbb{W}$.

Minimization formulation via LASSO

- for any given $\lambda_1 > 0$ one can apply simple algebra to express the original minimization as

$$\underset{\beta_S,\,\beta_J}{\textit{Minimize}} \quad \left\| \begin{pmatrix} \boldsymbol{Y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbb{X}_S & \mathbb{X}_J \\ \sqrt{\lambda_1}\mathbb{W}_1 & \sqrt{\lambda_1}\mathbb{W}_2 \end{pmatrix} \begin{pmatrix} \beta_S \\ \beta_J \end{pmatrix} \right\|_2^2 + \lambda_2 \|\beta_J\|_1$$

where $(\mathbb{W}_1, \mathbb{W}_2) = \mathbb{W}$.

- defining

$$\mathbb{H} = \begin{pmatrix} \mathbb{X}_S \\ \sqrt{\lambda_1}\mathbb{W}_1 \end{pmatrix} \left[\begin{pmatrix} \mathbb{X}_S \\ \sqrt{\lambda_1}\mathbb{W}_1 \end{pmatrix}^\top \begin{pmatrix} \mathbb{X}_S \\ \sqrt{\lambda_1}\mathbb{W}_1 \end{pmatrix} \right] \begin{pmatrix} \mathbb{X}_S \\ \sqrt{\lambda_1}\mathbb{W}_1 \end{pmatrix}^\top \text{ and}$$

$\mathbb{M} = (\mathbb{I} - \mathbb{H})$, we can express the solution $\mathbb{X}_S\widehat{\beta}_S + \mathbb{X}_J\widehat{\beta}_J$ of the original problem as $\mathbb{H} \begin{pmatrix} \boldsymbol{Y} \\ \mathbf{0} \end{pmatrix} + (\mathbb{I} - \mathbb{H}) \begin{pmatrix} \mathbb{X}_J \\ \sqrt{\lambda_1}\mathbb{W}_2 \end{pmatrix} \widehat{\beta}_J$, where $\widehat{\beta}_J$ solves

$$\underset{\beta_J}{\textit{Minimize}} \quad \left\| \mathbb{M} \begin{pmatrix} \boldsymbol{Y} \\ \mathbf{0} \end{pmatrix} - \mathbb{M} \begin{pmatrix} \mathbb{X}_J \\ \sqrt{\lambda_1}\mathbb{W}_2 \end{pmatrix} \beta_J \right\|_2^2 + \lambda_2 \|\beta_J\|_1$$

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**

- **Simultaneous Change-points**

- **Hierarchical Change-points**

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**
  - functions $s_0, \dots, s_{p-1}$ are not (mutually) related;
  - for every $s_j \Rightarrow$ a separate sequence of change-point locations $\xi_{j1}, \dots, \xi_{jk_j}$;

- **Simultaneous Change-points**

- **Hierarchical Change-points**

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**
  - multiple L_1 penalties - one for each level $(0, 1, \ldots, p-1)$;
  - penalty form: $\lambda_1 \|\beta_J^{(0)}\| + \cdots + \lambda_{p-1} \|\beta_J^{(p-1)}\|$;

- **Simultaneous Change-points**

- **Hierarchical Change-points**

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**
  - multiple L_1 penalties - one for each level $(0, 1, \ldots, p-1)$;
  - penalty form: $\lambda_1 \|\beta_J^{(0)}\| + \cdots + \lambda_{p-1}\|\beta_J^{(p-1)}\|$;

- **Simultaneous Change-points**
  - functions $s_0, \ldots, s_{p-1}$ are all connected through the change-point locations;
  - one sequence of locations $\xi_1, \ldots, \xi_k \Rightarrow$ in each ξ_ℓ every s_j has a "jump";

- **Hierarchical Change-points**

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**
  - multiple L_1 penalties - one for each level $(0, 1, \ldots, p-1)$;
  - penalty form: $\lambda_1 \|\beta_J^{(0)}\| + \cdots + \lambda_{p-1} \|\beta_J^{(p-1)}\|$;

- **Simultaneous Change-points**
  - Group LASSO penalty, where each group is defined by the location ξ_ℓ;
  - penalty form: $\lambda \sum_\ell \sqrt{\beta_{0\ell}^2 + \cdots + \beta_{(p-1)\ell}^2}$;

- **Hierarchical Change-points**

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**
  - multiple L_1 penalties - one for each level $(0, 1, \ldots, p-1)$;
  - penalty form: $\lambda_1 \|\beta_J^{(0)}\| + \cdots + \lambda_{p-1} \|\beta_J^{(p-1)}\|$;

- **Simultaneous Change-points**
  - Group LASSO penalty, where each group is defined by the location ξ_ℓ;
  - penalty form: $\lambda \sum_\ell \sqrt{\beta_{0\ell}^2 + \cdots + \beta_{(p-1)\ell}^2}$;

- **Hierarchical Change-points**
  - lower to higher order discontinuity is considered (change-point hierarchy);
  - if there is a jump in s_j, for some $j = 0, \ldots, p-1 \Rightarrow$ jump in all s_ℓ, for $\ell > j$;

Change-point Structure/Hierarchy

- various penalization concepts are possible for in real situations;
- different implementation and interpretations of change-point occurrences;

- **Mutually Independent Change-points**
  - multiple L_1 penalties - one for each level $(0, 1, \ldots, p-1)$;
  - penalty form: $\lambda_1 \|\beta_J^{(0)}\| + \cdots + \lambda_{p-1} \|\beta_J^{(p-1)}\|$;
- **Simultaneous Change-points**
  - Group LASSO penalty, where each group is defined by the location ξ_ℓ;
  - penalty form: $\lambda \sum_\ell \sqrt{\beta_{0\ell}^2 + \cdots + \beta_{(p-1)\ell}^2}$;
- **Hierarchical Change-points**
  - Overlap Group LASSO, where each group is defined by the location ξ_ℓ;
  - penalty form: $\lambda \sum_\ell \inf_g \mathcal{G}(\beta_\ell)$;

Group LASSO vs. Overlap LASSO

$$\mathcal{G}(\boldsymbol{\beta}_l) = \sqrt{\beta_{0l(a)}^2 + \beta_{1l(a)}^2 + \beta_{2l(a)}^2} + \sqrt{\beta_{0l(b)}^2 + \beta_{1l(b)}^2 + \beta_{2l(b)}^2} + \sqrt{\beta_{0l(c)}^2 + \beta_{1l(c)}^2 + \beta_{2l(c)}^2},$$

Group LASSO vs. Overlap LASSO

$$\mathcal{G}(\boldsymbol{\beta}_l) = \sqrt{\beta^2_{0l(a)} + \beta^2_{1l(a)} + \beta^2_{2l(a)}} + \sqrt{\beta^2_{0l(b)} + \beta^2_{1l(b)} + \beta^2_{2l(b)}} + \sqrt{\beta^2_{0l(c)} + \beta^2_{1l(c)} + \beta^2_{2l(c)}},$$

such, that

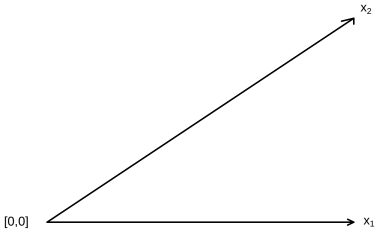
$$\begin{aligned} \beta_{0l} &= \beta_{0l(a)}, && \text{and} && \beta_{0l(b)} = \beta_{0l(c)} = 0, \\ \beta_{1l} &= \beta_{1l(a)} + \beta_{1l(b)}, && \text{and} && \beta_{1l(c)} = 0, \\ \beta_{2l} &= \beta_{2l(a)} + \beta_{2l(b)} + \beta_{2l(c)}. \end{aligned}$$

LARS Algorithm - geometry

- LARS - Least Angle Regression - Efron et al.(2004)
- straightforward modification to accommodate LASSO approach;

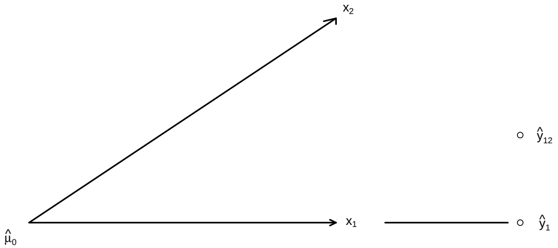
LARS Algorithm - geometry

- LARS - Least Angle Regression - Efron et al.(2004)
- straightforward modification to accommodate LASSO approach;



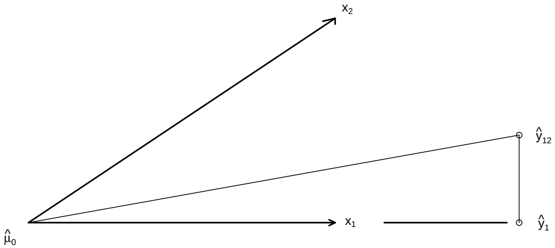
LARS Algorithm - geometry

- LARS - Least Angle Regression - Efron et al.(2004)
- straightforward modification to accommodate LASSO approach;



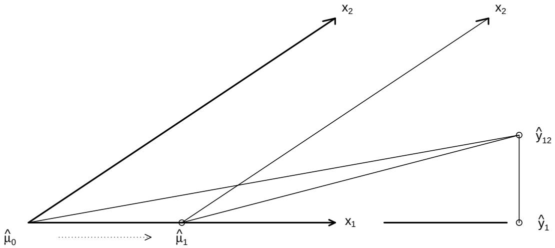
LARS Algorithm - geometry

- LARS - Least Angle Regression - Efron et al.(2004)
- straightforward modification to accommodate LASSO approach;



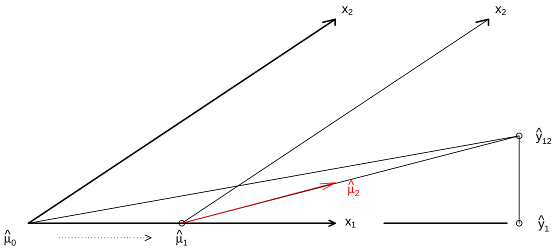
LARS Algorithm - geometry

- LARS - Least Angle Regression - Efron et al.(2004)
- straightforward modification to accommodate LASSO approach;

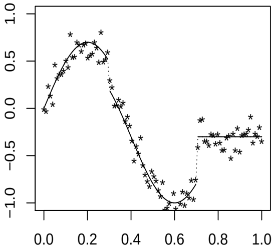


LARS Algorithm - geometry

- LARS - Least Angle Regression - Efron et al.(2004)
- straightforward modification to accommodate LASSO approach;



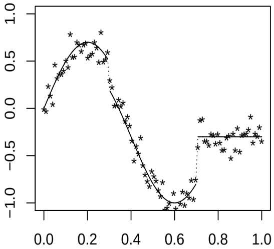
LARS Algorithm - example



- Data: $X_i \sim Unif[0,1]$, for $i = 1, \ldots, 100$;
  $Y_i = m_0(X_i) + \sum_{j=0}^{2} s_j(X_i) + \varepsilon_i$;
- Error: $\varepsilon \sim N(0, 1/400)$;
- Background functions:

$$s_0(x) = 0.49\mathbb{I}(x \geq 0.7) - 0.3\mathbb{I}(x \geq 0.3)$$
$$s_1(x) = -3.9x\mathbb{I}(x \geq 0.7)$$
$$s_2(x) = -30x^2\mathbb{I}(x \geq 0.7)$$

LARS Algorithm - example

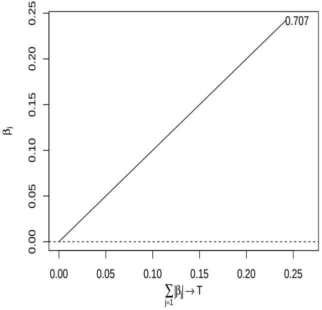


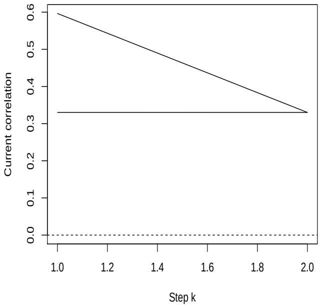
- Data: $X_i \sim Unif[0,1]$, for $i = 1, \dots, 100$;
  $Y_i = m_0(X_i) + \sum_{j=0}^{2} s_j(X_i) + \varepsilon_i$;
- Error: $\varepsilon \sim N(0, 1/400)$;
- Background functions:

$$s_0(x) = 0.49\mathbb{I}(x \geq 0.7) - 0.3\mathbb{I}(x \geq 0.3)$$
$$s_1(x) = -3.9x\mathbb{I}(x \geq 0.7)$$
$$s_2(x) = -30x^2\mathbb{I}(x \geq 0.7)$$

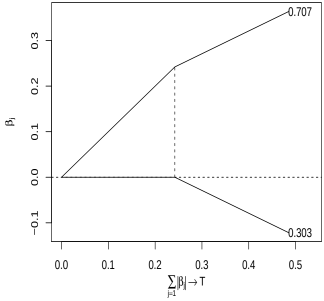
- Mutually Independent Change-points, for $\xi_{0i} = \xi_{1i} = X_i$, $i = 1, \dots, N$;
- Regularization parameters $\lambda_S > 0$ and $\lambda_0, \lambda_1 > 0$;
- LASSO Penalty: $\lambda_0 \|\beta_J^0\| + \lambda_1 \|\beta_J^1\| \longrightarrow$ LARS solution paths;

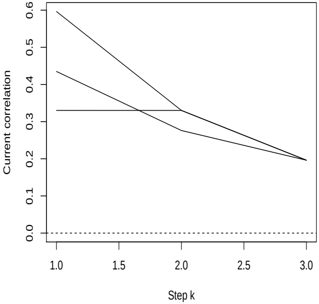
LARS Algorithm - Solution Paths



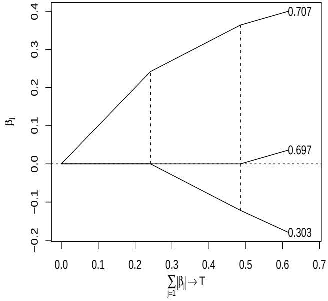


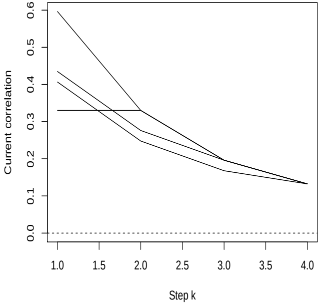
LARS Algorithm - Solution Paths



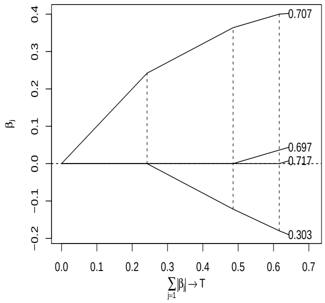


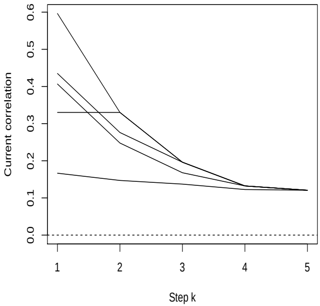
LARS Algorithm - Solution Paths



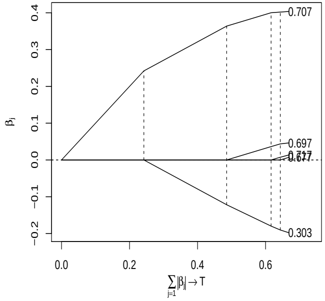


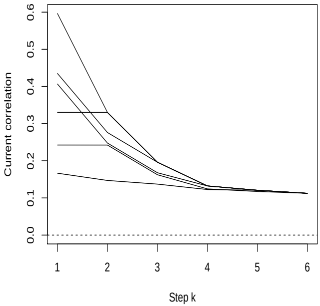
LARS Algorithm - Solution Paths



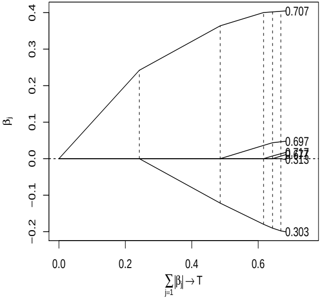


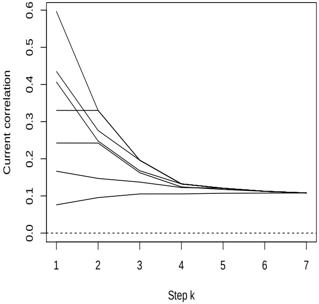
LARS Algorithm - Solution Paths



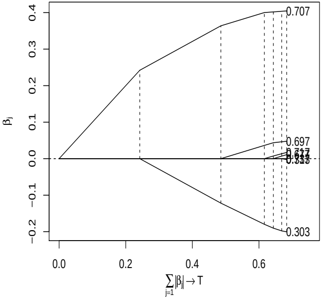


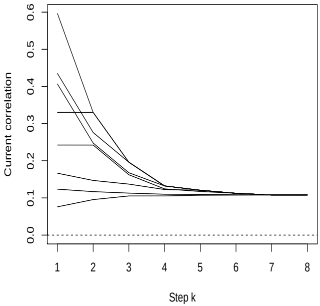
LARS Algorithm - Solution Paths



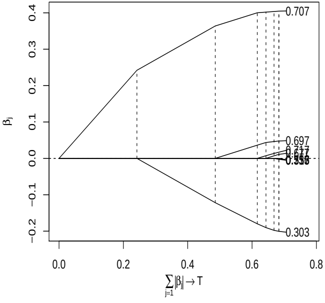


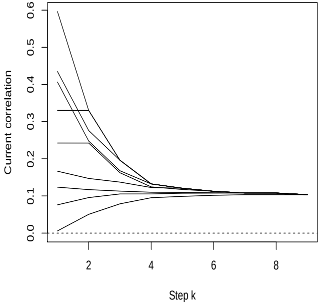
LARS Algorithm - Solution Paths



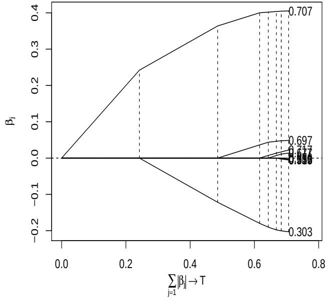


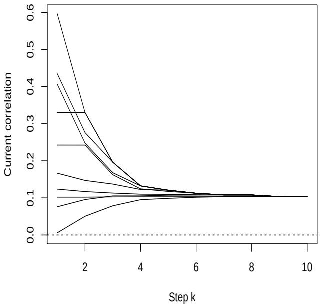
LARS Algorithm - Solution Paths



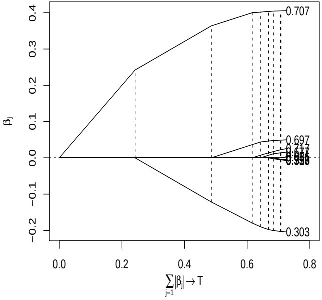


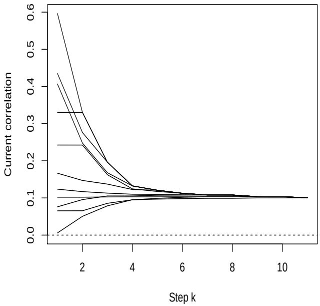
LARS Algorithm - Solution Paths



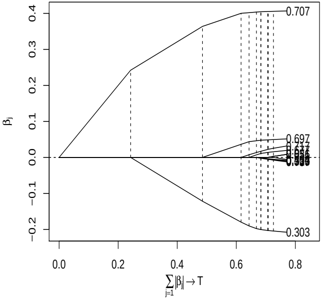


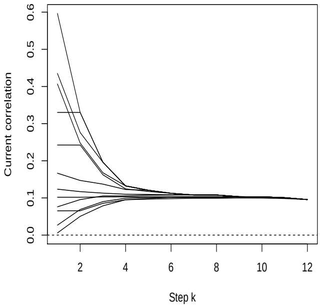
LARS Algorithm - Solution Paths



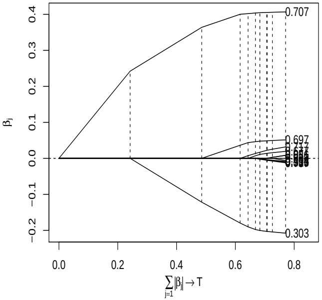


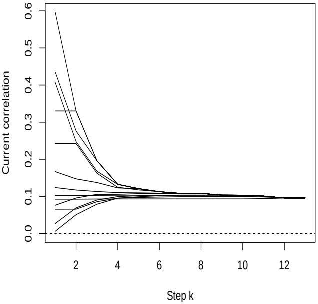
LARS Algorithm - Solution Paths



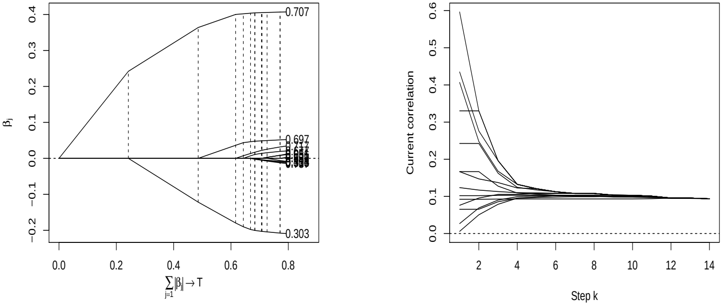


LARS Algorithm - Solution Paths

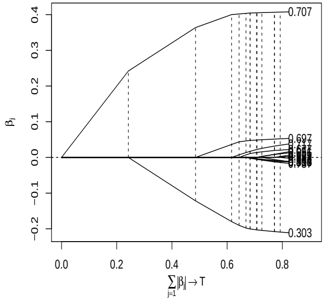


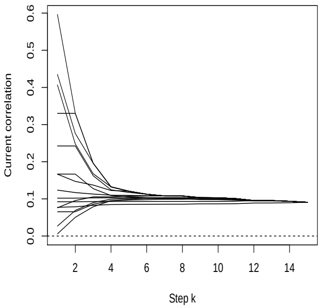


LARS Algorithm - Solution Paths

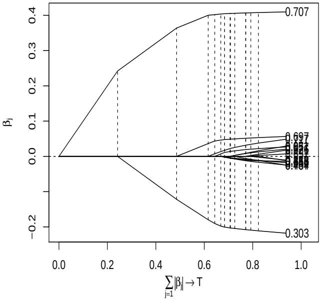


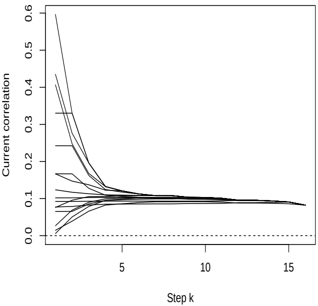
LARS Algorithm - Solution Paths



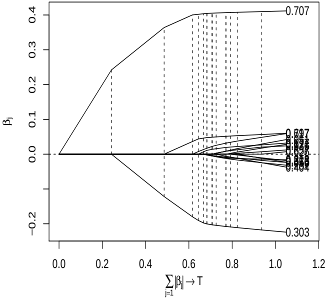


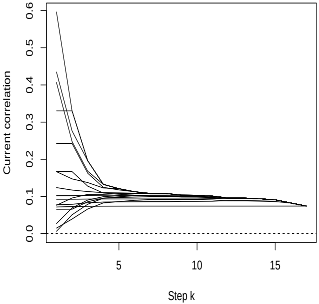
LARS Algorithm - Solution Paths



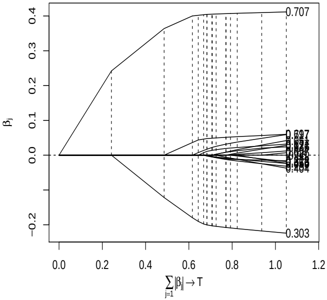


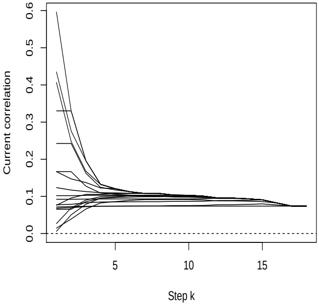
LARS Algorithm - Solution Paths



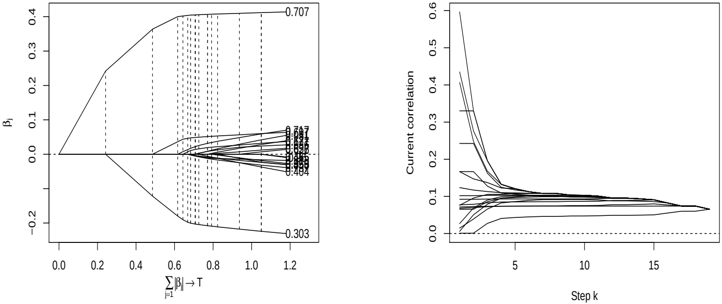


LARS Algorithm - Solution Paths

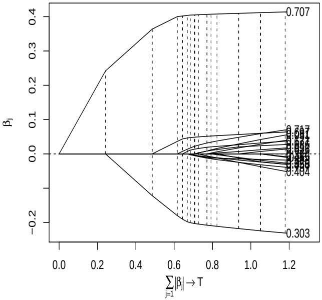


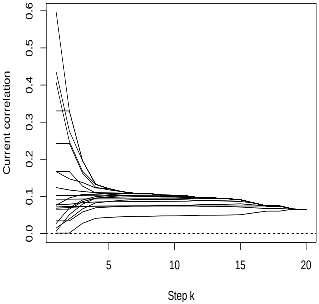


LARS Algorithm - Solution Paths

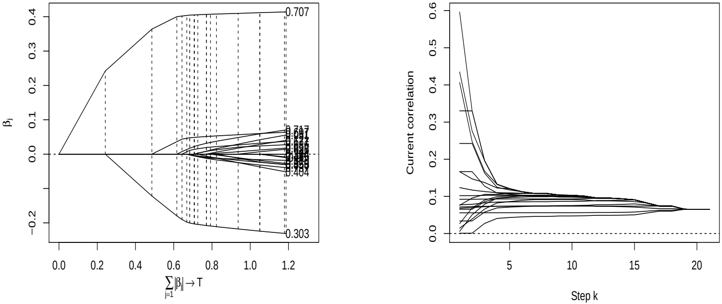


LARS Algorithm - Solution Paths

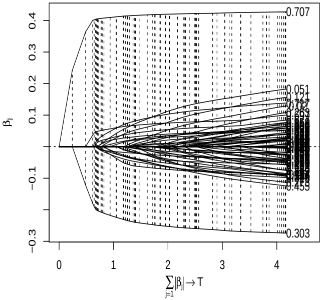


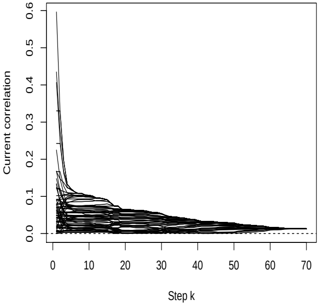


LARS Algorithm - Solution Paths

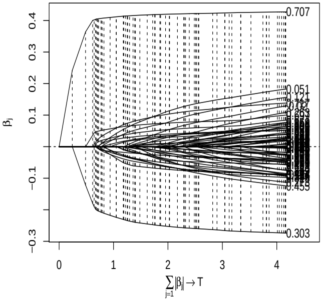


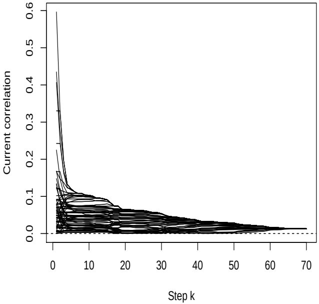
LARS Algorithm - Solution Paths





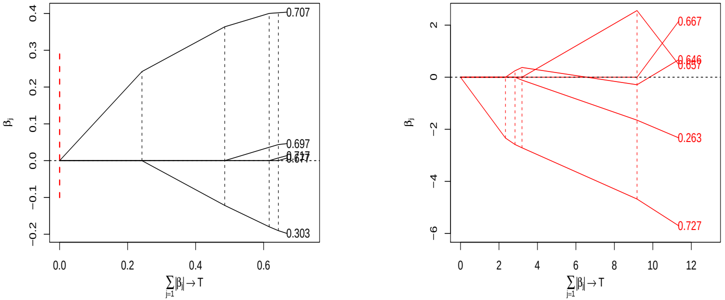
LARS Algorithm - Solution Paths



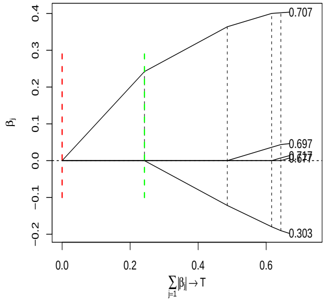


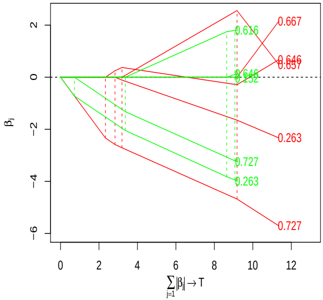
- piece-wise linear solution paths along a sequence $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$; (these "knot points" depend on $\boldsymbol{Y}$ and $\mathbb{X}$)
- piece-wise linear decrease in maximum (current) correlation $\mathbb{X}^\top(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_k)$;

Multiple Penalties

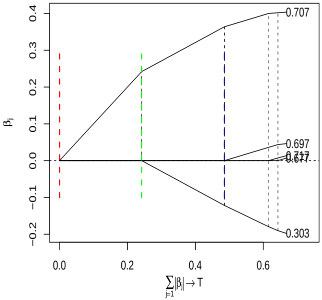


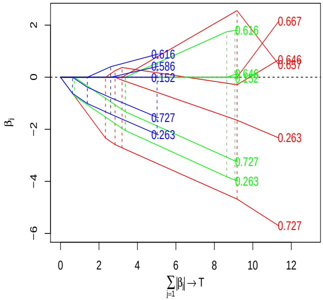
Multiple Penalties



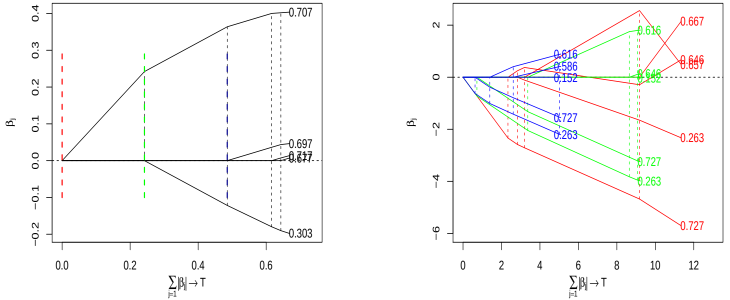


Multiple Penalties





Multiple Penalties



❑ How to choose the final model from the set of plausible ones?

A little bit of inference
on change-points

A little bit of inference

on change-points

- consistency;
- hypothesis tests;
- confidence regions;

Degrees of Freedom

- Degrees of freedom: $df(fit) = \frac{1}{\sigma^2}\sum_{i=1}^{n} Cov\left(\hat{Y}_i, Y_i\right)$;
- linear regression $\Rightarrow$ trace of the hat matrix $\Rightarrow$ number of parameters;
- smoothing splines $\Rightarrow$ trace of $\mathbb{X}\left(\mathbb{X}^\top\mathbb{X} - \sqrt{\lambda_1}\mathbb{W}_1^\top\mathbb{W}_1\right)^{-1}\mathbb{X}^\top$;
- LASSO regression $\Rightarrow$ average number of effective parameters;
  (result generalized by Tibshirani and Taylor (2012) even for $p \geq n$;
- splines with change-points: $\Rightarrow$ hat matrix trace + number of changes;

Degrees of Freedom

- Degrees of freedom: $df(fit) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov\left(\hat{Y}_i, Y_i\right)$;

- linear regression $\Rightarrow$ trace of the hat matrix $\Rightarrow$ number of parameters;
- smoothing splines $\Rightarrow$ trace of $\mathbb{X}\left(\mathbb{X}^\top\mathbb{X} - \sqrt{\lambda_1}\mathbb{W}_1^\top\mathbb{W}_1\right)^{-1}\mathbb{X}^\top$;
- LASSO regression $\Rightarrow$ average number of effective parameters;
  (result generalized by Tibshirani and Taylor (2012) even for $p \geq n$;
- splines with change-points: $\Rightarrow$ hat matrix trace + number of changes;

  - Mutually Independent Change-points: $df = |\mathcal{A}_0| + \cdots + |\mathcal{A}_{p-1}|$;
  - Simultaneous Change-points: $df = 3 \times |\mathcal{A}|$;
  - Hierarchical Change-points: $df = |\mathcal{A}_0| + \cdots + |\mathcal{A}_{p-1}|$;

Consistency of Estimates

- for now, only consistency with respect to change-points estimates; (considering a model $\widehat{\boldsymbol{\beta}}_J = Argmin\|\boldsymbol{Y} - \mathbb{X}_J\boldsymbol{\beta}_J\|^2 + \lambda\|\boldsymbol{\beta}_J\|_1$)
- restriction on the number of change-points (including their positions); (in general, we assume at most $\mathcal{K} \in \mathbb{N}$ change-points)

Consistency of Estimates

- for now, only consistency with respect to change-points estimates; (considering a model $\widehat{\boldsymbol{\beta}}_J = Argmin\|\boldsymbol{Y} - \mathbb{X}_J\boldsymbol{\beta}_J\|^2 + \lambda\|\boldsymbol{\beta}_J\|_1$)
- restriction on the number of change-points (including their positions); (in general, we assume at most $\mathcal{K} \in \mathbb{N}$ change-points)

Theorem (Consistency 1)

Under some common assumptions, for all $n \geq 1$ and $C > 2\sqrt{2}$, we have with a probability larger than $1 - n^{1-C^2/8}$, that

$$\left\|\mathbb{X}_J\left(\widehat{\boldsymbol{\beta}}_J(\lambda_n) - \boldsymbol{\beta}_J\right)\right\| \leq (2C\sigma\mathcal{K}\beta_{max})^{1/2} \cdot \left(\frac{\log n}{n}\right)^{1/4},$$

where $\lambda_n = C\sigma\sqrt{\log n/n}$, with an active set of pamameters $\mathcal{A}$.

- idea of the proof: extension of proof in Bickel, Ritov and Tsybakov (2009);

Consistency of Locations

- again, consistency with respect to change-points locations; (considering a model $\widehat{\boldsymbol{\beta}_J} = Argmin\|\boldsymbol{Y} - \mathbb{X}_J\boldsymbol{\beta}_J\|^2 + \lambda\|\boldsymbol{\beta}_J\|_1$)
- two change-point locations are not too much close to each other; (in general, we need enough data points to estimate each change-point)

Theorem (Consistency 2)

It can be shown that

$$\mathbb{P}\left(max_{1\leq k\leq|\widehat{\mathcal{A}}(\lambda_J)|}|\widehat{t}_k - t_k^\star| \leq n\delta_n\right) \overset{n\to\infty}{\longrightarrow} 1,$$

for some nonincreasing, positive sequence $\{\delta_n\}_{n\geq 1}$ *tending to zero, such that* $n\delta_n$

- generalization of the result of Harchaoui and Lévy-Leduc (2010)

Significance Test for LASSO

- classical theory based on RSS drop between two models not applicable;
  $\hookrightarrow$ test statistics: $R_j = (RSS_M - RSS_{M \cup \{j\}})/\sigma^2 \rightarrow \chi^2$ distribution
- in situations where $p \geq n$ the sets M and $M \cup \{j\}$ are not fixed any more;
  $\hookrightarrow$ using classical approach is way too far liberal (large type I. error)
- alternative approach must account for adaptivity of the LASSO procedure;
  $\hookrightarrow$ adaptiveness vs. shrinkage
- covariance test statistic proposed by Lockhart et at. (2013);
  $\hookrightarrow$ test statistics: $T_k = \left(\langle \boldsymbol{Y}, \mathbb{X}\widehat{\beta}(\lambda_{k+1}) \rangle - \langle \boldsymbol{Y}, \mathbb{X}_{\mathcal{A}}\widetilde{\beta}_{\mathcal{A}}(\lambda_{k+1}) \rangle \right) / \sigma^2$
- under the null hypothesis ($supp(\beta^\star) \subseteq \mathcal{A}$) it holds that:
  $\hookrightarrow$ test statistic $T_k \longrightarrow Exp(1)$ in distribution;

Confidence Regions

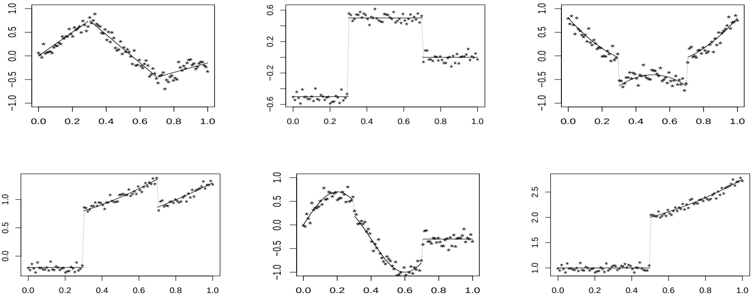
Point-wise Confidence Bands

- for the vector of parameters $(\beta_S^\top, \beta_J^\top)$ we have a pseudo design matrix;
- we can define a sandwich estimate for the covariance matrix;
- if variance σ^2 is unknown $\Rightarrow$ need for an estimate $\widehat{\sigma_n^2}$;

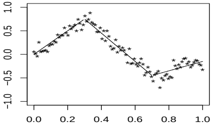
Uniform Confidence Bands

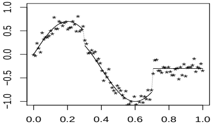
- the idea is to obtain a band $B_n(x)$ for m_0 ($s_0, \ldots, s_{p-1}$ resp.), such that $\mathbb{P}(f(x) \in B_{n,f}(x)) = 1 - \alpha$, for $f \in \{m_0, s_0, \ldots, s_{p-1}\}$;
- idea of the band construction: Hotelling (1939);
  (also Krivobokova et al. (2013) and Koenker (2011))
- however, requires continuity at least $\Rightarrow$ not applicable for s_0 yet;

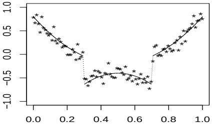
Some Simulation Results



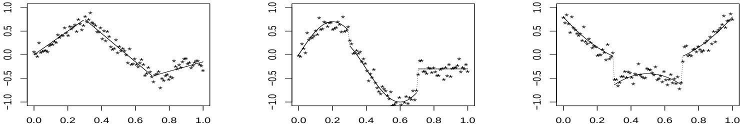
Some Simulation Results





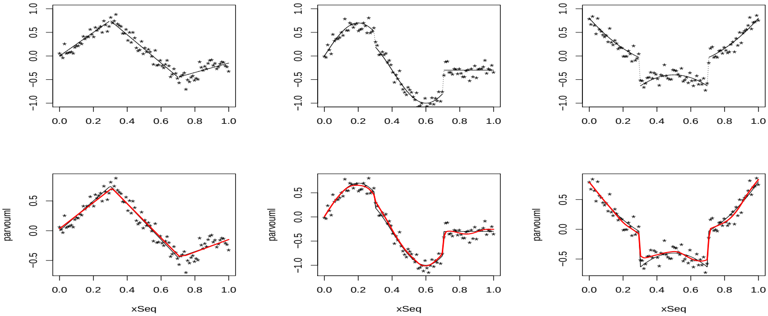


Some Simulation Results

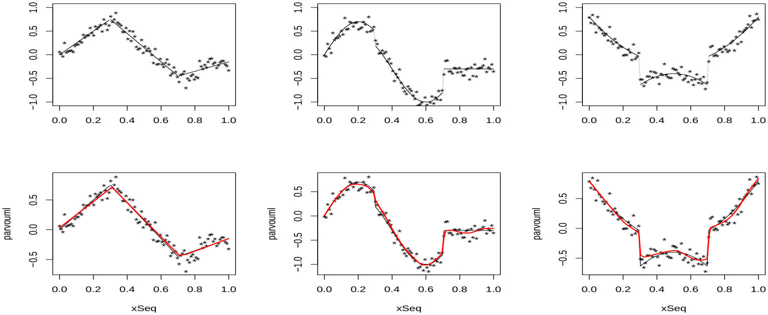


Mutually independent change-points

Some Simulation Results

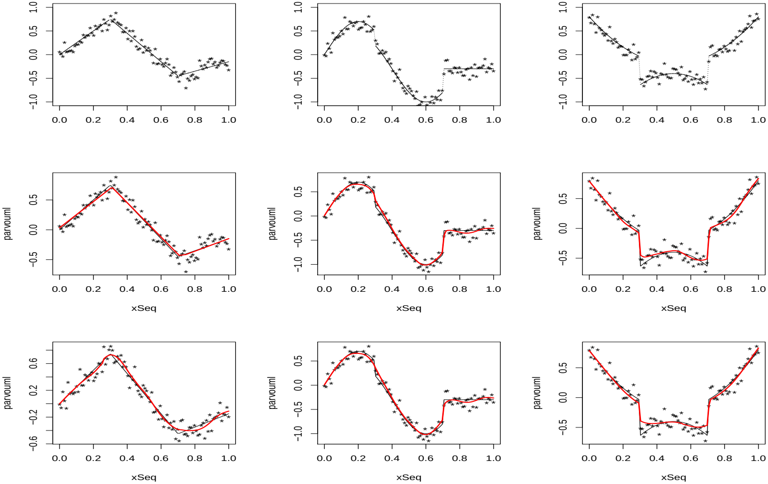


Some Simulation Results



Simultaneous change-points

Some Simulation Results



Independent Change-points

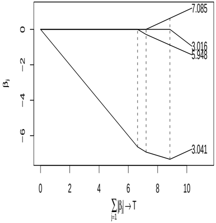
	Independent Change-points	$\sigma^2 = 0$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$	$\sigma^2 = 0.5$	$\sigma^2 = 1$
$\lambda_G = 0.1$	$\xi_1^{(0)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	3.8 | 1.8	49.4 | 25.8	69.6 | 38.5	94.5 | 49.3
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	69.2 | 31.8	85.8 | 58.5.0	100 | 69.2	100 | 68.1
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	100 | 12.2	100 | 31.5	100 | 37.6	100 | 52.9
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7,\ \xi_1^{(1)} = 0.5$	0.0 | 0.0	100 | 24.5	100 | 37.2	100 | 41.2	100 | 59.5
$\lambda_G = 0.01$	$\xi_1^{(0)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	2.6 | 1.0	38.5 | 20.18	65.8 | 35.4	82.1 | 49.1
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	71.0 | 33.3	87.2 | 60.0	100 | 63.6	100 | 68.5
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	100 | 30.4	100 | 55.4	100 | 60.2	100 | 66.7
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7,\ \xi_1^{(1)} = 0.5$	0.0 | 0.0	100 | 38.7	100 | 58.4	100 | 64.7	100 | 69.2
$\lambda_G = 0.001$	$\xi_1^{(0)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	3.3 | 1.4	52.3 | 27.0	77.9 | 43.7	97.6 | 57.0
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	82.2 | 56.8	92.2 | 70.8	100 | 75.3	100 | 77.3
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	100 | 58.7	100 | 73.7	100 | 75.4	100 | 78.2
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7,\ \xi_1^{(1)} = 0.5$	0.0 | 0.0	100 | 61.2	100 | 77.1	100 | 80.2	100 | 79.1
$\lambda_G = 0.0001$	$\xi_1^{(0)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	2.6 | 0.9	62.5 | 35.2	87.4 | 52.9	98.7 | 59.3
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7$	0.0 | 0.0	98.7 | 70.2	99.9 | 71.7	100 | 76.6	100 | 76.6
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	100 | 75.2	99.9 | 74.6	100 | 77.1	100 | 76.2
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3,\ \xi_2^{(0)} = 0.7,\ \xi_1^{(1)} = 0.5$	0.0 | 0.0	99.9 | 71.0	99.9 | 69.9	100 | 78.2	100 | 79.1

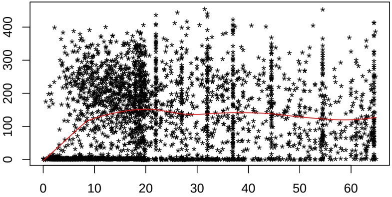
Mutually Related Change-points

	Mutually Related Change-points	$\sigma^2 = 0$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$	$\sigma^2 = 0.5$	$\sigma^2 = 1$
$\lambda_G = 0.1$	$\xi_1^{(0)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	4.1 | 2.8	51.4 | 27.7	70.6 | 40.5	90.0 | 50.3
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7, \xi_1^{(1)} = 0.5$	100	100	100	100	100
$\lambda_G = 0.01$	$\xi_1^{(0)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	3.8 | 3.1	49.1 | 29.3	72.1 | 40.5	93.7 | 51.9
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7, \xi_1^{(1)} = 0.5$	100	100	100	100	100
$\lambda_G = 0.001$	$\xi_1^{(0)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	4.0 | 2.2	55.5 | 29.2	74.2 | 45.5	97.2 | 54.0
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7, \xi_1^{(1)} = 0.5$	100	100	100	100	100
$\lambda_G = 0.0001$	$\xi_1^{(0)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7$	100	100	100	100	100
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = \xi_2^{(1)} = 0.7$	0.0 | 0.0	3.0 | 1.2	63.1 | 37.3	85.2 | 55.4	98.3 | 59.9
	$\xi_1^{(0)} = \xi_1^{(1)} = 0.3, \xi_2^{(0)} = 0.7, \xi_1^{(1)} = 0.5$	100	100	100	100	100

Back to Parvovirus B19 Data

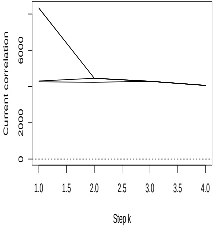
- ❑ intercept corrected smoothing B-spline basis + change-point basis;
  $\Rightarrow$ smoothness degree $p = 3$, change-points up to the order $p - 1 = 2$;
- ❑ mutually independent change-points assumed;
  $\Rightarrow$ four smoothing parameters $\lambda_S, \lambda_0, \lambda_1, \lambda_2 > 0$;

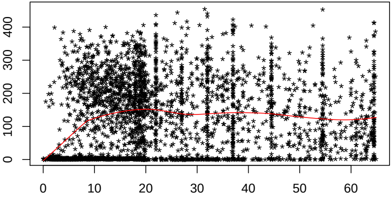




Back to Parvovirus B19 Data

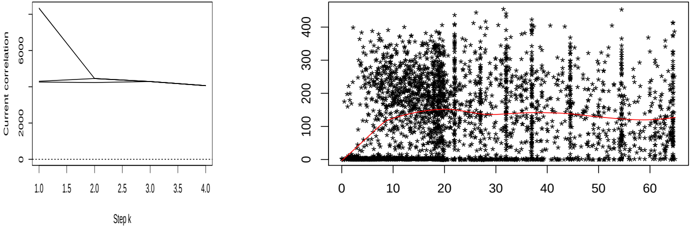
- intercept corrected smoothing B-spline basis + change-point basis;
  ⇒ smoothness degree $p = 3$, change-points up to the order $p - 1 = 2$;
- mutually independent change-points assumed;
  ⇒ four smoothing parameters $\lambda_S, \lambda_0, \lambda_1, \lambda_2 > 0$;



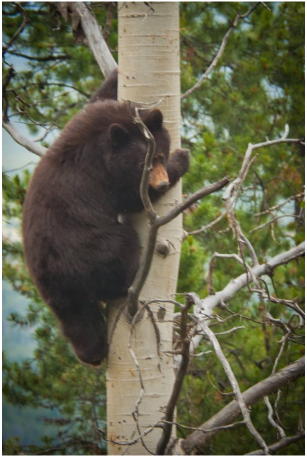


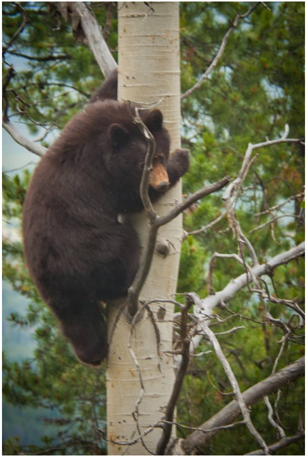
Back to Parvovirus B19 Data

- intercept corrected smoothing B-spline basis + change-point basis;
  ⇒ smoothness degree $p = 3$, change-points up to the order $p - 1 = 2$;
- mutually independent change-points assumed;
  ⇒ four smoothing parameters $\lambda_S, \lambda_0, \lambda_1, \lambda_2 > 0$;

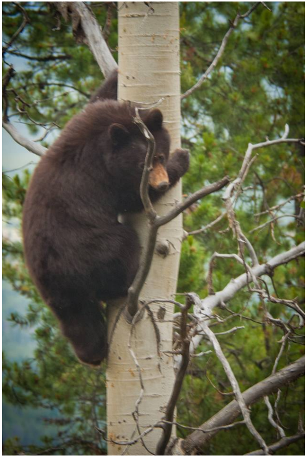


- one change-point in location (roughly at the age of 20);
- in addition, also a change-point in direction revealed (age 8 – 9);
  (even significant - p-value below 0.05)





Thank you...



Thank you...

...any questions?

✍ Efron, B., Hastie, T. and Tibshirani, R. (2004). Least Angle Regression, *Annals of Statistics, 32*, 407 – 499.

✍ Jacob, L., Obozinski, G. and Vert, J.P. (2009). Group Lassso with Overlap and Graph Lasso. *Proceedings of the 26th International Conference on Machine Learning (ICML 26)*, Montreal, Canada.

✍ Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2013). A significance Test for the Lasso. (preprint)

✍ Koenker, R. (2011). Additive Models for Quantile Regression: Model Selection and Confidence Bandaids. *Brazilian Journal of Probability and Statistics, 25*, No.3, 239 – 262.

✍ Tibshirani, R. and Taylor, J. (2012). Degrees of Freedom in Lasso Problems, *Annals of Statistics, 40*, 1198 – 1232.