Extreme Value Theory for SiZer

Jan Hannig

hannig@stat.colostate.edu

Department of Statistics

Colorado State University

Joint work with:

J.S. Marron, University of North Carolina - Chapel Hill

Question

- Consider two parameter Gaussian random field $X_{i,j}$ with mean 0, variance 1 and

$$\mathrm{Cov}(X_{i,k}, X_{j,l}) = e^{-(j-i)^2\tilde{\Delta}^2/(2h^2(d^{2k}+d^{2l}))}\left[1 - \frac{(j-i)^2\tilde{\Delta}^2}{h^2(d^{2l}+d^{2k})}\right]\left(\frac{2d^{k+l}}{d^{2k}+d^{2l}}\right)^{3/2}.$$

Question

- Consider two parameter Gaussian random field $X_{i,j}$ with mean 0, variance 1 and

$$\mathrm{Cov}(X_{i,k}, X_{j,l}) = e^{-(j-i)^2\tilde{\Delta}^2/(2h^2(d^{2k}+d^{2l}))}\left[1 - \frac{(j-i)^2\tilde{\Delta}^2}{h^2(d^{2l}+d^{2k})}\right]\left(\frac{2d^{k+l}}{d^{2k}+d^{2l}}\right)^{3/2}.$$

- Question of interest: Find (asymptotic) distribution of

$$\max_{0<j<g, 0<k<r} X_{j,k}, \quad g, r \to \infty$$

Question

- Consider two parameter Gaussian random field $X_{i,j}$ with mean 0, variance 1 and

$$\mathrm{Cov}(X_{i,k}, X_{j,l}) = e^{-(j-i)^2\tilde{\Delta}^2/(2h^2(d^{2k}+d^{2l}))}\left[1 - \frac{(j-i)^2\tilde{\Delta}^2}{h^2(d^{2l}+d^{2k})}\right]\left(\frac{2d^{k+l}}{d^{2k}+d^{2l}}\right)^{3/2}.$$

- Question of interest: Find (asymptotic) distribution of

$$\max_{0<j<g, 0<k<r} X_{j,k}, \quad g, r \to \infty$$

- This field is stationary in the first parameter but non-stationary in the second parameter.

Row-wise solutions

- Fix k and consider $T_j = X_{j,k}$ (T_j is stationary). Several approaches are possible here:

Row-wise solutions

- Fix k and consider $T_j = X_{j,k}$ (T_j is stationary). Several approaches are possible here:
  - Berman (1964) shows that $\max(T_1, ..., T_g)$ behaves asymptotically the same way as the max of g i.i.d. Gaussian random variables. (Convergence rate too slow, not useful here.)

Row-wise solutions

- Fix k and consider $T_j = X_{j,k}$ (T_j is stationary). Several approaches are possible here:
  - Berman (1964) shows that $\max\left(T_1, ..., T_g\right)$ behaves asymptotically the same way as the max of g i.i.d. Gaussian random variables. (Convergence rate too slow, not useful here.)
  - Rootzen (1983) gives a second order term to the approximation.

Row-wise solutions

- Fix k and consider $T_j = X_{j,k}$ (T_j is stationary). Several approaches are possible here:
  - Berman (1964) shows that $\max(T_1, ..., T_g)$ behaves asymptotically the same way as the max of g i.i.d. Gaussian random variables. (Convergence rate too slow, not useful here.)
  - Rootzen (1983) gives a second order term to the approximation.
  - Hsing, Husler and Reiss (1996) give an alternative approach using a triangular array with increasing correlation. (Works best for our application).

Main idea of Hsing et al's

- Embed the sequence T_i into a triangular array $\{\hat{T}_{j,g}\}$.

Main idea of Hsing et al's

- Embed the sequence T_i into a triangular array $\{\hat{T}_{j,g}\}$.
- $\hat{T}_{j,g},\ j = 1, 2, \ldots$ are Gaussian, mean zero, variance one, with $\rho_{j,g}$ satisfying $\log(g)\,(1 - \rho_{j,g}) \to \delta_j$ as $g \to \infty$, for all j.

Main idea of Hsing et al's

- Embed the sequence T_i into a triangular array $\{\hat{T}_{j,g}\}$.
- $\hat{T}_{j,g},\ j = 1, 2, \ldots$ are Gaussian, mean zero, variance one, with $\rho_{j,g}$ satisfying $\log(g)\,(1-\rho_{j,g}) \to \delta_j$ as $g \to \infty$, for all j.
- $\lim_{g\to\infty} P\left[\max_{i=1,\ldots,g} \hat{T}_{i,g} \le u(x)\right] = e^{-\vartheta e^{-x}}$, where $u(x) = \sqrt{2\log g} + \frac{x}{\sqrt{2\log g}} - \frac{\log\log g + \log 4\pi}{\sqrt{8\log g}}$ and

$$\vartheta = P\left[V/2 + \sqrt{\delta_k} H_k \le \delta_j \text{ for all } j \ge 1\right].$$

  V is exponential(1), H_k is a mean zero Gaussian process with $EH_iH_j = \frac{\delta_i + \delta_j - \delta_{|i-j|}}{2\sqrt{\delta_i\delta_j}}$, V and H_k are independent.

Extension: Hannig (2005+)

- Embed the sequence $T_{i,j}$ an isotropic stationary mean 0 and variance 1 Gausian random field embedded into a triangular array $\{\hat{T}_{i,j,g}\}$.

Extension: Hannig (2005+)

- Embed the sequence $T_{i,j}$ an isotropic stationary mean 0 and variance 1 Gausian random field embedded into a triangular array $\{\hat{T}_{i,j,g}\}$.
- Denote the correlation $\rho_{i,j,g} = E\hat{T}_{k,l,g}\hat{T}_{k+i,l+j,g}$ and assume that $\lim_{g\to\infty}(1-\rho_{i,j,g})\log g = \delta_{i,j} \in (0,\infty]$.

Extension: Hannig (2005+)

- Embed the sequence $T_{i,j}$ an isotropic stationary mean 0 and variance 1 Gausian random field embedded into a triangular array $\{\hat{T}_{i,j,g}\}$.
- Denote the correlation $\rho_{i,j,g} = E\hat{T}_{k,l,g}\hat{T}_{k+i,l+j,g}$ and assume that $\lim_{g\to\infty}(1-\rho_{i,j,g})\log g = \delta_{i,j} \in (0,\infty]$.
- $\lim_{g\to\infty} P\left(\max_{i=1,\ldots,g}\max_{j=1,\ldots,g}\hat{T}_{i,j,g} \le u_{g^2}(x)\right) = e^{-\theta e^{-x}}$, where

$$\theta = P\left(V/2 + \sqrt{\delta_{i,j}}H_{i,j} \le \delta_{i,j},\ (i,j) \in \{0,1,2,\ldots\}^2 \setminus \{(0,0)\}\right),$$

and $EH_{i,j}H_{k,l} = \frac{\delta_{i,j}+\delta_{k,l}-\delta_{|i-k|,|j-l|}}{2\sqrt{\delta_{i,j}\delta_{k,l}}}$.

Extension: Hannig (2005+)

- Embed the sequence $T_{i,j}$ an isotropic stationary mean 0 and variance 1 Gausian random field embedded into a triangular array $\{\hat{T}_{i,j,g}\}$.
- Denote the correlation $\rho_{i,j,g} = E\hat{T}_{k,l,g}\hat{T}_{k+i,l+j,g}$ and assume that $\lim_{g\to\infty}(1-\rho_{i,j,g})\log g = \delta_{i,j} \in (0,\infty]$.
- $\lim_{g\to\infty} P\left(\max_{i=1,\dots,g}\max_{j=1,\dots,g}\hat{T}_{i,j,g} \le u_{g^2}(x)\right) = e^{-\theta e^{-x}}$, where

$$\theta = P\left(V/2 + \sqrt{\delta_{i,j}}H_{i,j} \le \delta_{i,j},\ (i,j) \in \{0,1,2,\dots\}^2 \setminus \{(0,0)\}\right),$$

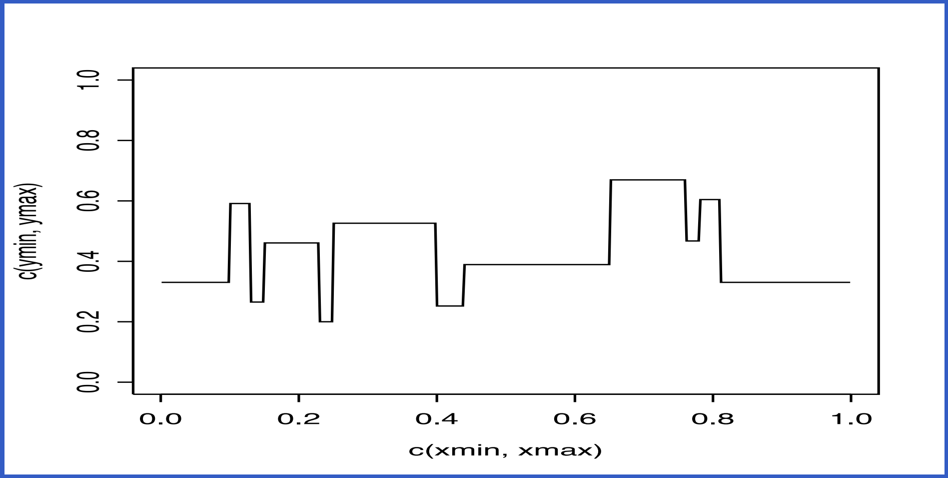
  and $EH_{i,j}H_{k,l} = \frac{\delta_{i,j}+\delta_{k,l}-\delta_{|i-k|,|j-l|}}{2\sqrt{\delta_{i,j}\delta_{k,l}}}$.
- Not fully satisfactory yet.

Introduction – Kernel based smoothing

- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.

Introduction – Kernel based smoothing

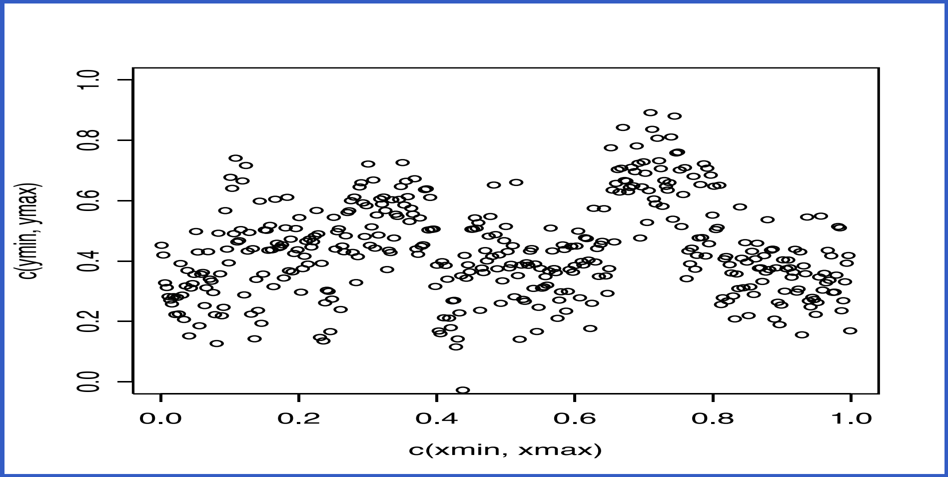
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



unknown function f

Introduction – Kernel based smoothing

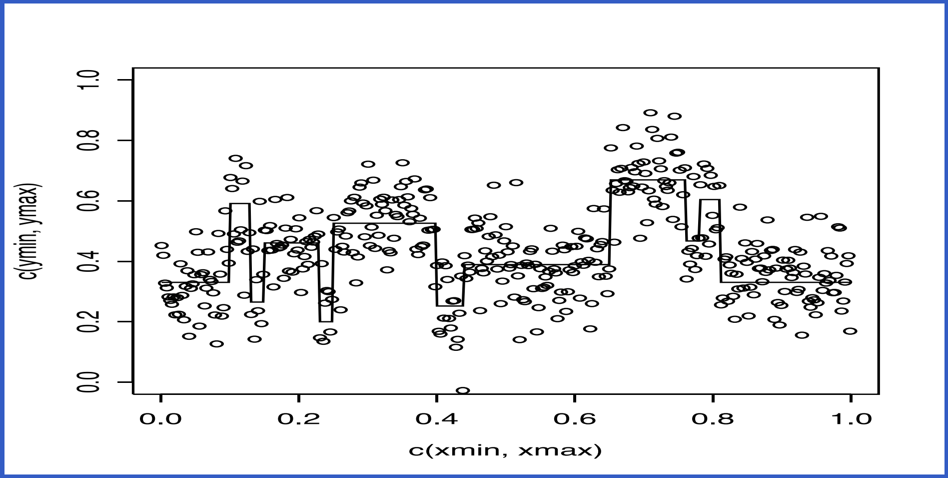
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



data

Introduction – Kernel based smoothing

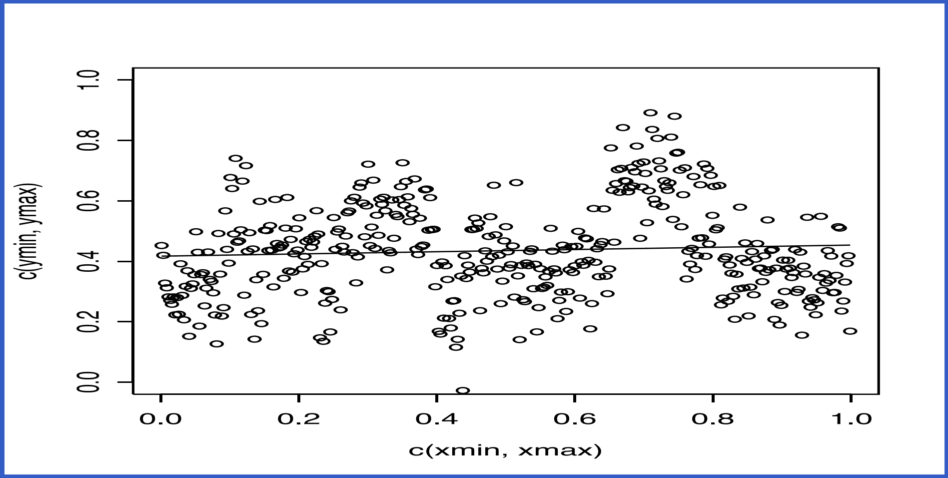
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



data and f

Introduction – Kernel based smoothing

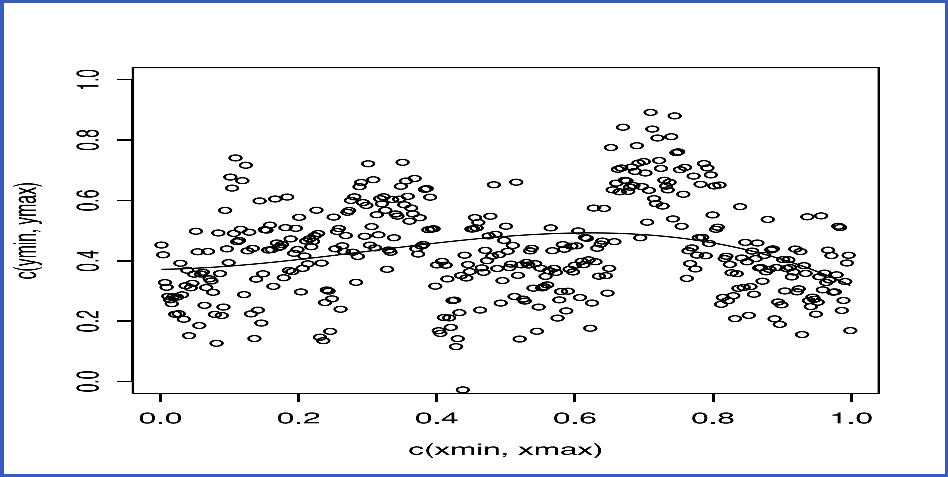
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



linear regression

Introduction – Kernel based smoothing

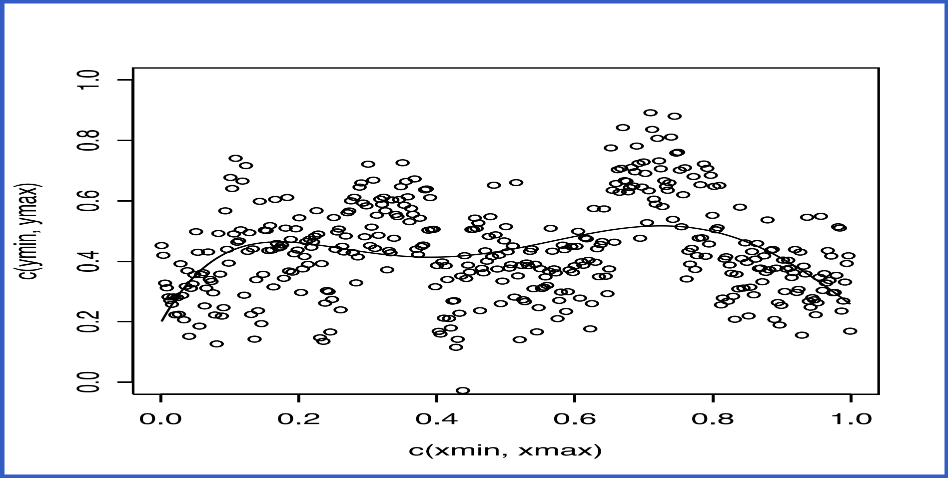
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



cubic regression

Introduction – Kernel based smoothing

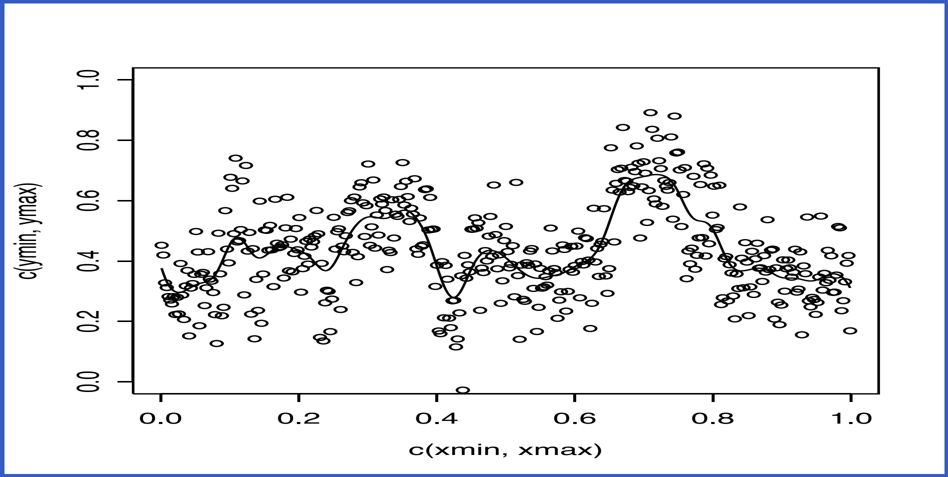
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



quintic regression

Introduction – Kernel based smoothing

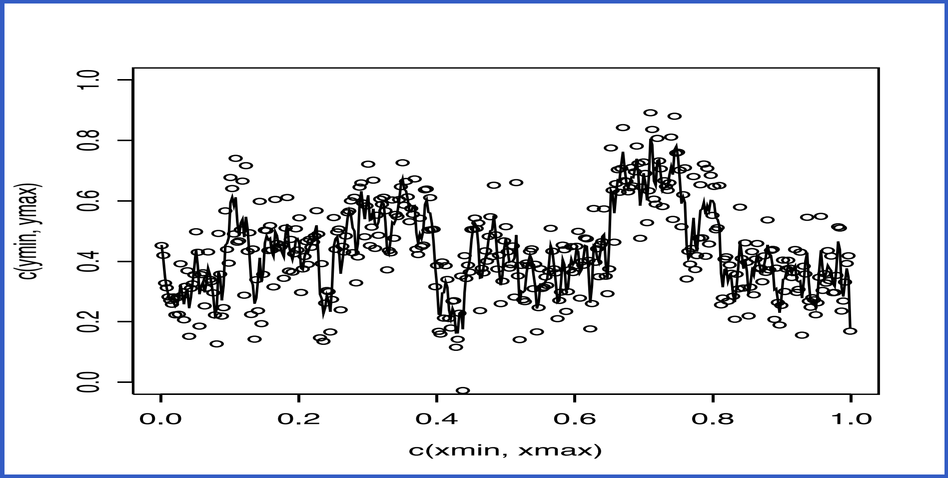
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



local linear with window width $b = .016$

Introduction – Kernel based smoothing

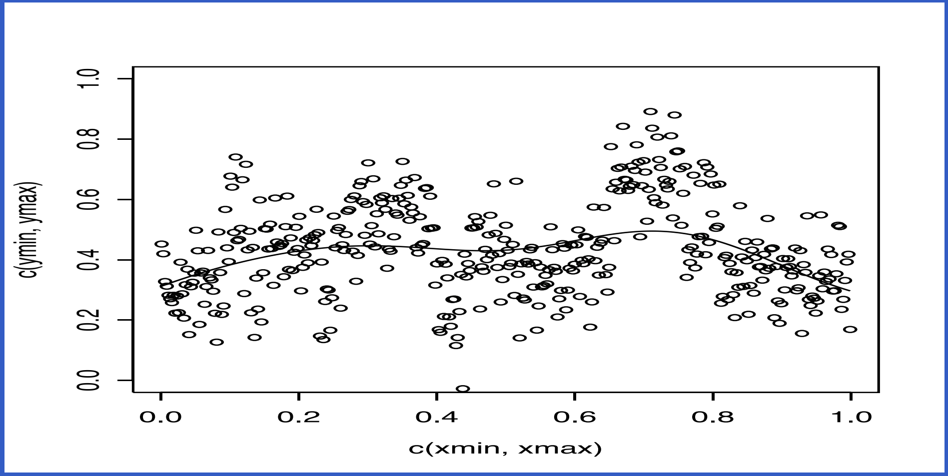
- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



local linear with window width $b = .002$

Introduction – Kernel based smoothing

- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:



local linear with window width $b = .128$

Introduction – Kernel based smoothing

- When given a data scatter-plot people commonly assume that $Y_j = f(X_j) + \varepsilon_j$ where X_j and Y_j are observed and ε_j is i.i.d. noise. We want to estimate f.
- As an example consider:
- The main issue for kernel based methods is the choice of window width b. No agreement on how properly do this has been reached among statistician.

Introduction — SiZer

- SiZer was introduced by Chaudhury and Marron (1999) as a tool for exploratory data analysis.

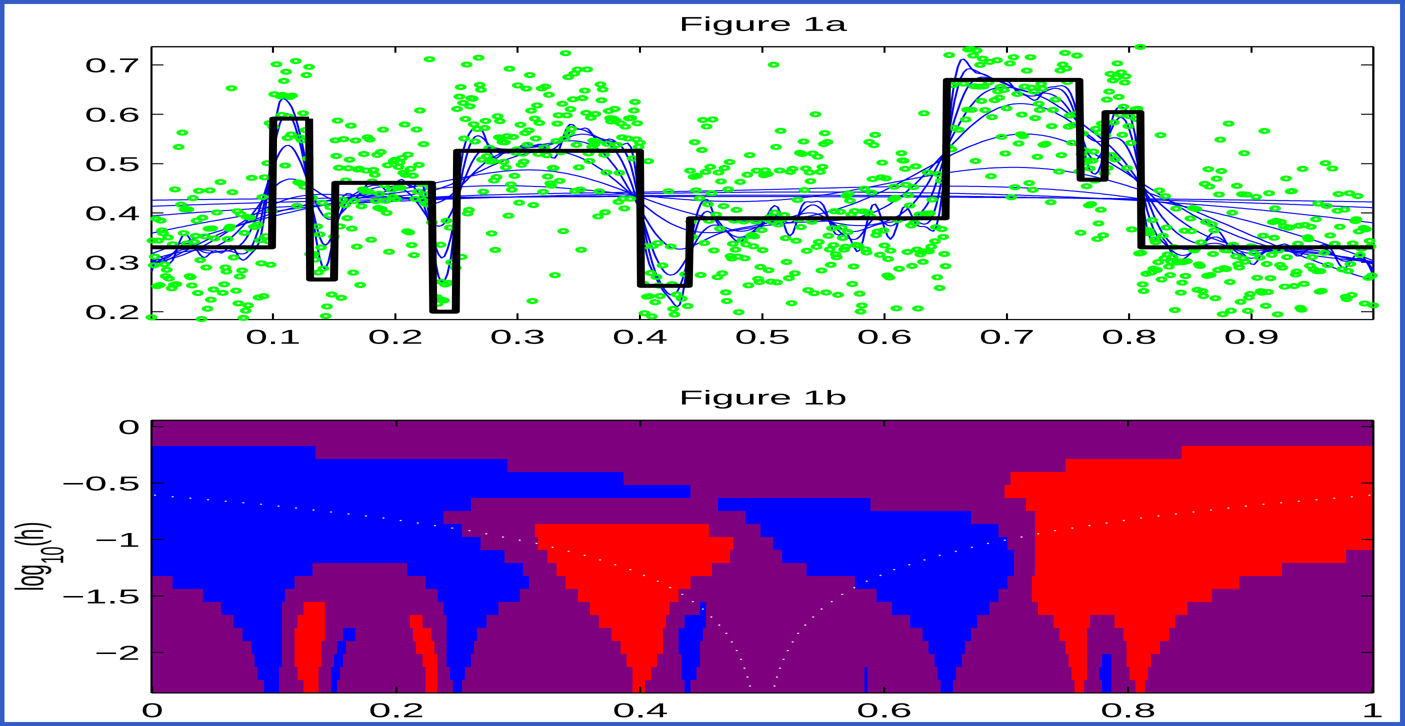
Introduction — SiZer

- SiZer was introduced by Chaudhury and Marron (1999) as a tool for exploratory data analysis.
- Instead of estimating f it addresses the question “what features are in the data” by:
  - Using several bandwidth to smooth the data.
  - Testing whether the derivative of the smoothed “true function” is positive or negative.

Introduction — SiZer

- SiZer was introduced by Chaudhury and Marron (1999) as a tool for exploratory data analysis.
- Instead of estimating f it addresses the question “what features are in the data” by:
  - Using several bandwidth to smooth the data.
  - Testing whether the derivative of the smoothed “true function” is positive or negative.
- The outcome is the SiZer color map that allows us to find “bumps in the data”.

Example



Conventional SiZer analysis of the Donoho - Johnstone Blocks regression, with high noise. True regression, data and scale space shown in Figure 1a. SiZer analysis in Figure 1b.

Size issue in conventional SiZer

- Every pixel on the SiZer map corresponds to a statistical test determining if the estimate of the first derivative is statistically different from 0. Columns correspond to locations and rows to bandwidths (kth row uses bandwidth hd^k).

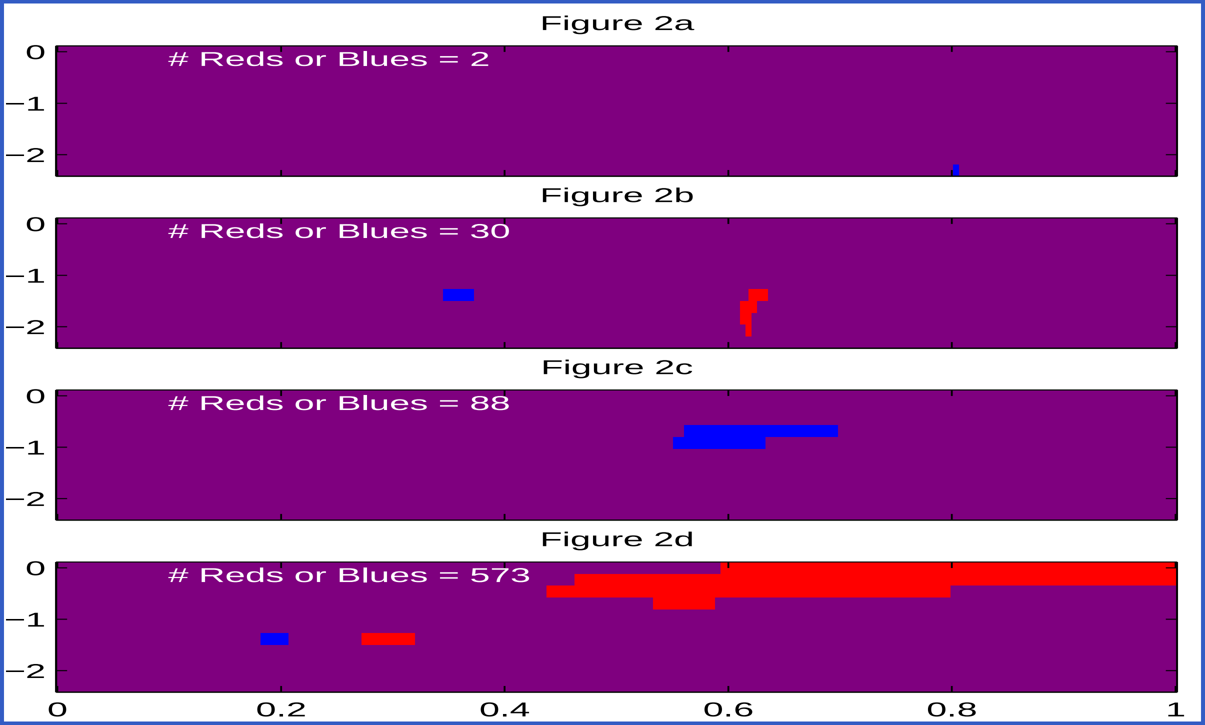
Size issue in conventional SiZer

- Every pixel on the SiZer map corresponds to a statistical test determining if the estimate of the first derivative is statistically different from 0. Columns correspond to locations and rows to bandwidths (kth row uses bandwidth hd^k).
- A multiple testing procedure is required. Originally SiZer used an ad-hoc multiple testing adjustment got way too many false positives.

Size issue in conventional SiZer

- Every pixel on the SiZer map corresponds to a statistical test determining if the estimate of the first derivative is statistically different from 0. Columns correspond to locations and rows to bandwidths (kth row uses bandwidth hd^k).
- A multiple testing procedure is required. Originally SiZer used an ad-hoc multiple testing adjustment got way too many false positives.
- If no signal is present, i.e., the data is just constant + noise, the SiZer map should be entirely purple. However that is often not the case.

Simulation - original SiZer



Conventional SiZer maps, based on simulated null distributions, for 1600 equally spaced regression data points. Figures 2a, b, c and d are for 0.5, 0.75, 0.85 and 0.95, respectively, quantiles of distribution.

SiZer distribution

- Each of the r rows of the SiZer map is created by a g of statistical tests. There are a total of $g \times r$ tests.

SiZer distribution

- Each of the r rows of the SiZer map is created by a g of statistical tests. There are a total of $g \times r$ tests.
- Under the null hypothesis of “no-signal”

$$T_{j,k} \approx -C \int_{-\infty}^{\infty} \phi' \left(\frac{j-x}{hd^k} \right) dB(x).$$

SiZer distribution

- Each of the r rows of the SiZer map is created by a g of statistical tests. There are a total of $g \times r$ tests.
- Under the null hypothesis of "no-signal"

$$T_{j,k} \approx -C \int_{-\infty}^{\infty} \phi' \left(\frac{j-x}{hd^k} \right) dB(x).$$

- $T_{1,1}, \ldots, T_{g,r}$ can be approximated by a mean 0, variance 1, Gaussian random field with

$$\mathrm{Cov}(T_{i,k}, T_{i+j,l}) = e^{-j^2\tilde{\Delta}^2/(2h^2(d^{2k}+d^{2l}))} \left[1 - \frac{j^2\tilde{\Delta}^2}{h^2(d^{2l}+d^{2k})} \right] \left(\frac{2d^{k+l}}{d^{2k}+d^{2l}} \right)^{3/2}.$$

Δ is the distance between two pixels in the SiZer map, hd^k is the bandwidth used for the kth row.

Solution

- To do the proper multiple adjustment we need to investigate the behavior of $\max_{j,k} T_{j,k}$. The idea is to use the quantile of this distribution to set up rejection region.

Solution

- To do the proper multiple adjustment we need to investigate the behavior of $\max_{j,k} T_{j,k}$. The idea is to use the quantile of this distribution to set up rejection region.
- Row-wise approach: Study a maximum of a fixed row. The goal is to have only $\alpha\%$ of rows to contain false positives.

Solution

- To do the proper multiple adjustment we need to investigate the behavior of $\max_{j,k} T_{j,k}$. The idea is to use the quantile of this distribution to set up rejection region.
- Row-wise approach: Study a maximum of a fixed row. The goal is to have only $\alpha\%$ of rows to contain false positives.
- Global approach: Study a maximum of the whole random field. The goal is to have only $\alpha\%$ of SiZer maps to contain false positives.

Row-wise solution

- Consider triangular array of SiZer rows. The number of pixels $g \to \infty$, correlation between pixels is

$$\mathrm{Cov}(\hat{T}_{i,g}, \hat{T}_{i+j,g}) = e^{-j^2 c^2/(4\log g)} \left[1 - \frac{j^2 c^2}{2\log g} \right]$$

This leads to $\delta_j = 3c^2 j^2/4$ and consequently $H_1 = H_2 = \cdots = H \sim N(0,1)$.

Row-wise solution

- Consider triangular array of SiZer rows. The number of pixels $g \to \infty$, correlation between pixels is

$$\mathrm{Cov}(\hat{T}_{i,g}, \hat{T}_{i+j,g}) = e^{-j^2c^2/(4\log g)}\left[1 - \frac{j^2c^2}{2\log g}\right]$$

  This leads to $\delta_j = 3c^2j^2/4$ and consequently $H_1 = H_2 = \cdots = H \sim N(0,1)$.

- To use Hsing et al's theorem we need to calculate

$$\vartheta = P\left[V/2 + j\sqrt{\frac{3c^2}{4}}H \leq \frac{3c^2}{4}j^2 \text{ for all } j \geq 1\right] = 2\Phi\left(\sqrt{3c^2/4}\right) - 1.$$

Row-wise solution

- Consider triangular array of SiZer rows. The number of pixels $g \to \infty$, correlation between pixels is

$$\mathrm{Cov}(\hat{T}_{i,g}, \hat{T}_{i+j,g}) = e^{-j^2c^2/(4\log g)} \left[1 - \frac{j^2c^2}{2\log g}\right]$$

  This leads to $\delta_j = 3c^2j^2/4$ and consequently $H_1 = H_2 = \cdots = H \sim N(0,1)$.

- To use Hsing et al's theorem we need to calculate

$$\vartheta = P\left[V/2 + j\sqrt{\frac{3c^2}{4}}H \le \frac{3c^2}{4}j^2 \text{ for all } j \ge 1\right] = 2\Phi\left(\sqrt{3c^2/4}\right) - 1.$$

- Then $\lim_{g\to\infty} P\left[\max_{i=1,\ldots,g} \hat{T}_{i,g} \le u(x)\right] = e^{-\vartheta e^{-x}}$.

Row-wise solution

- The way to use this in practice is

$$P[\max\left(\hat{T}_{1,g},...,\hat{T}_{g,g}\right) \leq x] \approx \Phi(x)^{\vartheta g},$$

Row-wise solution

- The way to use this in practice is

$$P[\max\left(\hat{T}_{1,g},...,\hat{T}_{g,g}\right) \leq x] \approx \Phi(x)^{\vartheta g},$$

- Approximate the max of a fixed SiZer row by:

$$P[\max\left(T_{1,k},...,T_{g,k}\right) \leq x] \approx \Phi(x)^{\theta_k g}, \quad \theta_k = 2\Phi\left(\sqrt{\frac{3\log(g)\Delta^2}{4h^2d^{2k}}}\right) - 1.$$

Row-wise solution

- The way to use this in practice is

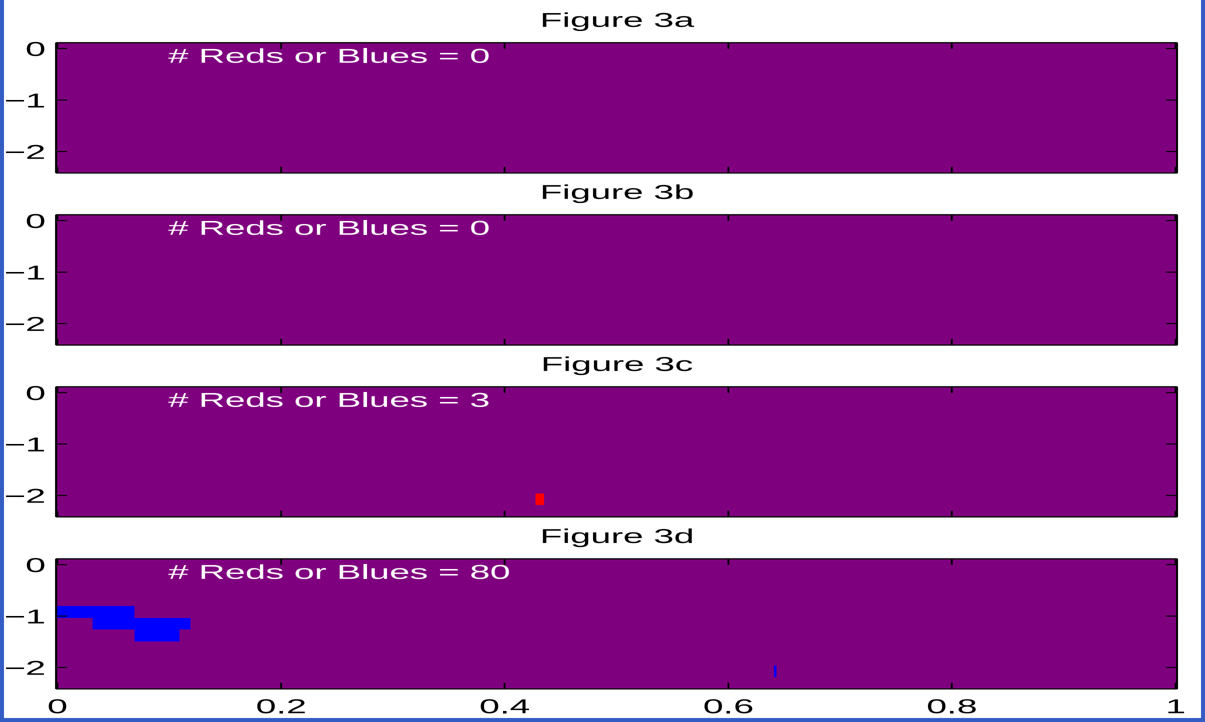
$$P[\max\left(\hat{T}_{1,g},...,\hat{T}_{g,g}\right) \leq x] \approx \Phi(x)^{\vartheta g},$$

- Approximate the max of a fixed SiZer row by:

$$P[\max(T_{1,k},...,T_{g,k}) \leq x] \approx \Phi(x)^{\theta_k g}, \quad \theta_k = 2\Phi\left(\sqrt{\frac{3\log(g)\Delta^2}{4h^2 d^{2k}}}\right) - 1.$$

- Define $C_R = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{1/(\theta(b)g)}\right)$ and color the pixel blue if the corresponding $T_i > C_R$ and red if $T_i < -C_R$.

Row-wise solution



SiZer maps for simulated null distributions, based on the new row-wise procedure. Figures 3a, b, c and d are for the 0.5, 0.75, 0.85 and 0.95, respectively, quantiles of the distribution.

Global Solutions

- For partial global solution compare the approximation to the SiZer map to a Gaussian random field with independent rows using Li and Shao (2002) improvement of Slepian’s inequality.

Global Solutions

- For partial global solution compare the approximation to the SiZer map to a Gaussian random field with independent rows using Li and Shao (2002) improvement of Slepian's inequality.
- The difference between distribution function of the maximum of the two Gaussian random fields is asymptotically negligible as $g \to \infty$ and r is fixed.

Global Solutions

- For partial global solution compare the approximation to the SiZer map to a Gaussian random field with independent rows using Li and Shao (2002) improvement of Slepian's inequality.
- The difference between distribution function of the maximum of the two Gaussian random fields is asymptotically negligible as $g \to \infty$ and r is fixed.
- The approximation for the maximum could be calculated using the random field with independent rows, i.e., $P[\max(T_{1,1}, ..., T_{g,r}) \leq x] \approx \Phi(x)^{(\theta_1+\cdots+\theta_r)g}$,

Size problem fixed

- According to our simulations the global procedure has false positive in a little less than 5% of the “no-signal” pictures.

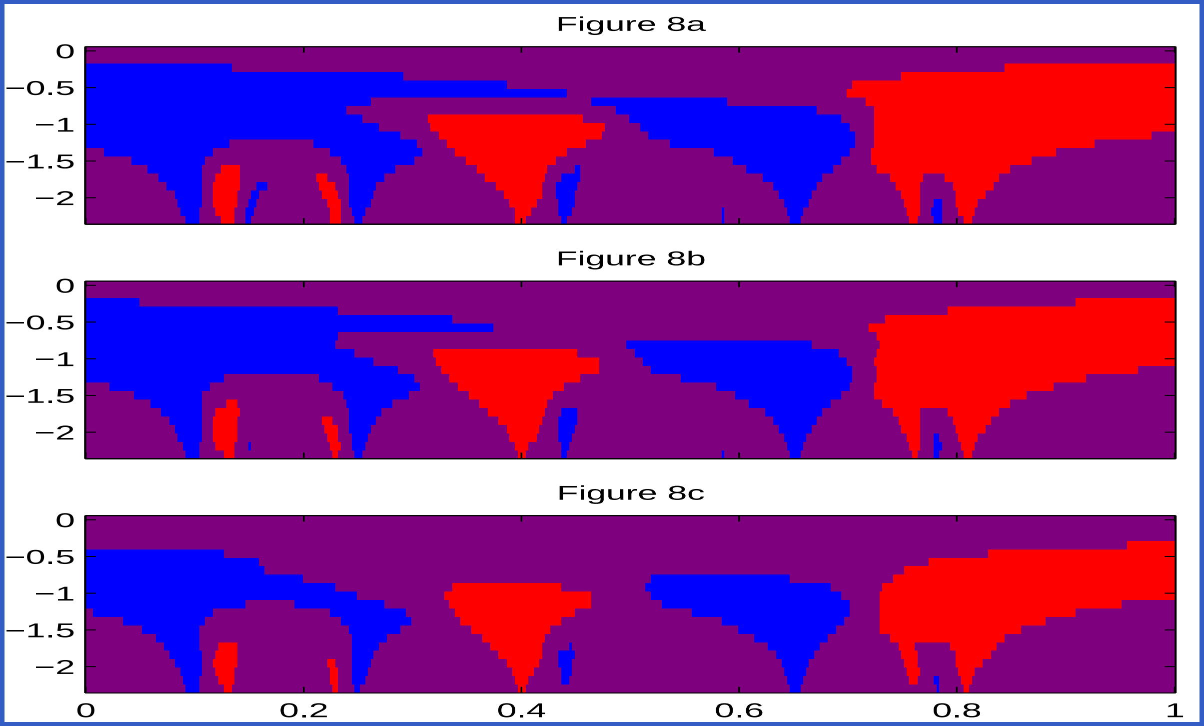
Size problem fixed

- According to our simulations the global procedure has false positive in a little less than 5% of the “no-signal” pictures.
- Row-wise procedure result are shown below however. Roughly 5% of the rows in “no-signal” pictures are entirely purple.

Size problem fixed

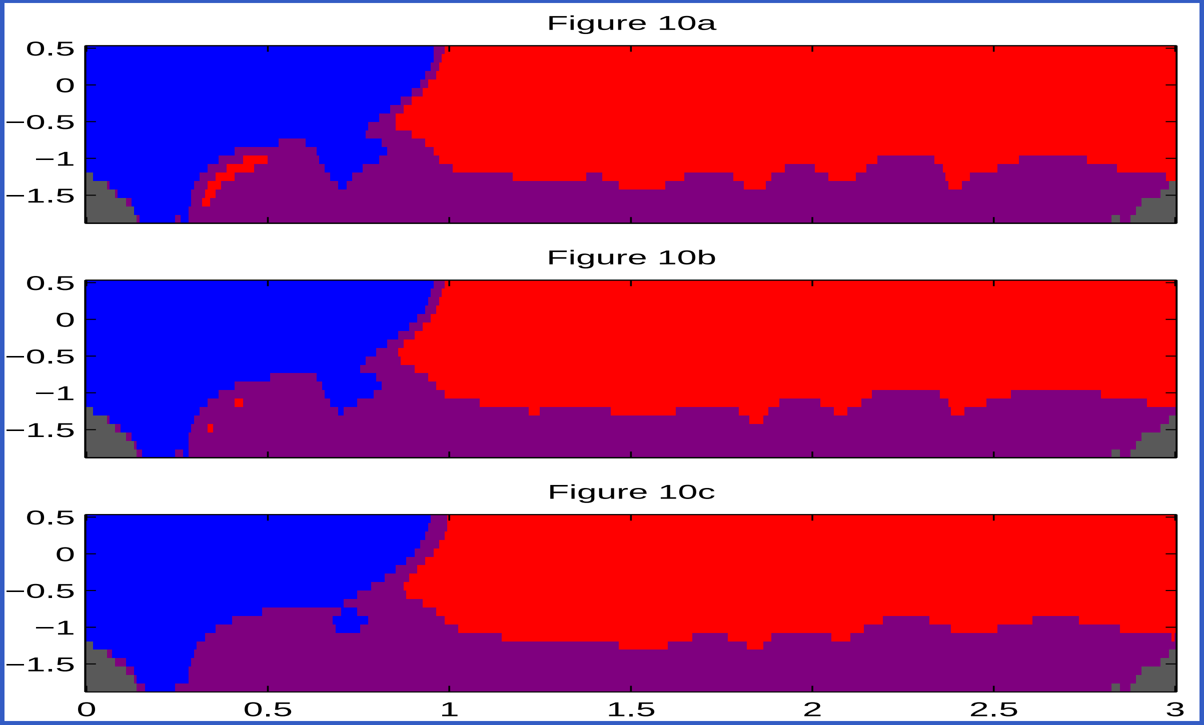
- According to our simulations the global procedure has false positive in a little less than 5% of the “no-signal” pictures.
- Row-wise procedure result are shown below however. Roughly 5% of the rows in “no-signal” pictures are entirely purple.
- Loss in power is not significant in the row-wise procedure. However, the global procedure exhibits some loss of power.

Comments on power



Full range of SiZer analyses of the Donoho - Johnstone Blocks regression, with high noise. Figures 8a, b and c show conventional, row-wise and global SiZer versions.

Comments on power



Full range of SiZer analyses of the British Family Incomes data. Figures 10a, b, c and d show conventional, row-wise and global SiZer versions.

Conclusions

- We proposed a new simultaneous adjustment procedure for SiZer.
- Because of the needed compromise between power and false positive rate we suggest that practitioners use the row-wise procedure.
- Some issues remain to be addressed. In particular there is a problem for small sample size/ small bandwidth caused by the fact that in that case the test statistics have approximately t distribution.
- A second order $g, r \to \infty$ approximation for the global maximum would be desirable.