Bump detection in heterogeneous Gaussian regression

Frank Werner[1,2]

joint with

Farida Enikeeva[3,4], Axel Munk[1,2]

[1] Statistical Inverse Problems in Biophysics group, MPIbpC
[2] University of Göttingen
[3] Université de Poitiers
[4] Russian Academy of Science

AMISTAT 2015 Prague

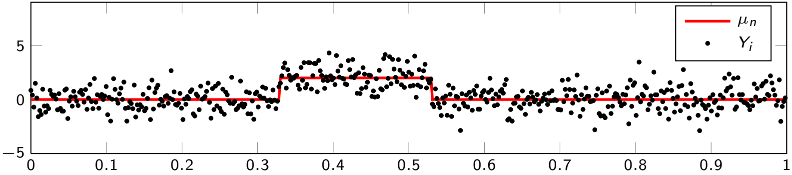
Bump detection in Gaussian regression

Consider a Gaussian regression model, i.e.

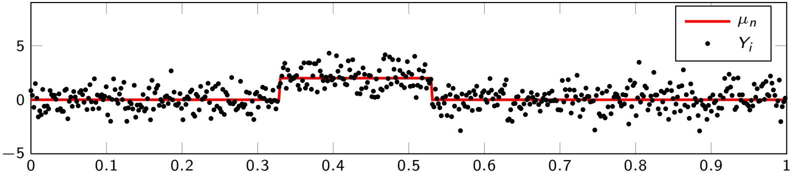
$$Y_i = \mu_n\left(\frac{i}{n}\right) + \sigma_0 Z_i, \qquad 1 \leq i \leq n$$

with i.i.d. Gaussian errors $Z_i \sim \mathcal{N}(0,1)$, $\sigma_0 > 0$ fixed and known.
Suppose the unknown function μ_n is a **bump**:

$$\mu_n(x) = \Delta_n 1_{I_n}(x) = \begin{cases} \Delta_n & \text{if } x \in I_n, \\ 0 & \text{otherwise.} \end{cases}$$



Bump detection in Gaussian regression (cont')



The asymptotic interface between detectable and undetectable signals is characterized by the detection boundary

$$\sqrt{n\,|I_n|}\Delta_n \asymp \sqrt{2}\sigma_0\sqrt{-\log\left(|I_n|\right)}.$$

Bump detection in Gaussian regression (cont')

$$\sqrt{n\,|I_n|}\Delta_n \asymp \sqrt{2}\sigma_0\sqrt{-\log\left(|I_n|\right)}.$$

Mathematical interpretation:

- If μ_n vanishes too fast, i.e.

$$\sqrt{n\,|I_n|}\Delta_n \precsim \left(\sqrt{2}\sigma_0 - \varepsilon_n\right)\sqrt{-\log\left(|I_n|\right)},$$

  then no test with level α can distinguish between μ_n and 0 with power $> \alpha$.

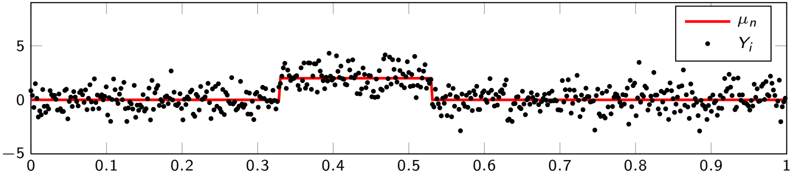
- If μ_n vanishes more slowly, i.e.

$$\sqrt{n\,|I_n|}\Delta_n \succsim \left(\sqrt{2}\sigma_0 + \varepsilon_n\right)\sqrt{-\log\left(|I_n|\right)},$$

  then there is a test with level α which can distinguish between μ_n and 0 with power $> \alpha$.

- (ε_n) is any sequence such that $\varepsilon_n \to 0, \varepsilon_n\sqrt{-\log\left(|I_n|\right)} \to \infty$.

Bump detection - some references



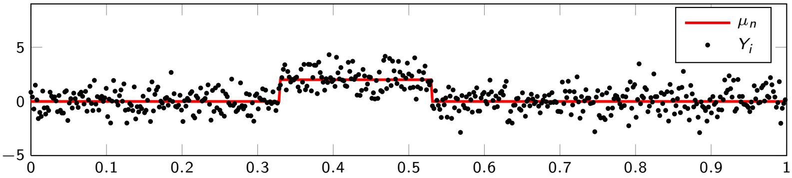
Minimax testing theory: Ingster '93, Tsybakov '09, ...

Detecting bumps and changes: Yao '88, Carlstein, Müller & Siegmund (eds.) '94, Siegmund & Venkatraman '95, Csörgo & Hovráth '97, Bai & Perron '98, Braun, Braun & Müller '00, Birgé & Massart '01, Lavielle '05, Harchaoui & Lévy-Leduc '10, Siegmund, Yakir & Zhang '11, Killick, Fearnhead & Eckley '12, Rigollet & Tsybakov '12, Rivera & Walther '13, Siegmund '13, Frick, Munk & Sieling '14, Du, Kao & Kou '15, ...

Minimax testing in bump detection: Dümbgen & Spokoiny 2001, Dümbgen & Walther '08, Jeng, Cai & Li '10, Chan & Walther '11, Korostelev & Korosteleva '11, Frick, Munk & Sieling 2014, ...

Heterogeneous bump detection

$$Y_i = \mu_n\left(\frac{i}{n}\right) + \sigma_0 Z_i, \qquad 1 \leq i \leq n$$

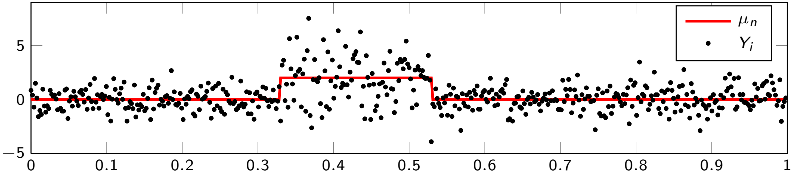


- variance function λ_n^2 is a bump function as well with the same "support" I_n:

$$\lambda_n^2(x) = \sigma_0^2\left(1 + \kappa_n^2 1_{I_n}(x)\right), \qquad x \in [0,1]$$

- if $\kappa_n^2 > 0$ this adds information to the model ...
- ... if $\kappa_n^2 = 0$ is possible, we loose information (variance as nuisance parameter)

Heterogeneous bump detection - applications and references



Applications: CGH array analysis (Muggeo & Adelfio '10), ion channel recordings with open channel noise (Sigworth '85, Schirmer '98), Econometrics (Bai & Perron '03), ...

Tests with variance as nuisance parameter: Huang & Chang '93, Venkatraman & Olshen '07, Muggeo & Adelfio '10, Arlot & Celisse '11, Boutahar '12, Pein, Munk & Sieling '15, ...

Identification in mixtures: Donoho & Jin '04, Cai, Jeng & Jin '11, Arias-Castro & Wang '13, Cai & Wu '14, ...

Minimax testing for $\kappa_n > 0$: this talk!

The setup

$$Y_i = \Delta_n 1_{I_n}\left(\frac{i}{n}\right) + \sigma_0 \sqrt{\left(1 + \kappa_n^2 1_{I_n}\left(\frac{i}{n}\right)\right)} Z_i, \qquad 1 \leq i \leq n$$

with $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

parameters: $\sigma_0 > 0$ (fixed and known), $\kappa_n \searrow 0$ (known), $|I_n| \searrow 0$ (known), $\Delta_n > 0$ (known, adaptation will be discussed)

TODO: provide lower detection bounds (no test can distinguish between zero signal and non-zero signal)

TODO: provide upper detection bounds (there is a test which can distinguish)

notation: (ε_n) is any sequence such that

$$\varepsilon_n \to 0, \qquad \varepsilon_n \min\left\{\kappa_n^{-2}, \sqrt{-\log(|I_n|)}\right\} \to \infty.$$

General lower detection bound

Theorem

No test can distinguish between the zero signal and non-zero signals with (asymptotic) level $\leq \alpha$ and (asymptotic) power $> \alpha$, if there exists a sequence $\delta_n \searrow 0$, such that for $n \to \infty$

$$\delta_n \left(\frac{n\,|I_n|\,\Delta_n^2}{2\sigma_0^2} + n\,|I_n|\,\frac{\kappa_n^4}{4} + \log\left(|I_n|\right) \right) + \delta_n^2 \left(\frac{n\,|I_n|\,\Delta_n^2}{2\sigma_0^2} + n\,|I_n|\,\frac{\kappa_n^4}{4} \right) \to -\infty$$

Proof: Techniques from Dümbgen & Spokoiny '01 generalized to non-central chi-squared likelihood ratios, Taylor expansion using $\kappa_n \searrow 0$.

General upper detection bound

Theorem

The likelihood ratio test can distinguish between the zero signal and non-zero signals with (asymptotic) level $\leq \alpha$ and (asymptotic) power $\geq 1-\alpha$, if for $n \to \infty$

$$\begin{aligned} & n\,|I_n| \left(\kappa_n^4 + 2\frac{\Delta_n^2}{\sigma_0^2} \right) + \kappa_n^2 \frac{\Delta_n^2 n\,|I_n|}{\sigma_0^2} \\ \geq \quad & 2\kappa_n^2 \log\left(\frac{1}{|I_n|}\right) + 2\kappa_n^2 \log\left(\frac{1}{\alpha}\right) + 2\sqrt{n\,|I_n| \left(\kappa_n^4 + 2\frac{\Delta_n^2}{\sigma_0^2}\right) \log\left(\frac{1}{\alpha\,|I_n|}\right)} \\ & + 2\left(1+\kappa_n^2\right) \sqrt{n\,|I_n| \left(\kappa_n^4 + 2\left(1+\kappa_n^2\right)\frac{\Delta_n^2}{\sigma_0^2}\right) \log\left(\frac{1}{\alpha}\right)}. \end{aligned}$$

Proof: Union bound, new chi-squared deviation inequality and straight forward analysis.

Regimes and phase transitions

$$\delta_n \left(\frac{n\,|I_n|\,\Delta_n^2}{2\sigma_0^2} + n\,|I_n|\,\frac{\kappa_n^4}{4} + \log\left(|I_n|\right) \right) + \delta_n^2 \left(\frac{n\,|I_n|\,\Delta_n^2}{2\sigma_0^2} + n\,|I_n|\,\frac{\kappa_n^4}{4} \right) \to -\infty$$

- Variance vanishes faster than signal
  $\rightsquigarrow$ **dominant signal regime (DSR)**: $\frac{\kappa_n^2}{\Delta_n} \to 0$
- Variance and signal vanish at the same rate
  $\rightsquigarrow$ **equilibrium regime (ER)**: $\frac{\kappa_n^2}{\Delta_n} \to \mathrm{const}$
- Signal vanishes faster than variance
  $\rightsquigarrow$ **dominant variance regime (DVR)**: $\frac{\kappa_n^2}{\Delta_n} \to \infty$

Dominant signal regime

$$\text{DSR: } \frac{\kappa_n^2}{\Delta_n} \to 0$$

Lower detection bound

No test can distinguish if

$$\sqrt{n\,|I_n|}\Delta_n \precsim \left(\sqrt{2}\sigma_0 - \varepsilon_n\right)\sqrt{-\log\left(|I_n|\right)}$$

Upper detection bound

The likelihood ratio test can distinguish if

$$\sqrt{n\,|I_n|}\Delta_n \succsim \left(\sqrt{2}\sigma_0 + \varepsilon_n\right)\sqrt{-\log\left(|I_n|\right)}$$

Equilibrium regime

$$\text{ER: } \frac{\kappa_n^2}{\Delta_n} \to \frac{c}{\sigma_0} \in (0, \infty)$$

Lower detection bound

No test can distinguish if

$$\sqrt{n\,|I_n|}\Delta_n \precsim (C - \varepsilon_n)\sqrt{-\log(|I_n|)}, \qquad C := \sqrt{2}\sigma_0\sqrt{\frac{2}{2+c^2}}$$

Upper detection bound

The likelihood ratio test can distinguish if

$$\sqrt{n\,|I_n|}\Delta_n \succsim (C + \varepsilon_n)\sqrt{-\log(|I_n|)}, \qquad C := \sqrt{2}\sigma_0\sqrt{\frac{2}{2+c^2}}$$

Equilibrium regime (alternative formulation)

$$\text{ER: } \frac{\kappa_n^2}{\Delta_n} \to \frac{c}{\sigma_0} \in (0, \infty)$$

Lower detection bound

No test can distinguish if

$$\sqrt{n\,|I_n|}\kappa_n^2 \lesssim (C - \varepsilon_n)\sqrt{-\log(|I_n|)}, \qquad C := 2\sqrt{\frac{c^2}{2 + c^2}}$$

Upper detection bound

The likelihood ratio test can distinguish if

$$\sqrt{n\,|I_n|}\kappa_n^2 \gtrsim (C + \varepsilon_n)\sqrt{-\log(|I_n|)}, \qquad C := 2\sqrt{\frac{c^2}{2 + c^2}}$$

Dominant variance regime

$$\text{DVR: } \frac{\kappa_n^2}{\Delta_n} \to \infty$$

Lower detection bound

No test can distinguish if

$$\sqrt{n\,|I_n|}\kappa_n^2 \precsim (2-\varepsilon_n)\sqrt{-\log(|I_n|)}$$

Upper detection bound

The likelihood ratio test can distinguish if

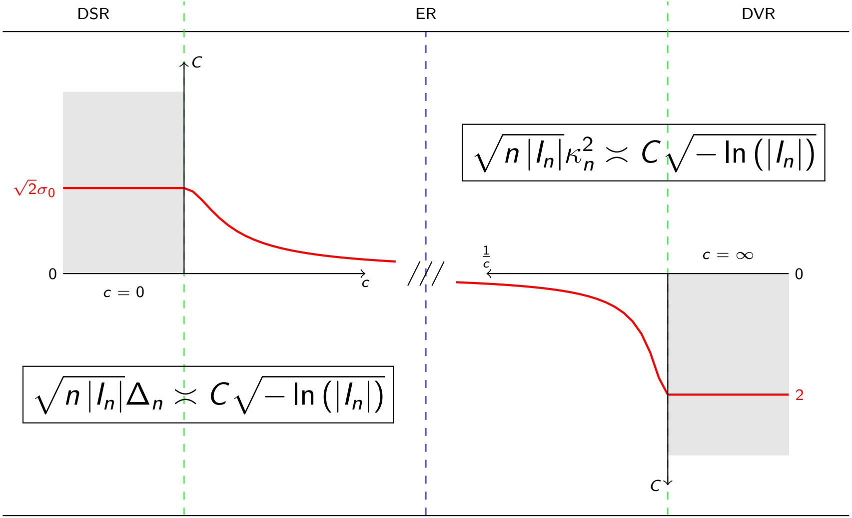
$$\sqrt{n\,|I_n|}\kappa_n^2 \succsim (2+\varepsilon_n)\sqrt{-\log(|I_n|)}$$

Overview

	rate	constant	
		lower bound	upper bound
DSR	$\sqrt{n\,\lvert I_n\rvert}\Delta_n \sim \sqrt{-\log\left(\lvert I_n\rvert\right)}$	$\sqrt{2}\sigma_0 - \varepsilon_n$	$\sqrt{2}\sigma_0 + \varepsilon_n$
ER	$\sqrt{n\,\lvert I_n\rvert}\Delta_n \sim \sqrt{-\log\left(\lvert I_n\rvert\right)}$	$\sqrt{2}\sigma_0\sqrt{\frac{2}{2+c^2}} - \varepsilon_n$	$\sqrt{2}\sigma_0\sqrt{\frac{2}{2+c^2}} + \varepsilon_n$
	$\sqrt{n\,\lvert I_n\rvert}\kappa_n^2 \sim \sqrt{-\log\left(\lvert I_n\rvert\right)}$	$2\sqrt{\frac{c^2}{2+c^2}} - \varepsilon_n$	$2\sqrt{\frac{c^2}{2+c^2}} + \varepsilon_n$
DVR	$\sqrt{n\,\lvert I_n\rvert}\kappa_n^2 \sim \sqrt{-\log\left(\lvert I_n\rvert\right)}$	$2 - \varepsilon_n$	$2 + \varepsilon_n$

The detection boundary

$$c := \lim_{n\to\infty} \sigma_0 \frac{\kappa_n^2}{\Delta_n} \in [0,\infty]$$



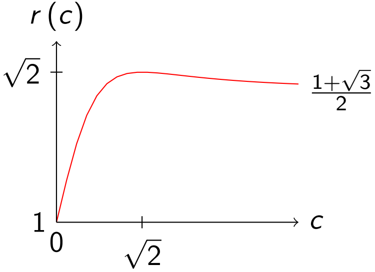
Adaptation: Δ_n unknown

- Lower bounds stay valid, but optimality of those is unclear
- Upper bounds: consider adaptive test, replace Δ_n by $(n\,|I_n|)^{-1}\sum_{i:i/n\in I_n} Y_i$.

Theorem

The adaptive likelihood ratio test can distinguish at the same rate but with possibly different constant. The ratio r of adaptive and non-adaptive constants yields the price for adaptation.

$$r(c) = \begin{cases} 1 & \text{DSR}, c = 0, \\ \frac{\sqrt{2+c^2}\left(c+\sqrt{2+3c^2}\right)}{2(1+c^2)} & \text{ER}, 0 < c < \infty, \\ \frac{1+\sqrt{3}}{2} & \text{DVR}, c = \infty, \end{cases}$$



Extensions

- $\rightarrow$ $\kappa_n \not\searrow 0$**:** Lower bounds available, but the constants involve logarithms of $\kappa := \lim_{n\to\infty} \kappa_n$. Upper bounds seem not sharp, as they do not involve logarithms of κ. Better chi-squared deviation bounds are necessary!
- $\rightarrow$ **adaptive upper bounds for unknown σ_0 or / and κ_n:** requires deviation bounds for fourth powers of Gaussians!
- $\rightarrow$ **adaptive upper bounds for unknown $|I_n|$:** requires structurally different tests!
- $\rightarrow$ **adaptive lower bounds in all cases:** are unclear so far!
- $\rightarrow$ **multiple bumps:** Lower and upper bounds are also interesting in that case!
- $\rightarrow$ **different model:** If we allow for $\kappa_n = 0$, does this really cause loss of information? What is the detection boundary in that case?

Conclusion

- Bump detection in Gaussian regression:
  - detection boundary in the homogeneous case well-known and investigated
  - in the heterogeneous case, we can derive it under certain restrictions
- improved detection power given the variance jumps as well
- adaptation to Δ_n has a cost, opposed to the homogeneous situation

F. Enikeeva, A. Munk and F. Werner
Bump detection in heterogeneous Gaussian regression.
Submitted, *arXiv:* 1504.0739.

Thank you for your attention!