The Power of Tests for Signal Detection in High Dimensional Data

Marc Ditzhaus

joint work with Arnold Janssen

Mathematical institute

Heinrich-Heine-University Düsseldorf

Wellington, December 2017

Motivation

genome analysis $\rightarrowtail$ early detection of common deceases

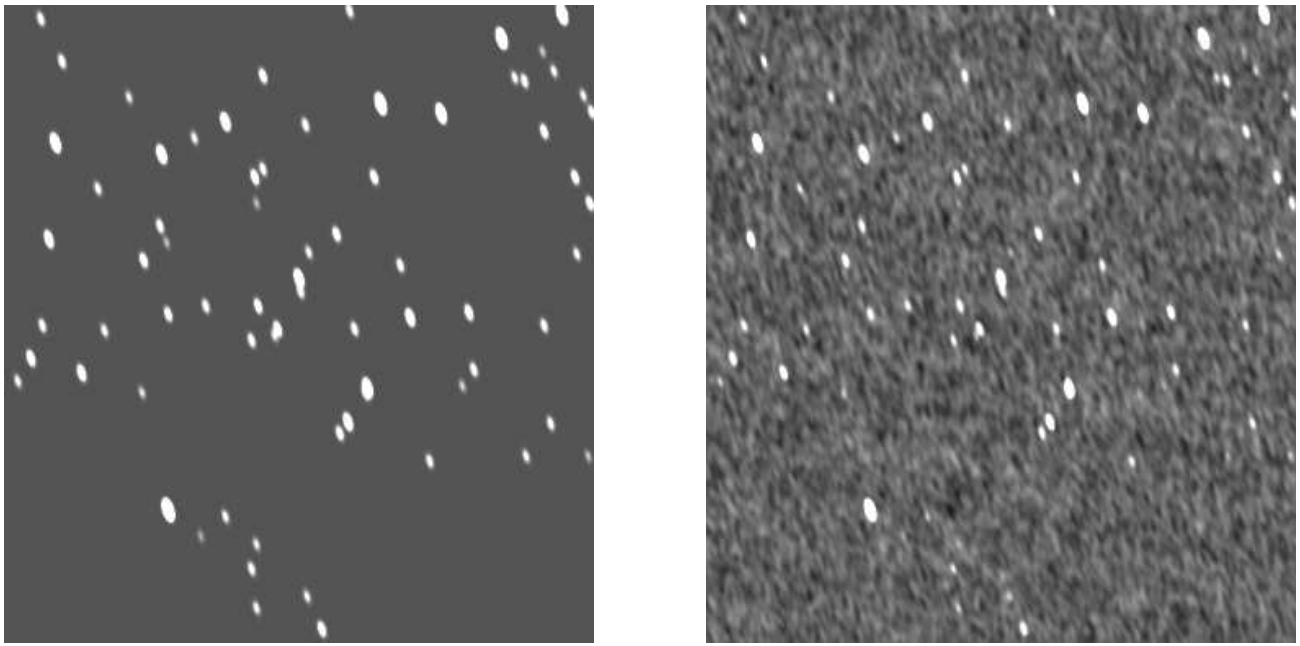
- high dimensional observation vector $\mathbf{X_n} := (\mathbf{X_{n,1}}, \ldots, \mathbf{X_{n,n}})$ for **each** patient
- healthy patient: $\mathbf{X}_n$ behaves some noisy background
- ill patient: $\mathbf{X}_n$ contains rare and weak signals

Motivation

genome analysis $\rightarrowtail$ early detection of common deceases

- high dimensional observation vector $\mathbf{X_n} := (\mathbf{X_{n,1}}, \ldots, \mathbf{X_{n,n}})$ for **each** patient
- healthy patient: $\mathbf{X}_n$ behaves some noisy background
- ill patient: $\mathbf{X}_n$ contains rare and weak signals
- Our interest: The asymptotic power of tests which decide whether there are any signals
- Other applications: astronomy, cosmology, disease surveillance,...

Astrophysics



Source: Hopkins et al. (2002)

Model

- Null: we observe

$$X_{n,i} = Z_{n,i} \sim P_{n,i},\ i = 1, \ldots, n \quad \text{(known noisy background)}.$$

Model

- Null: we observe

$$X_{n,i} = Z_{n,i} \sim P_{n,i},\ i = 1, \ldots, n \quad \text{(known noisy background)}.$$

- Alternative: we observe

$$X_{n,i} = \begin{cases} Y_{n,i} \sim \mu_{n,i} & \text{if } B_{n,i} = 1 \quad \text{(unknown signal)}. \\ Z_{n,i} & \text{if } B_{n,i} = 0 \quad \text{(noisy background)}, \end{cases}$$

where $\varepsilon_{n,i} = P(B_{n,i} = 1)$ is unknown.

Model

- Null: we observe

$$X_{n,i} = Z_{n,i} \sim P_{n,i},\ i = 1, \ldots, n \quad \text{(known noisy background)}.$$

- Alternative: we observe

$$X_{n,i} = \begin{cases} Y_{n,i} \sim \mu_{n,i} & \text{if } B_{n,i} = 1 \quad \text{(unknown signal)}. \\ Z_{n,i} & \text{if } B_{n,i} = 0 \quad \text{(noisy background)}, \end{cases}$$

  where $\varepsilon_{n,i} = P(B_{n,i} = 1)$ is unknown.

- In short: $X_{n,i} \sim (1 - \varepsilon_{n,i})P_{n,i} + \varepsilon_{n,i}\mu_{n,i} = Q_{n,i}$.
- Independence assumption!

The detection boundary: heterogeneous normal

$$\mathcal{H}_{0,n} : X_{n,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \qquad 1 \le i \le n$$

$$\mathcal{H}_{1,n} : X_{n,i} \overset{i.i.d.}{\sim} (1-\varepsilon_n)\mathcal{N}(0,1) + \varepsilon_n \mathcal{N}(\vartheta_n, 1), \qquad 1 \le i \le n.$$

$\varepsilon_n = n^{-\beta}$

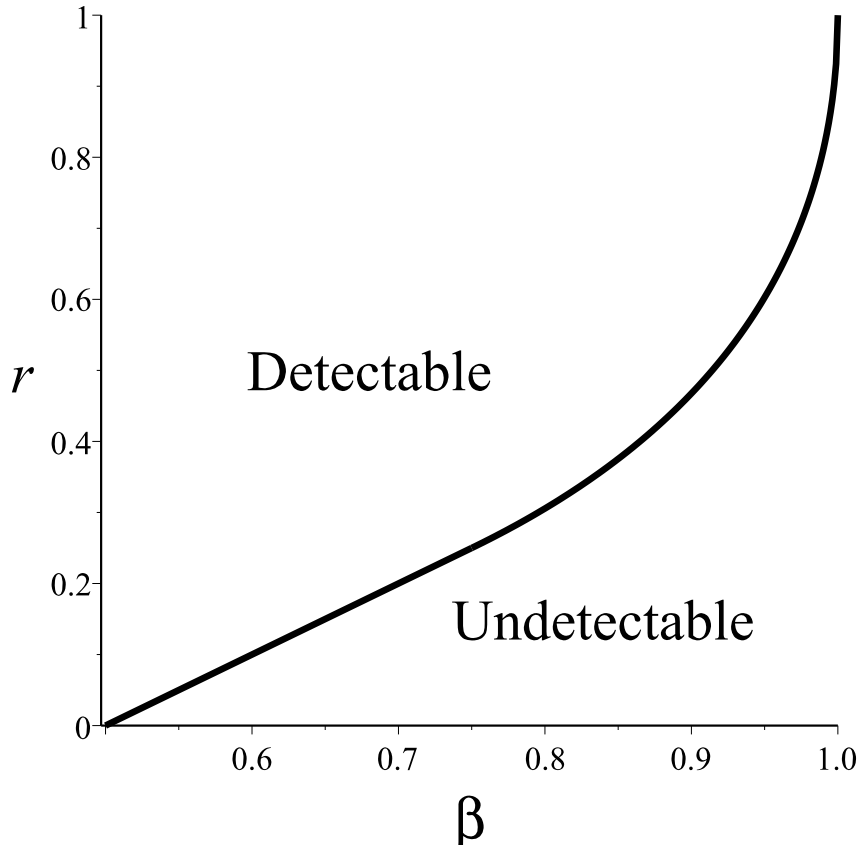
(signal probability)

$\vartheta_n = \sqrt{2r \log n}$

(signal strength)

see Ingster (1997),

Donoho & Jin (2004)



The detection boundary: heteroscedastic normal

$$\mathcal{H}_{0,n} : X_{n,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \qquad 1 \le i \le n$$

$$\mathcal{H}_{1,n} : X_{n,i} \overset{i.i.d.}{\sim} (1-\varepsilon_n)\mathcal{N}(0,1) + \varepsilon_n \mathcal{N}(\vartheta_n, \tau^2), \qquad 1 \le i \le n.$$

solid (—):

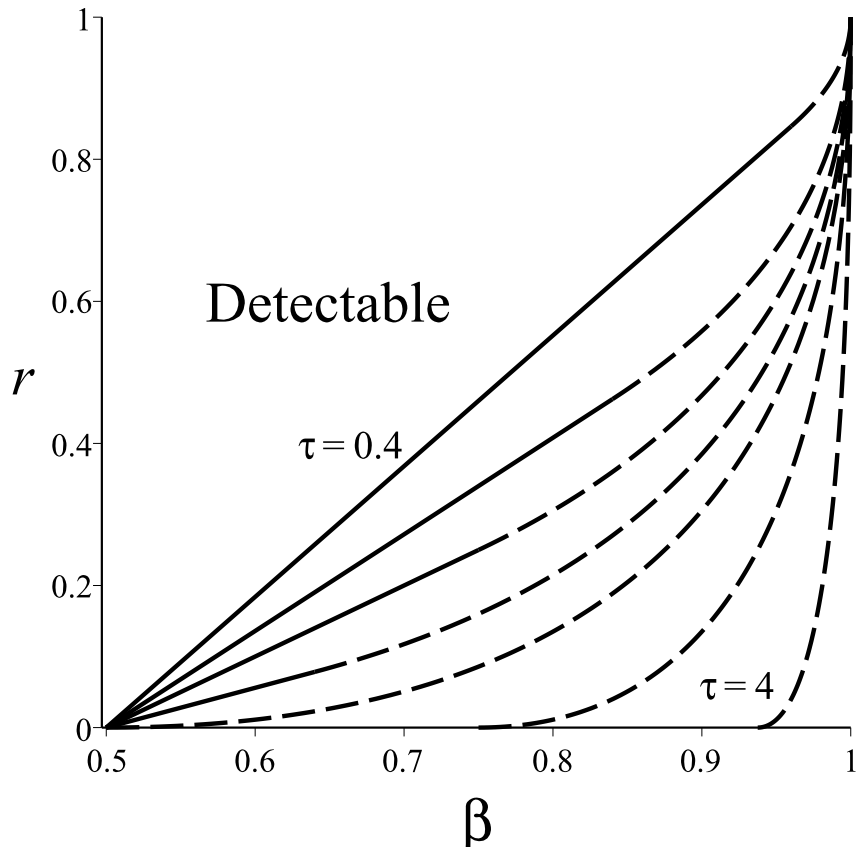
Gaussian limit

dashed (- - -):

non-Gaussian but

infinitely divisible

Cai, Jeng, Jin (2011)



- Problem in practice: the signal probability $\varepsilon_{n,i}$ and the signal distribution $\mu_{n,i}$ are unknown

  $\Rightarrow$ LLR cannot be used.

- Problem in practice: the signal probability $\varepsilon_{n,i}$ and the signal distribution $\mu_{n,i}$ are unknown

  $\Rightarrow$ LLR cannot be used.

- But *Tukey's higher criticism* (HC) is applicable and has the same detection area:
  - normal mixtures: Donoho and Jin (2004), Cai, Jeng, Jin (2011)
  - Poisson mixtures: Arias-Castro and Wang, M. (2015)
  - general class of exponential families: Cai and Wu (2014)
  - ...

Our aims

- Detectable and undetectable areas of LLRT and HC (in the literature: various results)
- LLRT on the boundary (few results, only normal distributions)
- HC on the boundary (no results until now)

General conditions

- triangular scheme $(\varepsilon_{n,i})_{n\in\mathbb{N}}$ in $[0,1]$ with $\max_{1\leq i\leq n}\varepsilon_{n,i} = \varepsilon_{n:n} \to 0$
- probability measures $\mu_{n,i} \ll P_{n,i}$ (Without loss of generality)

General testing problem

$$\mathcal{H}_{0,n}: \; P_{(n)} := \bigotimes_{i=1}^{n} P_{n,i} \quad \text{against}$$

$$\mathcal{H}_{1,n}: Q_{(n)} := \bigotimes_{i=1}^{n} Q_{n,i} \quad \text{with } Q_{n,i} := (1-\varepsilon_{n,i})P_{n,i} + \varepsilon_{n,i}\mu_{n,i}.$$

$$LLR_n = \log\left(\frac{\mathrm{d}Q_{(n)}}{\mathrm{d}P_{(n)}}(X_{n,1},\ldots,X_{n,n})\right)$$

$$LLR_n = \log\left(\frac{\mathrm{d}Q_{(n)}}{\mathrm{d}P_{(n)}}(X_{n,1},\ldots,X_{n,n})\right) \xrightarrow{\mathrm{d}} \begin{cases} \xi_1 \in \mathbb{R}\cup\{-\infty\} \text{ under } \mathcal{H}_0 \\ \xi_2 \in \mathbb{R}\cup\{\infty\} \text{ under } \mathcal{H}_1 \end{cases}$$

$$LLR_n = \log\left(\frac{\mathrm{d}Q_{(n)}}{\mathrm{d}P_{(n)}}(X_{n,1},\ldots,X_{n,n})\right) \xrightarrow{\mathrm{d}} \begin{cases} \xi_1 \in \mathbb{R}\cup\{-\infty\} \text{ under } \mathcal{H}_0 \\ \xi_2 \in \mathbb{R}\cup\{\infty\} \text{ under } \mathcal{H}_1 \end{cases}$$

Undetectable	$\xi_1 \equiv 0 \equiv \xi_2$
Detectable	$\xi_1 \equiv -\infty$, $\xi_2 \equiv \infty$
On the boundary (until now)	ξ_1, ξ_2 infinitely divisible on $\mathbb{R}$

Our tool

All three cases are uniquely determined by the two sums

$$I_{1,n}(x) = \sum_{i=1}^{n} \varepsilon_{n,i}\mu_{n,i}\left(\varepsilon_{n,i}\frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}} > x\right),$$

$$I_{2,n}(x) = \sum_{i=1}^{n} \varepsilon_{n,i}^2 \, E_{P_{n,i}}\left(\left(\frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}}\right)^2 \mathbf{1}\left\{\varepsilon_{n,i}\frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}} \leq x\right\} - 1\right).$$

Our tool

All three cases are uniquely determined by the two sums

$$I_{1,n}(x) = \sum_{i=1}^{n} \varepsilon_{n,i} E_{P_{n,i}} \left(\frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}} \mathbf{1} \left\{ \varepsilon_{n,i} \frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}} > x \right\} \right),$$

$$I_{2,n}(x) = \sum_{i=1}^{n} \varepsilon_{n,i}^2 \, E_{P_{n,i}} \left(\left(\frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}} \right)^2 \mathbf{1} \left\{ \varepsilon_{n,i} \frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}} \leq x \right\} - 1 \right).$$

Theorem (D & Janssen 2017)

(a) *Either* $\xi_1 \in \mathbb{R}$ *or* $\xi_1 \equiv -\infty$ *with probability one.*

(b) *Suppose* $\xi_1 \in \mathbb{R}$ *with probability one. Then we have:*

(i) $\xi_1 \sim \nu_1$ *is infinitely divisible.*

(ii) ξ_2 *is infinitely divisible on* $((-\infty, \infty], +)$*, i.e.*

$$\xi_2 \sim (1 - a)\, \epsilon_\infty + a\, \nu_2, \;\; a \in (0, 1],$$

where ν_2 *is an infinitely divisible probability measure on* $(\mathbb{R}, \mathcal{B})$ *and* $a = P(\xi_2 \in \mathbb{R})$*.*

Spike chimeric alternatives

Let $h : (0,1) \to [0,\infty)$ be measurable with

$$\int_0^1 h \, \mathrm{d}\lambda = 1 \text{ and } \int_0^1 h^2 \, \mathrm{d}\lambda \in (0,\infty).$$

Let $P_{n,i} = \lambda_{|(0,1)}$ and $\mu_{n,i}$ be defined by

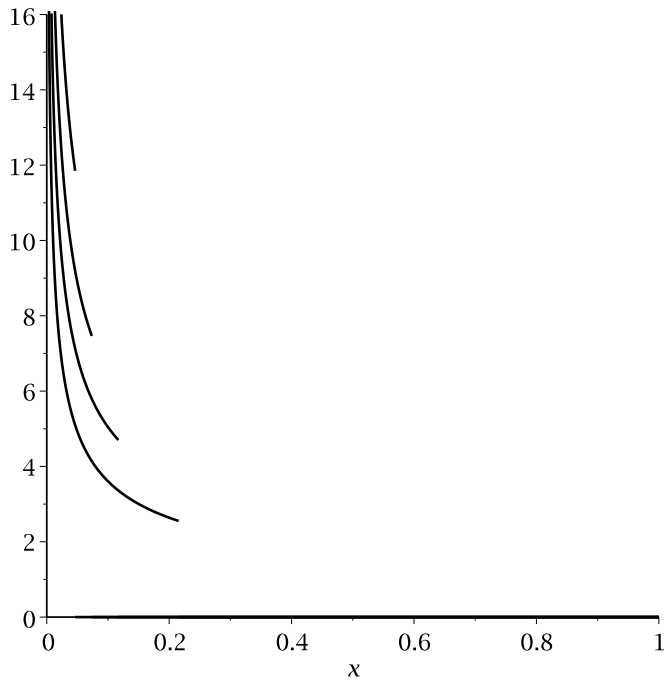
$$\frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}}(u) := \begin{cases} 0 & \text{if } u \in [\tau_{n,i}, 1) \\ \frac{1}{\tau_{n,i}} h\left(\frac{u}{\tau_{n,i}}\right) & \text{if } u \in (0, \tau_{n,i}) \end{cases},$$

where

$$\tau_{n,i} \in (0,1) \text{ and } \max_{1 \le i \le n} \tau_{n,i} \to 0 \text{ as } n \to \infty.$$

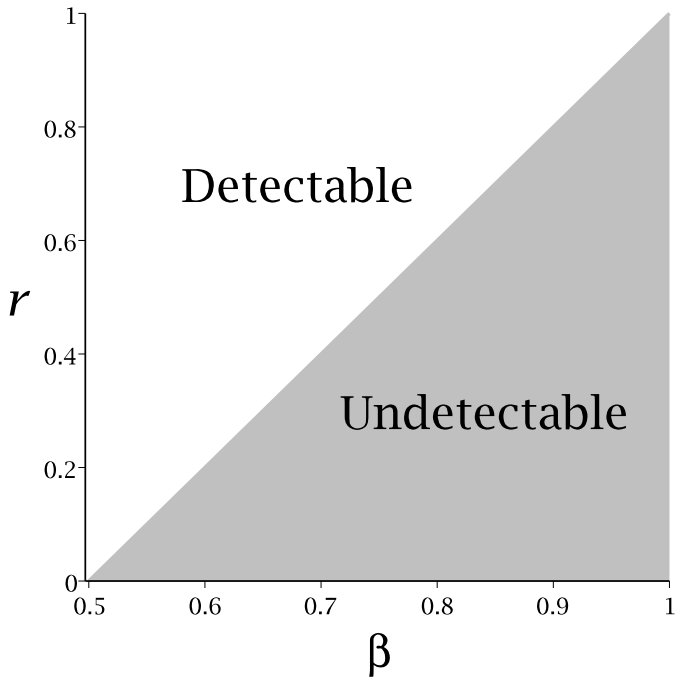
Literature: Khmaladze (1998).

Illustration of $x \mapsto \frac{\mathrm{d}\mu_{n,i}}{\mathrm{d}P_{n,i}(x)}$



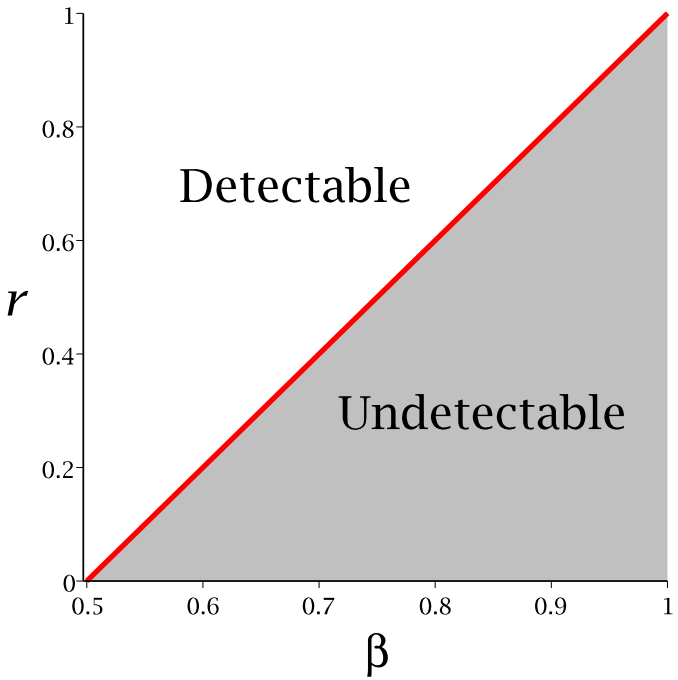
Nonparametric detection boundary

Let $\varepsilon_{n,i} = n^{-\beta}$ and $\tau_{n,i} = n^{-r}$ for $\beta \in (\frac{1}{2}, 1]$ and $r \in (0, 1]$.



Nonparametric detection boundary

Let $\varepsilon_{n,i} = n^{-\beta}$ and $\tau_{n,i} = n^{-r}$ for $\beta \in (\frac{1}{2}, 1]$ and $r \in (0, 1]$.

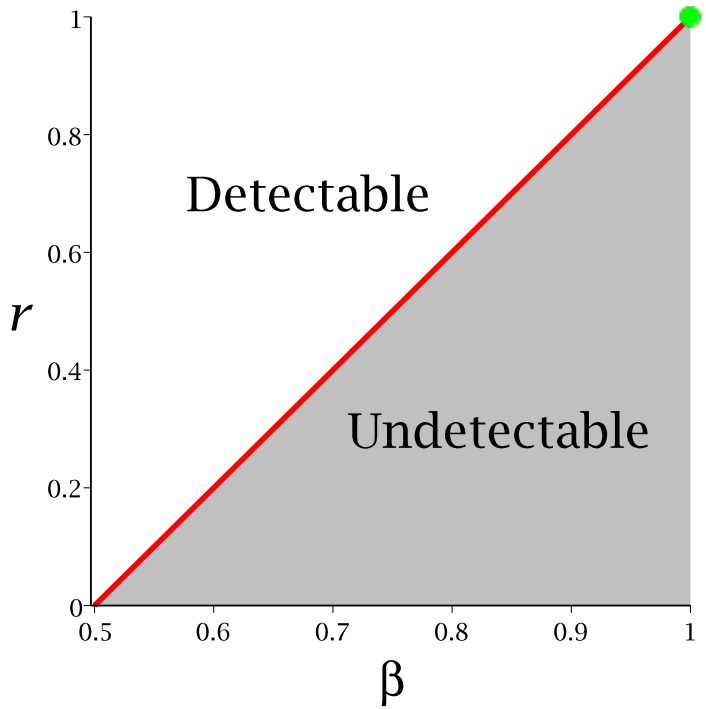


non-trivial power

— ξ_j is normal (only depend on $\int_0^1 h^2 \, \mathrm{d}\lambda$).

Nonparametric detection boundary

Let $\varepsilon_{n,i} = n^{-\beta}$ and $\tau_{n,i} = n^{-r}$ for $\beta \in (\frac{1}{2}, 1]$ and $r \in (0, 1]$.

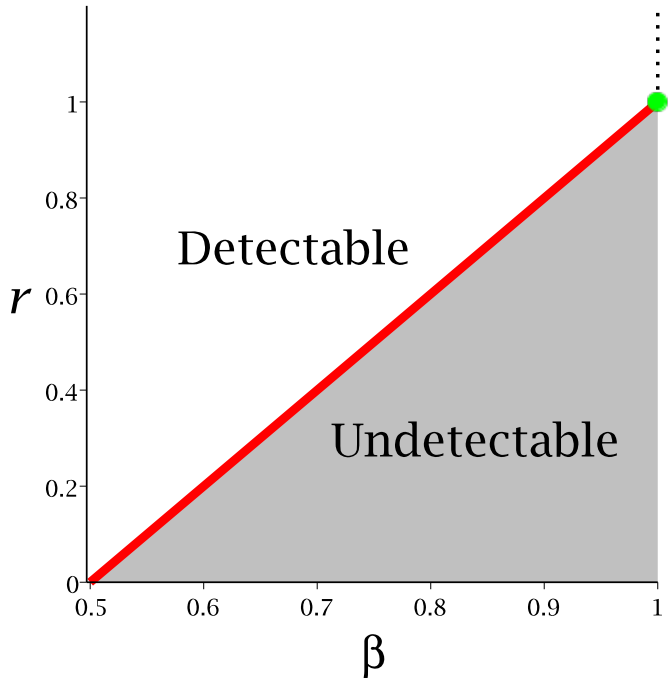


non-trivial power

— ξ_j is normal (only depend on $\int_0^1 h^2 \, \mathrm{d}\lambda$).

- ξ_j has non-trivial Lévy measure

Extended detection boundary



⋮ $\xi_1 \sim \epsilon_{-1}$ and $\xi_2 \sim e^{-1}\epsilon_{-1} + (1 - e^{-1})\epsilon_\infty$

Extended detection boundary

Figure: heteroscedastic normal

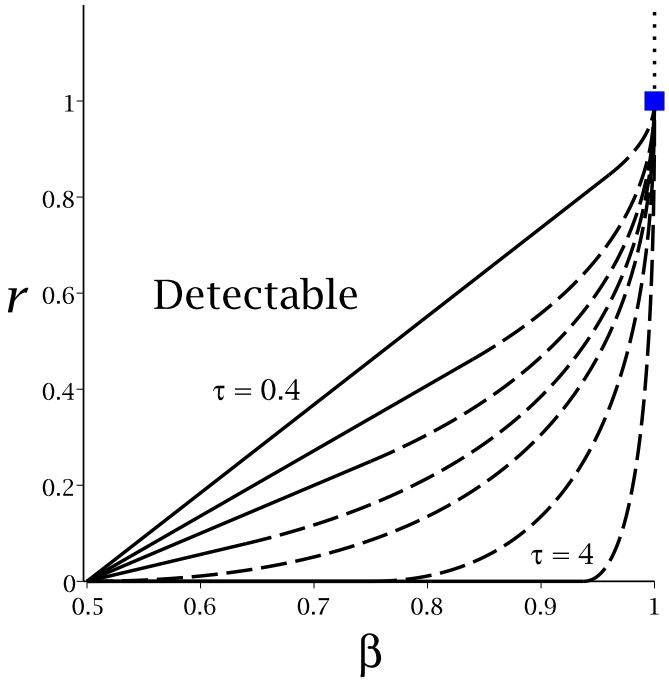
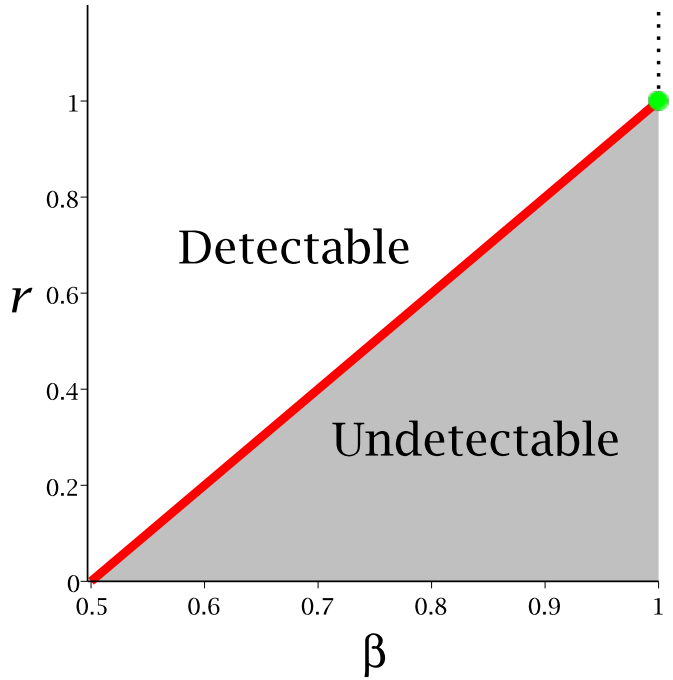
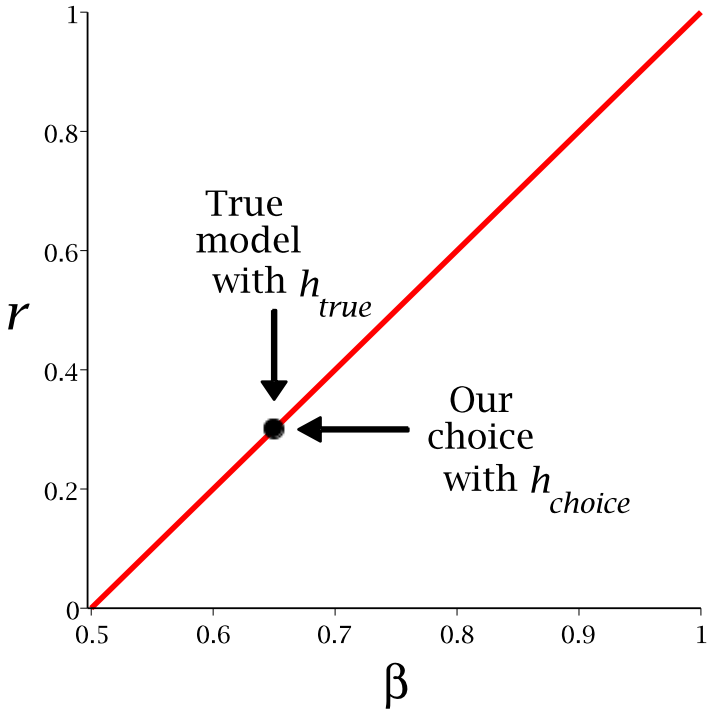


Figure: chimeric alternatives



- $\xi_1 \sim \epsilon_{-1}$ and $\xi_2 \sim e^{-1}\epsilon_{-1} + (1 - e^{-1})\epsilon_\infty$
- $\xi_1 \sim \epsilon_{-\frac{1}{2}}$ and $\xi_2 \sim e^{-\frac{1}{2}}\epsilon_{-\frac{1}{2}} + (1 - e^{-\frac{1}{2}})\epsilon_\infty$

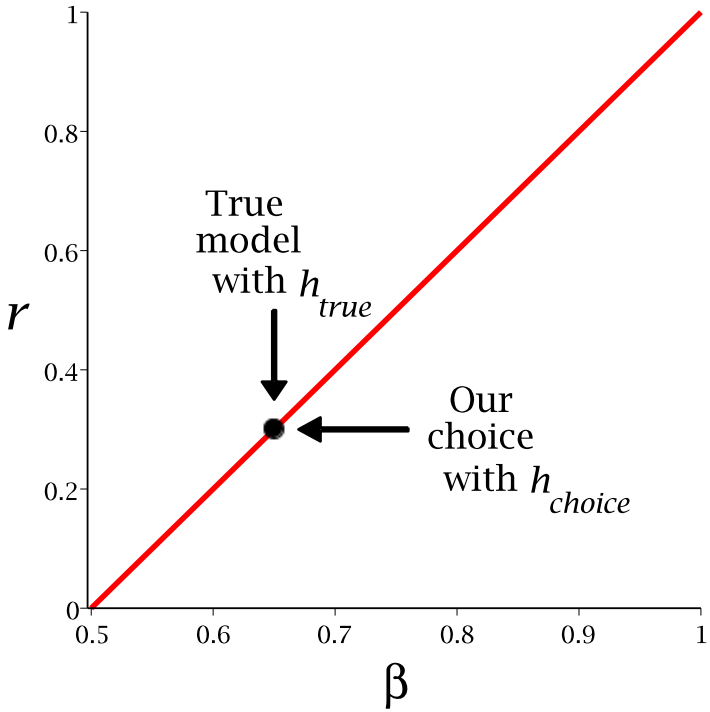
Pitman's asymptotic relative efficiency

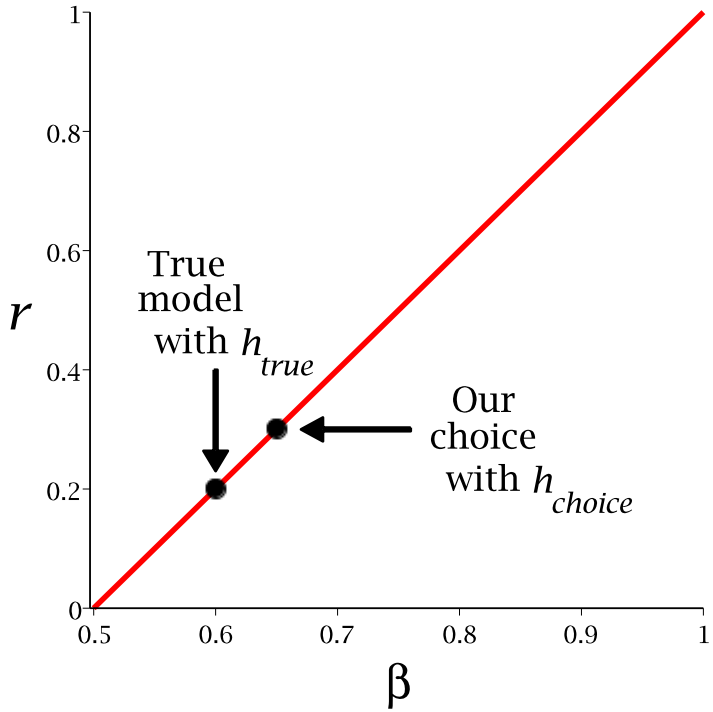


$$E_{Q_{(n),\text{true}}}(\varphi_{n,\text{choice}}) \to \Phi\left(u_\alpha + \sqrt{< h_t, h_t > \text{ARE}}\right),$$

$$\text{where ARE} = \frac{< h_t, h_c >^2}{< h_t, h_t >< h_t, h_c >}$$

Pitman's asymptotic relative efficiency





$$E_{Q_{(n),\text{true}}}(\varphi_{n,\text{choice}}) \to \Phi\left(u_\alpha + \sqrt{< h_t, h_t > \text{ARE}}\right),$$

$$\text{where ARE} = \frac{< h_t, h_c >^2}{< h_t, h_t >< h_t, h_c >} \mathbf{1}\{\beta_t = \beta_c\}$$

Tukey's *higher criticsm*

- suggested by Tukey ($\approx$ 1976)
- modified by Donoho and Jin 2004
- Multiple tests: $\mathcal{H}_{0,i} : P_{n,i}$ against $\mathcal{H}_{1,i} : Q_{n,i}$ for all $i = 1, \ldots, n$.
  Use p-values $p_{n,i} = T_{n,i}(X_{n,i})$ with $P_{n,i}^{T_{n,i}} = \lambda_{|(0,1)}$

$$HC_n := \sup_{0<\alpha<1} \sqrt{n}\, \frac{\widehat{F}_{n,p}(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)}},\ \alpha \in (0,1)$$

where $\widehat{F}_{n,p}$ denotes the empirical distribution function of the p-values

Tukey's *higher criticsm*

- Connection to stochastic processes
- Convergence under the null (see Jaeschke and Eicker, resp.):

$$a_n \, HC_n - b_n \xrightarrow{\mathrm{d}} Z \quad \text{(Gumbel distributed)}$$

$$a_n := \sqrt{2\log^2(n)} \text{ and } b_n := 2\log^2(n) + \frac{1}{2}\log^3(n) - \frac{1}{2}\log(\pi).$$

Tukey's *higher criticsm*

- Connection to stochastic processes
- Convergence under the null (see Jaeschke and Eicker, resp.):

$$a_n \, HC_n - b_n \xrightarrow{\mathrm{d}} Z \quad \text{(Gumbel distributed)}$$

$$a_n := \sqrt{2\log^2(n)} \;\text{ and }\; b_n := 2\log^2(n) + \frac{1}{2}\log^3(n) - \frac{1}{2}\log(\pi).$$

- Critical value $\approx \sqrt{2\log^2(n)}$
- Modifications of HC_n: absolute value, $(\frac{1}{n}, \alpha_0)$, $(\frac{1}{n}, 1 - \frac{1}{n}), \ldots$

Advantage: HC_n does not depend on the unknown parameters

For simplicity

We consider here only the rowwise identical case

$$\mu_{n,i} = \mu_n,\ T_{n,i} = T_n\,, \varepsilon_{n,i} = \varepsilon_n.$$

The first result holds also in the more general case.

Let $v \in (0, \frac{1}{2})$

$$H_n(v) = \sqrt{n}\,\varepsilon_n \left\{ \frac{\left|\mu_n^{T_n}(0, v] - v\right| + \left|\mu_n^{T_n}(1 - v, 1] - v\right|}{\sqrt{v}} \right\}.$$

Let $v \in (0, \frac{1}{2})$

$$H_n(v) = \sqrt{n}\,\varepsilon_n \left\{ \frac{\left|\mu_n^{T_n}(0, v] - v\right| + \left|\mu_n^{T_n}(1-v, 1] - v\right|}{\sqrt{v}} \right\}.$$

Theorem (D & Janssen 2017)

Under some more conditions we have:

(a) $a_n^{-1} H_n(v_n) \to \infty \quad \Rightarrow a_n\, HC_n - b_n \xrightarrow{Q_{(n)}} \infty$ *(complete separation)*

(b) $\sup_{v \in I_n} a_n H_n(v) \to 0 \Rightarrow a_n\, HC_n - b_n \xrightarrow{\mathrm{d}} Z$ *under* $Q_{(n)}$ *(no separation)*

Optimality of HC

Figure: heterogeneous normal

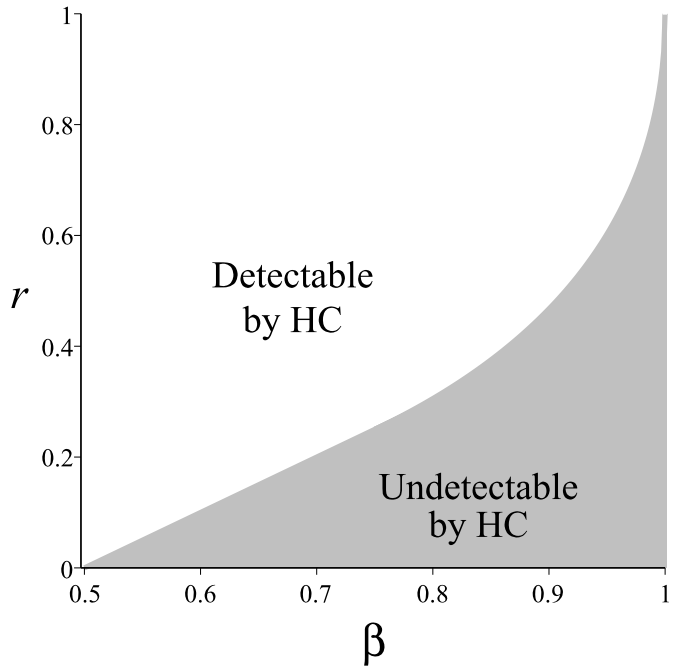
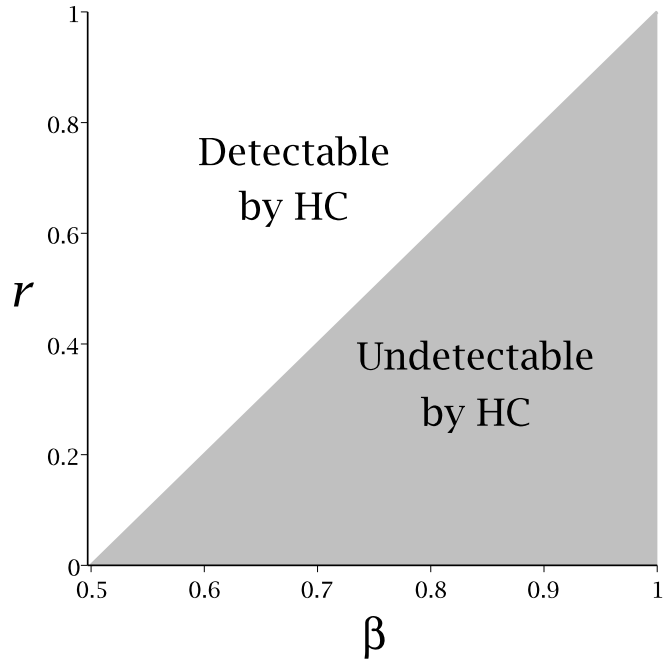


Figure: chimeric alternatives



Detection areas coincide for different (new) model assumptions, in particular we solved an open problem in Cai and Wu (2014)

Optimality of HC

Figure: heterogeneous normal

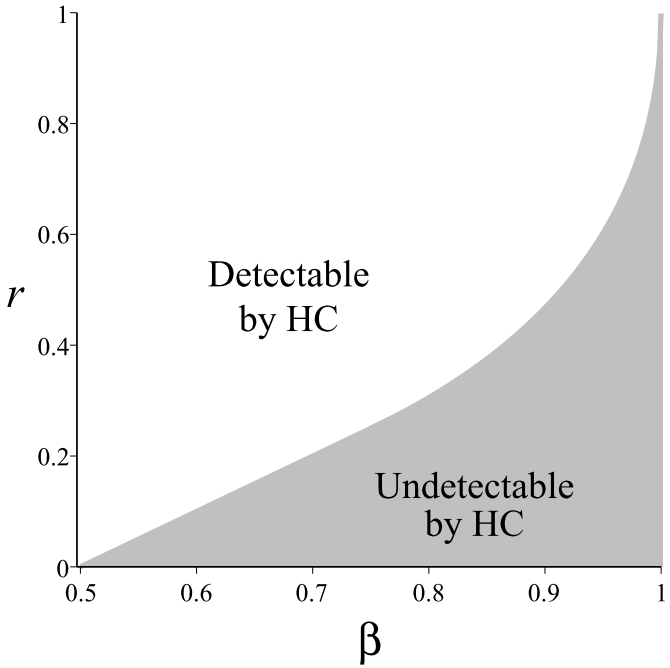
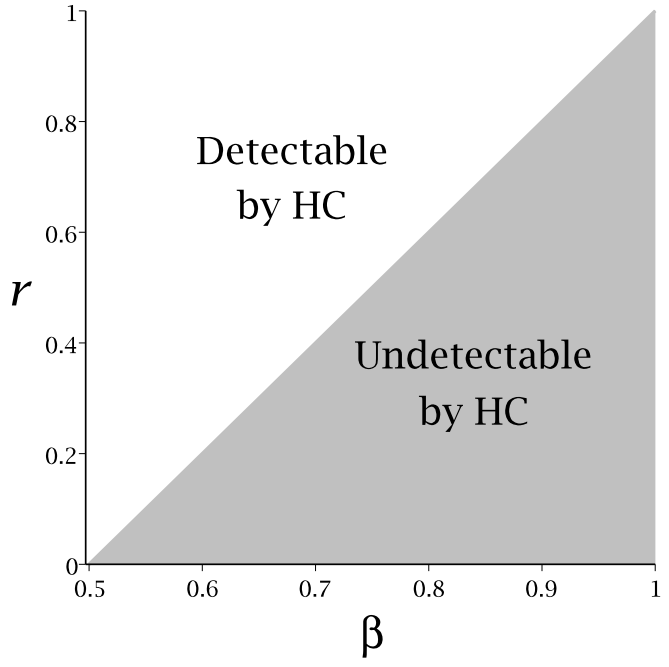


Figure: chimeric alternatives



Theorem (D & Janssen 2017)

HC has no power on the detection boundary (for all our examples).

Summary

- Fruitful theory for the detection boundary
- Tool to determine the limit distribution for all cases
- Extension of the detection boundary
  - New limit distributions ($P(\xi_2 \in \mathbb{R}) \in (0,1)$)
- Tools for HC for general distributions (in literature: mainly normal)
- Detection areas coincide.
- No power of HC on the boundary.

References I


ARIAS-CASTRO, E. AND WANG, M. (2015). The sparse Poisson means model. *Electron. J. Stat.* **9**, no. 2, 2170–2201.


CAI, T., JENG, J. AND JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, no. 5, 629–662.


CAI, T. AND WU, Y. (2014). Optimal Detection of Sparse Mixtures Against a Given Null Distribution. *IEEE Trans. Inform. Theory* **60**, no. 4, 2217-2232.


DITZHAUS, M. (2017). The power of tests for signal detection in high-dimensional data. *Dissertation, Heinrich Heine University Düsseldorf.* `https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=42808`


DONOHO, D. AND JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, no. 3, 962–994.


DONOHO, D. AND JIN, J. (2015). Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. *Statist. Sci.* **30** , no. 1, 1–25.


HOPKINS, A. M., MILLER, C. J., CONNOLLY, A. J., GENOVESE, C., NICHOL, R. C. AND WASSERMAN, L. (2002). A new source detection algorithm using the false-discovery rate. *Astron. J.* **123**, 1086–1094.

References II


INGSTER, Y. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6**, no. 1, 47–69.


JANSSEN, A., MILBRODT, H. AND STRASSER, H. (1985). *Infinitely divisible statistical experiments.* Lecture notes in Statistic **27**, Springer-Verlag, Berlin.


KHMALADZE, E.V. (1998). Goodness of fit tests for Chimeric alternatives.*Statist. Neerlandica* **52**, no. 1, 90–111.

DITZHAUS, M. AND JANSSEN, A. (2017). The power of big data sparse signal detection tests on nonparametric detection boundaries. *Submitted*. (arXiv 1709.07264)

Thank you for listening!

Figure: heteroscedastic normal

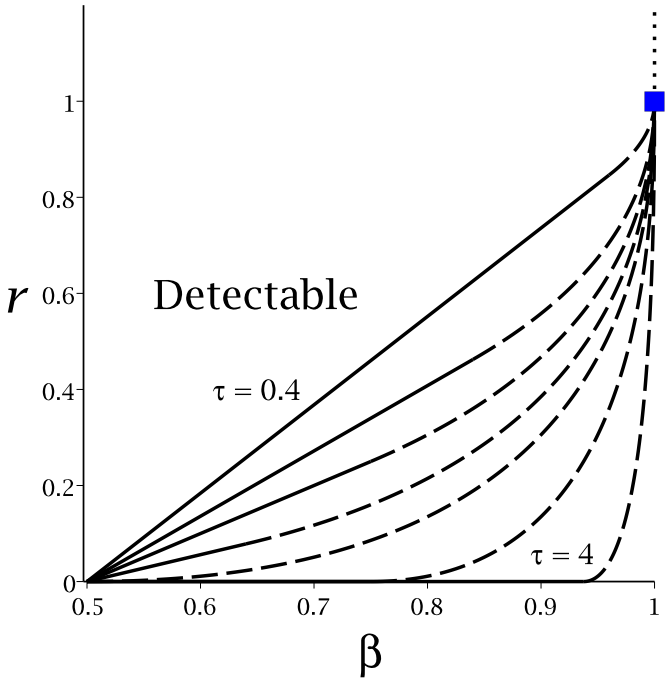
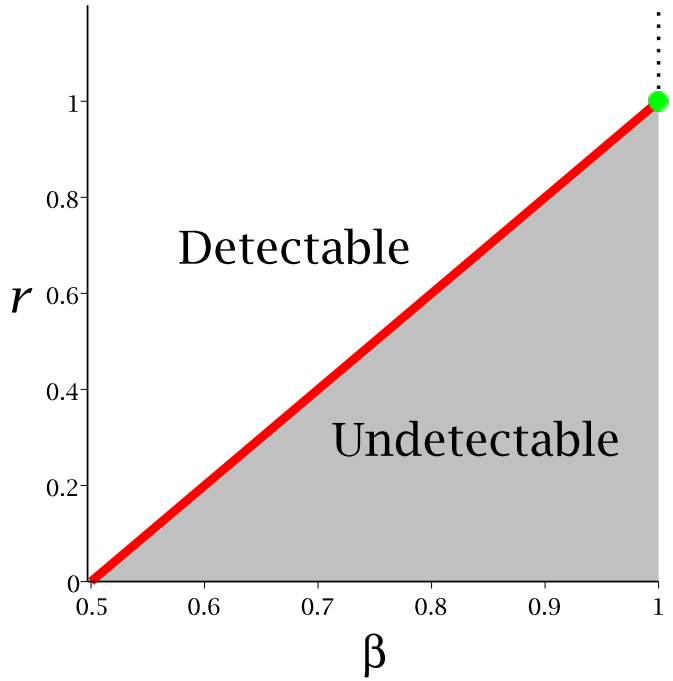


Figure: chimeric alternatives



DITZHAUS, M. AND JANSSEN, A. (2017). The power of big data sparse signal detection tests on nonparametric detection boundaries. *Submitted*. (arXiv 1709.07264)

Theorem (D & Janssen 2017)

(i) $I_{j,n}(x) \to 0$ *for some* $x > 0 \Leftrightarrow$ *Undetectable*

(ii) $I_{1,n}(x) \to \infty$ *or* $I_{2,n}(x) \to \infty$ *for some* $x \Leftrightarrow$ *Detectable*

Recall $\xi_1 \sim \nu_1$, where ν_1 has L-K-triplet $(\gamma_1, \sigma_1^2, \eta_1)$

$\xi_2 \sim (1 - a)\,\epsilon_\infty + a\,\nu_2$, where ν_2 has L-K-triplet $(\gamma_2, \sigma_2^2, \eta_2)$

Recall $\xi_1 \sim \nu_1$, where ν_1 has L-K-triplet $(\gamma_1, \sigma_1^2, \eta_1)$

$\xi_2 \sim (1-a)\,\epsilon_\infty + a\,\nu_2$, where ν_2 has L-K-triplet $(\gamma_2, \sigma_2^2, \eta_2)$

Theorem (D & Janssen 2017)

(iii) *Suppose that there is a finite measure M on $(0, \infty]$ and a dense subset $\mathcal{D}$ of $(0, \infty)$ such that for all $x \in \mathcal{D}$*

$$I_{1,n}(x) \to M(x, \infty] \quad \textit{and} \quad \sigma^2 = \lim_{\varepsilon \searrow 0} \begin{matrix} \limsup \\ \liminf \\ n \to \infty \end{matrix} I_{2,n}(\varepsilon).$$

Recall $\xi_1 \sim \nu_1$, where ν_1 has L-K-triplet $(\gamma_1, \sigma_1^2, \eta_1)$

$\xi_2 \sim (1-a)\,\epsilon_\infty + a\,\nu_2$, where ν_2 has L-K-triplet $(\gamma_2, \sigma_2^2, \eta_2)$

Theorem (D & Janssen 2017)

(iii) *Suppose that there is a finite measure M on $(0,\infty]$ and a dense subset $\mathcal{D}$ of $(0,\infty)$ such that for all $x \in \mathcal{D}$*

$$I_{1,n}(x) \to M(x,\infty] \quad \textit{and} \quad \sigma^2 = \lim_{\varepsilon \searrow 0} \begin{matrix} \limsup \\ \liminf \\ n\to\infty \end{matrix} I_{2,n}(\varepsilon).$$

Then $a = \exp(-M(\infty))$, $\eta_2 - \eta_1 = M_{|(0,\infty)}$, $\frac{\mathrm{d}\eta_2}{\mathrm{d}\eta_1} = \exp$, $\sigma_j^2 = \sigma^2$ *and*

$$\gamma_j := (-1)^j \frac{\sigma^2}{2} + \int_{(0,\infty)} \left(e^x - 1 + \frac{x}{1+x^2} \right) \mathrm{d}\eta_j(x)$$

h-model in general

Suppose that

$$\sum_{i=1}^{n} \frac{\varepsilon_{n,i}^2}{\tau_{n,i}} \;\to\; K \in [0,\infty] \text{ and } \sum_{i=1}^{n} \varepsilon_{n,i}^2 \;\to\; 0 \text{ as } n \to \infty.$$

Then,

- If $K = 0$ then $\xi_1 = \xi_2 = 0$ (Undetectable).
- If $K = \infty$ and $\limsup_{n\to\infty} \max_{1\le i\le n} \frac{\varepsilon_{n,i}}{\tau_{n,i}} < \infty$ then $\xi_1 = -\infty$ and $\xi_2 = \infty$ (Detectable).
- If $K \in (0,\infty)$ and $\max_{1\le i\le n} \frac{\varepsilon_{n,i}}{\tau_{n,i}} \to 0$ then $\xi_j \sim N\left((-1)^j \frac{\sigma^2}{2}, \sigma^2\right)$ with $\sigma^2 = Kc_2$ (non-trivial+Gaussian).