Resampling-based confidence regions in high dimension, from a non-asymptotic point of view

G. Blanchard[1]
Joint work with S. Arlot[2] and E. Roquain[3]

[1]Fraunhofer FIRST, Berlin, Germany

[2]Université de Paris-Sud, Orsay, France

[3]INRA, Jouy-en-Josas, France

Approximation and learning, Texas A&M 10/20/07

Plan

Plan

Setting

- Observation: a vector $Y \in \mathbb{R}^K$
- sample $\mathbf{Y}$ of n i.i.d. repetitions $\mathbf{Y} = (Y^1, \ldots, Y^n)$
- Unknown mean vector μ
- Unknown dependency between the coordinates
- "Small n large K" : $n \ll K$
- Goal 1: confidence region for μ?
- Goal 2: find coordinates $k : \mu_k \neq 0$? (Multiple testing)

Assumptions

- **(GA):** Y is Gaussian with known bound on coordinate variance $\sigma^2 \geq \max_k \operatorname{Var}[Y_k]$

or

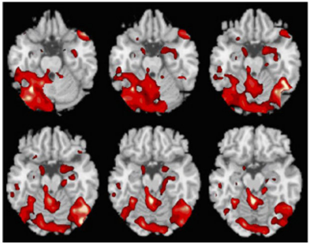
- **(BSA):** Y is bounded by known B and has a symmetric distribution

Some motivations

- Neuroimaging:
- small number of observations n of a noisy image with large number K of pixels
- want to detect where signal is present or obtain a confidence envelope about the signal
- strong spatial dependence with unknown structure (possibly non stationary, possible long-distance correlation. . .)
- Microarrays:
- detect significant differences (typical problem of multiple testing)
- completely unknown dependencies

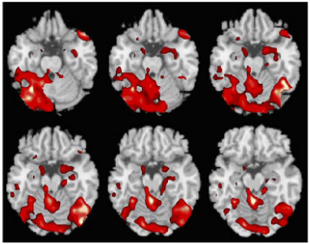
Some motivations

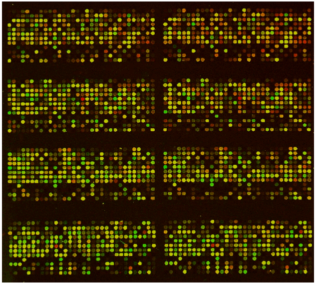
- Neuroimaging:
- small number of observations n of a noisy image with large number K of pixels
- want to detect where signal is present or obtain a confidence envelope about the signal
- strong spatial dependence with unknown structure (possibly non stationary, possible long-distance correlation. . .)
- Microarrays:
- detect significant differences (typical problem of multiple testing)
- completely unknown dependencies

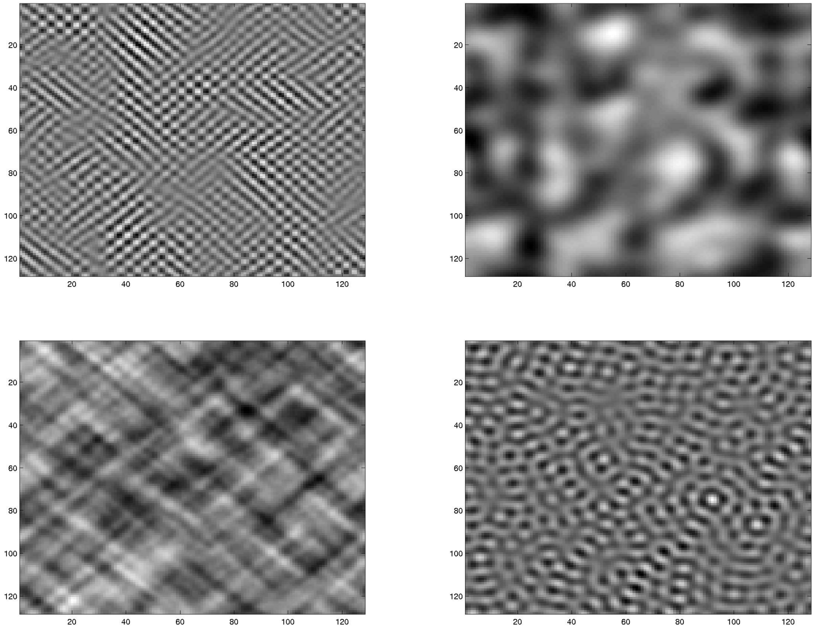


Some motivations

- Neuroimaging:
- small number of observations n of a noisy image with large number K of pixels
- want to detect where signal is present or obtain a confidence envelope about the signal
- strong spatial dependence with unknown structure (possibly non stationary, possible long-distance correlation. . .)
- Microarrays:
- detect significant differences (typical problem of multiple testing)
- completely unknown dependencies







20
40
60
80
100
120

Precising goals: confidence regions

- We are interested in "ψ-distance" uniform confidence regions based on the empirical mean $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y^i$, of the form

$$\left\{ \psi(\overline{\mathbf{Y}} - \mu) \leq t \right\} .$$

- Goal: find a threshold t as sharp as possible so that the above region has covering probability $1 - \alpha$.

Precising goals: multiple testing

- We want to test for every coordinate k the hypothesis $H_k : \mu_k = 0$ against the alternative $\mu_k \neq 0$
- Multiple testing procedure: rejects a subset of hypotheses

$$R(\mathbf{Y}) \subset \{1, \dots, K\}$$

- We want to control the family-wise error rate

$$\text{FWER}(R) = \mathbb{P}\left[\exists k \in R(\mathbf{Y}) | \mu_k = 0\right]$$

- Goal: $FWER(R) \leq \alpha$ while having high power, i.e. large $|R|$.
- Relationship to confidence region: if $R(\mathbf{Y}) = \left\{k : |\overline{\mathbf{Y}}_k| > t\right\}$,

$$\text{FWER}(R) \leq \mathbb{P}\left[\left\|\overline{\mathbf{Y}} - \mu\right\|_\infty > t\right]$$

- Note: can be used as a first step to control other type I error criteria such as FDP and FDR (Pacifico et al. 2004)

Bonferroni threshold

- under a Gaussian distribution and for $\psi(x) = \|x\|_\infty$, a union bound over coordinated gives the threshold

$$t^{Bonf} = \sigma \overline{\Phi}^{-1}(\alpha/(2K)) \, ,$$

where $\overline{\Phi}$ is the standard Gaussian cdf.

- deterministic threshold
- too conservative if there are strong dependencies between the coordinates
- to do better (and for more general ψ), take into account the observed dependencies.
- Note: $n \ll K$ essentially prevents us from using classical parametric methods, e.g. estimation of the covariance matrix.

Resampling

- Obviously, the ideal threshold is the $(1-\alpha)$-quantile q^*_α of $\psi(\overline{\mathbf{Y}}-\mu)$:
$$\mathbb{P}\left[\psi(\overline{\mathbf{Y}}-\mu) \leq q^*_\alpha\right] = \alpha \,.$$
- We want to use a resampling principle
- Usual (bootstrap) resampling: sample uniformly with replacement a n-sample $\widetilde{\mathbf{Y}}$ from the original sample $\mathbf{Y}$
- Resampling heuristics: the empirical process $\mathbb{P}_{\widetilde{\mathbf{Y}}} - \mathbb{P}_{\mathbf{Y}}$ conditional to $\mathbf{Y}$ "mimics" the empirical process $\mathbb{P}_{\mathbf{Y}} - \mathbb{P}$

Generalized resampling

- We consider more generally a reweighted sample scheme
- $W = (W_1, \ldots, W_n)$ vector of random weights independent of $\mathbf{Y}$ (but not necessarily jointly independent)
- Consider the reweighted sample $(Y^1, W_1), \ldots, (Y^n, W_n)$
  - Example 1: Efron's bootstrap: W is a multinomial $(n; n^{-1}, \ldots, n^{-1})$
  - Example 2: Rademacher weights: W_i i.i.d random signs
  - Example 3: Leave-one-out: $W_i = \mathbf{1}\{i = i_0\}$, $i_0 \sim \mathcal{U}(\{1 \ldots, n\})$.
- We consider in particular the reweighted mean

$$\overline{\mathbf{Y}}^{\langle W \rangle} = \sum_{i=1}^{n} W_i \mathbf{Y}^i$$

How to study resampling?

- K fixed and $n \to \infty$: asymptotic results (eg. van der Vaart and Wellner 1996); not adapted to our (typically non-asymptotic) setting.
- Idea 1: non-asymptotic results inspired from learning theory (for bounded random variables): Rademacher complexities (Koltchinskii 2001, Bartlett and Mendelson 2002), more general reweighting schemes (Fromont 2005). Based on concentration and comparison in expectation.
- Idea 2: try to estimate directly the quantile using ideas coming from exact (permutation) tests.

How to study resampling?

- K fixed and $n \to \infty$: asymptotic results (eg. van der Vaart and Wellner 1996); not adapted to our (typically non-asymptotic) setting.
- Idea 1: non-asymptotic results inspired from learning theory (for bounded random variables): Rademacher complexities (Koltchinskii 2001, Bartlett and Mendelson 2002), more general reweighting schemes (Fromont 2005). Based on concentration and comparison in expectation.
- Idea 2: try to estimate directly the quantile using ideas coming from exact (permutation) tests.

How to study resampling?

- K fixed and $n \to \infty$: asymptotic results (eg. van der Vaart and Wellner 1996); not adapted to our (typically non-asymptotic) setting.
- Idea 1: non-asymptotic results inspired from learning theory (for bounded random variables): Rademacher complexities (Koltchinskii 2001, Bartlett and Mendelson 2002), more general reweighting schemes (Fromont 2005). Based on concentration and comparison in expectation.
- Idea 2: try to estimate directly the quantile using ideas coming from exact (permutation) tests.

Plan

Result based on concentration

Theorem

Assume **(GA)**, ψ *is positive-homogeneous, subadditive and bounded by* $\|.\|_p$ *; W squared-integrable, exchangeable weight vector.*
Then, for any $\alpha \in (0,1)$ *:*

$$t_\alpha^{conc}(\mathbf{Y}) := \frac{\mathbb{E}_W\left[\psi(\overline{\mathbf{Y}}^{\langle W \rangle} - \overline{W\mathbf{Y}})\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{\sqrt{n}B_W} + 1\right]$$

satisfies

$$\mathbb{P}\left[\psi(\overline{\mathbf{Y}} - \mu) > t_\alpha^{conc}\right] \leq \alpha \, .$$

With $\sigma_k^2 = \operatorname{Var}\left[Y_k^1\right]$,

$$B_W = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})^2\right)^{\frac{1}{2}}\right] \; ; \qquad C_W = \left(\frac{n}{n-1}\mathbb{E}\left[(W_1 - \overline{W})^2\right]\right)^{\frac{1}{2}}$$

Main ingredients

- Comparison of expectations:

$$B_W \mathbb{E}\left[\psi(\overline{\mathbf{Y}} - \mu)\right] = \mathbb{E}\left[\psi(\overline{\mathbf{Y}}^{\langle W \rangle} - \overline{W}\overline{\mathbf{Y}})\right]$$

- Gaussian concentration theorem for Lipschitz functions of an i.i.d. Gaussian vector (Cirels'on, Ibragimov and Sudakov 1976)
  - for $\psi(\overline{\mathbf{Y}} - \mu)$: deviations bounded by a normal tail of standard deviation $\leq \|\sigma\|_p n^{-\frac{1}{2}}$;
  - for $\mathbb{E}_W \left[\psi(\overline{\mathbf{Y}}^{\langle W \rangle} - \overline{W}\overline{\mathbf{Y}})\right]$:
    standard deviation $\leq C_W \|\sigma\|_p n^{-1}$.

Additional remarks

- $C_W B_W^{-1} \approx 1$ for Rademacher weights and leave-one-out weights
- Can be generalized to more general weights, e.g. V-fold cross-validation weights (with $C_W B_W^{-1} \approx \sqrt{n/V}$), with calculation complexity V
- can be generalized (with larger constants) to **(BSA)** (symmetric, bounded random variables), see also Fromont (2005).
- if a deterministic threshold is known (for example Bonferroni's threshold for $\psi = \|.\|_\infty$), it can be combined with the resampling-based threshold, by considering a threshold that is "very close" to the minimum of the two.
- if the expectation cannot be computed exactly, a Monte-Carlo method can be used.

Plan

Symmetrization idea

- suppose the distribution of Y is symmetric (around μ) .
- the distribution of the centered sample $\mathbf{Y}-\mu=(\mathbf{Y}^1-\mu,\ldots,\mathbf{Y}^n-\mu)$ is invariant by reweighting with arbitrary signs $W_i\in\{-1,1\}$.
- define $q_\alpha^{quant}(\mathbf{Y})$ as the $(1-\alpha)$ quantile of

$$\mathcal{D}(\psi(\overline{\mathbf{Y}}^{\langle W\rangle})|\mathbf{Y})\,,$$

  where W is a vector of i.i.d. Rademacher weights.
- Using the invariance we have

$$\mathbb{P}\left[\psi(\overline{\mathbf{Y}}-\mu)>q_\alpha^{quant}(\mathbf{Y}-\mu)\right]\leq\alpha.$$

- For $\mu=0$ this can be computed exactly and is used in the framework of exact tests.

Empirically recentered quantiles

- What can we do for unknown μ? Use the resampling heuristic and replace μ by $\overline{\mathbf{Y}}$, i.e., consider

$$q_{\alpha}^{quant}(\mathbf{Y} - \overline{\mathbf{Y}})$$

- What kind of theoretical guarantee can we have for the empirically recentered quantile?

Theoretical guarantee for empirically recentered quantile

Theorem

Let $\alpha, \delta, \gamma \in]0,1[$ *and* f *a non-negative function such that*

$$\mathbb{P}\left[\psi(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y})\right] \leq \frac{\alpha\gamma}{2} \, ;$$

then the threshold

$$t_\alpha^{quant+f}(\mathbf{Y}) := q_{\alpha(1-\delta)(1-\gamma)}^{quant}(\mathbf{Y} - \overline{\mathbf{Y}}) + \sqrt{\frac{2\log(2/(\delta\alpha))}{n}} f(\mathbf{Y})$$

satisfies

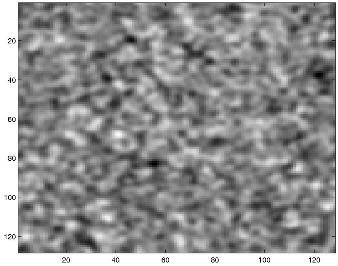
$$\mathbb{P}\left[\psi(\overline{\mathbf{Y}} - \mu) > t_\alpha^{quant+f}(\mathbf{Y})\right] \leq \alpha \, .$$

Remarks

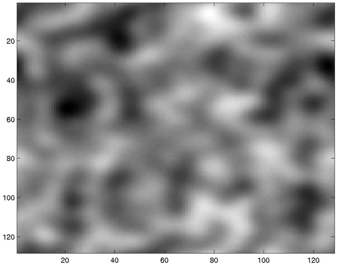
$$t_\alpha^{quant+f}(\mathbf{Y}) = q_{\alpha(1-\delta)(1-\gamma)}^{quant}(\mathbf{Y} - \overline{\mathbf{Y}}) + \sqrt{\frac{2\log(2/(\delta\alpha))}{n}} f(\mathbf{Y})$$

- the only assumption on Y is the symmetry of its distribution.
- the function f only appears as a second-order term.
- the theorem can be iterated, resulting in terms of increasing order.
- to obtain a computable threshold, we need to have a bound on some extreme quantile of the distribution.
- under additional assumptions (e.g. boundedness or Gausiannity) we can take f as one of the previous thresholds: $t_{\alpha\gamma/2}^{conc}$, $t_{\alpha\gamma/2}^{Bonf}\dots$
- the point is that the threshold used to define f does not have to be very sharp.
- if the quantile is computed approximately using a Monte-Carlo scheme with B repetitions, we lose at most $(B+1)^{-1}$ in the covering probablity.

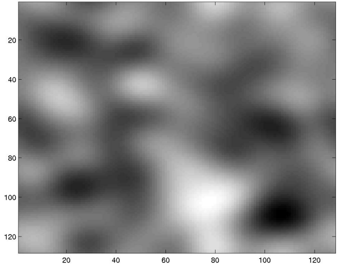
Plan


20
40
60
80
100
120

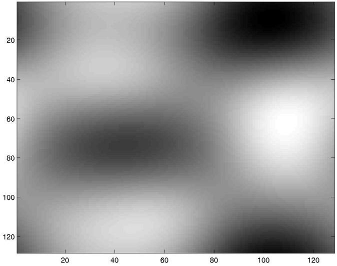
b=2


20
40
60
80
100
120

b=6

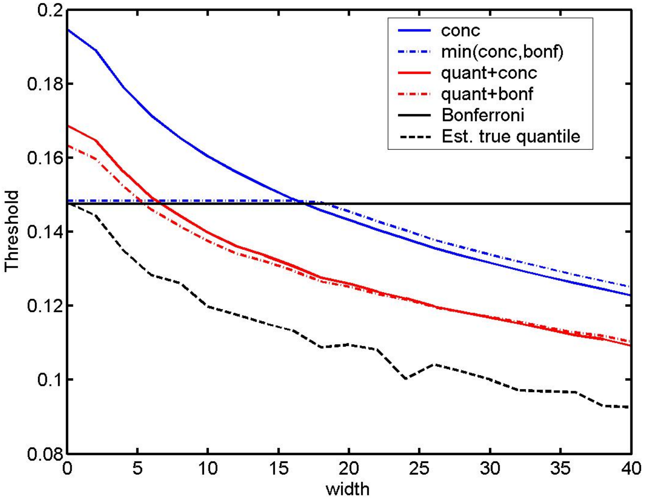

20
40
60
80
100
120

b=12

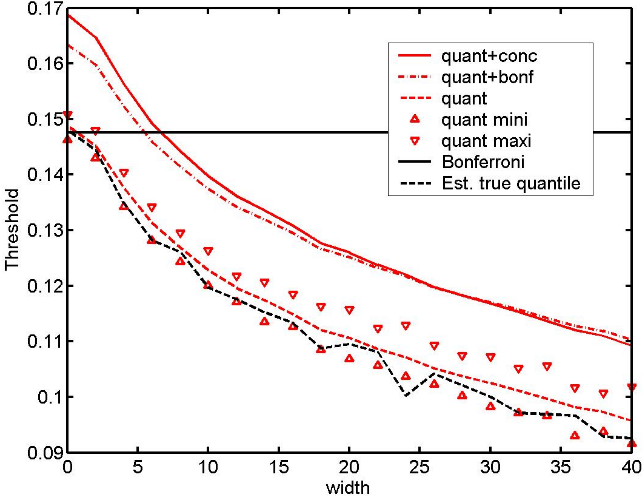

20
40
60
80
100
120

b=30

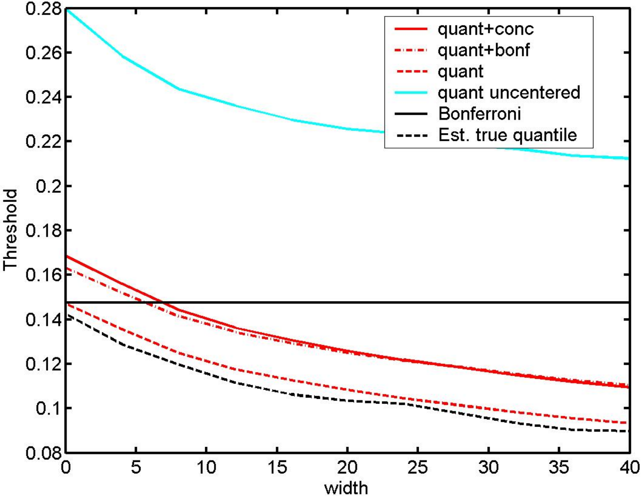
Simulations: n=1000 , $K = 128^2$, $\sigma = 1$



Simulations: without the additive term in quantiles?



Simulations: thresholds with non-zero means, $\mu_k \in [0,3]$



Plan

High points

We proposed two different methods to obtain non-asymptotic confidence regions for Gaussian random variables in high dimension with unknown correlations.

- concentration method inspired from learning theory: applicable to many different reweighting schemes.
- direct quantile estimation using symmetrization techniques

- non-asymptotic: valid for any K and n
- no knowledge on dependency structure required
- translation invariant (unlike classical symmetrized thresholds for testing)
- better than Bonferroni/Holm if there are strong correlations present
- can be used to accelerate classical step-down procedures when computation time is an issue

Perspectives for future work

- theoretical study of power/ asymptotic threshold optimality
- what about the quantile approach with other weights, with a non-symmetric distribution?
- application to model selection?
- application to adaptive testing?