Habilitation à diriger des recherches

Contributions to nonparametric hypotheses testing and statistical learning

Magalie Fromont

IRMAR / Université Rennes 2 (France)

2 décembre 2015
A theory in response to concrete challenges in various fields

- Laser vibrometry
- Public statistics
- Genetics
- Neuroscience
Single tests of single null hypotheses

Observed random variable: \( X \), defined on \((\Omega, \mathcal{A}, P)\), with distribution \( P \).

Possible set of distributions for \( X \) defined from a nonparametric model: \( \mathcal{P} \).

Single null hypothesis defined through \( \mathcal{P}_0 \subset \mathcal{P} \) as \((H_0)\) \( P \in \mathcal{P}_0 \).

Alternative hypothesis \((H_1)\) \( P \in \mathcal{P} \setminus \mathcal{P}_0 \).

A (single) nonrandomized test of \((H_0)\) against \((H_1)\) is a statistic \( \phi \) depending on \( X \):

- with value 1 when \( X \) leads to reject \((H_0)\) in favor of \((H_1)\),
- with value 0 otherwise.
Single tests of single null hypotheses
Nonasymptotic minimax testing

- First kind error requirement (Neyman-Pearson): given $\alpha$ in $(0, 1)$,
  $\sup_{P \in \mathcal{P}_0} P(\phi = 1) := \mathbb{P}(H_0)(\phi = 1) \leq \alpha$ (level $\alpha$ test).

- Second kind error requirement: given $\beta$ in $(0, 1)$,
  $\sup_{P \in \mathcal{P}_1} P(\phi = 0) \leq \beta$, with $\mathcal{P}_1 \subset \mathcal{P} \setminus \mathcal{P}_0$ as large as possible.

× In general, if $\alpha + \beta < 1$, $\mathcal{P}_1$ can not be equal to $\mathcal{P} \setminus \mathcal{P}_0$!

$\implies \mathcal{P}_1 = \{ P \in \mathcal{P}', \ d(P, \mathcal{P}_0) \geq r \}$, with $r$ as small as possible,

for some distance $d$ on $\mathcal{P}$, and (realistic ?) restricted class of probability distributions $\mathcal{P}' \subset \mathcal{P}$.

Let $\phi_\alpha$ be a level $\alpha$ test of $(H_0)$ against $(H_1)$.
The **uniform separation rate** of $\phi_\alpha$ over $\mathcal{P}'$ is defined as

$$SR^\beta_{d}(\phi_\alpha, \mathcal{P}') = \inf \left\{ r > 0, \ \sup_{P \in \mathcal{P}', d(P, \mathcal{P}_0) \geq r} P(\phi_\alpha = 0) \leq \beta \right\}.$$
Introduction

Aggregated tests: goodness-of-fit

Aggregated tests: two-sample problems

Multiple tests

Conclusion

\[
\sup_{P \in \mathcal{P}', d(P, P_0) \geq r} P(\phi_\alpha = 0)
\]
\[ \sup_{P \in \mathcal{P}', d(P, P_0) \geq r} \mathbb{P}(\phi_\alpha = 0) \]
\[ \sup_{P \in \mathcal{P}'} d(P, P_0) \geq r \quad P(\phi_\alpha = 0) \]

\[ 1 - \alpha \]
\[ \sup_{P \in \mathcal{P}', d(P, P_0) \geq r} P(\phi_\alpha = 0) \]

\[ P_0 \]

\[ \mathcal{P}' \]

\[ r \]
\[
\sup_{P \in \mathcal{P}', d(P, \mathcal{P}_0) \geq r} P(\phi_\alpha = 0)
\]
\[ \sup_{P \in \mathcal{P}', d(P, \mathcal{P}_0) \geq r} P(\phi_\alpha = 0) \]
$\sup_{P \in \mathcal{P}', d(P, \mathcal{P}_0) \geq r} \mathbb{P}(\phi_\alpha = 0)$

$\mathcal{P}'$

$\mathcal{P}_0$

$\text{SR}_d^\beta(\phi_\alpha, \mathcal{P}')$

$1 - \alpha$

$\beta$

$0$

$\text{SR}_d^\beta(\phi_\alpha, \mathcal{P}')$

$r$
Single tests of single null hypotheses
Nonasymptotic minimax testing

A second kind error related criterion which allows to:

- Compare two level $\alpha$ tests
- See whether a level $\alpha$ test is optimal over $\mathcal{P}'$, in the following minimax sense.

**The minimax separation rate** over $\mathcal{P}'$ is defined by

$$m\text{SR}_{d}^{\alpha,\beta} (\mathcal{P}') = \inf_{\{\phi_{\alpha} \text{ of level } \alpha\}} \text{SR}_{d}^{\beta} (\phi_{\alpha}, \mathcal{P}').$$

A level $\alpha$ test $\phi_{\alpha}$ is **minimax** over $\mathcal{P}'$, if

$$\text{SR}_{d}^{\beta} (\phi_{\alpha}, \mathcal{P}') \leq C(\alpha, \beta) m\text{SR}_{d}^{\alpha,\beta} (\mathcal{P}').$$

$\Rightarrow$ Parallel between the minimax hypothesis testing theory and the minimax estimation theory
Single tests of single null hypotheses
Nonasymptotic minimax testing: example in the density model

Density model: \( X = (X_1, \ldots, X_n) \) is a sample of \( n \) i.i.d. random variables with distribution \( P_f \) of density \( f \) with respect to the Lebesgue measure \( \lambda \) on \( X = \mathbb{R} \), \( \mathcal{P} = \{ P_f, f \in L_2(\mathbb{R}, \lambda) \} \).

Goodness-of-fit test: given a density \( f_0 \in L_2(\mathbb{R}, \lambda) \),

\[
(H_0) \quad f = f_0 \iff P_f \in \mathcal{P}_0 = \{ P_{f_0} \} \quad \text{against} \quad (H_1) \quad f \neq f_0 \iff P_f \notin \mathcal{P}_0 = \{ P_{f_0} \}
\]

Minimax separation rate: \( d_2(P_f, P_g) = \| f - g \|_2, B_{s,\infty,\infty}(\mathbb{R}) \) Hölder ball,

\[
mSR_{d_2}^{\alpha,\beta} \left( \{ P_f, f \in B_{s,\infty,\infty}(\mathbb{R}) \} \right) \approx n^{-2s/(4s+1)}
\]

Single tests of single null hypotheses
Nonasymptotic minimax testing: example in the density model

$S_m = \langle b_{m,k}, \ k \in \mathbb{Z} \rangle$, with $b_{m,k} = \sqrt{m} \mathbb{1}_{[k/m,(k+1)/m)}$ for $m \in \mathbb{N} \setminus \{0\}$,

$\Pi_{S_m}$ orthogonal projection onto $S_m$ w.r.t. $\langle ., . \rangle_2$

$(H_{0,m}) \ P_f \in \mathcal{P}_{0,m}$, with $\mathcal{P}_{0,m} = \{ P_f, \ \Pi_{S_m}(f - f_0) = 0 \} \supset \mathcal{P}_0$.

**Single test:** $\phi_{m,\alpha} = \mathbb{1}_{\{ T_m > F_m^{-1}(1-\alpha) \}}$, with

$$T_m = \frac{1}{n(n-1)} \sum_{k \in \mathbb{Z}} \sum_{i \neq j = 1}^n b_{m,k}(X_i) b_{m,k}(X_j) + \|f_0\|^2 - \frac{2}{n} \sum_{i = 1}^n f_0(X_i)$$

estimating $\|\Pi_{S_m}(f - f_0)\|^2$, $F_m = $ c.d.f. of $T_m$ under $(H_0)$

$\phi_{m,\alpha}$ is a level $\alpha$ test such that $P_f(\phi_{m,\alpha} = 0) \leq \beta$ as soon as

$$d_2^2(P_f, \mathcal{P}_0) > (1 + \varepsilon) \left\{ \|f - \Pi_{S_m}(f)\|^2_2 + C \left( \frac{\sqrt{m \ln(1/\alpha)}}{n} + \frac{m}{n^2} \right) \right\}.$$  


Tools: concentration inequalities ($U$ statistics of order 2, linear statistics)
Single tests of single null hypotheses

Nonasymptotic minimax testing: example in the density model

**Bias term:** for $s \in (0, 1]$, $f \in \mathcal{B}_{s,\infty,\infty}(R) \Rightarrow \|f - \Pi_{S_m}(f)\|^2 \leq C(s)R^2m^{-2s}$

**Minimax test:** Take $m$ such that $R^2m^{-2s} \sim \sqrt{m/n} \iff m \sim (R^2n)^{2/(4s+1)}$.

For $n$ large,

$$SR_{d_2}^\beta(\phi_{m,\alpha}, \{P_f, \ f \in \mathcal{B}_{s,\infty,\infty}(R) \cap L_\infty(R')\}) \leq C(s, \alpha, \beta, R')R^{\frac{1}{4s+1}}n^{\frac{-2s}{4s+1}}.$$

× Problem: the test depends on $s$! A priori **realistic** choice of $\mathcal{B}_{s,\infty,\infty}(R)$?

⇒ Test which does not depend on $s$ but which is minimax or nearly minimax over the class $\{P_f, \ f \in \mathcal{B}_{s,\infty,\infty}(R) \cap L_\infty(R')\}$ for every $s$?

A level $\alpha$ test $\phi_\alpha$ is **minimax adaptive** over a collection of classes $\mathcal{P}'$, if it is minimax or nearly minimax over all the classes $\mathcal{P}'$ in the collection.

⇒ Aggregation of tests
Aggregated tests

- Collection of subsets of $\mathcal{P}$: $\{ \mathcal{P}_{0,m}, \ m \in \mathcal{M} \}$, $\mathcal{P}_0 \subset \cap_{m \in \mathcal{M}} \mathcal{P}_{0,m}$
- Collection of hypotheses: $\{(H_{0,m}), \ m \in \mathcal{M}\}$, $(H_{0,m}) P \in \mathcal{P}_{0,m}$
- Collection of tests: $\Phi_\alpha = \{ \phi_{m,\alpha} = 1 \{ T_m > q_m(1-\alpha) \}, \ m \in \mathcal{M} \}$ with $\sup_{P \in \mathcal{P}_0} P(\phi_{m,\alpha} = 1) \leq \alpha$
- Collection of individual levels: $U_\alpha = \{ u_{m,\alpha}, \ m \in \mathcal{M} \}$

The **aggregated test** based on the collections $\Phi_\alpha$ and $U_\alpha$ is defined as

$$\bar{\Phi}_\alpha = \sup_{m \in \mathcal{M}} \phi_{m,u_{m,\alpha}} = \sup_{m \in \mathcal{M}} 1 \{ T_m > q_m(1-u_{m,\alpha}) \}.$$ 

$\Rightarrow$ Reject $(H_0)$ if at least one $(H_{0,m})$ is rejected with $\phi_{m,u_{m,\alpha}}$
Aggregated tests

Two concerns: level control + minimax adaptivity

- Minimax adaptivity: choice of $T_m$ (minimax single tests)
- Level control: choice of $q_m(1 - u_{m,\alpha})$

Four different cases can be distinguished ($Z$ is a statistic depending on $X$).

Notation:

- $L_{(H_0)}(T) = \text{distribution of } T \text{ given } Z$,
- $L_{(H_0)}(T|Z) = \text{conditional distribution of } T \text{ given } Z \text{ under } (H_0)$,
- $L(T|Z) = \text{conditional distribution of } T \text{ given } Z$

- **[KD]** (Known Distr.)
  - $L_{(H_0)}(T_m)$ is known (parameter free)
- **[UD1]** $L_{(H_0)}(T_m|Z)$ is known
- **[UD]** (Unknown Distr.)
  - $\exists T^*_m, L(T^*_m|Z) = L_{(H_0)}(T_m|Z)$
  - $\exists T^*_m, L_{(H_0)}(T^*_m|Z) = L_{(H_0)}(T_m|Z)$

Magalie Fromont

Habilitation à diriger des recherches 17 / 38
Aggregated tests: goodness-of-fit

In the density model ([KD])

\[ S_m = \langle b_{m,k}, k \in \mathbb{Z} \rangle, \text{ with } b_{m,k} = \sqrt{m} \mathbb{1}_{[k/m,(k+1)/m)} \text{ for } m \in \mathbb{N} \setminus \{0\} \]

- Collection of subsets of \( \mathcal{P} \): \( \{ \mathcal{P}_{0,m} = \{ P_f, \prod S_m(f - f_0) = 0 \}, m \in \mathcal{M} \} \)
- Collection of hypotheses: \( \{( H_{0,m}) P_f \in \mathcal{P}_{0,m}, m \in \mathcal{M} \} \)
- Collection of tests: \( \{ \phi_{m,\alpha} = \mathbb{1}_{\{ T_m > F_m^{-1}(1-\alpha) \}}, m \in \mathcal{M} \} \)
- Collection of individual levels: \( \{ u_{m,\alpha}, m \in \mathcal{M} \} \)

**Bonferroni choice:** \( u_{m,\alpha} = \alpha/\#\mathcal{M} \)

\[ \Phi_{\alpha}^{Bonf} = \sup_{m \in \mathcal{M}} \phi_{m,\alpha}/\#\mathcal{M} = \sup_{m \in \mathcal{M}} \mathbb{1}_{\{ T_m > F_m^{-1}(1-\alpha/\#\mathcal{M}) \}} \]

**FL choice:** \( u_{m,\alpha} = u_{\alpha} = \sup \{ u, \mathbb{P}(H_0) \left( \exists m \in \mathcal{M}, T_m > F_m^{-1}(1-u) \right) \leq \alpha \} \)

\[ \Phi_{\alpha}^{FL} = \sup_{m \in \mathcal{M}} \phi_{m,u_{\alpha}} = \sup_{m \in \mathcal{M}} \mathbb{1}_{\{ T_m > F_m^{-1}(1-u_{\alpha}) \}} \]

\( \Phi_{\alpha}^{Bonf} \) and \( \Phi_{\alpha}^{FL} \) are both of level \( \alpha \), and \( \Phi_{\alpha}^{FL} \) is less conservative than \( \Phi_{\alpha}^{Bonf} \).
**Introduction**

**Aggregated tests: goodness-of-fit**

**Aggregated tests: two-sample problems**

**Multiple tests**

**Conclusion**

---

**Aggregated tests: goodness-of-fit**

In the density model ([KDu])

\[
P_f(\Phi^{FL}_\alpha = 0) \leq \beta \text{ as soon as } \\
d_2^2(P_f, \mathcal{P}_0) > (1 + \varepsilon) \inf_{m \in \mathcal{M}} \left\{ \| f - \Pi_{S_m}(f) \|_2^2 + C \left( \frac{\sqrt{m \ln(\#\mathcal{M}/\alpha)}}{n} + \frac{m}{n^2} \right) \right\} \]


---

**Taking \( \mathcal{M} \) with \( \#\mathcal{M} \simeq \ln n \) \( \implies \) loss in \( \sqrt{\ln \ln n} \)

---

For \( n \) large enough, \( s \in (0, 1] \), \( \mathcal{M} = \{ 2^J, 0 \leq J \leq \log_2 \left( \frac{n^2}{(\ln \ln n)^3} \right) \} \),

\[
SR_{d_2}^\beta (\Phi^{FL}_\alpha, \{ P_f, f \in B_{s, \infty, \infty}(R) \cap L_{\infty}(R') \}) \leq C \ R^{\frac{1}{4s+1}} \left( \frac{\sqrt{\ln \ln n/n}}{n} \right)^{\frac{2s}{4s+1}}
\]

闰 \( \Phi^{FL}_\alpha \) is minimax adaptive with an unavoidable (闰 Ingster (2000)) loss the order of a \( \sqrt{\ln \ln n} \) factor.

闰 Extension to test that \( f \) belongs to a translation/scale family: similar results but with a loss of the order of a \( \sqrt{\ln n} \) factor
Aggregated tests: goodness-of-fit

In the Poisson model ([UD1])

\[ X = \{ X_1, \ldots, X_{N_X} \} \] is a Poisson process on \( X = [0, 1] \), with intensity \( f \) w.r.t. \( d\mu = nd\lambda \), whose distribution is denoted by \( P_f \), \( \mathcal{P} = \{ P_f, f \in \mathbb{L}_2(\mathbb{R}, \lambda) \} \).

Homogeneity test

\[
\begin{align*}
\text{(H}_0\text{)} & \quad P_f \in \mathcal{P}_0 = \{ P_f, \ f \text{ constant} \} \quad \text{against} \quad \text{(H}_1\text{)} & \quad P_f \not\in \mathcal{P}_0
\end{align*}
\]

Motivation: Detecting abnormal behaviors on the DNA sequence

Detecting alternative intensities with localized spikes

Minimax separation rate?

\[
d_2(P_f, P_g) = \| f - g \|_2 \quad \text{(w.r.t. } \lambda),
\]

\( \mathcal{B}_{s,2,\infty}(R) \) (strong) Besov body, \( w\mathcal{B}_{s'}(R') \) weak Besov body

defined from the Haar basis \( \{ \varphi_0, \psi(j,k), j \in \mathbb{N}, k \in \{0, \ldots, 2^j - 1\} \} \).

\[
\begin{align*}
\mathcal{B}_{s,2,\infty}(R) &= \left\{ f, \ \forall j \in \mathbb{N}, \ \sum_{k=0}^{2^j-1} \langle f, \psi(j,k) \rangle_2^2 \leq R^2 2^{-2js} \right\} \\
w\mathcal{B}_{s'}(R') &= \left\{ f, \ \forall t > 0, \ \sum_{j \in \mathbb{N}} \sum_{k=0}^{2^j-1} \langle f, \psi(j,k) \rangle_2^2 \ 1_{\langle f, \psi(j,k) \rangle_2^2 \leq t} \leq R'^2 t^{2s'/2s'+1} \right\}
\end{align*}
\]
Aggregated tests: goodness-of-fit

In the Poisson model ([UD1])

$$m_{\text{SR}}^{\alpha, \beta} \left( \{ P_f, f \in \mathcal{B}_{s, 2, \infty}(R) \cap w\mathcal{B}_{s'}(R') \cap \mathcal{L}_{\infty}(R'') \} \right)$$

☞ Fromont, Laurent, Reynaud-Bouret, Ann. IHP (2011)
Aggregated homogeneity tests

\[ S_m = \langle \varphi_0, \psi_{(j,k)}, (j, k) \in \mathcal{L}_m \rangle, \text{ with} \]
\[ \mathcal{L}_m \subset \{(j, k), j \in \mathbb{N}, k \in \{1, \ldots, 2^j - 1\}\}, \ m \in \mathcal{M} \]

- Collection of subsets of \( \mathcal{P} \): \( \{ \mathcal{P}_{0,m} = \{ P_f, \prod_{S_m(f)} \text{ is constant} \} , m \in \mathcal{M} \} \)
- Collection of hypotheses: \( \{ (H_{0,m}) P \in \mathcal{P}_{0,m}, \ m \in \mathcal{M} \} \)
- Collection of single tests: \( \{ \phi_{m, \alpha} = 1 \{ T_m > q_m^{N_X}(1-\alpha) \}, \ m \in \mathcal{M} \} \), with

\[ T_m = \frac{1}{n^2} \sum_{(j,k) \in \mathcal{L}_m} \sum_{i \neq i'=1}^{N_X} \psi_{(j,k)}(X_i)\psi_{(j,k)}(X_i') \rightarrow \| \prod_{\langle \psi_{(j,k)}, (j,k) \in \mathcal{L}_m \rangle (f) \|_2^2 \]

\[ q_m^{n_0} \text{ quantile function of } \mathcal{L}(H_0) \mid N_X = n_0, \text{ which is known since} \]

\[ \mathcal{L}(H_0) \mid N_X = n_0 = \mathcal{L} \left( \frac{1}{n^2} \sum_{(j,k) \in \mathcal{L}_m} \sum_{i \neq i'=1}^{n_0} \psi_{(j,k)}(U_i)\psi_{(j,k)}(U_i') \right), \text{ with} \]
\( (U_1, \ldots, U_{n_0}) \text{ i.i.d. uniformly distributed (case [UD1] with } Z = N_X) \)
Aggregated tests: goodness-of-fit

In the Poisson model ([UD1])

- Collection of individual levels: \( \{ u_m, \alpha, \ m \in \mathcal{M} \} \) ?

**FLR choice:** \( u_{m,\alpha} = u_{m,\alpha}^{\mathcal{N}_X} \), with

\[
u_{m,\alpha}^{n_0} = w_m \sup \left\{ u, \mathbb{P}(H_0) \left( \exists m \in \mathcal{M}, \ T_m > q_m^{(n_0)}(1 - w_m u) \bigg| \mathcal{N}_X = n_0 \right) \leq \alpha \right\},
\]

\((w_m)_{m \in \mathcal{M}}\) positive weights such that \( \sum_{m \in \mathcal{M}} w_m \leq 1 \)

\[
\Phi^{FLR}_{\alpha} = \sup_{m \in \mathcal{M}} \phi_{m,u_m,\alpha} = \sup_{m \in \mathcal{M}} 1 \left\{ T_m > q_m^{\mathcal{N}_X} \left( 1 - u_{m,\alpha}^{\mathcal{N}_X} \right) \right\}
\]

\( D_m = \dim(S_m), \ E_m = \sum_{j/(j,k) \in \mathcal{L}_m} 2^j. \)

Then \( \Phi^{FLR}_{\alpha} \) is of level \( \alpha \) and \( P_f(\Phi^{FLR}_{\alpha} = 0) \leq \beta \) as soon as

\[
d_2^2(P_f, P_0) > \inf_{m \in \mathcal{M}} \left\{ \| f - \Pi_{S_m}(f) \|_2^2 + C \left( \frac{\sqrt{D_m \ln(1/(w_m \alpha))}}{n} + \frac{\ln(1/(w_m \alpha))}{n} + \frac{E_m \ln^2(1/(w_m \alpha))}{n^2} \right) \right\}
\]

**Probabilistic tools:** concentration inequalities (\( U \) statistics of order 2)
In the Poisson model ([UD1])

\[ D_m = \dim(S_m), \quad E_m = \sum_{j/(j,k) \in \mathcal{L}_m} 2^j. \]

Then \( \Phi_{\alpha}^{FLR} \) is of level \( \alpha \) and \( P_f(\Phi_{\alpha}^{FLR} = 0) \leq \beta \) as soon as

\[ d_2^2(P_f, P_0) > \inf_{m \in \mathcal{M}} \left\{ \| f - \Pi_{S_m}(f) \|_2^2 + C \left( \frac{\sqrt{D_m \ln(1/(w_m \alpha))}}{n} + \frac{\ln(1/(w_m \alpha))}{n} + \frac{E_m \ln^2(1/(w_m \alpha))}{n^2} \right) \right\} \]

**Which choice for \( \{ S_m, \ m \in \mathcal{M} \} \) and \( (w_m)_{m \in \mathcal{M}} \)?**

- **Classical collection of nested spaces:** allows to detect intensities in \( \mathcal{B}_{s,2,\infty}(R) \), \( E_m \approx D_m \Rightarrow w_m = 1/\# \mathcal{M} \) possible \( \Rightarrow \Phi_{\alpha}^{FLR,\text{nest}} \) minimax adaptive with a loss \( \sim \sqrt{\ln \ln n} \) factor.

- **Need for a richer collection of nonnested spaces** to detect intensities in \( \mathcal{B}_{s,2,\infty}(R) \cap w\mathcal{B}_{s'}(R') \) with \( s \geq s'/(2s' + 1), s' > 1/2 \Rightarrow \# \mathcal{M} \) large \( \Rightarrow \) other choice for \( w_m \Rightarrow \Phi_{\alpha}^{FLR,\text{nonnest}} \) minimax adaptive without any loss.
Aggregated tests: goodness-of-fit

In the Poisson model ([UD1])

\[
\Phi_{\alpha}^{FLR,\text{nest}} \quad \text{(loss } \sqrt{\ln \ln n})
\]

\[
(i) \quad n^{-\frac{2s}{4s+1}}
\]

\[
(ii) \quad \left(\frac{\ln n}{n}\right)^{\frac{s'}{2s'+1}}
\]

\[
(iii) \quad s = s' \sqrt{\ln \ln n}
\]

\[
s = \frac{s'}{2s'+1}
\]
Aggregated tests: two-sample problems

In the Poisson model ([UD2])

Poisson model

\[ X = (X^1, X^2) \] is a pair of independent Poisson processes, observed on \( \mathbb{X} \subset \mathbb{R}^d \), with resp. intensities \( f_1 \) and \( f_2 \) (in \( L_1(\mathbb{X}, \lambda) \cap L_\infty(\mathbb{X}) \)), w.r.t \( d\mu = nd\lambda \).

\[ P_{f_1, f_2} = \text{joint distribution of} \ X = (X^1, X^2). \]

Motivations

- Differential analysis of replication origins peaks
- Spatial representativeness of services in public statistics

Notations

\[ X^1 = \{ X^1_1, \ldots, X^1_{N_1} \}, \ X^2 = \{ X^2_1, \ldots, X^2_{N_2} \} \] (\( N_1, N_2 \) random),

\[ \bar{X} = X^1 \cup X^2 = \{ X_1, \ldots, X_N \} \text{ with } N = N_1 + N_2. \]
Aggregated tests: two-sample problems

In the Poisson model ([UD2])

Single kernel based test

Considering as above a subspace $S_m = \langle b_l, l \in \mathcal{L}_m \rangle$ (orthonormal basis) of $\mathbb{L}_2(\mathbb{X}, \lambda)$, a natural idea is to introduce $(H_{0,m}) P_f \in \mathcal{P}_{0,m}$, with $\mathcal{P}_{0,m} = \{ P_f, \Pi_{S_m}(f_1 - f_2) = 0 \} \supset \mathcal{P}_0$.

Unbiased estimator of $n^2 \| \Pi_{S_m}(f_1 - f_2) \|^2_2$:

$$T_m = \sum_{i \neq j=1}^{N} \left( \sum_{l \in \mathcal{L}_m} b_l(X_i)b_l(X_j) \right) \varepsilon_i^0 \varepsilon_j^0,$$

where

$$\varepsilon_i^0 = 1 \text{ if } X_i \in \mathbb{X}^1,$$

$$\varepsilon_i^0 = -1 \text{ if } X_i \in \mathbb{X}^2.$$

Generalization to $T_m = \sum_{i \neq j=1}^{N} K_m(X_i, X_j) \varepsilon_i^0 \varepsilon_j^0$, where $K_m$ is a symmetric kernel s. t. $\int K_m^2(x, x')(f_1 + f_2)(x)(f_1 + f_2)(x')d\nu(x)d\nu(x') \leq D_m$

Unbiased estimator of $n^2 \langle K_m[f_1 - f_2], f_1 - f_2 \rangle_2$ with $K_m[f](x) = \langle K_m(., x), f \rangle_2$
Aggregated tests: two-sample problems

In the Poisson model ([UD2])

Possible choices for the kernel

- **[PK] projection kernel** $K_m(x, x') = \sum_{l \in L_m} b_l(x) b_l(x')$,
  $$\langle K_m[f_1 - f_2], f_1 - f_2 \rangle_2 = \|\Pi_{S_m}(f_1 - f_2)\|_2^2$$

- **[AK] approximation kernel** $K_m(x, x') = k_m\left(\frac{x_1 - x'_1}{h_1}, \ldots, \frac{x_d - x'_d}{h_d}\right)/\prod_{i=1}^d h_i$,
  $$\langle K_m[f_1 - f_2], f_1 - f_2 \rangle_2 = \langle k_m * (f_1 - f_2), f_1 - f_2 \rangle_2 / \prod_{i=1}^d h_i$$

- **[RK] reproducing kernel** $K_m(x, x') = \langle \theta_{K_m}(x), \theta_{K_m}(x') \rangle_{\mathcal{H}_{K_m}}$, $\theta_{K_m}$ and $\mathcal{H}_{K_m}$ feature function and RKHS space,
  $$\langle K_m[f_1 - f_2], f_1 - f_2 \rangle_2 = \|K_m[f_1] - K_m[f_2]\|_{\mathcal{H}_{K_m}}^2$$
  mean embeddings of $f_1$ and $f_2$ in the RKHS if they are densities.

**Single test:** $\phi_{m, \alpha} = 1\{T_m > q_m(1 - \alpha)\}$, $q_m$ to define

**Problem:** the distribution of $T_m$ is not free from $f_1 = f_2$ under $(H_0)$!

**Wild bootstrap approach**  

Wild bootstrap approach in the density model

- Classical Efron’s bootstrap
  - Empirical process: \((P_n - P)(h) \xrightarrow{d} (P_n^* - P_n)(h) = \frac{1}{n} \sum_{i=1}^{n} (M_{n,i} - 1) h(X_i)\)
    - \(\cong\) Giné, Zinn (1990, 1992)
  - Degenerate \(U\)-statistics: \(U_n(h) = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j)\)
    - \(\Rightarrow\) \(U_n^*(h) = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j)(M_{n,i} - 1)(M_{n,j} - 1)\)
    - \(\cong\) Arcones, Giné (1992)

- Wild bootstrap based on i.i.d. Rademacher variables \((\varepsilon_1, \ldots, \varepsilon_n)\)
  - Empirical process: \((P_n - P)(h) \xrightarrow{d} (P_n^* - P_n)(h) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(X_i)\)
    - \(\cong\) Mammen (1992)
  - Degenerate \(U\)-statistics: \(U_n(h) = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j)\)
    - \(\Rightarrow\) Fromont, Mach. Learn. (2007) for nonasymptotic results
  - \(\cong\) \(U_n^*(h) = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j)\varepsilon_i\varepsilon_j\)
    - \(\cong\) Dehling Mikosch (1994)
Wild bootstrap approach in the Poisson model

\[ T^*_m = \sum_{i \neq j} K_m(X_i, X_j) \varepsilon_i \varepsilon_j \Rightarrow \mathcal{L}(T^*_m|\bar{X}) = \mathcal{L}(H_0)(T_m|\bar{X}) \quad [UD2] \]

\[ q_m = q_{\bar{X}} = \text{quantile function of } \mathcal{L}(T^*_m|\bar{X}) \text{ (Monte Carlo)} \]

\[ \phi_{m,\alpha} = 1 \{ T_m > q_{\bar{X}}(1-\alpha) \} \text{ is of level } \alpha, \]

even when \( q_{\bar{X}}(1-\alpha) \) is approximated by a Monte Carlo method!


Tool: key exchangeability lemma

\( \equiv \) Romano, Wolf (2005)
Aggregated tests: two-sample problems
In the Poisson model ([UD2])

Aggregated two-sample tests

- Collection of subsets of $\mathcal{P}$:
  \[ \{ \mathcal{P}_{0,m} = \{ P_{f_1, f_2}, \langle K_m[f_1 - f_2], f_1 - f_2 \rangle_2 = 0 \}, m \in \mathcal{M} \} \]
- Collection of hypotheses: \( \{ (H_{0,m}) \ P \in \mathcal{P}_{0,m}, \ m \in \mathcal{M} \} \)
- Collection of single tests: \( \{ \phi_{m,\alpha} = 1 \{ T_m > q_m(1 - \alpha) \}, \ m \in \mathcal{M} \} \)
- Collection of individual levels: \( \{ u_{m,\alpha}, \ m \in \mathcal{M} \} \)

FLR choice: \( u_{m,\alpha} = u_{\bar{X}, m,\alpha} \), with
\[
\bar{X}_{m,\alpha} = w_m \sup \left\{ u, \mathbb{P}(H_0) \left( \exists m \in \mathcal{M}, T_m^* > q_m(1 - w_m u) \middle| \bar{X} \right) \leq \alpha \right\},
\]

\[
\Phi_{FLR\alpha} = \sup_{m \in \mathcal{M}} \phi_{m, u_{m,\alpha}} = \sup_{m \in \mathcal{M}} 1 \{ T_m > q_m(1 - u_{m,\alpha}) \}
\]
In the Poisson model ([UD2])

Oracle type result

The test $\Phi^{FLR}_\alpha$ is of level $\alpha$ and $P_{f_1,f_2}(\Phi^{FLR}_\alpha = 0) \leq \beta$, as soon as

$$\|f_1 - f_2\|_2^2 \geq \inf_{m \in M} \inf_{r > 0} \left\{ \| (f_1 - f_2) - r^{-1} K_m[f_1 - f_2] \|_2^2 + C \left( \frac{\sqrt{D_m} \ln(1/(w_m \alpha))}{rn} \right) \right\}.$$ 


Tools: concentration inequalities & exponential inequalities for Rademacher chaos

Minimax adaptivity properties over:

- $\{ P_{f_1,f_2}, (f_1 - f_2) \in B_{s,2,\infty}(R) \cap wB_s(R'), f_1, f_2 \in L_\infty(R'') \}$ (loss $\sim (\ln \ln n)$ in the case (i), no loss in the case (ii))
- subsets based on $d$ dim. Sobolev and anisotropic Nikol’skii-Besov balls (loss $\sim (\ln \ln n)$)

⇒ Parametric rate for the single tests based on characteristic kernels for the weak distance $\| K_m[f_1] - K_m[f_2] \|_{\mathcal{H}_m} \Rightarrow$ choice of the distance?
Aggregated tests: two-sample problems

In the density model ([$UD3$])

**Density model**

$X = (X^1, X^2)$ is a pair of independent sets of i.i.d. random variables, with respective densities $f_1$ and $f_2$, w.r.t. $\lambda$.

\[
(H_0) \quad f_1 = f_2 \iff P(f_1, f_2) \in \mathcal{P}_0 = \{ P(f_1, f_2), \ f_1 = f_2 \} \quad \text{against} \quad (H_1) \quad P(f_1, f_2) \notin \mathcal{P}_0
\]

**Aggregated tests based on kernels as in the Poisson process model**

\[T_m = \sum_{i \neq j=1}^N K_m(X_i, X_j) \varepsilon_i^0 \varepsilon_j^0,\]

where if $c_{N_1, N_2} = 1/(N_1N_2(N_1 + N_2 + 2))$,
\[
\varepsilon_i^0 = a_{N_1, N_2} = (1/(N_1(N_1 - 1)) - c_{N_1, N_2})^{1/2} \quad \text{if} \ X_i \in X^1,
\]
\[
\varepsilon_i^0 = b_{N_1, N_2} = -a_{N_2, N_1} \quad \text{if} \ X_i \in X^2.
\]

$T_m + c_{N_1, N_2} \sum_{i \neq j=1}^N K_m(X_i, X_j)$ unbiased estimator of $\langle K_m[f_1 - f_2], f_1 - f_2 \rangle_2$


**Another kind of possible (nonsymmetric) kernel based on $k_m$ nearest neighbors**:

$K_m(x, x') = 1_{\{x' \text{ $k_m$-nn of } x\}}$, with other marks

- less complex collections $\Rightarrow$ possible extension to functional data

Aggregated tests: two-sample problems

In the density model ([UD3])

**Bootstrap approach**

Wild bootstrap \( \Rightarrow \) asymptotically valid in the density model, but

Permutation \( \Rightarrow \) "exact" bootstrap approach in the density model

\[
\varepsilon_i = a_{N_1,N_2} \text{ if } \Pi_N(i) \in \{1, \ldots, N_1\}, \\
\varepsilon_i = b_{N_1,N_2} \text{ if } \Pi_N(i) \in \{N_1 + 1, \ldots, N\},
\]

\( \Pi_N \) random permutation uniformly distributed on \( \mathcal{S}_N \).

\[
T^*_m = \sum_{i \neq j} K_m(X_i, X_j)\varepsilon_i\varepsilon_j \quad \Rightarrow \quad \mathcal{L}(H_0)(T^*_m|\bar{X}) = \mathcal{L}(H_0)(T_m|\bar{X}) \quad [UD3]
\]

\( q_m = q^\bar{X}_m = \text{quantile function of } \mathcal{L}(T^*_m|\bar{X}) \) (Monte Carlo)

\( \phi_{m,\alpha} = 1\{T_m > q^\bar{X}_m(1-\alpha)\} \) is of level \( \alpha \),

even when \( q^\bar{X}_m(1-\alpha) \) is approximated by a Monte Carlo method!
Multiple tests
Parallel between aggregated tests and multiple tests

Collection of hypotheses: \( \{(H_{0,m}) \ P \in \mathcal{P}_{0,m}, \ m \in \mathcal{M}\} \)

Aggregated tests in the case \([KD]\)

\[
\text{Testing} \quad (H_0) \quad P \in \mathcal{P}_0 \subset \cap_{m \in \mathcal{M}} \mathcal{P}_{0,m} \quad \text{against} \quad (H_1) \quad P \notin \mathcal{P}_0
\]

Minimax adaptive level \(\alpha\) aggregated tests: \(\bar{\Phi}_{\alpha}^{\text{Bonf}}, \bar{\Phi}_{\alpha}^{\text{FL}}\) or \(\bar{\Phi}_{\alpha}^{\text{FLR}}\)

Multiple tests

\[
\text{Testing} \quad (H_{0,m}) \quad P \in \mathcal{P}_{0,m} \quad \text{simultaneously}
\]

Multiple tests whose FWER \(\leq \alpha\): \(R^{\text{Bonf}}, R^{\text{Holm}}, \) or \(R^{\text{minp}}\)

Under specific conditions,
\[
\bar{\Phi}_{\alpha}^{\text{Bonf}} = 1\{R^{\text{Bonf}} \neq \emptyset\} = 1\{R^{\text{Holm}} \neq \emptyset\} \quad \text{and} \quad \bar{\Phi}_{\alpha}^{\text{FL}} = 1\{R^{\text{minp}} \neq \emptyset\}
\]

Multiple tests
Multiple tests designed for particular concrete challenges

**Example:** Detecting synchronization periods between neural spike trains

→ **Multiple test based on permutation independence tests for point processes**

Case \([UD2]\)


**Perspectives:** Aggregation, study from the minimax point of view?

**Introduction of a minimax theory for multiple tests**


Allows to prove that when they are based on strongly dependent \(p\) values, \(R^{Bonf}\) can be clearly suboptimal, whereas \(R^{minp}\) is minimax adaptive...
**Conclusion**

Aggregated or multiple tests based on a collection of single tests, defined from test statistics $T_m$ (or $p$-values $p_m$) and associated critical values obtained from Monte Carlo or resampling methods, that are justified from a nonasymptotic point of view

⇒ implementable and adapted to moderate sample sizes

---

**[KD]**  (Known Distr.)

$\mathcal{L}(H_0)(T_m)$ is known (parameter free)

**[UD1]** $\mathcal{L}(H_0)(T_m|Z)$ is known

**[UD]**  (Unknown Distr.)

$\mathcal{L}(H_0)(T_m|Z)$ is known

**[UD2]** $\exists T^*_m, \mathcal{L}(T^*_m|Z) = \mathcal{L}(H_0)(T_m|Z)$

**[UD3]** $\exists T^*_m, \mathcal{L}(H_0)(T^*_m|Z) = \mathcal{L}(H_0)(T_m|Z)$
Conclusion

[KD] Goodness-of-fit tests in the density model
  Detection of atmospheric nitrogen deposition in ecology

[KD] Periodic signal detection tests in a regression model
  Target detection in laser vibrometry

[UD1] Homogenity tests in the Poisson model
  Detection of epidemics

[UD2] Two-sample tests in the Poisson model
  Differential analysis of replication origins peaks in genetics
  Spatial representativeness of services in public statistics

[UD2] Independence tests for point processes
  Detection of dependence periods between spike trains in neuroscience

[UD3] Two-sample tests in density and regression models
  Comparison of functional data (in progress)