

# Etude d'un modèle semi-paramétrique de contamination.

Pierre Vandekerkhove

Laboratoire d'Analyse et de Mathématiques Appliquées de  
l'université Paris-Est Marne-la-Vallée - CNRS UMR 8050

ENPC – 1er avril 2010

- Introduction sur les mélanges semi-paramétriques.
- Identifiabilité du modèle de contamination.
- Estimation, hypothèses, identifiabilité bis.
- Résultats asymptotiques : convergence et TCL fonctionnel.
- Technique de preuve :  $\delta$ -méthode.
- Application aux puces ADN.
- Tests d'adéquation.

Mélange de lois sur  $\mathbb{R}^p$  : fdr  $G$  définie par

$$G(\mathbf{x}) = \sum_{i=1}^k \lambda_i F_i(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^p.$$

Approche semi-paramétrique : Hall et Zhou (2003) pour  $k = 2$   
et

$$F_i(\mathbf{x}) = \prod_{j=1}^p F_{ij}(x_j), \quad \mathbf{x} = (x_1, \dots, x_p) \in \mathbb{R}^p, \quad i = 1, 2.$$

↔ Modèle identifiable pour  $p \geq 3$ .

**Idée** : estimation Bootstrap des fdr (Barbe & Bertail, 1995, et Hall & Presnel, 1999).

Dans [Bordes, Mottelet et Vandekerkhove \(2006\)](#), le mélange

$$g(x) = (1 - p)f(x - \mu_1) + pf(x - \mu_2), \quad x \in \mathbb{R}$$

est **identifiable** si  $p \in ]0, 1/2[$ ,  $\mu_1 \neq \mu_2$  et  $f \in \mathcal{F}$  (classe des densités symétriques).

Voir aussi [Hunter, Wang & Hettmansperger \(2007\)](#).

**Rappel** : soit  $\vartheta := (p, \mu_1, \mu_2) \in \Theta$ . On dit que le modèle ci-dessus est **identifiable semi-paramétriquement** sur  $\Theta \times \mathcal{F}$  si pour  $\vartheta = (p, \mu_1, \mu_2) \in \Theta$ ,  $\vartheta' = (p', \mu'_1, \mu'_2) \in \Theta$  and  $(f, f') \in \mathcal{F}^2$

$$(1 - p)f(\cdot - \mu_1) + pf(\cdot - \mu_2) = (1 - p')f_0(\cdot - \mu'_1) + p'f'(\cdot - \mu'_2) \quad \lambda - \text{p.p.},$$

nous avons  $\vartheta = \vartheta'$  et  $f = f'$   $\lambda$ -p.p. sur  $\mathbb{R}$ .

Dans Bordes, Delmas, et Vandekerkhove (2006), pour  $f_0$  connue, le mélange

### Modèle de contamination

$$g(x) = (1 - p)f_0(x) + pf(x - \mu), \quad x \in \mathbb{R},$$

quand  $f$  est la densité d'une distribution symétrique, n'est pas toujours identifiable.

**Rappel** : soit  $\vartheta := (p, \mu) \in \Theta$ . On dit que le modèle de contamination est **identifiable semi-paramétriquement** sur  $\Theta \times \mathcal{F}$  si pour  $\vartheta = (p, \mu) \in \Theta$ ,  $\vartheta' = (p', \mu') \in \Theta$  and  $(f, f') \in \mathcal{F}^2$

$$(1 - p)f_0(\cdot) + pf(\cdot - \mu) = (1 - p')f_0(\cdot) + p'f'(\cdot - \mu') \quad \lambda - \text{p.p.},$$

nous avons  $\vartheta = \vartheta'$  et  $f = f'$   $\lambda$ -p.p. sur  $\mathbb{R}$ .

# Identifiabilité du modèle de contamination

- $\vartheta_0 = (\rho_0, \mu_0)$  la vraie valeur du paramètre Euclidien  $\vartheta = (\rho, \mu)$ .
- On note  $m_0$  et  $m$  les moments d'ordre 2 de  $f_0$  et  $f$ .
- Soit l'ensemble

$$\Phi = \mathbb{R}^* \times ]0, +\infty[ \setminus \bigcup_{k \in \mathbb{N}^*} \Phi_k$$

$$\text{où } \Phi_k = \left\{ (\mu, m) \in \mathbb{R}^* \times ]0, +\infty[; m = m_0 + \mu^2 \frac{k+2}{3k} \right\}.$$

- On définit  $\mathcal{F}_q = \{f \in \mathcal{F}; \int_{\mathbb{R}} |x|^q f(x) dx < +\infty\} \subset \mathcal{F}$  pour  $q \geq 1$ .

**Conditions d'identifiabilité** :  $(f_0, f) \in \mathcal{F}_3^2$ ,  $\bar{f}_0 > 0$  et  $(\mu_0, m) \in \Phi_c$  où  $\Phi_c$  est un sous-ensemble compact de  $\Phi$ . De plus  $\vartheta_0 = (\rho_0, \mu_0) \in \Theta$  où  $\Theta$  est un sous-ensemble compact de  $(0, 1) \times \Xi$  où  $\Xi = \{\mu; (\mu, m) \in \Phi_c\}$ .

On remarque que

$$F(x) = \frac{1}{p} (G(x + \mu) - (1 - p)F_0(x + \mu)), \quad \forall x \in \mathbb{R}. \quad (1)$$

Comme  $F$  est la fdr d'une distribution symétrique, nous avons

$$F(x) = 1 - F(-x), \quad \forall x \in \mathbb{R}.$$

Soit  $\vartheta_0 = (p_0, \mu_0)$  est la vraie valeur du paramètre  $\vartheta$  inconnu.

On introduit pour  $x \in \mathbb{R}$ , les fonctions

$$H_1(x; \vartheta, G) = \frac{1}{p} G(x + \mu) - \frac{1 - p}{p} F_0(x + \mu),$$

$$H_2(x; \vartheta, G) = 1 - \frac{1}{p} G(-x + \mu) + \frac{1 - p}{p} F_0(-x + \mu).$$

Alors, en utilisant (1) et la symétrie de  $F$ ,

$$H(x; \vartheta_0, G) \equiv H_1(x; \vartheta_0, G) - H_2(x; \vartheta_0, G) = 0 \quad \forall x \in \mathbb{R}. \quad (2)$$

On considère le contraste

$$d(\vartheta) = \int_{\mathbb{R}} H^2(x; \vartheta, G) dG(x),$$

où clairement  $d(\vartheta) \geq 0$  pour tout  $\vartheta \in \Theta$  et  $d(\vartheta_0) = 0$ . La fdr  $G$  étant inconnue on la remplace par des estimations (empirique et régularisée) :

$$d_n(\vartheta) = \int_{\mathbb{R}} H^2(x; \vartheta, \tilde{G}_n) d\hat{G}_n(x) = \frac{1}{n} \sum_{i=1}^n H^2(X_i; \vartheta, \tilde{G}_n) \quad (3)$$

où  $\hat{G}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{X_i \leq x}$ , et  $\tilde{G}_n(x) = \int_{-\infty}^x \hat{g}_n(t) dt$ ,  $\forall x \in \mathbb{R}$ ,

où

$$\hat{g}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n q\left(\frac{x - X_i}{h_n}\right), \quad \forall x \in \mathbb{R}.$$

On estime les paramètres Euclidiens par

$$\hat{\vartheta}_n = (\hat{\rho}_n, \hat{\mu}_n) = \arg \min_{\vartheta \in \Theta} d_n(\vartheta).$$

En utilisant la formule d'inversion (1) on peut estimer, pour tout  $x \in \mathbb{R}$ ,  $F$  et  $f$  par :

$$\hat{F}_n(x) = \frac{1}{\hat{\rho}_n} \left( \hat{G}_n(x + \hat{\mu}_n) - (1 - \hat{\rho}_n) F_0(x + \hat{\mu}_n) \right),$$

$$\tilde{f}_n(x) = \frac{1}{\hat{\rho}_n} \left( \hat{g}_n(x + \hat{\mu}_n) - (1 - \hat{\rho}_n) f_0(x + \hat{\mu}_n) \right).$$

ou sa correction  $\hat{f}_n = \frac{1}{s_n} \tilde{f}_n \mathbb{I}_{\tilde{f}_n \geq 0}$ , avec  $s_n = \int_{\mathbb{R}} \tilde{f}_n(x) \mathbb{I}_{\tilde{f}_n(x) \geq 0} dx$ .

## Conditions sur le noyau.

- (i)  $q$  est pair, borné, uniformément continu, de carré intégrable et à variations bornées.
- (ii)  $q$  admet une dérivée au 1-er ordre  $q' \in L^1(\mathbb{R})$  et  $q'(x) \rightarrow 0$  quand  $|x| \rightarrow +\infty$ . De plus si  $\gamma$  est la racine carrée du module de continuité de  $q$ , nous avons

$$\int_0^1 (\log(1/u))^{1/2} d\gamma(u) < \infty.$$

## Conditions sur la fenêtre.

- (i)  $h_n \searrow 0$ ,  $nh_n \rightarrow +\infty$  et  $\sqrt{nh_n^2} = o(1)$ ,
- (ii)  $nh_n/|\log h_n| \rightarrow +\infty$ ,  $|\log h_n|/\log \log n \rightarrow +\infty$  et il existe un nombre réel  $c$  tel que  $h_n \leq ch_{2n}$  pour tout  $n \geq 1$ ,
- (iii)  $|\log h_n|/(nh_n^3) \rightarrow 0$ .

Exemple :  $h_n = n^{-1/4-\delta}$ , avec  $\delta \in (0, 1/8)$ .

# Identifiabilité et solution parasite

Exemple de non-identifiabilité :

$$(1 - p)\varphi(x) + pf(x - \mu) = (1 - \frac{p}{2})\varphi(x) + \frac{p}{2}\varphi(x - 2\mu), \quad \forall x \in \mathbb{R},$$

où  $\varphi$  est une densité paire,  $p \in (0, 1)$ , et pour tout  $a \in \mathbb{R}$

$$f(x) = (\varphi(x - a) + \varphi(x + a))/2.$$

**Situation parasite** : la valeur  $\mu = 0$  doit être rejetée de l'espace paramétrique.

En effet pour tout  $p \in (0, 1)$ ,  $\vartheta = (p, 0)$  est toujours solution de  $d(\vartheta) = 0$ .

Pour tout  $p \in (0, 1)$  et toute fdr  $G$  issue du modèle de contamination, nous avons :

$$H_1(x; (p, 0), G) = F_0(x), \quad H_2(x; (p, 0), G) = 1 - F_0(-x) = F_0(x).$$

## Lemme préliminaire

Sous les conditions d'identifiabilité, et si  $\Theta$  est un sous-ensemble compact de  $(0, 1) \times \Phi_c$ , nous avons alors :

- (i) la fonction  $d$  est continue sur  $\Theta$  ;
- (ii) si  $G$  est strictement croissante sur  $\mathbb{R}$  alors  $d$  est une fonction de contraste ;
- (iii) si  $F_0$  et  $F$  sont Lipschitz sur  $\mathbb{R}$ , alors  $d$  est Lipschitz sur tout compact de  $\mathbb{R}^2$  et pour tout  $\alpha > 0$ ,

$$\sup_{\vartheta \in \Phi_c} |d_n(\vartheta) - d(\vartheta)| = o_{p.s.}(n^{-1/2+\alpha});$$

- (iv) si  $\text{supp}(g) = \mathbb{R}$  alors

$$\ddot{d}(\vartheta_0) = 2 \int_{\mathbb{R}} \dot{H}(x; \vartheta_0, G) \dot{H}^T(x; \vartheta_0, G) dG(x) > 0.$$

## Théorème

- (i) Sous les conditions d'identifiabilité, si  $\Theta$  est un sous ensemble compact de  $(0, 1) \times \Phi_c$ , si  $G$  est strictement croissante sur  $\mathbb{R}$ , et que  $F_0$  et  $F$  sont Lipschitz sur  $\mathbb{R}$ , alors

$$\hat{\vartheta}_n \xrightarrow{p.s.} \vartheta_0, \quad \text{quand } n \rightarrow +\infty.$$

- (ii) Si de plus  $F_0$  et  $F$  sont  $\mathcal{C}^2$  avec des dérivées secondes dans  $L^1(\mathbb{R})$ , alors  $|\hat{\vartheta}_n - \vartheta_0| = o_{p.s.}(n^{-1/4+\alpha})$  pour tout  $\alpha > 0$ .

## Théorème

$$\sqrt{n} \left( \hat{\mu}_n - \mu_0, \hat{p}_n - p_0, \hat{F}_n(\cdot) - F(\cdot) \right)^T \rightsquigarrow \mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)^T,$$

dans  $\mathbb{R}^2 \times D(\mathbb{R})$ ,  $D(\mathbb{R})$  désignant l'espace des fonctions cad-lag sur  $\mathbb{R}$  et  $\mathcal{G}$  un processus gaussien de matrice de variance-covariance  $\Gamma$  que nous explicitons et savons estimer.

Par Taylor à l'ordre 1 sur  $\hat{d}_n$  au point  $\vartheta_0$  :

$$\ddot{d}_n(\theta_n^*)\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}\dot{d}_n(\theta_0),$$

$$\sqrt{n} \begin{pmatrix} \hat{p}_n - p_0 \\ \hat{\mu}_n - \mu_0 \\ \hat{F}_n - F \end{pmatrix} = L(\theta_0)\mathcal{J}^{-1}(\theta_0)\sqrt{n} \begin{pmatrix} \mathcal{U}_1(\hat{G}_n) \\ \mathcal{U}_2(\hat{G}_n) \\ \mathcal{U}_3(\hat{G}_n) \end{pmatrix} + o_{p.s.}(1)$$

où  $\mathcal{U}(G) = 0$  et pour une fdr  $V$

$$\mathcal{U}_1(V) = 2 \int_{\mathbb{R}} H(x; \theta_0, V) h_1(x) dV(x)$$

$$\mathcal{U}_2(V) = 2 \int_{\mathbb{R}} H(x; \theta_0, V) h_2(x) dV(x)$$

$$\mathcal{U}_3(V) = V(\cdot + \mu_0) - G(\cdot + \mu_0).$$

- **Théorème de Donsker** :  $\sqrt{n}(\hat{G}_n - G) \rightsquigarrow \mathcal{B}$ , où  $\mathcal{B}$  désigne un processus Gaussien de fonction de corrélation  $\rho(x, y) = G(x \wedge y)(1 - G(x \vee y))$ .

- **Théorème de Donsker** :  $\sqrt{n}(\hat{G}_n - G) \rightsquigarrow \mathcal{B}$ , où  $\mathcal{B}$  désigne un processus Gaussien de fonction de corrélation  $\rho(x, y) = G(x \wedge y)(1 - G(x \vee y))$ .
- **$\delta$ -méthode** :  
 $\sqrt{n}(\mathcal{U}(\hat{G}_n) - \mathcal{U}(G)) = \mathcal{U}'_G(\sqrt{n}(\hat{G}_n - G)) + o_{\mathbb{P}}(1)$  et

$$\sqrt{n}(\mathcal{U}(\hat{G}_n) - \mathcal{U}(G)) \rightsquigarrow \mathcal{U}'_G(\mathcal{B}),$$

où  $\mathcal{U}'_G$  est la différentielle au sens de Hadamard de  $\mathcal{U}$  en  $G$  vérifiant :

$$\left\| \frac{\mathcal{U}(G + tK_t) - \mathcal{U}(G)}{t} - \mathcal{U}'_G(K) \right\| \rightarrow 0, \quad t \rightarrow 0, \quad \forall K_t \rightarrow K.$$

- Les données brutes sont du type  $(A_{ijr})$  représentant le niveau d'expression (concentration de mRNA) lors de la  $r$ -ème répétition, pour le  $i$ -ème gène, sous la condition  $j$ .
- Deux transformations :

$$P_{ijr} = \frac{A_{ijr}}{\sum_{ijr} A_{ijr}}, \text{ et } X_{ijr} = \ln \left( \frac{P_{ijr}}{1 - P_{ijr}} \right).$$

- On pose  $X_{ij} = \frac{1}{R} \sum_{r=1}^R X_{ijr}$  et  $X_i = \frac{1}{JR} \sum_{j=1}^J \sum_{r=1}^R X_{ijr}$ .
- Sous de bonnes conditions expérimentales, il est admis qu'à  $r$  fixé :  $X_{ijr} \sim \mathcal{N}(m_{ij}, \sigma_{ij}^2)$ .

**Problème de test** : différence d'expression du gène  $i$  sous les  $J$  conditions ?

Revient à tester  $\mathbb{H}_0$  :

$$\left\{ m_{ij} = m_i, \sigma_{ij} = \sigma_i, \text{ où } m_i = \frac{1}{J} \sum_{j=1}^J m_{ij}, \text{ et } \sigma_i = \frac{1}{J} \sum_{j=1}^J \sigma_{ij} \right\}.$$

On considère la statistique de test usuelle  $S_i$  :

$$S_i = \frac{RJ(R-1)}{(J-1)} \times \frac{\sum_{j=1}^J (X_{ij} - X_i)^2}{\sum_{j=1}^J \sum_{r=1}^R (X_{ijr} - X_{ij})^2}.$$

Sous  $\mathbb{H}_0$  la statistique  $S_i \sim \mathcal{F}(J-1, JR-J)$ .

Pour  $J=2$  on considère

$$S_i = \frac{X_{i1} - X_{i2}}{\sqrt{\frac{\sum_{r=1}^R (X_{i1r} - X_{i1})^2 + \sum_{r=1}^R (X_{i2r} - X_{i2})^2}{R(R-1)}}} \underset{\mathbb{H}_0}{\sim} \mathcal{T}(2(R-1)).$$

# Lien avec le modèle de contamination

Sous l'alternative  $\mathbb{H}_1 = \{\exists j : m_{ij} \neq m_i \text{ ou } \sigma_{ij} \neq \sigma_i\}$ ,

la loi des  $S_i$  est inconnue !

Ainsi la loi des  $S_i$  peut être modélisée par :

$$g(x) = (1 - p)f_0(x) + pf(x - \mu),$$

où  $p$  est la proportion de statistiques sous  $\mathbb{H}_1$ ,  $f_0$  la densité sous  $\mathbb{H}_0$  (Student ou Fisher) et  $f$  la densité inconnue sous  $\mathbb{H}_1$ . La connaissance de  $(p, \mu)$  et de  $f$  permet de calculer :

$$\begin{aligned}\alpha(i) &= P(\text{gène } i \text{ est différemment exprimé} | S_i = s_i) \\ &= \frac{pf(s_i - \mu)}{(1 - p)f_0(s_i) + pf(s_i - \mu)}\end{aligned}$$

On cherche alors les gènes différemment exprimés parmi les  $[n\hat{p}_n]$  ayant les plus grands  $\hat{\alpha}_n(i)$ .

# Comparaison de 2 modes de gestation chez les bovins

- $n = 10214$  gènes.
- $J = 2$  (2 modes de gestation : naturel et in vitro).
- $R = 10$  répétitions.
- $S_i \sim \mathcal{T}(18)$ .
- $(p, \mu) \in [0.01, 0.1] \times [0.5, 1.5] := \Theta$ .

# Courbes de niveau du contraste empirique

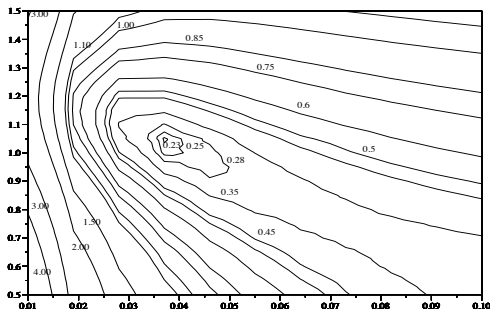


FIG.: Courbe de niveau  $(p, \mu) \mapsto d_n(p, \mu)$  pour  $(p, \mu) \in \Theta$ .

Nous obtenons  $(\hat{p}, \hat{\mu}) = (0.037, 1.05)$  avec  $d_n(\hat{p}, \hat{\mu}) = 0.2257$ .

# Reconstruction du mélange

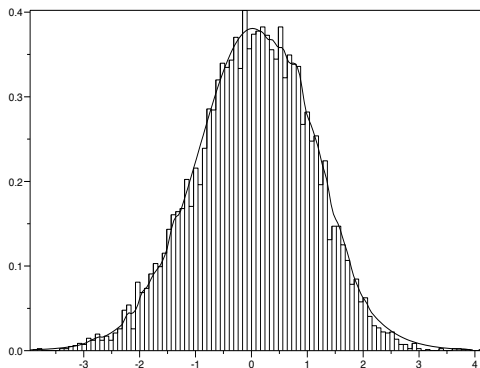
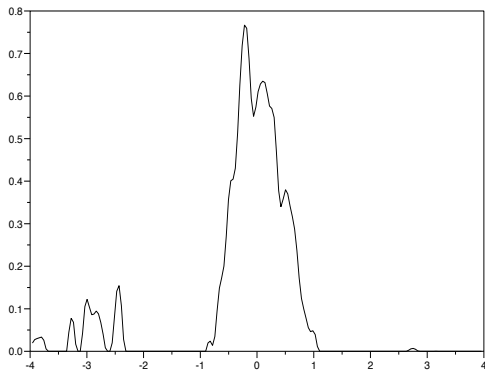


FIG.: Histogramme des données brutes et reconstruction de la loi au moyen de  $(1 - \hat{\rho})f_0(\cdot) + \hat{\rho}\hat{f}(\cdot - \hat{\mu})$ .

# Allure de la densité inconnue



# Application : test sur les paramètres Euclidiens et fonctionnel

On veut tester :

$$\mathbb{H}_0 := \{(\rho, \mu, F) = (\rho_0, \mu_0, F_*)\} \quad \text{vs} \quad \mathbb{H}_1 := \{(\rho, \mu, F) \neq (\rho_0, \mu_0, F_*)\},$$

Méthodologie : on fixe  $k$  réels  $s_1 < \dots < s_k$  tels que  $0 < F_*(s_1) < \dots < F_*(s_k) < 1$ . Soit le vecteur aléatoire

$$W_n = \sqrt{n} \begin{pmatrix} \hat{\mu}_n - \mu_0 \\ \hat{\rho}_n - \rho_0 \\ (\hat{F}_n - F_*)(s_1) \\ \vdots \\ (\hat{F}_n - F_*)(s_k) \end{pmatrix}.$$

D'après le TCL que nous venons d'énoncer :

$$W_n \xrightarrow{\mathcal{D}} \mathcal{N}(0_{\mathbb{R}^{k+2}}, V),$$

où  $V$  est la  $(k+2) \times (k+2)$  matrice de corrélation ayant pour entrées

$$v_{ij} = \Gamma_{ij}, \quad \text{for } 1 \leq i \leq j \leq 2,$$

$$v_{1j} = \Gamma_{13}(s_{j-2}) \text{ et } v_{2j} = \Gamma_{23}(s_{j-2}), \quad \text{pour } 3 \leq j \leq k+2,$$

et

$$v_{ij} = \Gamma_{33}(s_{i-2}, s_{j-2}), \quad \text{pour } 3 \leq i \leq j \leq k+2.$$

En utilisant une estimation consistante  $\hat{V}_n$  de  $V$ , nous avons

$$W_n^T \hat{V}_n^{-1} W_n \xrightarrow{\mathcal{D}} \chi_{k+2}^2,$$

Ainsi on rejète  $\mathbb{H}_0$  au niveau  $\alpha \in (0, 1)$  si  $W_n^T \hat{V}_n^{-1} W_n > \chi_{k+2, \alpha}^2$   
où  $\chi_{k+2, \alpha}^2$  est le quantile d'ordre  $1 - \alpha$  pour la loi du  $\chi_{k+2}^2$ .

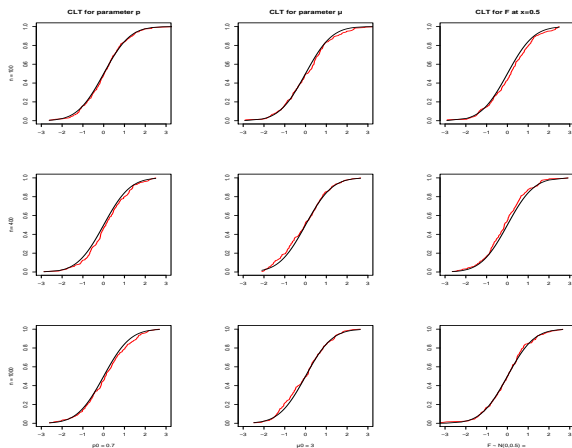
On étudie la qualité d'estimation sur le modèle :

$$G(x) = 0.3\Phi(\cdot) + 0.7\Phi(\cdot - 3/0.5), \quad \text{où } \Phi \text{ fdr de la loi } \mathcal{N}(0, 1).$$

$n$	$p = 0.7$	$\mu = 3$	$F(0.5) = 0.8413$
100	0.7106 (0.0498)	2.9912 (0.0757)	0.8415 (0.0378)
400	0.7048 (0.0277)	2.9959 (0.0355)	0.8390 (0.0177)
1000	0.7018 (0.0167)	2.9977 (0.0225)	0.8409 (0.0107)

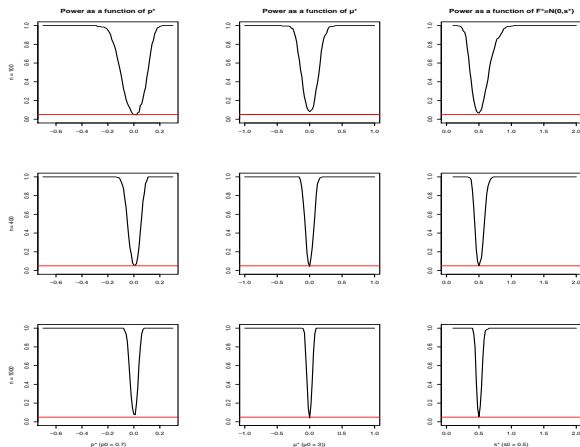
**TAB.:** Étude de Monte Carlo sur 200 estimateurs  $\hat{p}$ ,  $\hat{\mu}$  et  $\hat{F}(0.5)$  pour  $n = 100, 400$  et  $1000$ .

# Visualisation du TCL



**FIG.:** Comparaison de la fdr empirique pour 200 estimateurs de  $p$  (1-ère col.),  $\mu$  (2-ème col.) et  $F(0.5)$  (3-ème col.) avec la fdr  $\mathcal{N}(0, 1)$  et  $n = 100$  (1-ère l.), 400 (2-ème l.) et 1000 (3-ème l.).

# Puissance du test



**FIG.:** Calcul de la puissance pour  $n = 100$  (1-ère l.),  $400$  (2-ème l.) et  $1000$  (3-ème l.) pour tester  $p = p^*$  (1-ère col.),  $\mu = \mu^*$  (2-ème col.) et  $F(0.5) = F^*(0.5) \equiv \Phi(0.5/s^*)$  au niveau  $\alpha = 5\%$ . Sous  $\mathbb{H}_1$  on prend  $(p, \mu, F(0.5)) = (0.7, 3, \Phi(1)) \in ]0, 1[ \times ]2, 4[ \times \Phi(0.5/0.1, 2[$ .

# Application : test de symétrie sur la composante inconnue

On veut tester :

$$\mathbb{H}_0 : F(x) + F(-x) = 1, \text{ pour tout } x \in \mathbb{R}^+, \text{ vs}$$

$$\mathbb{H}_1 : \text{il existe } x \in \mathbb{R}^+ : F(x) + F(-x) \neq 1.$$

Méthodologie : on considère

$$Z_n = \sqrt{n} \begin{pmatrix} \hat{F}_n(-s_1) + \hat{F}_n(s_1) - 1 \\ \vdots \\ \hat{F}_n(-s_k) + \hat{F}_n(s_k) - 1 \end{pmatrix}.$$

Sous  $\mathbb{H}_0$  nous avons  $Z_n = AY_n$  où  $A = (I_k, I_k)$  et  $Y_n$  est le vecteur  $2k$  dimensionnel défini de la manière suivante.

$$Y_n = \sqrt{n} \begin{pmatrix} \hat{F}_n(s_1) - F(s_1) \\ \vdots \\ \hat{F}_n(s_k) - F(s_k) \\ \hat{F}_n(-s_1) - F(-s_1) \\ \vdots \\ \hat{F}_n(-s_k) - F(-s_k) \end{pmatrix}.$$

D'après le TCL fonctionnel

$$Y_n \rightsquigarrow \mathcal{N}(0, \Lambda),$$

où  $\Lambda$  est la  $2k \times 2k$  matrice de corrélation dont les entrées  $\lambda_{ij}$  sont définies par

$$\lambda_{ij} = \Gamma_{33}(s_i^*, s_j^*), \quad \text{for } 1 \leq i, j \leq 2k,$$

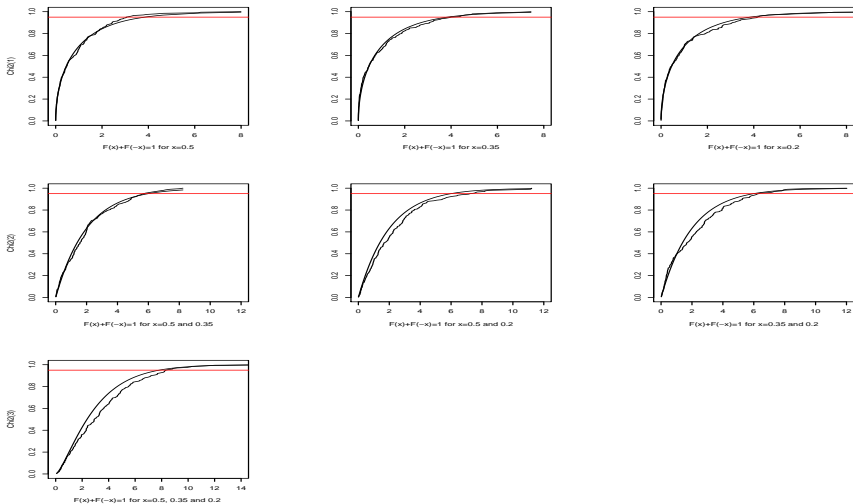
où  $s_i^* = s_i$  pour  $1 \leq i \leq k$  et  $s_i^* = -s_i$  pour  $k+1 \leq i \leq 2k$ .

En remplaçant  $\Lambda$  par  $\hat{\Lambda}$  (c.a.d  $\Gamma$  par  $\hat{\Gamma}$ ), on obtient

$$Z_n^T (A\hat{\Lambda}A^T)^{-1} Z_n \rightsquigarrow \chi_k^2.$$

Ainsi on rejette  $\mathbb{H}_0$  au niveau  $\alpha \in (0, 1)$  si

$$Z_n^T (A\hat{\Lambda}A^T)^{-1} Z_n > \chi_{k,1-\alpha}^2.$$



**FIG.:** fdr empirique de la statistique de test de symétrie calculée pour  $n = 1000$  et répliquée 200 fois. Lignes 1,2,3 : test sur 1,2,3 point(s). La ligne horizontale correspond au niveau 95%.