Troncatures dans les modèles linéaires simples et à effets mixtes sous R

Lyon, 27 et 28 juin 2013

D.Thiam, J.C Thalabard, G.Nuel

Université Paris Descartes
MAP5 UMR CNRS 8145
IRD UMR 216

2èmes ℝencontres 

Motivations

Le modèle linéaire simple

	y	x1	x2
1	14.116	2.785	4.828
2	15.206	3.065	5.185
3	15.179	2.966	5.333
4	15.437	5.128	5.212
5	13.720	2.259	4.891
6	12.583	1.904	5.018
7	13.755	3.038	4.789
8	13.940	3.310	4.785
9	15.310	3.437	5.265
10	12.952	2.542	4.695

- y: outcome, x_1, x_2 covariables
- Le modèle: $Y = X\beta + \varepsilon$
- $(1, X = X_1, X_2)$; $\beta = (\beta_0, \beta_1, \beta_2)$
- $\Theta = (\beta, \sigma)$; σ=`sd`(ε)
- $\Theta = (1, 1, 2, 0.8)$; $n = 1000$

Maximum de vraisemblance

$$L(\Theta; y) = \sum_{1}^{n} f(y_i)$$

$$\widehat{\Theta} = \arg\max_{\Theta} L(\Theta; y)$$

- fonction R `lm`
- `lm`$(y \sim x_1 + x_2)$
- $\widehat{\Theta} = (0.66, 0.98, 2.08, 0.79)$

Motivations

Le modèle linéaire simple

	y	x1	x2
1	14.116	2.785	4.828
2	15.206	3.065	5.185
3	15.179	2.966	5.333
4	15.437	5.128	5.212
5	13.720	2.259	4.891
6	12.583	1.904	5.018
7	13.755	3.038	4.789
8	13.940	3.310	4.785
9	15.310	3.437	5.265
10	12.952	2.542	4.695

- y: outcome, x_1, x_2 covariables
- Le modèle: $Y = X\beta + \varepsilon$
- $(1, X = X_1, X_2)$; $\beta = (\beta_0, \beta_1, \beta_2)$
- $\Theta = (\beta, \sigma)$; σ=`sd`(ε)
- $\Theta = (1, 1, 2, 0.8)$; $n = 1000$

Maximum de vraisemblance

$L =$`sum(dnorm(y,mean=X`β`,sd=`σ`,log=T))`

$\widehat{\Theta} = \arg\max_{\Theta} L$

- fonction R `lm`
- `lm`$(y \sim x_1 + x_2)$
- $\widehat{\Theta} = (0.66, 0.98, 2.08, 0.79)$

Motivations

Le modèle linéaire avec troncatures

	y	x1	x2
1	14.116	2.785	4.828
2	15.206	3.065	5.185
3	15.179	2.966	5.333
4	High	5.128	5.212
5	13.720	2.259	4.891
6	Low	1.904	5.018
7	13.755	3.038	4.789
8	13.940	3.310	4.785
9	15.310	3.437	5.265
10	Low	2.542	4.695

- y: outcome, x_1, x_2 covariables
- $Y = X\beta + \varepsilon$ si $l \leq y_i < u$
- Low=$y_i \leq 13$
- High=$y_i \geq 15.4$

Maximum de vraisemblance

$$L(\Theta; y) = \sum f(y_i) + \sum \Phi(l) + \sum (1 - \Phi(u))$$

- Fonctions R : `tobit`,`censreg`
- `tobit(`$y \sim x_1 + x_2$`,left=13,right=15.4)`
- $\widehat{\Theta} = (0.6751, 0.9298, 2.1153, 0.728)$

Motivations

Le modèle linéaire avec troncatures

	y	x1	x2
1	14.116	2.785	4.828
2	15.206	3.065	5.185
3	15.179	2.966	5.333
4	High	5.128	5.212
5	13.720	2.259	4.891
6	Low	1.904	5.018
7	13.755	3.038	4.789
8	13.940	3.310	4.785
9	15.310	3.437	5.265
10	Low	2.542	4.695

- y: outcome, x_1, x_2 covariables
- $Y = X\beta + \varepsilon$ si $l \leq y_i < u$
- Low=$y_i \leq 13$
- High=$y_i \geq 15.4$

Maximum de vraisemblance

$L =$sum(dnorm(y,mean=$X\beta$,sd=σ,log=T))
+sum(pnorm(t,mean=Xβ,sd=σ,log=T))
+sum(pnorm(u,mean=Xβ,sd=σ,log=T,
lower.tail=F))

- Fonctions R :
  tobit,censreg
- tobit($y \sim x_1 + x_2$,left=13,right=15.4)
- $\widehat{\Theta} = (0.6751, 0.9298, 2.1153, 0.728)$

Motivations

Le modèle linéaire tronqué: extension

	y	x1	x2
1	14.116	2.785	4.828
2	15.206	3.065	5.185
3	15.179	2.966	5.333
4	High_1	5.128	5.212
5	Low_2	2.259	4.891
6	Low_1	1.904	5.018
7	13.755	3.038	4.789
8	13.940	3.310	4.785
9	Low_2	3.437	5.265
10	Low_1	2.542	4.695

- y: outcome, x_1, x_2 covariables
- $Y = X\beta + \varepsilon$ si $l \leq y_i < u$
- $\text{Low}_1 = y_i \leq 13$; $\text{Low}_2 = y_i \leq 13.72$!

Maximum de vraisemblance

$$L(\Theta; y) = \sum f(y_i) + \sum \Phi(l_i) + \sum (1 - \Phi(u_i))$$

Limites `tobit,censreg`

- Pas de seuils multiples
- Résidus calculés par imputation simple

Alternative:

- Implémentation d'un algorithme EM adapté

Motivations

Le modèle linéaire tronqué: extension

	y	x1	x2
1	14.116	2.785	4.828
2	15.206	3.065	5.185
3	15.179	2.966	5.333
4	High_1	5.128	5.212
5	Low_2	2.259	4.891
6	Low_1	1.904	5.018
7	13.755	3.038	4.789
8	13.940	3.310	4.785
9	Low_2	3.437	5.265
10	Low_1	2.542	4.695

- y: outcome, x_1, x_2 covariables
- $Y = X\beta + \varepsilon$ si $l \leq y_i < u$
- $\text{Low}_1 = y_i \leq 13$; $\text{Low}_2 = y_i \leq 13.72$!

Maximum de vraisemblance

$L =$sum(dnorm(y,mean=$X\beta$,sd=σ,log=T))
+sum(pnorm(t_i,mean=Xβ,sd=σ,log=T))
+sum(pnorm(u_i,mean=Xβ,sd=σ,log=T,
lower.tail=F))

Limites `tobit,censreg`

- Pas de seuils multiples
- Résidus calculés par imputation simple

Alternative:

- Implémentation d'un algorithme EM adapté

Notre approche: EM flexible

Algorithe EM

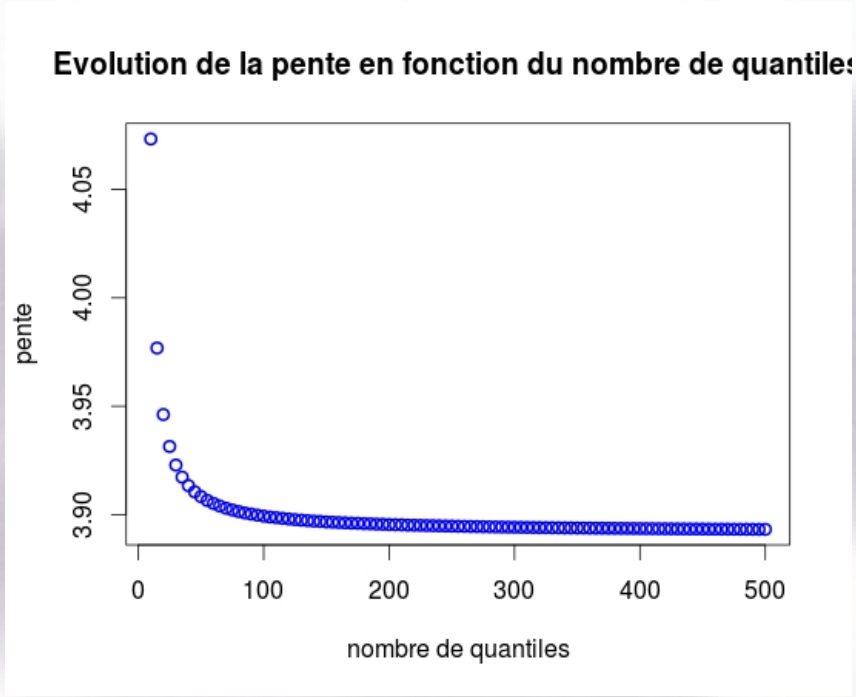
- y^O données observées;
- y^T données tronquées
- $$M(\theta) = \arg\max_{\theta'} \underbrace{\int_{y_T} \mathbb{P}(y_T|y_O;\theta) \log \mathbb{P}(y_O, y_T;\theta') \, \mathrm{dz} \, \mathrm{dy}_T}_{Q(\theta'|\theta)}$$

- $y_T|y_O$ loi normale tronquée
- $\log \mathbb{P}(y_O, y_T;\theta')$ donné par `lm` à y_T fixé.
- $\mathbb{P}(y_T|y_O;\theta)$ joue le rôle de poids

EM flexlible pour Tobit

1) Générer N quantiles de la loi $y_T|y_O$
2) Maximiser la vraisemblance des données complètes avec `lm`+ option `weights`
3) Alterner les deux étapes jusqu'à convergence

Nombre optimal de quantiles



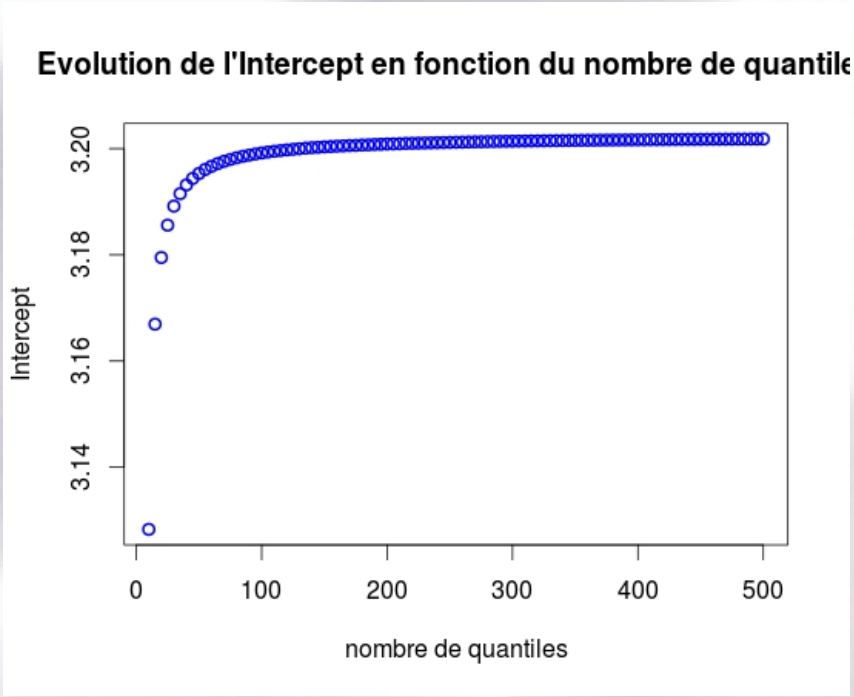


Figure: Etude de la variabilité de l'estimation en fonction du nombre de paramètres

Extension au modèle mixte

Le modèle

- $y_{ij} = \beta x_{ij} + z_i + \varepsilon_{ij}$; (`lmer`)
- $l_{ij} \leq y_{ij} < u_{ij}$
- $z \sim \mathcal{N}(0, \eta)$; $\varepsilon \sim \mathcal{N}(0, \sigma)$

Algorithme EM

$$M(\theta) = \arg\max_{\theta'} \underbrace{\int_z \int_{y_T} \mathbb{P}(z, y_T | y_O; \theta) \log \mathbb{P}(y_O, z, y_T; \theta') \, \mathrm{dz} \, \mathrm{d}y_T}_{Q(\theta'|\theta)}$$

- Introduction de variables latentes continues
  - y_O : données observées
  - y_T variable latente continue: données tronquées
  - z variable latente continue: effets aléatoires
- Approximation de la vraisemblance des données complètes
- Maximisation de cette vraisemblance avec les fonctions R usuelles

Extension au modèle mixte

Remarque:

- y_T fixé $=$ `lmer`
- z fixé $=$ `tobit`

Solution Alernative:

- Modèle Tobit avec effets aléatoires
- Modèle mixte avec troncatures multiples

Solution alternatives

1) EM combiné avec `lmer` du package `lme4`

$$M_{\text{lmer}}(\theta) = \arg\max_{\theta'} \underbrace{\int_{y_T} \mathbb{P}(y_T|y_O;\theta) \log \mathbb{P}(y_O, y_T;\theta') \, \mathrm{d}y_T}_{Q(\theta'|\theta)}$$

2) EM combiné avec `tobit` des packages `censReg` ou `AER`

$$M_{\text{tobit}}(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathbf{z}} \mathbb{P}(z|y_O;\theta) \log \mathbb{P}(y_O, z;\theta') \, \mathrm{d}z}_{Q(\theta'|\theta)}$$

Extension au modèle mixte

Etapes de l'algorithme version `lmer`

- 1) Création des données complètes en remplaçant les données tronquées par N quantiles $q_{ijk} = (q_{ij1}, \ldots, q_{ijN})$
- 2) Estimation des paramètres du modèle mixte avec `lmer`
- 3) Mise à jour des paramètres et répétion jusqu'à convergence

Lois conditionnelles

Pour chaque bloc i: $\forall i$, $y_{i,O}$, $y_{i,T}$:

$y_{i,\bullet}$ est un vecteur Gaussien avec pour loi $\sim \mathcal{N}(\beta x_{i\bullet}, \Sigma)$

$$\Sigma = \begin{pmatrix} \sigma^2 + \eta^2 & \eta^2 & \eta^2 \\ \eta^2 & \ddots & \vdots \\ \vdots & \eta^2 & \sigma^2 + \eta^2 \end{pmatrix}$$

- $y_{i,T} | y_{i,O} \sim \mathcal{N}_t\left(\overline{\mu_i}, \overline{\Sigma}_i, \text{lower} = t\right)$
- $\mathcal{N}_t$ la loi normale tronqué
- $\overline{\mu}_i = \beta x_{i,T} + \Sigma_{i,T,O}\Sigma_{i,O,O}^{-1}\left(y_{O,i} - \mu_{i,O}\right)$
- $\overline{\Sigma}_i = \Sigma_{i,T,T} - \Sigma_{i,T,O}\Sigma_{i,O,O}^{-1}\Sigma_{i,O,T}$

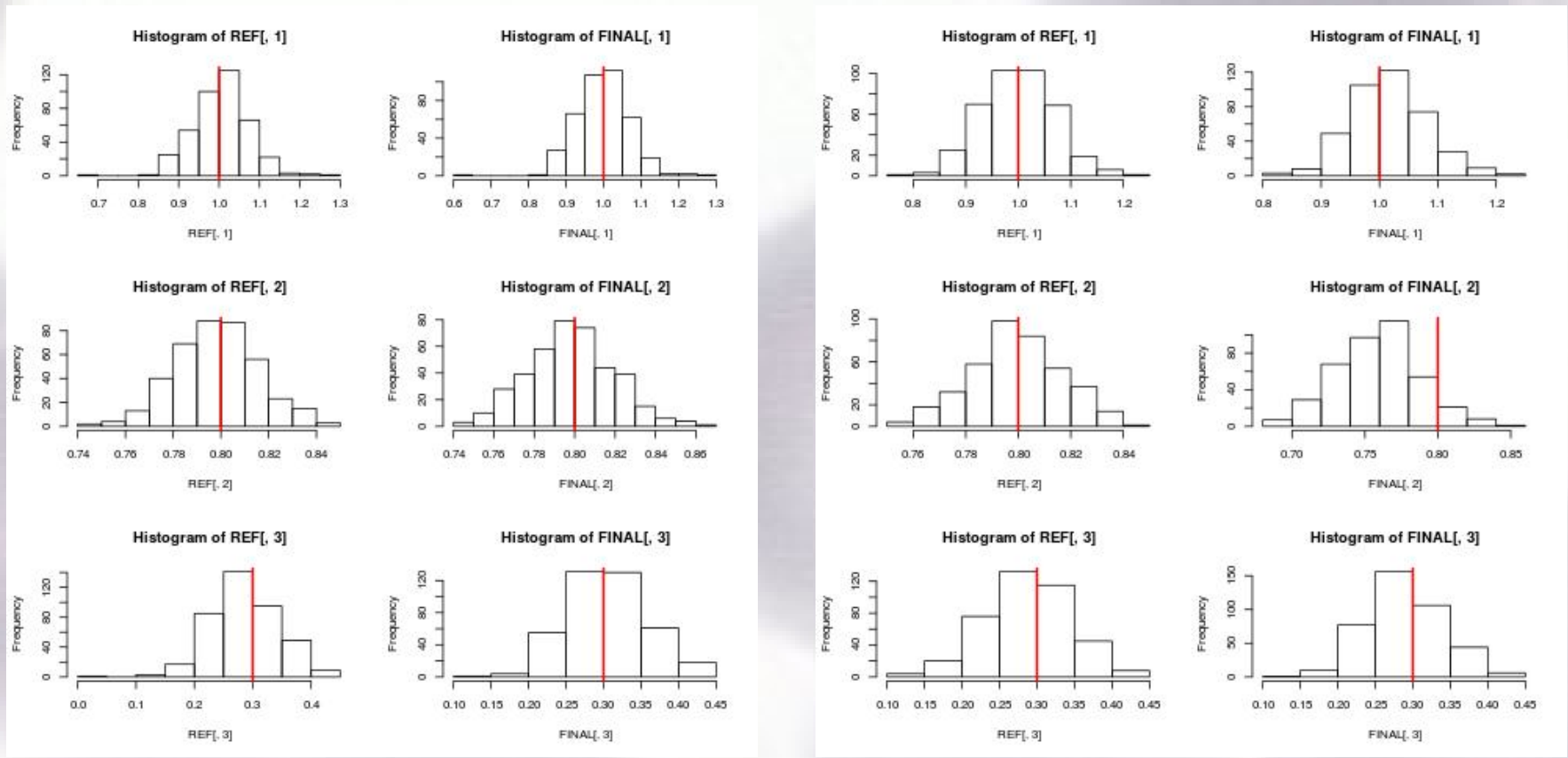
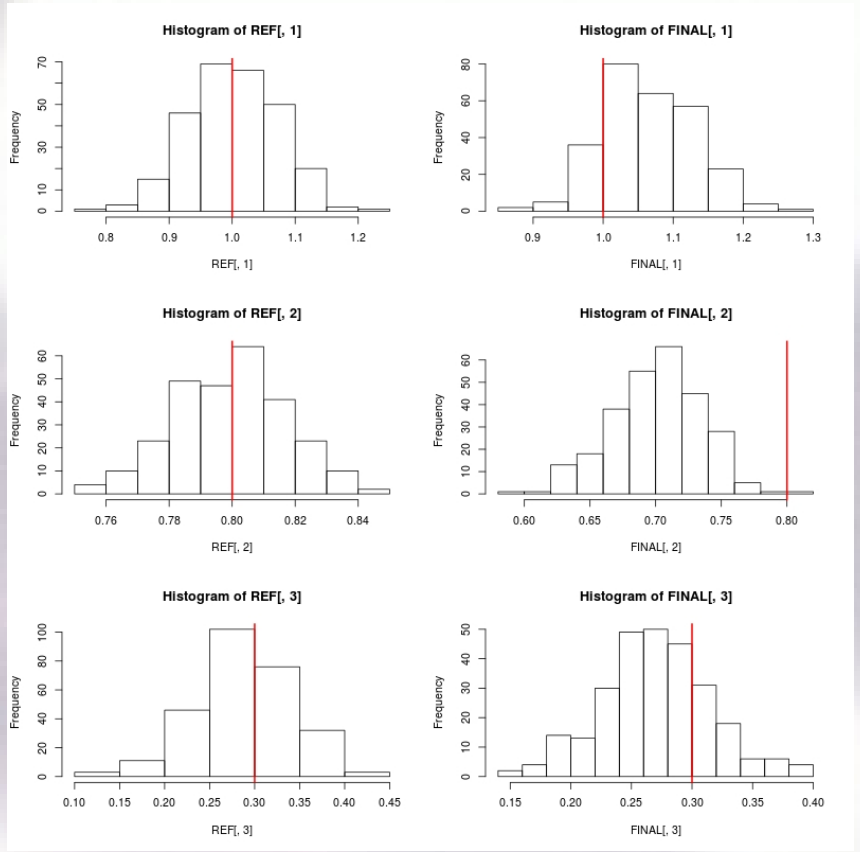


Figure: Comparaison de la stabilité des paramètres en itérant les calculs en modifiant le taux de troncatures. 30% par FEM vs 40% par FEM



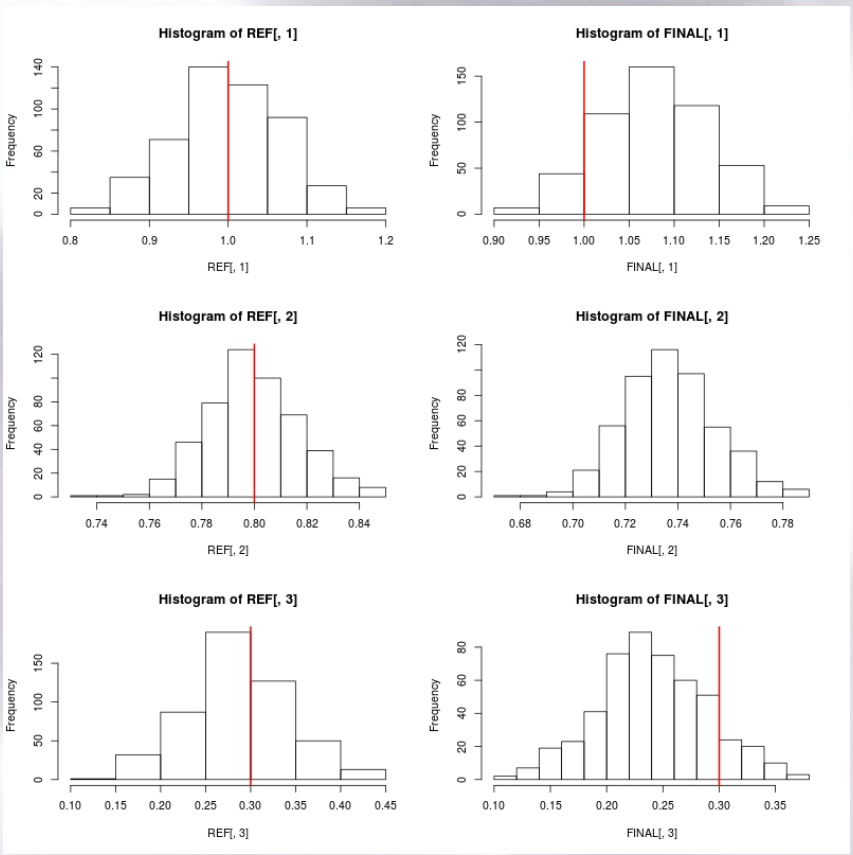


Figure: Comparaison de la stabilité des paramètres en itérant les calculs. 45% de troncatures par FEM; vs 10% de troncatures par imputation simple

Algorithme EM flexible : avantages

- Gestion externe des variables latentes
- Méthodologie flexible
- Utilisation des packages R familiers
- Seuils multiples autorisés
- Convergence rapide

Algorithme EM flexible : extensions

- Modèle linéaire généralisé
- Introduction de classes latentes

Recommendations

- Option `weights`!!

Applications

- Application 1: mesures ELISA en immunologie (UMR 216 IRD)
- Application 2: Méta-analyse sur test de diagnostic à données corrélées avec statut réel du patient (Malade/non malade) latent(J.C. Thalabard).