Algorithme EM régularisé pour données longitudinales

Jocelyn CHAUVET
Catherine TROTTIER, Xavier BRY

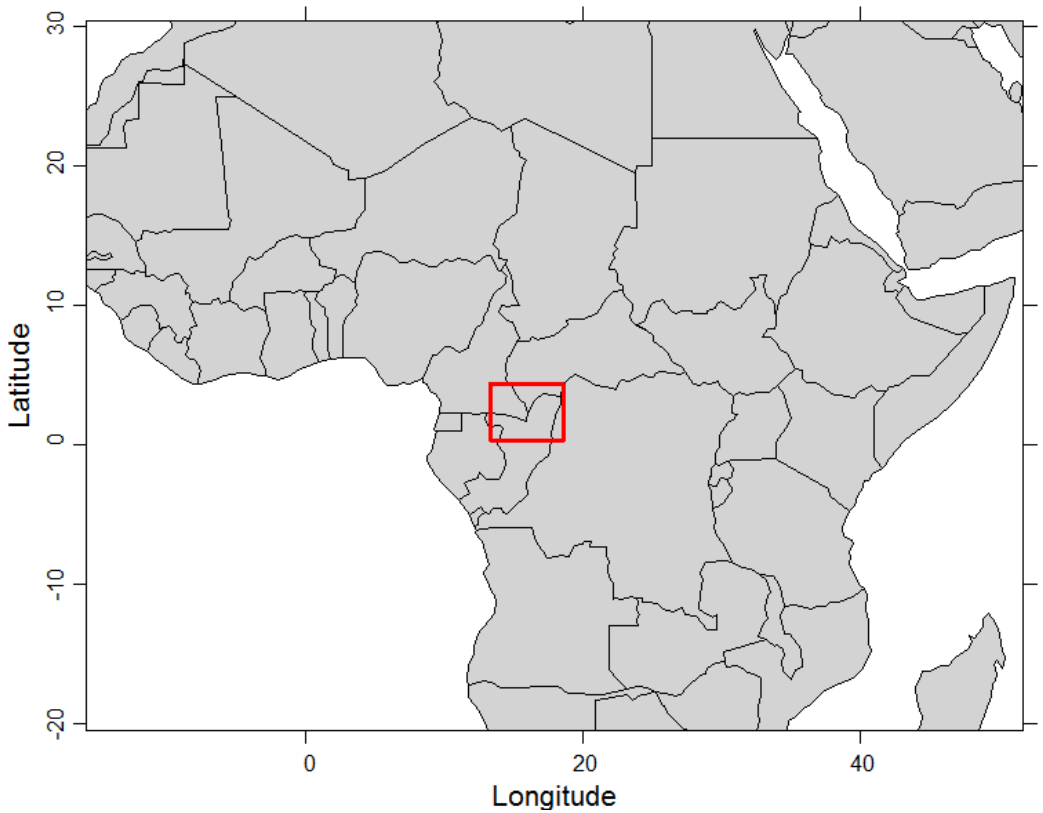
RJS 2017, Porquerolles
6 avril 2017

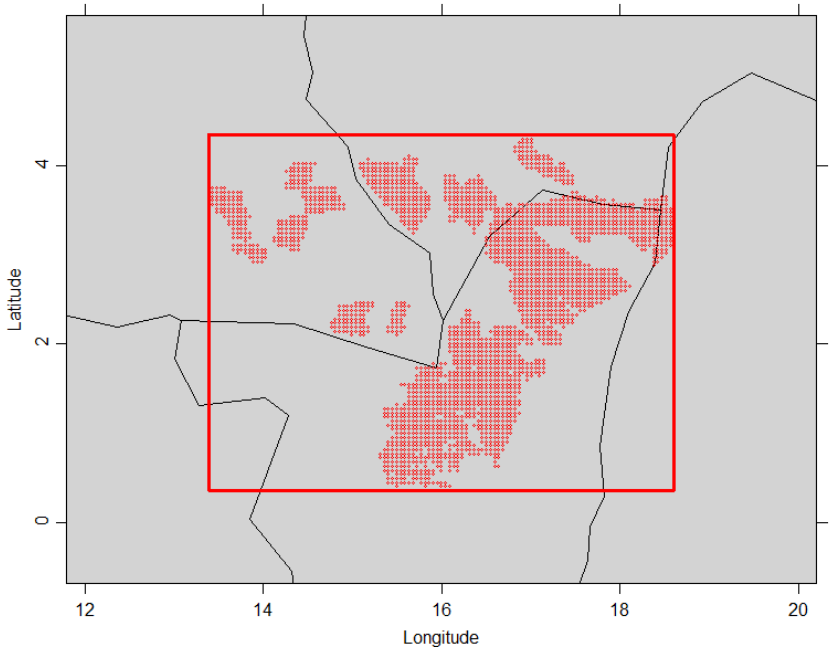
1. Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR)
2. Ridge GL2M pour données longitudinales
3. Bilan et perspectives

1. Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR)
   - Motivation
   - Modèle et méthode d'estimation (première composante)
   - Diagnostics graphiques sur données réelles

Problème

Modéliser et prédire les **distributions d'abondance d'espèces d'arbres** dans les forêts tropicales du bassin du Congo





Sur chaque parcelle :

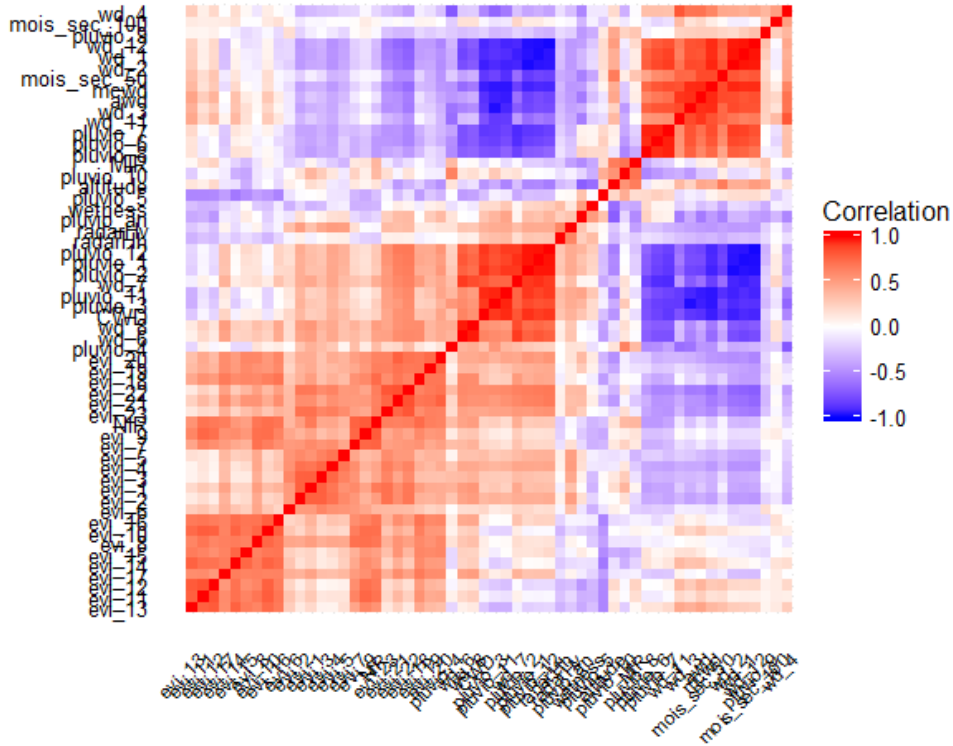
- $q = 94$ **espèces d'arbres** à prédire (abondance : données de comptage)

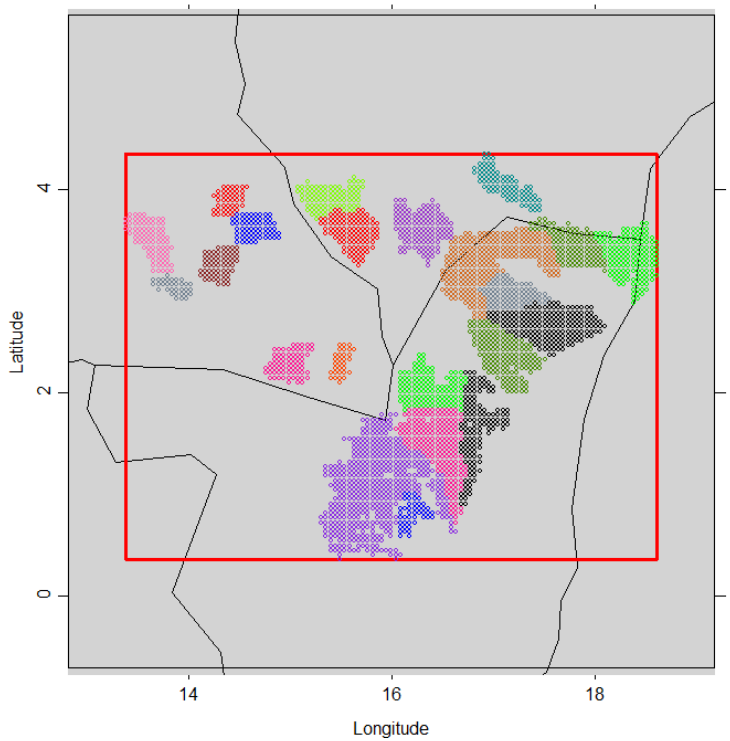
Dans le but de modéliser ces distributions d'espèces :

- $p = 56$ **variables explicatives**

Difficultés

- Variables explicatives très corrélées
  - ↪ **Régularisation** nécessaire
- Observations géo-référencées
  - ↪ **Structure de dépendance** plus complexe





SCGLR

- Gère les fortes redondances au sein des variables explicatives
  - ↪ **Régularisation par construction de composantes supervisées**

Mixed-SCGLR

- Prends en compte la structure de dépendance
  - ↪ Dépendance intra-groupe modélisée par un **effet aléatoire**
  - ↪ **GLMM multivarié**

1. Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR)
   - Motivation
   - Modèle et méthode d'estimation (première composante)
   - Diagnostics graphiques sur données réelles

Notations :

- $\boldsymbol{Y_{n\times q}}$: matrice des q réponses $y^1, \ldots, y^q$
- $\boldsymbol{X_{n\times p}}$: variables explicatives (redondantes)
- $\boldsymbol{U_{n\times N}}$: matrice de design des effets aléatoires

Notations :

- $\boldsymbol{Y_{n\times q}}$: matrice des q réponses $y^1, \ldots, y^q$
- $\boldsymbol{X_{n\times p}}$: variables explicatives (redondantes)
- $\boldsymbol{U_{n\times N}}$: matrice de design des effets aléatoires

Définition des prédicteurs linéaires : $\forall k \in \{1, \ldots, q\}$,

$$\boldsymbol{\eta^k = g\left(\mathbb{E}\left(y^k \mid \xi_k\right)\right)}$$

Modélisation classique

$$\boldsymbol{\eta^k = X\beta_k + U\xi_k}$$

Notations :

- $\boldsymbol{Y_{n\times q}}$: matrice des q réponses $y^1, \ldots, y^q$
- $\boldsymbol{X_{n\times p}}$: variables explicatives (redondantes)
- $\boldsymbol{U_{n\times N}}$: matrice de design des effets aléatoires

Définition des prédicteurs linéaires : $\forall k \in \{1, \ldots, q\}$,

$$\boldsymbol{\eta^k = g\left(\mathbb{E}\left(y^k \mid \xi_k\right)\right)}$$

Régularisation

$$\boldsymbol{\eta^k = (Xu)\gamma_k + U\xi_k}$$

Comment calcule-t-on la composante $f = Xu$?

Bry et Verron (2015)

$$\max_{||u||=1} \left\{ C(u) = [\psi(u)]^{1-s} \times [\phi(u)]^{s} \right\}$$

Comment calcule-t-on la composante $f = Xu$?

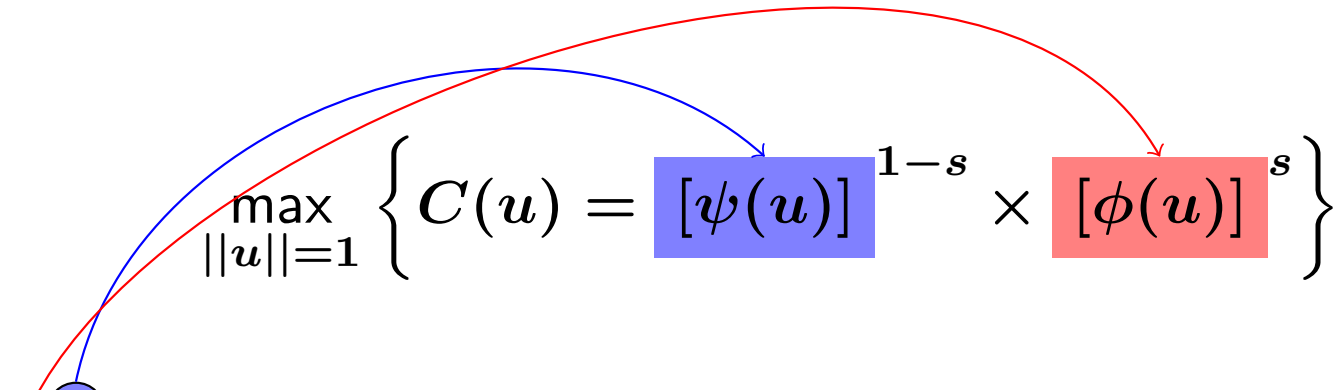
Bry et Verron (2015)

$$\max_{||u||=1} \left\{ C(u) = [\psi(u)]^{1-s} \times [\phi(u)]^{s} \right\}$$

- Critère de qualité d'ajustement : mesure le degré d'ajustement de l'ensemble des prédicteurs linéaires η^k sur la composante $f = Xu$

Comment calcule-t-on la composante $f = Xu$?

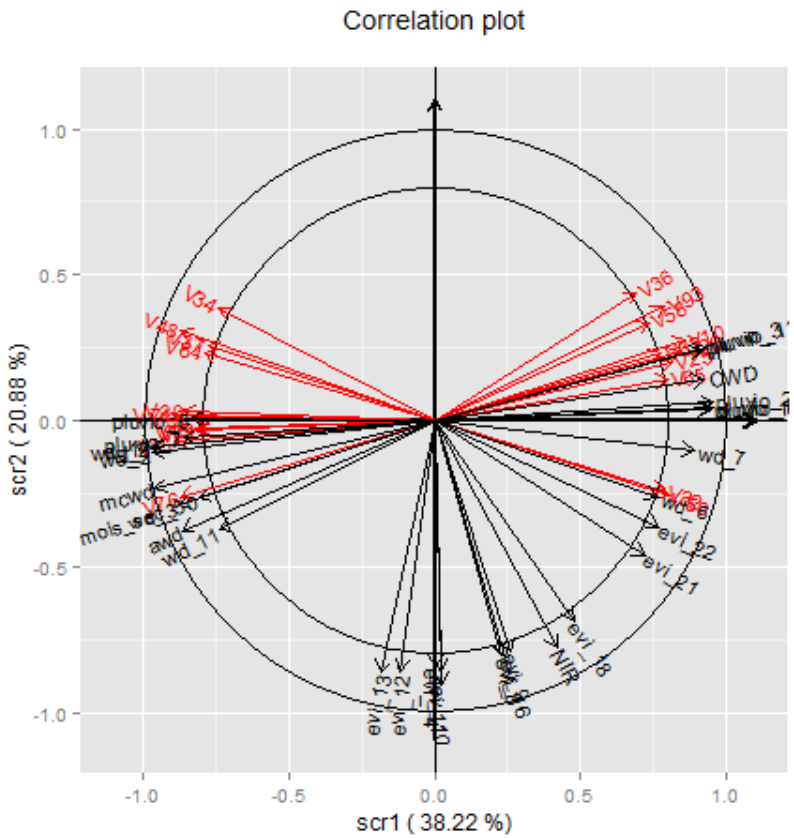
Bry et Verron (2015)

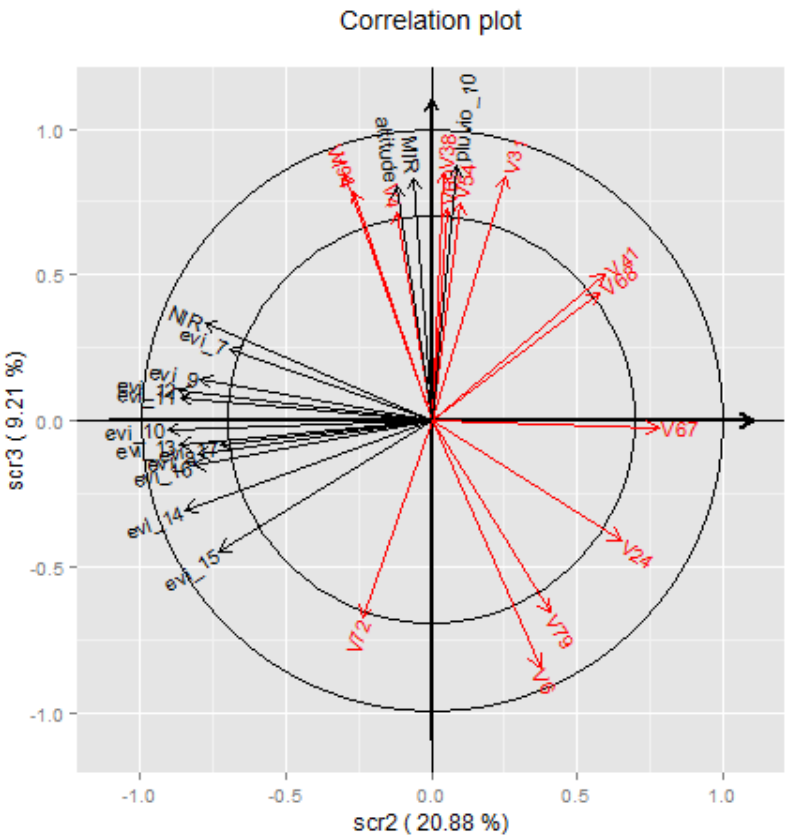


- Critère de qualité d'ajustement : mesure le degré d'ajustement de l'ensemble des prédicteurs linéaires η^k sur la composante $f = Xu$
- Critère de pertinence structurelle : mesure à quel point la composante est proche des structures fortes de X

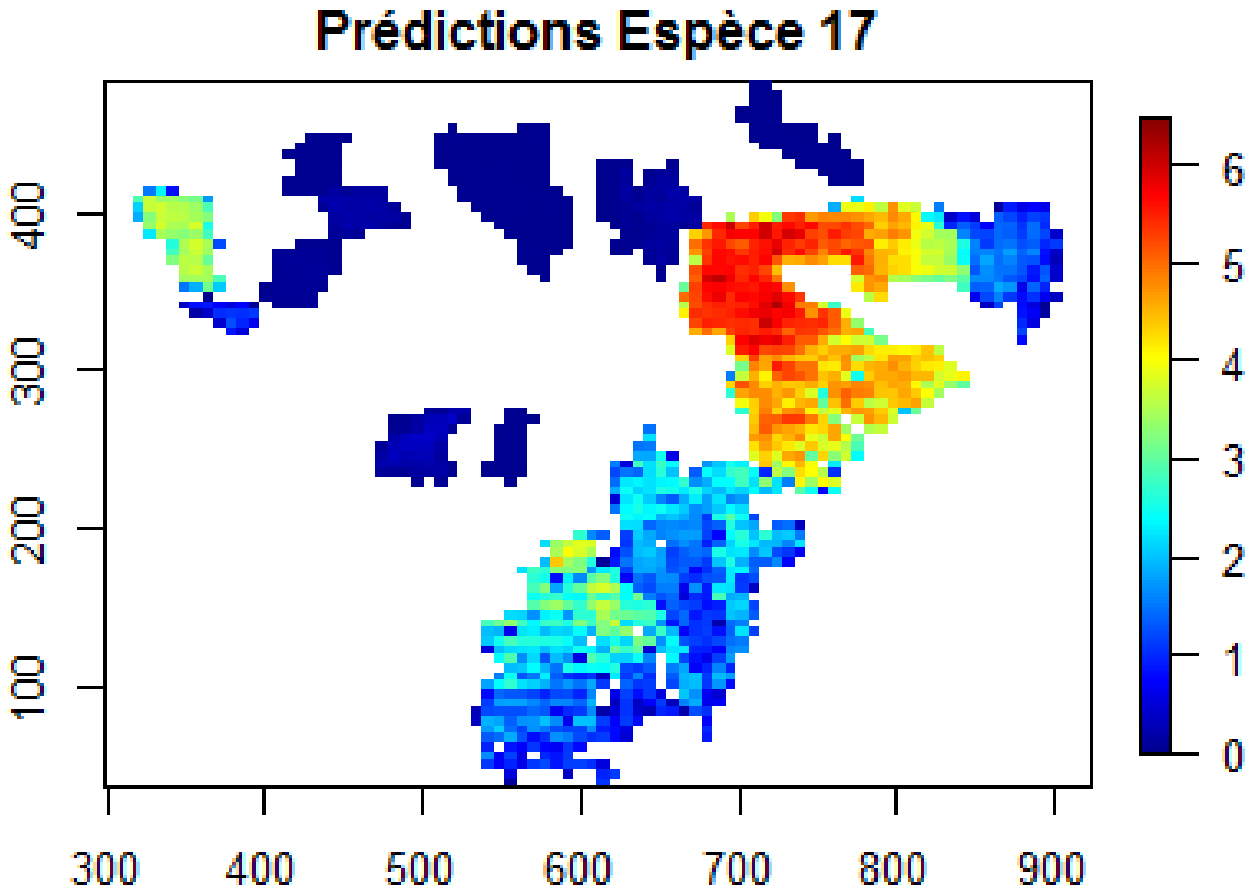
1. Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR)
   - Motivation
   - Modèle et méthode d'estimation (première composante)
   - Diagnostics graphiques sur données réelles

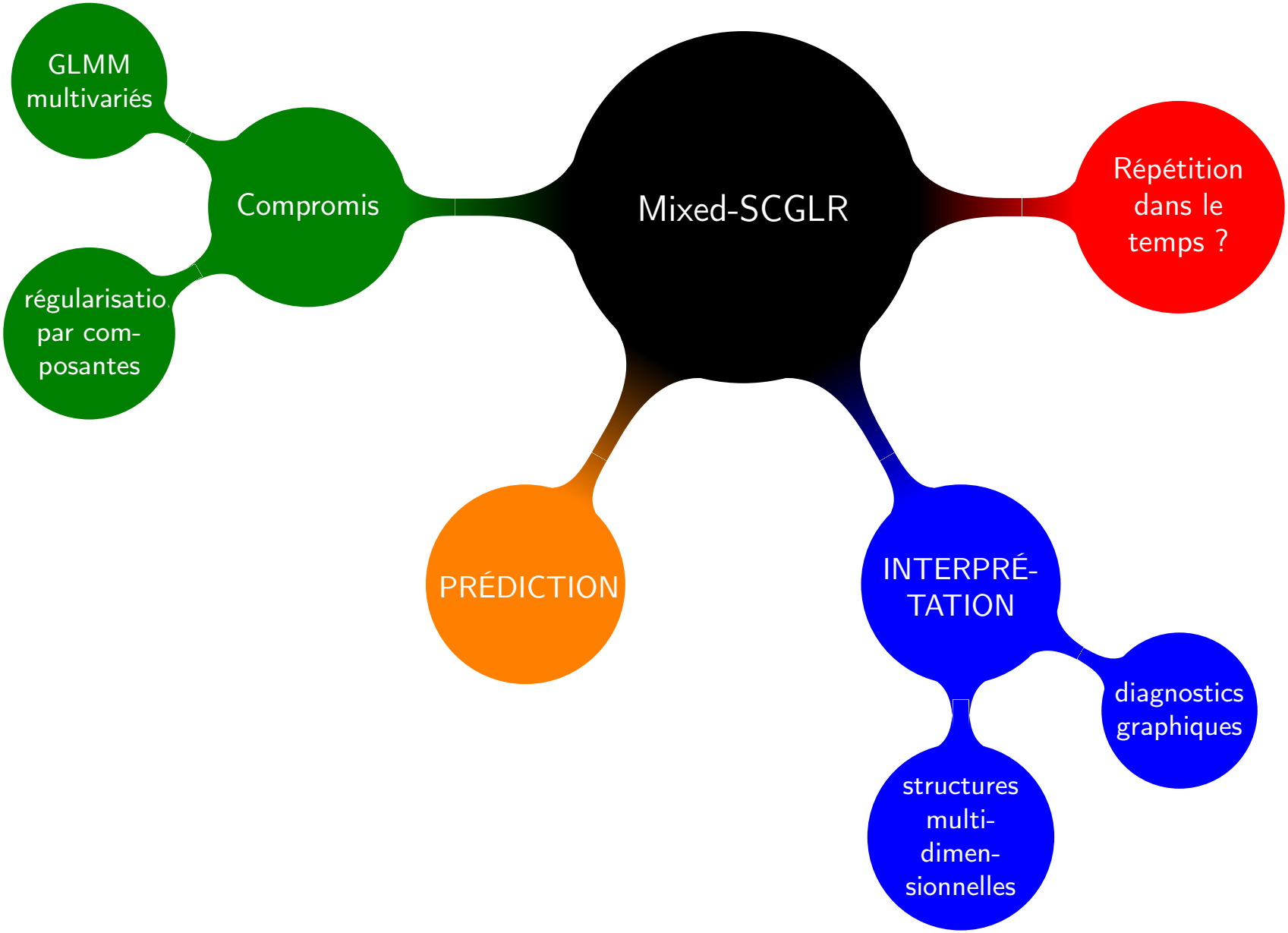
Qualités interprétatives
Plans factoriels (1,2) & (2,3)





Cartes prédictives




GLMM multivariés
Compromis
régularisatio par com-posantes
Mixed-SCGLR
Répétition dans le temps ?
PRÉDICTION
INTERPRÉ-TATION
diagnostics graphiques
structures multi-dimen-sionnelles

2. Ridge GL2M pour données longitudinales
   - L'existant : L2M-Ridge
   - Adaptation aux données longitudinales : ajout d'un effet aléatoire temporel
   - Extension aux GL2M
   - Simulations et données réelles

L2M pour données répétées : $\boxed{\boldsymbol{y = X\beta + U\xi + \varepsilon}}$

- $y = (y_{11}, \ldots, y_{1R}, y_{21}, \ldots, y_{2R}, \ldots\ldots\ldots, y_{N1}, \ldots, y_{NR})^T$
- $\boldsymbol{\beta}$ vecteur des effets fixes et X leur matrice de design
- $\xi = (\xi_1, \xi_2, \ldots, \xi_N)^T$ vecteur de l'effet aléatoire "individu" et $U = I_N \otimes \mathbf{1}_R$ matrice de design correspondante
- $\varepsilon \sim \mathcal{N}\left(0, \boldsymbol{\sigma_0^2} I_{NR}\right) \quad \perp \quad \xi \sim \mathcal{N}(0, \boldsymbol{\sigma_1^2} I_N)$

L2M pour données répétées : $\boxed{y = X\beta + U\xi + \varepsilon}$

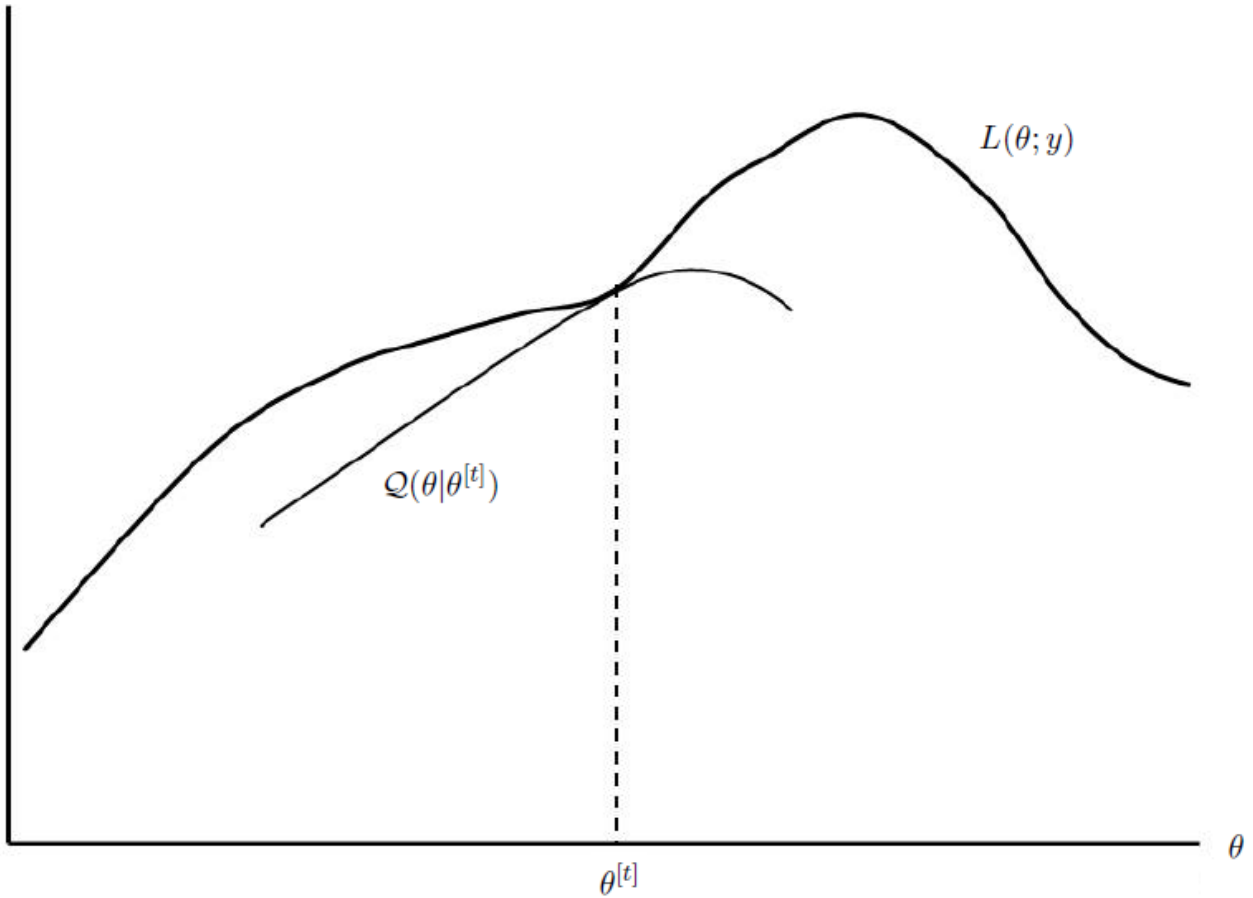
- $y = (y_{11}, \ldots, y_{1R}, y_{21}, \ldots, y_{2R}, \ldots\ldots\ldots, y_{N1}, \ldots, y_{NR})^T$
- β vecteur des effets fixes et X leur matrice de design
- $\xi = (\xi_1, \xi_2, \ldots, \xi_N)^T$ vecteur de l'effet aléatoire "individu" et $U = I_N \otimes \mathbf{1}_R$ matrice de design correspondante
- $\varepsilon \sim \mathcal{N}\left(0, \boldsymbol{\sigma_0^2} I_{NR}\right) \quad \perp \quad \xi \sim \mathcal{N}(0, \boldsymbol{\sigma_1^2} I_N)$

Estimation classique par algorithme EM, $\boldsymbol{\theta = (\beta^T, \sigma_0^2, \sigma_1^2)^T}$

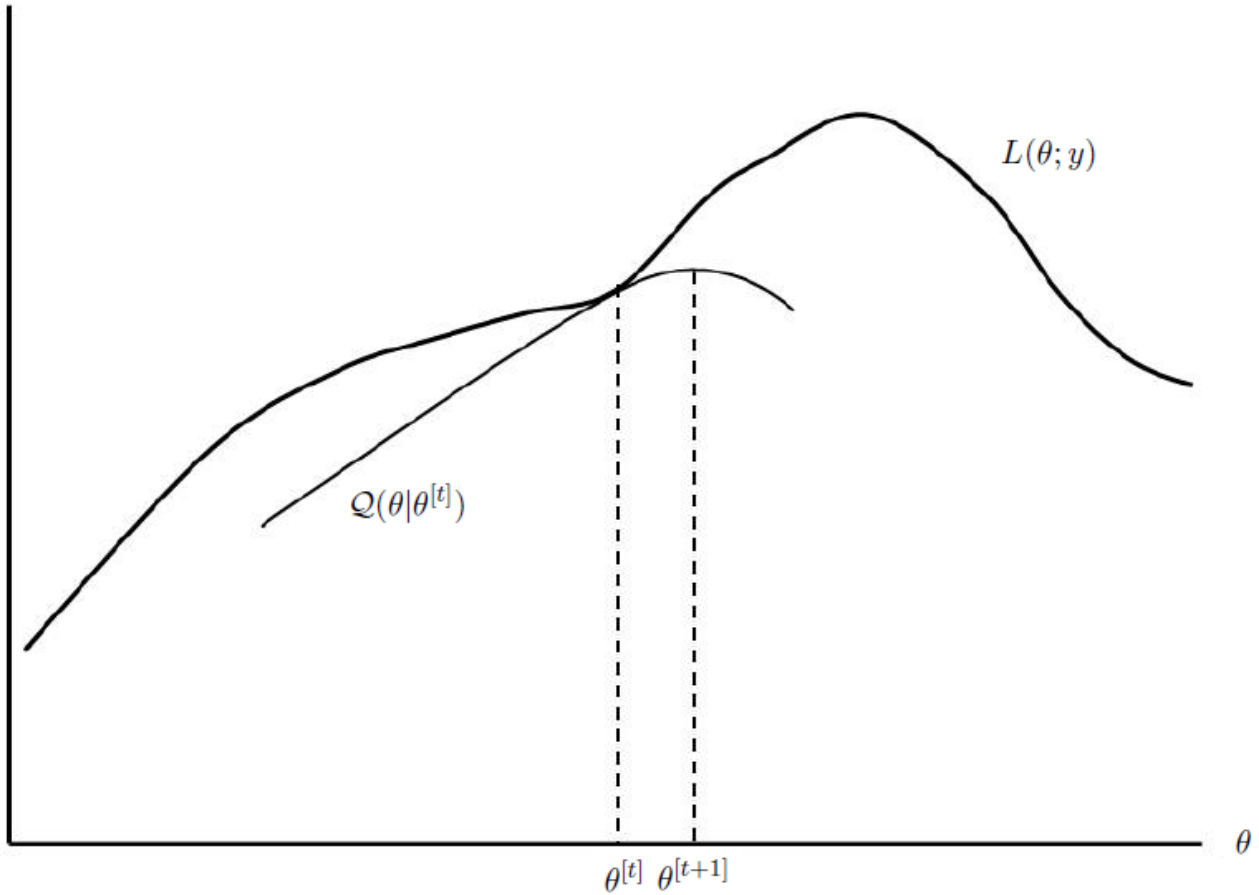
$$\mathbf{E} : \mathcal{Q}\left(\theta \,|\, \theta^{[t]}\right) := \mathbb{E}_{\xi|y}\left[L(\theta; y, \xi) \,|\, \theta^{[t]}\right]$$

$$\mathbf{M} : \theta^{[t+1]} \longleftarrow \arg\max_{\theta} \mathcal{Q}\left(\theta \,|\, \theta^{[t]}\right)$$

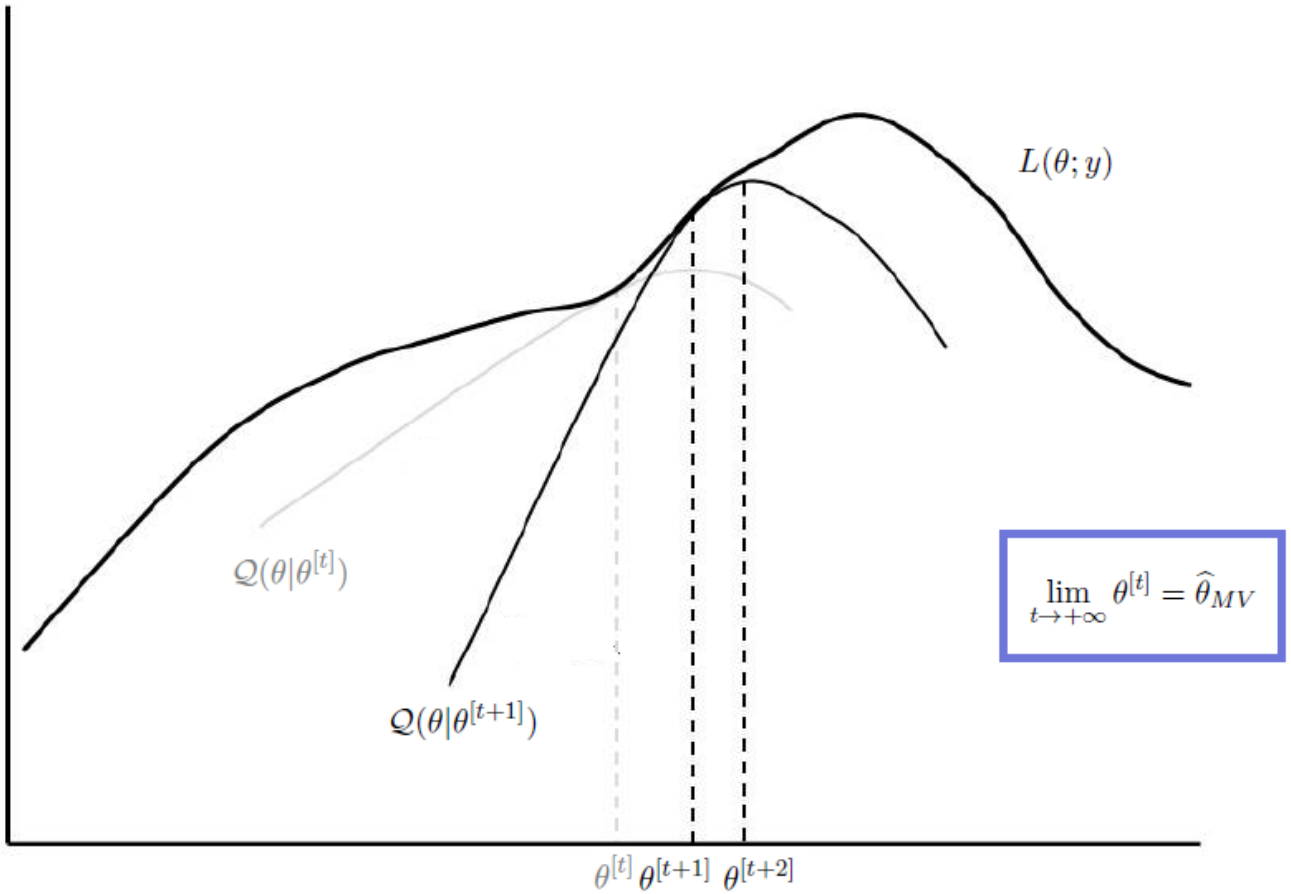
Algorithme EM classique



Algorithme EM classique



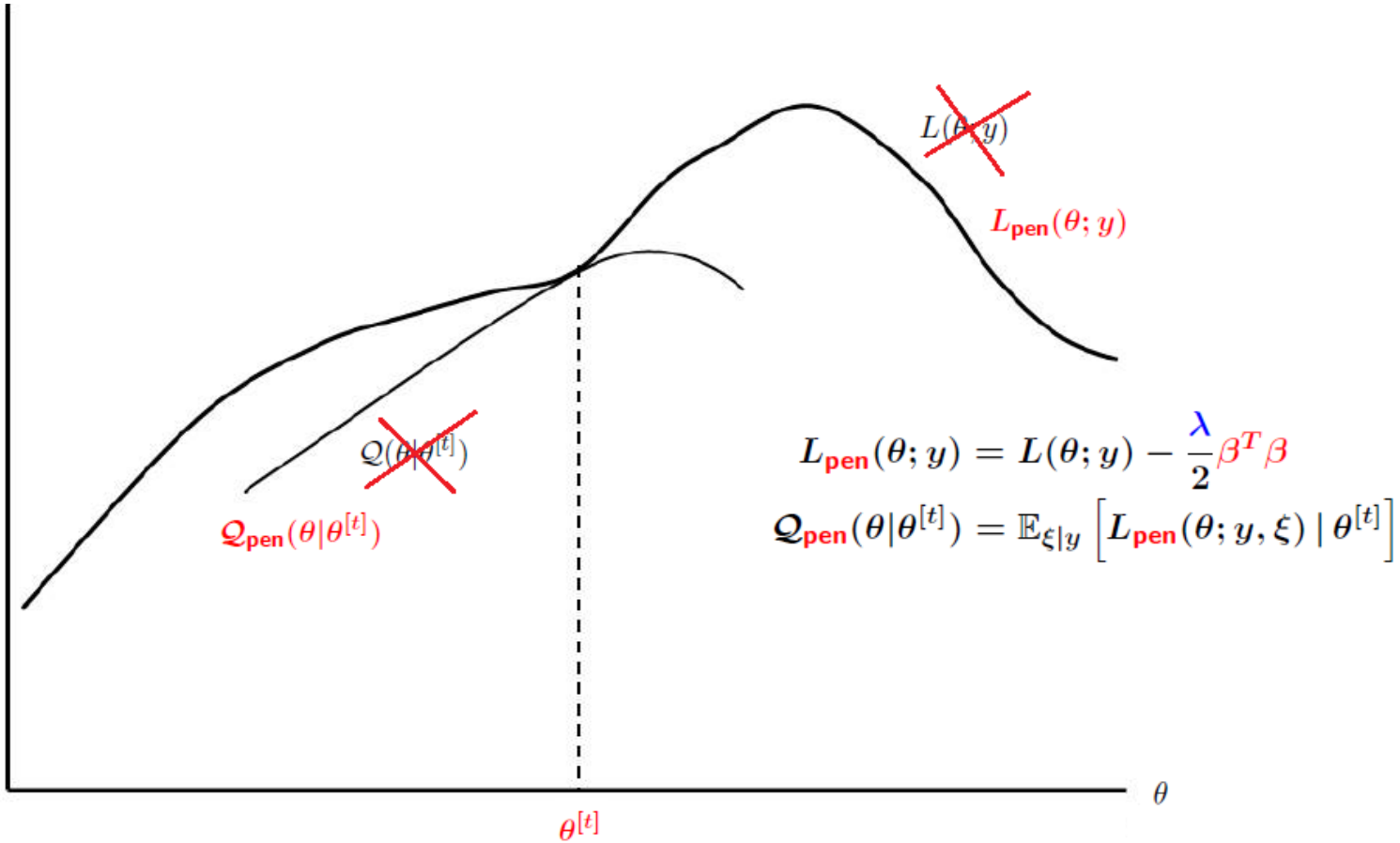
Algorithme EM classique



$$\lim_{t \to +\infty} \theta^{[t]} = \widehat{\theta}_{MV}$$

Algorithme EM avec pénalité RIDGE

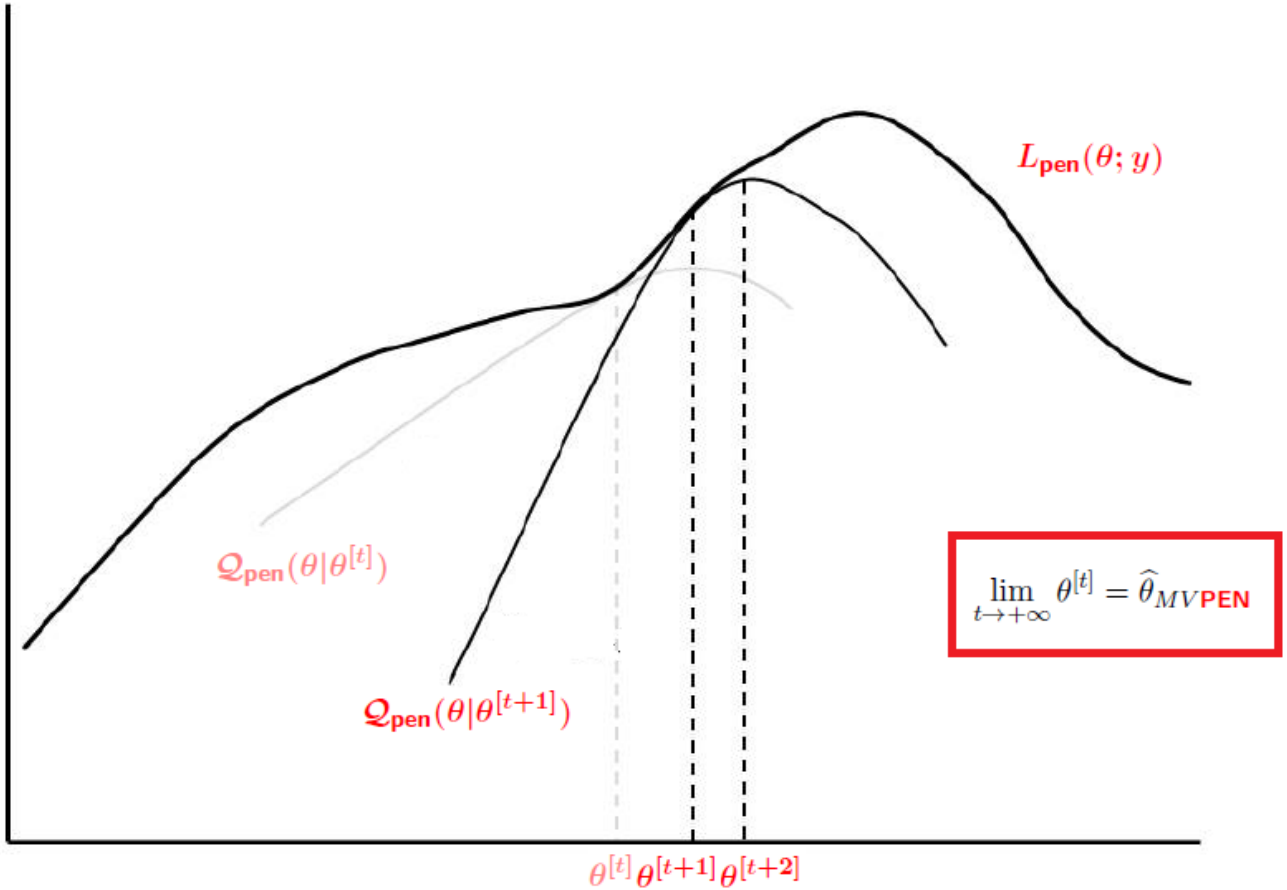
Green (1990), Eliot et al (2011)



$$L_{\text{pen}}(\theta; y) = L(\theta; y) - \frac{\lambda}{2}\beta^T\beta$$

$$\mathcal{Q}_{\text{pen}}(\theta|\theta^{[t]}) = \mathbb{E}_{\xi|y}\left[L_{\text{pen}}(\theta; y, \xi) \mid \theta^{[t]}\right]$$

Algorithme EM avec pénalité RIDGE



Estimation par algorithme EM pénalisé

- Comment calibrer le paramètre de shrinkage λ ?

Estimation par algorithme EM pénalisé

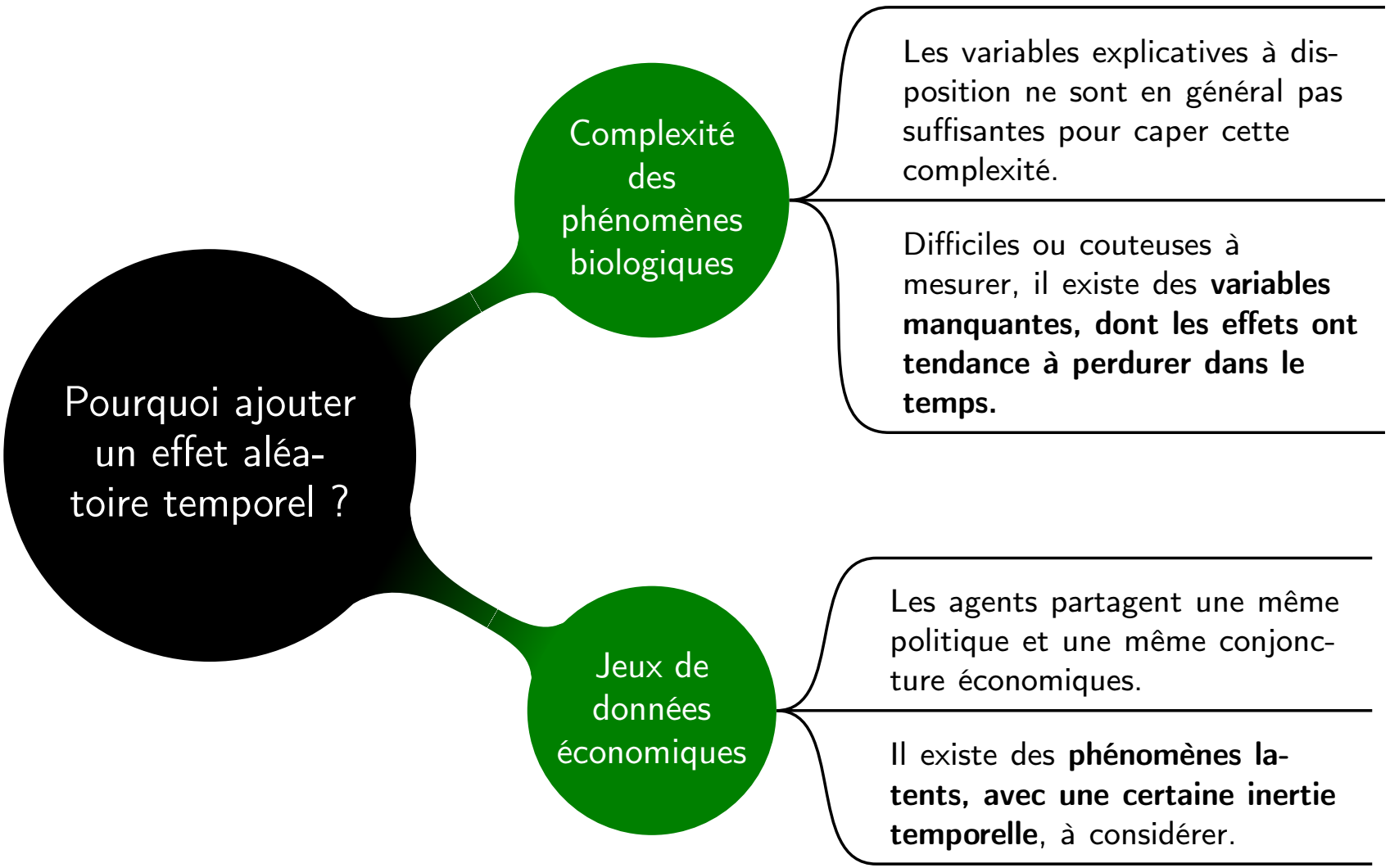
- Comment calibrer le paramètre de shrinkage $\boldsymbol{\lambda}$?
  - ↪ **Validation Croisée Généralisée (GCV)** à chaque étape de l'EM

$$\textbf{GCV} : \boldsymbol{\lambda}^{[t]} \longleftarrow \arg\min_{\lambda} \left\{ \text{GCV}(\lambda) = \frac{n^{-1} \left\| y - H_{\lambda}^{[t]} y \right\|^2}{\left[1 - n^{-1} \text{tr}\left(H_{\lambda}^{[t]} \right) \right]^2} \right\}$$

$$\textbf{E} : \mathcal{Q}_{\textbf{pen}}\left(\theta \,|\, \theta^{[t]} \right) := \mathbb{E}_{\xi|y} \left[L(\theta; y, \xi) - \frac{\boldsymbol{\lambda}^{[t]}}{2} \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{\beta} \,|\, \theta^{[t]} \right]$$

$$\textbf{M} : \theta^{[t+1]} \longleftarrow \arg\max_{\theta} \mathcal{Q}_{\textbf{pen}}\left(\theta \,|\, \theta^{[t]} \right)$$

2. Ridge GL2M pour données longitudinales
   - L'existant : L2M-Ridge
   - Adaptation aux données longitudinales : ajout d'un effet aléatoire temporel
   - Extension aux GL2M
   - Simulations et données réelles


Pourquoi ajouter un effet aléatoire temporel ?
Complexité des phénomènes biologiques
Les variables explicatives à disposition ne sont en général pas suffisantes pour caper cette complexité.
Difficiles ou couteuses à mesurer, il existe des **variables manquantes, dont les effets ont tendance à perdurer dans le temps.**
Jeux de données économiques
Les agents partagent une même politique et une même conjoncture économiques.
Il existe des **phénomènes latents, avec une certaine inertie temporelle**, à considérer.

L2M pour données longitudinales :

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{U_1}\boldsymbol{\xi}^{\mathbf{1}} + \boldsymbol{U_2}\boldsymbol{\xi}^{\mathbf{2}} + \varepsilon}$$

$$\begin{aligned} y = (&y_{\mathbf{11}}, y_{\mathbf{12}}, \ldots, y_{\mathbf{1R}}, \\ &y_{\mathbf{21}}, y_{\mathbf{22}}, \ldots, y_{\mathbf{2R}}, \ldots\ldots \\ &y_{\mathbf{N1}}, y_{\mathbf{N2}}, \ldots, y_{\mathbf{NR}})^T \end{aligned}$$

- β vecteur des effets fixes
- $\boldsymbol{\xi}^{\mathbf{1}} = \left(\boldsymbol{\xi}^{\mathbf{1}}_{\mathbf{1}}, \boldsymbol{\xi}^{\mathbf{1}}_{\mathbf{2}}, \ldots, \boldsymbol{\xi}^{\mathbf{1}}_{\boldsymbol{N}}\right)^{\boldsymbol{T}}$ vecteur de l'effet aléatoire "individu", $U_1 = I_N \otimes \mathbf{1}_R$ matrice de design correspondante
- $\boldsymbol{\xi}^{\mathbf{2}} = \left(\boldsymbol{\xi}^{\mathbf{2}}_{\mathbf{1}}, \boldsymbol{\xi}^{\mathbf{2}}_{\mathbf{2}}, \ldots, \boldsymbol{\xi}^{\mathbf{2}}_{\boldsymbol{R}}\right)^{\boldsymbol{T}}$ vecteur de l'effet aléatoire "temporel", $U_2 = \mathbf{1}_N \otimes I_R$ matrice de design correspondante

Modélisation des effets aléatoires

- $\varepsilon \sim \mathcal{N}\left(0, \boldsymbol{\sigma_0^2} I_{NR}\right) \quad \perp \quad \xi^1 \sim \mathcal{N}(0, \boldsymbol{\sigma_1^2} I_N)$
- $\xi^2 \sim \mathcal{N}(0, \Sigma)$, avec $\Sigma_{ij} = \boldsymbol{\sigma_2^2} \dfrac{\boldsymbol{\rho}^{|i-j|}}{1-\boldsymbol{\rho}^2}$, $\xi^2 \perp \xi^1$ et $\xi^2 \perp \varepsilon$

Modélisation des effets aléatoires

- $\varepsilon \sim \mathcal{N}\left(0, \boldsymbol{\sigma_0^2} I_{NR}\right) \quad \perp \quad \xi^1 \sim \mathcal{N}(0, \boldsymbol{\sigma_1^2} I_N)$
- $\xi^2 \sim \mathcal{N}(0, \Sigma)$, avec $\Sigma_{ij} = \boldsymbol{\sigma_2^2} \dfrac{\boldsymbol{\rho}^{|i-j|}}{1-\boldsymbol{\rho}^2}$, $\xi^2 \perp \xi^1$ et $\xi^2 \perp \varepsilon$

EM Généralisé, avec

$\boldsymbol{\theta} = \left(\boldsymbol{\beta^T}, \boldsymbol{\sigma_0^2}, \boldsymbol{\sigma_1^2}, \boldsymbol{\sigma_2^2}, \boldsymbol{\rho}\right)^{\boldsymbol{T}}$ et $\boldsymbol{\xi} = \left(\boldsymbol{\xi^{1T}}, \boldsymbol{\xi^{2T}}\right)^{\boldsymbol{T}}$

GCV : $\boldsymbol{\lambda}^{[t]} \longleftarrow \arg\min_{\lambda} \text{GCV}(\lambda)$

E : $\mathcal{Q}_{\textbf{pen}}\left(\theta \,|\, \theta^{[t]}\right) := \mathbb{E}_{\xi|y}\left[L(\theta; y, \xi) - \dfrac{\boldsymbol{\lambda}^{[t]}}{2}\boldsymbol{\beta^T}\boldsymbol{\beta} \,|\, \theta^{[t]}\right]$

M : $\theta^{[t+1]}$ tel que : $\mathcal{Q}_{\textbf{pen}}\left(\theta^{[t+1]} \,|\, \theta^{[t]}\right) \geqslant \mathcal{Q}_{\textbf{pen}}\left(\theta^{[t]} \,|\, \theta^{[t]}\right)$

2 Ridge GL2M pour données longitudinales

- L'existant : L2M-Ridge
- Adaptation aux données longitudinales : ajout d'un effet aléatoire temporel
- Extension aux GL2M
- Simulations et données réelles

Extension aux GLMM

$$Y_i \,|\, \xi \overset{\text{iid}}{\sim} F \text{ de la famille exponentielle}$$
$$g(\underbrace{\mathbb{E}\left(Y \,|\, \xi\right)}_{\mu}) = \eta = X\beta + U_1\xi^1 + U_2\xi^2$$

Méthodes d'estimation dans les GLMM

- Approximations d'intégrales : Laplace, Gauss-Hermite
- Méthodes de Monte Carlo : MCMC, MCEM (McCulloch (1997)), MCML (Knudson (2016))
- Méthodes de linéarisation : Schall (1991), PQL (Breslow et al. (1993))

La méthode

Étape de LINÉARISATION

- Linéarisation à l'ordre 1 de y_i au voisinage de μ_i :

$$y_i \simeq z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$
$$z_i = \eta_i + e_i$$

- Définition du modèle linéarisé :

$$\mathcal{M} : z = X\beta + U_1\xi^1 + U_2\xi^2 + e, \quad \text{avec } \mathbb{V}(e) = \Gamma$$

Étape d'ESTIMATION

Algorithme EM pénalisé sur le modèle $\mathcal{M}$

Itération générique

Linéarisation

$$\mathcal{M} : z^{[t]} = X\beta + U_1\xi^1 + U_2\xi^2 + e, \quad \text{avec } \mathbb{V}(e) = \Gamma^{[t]}$$

Estimation

GCV : $\boldsymbol{\lambda}^{[t]} \longleftarrow \arg\min_{\lambda} \text{GCV}(\lambda)$

E : $\mathcal{Q}_{\textbf{pen}}\left(\theta \,|\, \theta^{[t]}\right) := \mathbb{E}_{\xi|z}\left[L(\theta; z, \xi) - \frac{\boldsymbol{\lambda}^{[t]}}{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta} \,|\, \theta^{[t]}\right]$

M : $\theta^{[t+1]}$ tel que : $\mathcal{Q}_{\textbf{pen}}\left(\theta^{[t+1]} \,|\, \theta^{[t]}\right) \geqslant \mathcal{Q}_{\textbf{pen}}\left(\theta^{[t]} \,|\, \theta^{[t]}\right)$

Mise à jour

Définir $\xi^{[t+1]}, z^{[t+1]}, \Gamma^{[t+1]}$

2. Ridge GL2M pour données longitudinales
   - L'existant : L2M-Ridge
   - Adaptation aux données longitudinales : ajout d'un effet aléatoire temporel
   - Extension aux GL2M
   - Simulations et données réelles

Plan de simulation

Données Poissonniennes

- $y \sim \mathcal{P}\left(\exp(X\beta + U_1\xi^1 + U_2\xi^2)\right)$
- $X = [\underbrace{x^1, x^2, x^3, x^4, x^5}_{\text{corr. 2 à 2 = 0.9}} \mid \underbrace{x^6, x^7, x^8, x^9, x^{10}}_{\text{indépendantes}}]$

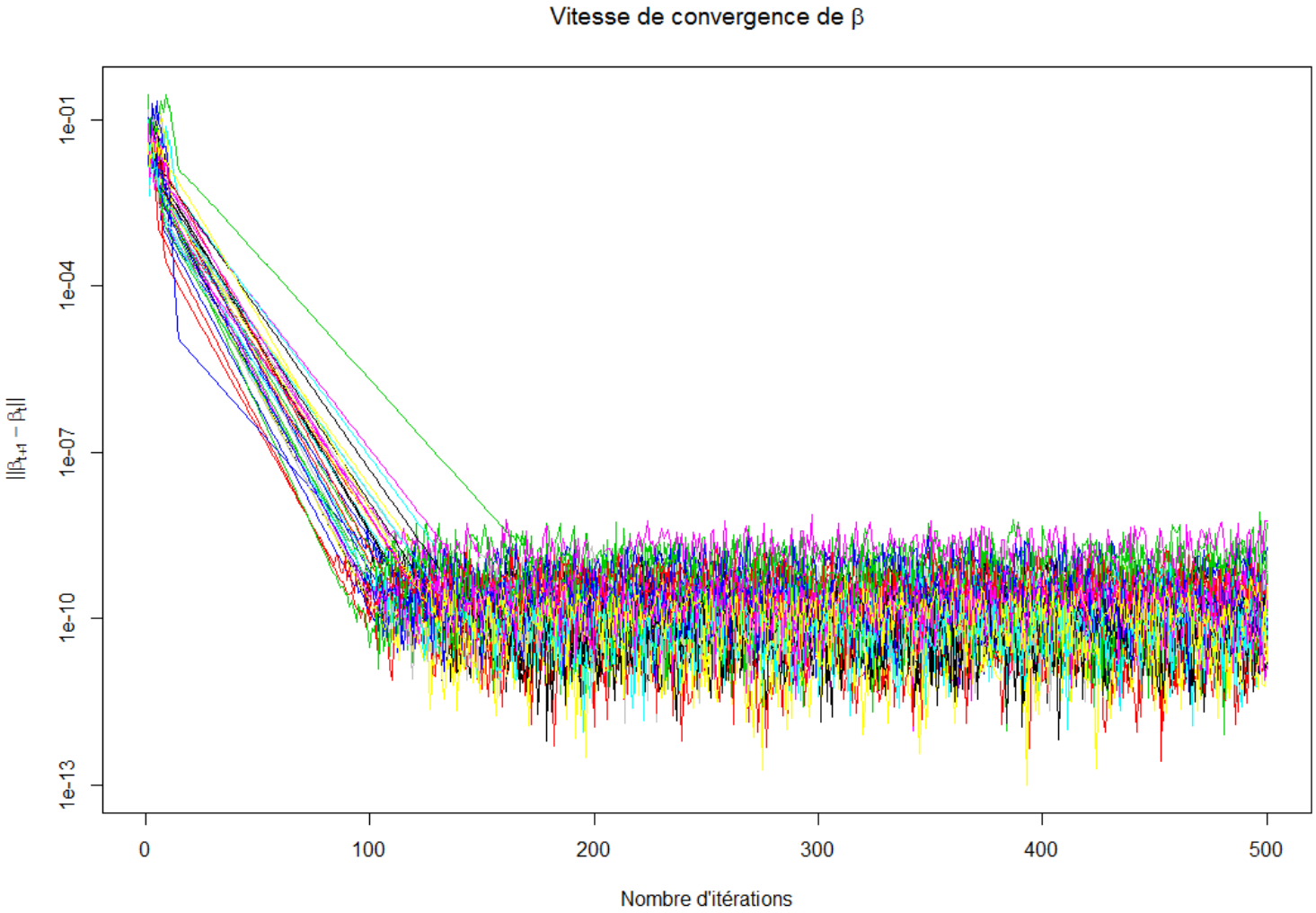
Paramètres choisis

- $\beta = (0.5, 0.5, 0.5, 0.5, 0.5,\ \ 0, 0, 0, 0, 0)$
- $\sigma_1^2 = \sigma_2^2 = \rho = 0.5$

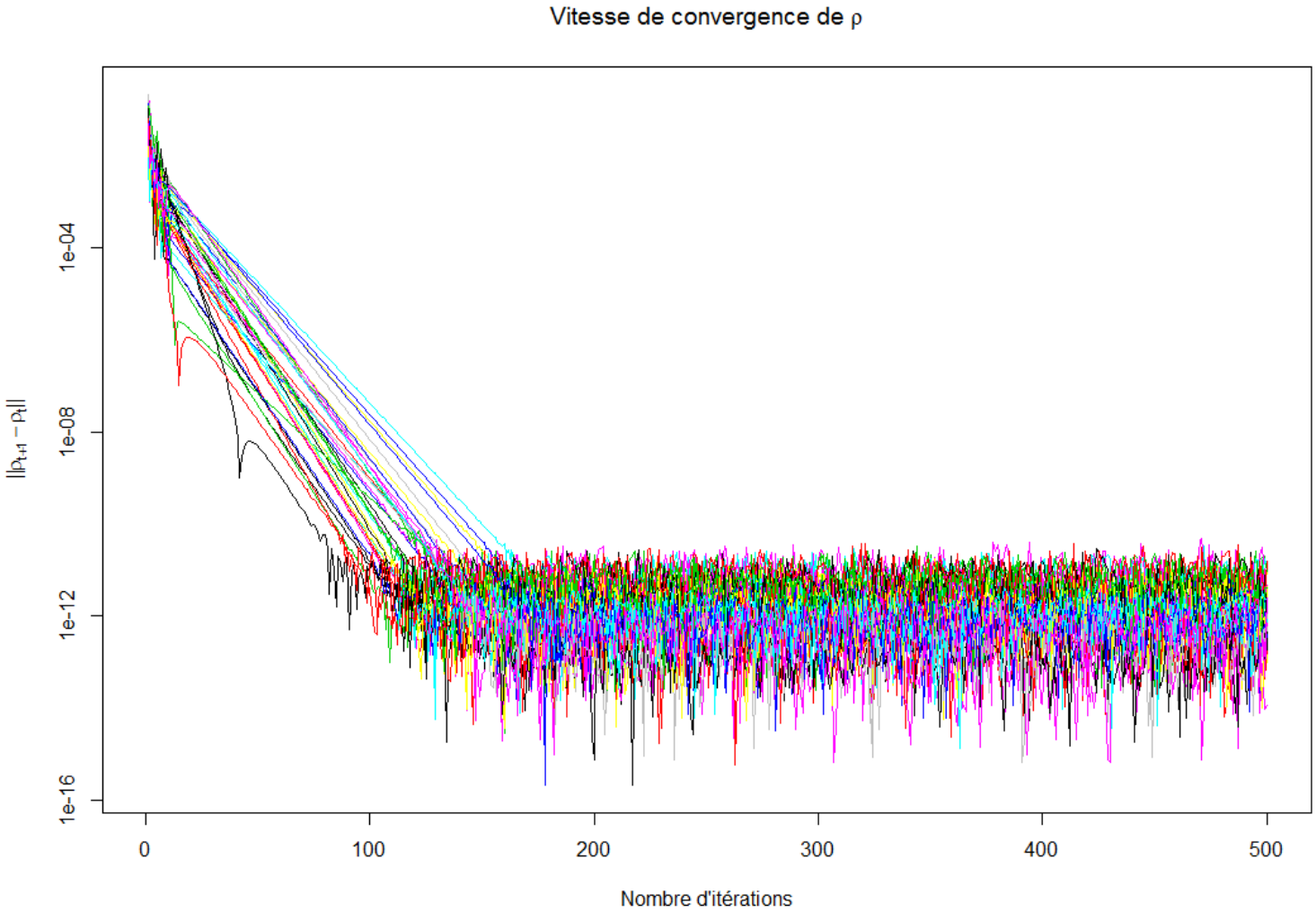
Individus – Répétitions

- $N = 10$
- $R \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$

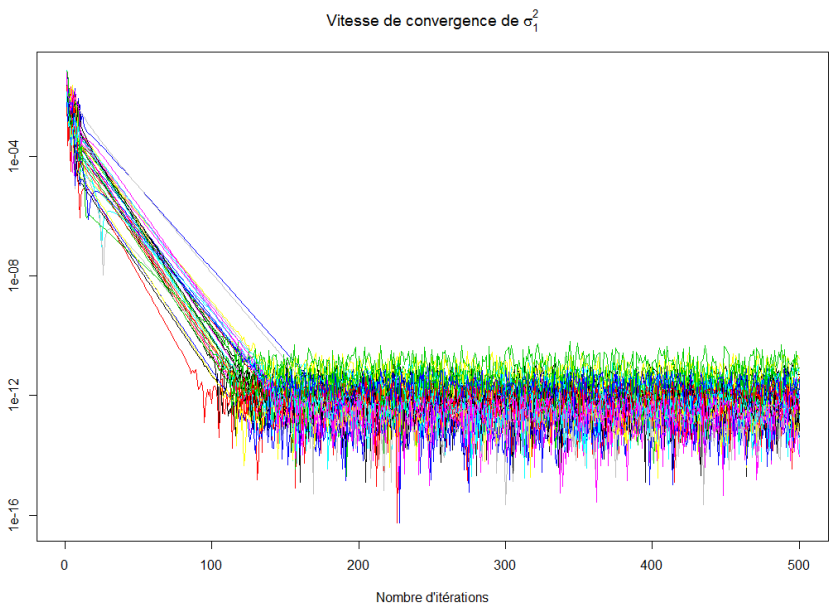
La méthode est-elle convergente ?
Effets fixes β

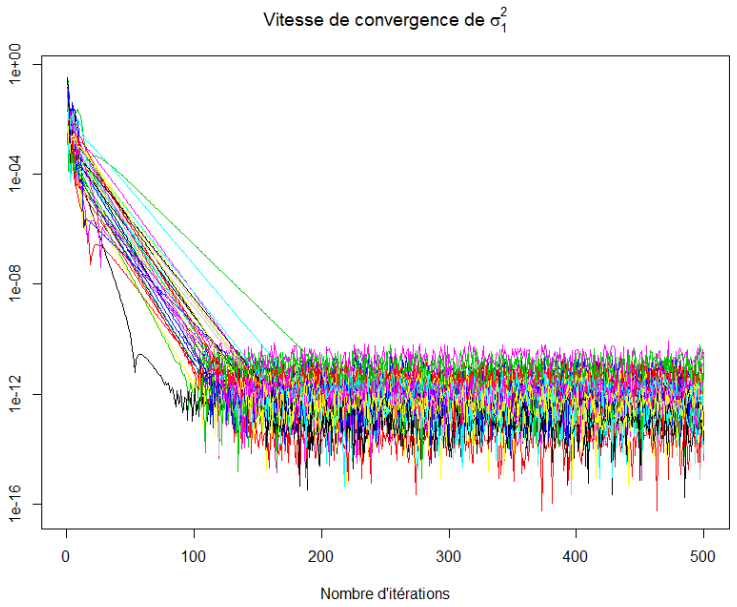


La méthode est-elle convergente ?
Paramètre du processus autorégressif ρ

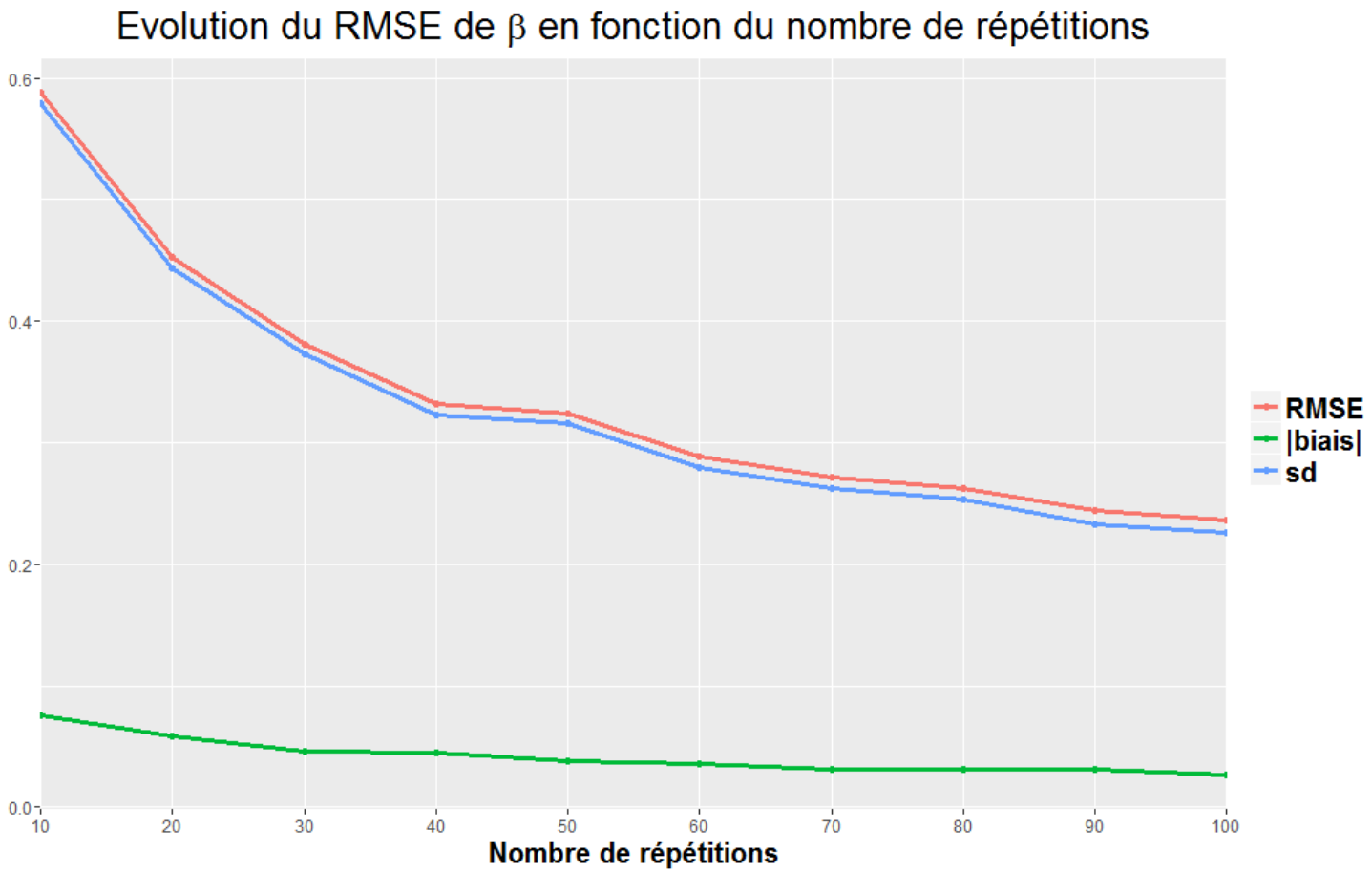


La méthode est-elle convergente ?
Paramètres de variances σ_1^2 et σ_2^2

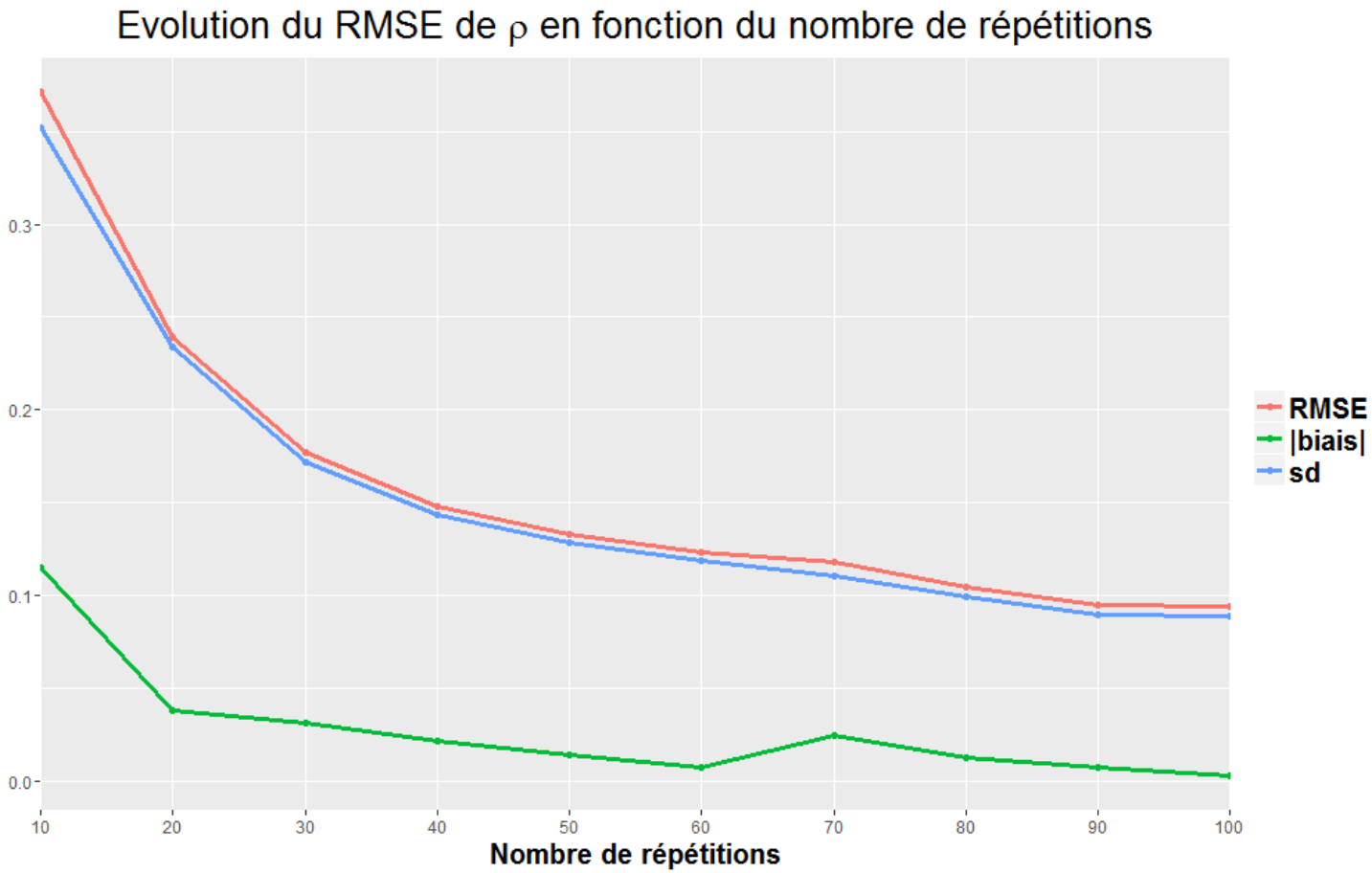




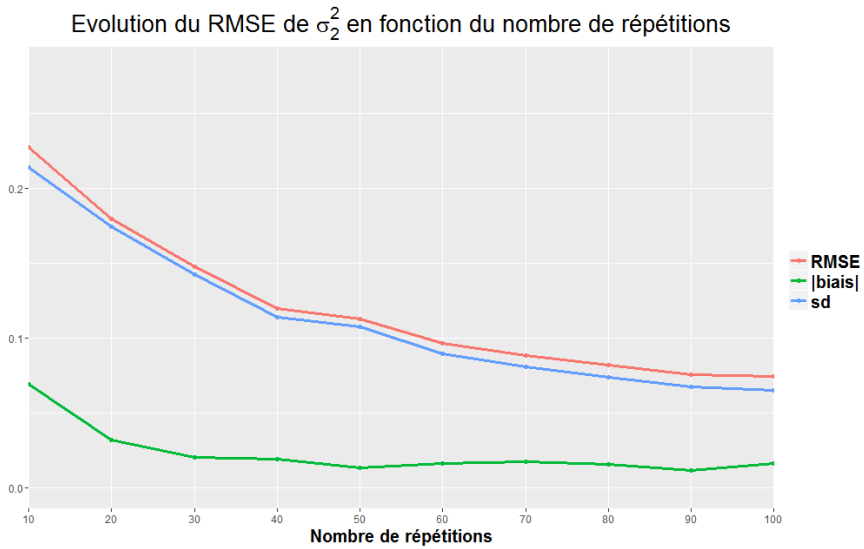
Les paramètres sont-ils bien estimés ?
Effets fixes β

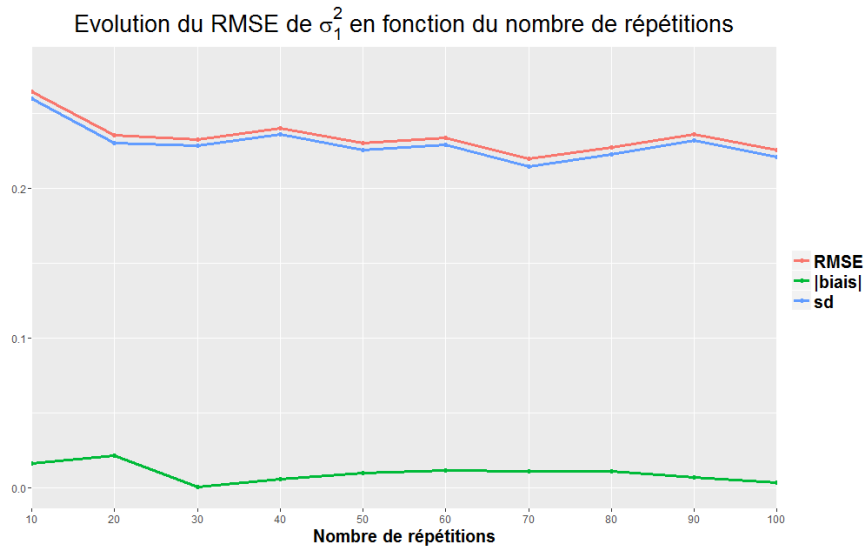


Les paramètres sont-ils bien estimés ?
Paramètre du processus autorégressif ρ

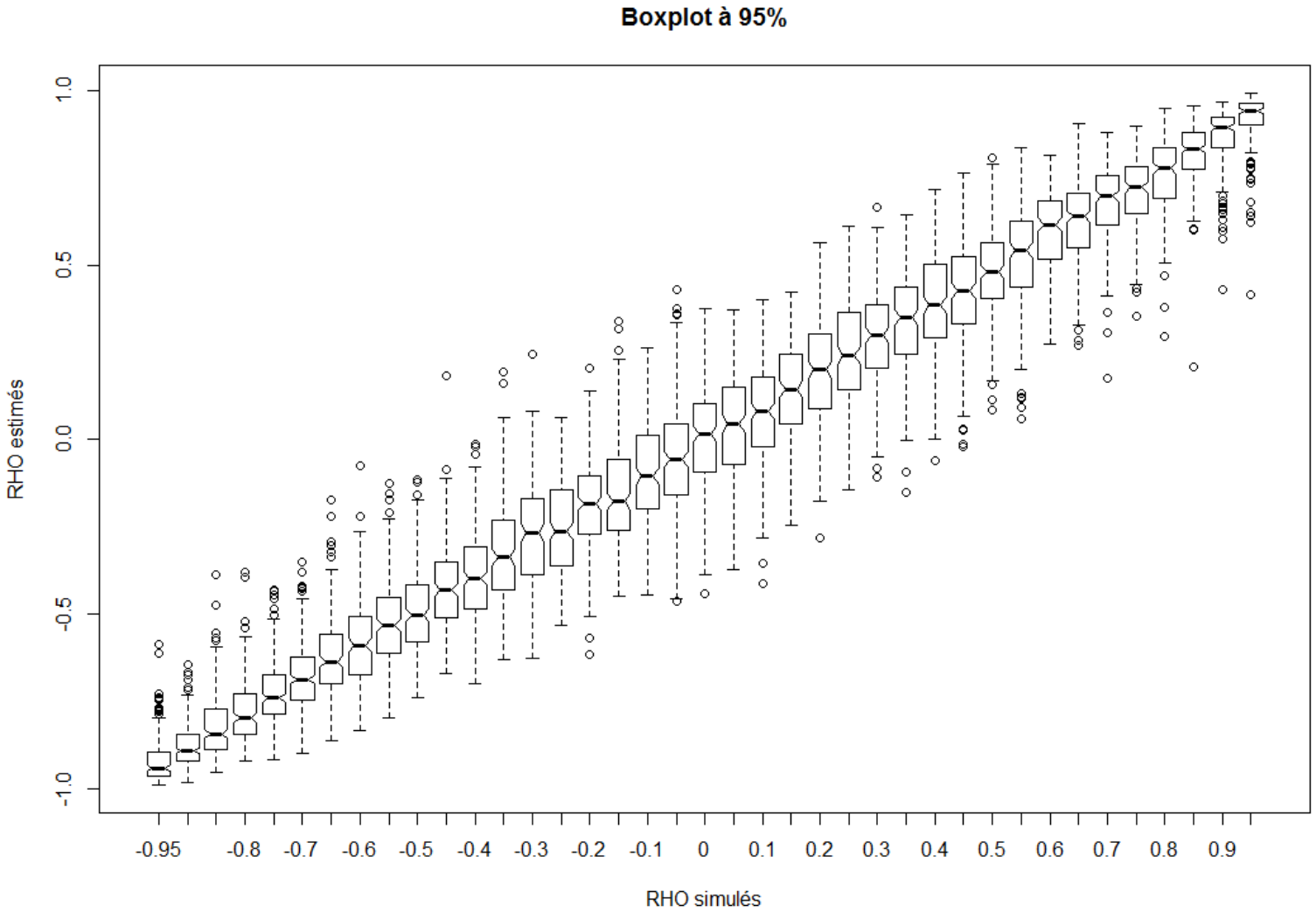


Les paramètres sont-ils bien estimés ?
Paramètres de variances σ_1^2 et σ_2^2





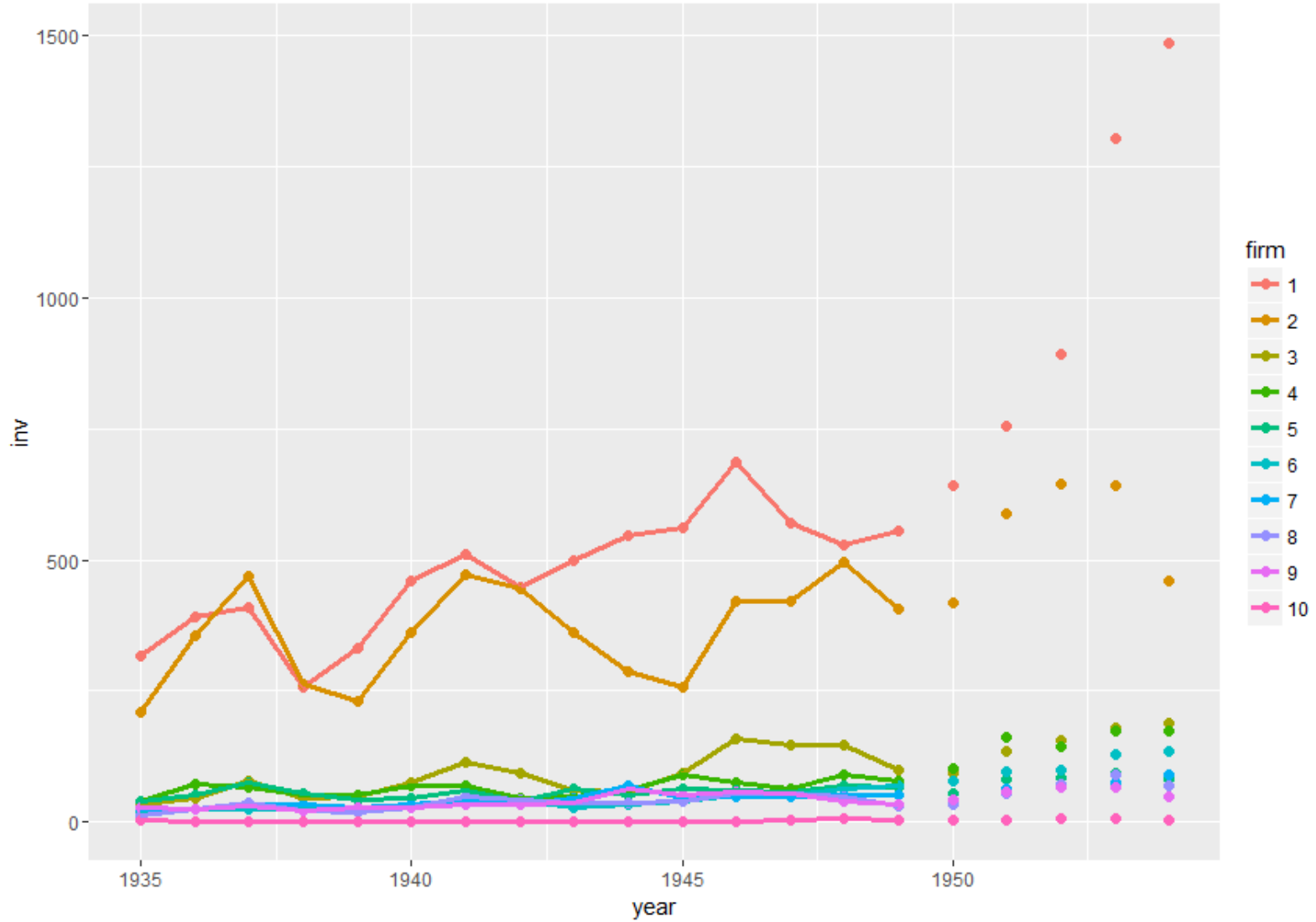
La méthode est-elle sensible à la valeur de ρ ?



Données réelles : "GRUNFELD"

- Observations annuelles de 10 grandes entreprises des USA sur 20 ans (1935–1954).
- Variable réponse : "inv" (somme des investissements réalisés par l'entreprise)
- Variables explicatives : "value" (valeur de l'entreprise sur le marché) et "capital" (capital de l'entreprise).

Modèle calibré sur les 15 premières années



Comparaison des estimations

Packape "PLM"

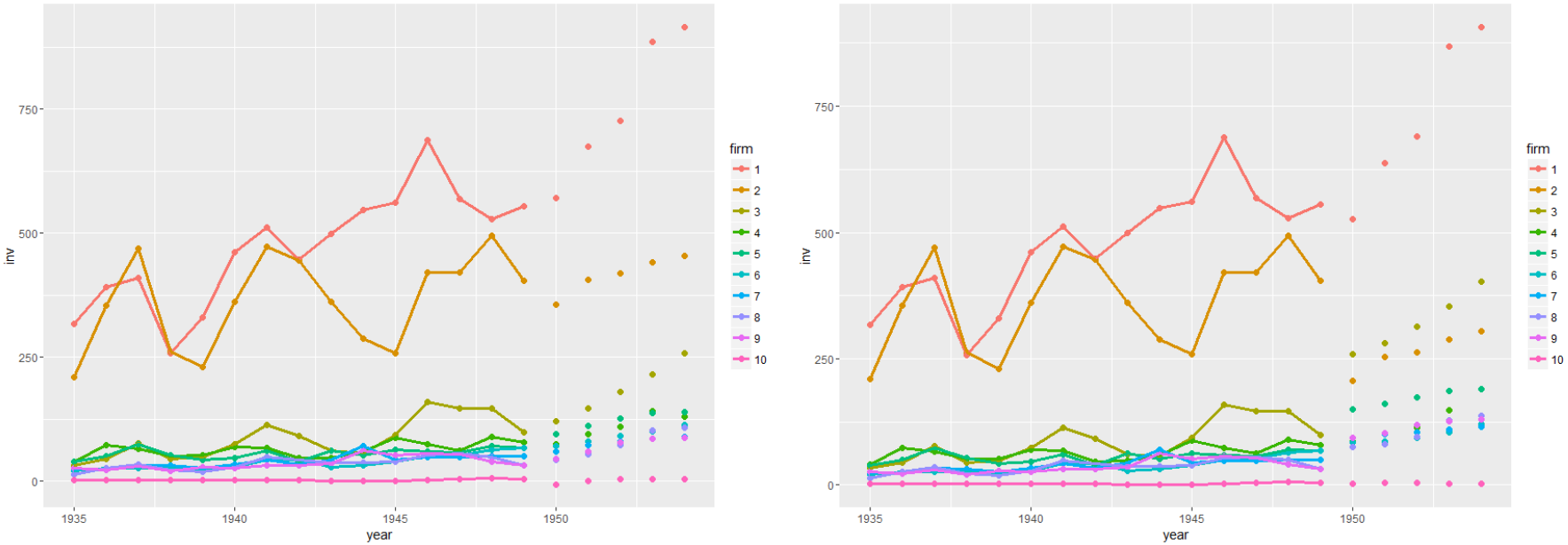
- $\widehat{\beta} = (-4.5, 0.08, 0.20)^T$
- $\widehat{\sigma_1^2} = 6197$
- $\widehat{\sigma_2^2} = 95$

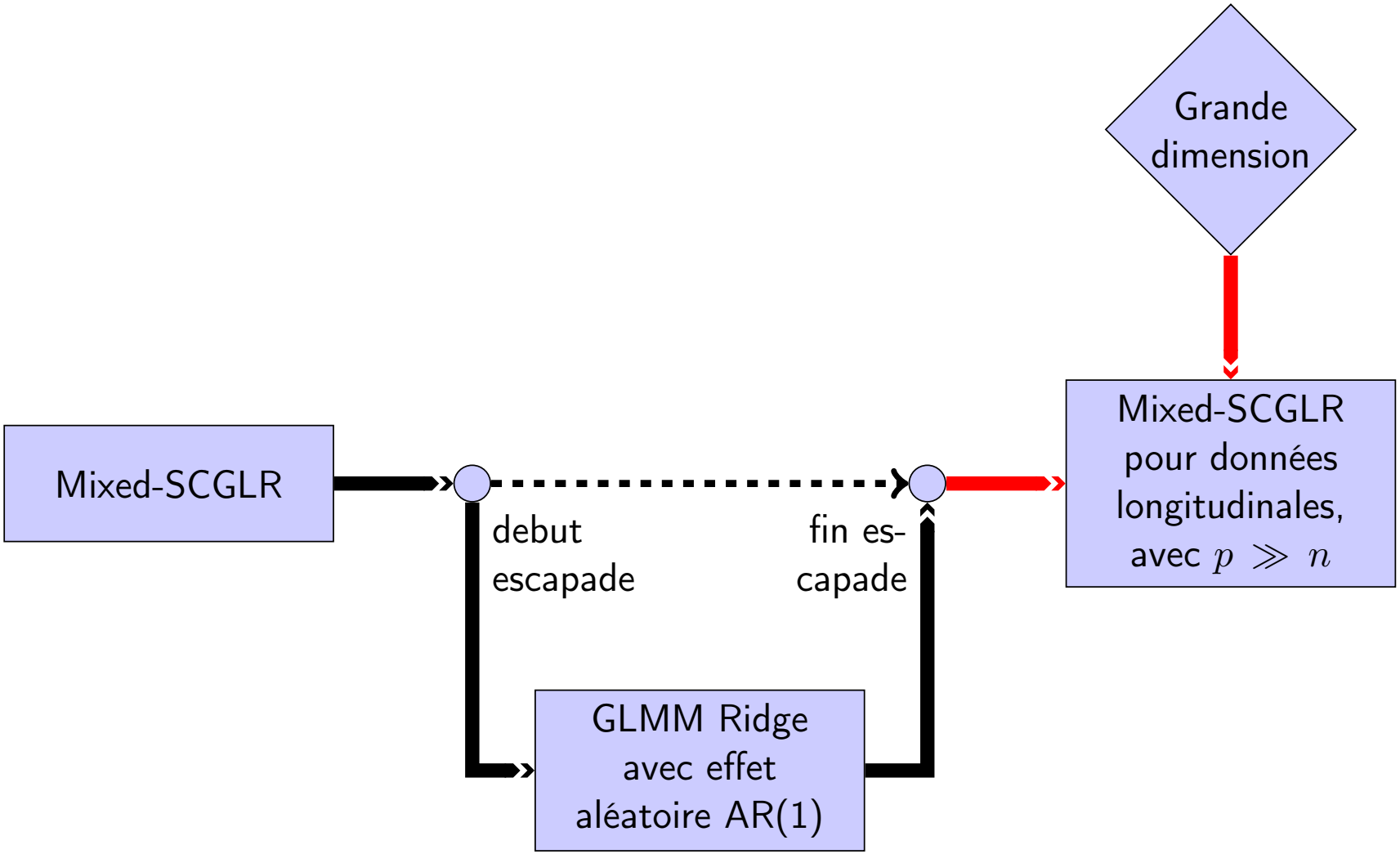
Ma méthode

- $\widehat{\beta} = (0.005, 0.07, 0.17)^T$
- $\widehat{\sigma_1^2} = 5740$
- $\widehat{\sigma_2^2} = 193$
- $\widehat{\rho} = 0.49$

Comparaison des prédictions

Mes prédictions vs Leurs prédictions




Grande dimension
Mixed-SCGLR
Mixed-SCGLR pour données longitudinales, avec $p \gg n$
debut escapade
fin escapade
GLMM Ridge avec effet aléatoire AR(1)