# Validation de code de calcul par l'estimation du mélange des modèles

Negar Soleimani

Département de Mathématiques Hadamard
Université Paris-Saclay

Encadrants : Pierre Barbillon et Kaniav Kamary

Workshop Statistiques au sommet de Rochebrune — 2026

25 mars 2026

AgroParisTech

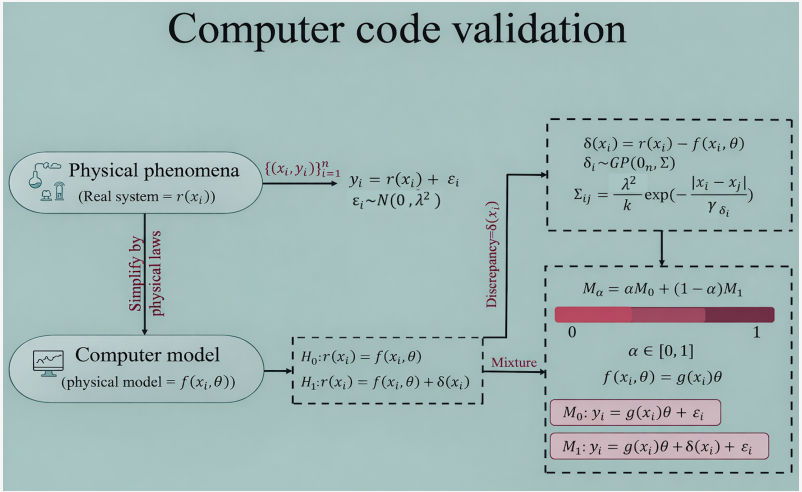
MIA
PARIS-SACLAY

# Plan de la présentation

**Structure**

- Objectifs et cadre méthodologique
- Modélisation par mélange et choix a priori
- Simulation sous le modèle $M_0$
- Simulation sous le modèle $M_1$
- Diagnostic local par seuil
- Inférence sur les données réelles
- Conclusion et perspectives
- Références

# Objectifs et cadre méthodologique



# Objectifs et cadre méthodologique

**Prior choice :**

- Une loi beta pour le poids $\alpha$

$$\alpha \sim \mathcal{B}eta(0.5, 0.5)$$

- une loi impropre pour $\theta$ et $\lambda$

$$\pi(\boldsymbol{\theta}, \lambda^2) \propto \frac{1}{\left(\lambda^2\right)^{\frac{d}{2}+1}}, \qquad d = 2,$$

- un processus gaussien pour $\delta$

$$\delta(X) \sim \mathcal{GP}(0_n, \Sigma_\delta); \quad \Sigma_\delta = (\frac{\lambda^2}{k})\mathsf{Corr}_{\gamma_\delta}(x_i, x_{i'})$$

$$\mathsf{Corr}_{\gamma_\delta}(x_i, x_{i'}) = \exp\left(-\frac{|x_i - x_{i'}|}{\gamma_\delta}\right).$$

- pour les hyperparamètres $k$ and $\gamma_\delta$

$$k \sim \mathcal{B}(1, 1), \quad \gamma_\delta \sim \mathcal{U}(0.1, 1)$$

# Objectifs et cadre méthodologique

**Inférence bayésienne du modèle**

- Après estimation du modèle de mélange, la composante ayant le poids a posteriori le plus important est considérée comme celle qui s'ajuste le mieux aux données.
- De façon équivalente, la postérieure de $\alpha$ porte la décision : une masse a posteriori proche de $1$ soutient $M_0$, tandis qu'une masse proche de $0$ soutient $M_1$.
- Le terme de discrépance $\delta(x) = r(x) - f(x, \theta)$ mesure dans quelle mesure le code reproduit correctement le système réel en présence du bruit de mesure $\epsilon$.

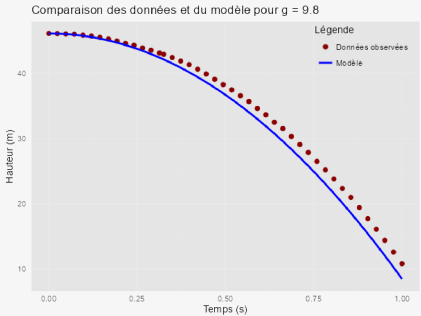
# Inférence sur les données réelles et le modèle physique

$$f(t, \theta = (g, h_0)) = h_0 - \frac{1}{2} g t^2$$

$$\begin{cases} h_0 : \text{hauteur initiale (en mètres).} \\ g : \text{accélération de la pesanteur } (\mathrm{m/s^2}). \end{cases}$$

## Jeu de données

- Source : *Blue Basketball*
- Nombre d'observations : $n = 45$
- Temps $t \in [0.00,\ 2.77]$ s $\Rightarrow$ normalisé en $[0, 1]$
- Hauteur $y \in [10.83,\ 46.45]$ m
- Inférence par MCMC : $10\,000$ itérations




Ce jeu de données réel sert à comparer le comportement du modèle sans seuil et avec seuil.

# Illustrations et données synthétiques

**Simulation à partir de modèle $M_0$ (sans discrépance) :**

- Données synthétiques

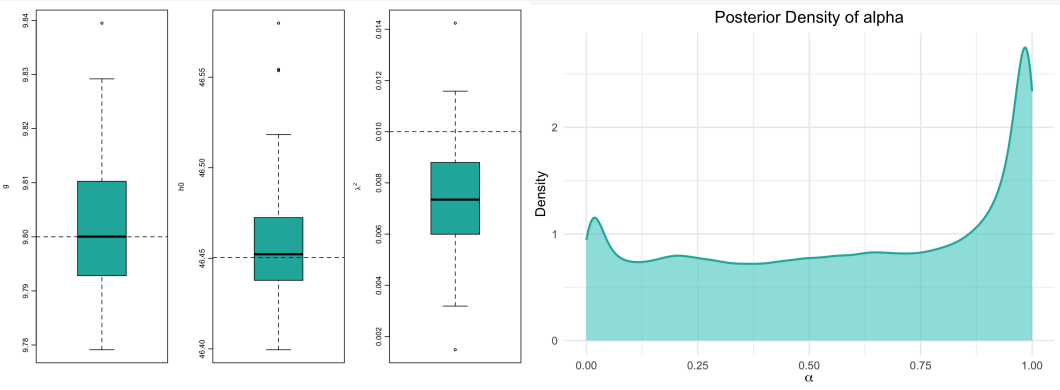
$$y_i = f(t_i)\theta^* + \epsilon_i, \quad \theta^* = (g = 9.8, h_0 = 46.45), \quad \lambda^{2*} = 0.01$$

- Génération de **50** jeux de données indépendantes de taille $n = 45$.
- Pour chaque jeu de données,

Chaîne de $10\,000$ itérations de Metropolis-within-Gibbs

# Illustrations et données synthétiques

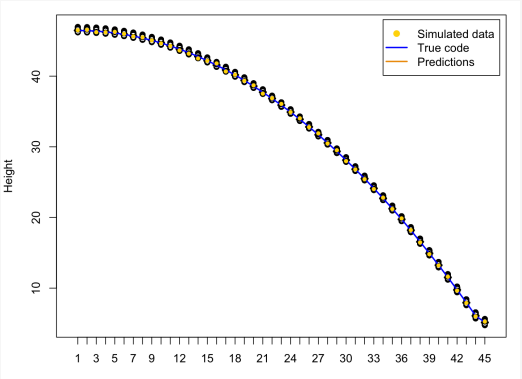
**Simulation à partir de modèle $M_0$ (sans discrépance)**



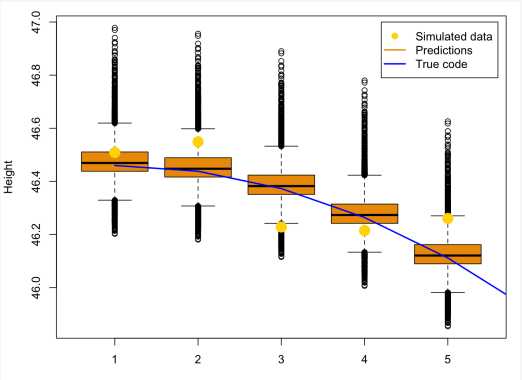

Estimations ponctuelles (moyenne a posteriori) des paramètres pour 50 jeux de données.

# Illustration et données synthétiques

**Simulation à partir de modèle $M_0$ (sans discrépance)**




Comparaison entre un jeu de données simulées avec les distributions des prédictions a posteriori.




Zoom du même graphe présenté à gauche pour les 5 premières données

# Illustration et données synthétiques

**Simulation à partir de modèle $M_1$ (avec discrépance)**

- **Données simulées selon**

$$y_i = f(t_i, \theta^*) + \delta(t_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \lambda^2)$$

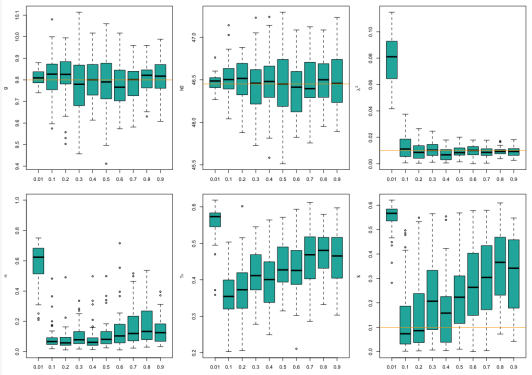
$$\delta \sim \mathcal{GP}(0, \Sigma_\delta), \qquad \Sigma_\delta = \frac{\lambda^{2*}}{k^*} \exp\left(-\frac{|t_i - t_j|}{\gamma_\delta^*}\right).$$

- **Paramètres utilisés :**
  - $\rightarrow$ $k = 0.1, \lambda^2 = 0.01 \implies \sigma_\delta^2 = \frac{0.01}{0.1} = 0.1.$
  - $\rightarrow$ Vecteur $\gamma_\delta \in \{0.01, 0.1, 0.2, \ldots, 0.9\}$ (10 valeurs).

- **Protocole de simulation :**
  - $\rightarrow$ Génération de $50$ jeux de données indépendants, chacun de taille $n = 45$.
  - $\rightarrow$ Estimation par MCMC : $10\,000$ itérations.

# Illustration et données synthétiques

**Simulation à partir de modèle $M_1$ (avec discrépance)**



Pour des valeurs modérées de $\gamma_\delta^*$, la postérieure de $\alpha$ se concentre près de $0$, ce qui favorise le modèle avec discrépance.

# Diagnostic local par seuil

**Seuil pour $\delta$ :**

$$\delta = \begin{cases} \delta & \text{si } |\delta| > s \\ 0 & \text{sinon} \end{cases}$$

Ce seuil est appliqué pour éliminer les petites variations dans les données.

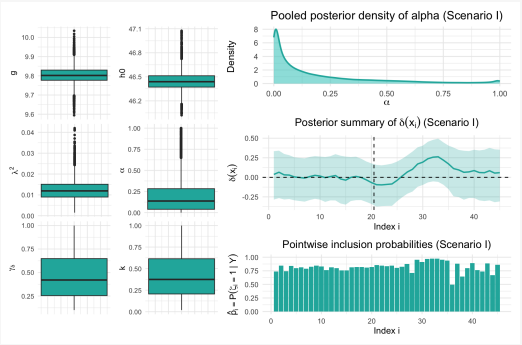
**Sélection du modèle :**

$$P(\zeta_i = 0) = \alpha' \quad \text{et} \quad P(\zeta_i = 1) = 1 - \alpha'$$

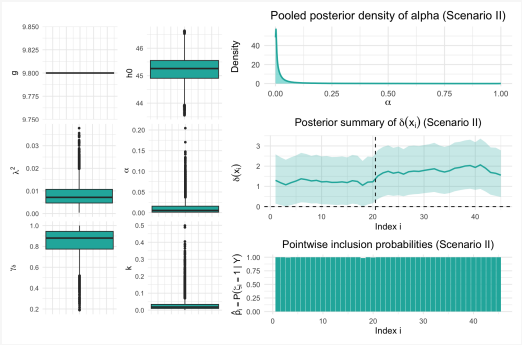
**Application du seuil à $\zeta_i$ :** $\begin{cases} P(\zeta_i = 0|\delta) = \alpha' \cdot \mathbf{1}_{\{|\delta(x_i)|>s\}} + \mathbf{1}_{\{|\delta(x_i)|\leq s\}} \\ P(\zeta_i = 1|\delta) = (1-\alpha') \cdot \mathbf{1}_{\{|\delta(x_i)|>s\}} \end{cases}$

Cela signifie que si $|\delta(x_i)| > s$, alors avec probabilité $\alpha'$, on choisit le modèle avec erreur (c'est-à-dire $\zeta_i = 1$) ; sinon, $\zeta_i = 0$ avec une probabilité de 1 lorsque $|\delta(x_i)| \leq s$.

# Simulation : sans seuil



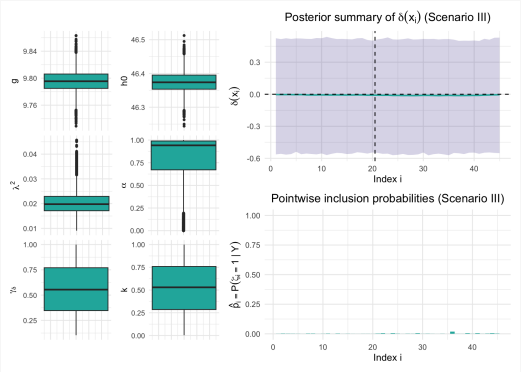

Tous les paramètres sont libres.



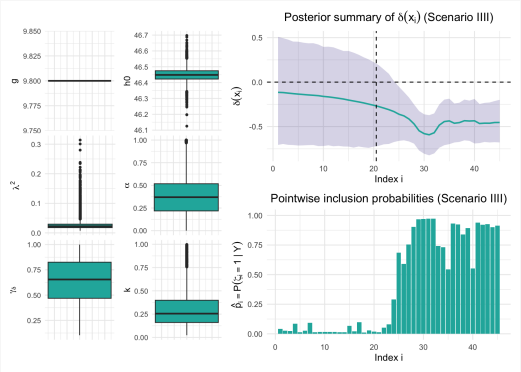

g = fixe, seuil = FALSE

Sans seuil : la discrépance est détectée globalement, mais reste mal localisée.

# Simulation : avec seuil




g = libre, seuil = TRUE

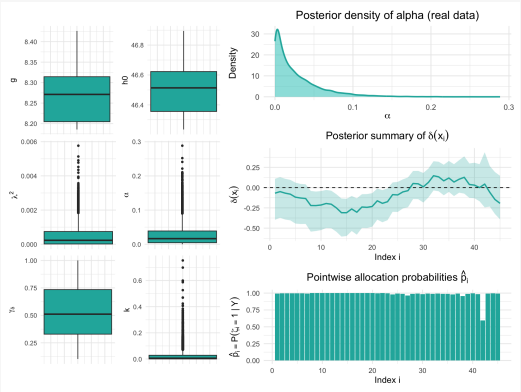



g = fixe, seuil = TRUE

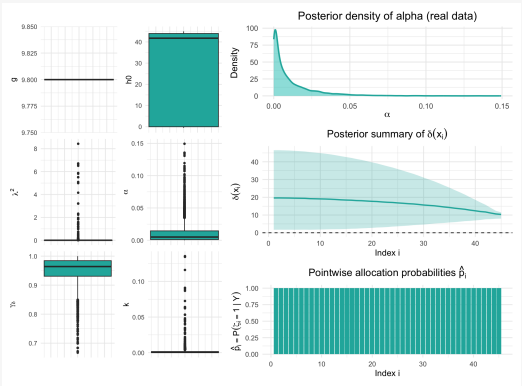
Avec seuil : les probabilités $\hat{p}_i$ localisent mieux la zone de défaut.

# Inférence sur les données réelles

## Données réelles : Sans seuil



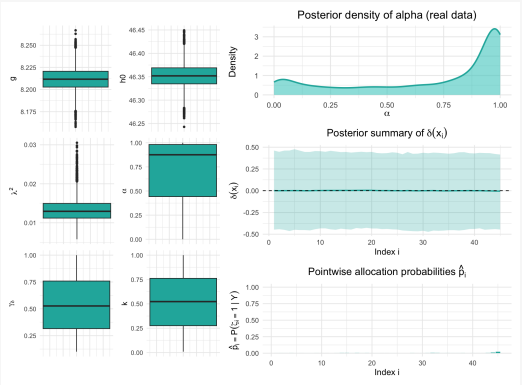

Tous les paramètres sont libres.



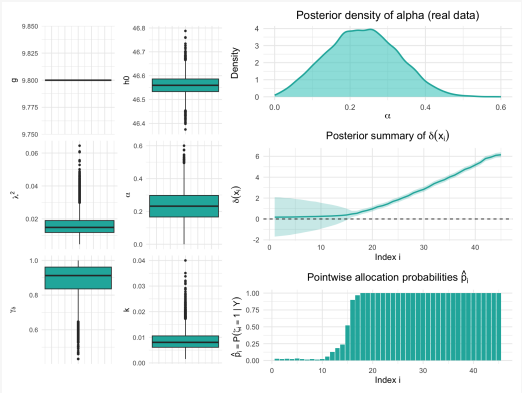

g = fixe, seuil = FALSE

# Inférence sur les données réelles

## Données réelles : avec seuil




g = libre, seuil = TRUE.




g = fixe, seuil = TRUE

# Conclusion

- L'estimation par modèle de mélange permet une validation bayésienne flexible du code.
- En simulation, la méthode discrimine correctement entre $M_0$ et $M_1$.
- Le seuil est surtout utile comme diagnostic local pour repérer où le modèle physique échoue.

**Perspective :**

- extension aux codes non linéaires et coûteux en temps de calcul.

# Références I


[1] Guillaume Damblin et al. « Bayesian Model Selection for the Validation of Computer Codes ». In : *Quality and Reliability Engineering International* 32.6 (2016). Published online 13 July 2016 in Wiley Online Library, p. 1981-2000. doi : 10.1002/qre.2036. url : https://doi.org/10.1002/qre.2036.

[2] Kamari Kamary, J. E. Lee et C. P. Robert. *Weakly informative reparameterisations for location–scale mixtures*. arXiv preprint arXiv :1601.01178 [stat.ME]. 2017.

[3] Kamari Kamary et al. *Testing hypotheses via a mixture estimation model*. arXiv preprint arXiv :1412.2044. 2014.

[4] Jean-Michel Marin et Christian P. Robert. *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. 1re éd. Springer Science & Business Media, 2007.

[5] M. Plumlee. « Bayesian calibration of inexact computer models ». In : *Journal of the American Statistical Association* 112 (2017), p. 1274-1285. doi : 10.1080/01621459.2016.1211016. url : https://doi.org/10.1080/01621459.2016.1211016.

[6] Christian P. Robert. *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation*. 2e éd. New York : Springer-Verlag, 2007.