Problèmes inverses et test bayésien d'adéquation du modèle

Pierre Palud

doctorant en 3e année sous la supervision de
Pierre Chainais, Franck Le Petit

avec la collaboration de
Emeric Bron, Pierre-Antoine Thouvenin

Ecole Centrale de Lille, CRIStAL, LERMA
funded by CNRS via 80|Prime

Vue d'ensemble

modèle génératif / modèle d'observation $\mathcal{M}$

- ★ dépend d'un paramètre Θ inféré à partir d'obs. Y
- ★ génère des reproduction d'obs. $\widetilde{Y} \sim \pi(\widetilde{Y}|\Theta, \mathcal{M})$
- ★ évalue la pdf / le score $\pi(\widetilde{Y}|\Theta, \mathcal{M})$

Générer un modèle $\mathcal{M}$: souvent coûteux.

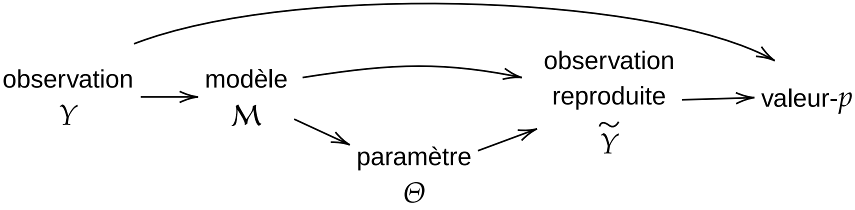
Le modèle $\mathcal{M}$ est-il valide ?

Permet-il de reproduire des observations $\widetilde{Y}$ aussi vraisemblables que les vraies observations Y ?

Vue d'ensemble

$$
\begin{tikzcd}
\text{observation } Y \arrow[r] \arrow[rrrr, bend left] & \text{modèle } \mathcal{M} \arrow[rr, bend left] \arrow[dr] & & \text{observation reproduite } \widetilde{Y} \arrow[r] & \text{valeur-}p \\
& & \text{paramètre } \Theta \arrow[ur] & &
\end{tikzcd}
$$

Vue d'ensemble



utilisation de la valeur-p :

- ★ hypothèse nulle $\mathcal{H}_0$: "$\mathcal{M}$ permet de reproduire Y"
- ★ déf. a priori d'un niveau de confiance α, par ex. 0.05
- ★ calcul valeur-p
- ★ si $p \leq \alpha$: rejet $\mathcal{H}_0$ (niveau de confiance $1 - \alpha$)

Distribution a posteriori prédictive

- observations reproduites : $\widetilde{Y}|\Theta \sim \pi(Y|\Theta, \mathcal{M})$

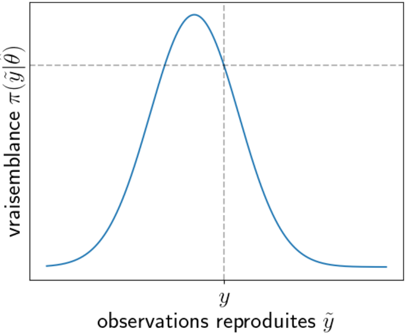
- distribution a posteriori prédictive :

$$\pi\left(\widetilde{Y}|Y, \mathcal{M}\right) = \int_{\Theta} \pi\left(\widetilde{Y}|\Theta, \mathcal{M}\right) \pi\left(\Theta|Y, \mathcal{M}\right) d\Theta$$

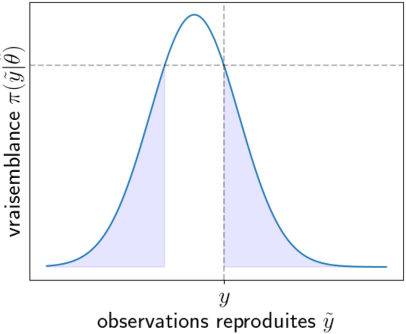
- si $\pi\left(\Theta|Y, \mathcal{M}\right) = \delta_{\widehat{\Theta}}(\Theta)$:

$$\pi\left(\widetilde{Y}|Y, \mathcal{M}\right) = \pi\left(\widetilde{Y}|\Theta, \mathcal{M}\right)$$

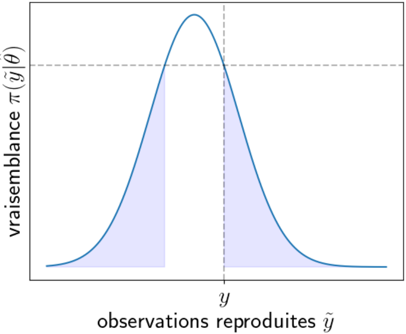
$\pi(\widetilde{Y} = Y | Y, \mathcal{M})$ sans rejet de modèle



$\pi(\widetilde{Y} = Y|Y, \mathcal{M})$ sans rejet de modèle



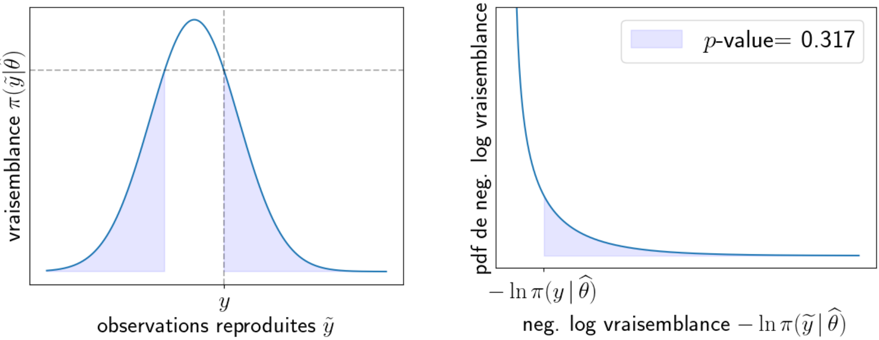
$\pi(\widetilde{Y} = Y | Y, \mathcal{M})$ sans rejet de modèle



$$p = \int_{\widetilde{Y}} 1_I(\widetilde{Y}, \widehat{\Theta})\, \pi\left(\widetilde{Y}|\widehat{\Theta}, \mathcal{M}\right) d\widetilde{Y}$$

avec $I = \{(\widetilde{Y}, \Theta) \,|\, \pi(\widetilde{Y}|\Theta, \mathcal{M}) \leq \pi(Y|\Theta, \mathcal{M})\}$

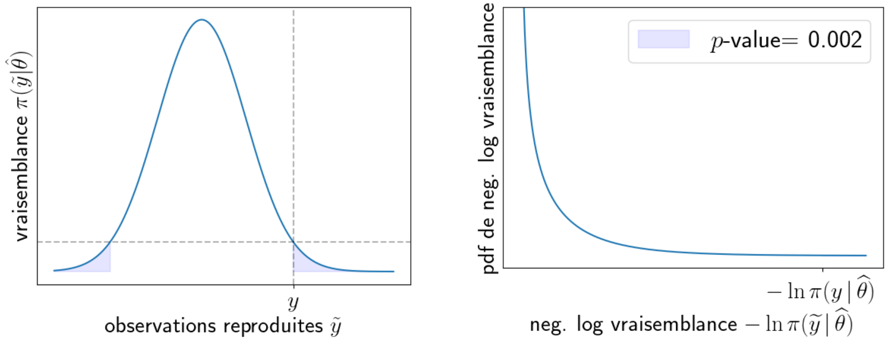
$\pi(\widetilde{Y} = Y | Y, \mathcal{M})$ sans rejet de modèle



$$p = \int_{\widetilde{Y}} 1_I(\widetilde{Y}, \widehat{\Theta}) \, \pi\left(\widetilde{Y} | \widehat{\Theta}, \mathcal{M}\right) d\widetilde{Y}$$

avec $I = \{(\widetilde{Y}, \Theta) \,|\, \pi(\widetilde{Y}|\Theta, \mathcal{M}) \leq \pi(Y|\Theta, \mathcal{M})\}$

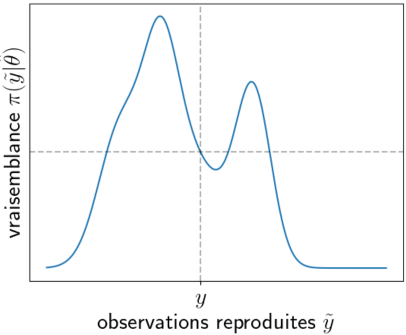
$\pi(\widetilde{Y} = Y | Y, \mathcal{M})$ avec rejet de modèle



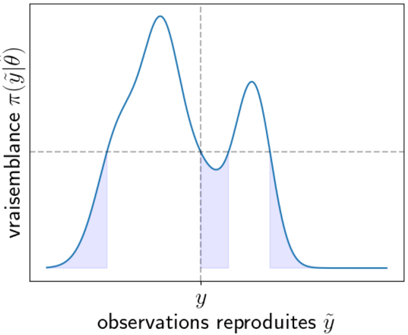
$$p \;=\; \int_{\widetilde{Y}} 1_I(\widetilde{Y}, \widehat{\Theta})\, \pi\left(\widetilde{Y}|\widehat{\Theta}, \mathcal{M}\right)\, d\widetilde{Y}$$

avec $I = \{(\widetilde{Y}, \Theta) \,|\, \pi(\widetilde{Y}|\Theta, \mathcal{M}) \leq \pi(Y|\Theta, \mathcal{M})\}$

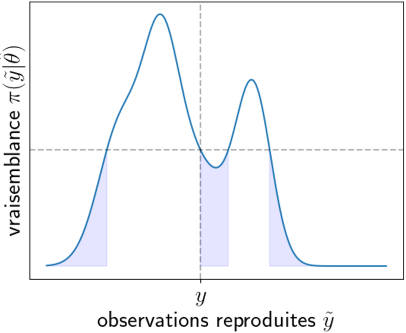
Cas général


vraisemblance $\pi(\tilde{y}|\hat{\theta})$
y
observations reproduites $\tilde{y}$

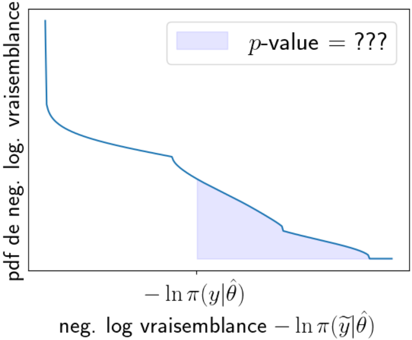
Cas général


vraisemblance $\pi(\tilde{y}|\hat{\theta})$
y
observations reproduites $\tilde{y}$

Cas général

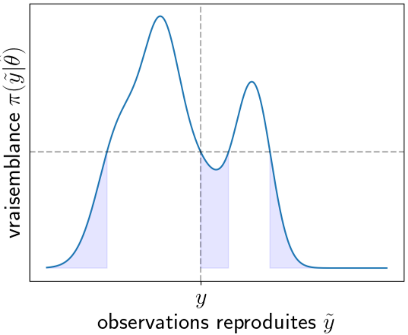


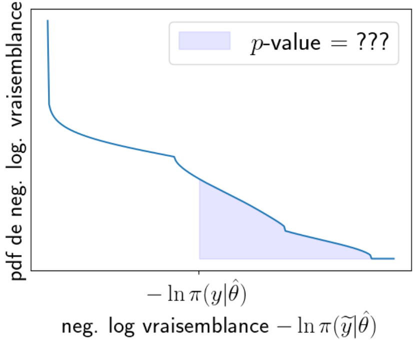
vraisemblance $\pi(\tilde{y}|\hat{\theta})$
y
observations reproduites $\tilde{y}$



pdf de neg. log. vraisemblance
p-value $= ???$
$-\ln \pi(y|\hat{\theta})$
neg. log vraisemblance $-\ln \pi(\tilde{y}|\hat{\theta})$

Cas général





comment calculer valeur-p si modèle non gaussien ?

ou si estimateur non ponctuel ?

Cas général : valeur-p bayésienne

★ si $\pi\left(\Theta|Y,\mathcal{M}\right) \neq \delta_{\widehat{\Theta}}(\Theta)$, valeur-$p$ bayésienne

$$p \;=\; \int_{\widetilde{Y}} \int_{\Theta} 1_I(\widetilde{Y},\Theta)\, \pi\left(\widetilde{Y}|\Theta,\mathcal{M}\right)\, \pi\left(\Theta|Y,\mathcal{M}\right)\, d\Theta\; d\widetilde{Y}$$

avec $I = \{(\widetilde{Y},\Theta) \mid \pi(\widetilde{Y}|\Theta,\mathcal{M}) \leq \pi(Y|\Theta,\mathcal{M})\}$

Cas général : valeur-p bayésienne

- ★ si $\pi\left(\Theta|Y,\mathcal{M}\right) \neq \delta_{\widehat{\Theta}}(\Theta)$, valeur-$p$ bayésienne

$$p \;=\; \int_{\widetilde{Y}}\int_{\Theta} 1_I(\widetilde{Y},\Theta)\;\pi\left(\widetilde{Y}|\Theta,\mathcal{M}\right)\;\pi\left(\Theta|Y,\mathcal{M}\right)\,d\Theta\;d\widetilde{Y}$$

avec $I = \{(\widetilde{Y},\Theta) \mid \pi(\widetilde{Y}|\Theta,\mathcal{M}) \leq \pi(Y|\Theta,\mathcal{M})\}$

- ★ estimateur Monte Carlo $\widehat{p}^{(t)}$ (Gelman et al., 1996)

$$\widehat{p}^{(t)} \;=\; \frac{1}{t}\sum_{\tau=1}^{t} 1_I\left(\widetilde{Y}^{(\tau)},\Theta^{(\tau)}\right)$$

avec $\Theta^{(\tau)} \sim \pi\left(\Theta|Y,\mathcal{M}\right)$ et $\widetilde{Y}^{(\tau)} \sim \pi(\widetilde{Y}|\Theta^{(\tau)},\mathcal{M})$

Cas général : valeur-p bayésienne

- si $\pi\left(\Theta|Y,\mathcal{M}\right) \neq \delta_{\widehat{\Theta}}(\Theta)$, valeur-$p$ bayésienne

$$p \;=\; \int_{\widetilde{Y}}\int_{\Theta} 1_I(\widetilde{Y},\Theta)\;\pi\left(\widetilde{Y}|\Theta,\mathcal{M}\right)\;\pi\left(\Theta|Y,\mathcal{M}\right)\,d\Theta\;d\widetilde{Y}$$

avec $I = \{(\widetilde{Y},\Theta)\mid \pi(\widetilde{Y}|\Theta,\mathcal{M}) \le \pi(Y|\Theta,\mathcal{M})\}$

- estimateur Monte Carlo $\widehat{p}^{(t)}$ (Gelman et al., 1996)

$$\widehat{p}^{(t)} \;=\; \frac{1}{t}\sum_{\tau=1}^{t} 1_I\left(\widetilde{Y}^{(\tau)},\Theta^{(\tau)}\right)$$

avec $\Theta^{(\tau)} \sim \pi\left(\Theta|Y,\mathcal{M}\right)$ et $\widetilde{Y}^{(\tau)} \sim \pi(\widetilde{Y}|\Theta^{(\tau)},\mathcal{M})$

- intégrale haute dim. $\rightarrow$ **erreur** sur estimation ?

Erreur sur l'estimation MC de la valeur-p

★ $p = \mathbb{P}_{(\widetilde{Y},\Theta)}\left[1_I\left(\widetilde{Y},\Theta\right)=1\right] \rightarrow$ test Bernoulli sur 1_I

★ loi Bêta-binomiale pour quantifier l'incertitude sur p [1]

$$p^{(t)}\Big|\left\{\widetilde{Y}^{(t)},\Theta^{(t)}\right\} \sim \text{Bêta}\left(1+T^{(t)}\,\widehat{p}^{(t)}, 1+T^{(t)}(1-\widehat{p}^{(t)})\right)$$

★ proba. rejet $= \mathbb{P}_{p^{(t)}\big|\left\{\widetilde{Y}^{(t)},\Theta^{(t)}\right\}}\left[p^{(t)} \leq \alpha\right] = \text{cdf}(\alpha)$

facile à calculer ✓

[1] pour un estimateur ponctuel $\widehat{\Theta}$, $T^{(t)} = t$
pour considérer la corrélation dans MCMC $T^{(t)} = \text{ESS}^{(t)}$

Test robuste

★ seuil $\delta \in [0, \frac{1}{2}] \to$ 3 cas

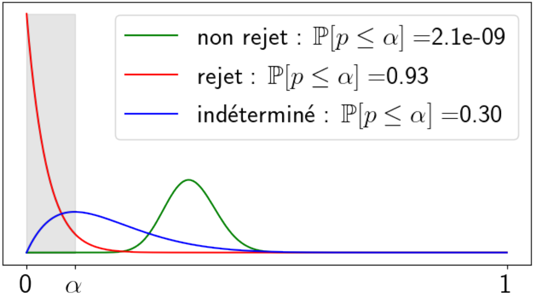
$$\begin{cases} \mathbb{P}\left[p^{(t)} \leq \alpha\right] = \mathrm{cdf}(\alpha) \leq \delta & \Longrightarrow \text{ non rejet ✓} \\ \mathbb{P}\left[p^{(t)} \leq \alpha\right] = \mathrm{cdf}(\alpha) \geq 1 - \delta & \Longrightarrow \text{ rejet ✗} \\ \text{otherwise} & \Longrightarrow \text{ indéterminé ?} \end{cases}$$

Test robuste

★ seuil $\delta \in [0, \frac{1}{2}] \to$ 3 cas

$$\begin{cases} \mathbb{P}\left[p^{(t)} \leq \alpha\right] = \mathrm{cdf}(\alpha) \leq \delta & \Longrightarrow \text{ non rejet ✓} \\ \mathbb{P}\left[p^{(t)} \leq \alpha\right] = \mathrm{cdf}(\alpha) \geq 1 - \delta & \Longrightarrow \text{ rejet ✗} \\ \text{otherwise} & \Longrightarrow \text{ indéterminé ?} \end{cases}$$

★ Par exemple, pour $\delta = 0.1$,



Application : cas synthétique astro.

Introduit et résolu dans Palud et al. (2023)

★ $\Theta = (\boldsymbol{\theta}_n) \in \mathbb{R}^{100 \times 4}, \quad Y = (\boldsymbol{y}_n) \in \mathbb{R}^{100 \times 10}$

vraisemblance $\pi(Y | \Theta, \mathcal{M})$

★ modèle direct $\mathbf{f} : \boldsymbol{\theta} \in \mathbb{R}^4 \mapsto \mathbf{f}(\boldsymbol{\theta}) \in \mathbb{R}^{10}$ non linéaire

★ modèle bruit : non gaussien, **i.i.d. par pixel**, censure

distribution a priori $\pi(\Theta)$

★ régularisation spatiale + ensemble validité sur $\boldsymbol{\theta}_n$

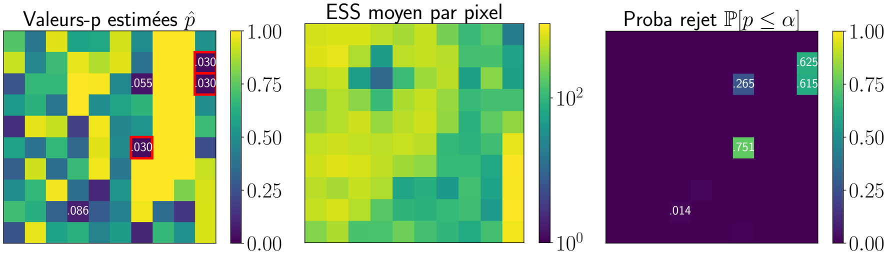
distribution a posteriori $\pi(\Theta | Y, \mathcal{M})$

★ neg log pdf : $\mathscr{C}^2$, non gradient-Lipschitz

★ non convex, multimodale

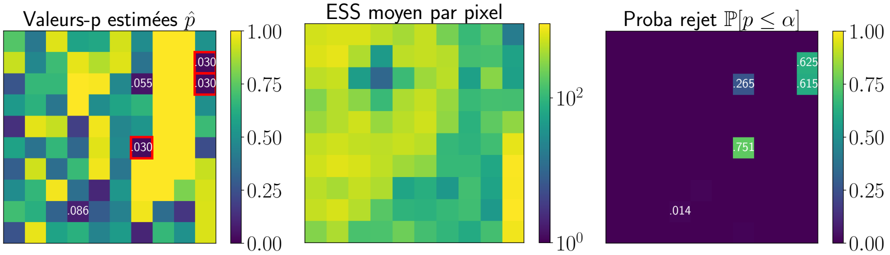
Application : cas synthétique astro.

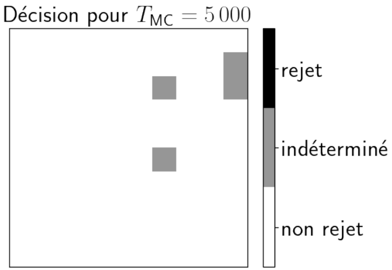
Chaîne avec $T_{\mathrm{MC}} = 5\,000$



Application : cas synthétique astro.

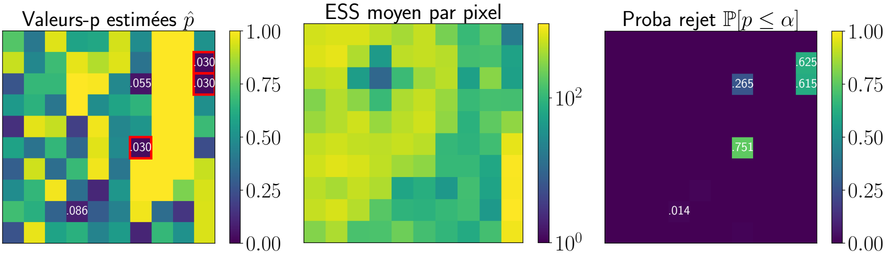
Chaîne avec $T_{\text{MC}} = 5\,000$

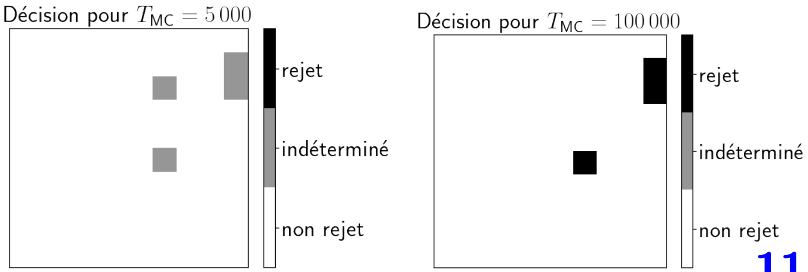




Application : cas synthétique astro.

Chaîne avec $T_{\text{MC}} = 5\,000$





Conclusion

valeur-p bayésienne et estimateur Monte Carlo

vérifier **validité modèle** $\mathcal{M}$ et $\pi(Y|\Theta, \mathcal{M})$, par ex.,

- ★ un modèle génératif
- ★ un modèle d'obs. dans un problème inverse

incertitude liée à estimateur Monte Carlo

- ★ loi Bêta-binomiale — calculs simples ✓
- ★ cas "indéterminé" — test robuste aux cas limites ✓
- ★ valeurs-p sur $\widetilde{\boldsymbol{y}}_n|\Theta$ — diagnostic + fin[2] ✓

application : problème inverse complexe synthétique ✓

[2] peut être pertinent ou non, selon le modèle de bruit

Conclusion

Once we have accomplished the first two steps of a Bayesian analy sis – constructing a probability model and computing the posterior distribution of all estimands – **we should not ignore the relatively easy step of assessing the fit of the model to the data and to our substantive knowledge**.

(Gelman et al., 2015)

References I

- Gelman, Andrew, Xiao-Li Meng, and Hal Stern (1996). "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies". In: *Statistica Sinica* 6.4, pp. 733–760.
- Gelman, Andrew et al. (July 2015). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC.
- Palud, Pierre, Pierre-Antoine Thouvenin, Pierre Chainais, Emeric Bron, and Franck Le Petit (2023). "Efficient Sampling of Non Log-Concave Posterior Distributions With Mixture of Noises". In: *IEEE Transactions on Signal Processing* 71, pp. 2491–2501.