Approches Bayésiennes pour les problèmes inverses

PIM 2024 - Angèle Niclas

Université Paris Cité - MAP5

Plan du cours

Références

- H. Kekkonen, Bayesian inverse problems (2019)
  Cours donné à Cambridge
- R. Scheichl, J. Zech, Numerical Methods for Bayesian Inverse Problems (2021)
  Cours donné à Heidelberg
- A. Tarantola, Inverse Problem Theory (2005)

Introduction et Motivation

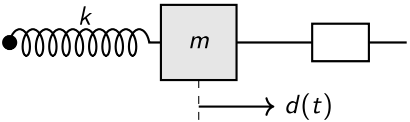
Historique

- Thomas Bayes (1763), "An Essay Towards Solving a Problem in the Doctrine of Chances"
- Pierre-Simon Laplace (1786), ratio des sexes
- Marginalisation au 19ème et 20ème siècle en faveur de l'approche fréquentiste (Ronald Fisher)
- Renaissance autour de 1960 (Harold Jeffreys)
- Essor des algorithmes MCMC autour de 1990

Applications :

- Prise de décision (essais cliniques)
- IA (classification et reconnaissance de motifs, spam)
- Finance (optimisation des portefeuilles)
- Optimisation numérique

Oscillateur harmonique amorti



- k raideur du ressort
- m masse de l'objet
- τ temps de relaxation

Si $\omega_0 = \sqrt{k/m}$, on a pour $t \geq 0$,

$$d''(t) + \frac{2}{\tau}\, d'(t) + {\omega_0}^2\, d(t) = 0$$

Il existe $A, \phi \in \mathbb{R}^2$ tels que $d(t) =$

$$Ae^{-t/\tau} \cos\left(\omega_0 \sqrt{1 - \frac{1}{\tau^2 {\omega_0}^2}}\, t + \phi\right)$$

On mesure

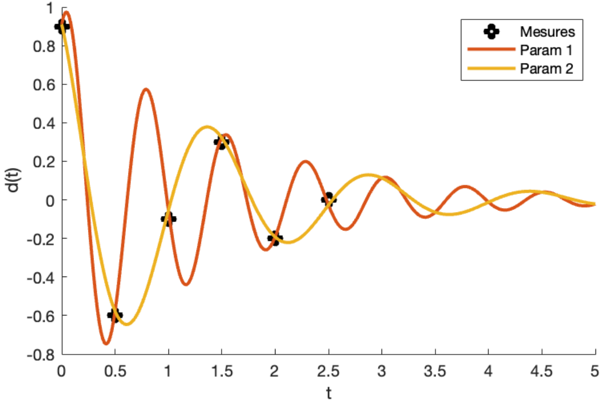
$$\begin{array}{ll} d(0) = 0.9 & d(0.5) = -0.6 \\ d(1) = -0.1 & d(1.5) = 0.3 \\ d(2) = -0.2 & d(2.5) = 0 \end{array}$$

On sait que

- Les capteurs ont une incertitude de 0.05
- $\omega_0 \approx 4$ rad/s
- τ est inférieur à 3 s

Objectif :

Retrouver ω_0 et τ.



- Pas unicité de la solution
- Incorporer les informations supplémentaires
- Tenir compte du bruit de mesure
- Quantifier la précision du résultat

Cadre général (déterministe)

Travail en dimension finie.

Définition

On s'intéresse à résoudre le problème inverse $y = A(x)$, où :

- Inconnues : $x \in \mathbb{R}^n$, où n est le nombre total d'inconnues
- Mesures : $y \in \mathbb{R}^m$, où m représente le nombre de mesures
- Modèle : opérateur $A : \mathbb{R}^n \to \mathbb{R}^m$.

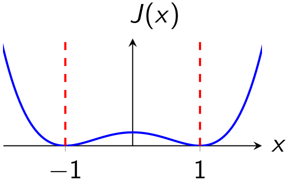
Exercice

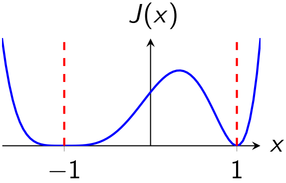
Dans le cas de l'oscillateur harmonique amorti, identifiez les vecteurs x, y et l'opérateur A.

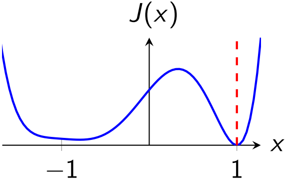
Minimisation par moindres carrés :

$$x \in \text{argmin}_{x\in\mathbb{R}^n} J(x), \qquad J(x) = \|A(x) - y\|$$

Quelques exemples :







Pénalisation avec un paramètre λ :

$$\widetilde{J}(x) = J(x) + \lambda\|x_4 - 4\|$$

Rappels de probabilité

Espace probabilisé

Espace probabilisé

Un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ est constitué de

- Un univers Ω qui représente l'ensemble des résultats possibles de l'expérience considérée.
- Une famille $\mathcal{F}$ de parties de Ω, appelée σ-algèbre ou tribu.
- Une fonction $\mathbb{P} : \mathcal{F} \to [0,1]$, appelée mesure de probabilité, telle que $\mathbb{P}(\Omega) = 1$ et qui vérifie la propriété suivante : pour tout $(A_n)_{n\in\mathbb{N}} \in \mathcal{F}^{\mathbb{N}}$ avec $A_i \cap A_j$ si $i \neq j$, alors

$$\mathbb{P}\left(\cup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

Exercice

Proposer un univers Ω adapté à l'exemple de l'oscillateur harmonique amorti.

Choix pour $\mathcal{F}$: si $\Omega = \mathbb{R}^n$ (ou un sous-ensemble de $\mathbb{R}^n$), la tribu la plus couramment utilisée est la tribu borélienne $\mathcal{B}(\mathbb{R}^n)$.

Choix pour $\mathbb{P}$:

Densité de probabilité

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction positive, mesurable et telle que son intégrale sur $\mathbb{R}^n$ soit égale à 1. On définit une probabilité $\mathbb{P}$ en posant

$$\forall A \in \mathcal{B}(\mathbb{R}^n) \qquad \mathbb{P}(A) = \int_A f(x)\mathrm{d}x$$

Une telle probabilité est dite à densité, et f est appelée la densité de probabilité de $\mathbb{P}$.

Exercice

Donner la densité de probabilité pour une loi uniforme sur l'intervalle $[a, b]$, notée $\mathcal{U}(a, b)$. Idem pour une loi normale d'espérance μ et d'écart type σ, notée $\mathcal{N}(\mu, \sigma)$.

Presque partout

Une proposition est dite vraie presque partout (abrégé p.p.) si elle est vraie sur l'ensemble $\Omega \setminus A$, où $A \in \mathcal{F}$ est un ensemble tel que la mesure de probabilité de A est nulle, c'est-à-dire $\mathbb{P}(A) = 0$.

Exemple : La fonction $x \mapsto 1/x$ est définie presque partout dans $\mathbb{R}$.

Variable aléatoire

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et V un espace de Banach. Une fonction mesurable $X : \Omega \to V$ est appelée une variable aléatoire (VA).

On définit une nouvelle probabilité $\mathbb{P}_X$ par

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) \in B\})$$

On dit que X suit une loi μ (noté $X \sim \mu$) si $\mathbb{P}_X = \mu$. Si $\mathbb{P}_X$ est à densité, on note sa densité par π_X.

Loi jointe et marginales

Pour deux VA X et Y, la loi jointe de (X, Y) est définie par :

$$\mathbb{P}_{(X,Y)}(A \times B) = \mathbb{P}(X \in A, Y \in B)$$

Les lois marginales de X et Y sont les lois obtenues en considérant les distributions individuelles de X et Y respectivement.

Espérance et Variance

Soit $X : \Omega \to V$ une VA. L'espérance de X existe si $\int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega) < +\infty$ et l'espérance de X, notée $\mathbb{E}(X)$, vaut

$$\mathbb{E}(X) = \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_V x \,\mathrm{d}\mathbb{P}_X(x)$$

Si elle existe, la variance de X, notée $V(X)$ ou $\mathrm{Var}(X)$, vaut

$$V(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$$

Probabilités conditionnelles

Probabilité conditionnelle

Soient $A, B \in \mathcal{F}$ avec $\mathbb{P}(B) > 0$. La probabilité conditionnelle de A sachant B est définie par :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Si $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, B n'a pas d'influence sur A. On dit que :

Indépendance

Deux événements A et B dans $\mathcal{F}$ sont indépendants si :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Exercice

Soit Y une variable aléatoire qui prend les valeurs $1/3$ ou $2/3$ avec probabilité $1/2$. Soit X une variable aléatoire telle que X prend la valeur 1 avec probabilité Y et la valeur 0 avec probabilité $1 - Y$. Que vaut $\mathbb{P}(X = 1 \mid Y = 1/3)$? Que dire de $\mathbb{P}(X = 1 \mid Y = 1/2)$? Et si $Y \sim \mathcal{U}(0,1)$?

Version régulière de la probabilité conditionnelle

Soit $X : \Omega \to V$, $Y : \Omega \to W$ deux VA sur des espaces de Banach séparables V et W. Une fonction $\pi_{X|Y} : \mathcal{B}(V) \times W \to [0, 1]$ est appelée la version régulière de la probabilité conditionnelle de X sachant Y si elle satisfait les conditions suivantes :

- $y \mapsto \pi_{X|Y}(B, y)$ est mesurable par rapport à $\mathcal{B}(W)$.
- $B \mapsto \pi_{X|Y}(B, y)$ est une distribution de probabilité sur $(V, \mathcal{B}(V))$.
- $\mathbb{P}(X \in B, Y \in A) = \displaystyle\int_A \pi_{X|Y}(B, y) \mathrm{d}\mathbb{P}_Y(y)$

On note $\mathbb{P}(X \in B \mid Y = y) = \pi_{X|Y}(B, y)$.

Formule de Bayes

Soient $A, B \in \mathcal{F}$ deux événements tels que $\mathbb{P}(B) > 0$, alors :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

Formule de Bayes (version régulière)

Soit $X : \Omega \to \mathbb{R}^n$, $Y : \Omega \to \mathbb{R}^m$ deux VA avec densité jointe. Si $\pi_{X|Y}(\cdot \mid y)$ et $\pi_{Y|X}(\cdot \mid x)$ désignent les densités respectives de $\pi_{X|Y}(\cdot, y)$ et $\pi_{Y|X}(\cdot, x)$, alors $\mathbb{P}_Y$-presque partout :

$$\pi_{X|Y}(x \mid y) = \frac{\pi_{Y|X}(y \mid x)\pi_X(x)}{\pi_Y(y)}$$

Cadre Bayésien

Problème inverse (version bayésienne) :

On considère le problème $Y = A(X) + E$, avec les éléments suivants :

- $X : \Omega \to V$ est une VA représentant les paramètres inconnus du problème. Elle suit une loi de densité π_X.
- $E : \Omega \to W$ est une VA modélisant le bruit sur les mesures. Elle est supposée indépendante de X et suit une densité π_E.
- $Y : \Omega \to W$ est une VA correspondant aux mesures effectuées. Elle suit une densité π_Y.
- $A : V \to W$ est un opérateur, appelé opérateur direct.

Prior - vraisemblance - posterior

- Distribution a priori (prior) : $\mathbb{P}_X$ est appelée distribution a priori. Elle représente l'information disponible sur X avant d'effectuer les mesures ou les observations. Sa densité est $\pi_X(x)$.
- Vraisemblance : $\mathbb{P}(Y \mid X = x) = \pi_{Y|X}(\cdot, x)$ est appelée la vraisemblance. Elle modélise la probabilité d'observer les données Y en fonction des valeurs des paramètres X. On suppose qu'elle a pour densité $\pi_{Y|X}(y \mid x)$.
- Distribution a posteriori (posterior) : $\mathbb{P}(X \mid Y = y) = \pi_{X|Y}(\cdot, y)$ est appelée distribution a posteriori. Elle met à jour la distribution de X en tenant compte de la nouvelle observation $Y = y$. On suppose qu'elle a pour densité $\pi_{X|Y}(x \mid y)$.

Estimateurs

Estimateurs

- Maximum de vraisemblance (Maximum Likelihood, ML) :

$$x_{ML} \in \operatorname{argmax}_{x \in V} \pi_{Y|X}(y \mid x)$$

- Maximum a posteriori (MAP) :

$$x_{MAP} \in \operatorname{argmax}_{x \in V} \pi_{X|Y}(x \mid y)$$

- Moyenne conditionnelle (Conditional Mean, CM) :

$$x_{CM} = \mathbb{E}(X \mid Y = y) = \int_V x \, \pi_{X|Y}(x \mid y) \mathrm{d}x$$

Exercice

Soit la fonction $\phi(x)$ définie par :

$$\phi(x) = \mathbf{1}_{0 \leq x \leq 1}(1 - x) + \mathbf{1}_{-1 \leq x < 0}(1 + x)$$

et supposons que, pour un paramètre $y \in (0, 1)$ et des écarts types $\sigma_1, \sigma_2 \in (0, 0.5)$, la densité conditionnelle de X sachant $Y = y$ soit donnée par :

$$\pi_{X|Y}(x \mid y) = \frac{y}{\sigma_1}\phi\left(\frac{x}{\sigma_1}\right) + \frac{1 - y}{\sigma_2}\phi\left(\frac{x - 1}{\sigma_2}\right)$$

Calculer x_{CM} et x_{MAP} en fonction de y, σ_1, et σ_2.
Calculer la variance σ^2 de la distribution postérieure $\pi_{X|Y}$.

Expression de la loi à postériori

Avec les hypothèses du théorème de Bayes, on a $\mathbb{P}_Y$-presque partout :

$$\pi_{X|Y}(x \mid y) = \frac{1}{Z(y)} \pi_E(y - A(x)) \pi_X(x)$$

où

$$Z(y) = \int_{\mathbb{R}^n} \pi_E(y - A(x)) \pi_X(x) \mathrm{d}x$$

Expression de la loi à postériori

Avec les hypothèses du théorème de Bayes, on a $\mathbb{P}_Y$-presque partout :

$$\pi_{X|Y}(x \mid y) = \frac{1}{Z(y)}\pi_E(y - A(x))\pi_X(x)$$

où

$$Z(y) = \int_{\mathbb{R}^n} \pi_E(y - A(x))\pi_X(x)\mathrm{d}x$$

Exercice

Soit $E \sim \mathcal{N}(0, I_m)$ et supposons que l'information a priori sur X est $X \sim \mathcal{N}(0, I_n/\alpha)$, où $\alpha > 0$. Déterminez les expressions des estimateurs x_{ML}, x_{MAP} et x_{CM}.

Choix de priors

Priors informatifs : Ces priors reposent sur une connaissance préalable spécifique de l'objet étudié.

- Prior gaussien
- Prior Gamma : $\pi_X(x) \propto x^{\alpha-1} \exp(-x/\beta)$
- Prior de Wishart

Priors non (ou peu) informatifs : Ces priors sont utilisés lorsque l'on dispose de peu ou pas d'informations préalables sur les paramètres à estimer.

- Prior uniforme
- Prior de Jeffreys

Autres considérations : Priors conjugués

Exercice

Proposer des priors dans le cas de l'oscillateur harmonique amorti.

Méthodes MCMC

Distribution cible : Π, de densité π, avec une version non normalisée μ

Objectif :

- Visualiser Π sans trop d'évaluations de μ
- Calculer x_{CM}

Méthode de Monte-Carlo

Si des points x_j sont tirés selon la loi Π, alors

$$\int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j)$$

Chaîne de Markov

Chaîne de Markov

Une chaîne de Markov sur $\mathbb{R}^n$ est une suite de VA $X_j : \Omega \to \mathbb{R}^n$ telle que, pour tout $j \in \mathbb{N}$ et pour tout $B \in \mathcal{B}(\mathbb{R}^n)$,

$$\mathbb{P}(X_{j+1} \in B \mid X_1, \cdots, X_j) = \mathbb{P}(X_{j+1} \in B \mid X_j) \qquad \mathbb{P} - \text{p.p.}$$

Noyau de transition

Une fonction $K : \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$ est appelée noyau de transition si :

- $x \mapsto K(x, B)$ est mesurable.
- $B \mapsto K(x, B)$ définit une distribution de probabilité sur $\mathcal{B}(\mathbb{R}^n)$.

Une chaîne de Markov $(X_j)_{j\in\mathbb{N}}$ est dite homogène s'il existe un noyau de transition K tel que, pour tout $j \in \mathbb{N}$, $x \in \mathbb{R}^d$ et $B \in \mathcal{B}(\mathbb{R}^d)$, on ait :

$$\mathbb{P}(X_{j+1} \in B \mid X_j = x) = K(x, B) \qquad \mathbb{P} - p.p.$$

Dans la suite, on suppose que le noyau K admet une densité k :

$$\forall B \in \mathcal{B}(\mathbb{R}^d) \quad K(x, B) = \int_B k(x, y)\mathrm{d}y$$

Définition

Soit K un noyau de transition. On définit la probabilité ΠK sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ par :

$$\forall B \in \mathcal{B}(\mathbb{R}^d) \quad (\Pi K)(B) = \int_{\mathbb{R}^d} K(x, B)\pi(x)\mathrm{d}x$$

On définit également, pour $j > 0$, le noyau de transition K^j par récurrence :

$$K^j(x, B) = \int_{\mathbb{R}^d} K^{j-1}(y, B)\, k(x, y)\mathrm{d}y$$

De manière équivalente, si $X_1 \sim \Pi$, alors $X_j \sim \Pi K^{j-1}$.

Invariance - irréductibilité - périodicité

- Π est dite invariante par rapport à K si

$$\Pi = \Pi K$$

- K est dit irréductible par rapport à Π si, pour tout $x \in \mathbb{R}^n$ et tout $B \in \mathcal{B}(\mathbb{R}^n)$ tel que $\Pi(B) > 0$, il existe un entier $j \in \mathbb{N}$ tel que :

$$K^j(x, B) > 0.$$

- K est dit périodique s'il existe un entier $m \geq 2$ et des ensembles disjoints non vides $\{E_1, \ldots, E_m\} \subset \mathbb{R}^n$ tels que :

$$\forall j = 1, \ldots, m, \quad \forall x \in E_j, \quad K(x, E_{(j+1) \mod m}) = 1.$$

Proposition

Soit $X_{j_{j\geq 1}}$ une chaîne de Markov homogène de noyau K. Supposons que Π est invariante par rapport à K et que K est irréductible et apériodique. Alors, pour tout $X \in \mathbb{R}^n$ et tout $B \in \mathcal{B}(\mathbb{R}^n)$, on a :

$$\lim_{N\to+\infty} K^N(X,B) = \Pi(B)$$

De plus, pour toute fonction $f \in \mathsf{L}^1_\Pi(\mathbb{R}^n)$,

$$\lim_{N\to+\infty} \frac{1}{N}\sum_{j=1}^{N} f(X_j) = \int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \qquad \text{p.s.}$$

Réversibilité

Un noyau de transition K est dit Π-réversible s'il satisfait la condition de balance détaillée :

$$\forall x, y \in \mathbb{R}^n, \quad k(x, y)\pi(x) = k(y, x)\pi(y)$$

Proposition

Si K est un noyau de transition Π-réversible, alors Π est invariante par rapport à K.

Metropolis-Hastings

Entrées : Noyau K de densité k, valeur de $x_1 \in \mathbb{R}^n$
Sorties : Réalisations $(x_j)_{j \geq 0}$ de la chaîne de Markov $(X_j)_{j \geq 0}$

Algorithme : Pour $j = 1, 2, \ldots$

- A partir de $x = x_j$, tirer y selon la loi $K(x, \cdot)$
- Calculer la probabilité d'acceptation :

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)k(y, x)}{\pi(x)k(x, y)}\right)$$

- Tirer $t \sim \mathcal{U}(0, 1)$ et définir

$$x_{j+1} = \begin{cases} y & \text{si } t \leq \alpha(x, y) \\ x & \text{sinon} \end{cases}$$

Proposition

Le noyau de la chaîne de Markov associée à cet algorithme, noté $\tilde{K} : \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$, admet une densité $\tilde{k}$ définie par :

$$\tilde{k}(x,y) = \alpha(x,y)k(x,y) + \left(1 - \int_{\mathbb{R}^n} \alpha(x,z)k(x,z)\mathrm{d}z\right) \mathbf{1}_{x=y}$$

Le noyau $\tilde{K}$ est Π-réversible, et si K est irréductible et apériodique, alors $\tilde{K}$ l'est également.

TP d’application