Inférence bayésienne et méthodes de Monte Carlo

Une introduction

Nicolas Dobigeon

University of Toulouse, IRIT/INP-ENSEEIHT
Institut Universitaire de France (IUF)
http://dobigeon.perso.enseeiht.fr

Rencontre GdR ISIS-OG, 8 Octobre 2018

Vous avez dit "*bayésien*" ?

"*Everyone uses Bayesian inference when it is clearly appropriate.*
A Bayesian is someone who uses Bayesian inference
***even** when it might seem inappropriate*."

A. Gelman & C. P. Robert,
"Not Only Defended But Also Applied":
The Perceived Absurdity of Bayesian Inference",
The American Statistician, Vol. 67, No. 1, Feb. 2013.

Plan

Plan

Formulation du problème

Soit une variable aléatoire Y dont la loi[1] $f(y|\theta)$ dépend d'un paramètre inconnu θ.

Définitions
Un ***n*-échantillon** de Y est un n-uplet $(Y_1, Y_2, \ldots, Y_n)$ tel que les Y_i ont la même loi que Y et sont indépendantes.
Une **réalisation** de l'échantillon est alors un n-uplet $(Y_1, Y_2, \ldots, Y_n)$ de valeurs prises par l'échantillon.

Problème
A l'aide d'un échantillon issu de Y, on cherche à déterminer au mieux la vraie valeur du paramètre θ.

[1] Dans cet exposé, par soucis de concision, f désignera une densité de probabilité définissant une variable aléatoire réelle. Les résultats s'étendent simplement à une variable aléatoire discrète définie par des probabilités.

Estimateur et estimation

Définition
Un **estimateur** $\hat{\theta}$ est une variable aléatoire $\hat{\theta}(Y_1, \ldots, Y_n) = \hat{\theta}_n$, fonction (suffisamment) régulière des variables aléatoires $Y_1, \ldots, Y_n$.

$\rightsquigarrow$ La valeur de $\hat{\theta}$ obtenue en remplaçant les Y_i par les réalisations Y_i (valeurs observées) est censée donner une valeur approchée du paramètre θ.
C'est une **estimation** de θ.

Example
Soient $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. Alors

$$\hat{m} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

est **un** estimateur de m.

Estimateur et estimation

Définition
Un **estimateur** $\hat{\theta}$ est une variable aléatoire $\hat{\theta}(Y_1, \ldots, Y_n) = \hat{\theta}_n$, fonction (suffisamment) régulière des variables aléatoires $Y_1, \ldots, Y_n$.

$\rightsquigarrow$ La valeur de $\hat{\theta}$ obtenue en remplaçant les Y_i par les réalisations Y_i (valeurs observées) est censée donner une valeur approchée du paramètre θ.
C'est une **estimation** de θ.

Example
Soient $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. Alors

$$\hat{m} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

est **un** estimateur de m.

Estimateur et estimation

Définition
Un **estimateur** $\hat{\theta}$ est une variable aléatoire $\hat{\theta}(Y_1, \ldots, Y_n) = \hat{\theta}_n$, fonction (suffisamment) régulière des variables aléatoires $Y_1, \ldots, Y_n$.

$\rightsquigarrow$ La valeur de $\hat{\theta}$ obtenue en remplaçant les Y_i par les réalisations Y_i (valeurs observées) est censée donner une valeur approchée du paramètre θ.
C'est une **estimation** de θ.

Example
Soient $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. Alors

$$\hat{m} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

est **un** estimateur de m.

Qualités d’un estimateur

- Absence de biais

$$\mathrm{E}\left[\hat{\theta}_n\right] = \theta$$

  $\rightsquigarrow$ absence d’erreur systématique

- Convergence, qu’on peut traduire par

$$\lim_{n \to +\infty} \mathrm{var}\left[\hat{\theta}_n\right] = 0$$

  $\rightsquigarrow$ plus on a d’observations, plus on a de certitude

Erreur quadratique moyenne

$$\begin{aligned} e_n^2(\theta) &= \mathrm{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] \\ &= \mathrm{var}\left[\hat{\theta}_n\right] + \left(\mathrm{E}\left[\hat{\theta}_n\right] - \theta\right)^2 \end{aligned}$$

Qualités d’un estimateur

Example

Soient $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. Soient deux estimateurs de m

$$\hat{m}_1 = Y_1 \qquad \text{et} \qquad \hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Alors

$$\mathrm{E}\left[\hat{m}_1\right] = m \quad \mathrm{var}\left[\hat{m}_1\right] = \sigma^2$$
$$\mathrm{E}\left[\hat{m}_2\right] = m \quad \mathrm{var}\left[\hat{m}_2\right] = \frac{\sigma^2}{n}$$

$\rightarrow \hat{m}_2$ “meilleur” que $\hat{m}_1$.

Inégalité de Cramer-Rao

Propriété

Soit $\hat{\theta}$ un estimateur sans biais de θ. Alors

$$\text{var}\left[\hat{\theta}\right] \geq \text{BCR}_n(\theta) \triangleq \mathcal{I}(\theta)^{-1}$$

où $\mathcal{I}(\theta)$ est l'**information de Fisher**

$$\mathcal{I}(\theta) = \text{E}\left[-\frac{\partial^2 \log f\left(Y_1, \ldots, Y_n|\theta\right)}{\partial \theta^2}\right].$$

Définition

Un estimateur sans biais est dit **efficace** si $\text{var}\left[\hat{\theta}\right] = \text{BCR}_n(\theta)$.

Propriété

L'estimateur efficace est unique.

Inégalité de Cramer-Rao

Propriété

Soit $\hat{\theta}$ un estimateur sans biais de θ. Alors

$$\text{var}\left[\hat{\theta}\right] \geq \text{BCR}_n(\theta) \triangleq \mathcal{I}(\theta)^{-1}$$

où $\mathcal{I}(\theta)$ est l'**information de Fisher**

$$\mathcal{I}(\theta) = \text{E}\left[-\frac{\partial^2 \log f\left(Y_1, \ldots, Y_n|\theta\right)}{\partial \theta^2}\right].$$

Définition

Un estimateur sans biais est dit **efficace** si $\text{var}\left[\hat{\theta}\right] = \text{BCR}_n(\theta)$.

Propriété

L'estimateur efficace est unique.

Inégalité de Cramer-Rao

Propriété

Soit $\hat{\theta}$ un estimateur sans biais de θ. Alors

$$\text{var}\left[\hat{\theta}\right] \geq \text{BCR}_n(\theta) \triangleq \mathcal{I}(\theta)^{-1}$$

où $\mathcal{I}(\theta)$ est l'**information de Fisher**

$$\mathcal{I}(\theta) = \text{E}\left[-\frac{\partial^2 \log f\left(Y_1, \ldots, Y_n|\theta\right)}{\partial \theta^2}\right].$$

Définition

Un estimateur sans biais est dit **efficace** si $\text{var}\left[\hat{\theta}\right] = \text{BCR}_n(\theta)$.

Propriété

L'estimateur efficace est unique.

Estimateur du maximum de vraisemblance

Définition
Soient $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim f(y|\theta)$. La **vraisemblance** est la fonction $\mathcal{L}_n$ définie par

$$\mathcal{L}_n(Y_1, \ldots, Y_n; \theta) = f(Y_1, \ldots, Y_n|\theta).$$

Remarque
Soit $(Y_1, Y_2, \ldots, Y_n)$ une réalisation. Alors $\mathcal{L}_n(Y_1, \ldots, Y_n; \theta)$ mesure l'adéquation des observations au modèle statistique.

Définition
L'estimateur du maximum de vraisemblance $\hat{\theta}_{\mathrm{ML}}$ du paramètre θ est

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_n(Y_1, \ldots, Y_n; \theta)$$

Estimateur du maximum de vraisemblance

Définition
Soient $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim f(y|\theta)$. La **vraisemblance** est la fonction $\mathcal{L}_n$ définie par

$$\mathcal{L}_n(Y_1, \ldots, Y_n; \theta) = f(Y_1, \ldots, Y_n|\theta).$$

Remarque
Soit $(Y_1, Y_2, \ldots, Y_n)$ une réalisation. Alors $\mathcal{L}_n(Y_1, \ldots, Y_n; \theta)$ mesure l'adéquation des observations au modèle statistique.

Définition
L'estimateur du maximum de vraisemblance $\hat{\theta}_{\mathrm{ML}}$ du paramètre θ est

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_n(Y_1, \ldots, Y_n; \theta)$$

Estimateur du maximum de vraisemblance

Propriétés

- Asymptotiquement non-biaisé

$$\lim_{n \to +\infty} \mathrm{E}\left[\hat{\theta}_n\right] = \theta$$

- Convergent

$$\lim_{n \to +\infty} \mathrm{var}\left[\hat{\theta}_n\right] = 0$$

- Asymptotiquement efficace

$$\lim_{n \to +\infty} \frac{\mathrm{var}\left[\hat{\theta}_n\right]}{\mathrm{BCR}_n\left(\theta\right)} = 1$$

- Normalité asymptotique

Estimateur du maximum de vraisemblance

En pratique

En général, la recherche de l'estimateur du maximum de vraisemblance $\hat{\theta}_{\mathrm{ML}}$ se fait par minimisation de la neg-log-vraisemblance

$$\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_n(Y_1, \ldots, Y_n; \theta) = \underset{\theta}{\operatorname{argmin}}\, (-\log f(Y_1, \ldots, Y_n|\theta))$$

Exemple

Soit $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. Alors

$$\mathcal{L}_n(Y_1, \ldots, Y_n; m) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (Y_i - m)^2}{2\sigma^2}\right).$$

Maximiser la vraisemblance peut se réécrire comme

$$\underset{m}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - m)^2$$

qui fournit

$$\hat{m}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Estimateur du maximum de vraisemblance

En pratique

En général, la recherche de l'estimateur du maximum de vraisemblance $\hat{\theta}_{\mathrm{ML}}$ se fait par minimisation de la neg-log-vraisemblance

$$\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_n(Y_1, \ldots, Y_n; \theta) = \underset{\theta}{\operatorname{argmin}}\, (-\log f(Y_1, \ldots, Y_n|\theta))$$

Exemple

Soit $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. Alors

$$\mathcal{L}_n(Y_1, \ldots, Y_n; m) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (Y_i - m)^2}{2\sigma^2}\right).$$

Maximiser la vraisemblance peut se réécrire comme

$$\underset{m}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - m)^2$$

qui fournit

$$\hat{m}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Plan

Paradigme bayésien
Une philosophie...

Principe
L'estimation bayésienne consiste à considérer le paramètre inconnu θ comme la réalisation d'une variable aléatoire. Cette variable aléatoire est décrite par une loi de probabilité $f(\theta)$, appelée **loi a priori**, qui résume l'information concernant le paramètre θ disponible avant toute expérience/mesure.

Une loi a priori ?

- un moyen d'exploiter de l'information disponible pour le problème
- une connaissance incertaine modélisée de façon probabiliste
- une description personnelle ou subjective de cette connaissance
- une nécessité lorsque l'estimation non-bayésienne échoue (cf. plus loin)
  $\rightarrow$ correction de l'information apportée par les observations
- une interprétation probabiliste de la régularisation/pénalisation pour éviter le sur-apprentissage

... mais (d'accord), aussi un **prétexte** pour entrer dans le "cadre bayésien".

Paradigme bayésien
Une philosophie...

Principe
L'estimation bayésienne consiste à considérer le paramètre inconnu θ comme la réalisation d'une variable aléatoire. Cette variable aléatoire est décrite par une loi de probabilité $f(\theta)$, appelée **loi a priori**, qui résume l'information concernant le paramètre θ disponible avant toute expérience/mesure.

Une loi a priori ?

- un moyen d'exploiter de l'information disponible pour le problème
- une connaissance incertaine modélisée de façon probabiliste
- une description personnelle ou subjective de cette connaissance
- une nécessité lorsque l'estimation non-bayésienne échoue (cf. plus loin)
  $\rightarrow$ correction de l'information apportée par les observations
- une interprétation probabiliste de la régularisation/pénalisation pour éviter le sur-apprentissage

... mais (d'accord), aussi un **prétexte** pour entrer dans le "cadre bayésien".

Paradigme bayésien
Une philosophie...

Principe
L'estimation bayésienne consiste à considérer le paramètre inconnu θ comme la réalisation d'une variable aléatoire. Cette variable aléatoire est décrite par une loi de probabilité $f(\theta)$, appelée **loi a priori**, qui résume l'information concernant le paramètre θ disponible avant toute expérience/mesure.

Une loi a priori ?

- un moyen d'exploiter de l'information disponible pour le problème
- une connaissance incertaine modélisée de façon probabiliste
- une description personnelle ou subjective de cette connaissance
- une nécessité lorsque l'estimation non-bayésienne échoue (cf. plus loin)
  $\rightarrow$ correction de l'information apportée par les observations
- une interprétation probabiliste de la régularisation/pénalisation pour éviter le sur-apprentissage

... mais (d'accord), aussi un **prétexte** pour entrer dans le "cadre bayésien".

Paradigme bayésien
... mais pas une croyance !

"*Once again, for Bayesians as much as for any other statistician, parameters are (typically) fixed but unknown. It is the knowledge about these unknowns that Bayesians model as random*."

A. Gelman & C. P. Robert,
"Not Only Defended But Also Applied":
The Perceived Absurdity of Bayesian Inference",
The American Statistician, Vol. 67, No. 1, Feb. 2013.

Construction des estimateurs bayésiens

Principe
Un estimateur bayésien minimise un risque bayésien

$$\hat{\theta}_{\text{Bayes}} = \min_{\hat{\theta}} \mathrm{E}\left[c\left(\theta, \hat{\theta}\right)\right]$$

où $c(\theta|\cdot)$ est une fonction de coût qui quantifie l'erreur commise lors de l'estimation de θ par l'estimateur $\hat{\theta}$.

Remarque
Ici, $\mathrm{E}[\cdot]$ est l'espérance sur les variables $Y_1, \ldots, Y_n$ **et** sur le paramètre θ (considéré comme aléatoire)

$$\mathrm{E}\left[c\left(\theta, \hat{\theta}\right)\right] = \mathrm{E}_\theta\left[\mathrm{E}_{Y_1,\ldots,Y_n}\left[c\left(\theta, \hat{\theta}\right)|\theta\right]\right]$$

Estimateurs bayésiens classiques

Estimateur minimisant l'erreur quadratique moyenne

On choisit un coût quadratique

$$c(\theta, \hat{\theta}) = \left(\theta - \hat{\theta}\right)^2.$$

Le risque est l'erreur quadratique moyenne (MSE, mean square error)

$$\mathrm{MSE} = \mathrm{E}\left[\left(\theta - \hat{\theta}\right)\right]^2.$$

On montre que l'estimateur qui minimise l'erreur quadratique moyenne (MMSE) est

$$\hat{\theta}_{\mathrm{MMSE}} = \mathrm{E}\left[\theta | Y_1, \ldots, Y_n\right]$$

c'est-à-dire la moyenne a posteriori

$$\mathrm{E}\left[\theta | Y_1, \ldots, Y_n\right] = \int \theta f\left(\theta | Y_1, \ldots, Y_n\right) d\theta.$$

Estimateurs bayésiens classiques

Estimateur de la moyenne a posteriori

On choisit un coût 0/1, défini pour δ arbitrairement petit par

$$c(\theta, \hat{\theta}) = \begin{cases} 1, & \text{si } |\theta - \hat{\theta}| \geq \delta; \\ 0, & \text{sinon.} \end{cases}$$

On montre que l'estimateur qui minimise le risque associé est

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\, f(\theta | Y_1, \ldots, Y_n)$$

c'est-à-dire le mode de la loi a posteriori.

Loi a posteriori

Estimateurs bayésiens classiques

$$\hat{\theta}_{\text{MMSE}} = \int \theta f(\theta|Y_1,\ldots,Y_n)\, d\theta$$

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\, f(\theta|Y_1,\ldots,Y_n)$$

La loi a posteriori s'écrit grâce à la loi de Bayes

$$f(\theta|Y_1,\ldots,Y_n) = \frac{f(Y_1,\ldots,Y_n|\theta)f(\theta)}{f(Y_1,\ldots,Y_n)}$$

et fait la synthèse des informations fournie par

- les observations, via la vraisemblance $f(Y_1,\ldots,Y_n|\theta)$
- le modèle a priori, via la loi a priori $f(\theta)$

La quantité $f(Y_1,\ldots,Y_n) = \int f(Y_1,\ldots,Y_n|\theta)f(\theta)d\theta$ est la vraisemblance marginale, généralement difficile à calculer.

Loi a posteriori

Estimateurs bayésiens classiques

$$\hat{\theta}_{\text{MMSE}} = \int \theta f(\theta|Y_1,\ldots,Y_n)\, d\theta$$

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\, f(\theta|Y_1,\ldots,Y_n)$$

La loi a posteriori s'écrit grâce à la loi de Bayes

$$f(\theta|Y_1,\ldots,Y_n) = \frac{f(Y_1,\ldots,Y_n|\theta)f(\theta)}{f(Y_1,\ldots,Y_n)}$$

et fait la synthèse des informations fournie par

- les observations, via la vraisemblance $f(Y_1,\ldots,Y_n|\theta)$
- le modèle a priori, via la loi a priori $f(\theta)$

La quantité $f(Y_1,\ldots,Y_n) = \int f(Y_1,\ldots,Y_n|\theta)f(\theta)d\theta$ est la vraisemblance marginale, généralement difficile à calculer.

Loi a posteriori

Estimateurs bayésiens classiques

$$\hat{\theta}_{\text{MMSE}} = \int \theta f(\theta|Y_1,\ldots,Y_n)\, d\theta$$

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\, f(\theta|Y_1,\ldots,Y_n)$$

La loi a posteriori s'écrit grâce à la loi de Bayes

$$f(\theta|Y_1,\ldots,Y_n) = \frac{f(Y_1,\ldots,Y_n|\theta)f(\theta)}{f(Y_1,\ldots,Y_n)}$$

et fait la synthèse des informations fournie par

- les observations, via la vraisemblance $f(Y_1,\ldots,Y_n|\theta)$
- le modèle a priori, via la loi a priori $f(\theta)$

La quantité $f(Y_1,\ldots,Y_n) = \int f(Y_1,\ldots,Y_n|\theta)f(\theta)d\theta$ est la vraisemblance marginale, généralement difficile à calculer.

Quelle loi a priori ?

- un choix guidé par la problématique visée
  $\rightsquigarrow$ pour exploiter une connaissance a priori concernant le paramètre à estimer
  - parcimonie, régularité,...
  - contraintes de support (e.g., positivité)

  $\rightsquigarrow$ pour exploiter des données d’apprentissage
  - approche bayésienne empirique
    (“empirical Bayesian approach” vs. “fully Bayesian approach”)
- loi non-informative (ou vague ou objective)
  $\rightsquigarrow$ le paradoxe... pour entrer dans le “cadre bayésien” !
  - loi standard avec hyperparamètres choisis pour assurer une grande variance
  - loi de Jeffreys (invariante par reparamétrisation), définie par

    $$p(\theta) \propto \sqrt{I_n(\theta)}, \quad \text{où } I_n(\theta) \text{ est l’information de Fisher}$$

  - reference prior de Bernardo
  - principe du maximum d’entropie
- une loi choisie pour faciliter les calculs (et alléger le coût computationel)
  - loi conjuguée, c’est-à-dire choisie pour que la loi a posteriori soit simple
    (loi paramétrée, pour avoir plus de “flexibilité”)

Quelle loi a priori ?

- un choix guidé par la problématique visée
  $\rightsquigarrow$ pour exploiter une connaissance a priori concernant le paramètre à estimer
  - parcimonie, régularité,...
  - contraintes de support (e.g., positivité)

  $\rightsquigarrow$ pour exploiter des données d’apprentissage
  - approche bayésienne empirique
    (“empirical Bayesian approach” vs. “fully Bayesian approach”)
- loi non-informative (ou vague ou objective)
  $\rightsquigarrow$ le paradoxe... pour entrer dans le “cadre bayésien” !
  - loi standard avec hyperparamètres choisis pour assurer une grande variance
  - loi de Jeffreys (invariante par reparamétrisation), définie par

    $$p(\theta) \propto \sqrt{I_n(\theta)}, \quad \text{où } I_n(\theta) \text{ est l’information de Fisher}$$

  - reference prior de Bernardo
  - principe du maximum d’entropie
- une loi choisie pour faciliter les calculs (et alléger le coût computationel)
  - loi conjuguée, c’est-à-dire choisie pour que la loi a posteriori soit simple
    (loi paramétrée, pour avoir plus de “flexibilité”)

Quelle loi a priori ?

- un choix guidé par la problématique visée
  $\rightsquigarrow$ pour exploiter une connaissance a priori concernant le paramètre à estimer
  - parcimonie, régularité,...
  - contraintes de support (e.g., positivité)

  $\rightsquigarrow$ pour exploiter des données d’apprentissage
  - approche bayésienne empirique
    (“empirical Bayesian approach” vs. “fully Bayesian approach”)
- loi non-informative (ou vague ou objective)
  $\rightsquigarrow$ le paradoxe... pour entrer dans le “cadre bayésien” !
  - loi standard avec hyperparamètres choisis pour assurer une grande variance
  - loi de Jeffreys (invariante par reparamétrisation), définie par

    $$p(\theta) \propto \sqrt{I_n(\theta)}, \quad \text{où } I_n(\theta) \text{ est l’information de Fisher}$$

  - reference prior de Bernardo
  - principe du maximum d’entropie
- une loi choisie pour faciliter les calculs (et alléger le coût computationel)
  - loi conjuguée, c’est-à-dire choisie pour que la loi a posteriori soit simple
    (loi paramétrée, pour avoir plus de “flexibilité”)

Exemple

Soit $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. La vraisemblance est

$$f(Y_1, \ldots, Y_n | m) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n} (Y_i - m)^2}{2\sigma^2} \right).$$

On cherche à estimer le paramètre m.

Estimateur du maximum de vraisemblance

$$\hat{m}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{1}$$

Loi a priori conjuguée
On munit m de la loi a priori $\mathcal{N}(m_0, \sigma_0^2)$, i.e.,

$$f(m) = \left(\frac{1}{2\pi\sigma_0^2} \right)^{\frac{1}{2}} \exp\left(-\frac{(m - m_0)^2}{2\sigma_0^2} \right).$$

Exemple

Soit $Y_1, \ldots, Y_n$ un échantillon tel que $Y_i \sim \mathcal{N}(m, \sigma^2)$. La vraisemblance est

$$f(Y_1, \ldots, Y_n | m) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left(- \frac{\sum_{i=1}^{n} (Y_i - m)^2}{2\sigma^2} \right).$$

On cherche à estimer le paramètre m.

Estimateur du maximum de vraisemblance

$$\hat{m}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{1}$$

Loi a priori conjuguée

On munit m de la loi a priori $\mathcal{N}(m_0, \sigma_0^2)$, i.e.,

$$f(m) = \left(\frac{1}{2\pi\sigma_0^2} \right)^{\frac{1}{2}} \exp\left(- \frac{(m - m_0)^2}{2\sigma_0^2} \right).$$

Exemple

On montre que la loi a posteriori est

$$f(m|Y_1,\ldots,Y_n) = \left(\frac{1}{2\pi\eta^2}\right)^2 \exp\left(-\frac{(m-\mu)^2}{2\eta^2}\right)$$

avec

$$\eta^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \quad \text{et} \quad \mu = \frac{\left(\sum_{i=1}^n Y_i\right)\sigma_0^2 + m_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Donc

$$\hat{m}_{\text{MMSE}} = \hat{m}_{\text{MAP}} = \mu = \frac{\left(\sum_{i=1}^n Y_i\right)\sigma_0^2 + m_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Comportements limites

- Si $n \to 0$, pas de données observées, on fait confiance à la loi a priori

$$\begin{aligned} \hat{m}_{\text{MMSE}} &\to \mathrm{E}[m] = m_0 \\ \hat{m}_{\text{MAP}} &\to \operatorname*{argmax}_m f(m) = m_0 \end{aligned}$$

- Si $n \to \infty$, on fait confiance aux données

$$\hat{m}_{\text{MAP}} \to \operatorname*{argmax}_m f(Y_1,\ldots,Y_n|m) = \hat{m}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^n Y_i$$

Exemple

On montre que la loi a posteriori est

$$f(m|Y_1,\ldots,Y_n) = \left(\frac{1}{2\pi\eta^2}\right)^2 \exp\left(-\frac{(m-\mu)^2}{2\eta^2}\right)$$

avec

$$\eta^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \quad \text{et} \quad \mu = \frac{\left(\sum_{i=1}^n Y_i\right)\sigma_0^2 + m_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Donc

$$\hat{m}_{\text{MMSE}} = \hat{m}_{\text{MAP}} = \mu = \frac{\left(\sum_{i=1}^n Y_i\right)\sigma_0^2 + m_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Comportements limites

- Si $n \to 0$, pas de données observées, on fait confiance à la loi a priori

$$\begin{aligned} \hat{m}_{\text{MMSE}} &\to \mathrm{E}[m] = m_0 \\ \hat{m}_{\text{MAP}} &\to \operatorname*{argmax}_m f(m) = m_0 \end{aligned}$$

- Si $n \to \infty$, on fait confiance aux données

$$\hat{m}_{\text{MAP}} \to \operatorname*{argmax}_m f(Y_1,\ldots,Y_n|m) = \hat{m}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^n Y_i$$

Exemple

On montre que la loi a posteriori est

$$f(m|Y_1,\ldots,Y_n) = \left(\frac{1}{2\pi\eta^2}\right)^2 \exp\left(-\frac{(m-\mu)^2}{2\eta^2}\right)$$

avec

$$\eta^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \quad \text{et} \quad \mu = \frac{\left(\sum_{i=1}^n Y_i\right)\sigma_0^2 + m_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Donc

$$\hat{m}_{\text{MMSE}} = \hat{m}_{\text{MAP}} = \mu = \frac{\left(\sum_{i=1}^n Y_i\right)\sigma_0^2 + m_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Comportements limites

- Si $n \to 0$, pas de données observées, on fait confiance à la loi a priori

$$\begin{aligned} \hat{m}_{\text{MMSE}} &\to \mathrm{E}[m] = m_0 \\ \hat{m}_{\text{MAP}} &\to \operatorname*{argmax}_m f(m) = m_0 \end{aligned}$$

- Si $n \to \infty$, on fait confiance aux données

$$\hat{m}_{\text{MAP}} \to \operatorname*{argmax}_m f(Y_1,\ldots,Y_n|m) = \hat{m}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^n Y_i$$

Modèles bayésiens hiérarchiques
Ou comment choisir les hyperparamètres ?

Problème
La loi a priori $f(\theta|\phi)$ peut dépendre d'un ou plusieurs hyperparamètres inconnus ϕ.

Deux stratégies

- on les fixe arbitrairement pour obtenir une loi informative ou, au contraire, vague (i.e., de grande variance)
- on essaye de les estimer

Trois approches

- approche bayésienne empirique
  $\rightarrow$ on estime ϕ sur des données d'apprentissage
- approche fréquentiste
  $\rightarrow$ on estime ϕ au sens du maximum de vraisemblance (e.g., algo. EM)
- approche "fully bayesian"
  $\rightarrow$ on introduit un deuxième niveau dans la hiérarchie bayésienne

Modèles bayésiens hiérarchiques
Ou comment choisir les hyperparamètres ?

Problème

La loi a priori $f(\theta|\phi)$ peut dépendre d'un ou plusieurs hyperparamètres inconnus ϕ.

Deux stratégies

- on les fixe arbitrairement pour obtenir une loi informative ou, au contraire, vague (i.e., de grande variance)
- on essaye de les estimer

Trois approches

- approche bayésienne empirique
  $\rightarrow$ on estime ϕ sur des données d'apprentissage
- approche fréquentiste
  $\rightarrow$ on estime ϕ au sens du maximum de vraisemblance (e.g., algo. EM)
- approche "fully bayesian"
  $\rightarrow$ on introduit un deuxième niveau dans la hiérarchie bayésienne

Modèles bayésiens hiérarchiques
Ou comment choisir les hyperparamètres ?

Problème
La loi a priori $f(\theta|\phi)$ peut dépendre d'un ou plusieurs hyperparamètres inconnus ϕ.

Deux stratégies

- on les fixe arbitrairement pour obtenir une loi informative ou, au contraire, vague (i.e., de grande variance)
- **on essaye de les estimer**

Trois approches

- approche bayésienne empirique
  $\rightarrow$ on estime ϕ sur des données d'apprentissage
- approche fréquentiste
  $\rightarrow$ on estime ϕ au sens du maximum de vraisemblance (e.g., algo. EM)
- approche "fully bayesian"
  $\rightarrow$ on introduit un deuxième niveau dans la hiérarchie bayésienne

Modèles bayésiens hiérarchiques

Principe

L'hyperparamètre est lui aussi considéré comme (la réalisation d') une variable aléatoire munie d'une loi a priori $f(\phi)$.

Inférence bayésienne hiérarchique

Le théorème de Bayes permet d'écrire la loi jointe des paramètres et hyperparamètres inconnus

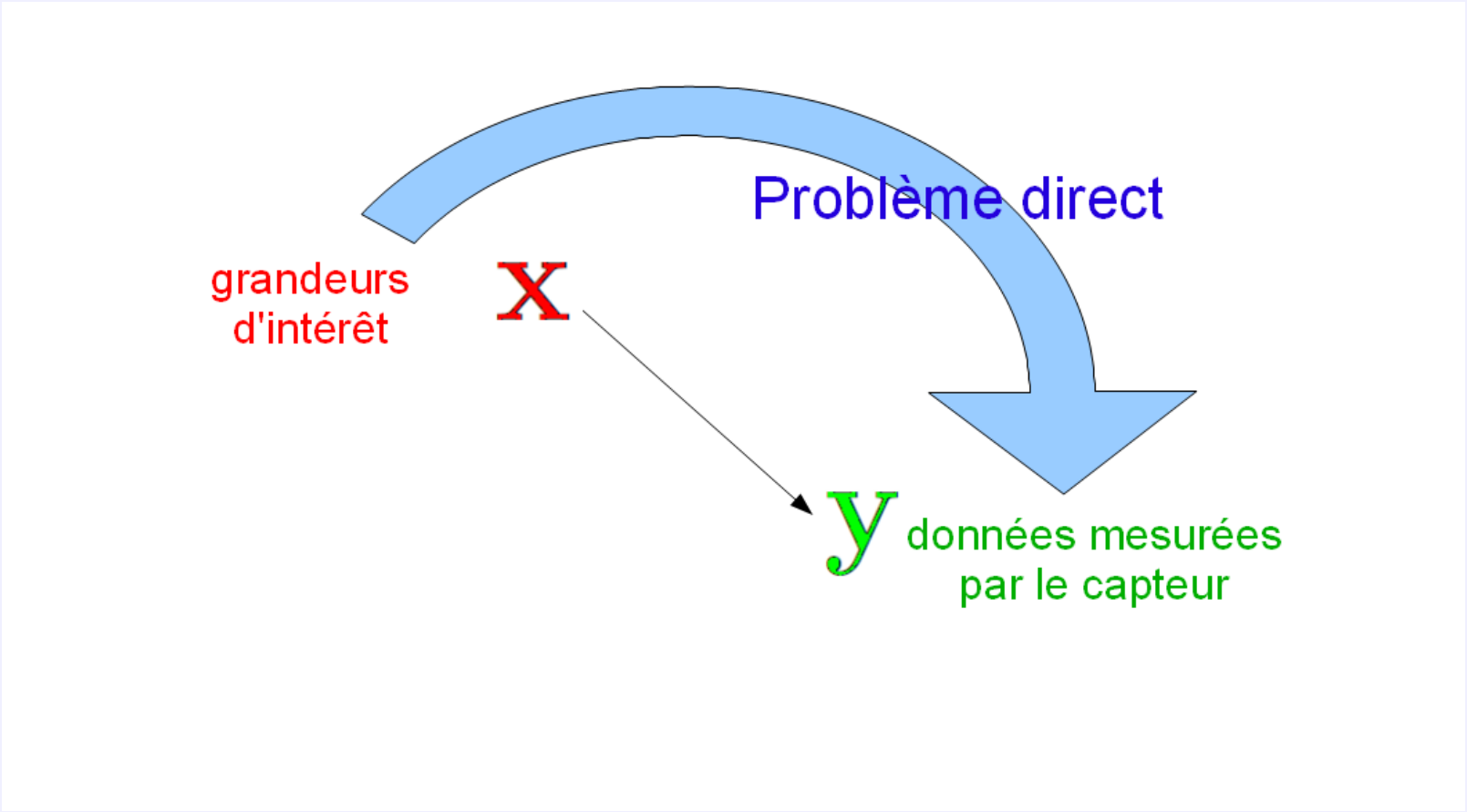
$$f(\theta, \phi | Y_1, \ldots, Y_n) \propto f(Y_1, \ldots, Y_n | \theta) f(\theta | \phi) f(\phi)$$

où la constante de normalisation est $f(Y_1, \ldots, Y_n)^{-1}$.

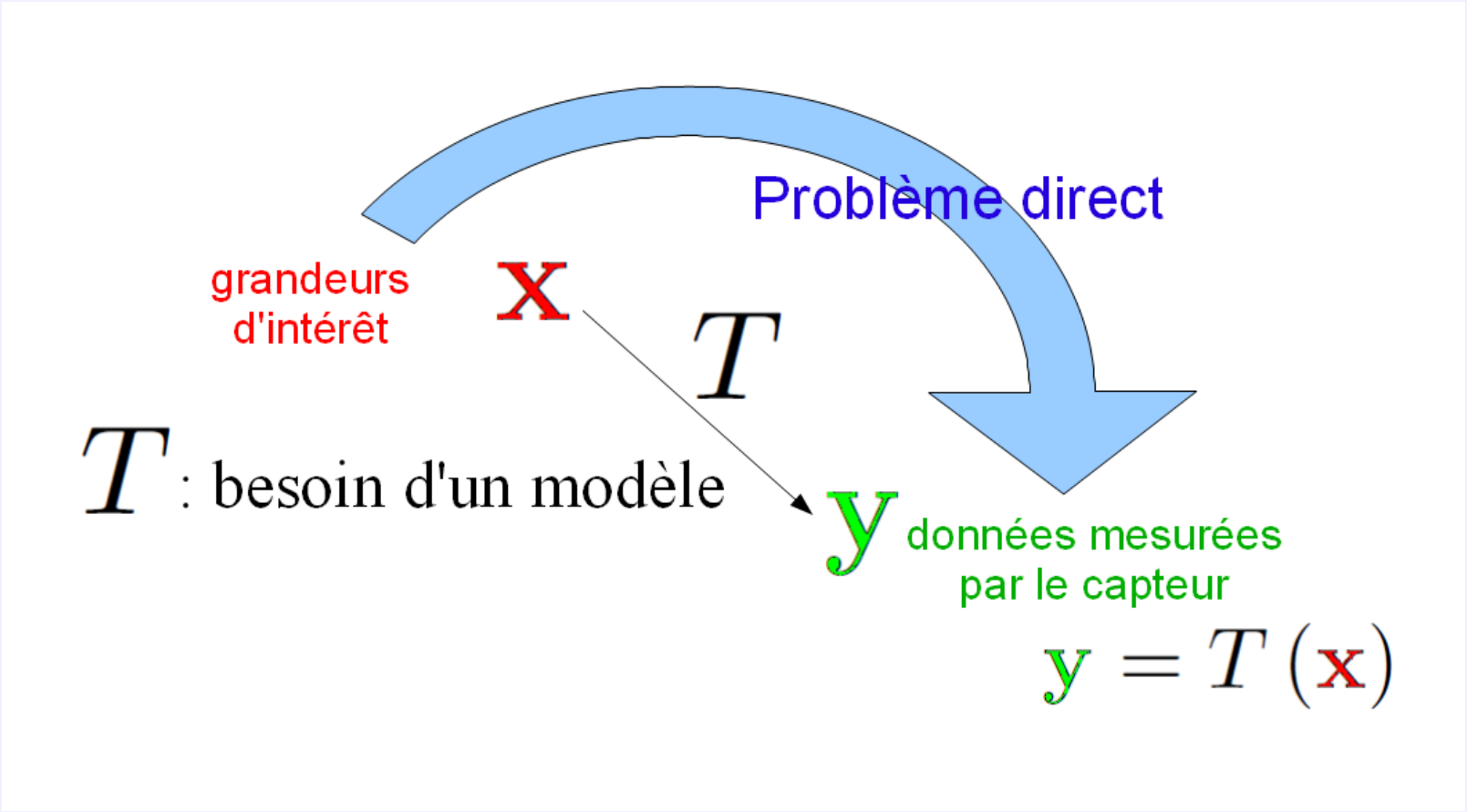
Estimation conjointe de θ et ϕ

Plan

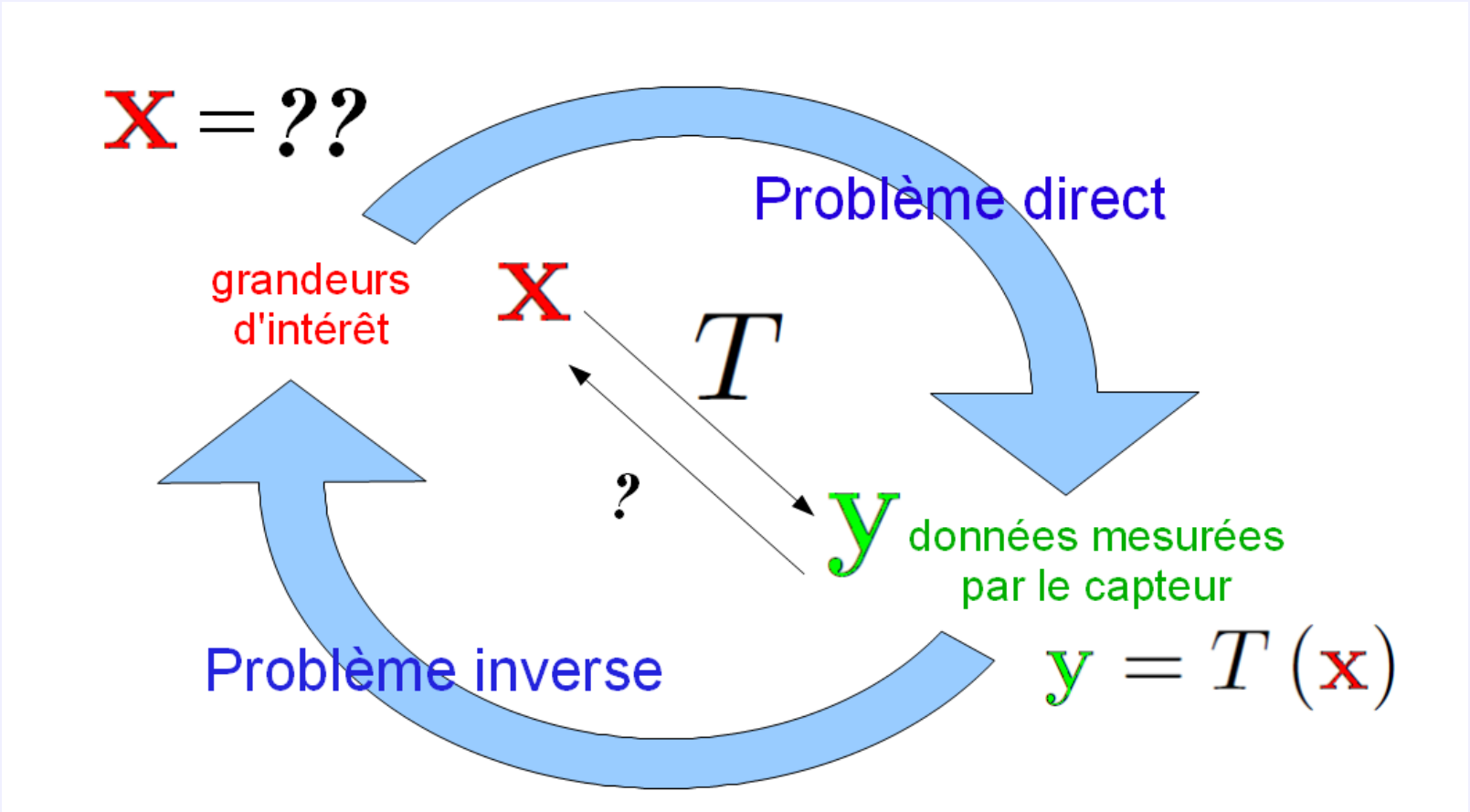
Problèmes inverses, myopes et aveugles



Problèmes inverses, myopes et aveugles



Problèmes inverses, myopes et aveugles



Problèmes inverses, myopes et aveugles

$$\begin{array}{rccc} T: & \mathbb{R}^M & \rightarrow & \mathbb{R}^N \\ & \mathbf{x} & \mapsto & \mathbf{y} \approx T(\mathbf{x}) \end{array}$$

Hypothèses relatives au modèle T

- parfaitement connu : problème inverse
  T non-inversible: problème mal-posé
  $\Rightarrow$ besoin d'une régularisation
- partiellement connu : problème myope (semi-aveugle)
  T connu via une forme paramétrique $T_{\boldsymbol{\theta}}(\mathbf{x})$, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]^T$
  $\Rightarrow$ estimation conjointe de $\mathbf{x}$ et $\boldsymbol{\theta}$
- totalement inconnu : problème aveugle
  $\Rightarrow$ besoin d'hypothèses supplémentaires
  exemple : T linéaire
  - déconvolution $y(n) \approx h(n) * x(n)$
  - séparation aveugle de source $\mathbf{y}(n) \approx \mathbf{A}\mathbf{x}(n)$
  - factorisation matricielle $\mathbf{Y} \approx \mathbf{M}\mathbf{X}$

Problèmes inverses, myopes et aveugles

$$\begin{array}{rccc} T: & \mathbb{R}^M & \rightarrow & \mathbb{R}^N \\ & \mathbf{x} & \mapsto & \mathbf{y} \approx T(\mathbf{x}) \end{array}$$

Hypothèses relatives au modèle T

- parfaitement connu : problème inverse
  T non-inversible: problème mal-posé
  $\Rightarrow$ besoin d'une régularisation
- partiellement connu : problème myope (semi-aveugle)
  T connu via une forme paramétrique $T_{\boldsymbol{\theta}}(\mathbf{x})$, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]^T$
  $\Rightarrow$ estimation conjointe de $\mathbf{x}$ et $\boldsymbol{\theta}$
- totalement inconnu : problème aveugle
  $\Rightarrow$ besoin d'hypothèses supplémentaires
  exemple : T linéaire
  - déconvolution $y(n) \approx h(n) * x(n)$
  - séparation aveugle de source $\mathbf{y}(n) \approx \mathbf{A}\mathbf{x}(n)$
  - factorisation matricielle $\mathbf{Y} \approx \mathbf{M}\mathbf{X}$

Problèmes inverses, myopes et aveugles

$$\begin{array}{rccc} T: & \mathbb{R}^M & \rightarrow & \mathbb{R}^N \\ & \mathbf{x} & \mapsto & \mathbf{y} \approx T(\mathbf{x}) \end{array}$$

Hypothèses relatives au modèle T

- parfaitement connu : problème inverse
  T non-inversible: problème mal-posé
  $\Rightarrow$ besoin d'une régularisation
- partiellement connu : problème myope (semi-aveugle)
  T connu via une forme paramétrique $T_{\boldsymbol{\theta}}(\mathbf{x})$, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]^T$
  $\Rightarrow$ estimation conjointe de $\mathbf{x}$ et $\boldsymbol{\theta}$
- totalement inconnu : problème aveugle
  $\Rightarrow$ besoin d'hypothèses supplémentaires
  exemple : T linéaire
  - déconvolution $y(n) \approx h(n) * x(n)$
  - séparation aveugle de source $\mathbf{y}(n) \approx \mathbf{A}\mathbf{x}(n)$
  - factorisation matricielle $\mathbf{Y} \approx \mathbf{M}\mathbf{X}$

Inversion et régularisation
Formulation probabiliste bayésienne

Comment choisir une solution dans l'espace des solutions admissibles ?

Inversion et régularisation
Formulation probabiliste bayésienne

Comment choisir une solution dans l'espace des solutions admissibles ?

$\Downarrow$

introduction de pénalités et/ou de contraintes
(guidé par l'application visée)

Construction du critère

- $\mathbf{y}|\mathbf{x} \sim f(\mathbf{y}|\mathbf{x})$: terme d'attache aux données
- $\mathbf{x} \sim f(\mathbf{x})$: pénalités et contraintes

Calcul d'un estimateur bayésien à partir de la loi *a posteriori*

$$f(\mathbf{x}|\mathbf{y}) = \frac{1}{f(\mathbf{y})} f(\mathbf{y}|\mathbf{x})\, f(\mathbf{x})$$

- $\hat{\mathbf{x}}_{\mathrm{MMSE}} = \mathrm{E}[\mathbf{x}|\mathbf{y}] = \int \mathbf{x} f(\mathbf{x}|\mathbf{y})\, d\mathbf{x}$
- $\hat{\mathbf{x}}_{\mathrm{MAP}} = \mathrm{argmax}_{\mathbf{x}}\, f(\mathbf{x}|\mathbf{y})$

Modèle direct linéaire

On considère le modèle d'observation défini par

$$\mathrm{E}\left[\mathbf{y}|\mathbf{x}\right] = \mathbf{H}\mathbf{x}$$

où

- $\mathbf{x} \in \mathbb{R}^M$ est le vecteur inconnu,
- $\mathbf{y} \in \mathbb{R}^N$ est le vecteur des observations/mesures,
- $\mathbf{H} \in \mathbb{R}^{N \times M}$ est un opérateur d'acquisition (sous-échantillonnage, convolution, etc...),

Description probabilisite du bruit

L'application visée fournit un modèle aléatoire naturel pour préciser la nature de $\mathrm{E}\left[\cdot\right]$:

- bruit additif gaussien (la plupart des cas d'étude) :

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b} \quad \text{avec} \quad \mathbf{b} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}\right)$$

- bruit poissonien (acquisition faible flux)

$$y_i|\mathbf{x} \sim \mathcal{P}([\mathbf{H}\mathbf{x}]_i), \; i = 1, \ldots, N$$

- bruit multiplicatif (speckle)

$$y_i = [\mathbf{H}\mathbf{x}]_i \, b_i, \; i = 1, \ldots, N \quad \text{avec} \quad b_i \sim \mathcal{G}\left(\alpha, 1/\alpha\right)$$

Modèle direct linéaire

On considère le modèle d'observation défini par

$$\mathrm{E}\left[\mathbf{y}|\mathbf{x}\right] = \mathbf{H}\mathbf{x}$$

où

- $\mathbf{x} \in \mathbb{R}^M$ est le vecteur inconnu,
- $\mathbf{y} \in \mathbb{R}^N$ est le vecteur des observations/mesures,
- $\mathbf{H} \in \mathbb{R}^{N \times M}$ est un opérateur d'acquisition (sous-échantillonnage, convolution, etc...),

Description probabilisite du bruit

L'application visée fournit un modèle aléatoire naturel pour préciser la nature de $\mathrm{E}\left[\cdot\right]$:

- bruit additif gaussien (la plupart des cas d'étude) :

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b} \quad \text{avec} \quad \mathbf{b} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}\right)$$

- bruit poissonien (acquisition faible flux)

$$y_i|\mathbf{x} \sim \mathcal{P}([\mathbf{H}\mathbf{x}]_i), \ i = 1, \ldots, N$$

- bruit multiplicatif (speckle)

$$y_i = [\mathbf{H}\mathbf{x}]_i \, b_i, \ i = 1, \ldots, N \quad \text{avec} \quad b_i \sim \mathcal{G}\left(\alpha, 1/\alpha\right)$$

Problème inverse linéaire gaussien

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b} \quad \text{avec} \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_b})$$

Le terme **b** représente un bruit de mesure ou une erreur de modèle

- de moyenne nulle, $\mathrm{E}[\mathbf{b}] = \mathbf{0}$
- de matrice de covariance $\mathbf{\Sigma_b}$

Problème
Retrouver/reconstruire **x** étant donné le vecteur observé **y**.

Fonction de vraisemblance
Compte tenu de la statistique du bruit,

$$f(\mathbf{y}|\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{n/2} |\mathbf{\Sigma_b}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T \mathbf{\Sigma_b}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right]$$

Problème inverse linéaire gaussien

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b} \quad \text{avec} \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_b})$$

Le terme **b** représente un bruit de mesure ou une erreur de modèle

- de moyenne nulle, $\mathrm{E}[\mathbf{b}] = \mathbf{0}$
- de matrice de covariance $\mathbf{\Sigma_b}$

Problème
Retrouver/reconstruire **x** étant donné le vecteur observé **y**.

Fonction de vraisemblance
Compte tenu de la statistique du bruit,

$$f(\mathbf{y}|\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{n/2} |\mathbf{\Sigma_b}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T \mathbf{\Sigma_b}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right]$$

Problème inverse linéaire gaussien

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b} \quad \text{avec} \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}})$$

Le terme **b** représente un bruit de mesure ou une erreur de modèle

- de moyenne nulle, $\mathrm{E}[\mathbf{b}] = \mathbf{0}$
- de matrice de covariance $\boldsymbol{\Sigma}_{\mathbf{b}}$

Problème
Retrouver/reconstruire **x** étant donné le vecteur observé **y**.

Fonction de vraisemblance
Compte tenu de la statistique du bruit,

$$f(\mathbf{y}|\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{n/2} \left|\boldsymbol{\Sigma}_{\mathbf{b}}\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{H}\mathbf{x})^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x})\right]$$

Problème inverse linéaire gaussien
Estimation du maximum de vraisemblance

Maximiser la (log-)vraisemblance conduit à

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2_{\boldsymbol{\Sigma}_\mathbf{b}^{-1}} = \left(\mathbf{H}^T \boldsymbol{\Sigma}_\mathbf{b}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \boldsymbol{\Sigma}_\mathbf{b}^{-1} \mathbf{y}$$

Cas particuliers

- $\boldsymbol{\Sigma}_\mathbf{b} = \sigma^2_\mathbf{b} \boldsymbol{I}$, alors $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T = \mathbf{H}^\dagger$ ($= \mathbf{H}^{-1}$ si $\mathbf{H}$ inversible)

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 = \mathbf{H}^\dagger \mathbf{y} = \hat{\mathbf{x}}_{\mathrm{LS}}$$

  $\rightarrow$ solution des moindres carrés

- $\boldsymbol{\Sigma}_\mathbf{b} = \operatorname{diag}\left(\sigma^2_{\mathbf{b},1}, \ldots, \sigma^2_{\mathbf{b},N}\right)$ alors

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(y_i - [\mathbf{H}\mathbf{x}]_i\right)^2 = \hat{\mathbf{x}}_{\mathrm{WLS}}$$

  $\rightarrow$ solution des moindres carrés pondérés

Problème inverse linéaire gaussien
Estimation bayésienne

Problème de la solution $\hat{\mathbf{x}}_{\mathrm{ML}}$

- problème mal-posé (inversibilité de $\left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{H}\right)$ non garantie)
- $\mathbf{H}$ mal conditionnée (très forte sensibilité au bruit)

Modèle a priori

Le vecteur $\mathbf{x}$ est muni d'une loi a priori $f(\mathbf{x}|\phi)$, ici choisie normale conjuguée $\mathbf{x}|\phi \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})$:

$$f(\mathbf{x}|\phi) = \left(\frac{1}{2\pi}\right)^{m/2} |\boldsymbol{\Sigma}_{\mathbf{x}}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\right]$$

avec $\phi = \{\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\}$ l'ensemble des hyperparamètres.

Problème inverse linéaire gaussien
Estimation bayésienne

Problème de la solution $\hat{\mathbf{x}}_{\mathrm{ML}}$

- problème mal-posé (inversibilité de $\left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{H}\right)$ non garantie)
- $\mathbf{H}$ mal conditionnée (très forte sensibilité au bruit)

Modèle a priori

Le vecteur $\mathbf{x}$ est muni d'une loi a priori $f(\mathbf{x}|\boldsymbol{\phi})$, ici choisie normale conjuguée $\mathbf{x}|\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})$:

$$f(\mathbf{x}|\boldsymbol{\phi}) = \left(\frac{1}{2\pi}\right)^{m/2} \left|\boldsymbol{\Sigma}_{\mathbf{x}}\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\right]$$

avec $\boldsymbol{\phi} = \{\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\}$ l'ensemble des hyperparamètres.

Problème inverse linéaire gaussien
Estimation bayésienne

Loi a posteriori

D'après la loi de Bayes $f(\mathbf{x}|\mathbf{y},\phi) \propto f(\mathbf{y}|\mathbf{x})\, f(\mathbf{x}|\phi)$, on montre que

$$\mathbf{x}|\mathbf{y},\phi \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right)$$

$$\text{avec} \quad \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}^{-1} = \left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\right) \quad \text{et} \quad \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{y} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}\right)$$

Estimateurs ML, MMSE et MAP

On considère le cas

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{b}} &= \sigma_{\mathbf{b}}^2 \boldsymbol{I} \quad \text{(bruit i.i.d.)} \\ \boldsymbol{\Sigma}_{\mathbf{x}} &= \sigma_{\mathbf{x}}^2 \boldsymbol{I} \end{aligned}$$

alors

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{ML}} &= \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{y} \\ \hat{\mathbf{x}}_{\mathrm{MAP}} = \hat{\mathbf{x}}_{\mathrm{MMSE}} &= \left(\mathbf{H}^T \mathbf{H} + \lambda \boldsymbol{I}\right)^{-1} \left(\mathbf{H}^T \mathbf{y} + \lambda \boldsymbol{\mu}_{\mathbf{x}}\right) \end{aligned}$$

avec $\lambda = \frac{\sigma_{\mathbf{b}}^2}{\sigma_{\mathbf{x}}^2}$ (> 0).

Remarque

Inversibilité de $\left(\mathbf{H}^T \mathbf{H} + \lambda \boldsymbol{I}\right)$ garantie pour λ suffisamment grand...

Problème inverse linéaire gaussien
Estimation bayésienne

Loi a posteriori

D'après la loi de Bayes $f(\mathbf{x}|\mathbf{y},\phi) \propto f(\mathbf{y}|\mathbf{x})\, f(\mathbf{x}|\phi)$, on montre que

$$\mathbf{x}|\mathbf{y},\phi \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right)$$

$$\text{avec} \quad \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}^{-1} = \left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\right) \quad \text{et} \quad \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{y} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}\right)$$

Estimateurs ML, MMSE et MAP

On considère le cas

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{b}} &= \sigma_{\mathbf{b}}^2 \boldsymbol{I} \quad \text{(bruit i.i.d.)} \\ \boldsymbol{\Sigma}_{\mathbf{x}} &= \sigma_{\mathbf{x}}^2 \boldsymbol{I} \end{aligned}$$

alors

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{ML}} &= \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{y} \\ \hat{\mathbf{x}}_{\mathrm{MAP}} = \hat{\mathbf{x}}_{\mathrm{MMSE}} &= \left(\mathbf{H}^T \mathbf{H} + \lambda \boldsymbol{I}\right)^{-1} \left(\mathbf{H}^T \mathbf{y} + \lambda \boldsymbol{\mu}_{\mathbf{x}}\right) \end{aligned}$$

avec $\lambda = \frac{\sigma_{\mathbf{b}}^2}{\sigma_{\mathbf{x}}^2}$ (> 0).

Remarque

Inversibilité de $\left(\mathbf{H}^T \mathbf{H} + \lambda \boldsymbol{I}\right)$ garantie pour λ suffisamment grand...

Problème inverse linéaire gaussien
Estimation bayésienne

Loi a posteriori

D'après la loi de Bayes $f(\mathbf{x}|\mathbf{y},\phi) \propto f(\mathbf{y}|\mathbf{x})\, f(\mathbf{x}|\phi)$, on montre que

$$\mathbf{x}|\mathbf{y},\phi \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right)$$

$$\text{avec} \quad \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}^{-1} = \left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\right) \quad \text{et} \quad \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \left(\mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{b}}^{-1} \mathbf{y} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}\right)$$

Estimateurs ML, MMSE et MAP

On considère le cas

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{b}} &= \sigma_{\mathbf{b}}^2 \boldsymbol{I} \quad \text{(bruit i.i.d.)} \\ \boldsymbol{\Sigma}_{\mathbf{x}} &= \sigma_{\mathbf{x}}^2 \boldsymbol{I} \end{aligned}$$

alors

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{ML}} &= \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{y} \\ \hat{\mathbf{x}}_{\mathrm{MAP}} = \hat{\mathbf{x}}_{\mathrm{MMSE}} &= \left(\mathbf{H}^T \mathbf{H} + \lambda \boldsymbol{I}\right)^{-1} \left(\mathbf{H}^T \mathbf{y} + \lambda \boldsymbol{\mu}_{\mathbf{x}}\right) \end{aligned}$$

avec $\lambda = \frac{\sigma_{\mathbf{b}}^2}{\sigma_{\mathbf{x}}^2}$ (> 0).

Remarque

Inversibilité de $\left(\mathbf{H}^T \mathbf{H} + \lambda \boldsymbol{I}\right)$ garantie pour λ suffisamment grand...

Problème inverse linéaire gaussien
Estimation MAP et pénalisation

Maximiser la (log-)distribution a posteriori conduit à

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x}}{\text{argmin}} \quad \underbrace{\log f\left(\mathbf{y}|\mathbf{x}, \phi\right)}_{\text{attache aux données}} + \underbrace{\log f\left(\mathbf{x}|\phi\right)}_{\text{pénalisation}}$$

Ici

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x}}{\text{argmin}} \, \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda \, \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\|_2^2$$

$\rightarrow$ optimisation d'un critère ℓ_2–ℓ_2 (régularization de Tikhonov, aka "*ridge regression*")

Avantage

Le "cadre" bayésien fournit une interprétation probabiliste du paramètre de régularisation/pénalisation

$$\lambda = \frac{\sigma_{\mathbf{b}}^2}{\sigma_{\mathbf{x}}^2}.$$

Remarque

$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MMSE}}$ pour les lois a posteriori symétriques...

Problème inverse et inversion bayésienne

Difficultés

- Choix des hyperparamètres ϕ caractérisant le modèle *a priori*

Solutions

- Introduction d'un deuxième niveau dans le modèle bayésien,

$$\phi \sim f(\phi)$$

puis marginalisation :

$$f(\mathbf{x}|\mathbf{y}) = \frac{1}{f(\mathbf{y})} \int f(\mathbf{y}|\mathbf{x}) f(\mathbf{x}|\phi) f(\phi) d\phi$$

Problème inverse et inversion bayésienne

Difficultés

- Choix des hyperparamètres ϕ caractérisant le modèle *a priori*
- Optimisation (MAP) ou intégration (MMSE) du critère

Solutions

- Introduction d'un deuxième niveau dans le modèle bayésien,

$$\phi \sim f(\phi)$$

puis marginalisation :

$$f(\mathbf{x}|\mathbf{y}) = \frac{1}{f(\mathbf{y})} \int f(\mathbf{y}|\mathbf{x}) f(\mathbf{x}|\phi) f(\phi) d\phi$$

- Si difficile, recours à des algorithmes
  - d'optimisation pour approcher $\hat{\mathbf{x}}_{\mathrm{MAP}}$
  - de simulation stochastique pour approcher $\hat{\mathbf{x}}_{\mathrm{MMSE}}$

$$\hat{\mathbf{x}}_{\mathrm{MMSE}} = \mathrm{E}[\mathbf{x}|\mathbf{y}] \approx \frac{1}{N_{\mathrm{MC}}} \sum_{t=1}^{N_{\mathrm{MC}}} \tilde{\mathbf{x}}^{(t)} \quad \text{avec} \quad \tilde{\mathbf{x}}^{(t)} \sim f(\mathbf{x}|\mathbf{y})$$

Plan

Problématique

Pour une variable aléatoire $\theta \sim f(\theta)$, on cherche à évaluer la moyenne

$$\mathrm{E}\left[G(\theta)\right] = \int G(\theta) f(\theta) d\theta$$

Remarque
Par exemple, dans le cadre de l'estimation bayésienne (cf. précédemment)

$$\begin{aligned} G(\theta) &= \mathbf{x} \\ f(\theta) &= f(\mathbf{x}|\mathbf{y}) \end{aligned}$$

alors

$$\begin{aligned} \mathrm{E}\left[G(\theta)\right] &= \mathrm{E}\left[\mathbf{x}|\mathbf{y}\right] \\ &= \int \mathbf{x} f(\mathbf{x}|\mathbf{y}) d\mathbf{x} \\ &= \hat{\mathbf{x}}_{\mathrm{MMSE}} \end{aligned}$$

Problématique

Pour une variable aléatoire $\theta \sim f(\theta)$, on cherche à évaluer la moyenne

$$\mathrm{E}\left[G(\theta)\right] = \int G(\theta) f(\theta) d\theta$$

Remarque
Par exemple, dans le cadre de l'estimation bayésienne (cf. précédemment)

$$\begin{aligned} G(\theta) &= \mathbf{x} \\ f(\theta) &= f(\mathbf{x}|\mathbf{y}) \end{aligned}$$

alors

$$\begin{aligned} \mathrm{E}\left[G(\theta)\right] &= \mathrm{E}\left[\mathbf{x}|\mathbf{y}\right] \\ &= \int \mathbf{x} f(\mathbf{x}|\mathbf{y}) d\mathbf{x} \\ &= \hat{\mathbf{x}}_{\mathrm{MMSE}} \end{aligned}$$

Intégration de Monte Carlo

Solution

Générer un échantillon $\theta^{(1)}, \ldots, \theta^{(T)}$ distribué selon $f(\theta)$ puis approcher $\mathrm{E}\left[G(\theta)\right]$

$$\bar{G}_T = \frac{1}{T} \sum_{t=1}^{T} G\left(\theta^{(t)}\right)$$

“Preuve” : loi forte des grands nombres.

Propriété

Asymptotiquement, on a (dans un sens qu’il faudrait spécifier...)

$$\bar{G}_T \sim \mathcal{N}\left(\mathrm{E}\left[G(\theta)\right], \nu_T^2\right)$$

$\rightarrow$ fournit des intervalles de confiance...

Difficulté

Génération de $\theta^{(1)}, \ldots, \theta^{(T)} \sim f(\theta)$

Intégration de Monte Carlo

Solution

Générer un échantillon $\theta^{(1)}, \ldots, \theta^{(T)}$ distribué selon $f(\theta)$ puis approcher $\mathrm{E}\left[G(\theta)\right]$

$$\bar{G}_T = \frac{1}{T} \sum_{t=1}^{T} G\left(\theta^{(t)}\right)$$

“Preuve” : loi forte des grands nombres.

Propriété

Asymptotiquement, on a (dans un sens qu’il faudrait spécifier...)

$$\bar{G}_T \sim \mathcal{N}\left(\mathrm{E}\left[G(\theta)\right], \nu_T^2\right)$$

$\rightarrow$ fournit des intervalles de confiance...

Difficulté

Génération de $\theta^{(1)}, \ldots, \theta^{(T)} \sim f(\theta)$

Intégration de Monte Carlo

Solution
Générer un échantillon $\theta^{(1)}, \ldots, \theta^{(T)}$ distribué selon $f(\theta)$ puis approcher $\mathrm{E}\left[G(\theta)\right]$

$$\bar{G}_T = \frac{1}{T} \sum_{t=1}^{T} G\left(\theta^{(t)}\right)$$

“Preuve” : loi forte des grands nombres.

Propriété
Asymptotiquement, on a (dans un sens qu’il faudrait spécifier...)

$$\bar{G}_T \sim \mathcal{N}\left(\mathrm{E}\left[G(\theta)\right], \nu_T^2\right)$$

$\rightarrow$ fournit des intervalles de confiance...

Difficulté
Génération de $\theta^{(1)}, \ldots, \theta^{(T)} \sim f(\theta)$

Génération de variables aléatoires

Génération explicite

Pour des lois univariées et multivariées simples :

- génération par inversion de la fonction de réparation
  (e.g., loi exponentielle)
- génération par changement de variable
  (e.g., méthode de Box-Muller pour une loi normale)
- ...

Génération non explicite

Nécessaire lorsqu'une difficulté est rencontrée

- loi univariée non standard
  (e.g., loi tronquée)
- loi multivariée non standard
  (e.g., loi a posteriori dans un modèle bayésien hiérarchique)

On fait comment ?

Recours à d'autres stratégies...

Génération de variables aléatoires

Génération explicite

Pour des lois univariées et multivariées simples :

- génération par inversion de la fonction de réparation
  (e.g., loi exponentielle)
- génération par changement de variable
  (e.g., méthode de Box-Muller pour une loi normale)
- ...

Génération non explicite

Nécessaire lorsqu'une difficulté est rencontrée

- loi univariée non standard
  (e.g., loi tronquée)
- loi multivariée non standard
  (e.g., loi a posteriori dans un modèle bayésien hiérarchique)

On fait comment ?

Recours à d'autres stratégies...

Echantillonnage d'importance

Hypothèse

On ne sait pas générer θ suivant $f(\cdot)$ mais on sait générer θ suivant $q(\cdot)$ telle que $\text{supp}(q) \supset \text{supp}(f)$

Reformulation

$$\begin{aligned} \mathrm{E}_f\left(G(\theta)\right) &= \int G(\theta) f(\theta) d\theta \\ &= \int G(\theta) \frac{f(\theta)}{q(\theta)} q(\theta) d\theta \\ &= \mathrm{E}_q\left(G(\theta) \frac{f(\theta)}{q(\theta)}\right) \end{aligned}$$

Donc on peut approcher $\mathrm{E}_f\left(G(\theta)\right)$ par

$$\bar{G}_T = \frac{1}{T} \sum_{t=1}^{T} G(\theta^{(t)}) \frac{f(\theta^{(t)})}{q(\theta^{(t)})}$$

avec $\theta^{(1)}, \ldots, \theta^{(T)}$ distribué suivant $q(\theta)$, appelée loi instrumentale.

Algorithme d'acceptation-rejet

Hypothèse

On ne sait pas générer θ suivant $f(\cdot)$ mais on sait générer θ suivant $q(\cdot)$ pour laquelle il existe $M > 0$ tel que $f(x) \leq Mq(x)$ $(\forall x)$.

Algorithme

while $t < T$ **do**
  générer $z \sim q(z)$
  générer $w \sim \mathcal{U}(0,1)$
  calculer $\rho = \frac{f(z)}{Mq(z)}$
  if $w < \rho$ **then**
    $\theta^{(t)} \leftarrow z$ (accepter)
    $t \leftarrow t + 1$
  else
    rejeter
  end if
end while

Propriétés

- La probabilité moyenne d'accepter est $\frac{1}{M}$.
- Les variables $\theta^{(1)}, \ldots, \theta^{(T)}$ sont distribuées suivant $f(\theta)$
- Les variables $\theta^{(1)}, \ldots, \theta^{(T)}$ sont indépendantes.

Algorithme d'acceptation-rejet

Hypothèse

On ne sait pas générer θ suivant $f(\cdot)$ mais on sait générer θ suivant $q(\cdot)$ pour laquelle il existe $M > 0$ tel que $f(x) \leq Mq(x)$ $(\forall x)$.

Algorithme

while $t < T$ **do**
  générer $z \sim q(z)$
  générer $w \sim \mathcal{U}(0,1)$
  calculer $\rho = \frac{f(z)}{Mq(z)}$
  if $w < \rho$ **then**
    $\theta^{(t)} \leftarrow z$ (accepter)
    $t \leftarrow t + 1$
  else
    rejeter
  end if
end while

Propriétés

- La probabilité moyenne d'accepter est $\frac{1}{M}$.
- Les variables $\theta^{(1)}, \ldots, \theta^{(T)}$ sont distribuées suivant $f(\theta)$
- Les variables $\theta^{(1)}, \ldots, \theta^{(T)}$ sont indépendantes.

Algorithmes de Monte Carlo par Chaîne de Markov

Algorithme de Metropolis-Hastings

Hypothèse

On ne sait pas générer θ suivant $f(\cdot)$ mais on sait générer θ suivant une loi $q(\cdot)$.

Algorithme itératif

Initialisation: $\theta^{(0)}$
for $t = 1$ to T **do**
  générer $z \sim q(z|\theta^{(t-1)})$
  générer $w \sim \mathcal{U}(0,1)$
  calculer $\rho = \min\left\{1, \frac{f(z)}{f(\theta^{(t-1)})} \frac{q(\theta^{(t-1)}|z)}{q(z|\theta^{(t-1)})}\right\}$
  if $w < \rho$ **then**
    $\theta^{(t)} \leftarrow z$
  else
    $\theta^{(t)} \leftarrow \theta^{(t-1)}$
  end if
end for

Propriétés

- Après convergence, Les variables $\theta^{(T_0)}, \ldots, \theta^{(T)}$ sont distribuées suivant $f(\cdot)$.
- Les variables $\theta^{(1)}, \ldots, \theta^{(T)}$ sont dépendantes (elles forment une chaîne de Markov).

Algorithmes de Monte Carlo par Chaîne de Markov
Algorithme de Metropolis-Hastings

Remarques

- cas symétrique : $q(z|\theta^{(t-1)}) = q(\theta^{(t-1)}|z)$, alors

$$\rho = \min\left\{1, \frac{f(z)}{f(\theta^{(t-1)})}\right\}$$

  (e.g., marche aléatoire gaussienne)

- cas indépendant : $q(z|\theta^{(t-1)}) = q(z)$
- on peut avoir $\theta^{(t)} = \theta^{(t-1)}$
- la loi $f(\cdot)$ peut être définie à une constante près. Bonne nouvelle ! car dans le cas de l'estimation bayésienne :

$$f(\theta|Y_1, \ldots, Y_n) = \frac{f(Y_1, \ldots, Y_n|\theta)f(\theta)}{f(Y_1, \ldots, Y_n)}$$

  où $f(Y_1, \ldots, Y_n)$ est difficile à calculer.

Algorithmes de Monte Carlo par Chaîne de Markov
Echantillonneur de Gibbs

Principe

Soit une loi multivariée $f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$.

Echantillonnage itérative selon les lois conditionnelles :

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t+1)} &\sim f\left(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)}\right) \\
\boldsymbol{\theta}_2^{(t+1)} &\sim f\left(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)}\right) \\
&\vdots \\
\boldsymbol{\theta}_N &\sim f\left(\boldsymbol{\theta}_N | \boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_{N-1}^{(t+1)}\right)
\end{aligned}$$

Propriétés

- L'ensemble des T variables $\left\{\left(\theta_1^{(t)}, \ldots, \theta_N^{(t)}\right)\right\}_{t=1}^{T}$ asymptotiquement distribuées suivant la loi jointe $f(\theta_1, \ldots, \theta_N)$
- L'ensemble des T variables $\left\{\theta_j^{(t)}\right\}_{t=1}^{T}$ asymptotiquement distribuées suivant les lois marginales $f(\theta_j)$
- *Metropolis-within-Gibbs :* une des étapes peut être réalisée à l'aide d'une étape de Metropolis-Hastings.

Algorithmes de Monte Carlo par Chaîne de Markov
Echantillonneur de Gibbs

Principe

Soit une loi multivariée $f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$.

Echantillonnage itérative selon les lois conditionnelles :

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t+1)} &\sim f\left(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)}\right) \\
\boldsymbol{\theta}_2^{(t+1)} &\sim f\left(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)}\right) \\
&\vdots \\
\boldsymbol{\theta}_N &\sim f\left(\boldsymbol{\theta}_N | \boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_{N-1}^{(t+1)}\right)
\end{aligned}$$

Propriétés

- L'ensemble des T variables $\left\{\left(\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)}\right)\right\}_{t=1}^{T}$ asymptotiquement distribuées suivant la loi jointe $f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$
- L'ensemble des T variables $\left\{\boldsymbol{\theta}_j^{(t)}\right\}_{t=1}^{T}$ asymptotiquement distribuées suivant les lois marginales $f(\boldsymbol{\theta}_j)$
- *Metropolis-within-Gibbs :* une des étapes peut être réalisée à l'aide d'une étape de Metropolis-Hastings.

Algorithmes de Monte Carlo par Chaîne de Markov
Echantillonneur de Gibbs

Application à un modèle bayésien hiérarchique

- vraisemblance $f(\mathbf{y}|\boldsymbol{\theta})$
- loi a priori des paramètres $f(\boldsymbol{\theta}|\boldsymbol{\phi})$
- loi a priori des hyperparamètres $f(\boldsymbol{\phi})$

L'algorithme s'écrit

$$\begin{aligned} \boldsymbol{\theta}^{(t)} &\sim f\left(\boldsymbol{\theta}|\boldsymbol{\phi}^{(t-1)}, \mathbf{y}\right) \\ \boldsymbol{\phi}^{(t)} &\sim f\left(\boldsymbol{\phi}|\boldsymbol{\theta}^{(t)}\right) \end{aligned}$$

On a alors

$$\hat{\boldsymbol{\theta}}_{\mathrm{MMSE}} \approx \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)} \quad \text{et} \quad \hat{\boldsymbol{\phi}}_{\mathrm{MMSE}} \approx \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}^{(t)} \tag{2}$$

Application à l'augmentation de modèle

La loi cible $f(\boldsymbol{\theta})$

- n'est pas simple à échantillonner
- mais s'écrit $f(\boldsymbol{\theta}) = \int f(\boldsymbol{\theta}, \mathbf{z})\, d\mathbf{z}$

avec $f(\boldsymbol{\theta}|\mathbf{z})$ et $f(\mathbf{z}|\boldsymbol{\theta})$ simples à échantillonner.

Plan

Limitations

Algorithme de Metropolis-Hastings (MH)

Soumis au choix de la loi instrumentale $q(\cdot)$

- assurer de bonnes propriétés de mélange (limiter la corrélation entre les échantillons)
- assurer un taux d'acceptation raisonnable (e.g., entre 0.4 et 0.6)

et ne tient pas compte de la nature du critère : où sont les régions à haute densité ?

Algorithme de Gibbs

Echantillonnage composante par composante :

- diminution de l'espace à explorer
- choix plus aisée de la loi instrumentale $q(\cdot)$ si recours à une étape de MH

Mais

- (nécessite d'avoir accès à toutes les lois conditionnelles)
- mauvaises propriétés de mélanges (forte corrélation entre les échantillons)
- coûteux en temps de calcul

Alternatives

- recours à des modèles de la physique computationnelle/statistique
  diffusion de Langevin, dynamique Hamiltonienne
- idées empruntées aux avancées récentes en optimisation

Limitations

Algorithme de Metropolis-Hastings (MH)

Soumis au choix de la loi instrumentale $q(\cdot)$

- assurer de bonnes propriétés de mélange (limiter la corrélation entre les échantillons)
- assurer un taux d'acceptation raisonnable (e.g., entre 0.4 et 0.6)

et ne tient pas compte de la nature du critère : où sont les régions à haute densité ?

Algorithme de Gibbs

Echantillonnage composante par composante :

- diminution de l'espace à explorer
- choix plus aisée de la loi instrumentale $q(\cdot)$ si recours à une étape de MH

Mais

- (nécessite d'avoir accès à toutes les lois conditionnelles)
- mauvaises propriétés de mélanges (forte corrélation entre les échantillons)
- coûteux en temps de calcul

Alternatives

- recours à des modèles de la physique computationnelle/statistique
  diffusion de Langevin, dynamique Hamiltonienne
- idées empruntées aux avancées récentes en optimisation

Exact Perturbation Optimization (E-PO)

Cadre : modèle linéaire gaussien

Hypothèses :

- Vraisemblance : $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ avec $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$
- Loi a priori : $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$

On cherche à échantillonner suivant $f(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \mathbf{Q}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]$ avec

$$\mathbf{Q} = \mathbf{H}^T \boldsymbol{\Sigma}_\mathbf{w}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \tag{3}$$

$$\mathbf{Q}\boldsymbol{\mu} = \mathbf{H}^T \boldsymbol{\Sigma}_\mathbf{w}^{-1} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{w}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}} \tag{4}$$

Algorithme E-PO

- Générer $\mathbf{z}_\mathbf{w} \sim \mathcal{N}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$
- Générer $\mathbf{z}_{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$
- Poser $\boldsymbol{\eta} = \mathbf{H}^T \boldsymbol{\Sigma}_\mathbf{w}^{-1} \mathbf{z}_\mathbf{w} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z}_{\boldsymbol{\theta}}$
- Résoudre $\boldsymbol{\theta} = \mathbf{Q}^{-1} \boldsymbol{\eta}$

Truncated & Reversible Jump Perturbation Optimization (T-PO & RJ-PO)

Difficulté
Résolution du problème linéaire

$$\boldsymbol{\theta} = \mathbf{Q}^{-1}\boldsymbol{\eta}$$

avec $\mathbf{Q} = \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

Truncated Perturbation Optimization (T-PO)

- Utilisation d'un *solver* itératif
- Résolution approchée $\boldsymbol{\theta}^*$

Reversible Jump Perturbation Optimization (RJ-PO)

- Correction de l'approximation par une étape de Metropolis-Hastings avec

$$\rho = \min\left\{1, \exp\left[\left(\mathbf{Q}\boldsymbol{\theta}^* - \boldsymbol{\mu}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)\right]\right\} \tag{5}$$

Remarque : d'autres méthodes efficaces si $\mathbf{Q} = \mathbf{Q}_1 + \mathbf{Q}_2$ avec $\mathbf{Q}_1$ et $\mathbf{Q}_2$ de structures particulières (e.g., diagonalisables dans la même base).

Truncated & Reversible Jump Perturbation Optimization (T-PO & RJ-PO)

Difficulté

Résolution du problème linéaire

$$\boldsymbol{\theta} = \mathbf{Q}^{-1}\boldsymbol{\eta}$$

avec $\mathbf{Q} = \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

Truncated Perturbation Optimization (T-PO)

- Utilisation d'un *solver* itératif
- Résolution approchée $\boldsymbol{\theta}^*$

Reversible Jump Perturbation Optimization (RJ-PO)

- Correction de l'approximation par une étape de Metropolis-Hastings avec

$$\rho = \min\left\{1, \exp\left[\left(\mathbf{Q}\boldsymbol{\theta}^* - \boldsymbol{\mu}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)\right]\right\} \tag{5}$$

Remarque : d'autres méthodes efficaces si $\mathbf{Q} = \mathbf{Q}_1 + \mathbf{Q}_2$ avec $\mathbf{Q}_1$ et $\mathbf{Q}_2$ de structures particulières (e.g., diagonalisables dans la même base).

Hamiltonian Monte Carlo

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$.

Principe : modèle étendu

- On considère le hamiltonien associé à $U(\cdot)$

$$H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) \tag{6}$$

  où $K(\mathbf{p})$ est une fonction d'énergie cinétique, e.g., $K(\mathbf{p}) = \frac{\mathbf{p}^T\mathbf{p}}{2}$
- Simulation de la dynamique Hamiltonienne après discrétisation de l'équation différentielle associée
- Échantillonnage de la loi jointe $f(\boldsymbol{\theta})g(\mathbf{p})$ à l'aide d'une étape de MH

Génération du candidat $(\boldsymbol{\theta}^*, \mathbf{p}^*)$

- Initialisation du candidat $\mathbf{p}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:
- Leapfrogs par N_{LF} itérations du schéma :

$$\mathbf{p}^{(n+\frac{1}{2})} = \mathbf{p}^{(n)} - \frac{\epsilon}{2}\nabla U(\boldsymbol{\theta}^{(n)}) \tag{7}$$

$$\boldsymbol{\theta}^{(n+\frac{1}{2})} = \boldsymbol{\theta}^{(n)} + \epsilon\mathbf{p}^{(n+\frac{1}{2})} \tag{8}$$

$$\mathbf{p}^{(n+1)} = \boldsymbol{\theta}^{(n+\frac{1}{2})} - \frac{\epsilon}{2}\nabla U(\boldsymbol{\theta}^{(n+1)}) \tag{9}$$

- On choisit $(\boldsymbol{\theta}^*, \mathbf{p}^*) = (\boldsymbol{\theta}^{(N_{\mathrm{LF}})}, \mathbf{p}^{(N_{\mathrm{LF}})})$

Hamiltonian Monte Carlo

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$.

Principe : modèle étendu

- On considère le hamiltonien associé à $U(\cdot)$

$$H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) \tag{6}$$

  où $K(\mathbf{p})$ est une fonction d'énergie cinétique, e.g., $K(\mathbf{p}) = \frac{\mathbf{p}^T\mathbf{p}}{2}$
- Simulation de la dynamique Hamiltonienne après discrétisation de l'équation différentielle associée
- Échantillonnage de la loi jointe $f(\boldsymbol{\theta})g(\mathbf{p})$ à l'aide d'une étape de MH

Génération du candidat $(\boldsymbol{\theta}^*, \mathbf{p}^*)$

- Initialisation du candidat $\mathbf{p}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:
- Leapfrogs par N_{LF} itérations du schéma :

$$\mathbf{p}^{(n+\frac{1}{2})} = \mathbf{p}^{(n)} - \frac{\epsilon}{2}\nabla U(\boldsymbol{\theta}^{(n)}) \tag{7}$$

$$\boldsymbol{\theta}^{(n+\frac{1}{2})} = \boldsymbol{\theta}^{(n)} + \epsilon\mathbf{p}^{(n+\frac{1}{2})} \tag{8}$$

$$\mathbf{p}^{(n+1)} = \boldsymbol{\theta}^{(n+\frac{1}{2})} - \frac{\epsilon}{2}\nabla U(\boldsymbol{\theta}^{(n+1)}) \tag{9}$$

- On choisit $(\boldsymbol{\theta}^*, \mathbf{p}^*) = (\boldsymbol{\theta}^{(N_{\mathrm{LF}})}, \mathbf{p}^{(N_{\mathrm{LF}})})$

Hamiltonian Monte Carlo

Propriété

On échantillonne suivant la loi jointe $f(\boldsymbol{\theta})g(\mathbf{p})$ donc les échantillons $\left\{\boldsymbol{\theta}^{(t)}\right\}_{t=1}^{T}$ sont distribuées suivant $f(\cdot)$.

Metropolis Adjusted Langevin Algorithm (MALA)

Si $N_{\mathrm{LF}} = 1$, le candidat s'écrit

$$\boldsymbol{\theta}^{*} = \boldsymbol{\theta}^{(t)} - \delta \nabla U(\boldsymbol{\theta}^{(t)}) + \sqrt{2\delta}\mathbf{p} \tag{10}$$

avec $\delta = \frac{\epsilon^2}{2}$ et $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

- Approximation à temps discret du processus de diffusion de Langevin
- Version non-ajustée (ULA) : $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{*}$

Remarques

- Proposition d'un candidat θ^{*} dans les régions de haut potentiel
- Choix parfois difficile de ϵ et N_{LF}
- Nécessite la différentiabilité du potentiel $U(\cdot)$
- Généralisations de HMC et MALA avec pré-conditionnement, e.g.,

$$\theta^{*} = \theta^{(t)} - \delta \mathbf{G}^{-1}(\theta^{(t)}) \nabla U(\theta^{(t)}) + \sqrt{2\delta \mathbf{G}^{-1}(\theta^{(t)})}\mathbf{p} \tag{11}$$

où $\mathbf{G}(\cdot)$ définit une variété d'intérêt (MIF, Hessien, ...).

Hamiltonian Monte Carlo

Propriété

On échantillonne suivant la loi jointe $f(\boldsymbol{\theta})g(\mathbf{p})$ donc les échantillons $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$ sont distribuées suivant $f(\cdot)$.

Metropolis Adjusted Langevin Algorithm (MALA)

Si $N_{\mathrm{LF}} = 1$, le candidat s'écrit

$$\boldsymbol{\theta}^{*} = \boldsymbol{\theta}^{(t)} - \delta \nabla U(\boldsymbol{\theta}^{(t)}) + \sqrt{2\delta}\mathbf{p} \tag{10}$$

avec $\delta = \frac{\epsilon^2}{2}$ et $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

- Approximation à temps discret du processus de diffusion de Langevin
- Version non-ajustée (ULA) : $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{*}$

Remarques

- Proposition d'un candidat $\boldsymbol{\theta}^{*}$ dans les régions de haut potentiel
- Choix parfois difficile de ϵ et N_{LF}
- Nécessite la différentiabilité du potentiel $U(\cdot)$
- Généralisations de HMC et MALA avec pré-conditionnement, e.g.,

$$\boldsymbol{\theta}^{*} = \boldsymbol{\theta}^{(t)} - \delta \mathbf{G}^{-1}(\boldsymbol{\theta}^{(t)}) \nabla U(\boldsymbol{\theta}^{(t)}) + \sqrt{2\delta \mathbf{G}^{-1}(\boldsymbol{\theta}^{(t)})}\mathbf{p} \tag{11}$$

où $\mathbf{G}(\cdot)$ définit une variété d'intérêt (MIF, Hessien, ...).

Proximal Metropolis Adjusted Langevin Algorithm (P-MALA)

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$ où $U(\cdot)$ définit un potentiel convexe mais non lisse.

Principe : approximation de Moreau de $f(\cdot)$

La loi cible $f(\cdot)$ est approchée par

$$f_\lambda(\boldsymbol{\theta}) \propto \exp\left[-U_\lambda(\boldsymbol{\theta})\right] \tag{12}$$

où $U_\lambda(\cdot)$ est l'enveloppe de Moreau-Yoshida de $U(\cdot)$

$$U_\lambda(\boldsymbol{\theta}) = \inf_{\mathbf{v}} \left\{ U(\boldsymbol{\theta}) - \frac{1}{2\lambda} \|\boldsymbol{\theta} - \mathbf{v}\|_2^2 \right\} \tag{13}$$

et λ contrôle la qualité de l'approximation (convergence ponctuelle):

$$\lim_{\lambda \to 0} f_\lambda(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) \qquad (\forall \boldsymbol{\theta}). \tag{14}$$

Proximal Metropolis Adjusted Langevin Algorithm (P-MALA)

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$ où $U(\cdot)$ définit un potentiel convexe mais non lisse.

Principe : approximation de Moreau de $f(\cdot)$

La loi cible $f(\cdot)$ est approchée par

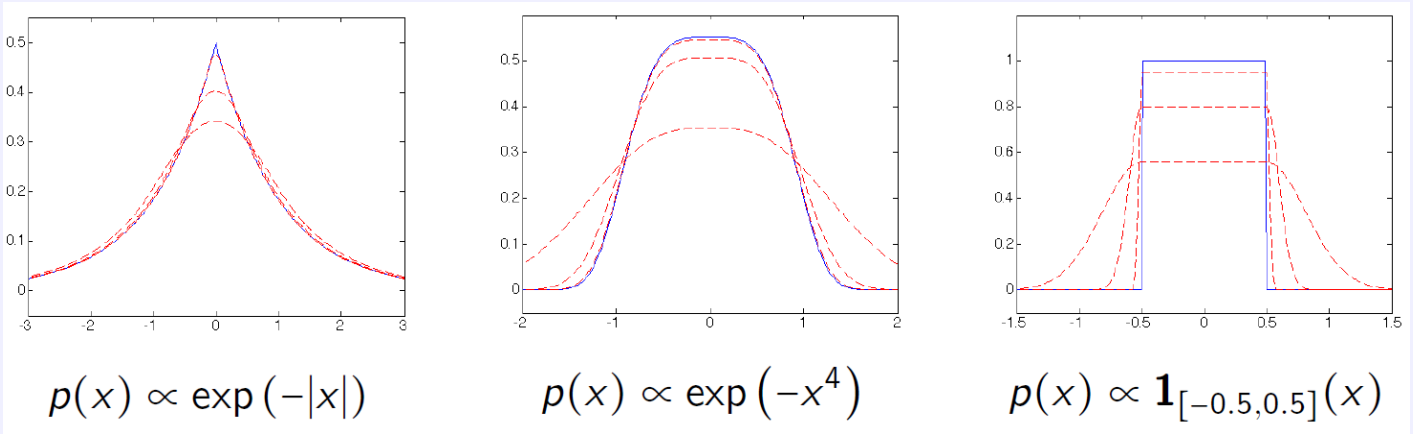
$$f_\lambda(\boldsymbol{\theta}) \propto \exp\left[-U_\lambda(\boldsymbol{\theta})\right] \tag{12}$$

où $U_\lambda(\cdot)$ est l'enveloppe de Moreau-Yoshida de $U(\cdot)$

$$U_\lambda(\boldsymbol{\theta}) = \inf_{\mathbf{v}} \left\{ U(\boldsymbol{\theta}) - \frac{1}{2\lambda} \|\boldsymbol{\theta} - \mathbf{v}\|_2^2 \right\} \tag{13}$$

et λ contrôle la qualité de l'approximation (convergence ponctuelle):

$$\lim_{\lambda \to 0} f_\lambda(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) \qquad (\forall \boldsymbol{\theta}). \tag{14}$$



Extrait de [Pereyra2016].

Proximal Metropolis Adjusted Langevin Algorithm (P-MALA)

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$ où $U(\cdot)$ définit un potentiel convexe mais non lisse.

Principe : approximation de Moreau de $f(\cdot)$

La loi cible $f(\cdot)$ est approchée par

$$f_\lambda(\boldsymbol{\theta}) \propto \exp\left[-U_\lambda(\boldsymbol{\theta})\right] \tag{12}$$

où $U_\lambda(\cdot)$ est l'enveloppe de Moreau-Yoshida de $U(\cdot)$

$$U_\lambda(\boldsymbol{\theta}) = \inf_{\mathbf{v}} \left\{ U(\boldsymbol{\theta}) - \frac{1}{2\lambda} \|\boldsymbol{\theta} - \mathbf{v}\|_2^2 \right\} \tag{13}$$

et λ contrôle la qualité de l'approximation (convergence ponctuelle):

$$\lim_{\lambda \to 0} f_\lambda(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) \qquad (\forall \boldsymbol{\theta}). \tag{14}$$

Propriétés

- $f_\lambda(\cdot)$ définit bien une densité de probabilité
- $U_\lambda(\cdot)$ est gradient Lipchitz avec

$$\nabla U_\lambda(\boldsymbol{\theta}) = \frac{1}{\lambda} \left\{ \boldsymbol{\theta} - \mathrm{prox}_U^\lambda(\boldsymbol{\theta}) \right\} \tag{15}$$

Proximal Metropolis Adjusted Langevin Algorithm (P-MALA)

Algorithme de Langevin ajusté (MALA) appliqué à $f_\lambda(\cdot)$

En utilisant l'identité

$$\nabla U_\lambda(\boldsymbol{\theta}) = \frac{1}{\lambda}\left(\boldsymbol{\theta} + \mathrm{prox}_U^\lambda(\boldsymbol{\theta})\right) \tag{16}$$

il vient

$$\boldsymbol{\theta}^* = \left(1 - \frac{\delta}{\lambda}\right)\boldsymbol{\theta}^{(t)} - \frac{\delta}{\lambda}\mathrm{prox}_U^\lambda(\boldsymbol{\theta}^{(t)}) + \sqrt{2\delta}\mathbf{p} \tag{17}$$

avec $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Remarques

- Permet d'échantillonner une distribution de potentiel non-lisse
- Nécessite la connaissance de $\mathrm{prox}_U^\lambda(\cdot)$
  - pas direct, notamment si $U(\cdot) = E(\cdot) + R(\cdot)$
  - approximation possible si $E(\cdot)$ est lisse:

$$\boldsymbol{\theta}^* = \mathrm{prox}_R^\lambda\left(\boldsymbol{\theta}^{(t)} - \delta\nabla E(\boldsymbol{\theta}^{(t)})\right) + \sqrt{2\delta}\mathbf{p} \tag{18}$$

  (algorithme de type *forward-backard* perturbé)
- Nécessite une étape de Metropolis-Hastings pour corriger l'approximation
  - coûteux en temps de calcul

Proximal Metropolis Adjusted Langevin Algorithm (P-MALA)

Algorithme de Langevin ajusté (MALA) appliqué à $f_\lambda(\cdot)$

En utilisant l'identité

$$\nabla U_\lambda(\boldsymbol{\theta}) = \frac{1}{\lambda}\left(\boldsymbol{\theta} + \mathrm{prox}_U^\lambda(\boldsymbol{\theta})\right) \tag{16}$$

il vient

$$\boldsymbol{\theta}^* = \left(1 - \frac{\delta}{\lambda}\right)\boldsymbol{\theta}^{(t)} - \frac{\delta}{\lambda}\mathrm{prox}_U^\lambda(\boldsymbol{\theta}^{(t)}) + \sqrt{2\delta}\mathbf{p} \tag{17}$$

avec $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Remarques

- Permet d'échantillonner une distribution de potentiel non-lisse
- Nécessite la connaissance de $\mathrm{prox}_U^\lambda(\cdot)$
  - · pas direct, notamment si $U(\cdot) = E(\cdot) + R(\cdot)$
  - · approximation possible si $E(\cdot)$ est lisse:

$$\boldsymbol{\theta}^* = \mathrm{prox}_R^\lambda\left(\boldsymbol{\theta}^{(t)} - \delta\nabla E(\boldsymbol{\theta}^{(t)})\right) + \sqrt{2\delta}\mathbf{p} \tag{18}$$

  (algorithme de type *forward-backard* perturbé)
- Nécessite une étape de Metropolis-Hastings pour corriger l'approximation
  - · coûteux en temps de calcul

Moreau-Yoshida regularised ULA (MYULA)

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$ avec $U(\cdot) = E(\cdot) + R(\cdot)$, où $E(\cdot)$ et $R(\cdot)$ sont convexes, $E(\cdot)$ est L_E-gradient Lipschitz mais $R(\cdot)$ non lisse.

Principe : approximation de Moreau-Yoshida de $R(\cdot)$

La loi cible $f(\cdot)$ est approchée par

$$f_\lambda(\boldsymbol{\theta}) \propto \exp\left[-E(\boldsymbol{\theta}) - R_\lambda(\boldsymbol{\theta})\right] \tag{19}$$

où $R_\lambda(\cdot)$ est l'enveloppe de Moreau-Yoshida de $R(\cdot)$

$$R_\lambda(\boldsymbol{\theta}) = \inf_{\mathbf{v}} \left\{ R(\boldsymbol{\theta}) - \frac{1}{2\lambda} \|\boldsymbol{\theta} - \mathbf{v}\|_2^2 \right\} \tag{20}$$

et λ contrôle la qualité de l'approximation :

$$\lim_{\lambda \to 0} \|f - f_\lambda\|_{\mathrm{TV}} = 0. \tag{21}$$

Propriétés

- $f_\lambda(\cdot)$ définit bien une densité de probabilité
- $U_\lambda(\cdot) = E(\cdot) + R_\lambda(\cdot)$ est gradient Lipchitz avec

$$\nabla U_\lambda(\boldsymbol{\theta}) = \nabla E(\boldsymbol{\theta}) + \frac{1}{\lambda}\left\{\boldsymbol{\theta} - \mathrm{prox}_R^\lambda(\boldsymbol{\theta})\right\} \tag{22}$$

Moreau-Yoshida regularised ULA (MYULA)

Hypothèse

On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$ avec $U(\cdot) = E(\cdot) + R(\cdot)$, où $E(\cdot)$ et $R(\cdot)$ sont convexes, $E(\cdot)$ est L_E-gradient Lipschitz mais $R(\cdot)$ non lisse.

Principe : approximation de Moreau-Yoshida de $R(\cdot)$

La loi cible $f(\cdot)$ est approchée par

$$f_\lambda(\boldsymbol{\theta}) \propto \exp\left[-E(\boldsymbol{\theta}) - R_\lambda(\boldsymbol{\theta})\right] \tag{19}$$

où $R_\lambda(\cdot)$ est l'enveloppe de Moreau-Yoshida de $R(\cdot)$

$$R_\lambda(\boldsymbol{\theta}) = \inf_{\mathbf{v}} \left\{ R(\boldsymbol{\theta}) - \frac{1}{2\lambda} \|\boldsymbol{\theta} - \mathbf{v}\|_2^2 \right\} \tag{20}$$

et λ contrôle la qualité de l'approximation :

$$\lim_{\lambda \to 0} \|f - f_\lambda\|_{\mathrm{TV}} = 0. \tag{21}$$

Propriétés

- $f_\lambda(\cdot)$ définit bien une densité de probabilité
- $U_\lambda(\cdot) = E(\cdot) + R_\lambda(\cdot)$ est gradient Lipchitz avec

$$\nabla U_\lambda(\boldsymbol{\theta}) = \nabla E(\boldsymbol{\theta}) + \frac{1}{\lambda}\left\{\boldsymbol{\theta} - \mathrm{prox}_R^\lambda(\boldsymbol{\theta})\right\} \tag{22}$$

Moreau-Yoshida regularised ULA (MYULA)

Algorithme de Langevin non-ajusté (ULA) appliqué à $f_\lambda(\cdot)$

En utilisant l'identité

$$\nabla R_\lambda(\boldsymbol{\theta}) = \frac{1}{\lambda}\left(\boldsymbol{\theta} - \mathrm{prox}_R^\lambda(\boldsymbol{\theta})\right) \tag{23}$$

il vient

$$\boldsymbol{\theta}^{(t+1)} = \left(1 - \frac{\delta}{\lambda}\right)\boldsymbol{\theta}^{(t)} - \delta\nabla E(\boldsymbol{\theta}^{(t)}) + \frac{\delta}{\lambda}\mathrm{prox}_R^\lambda(\boldsymbol{\theta}^{(t)}) + \sqrt{2\delta}\mathbf{p} \tag{24}$$

avec $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Remarques

- Permet d'échantillonner une distribution de potentiel non-lisse
  mais le potentiel $E(\cdot)$ doit l'être
- Propriétés de mélange guidée par la constante $L_E + \lambda^{-1}$
  compromis entre qualité de l'approximation et propriétés de mélange
- Si non convexité de $E(\cdot)$ et $R(\cdot) \rightarrow$ *Forward-Backward Langevin Algorithm*

$$\boldsymbol{\theta}^{(t+1)} = \left(1 - \frac{\delta}{\lambda}\right)\boldsymbol{\theta}^{(t)} + \frac{\delta}{\lambda}\mathrm{prox}_R^\lambda\left(\boldsymbol{\theta}^{(t)} - \lambda\nabla E(\boldsymbol{\theta}^{(t)})\right) + \sqrt{2\delta}\mathbf{p} \tag{25}$$

  interprété comme un algorithme *forward-backward proximal splitting* perturbé

Split-and-augmented Gibbs sampler

Hypothèse
On cherche à échantillonner suivant $f(\boldsymbol{\theta}) \propto \exp\left[-U(\boldsymbol{\theta})\right]$ avec $U(\cdot) = E(\cdot) + R(\cdot)$, où $E(\cdot)$ et $R(\cdot)$ sont convexes, $E(\cdot)$ est L_E-gradient Lipschitz mais $R(\cdot)$ non lisse.

Formalisme bayésien et variable splitting
L'estimateur MAP est calculé par résolution du problème

$$\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}) + R(\boldsymbol{\theta}). \tag{26}$$

Un problème équivalent est

$$\min_{\boldsymbol{\theta},\mathbf{z}} E(\boldsymbol{\theta}) + R(\mathbf{z}) \qquad \text{s.c.} \qquad \boldsymbol{\theta} = \mathbf{z} \tag{27}$$

ou

$$\min_{\boldsymbol{\theta},\mathbf{z}} E(\boldsymbol{\theta}) + R(\mathbf{z}) + \frac{1}{2\rho^2} \|\boldsymbol{\theta} - \mathbf{z}\|_2^2 \tag{28}$$

et, e.g., résolution par *alternating direction method of multipliers* (ADMM).

Split-and-augmented Gibbs sampler

Principe : échantillonnage de la split distribution

La loi cible $f(\cdot)$ est remplacée par la loi jointe

$$f_\rho(\boldsymbol{\theta}, \mathbf{z}) \propto \exp\left[-E(\boldsymbol{\theta}) - R(\mathbf{z}) - \frac{1}{2\rho^2}\|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right] \tag{29}$$

puis échantillonnée à l'aide d'un échantillonneur de Gibbs

- $\boldsymbol{\theta}|\mathbf{z} \sim \exp\left(-E(\boldsymbol{\theta}) - \frac{1}{2\rho^2}\|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right)$
- $\mathbf{z}|\boldsymbol{\theta} \sim \exp\left(-R(\mathbf{z}) - \frac{1}{2\rho^2}\|\boldsymbol{\theta} - \mathbf{z}\|_2^2\right)$

Propriétés

- Echantillonnage plus simple et/ou plus efficace suivant chaque loi conditionnelle
  E-PO, MYULA...
- Soit $\tilde{f}_\rho(\boldsymbol{\theta}) = \int f_\rho(\boldsymbol{\theta}, \mathbf{z}) d\mathbf{z}$ la marginale d'intérêt. On a $\lim_{\rho \to 0} \left\| f - \tilde{f}_\rho \right\|_{\mathrm{TV}} = 0$.
- Stratégie qui se généralise à des lois cibles de la forme

$$f(\boldsymbol{\theta}) \propto \exp\left[-\sum_{i=1}^{n} E(\mathbf{h}_i^T \boldsymbol{\theta}) - R(\boldsymbol{\theta})\right]$$

- Permet de distribuer l'échantillonnage (efficacité, *data privacy*).

Plan

Conclusions

Inférence bayésienne

Moyen élégant

- de décrire une connaissance a priori
- de régulariser un problème mal posé ou mal conditionné
- d'interpréter (et d'estimer) des hyperparamètres

Mais conduit généralement à des estimateurs difficiles à calculer.

Méthodes de Monte Carlo

- description complète (et non uniquement ponctuelle) de la loi cible
- approximation d'un estimateur ponctuel (e.g., MMSE)
- intervalles de crédibilité

Mais souvent coûteuses en temps de calcul.

Non abordées aujourd'hui

- *Approximate Bayesian Computation* (ABC)
- méthodes de quasi-Monte Carlo
- méthodes bayesiennes variationnelles
- méthodes de Monte Carlo séquentielles

Références

Inférence bayésienne et méthodes de Monte Carlo

- J. Bernardo and A. Smith, *Bayesian Theory*. New York: Wiley, 1994.
- W. Gilks, S. Richardson, and D. Spiegehalter, *Markov Chain Monte Carlo in Practice*. London, UK: Chapman & Hall, 1999.
- C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York, NY, USA: Springer, 2004.
- J.-M. Marin and C. P. Robert, *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York, NY, USA: Springer, 2007.
- C. P. Robert, *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer-Verlag, 2007.
- J. Idier, Ed., *Bayesian Approach to Inverse Problems*, ser. Digital Signal and Image Processing. Hoboken, NJ: Wiley-ISTE, 2008.
- S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, Ed., *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL, USA: CRC, 2011.

Références

Simulation de lois normales à grande dimension

- G. Papandreou and A. Yuille, “Gaussian sampling by local perturbations,” *Adv. in Neural Information Processing Systems (NIPS)*, vol. 23, 1858-1866, 2010.
- F. Orieux, O. Féron, and J.-F. Giovannelli, “Sampling high-dimensional Gaussian distributions for general linear inverse problems,” *IEEE Signal Process. Letters*, vol. 19, no. 5, pp. 251-254, May 2012.
- C. Gilavert, S. Moussaoui, and J. Idier, “Efficient Gaussian sampling for solving large-scale inverse problems using MCMC,” *IEEE Trans. Signal Processing*, vol. 63, no. 1, pp. 70-80, Jan. 2015.
- O. Féron, F. Orieux, and J.-F. Giovannelli, “Gradient scan Gibbs sampler: an efficient algorithm for high-dimensional Gaussian distributions,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 343-352, 2016.
- Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia and J.-C. Pesquet, “An auxiliary variable method for MCMC algorithms in high dimension,” *Entropy*, Vol. 20, No. 110, 2018.

Références

Hamiltonian Monte Carlo et algorithmes de Langevin

- S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, “Hybrid Monte Carlo,” *Phys. Lett. B*, vol. 195, no. 2, pp. 216-222, Sept. 1987.
- G. Roberts and O. Stramer, “Langevin diffusions and Metropolis-Hastings algorithms,” *Methodol. Comput. Appl. Probabil.*, vol. 4, pp. 337-358, 2003.
- C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, “An introduction to MCMC for machine learning,” *Mach. Learning*, vol. 50, no. 1, pp. 3-43, Jan. 2003.
- R. M. Neal, “MCMC using Hamiltonian dynamics,” in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, Eds. Boca Raton, FL, USA: CRC, 2011, pp. 93-112.
- M. Girolami and B. Calderhead, “Riemann manifold Langevin and Hamiltonian Monte Carlo methods,” *J. Roy. Stat. Soc. Ser. B*, vol. 73, pp. 123-214, 2011.
- C. Vacar, J.-F. Giovannelli and Y. Berthoumieu, “Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance,” in *Proc. IEEE ICASSP*, Prague, 2011, pp. 3964-3967.

Références

Méthodes de Monte Carlo proximales

- M. Pereyra, “Proximal Markov chain Monte Carlo algorithms,” Statistics and Computing, vol. 26, no. 4, pp 745-760, Jul. 2016
- T. Duy Luu, J. Fadili and C. Chesneau, “Sampling from non-smooth distribution through Langevin diffusion,” 2017. [Online]. Available: `https://www.archives-ouvertes.fr/hal-01492056`.
- A. Durmus, E. Moulines and M. Pereyra, “Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau,” *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, 473-506. Mar. 2018.

Références

Splitting/augmented-based Monte Carlo methods

- D. Geman and G. Reynolds, “Constrained restoration and the recovery of discontinuities,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367-383, March 1992.
- D. Geman and C. Yang, “Nonlinear image recovery with half-quadratic regularization,” *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932-946, July 1995.
- D. A. van Dyk and X.-L. Meng, “The art of data augmentation,” *J. Comput. Graph. Stat.*, vol. 10, no. 1, pp. 1-50, 2001.
- Bardenet, A. Doucet and C. Holmes, “On Markov chain Monte Carlo methods for tall data,” *J. Machine Learning Research*, 2017.
- M. Vono, N. Dobigeon and P. Chainais, “Split-and-Augmented Gibbs sampler - Application to large scale inverse problems,” 2018. [Online]. Available: `http://arxiv.org/abs/1804.05809/`.
- M. Vono, N. Dobigeon and P. Chainais, ”Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler,” in *Proc. IEEE Int. Workshop Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, Sept. 2018.
- L. J. Rendell, A. M. Johansen, A. Lee and Nick Whiteley, “Global consensus Monte Carlo,” 2018. [Online]. Available: `http://arxiv.org/abs/1807.09288/`.

Inférence bayésienne et méthodes de Monte Carlo

Une introduction

Nicolas Dobigeon

University of Toulouse, IRIT/INP-ENSEEIHT
Institut Universitaire de France (IUF)
http://dobigeon.perso.enseeiht.fr

Rencontre GdR ISIS-OG, 8 Octobre 2018