Statistique Bayésienne

Le CEPE

J-B Salomond

jean-bernard.salomond@u-pec.fr

8, 9 et 10 octobre 2018

Chap.1

Introduction

Exemple introductif

Formule de Bayes
- Retour sur les probabilités
- Combiner les informations
- Loi a posteriori

Loi a priori et a posteriori
- Exemple détaillé
- Changer la loi a priori

Choix des lois a priori
- Comment choisir une loi a priori
- Zoologie des lois a priori

Exemple introductif

Introduction

On s'intéresse au (log) salaire des diplômés d'un M2 en Data Science. On va se donner un modèle probabiliste pour ces données.

θ Le log salaire moyen

Y Le log du salaire d'un individu : $Y \sim \mathcal{N}(\theta, \sigma^2)$

Exemple introductif

Introduction

On s'intéresse au (log) salaire des diplômés d'un M2 en Data Science. On va se donner un modèle probabiliste pour ces données.

θ Le log salaire moyen

Y Le log du salaire d'un individu : $Y \sim \mathcal{N}(\theta, \sigma^2)$

On va observer le salaire de n anciens étudiants, on aura donc accès à la réalisation de n variables aléatoires $Y_1, \ldots, Y_n$.

Exemple introductif

Introduction

On s'intéresse au (log) salaire des diplômés d'un M2 en Data Science. On va se donner un modèle probabiliste pour ces données.

- θ Le log salaire moyen
- Y Le log du salaire d'un individu : $Y \sim \mathcal{N}(\theta, \sigma^2)$

On va observer le salaire de n anciens étudiants, on aura donc accès à la réalisation de n variables aléatoires $Y_1, \ldots, Y_n$. De ce modèle on peut en déduire une fonction de vraisemblance

$$\mathcal{L}(\theta | y_1, \ldots, y_n) = \frac{1}{(2\pi)^{n/2}} e^{\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2}$$

Exemple introductif

- L'approche classique des statistique consiste à trouver le paramètre inconnu du modèle θ à partir des données en maximisant la vraisemblance, par la méthode des moments, etc..

Exemple introductif

- L'approche classique des statistique consiste à trouver le paramètre inconnu du modèle θ à partir des données en maximisant la vraisemblance, par la méthode des moments, etc..
- L'approche Bayésienne est différente : On modélise l'incertitude sur θ par une distribution de probabilité π appelée distribution a priori.

Exemple introductif

- L'approche classique des statistique consiste à trouver le paramètre inconnu du modèle θ à partir des données en maximisant la vraisemblance, par la méthode des moments, etc..
- L'approche Bayésienne est différente : On modélise l'incertitude sur θ par une distribution de probabilité π appelée distribution a priori. La vraisemblance s'interprète comme la distribution des données conditionnellement au paramètre θ.

Exemple introductif

- L'approche classique des statistique consiste à trouver le paramètre inconnu du modèle θ à partir des données en maximisant la vraisemblance, par la méthode des moments, etc..
- L'approche Bayésienne est différente : On modélise l'incertitude sur θ par une distribution de probabilité π appelée distribution a priori. La vraisemblance s'interprète comme la distribution des données conditionnellement au paramètre θ. Par la formule de Bayes on inverse le conditionnement et on obtient la loi du paramètre sachant les observations

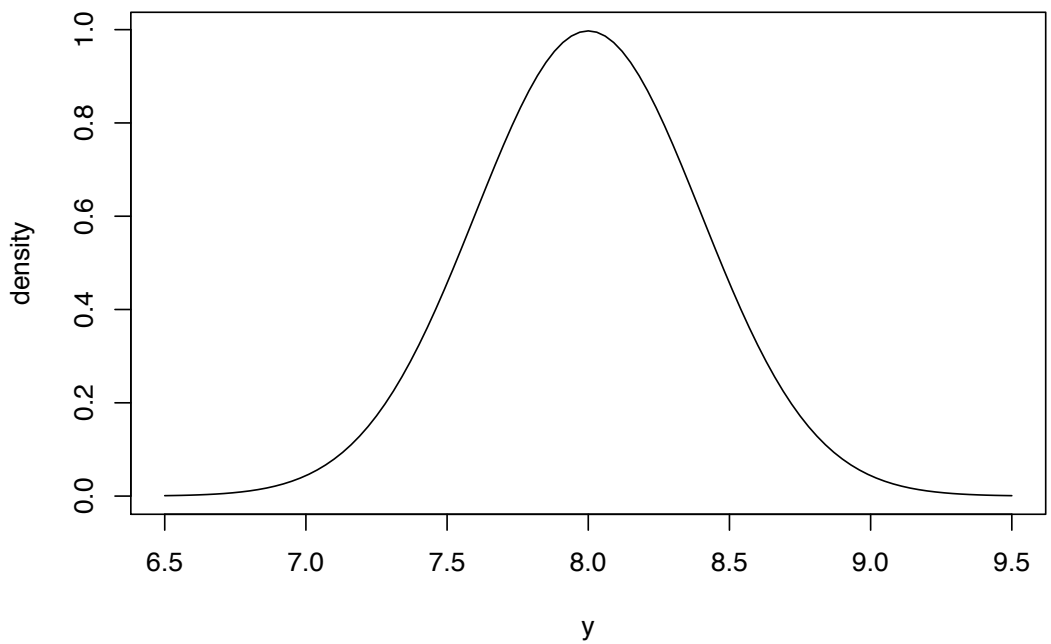
$$\pi(\theta|Y_1,\ldots,Y_n) = \frac{\pi(\theta)\mathcal{L}(\theta|Y_1,\ldots,Y_n)}{\int_\Theta \pi(\theta)\mathcal{L}(\theta|Y_1,\ldots,Y_n)d\theta}.$$

Exemple introductif

Dans notre exemple le calcul de l'estimateur du maximum de vraisemblance donne $\widehat{\theta}^{MV} = \bar{Y}_n$.

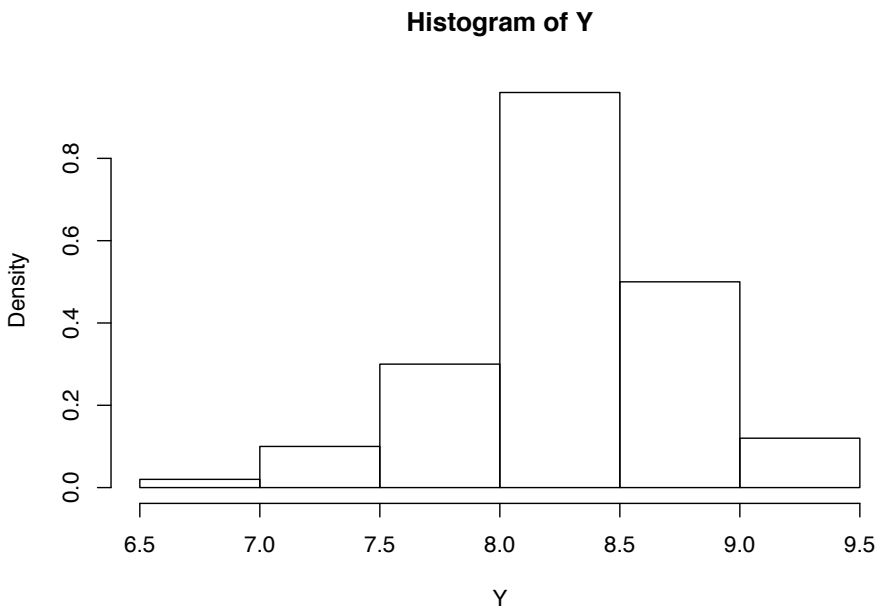
Exemple introductif

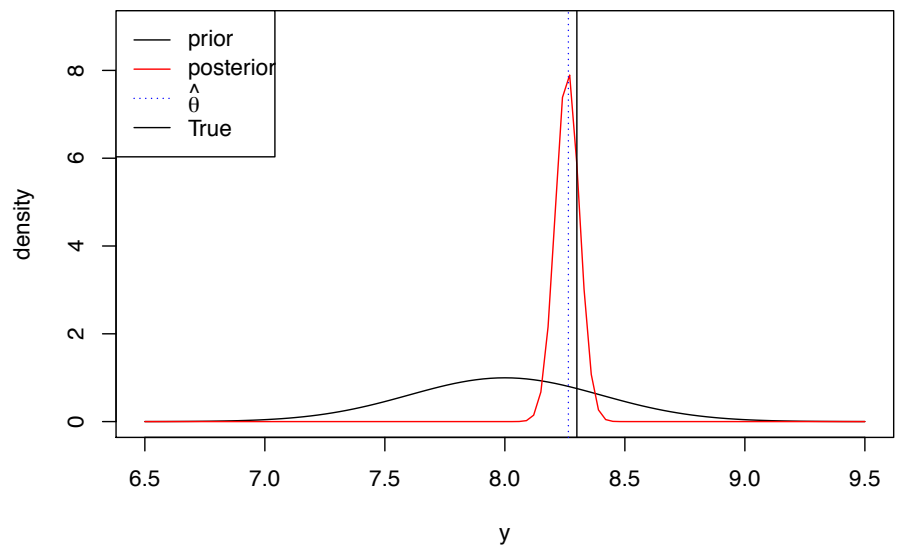
Dans notre exemple le calcul de l'estimateur du maximum de vraisemblance donne $\widehat{\theta}^{MV} = \bar{Y}_n$.
Pour utiliser l'approche Bayésienne on va devoir modéliser notre incertitude sur le paramètre. Connaissant un peu le marché de l'emploi pour les Data Scientists, on se dit que le log-salaire moyen devrait être autour de 8, on modélise notre incertitude par une loi gaussienne



Exemple introductif

On peut trouver la loi a posteriori et on obtient les valeurs suivantes pour $n = 100$





Probabilités subjectives vs. fréquences

- Le log salaire moyen correspond à un événement unique

Probabilités subjectives vs. fréquences

- Le log salaire moyen correspond à un événement unique
- Pas de notion de tirage aléatoire

Probabilités subjectives vs. fréquences

- Le log salaire moyen correspond à un événement unique
- Pas de notion de tirage aléatoire
- Interprétation subjective des probabilités distincte de l'interprétation fréquentiste

Probabilités subjectives vs. fréquences

- Le log salaire moyen correspond à un événement unique
- Pas de notion de tirage aléatoire
- Interprétation subjective des probabilités distincte de l'interprétation fréquentiste
- On peut attribuer une probabilité fréquentiste pour un évenement de type : "le lancer de dé donnera un 1", pas à "BNP-Paribas fera des pertes cette année"

Formule de Bayes

- Les probabilités “subjectives” sont plus générales que les probabilités “fréquentistes”
- Lois de la probabilité comme extension de la logique au domaine de l’incertain, voir le livre de E.T. Jaynes *Probability : The Logic of Science*
- Théorème de Cox : les lois de la probabilité sont les seuls qui combinent des informations de manière raisonnable (vérifient certains axiomes)

Formule de Bayes

Théorème de Bayes

Soient $\mathbb{P}$ un mesure de probabilité et A et B deux événements le théorème de Bayes donne

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

Théorème de Bayes

Soient $\mathbb{P}$ un mesure de probabilité et A et B deux événements le théorème de Bayes donne

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

Le même théorème s'applique pour les *loi de probabilités conditionnelles*

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)f_Y(y)}{f_X(x)}$$

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

θ est le paramètre inconnu du modèle

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

θ est le paramètre inconnu du modèle

$\mathbf{X}^n$ sont les observations

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

θ est le paramètre inconnu du modèle

$\mathbf{X}^n$ sont les observations

$f(\mathbf{X}^n|\theta)$ est la loi des observations sachant le paramètre (vraisemblance)

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

θ est le paramètre inconnu du modèle

$\mathbf{X}^n$ sont les observations

$f(\mathbf{X}^n|\theta)$ est la loi des observations sachant le paramètre (vraisemblance)

$\pi(\theta)$ est la loi a priori sur le paramètre

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

θ	est le paramètre inconnu du modèle
$\mathbf{X}^n$	sont les observations
$f(\mathbf{X}^n\|\theta)$	est la loi des observations sachant le paramètre (vraisemblance)
$\pi(\theta)$	est la loi a priori sur le paramètre
$\pi(\theta\|\mathbf{X}^n)$	est la loi a posteriori

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

- θ est le paramètre inconnu du modèle
- $\mathbf{X}^n$ sont les observations
- $f(\mathbf{X}^n|\theta)$ est la loi des observations sachant le paramètre (vraisemblance)
- $\pi(\theta)$ est la loi a priori sur le paramètre
- $\pi(\theta|\mathbf{X}^n)$ est la loi a posteriori
- $f(\mathbf{X}^n) = \int_\Theta f(\mathbf{X}^n|\theta)\pi(\theta)d\theta$ est la vraisemblance marginale

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

- θ est le paramètre inconnu du modèle
- $\mathbf{X}^n$ sont les observations
- $f(\mathbf{X}^n|\theta)$ est la loi des observations sachant le paramètre (vraisemblance)
- $\pi(\theta)$ est la loi a priori sur le paramètre
- $\pi(\theta|\mathbf{X}^n)$ est la loi a posteriori
- $f(\mathbf{X}^n) = \int_\Theta f(\mathbf{X}^n|\theta)\pi(\theta)d\theta$ est la vraisemblance marginale

Formule de Bayes

Le théorème de Bayes nous donne la loi a posteriori du paramètre :

Loi a posteriori

$$\pi(\theta|\mathbf{X}^n) = \frac{f(\mathbf{X}^n|\theta)\pi(\theta)}{f(\mathbf{X}^n)}$$

θ est le paramètre inconnu du modèle

$\mathbf{X}^n$ sont les observations

$f(\mathbf{X}^n|\theta)$ est la loi des observations sachant le paramètre (vraisemblance)

$\pi(\theta)$ est la loi a priori sur le paramètre

$\pi(\theta|\mathbf{X}^n)$ est la loi a posteriori

$f(\mathbf{X}^n) = \int_\Theta f(\mathbf{X}^n|\theta)\pi(\theta)d\theta$ est la vraisemblance marginale

On peut de plus définir la densité d'une nouvelle observation sachant les précédentes. On appelle cette loi a distribution prédictive

$$f(y|\mathbf{X}^n) = \int_\Theta f(y|\theta)\pi(\theta|\mathbf{X}^n)d\theta$$

Loi a priori et a posteriori

Example détaillé

- On fait passer un questionnaire de 10 questions à une personne.
- Hypothèse : elle a une probabilité θ de répondre correctement à chaque question.
- $\mathbf{y} = [0, 1, 0, 0, 1, 1, 0, 0, 1, 1]$ soit 6 réponses correctes sur 10.
- Comment calculer une distribution a posteriori sur θ ?

- Ici, on procède comme en stat. classique et on postule un modèle
- Par exemple : toutes les réponses sont IID, donc

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{10} p(y_i|\theta) = \prod \theta^{y_i} (1-\theta)^{1-y_i} \\ &= \theta^{\sum y_i} (1-\theta)^{10-\sum y_i} \end{aligned}$$

Principe d'indifférence de Laplace : en l'absence de toute information, toutes les valeurs sont également probables.

$$p(\theta) = 1$$

(Mais on a souvent de l'information, voir plus loin)

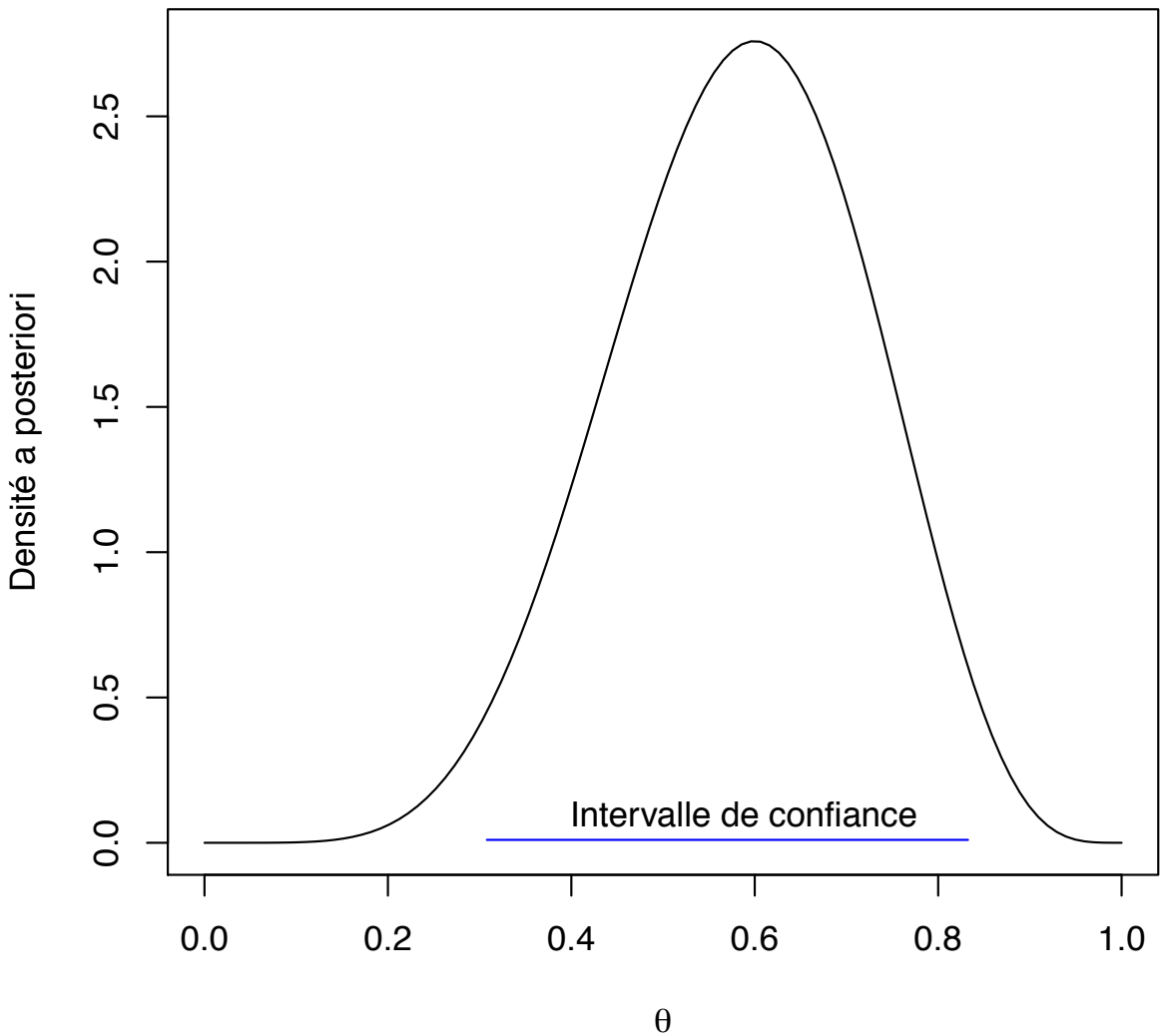
La loi a posteriori s'écrit donc :

$$\begin{aligned} p(\theta|\mathbf{y}) &= \frac{p(\mathbf{y}|\theta)\, p(\theta)}{p(\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\theta)}{p(\mathbf{y})} \\ &\propto \theta^{\sum y_i} (1-\theta)^{10-\sum y_i} \end{aligned}$$

où l'on a utilisé la notation $\propto$ pour "est proportionnel à".

Loi a priori et a posteriori

Loi a posteriori



Loi a priori et a posteriori

Changer la loi a priori

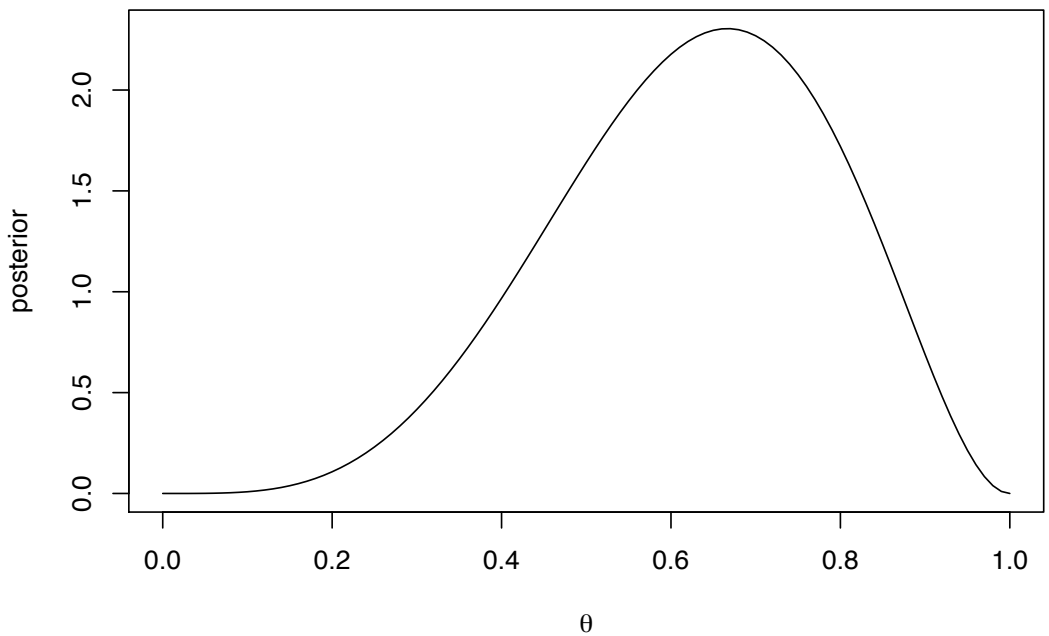
- On peut très souvent faire mieux que la loi uniforme
- Exemple pour un QCM : si les gens ne savent rien, il répondront au hasard
  - Pour un QCM à quatre choix, taux de hasard 25%
  - On peut imaginer quelqu'un qui répond intentionnellement de travers (prob. correct ¡ 25%), mais c'est *a priori* peu probable
  - On peut donc ajuster la loi a priori pour réfléter ce fait.

Loi a priori et a posteriori

Remarque

La loi a priori a un impacte sur la loi a posteriori !

Dans l'exemple précédent si on prend comme loi a priori une $\beta(5, 3)$

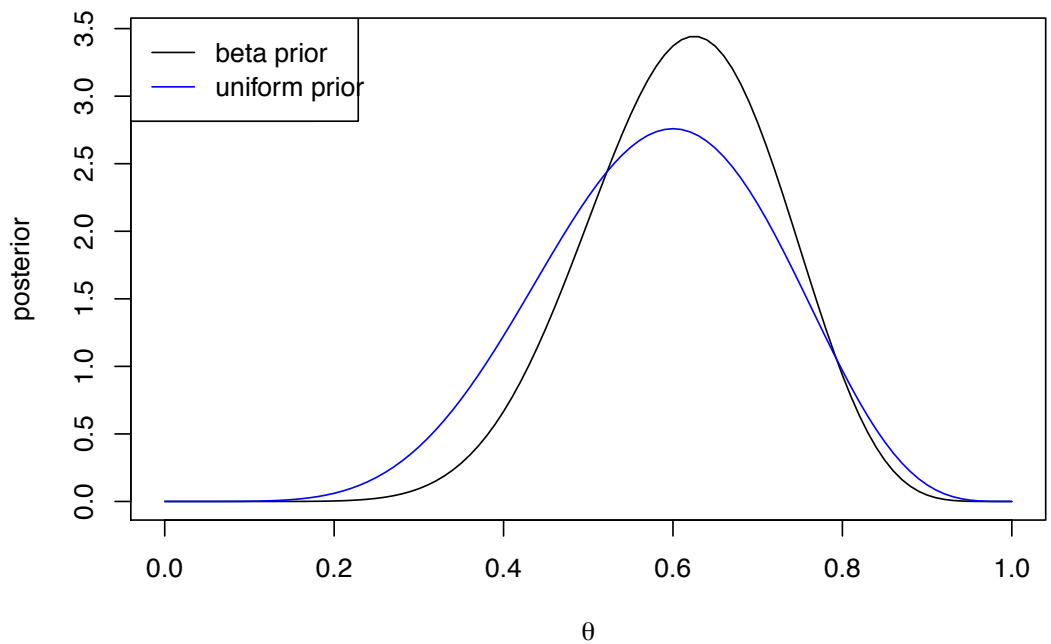


Loi a priori et a posteriori

Remarque

La loi a priori a un impacte sur la loi a posteriori !

Dans l'exemple précédent si on prend comme loi a priori une $\beta(5, 3)$ on obtient



Loi a priori et a posteriori

- En général l'influence de la loi a priori disparaît lorsqu'on ajoute des données.
- La loi a posteriori est une combinaison des information a priori et de l'information apportée par les observations.
- La loi a priori doit refléter notre connaissance du paramètre a priori (études précédentes, connaissance du problème, etc.).
- Le choix de l'a priori est explicité, l'utilisateur de l'étude aura donc toutes les informations.

Choix des lois a priori

Analyse "one-shot"

- Exemple : vous travaillez pour un institut de sondage, et on vous demande d'estimer le positionnement politique des *gendarmes mobiles* (cas réel)
- Paramètre à estimer : proportion de gendarmes mobiles ayant voté pour le candidat XYZ à la dernière présidentielle
- Données : résultats des bureaux de vote situés à proximité des casernes (données relativement parcellaires)
- Objectif de la loi a priori : *combiner un maximum d'informations pertinentes pour ce cas particulier* (ex., votes des militaires en général)

Choix des lois a priori

Analyses répétées (développer une méthode)

- Objectif : produire un logiciel/une méthode à destination d'utilisateurs non-statisticiens, qui permet d'estimer une quantité θ à partir de données $\mathbf{y}$.
- Exemple réel : mesures de seuil (perceptifs, toxicologie)
  - Seuil perceptif = intensité à partir de laquelle un certain stimulus devient détectable
- On va essayer de déterminer une loi a priori qui garantit une bonne performance moyenne
- On est proche d'un raisonnement fréquentiste, et on peut s'aider de mesures réelles de la variabilité de θ dans la population.

Choix des lois a priori

Autres considérations

- Tactique :
  - Quand il s'agit de produire une analyse pour convaincre quelqu'un, évitez de mettre quoi que ce soit dans l'a-priori qui pourrait favoriser vos conclusions (même si c'est parfaitement raisonnable)
- Calculatoire :
  - En pratique les gens adoptent souvent des lois a priori simples parce qu'elles facilitent les calculs, même si elles ne sont pas forcément optimales

Choix des lois a priori

Loi a priori conjuguées

On appelle loi a priori conjuguée une loi telle que l'a priori et l'a posteriori appartiennent à la même famille de lois

Exemple

Voici quelques exemples de lois a priori conjuguées pour des modèles classiques

Model	Prior pour θ
$\mathcal{N}(\theta, \sigma)$	$\mathcal{N}(a, b)$
$\mathcal{B}(n, \theta)$	$\beta(a, b)$
$\Gamma(k, \theta)$	$\Gamma(a, b)$
$\mathcal{N}(m, \theta)$	$I\Gamma(a, b)$

Exercice : Trouver la loi a posteriori pour ces modèles.

Choix des lois a priori

Loi a priori conjuguées

- Les loi a priori conjuguées sont très utiles d'un point de vue pratique car l'a posteriori est très facile a obtenir.
- C'est souvent un choix par défaut assez satisfaisant
- Elle n'existe que pour un nombre assez limités de modèles (les modèles exponentiels)

Choix des lois a priori

Lois impropres

Comme on l'a vu la loi a priori provient de considérations subjectives et il peut arriver qu'on veuille utiliser une mesure σ-finie à la place d'une loi de probabilité.

Exemple

On cherche à estimer la moyenne d'une loi normale $\mathcal{N}(\theta, 1)$. Cependant notre connaissance a priori du problème nous dit que toutes les valeurs possibles pour $\theta \in \mathbb{R}$ sont équiprobables...

Choix des lois a priori

Lois impropres

Comme on l'a vu la loi a priori provient de considérations subjectives et il peut arriver qu'on veuille utiliser une mesure σ-finie à la place d'une loi de probabilité.

Exemple

On cherche à estimer la moyenne d'une loi normale $\mathcal{N}(\theta, 1)$. Cependant notre connaissance a priori du problème nous dit que toutes les valeurs possibles pour $\theta \in \mathbb{R}$ sont équiprobables...

$$\pi(\theta) \propto 1$$

Choix des lois a priori

Lois impropres

Comme on l'a vu la loi a priori provient de considérations subjectives et il peut arriver qu'on veuille utiliser une mesure σ-finie à la place d'une loi de probabilité.

Exemple

On cherche à estimer la moyenne d'une loi normale $\mathcal{N}(\theta, 1)$. Cependant notre connaissance a priori du problème nous dit que toutes les valeurs possibles pour $\theta \in \mathbb{R}$ sont équiprobables...

$$\pi(\theta) \propto 1 \text{ ???}$$

Si π est une mesure σ-finie telle que $\pi(\theta) f(\mathbf{X}^n|\theta)$ est intégrable alors on définira

$$\pi(\theta|\mathbf{X}^n) = \frac{\pi(\theta) f(\mathbf{X}^n|\theta)}{\int_\Theta \pi(\theta) f(\mathbf{X}^n|\theta)}$$

Choix des lois a priori

Lois impropres

Comme on l'a vu la loi a priori provient de considérations subjectives et il peut arriver qu'on veuille utiliser une mesure σ-finie à la place d'une loi de probabilité.

Exemple

On cherche à estimer la moyenne d'une loi normale $\mathcal{N}(\theta, 1)$. Cependant notre connaissance a priori du problème nous dit que toutes les valeurs possibles pour $\theta \in \mathbb{R}$ sont équiprobables...

$$\pi(\theta) \propto 1 \text{ ???}$$

Si π est une mesure σ-finie telle que $\pi(\theta) f(\mathbf{X}^n|\theta)$ est intégrable alors on définira

$$\pi(\theta|\mathbf{X}^n) = \frac{\pi(\theta) f(\mathbf{X}^n|\theta)}{\int_\Theta \pi(\theta) f(\mathbf{X}^n|\theta)}$$

On parlera de loi *a priori impropre*.

Choix des lois a priori

A priori non informatifs

- Il existe des méthodes pour construire des a priori non-informatifs

Choix des lois a priori

A priori non informatifs

- Il existe des méthodes pour construire des a priori non-informatifs
- Ces approches reposent sur des considération statistiques qui dépassent le cadre de cette formation (voir C.P. Robert *Le Choix Bayésien* pour plus de détails)

Choix des lois a priori

A priori non informatifs

- Il existe des méthodes pour construire des a priori non-informatifs
- Ces approches reposent sur des considération statistiques qui dépassent le cadre de cette formation (voir C.P. Robert *Le Choix Bayésien* pour plus de détails)
- L'a priori de Jeffrey est de loin l'a priori le plus courant.

Remarque

Lorsque l'on dispose d'une information a priori sur le paramètre le mieux est de l'utiliser !

Remarque

Lorsque l'on dispose d'une information a priori sur le paramètre le mieux est de l'utiliser !

Exemple (Modèles parcimonieux)

Le cas de la régression en grande dimension avec un vecteur de paramètre sparse est un cas très parlant d'introduction d'information a priori. Ce modèle est très utilisé et on peut rapprocher ces méthodes des approches par maximum de vraisemblance pénalisé type LASSO.

Chap.2

Loi a posteriori

Pourquoi des méthodes computationnelles ?

Introduction aux méthodes MCMC

Mise en œuvre pratique

Pourquoi des méthodes computationnelles ?

Why indeed ?

- La distribution *a posteriori* n'est généralement non calculable (sauf modèles simples ou conjugués)
- Les simulations iid sont aussi difficiles à obtenir
- On fait appelle à des échantillonneurs permettant de générer des chaines de Markov **approchant** la loi.

Pourquoi des méthodes computationnelles ?

Example 1

- Modèle linéaire gaussien avec un prior Student :

$$\begin{aligned} L(y_{1,\cdots,n}|\theta) &= \prod_{i=1}^{n} \phi(y_i; x_i\theta, \sigma^2 I) \\ \pi(\theta) &= \prod_{i=1}^{p} t(\theta_i|0, 10, \nu). \end{aligned}$$

Contrairement au cas du prior gaussien le modèle n'est plus conjugué....

- **But** : Simuler selon

$$\begin{aligned} \pi(\theta|y_{1,\cdots,n}) &\propto L(y_{1,\cdots,n}|\theta)\pi(\theta) \\ &= \prod_{i=1}^{n} \phi(y_i; x_i\theta, \sigma^2 I) \prod_{i=1}^{p} t(\theta_i|0, 10, \nu) \end{aligned}$$

Introduction aux méthodes MCMC

Simulations : une méthode génerique

- **But** : Simuler selon une chaine de Markov avec pour invariant la loi cible $\pi(\theta|y_{1,\cdots,n})$.
- Soit $\theta_1, \cdots, \theta_M$ un M échantillon issu de cette chaine de Markov, sous certaines conditions
  - La loi des grands nombres $\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} f(\theta_i) = \mathbb{E}_\pi \left[f(\theta)|y_{1,\cdots,M} \right]$.
  - Un TCL $\sqrt{M} \left\{ \sum_{i=1}^{M} f(\theta_i) - \mathbb{E}_\pi \left[f(\theta)|y_{1,\cdots,M} \right] \right\} \to \mathcal{N}(0, \psi)$

Introduction aux méthodes MCMC

Simulations : un exemple d'algorithme Metropolis-Hastings

- On définit une loi de proposition : $\theta'_t \sim q(.|\theta_{t-1})$
- On accepte θ'_t (θ_t est défini comme θ'_t) avec probabilité $\alpha(\theta_{t-1}, \theta'_t)$.
- Sinon θ_t est défini comme θ_{t-1}.

Algorithm Metropolis Hastings algorithm

Input : θ_0, M

Output : $(\theta_t)_{t\geq 0}$

For $t \in \{1, \cdots, M\}$

a. Sample $\theta_{\text{prop}} \sim q(.|\theta_{t-1})$.

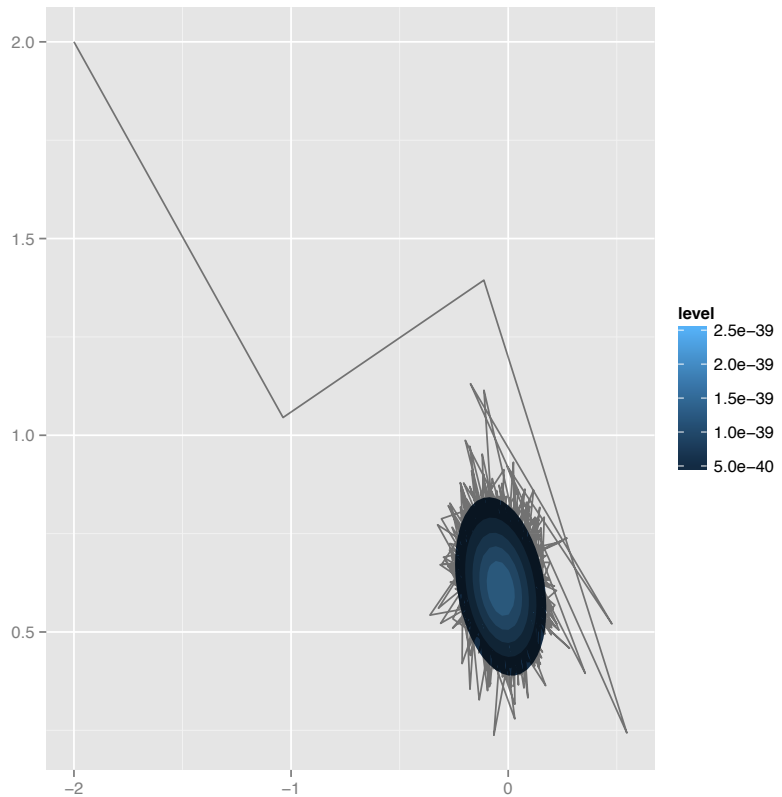
b. Sample $U \sim \mathcal{U}nif$.

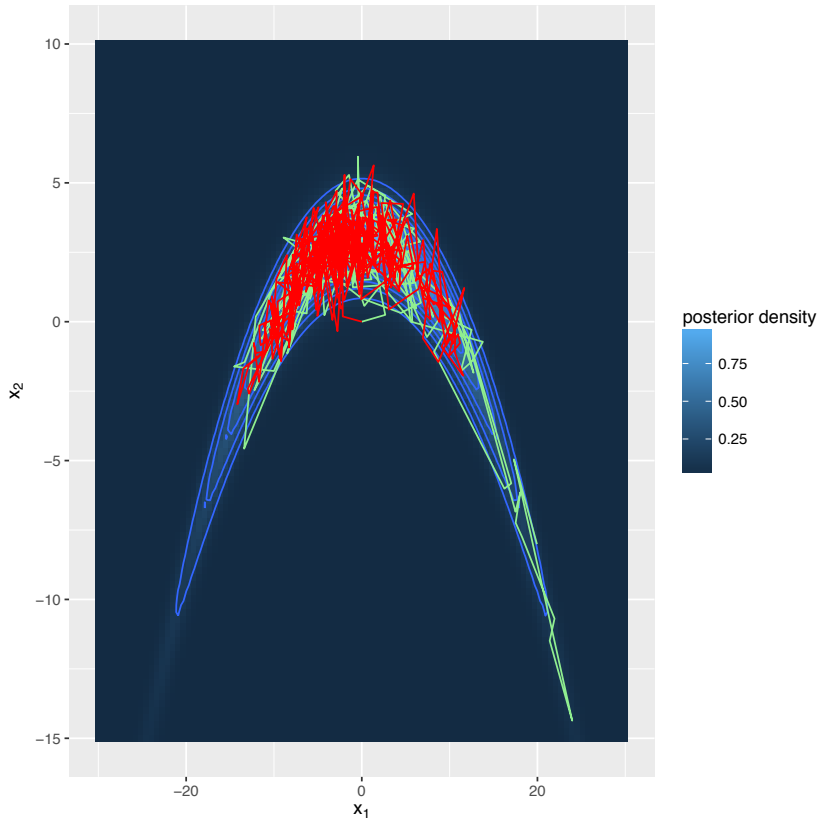
c. If $U \leq \frac{\pi(\theta_{\text{prop}}|y_1,\cdots,y_n)q(\theta_{t-1}|\theta_{\text{prop}})}{\pi(\theta_{t-1}|y_1,\cdots,y_n)q(\theta_{\text{prop}}|\theta_{t-1})}$, set $\theta_t \leftarrow \theta_{\text{prop}}$, otherwise set $\theta_t \leftarrow \theta_{t-1}$.

End For

Introduction aux méthodes MCMC

Simulations : un exemple d'algorithme Metropolis-Hastings





Mise en œuvre pratique

Outils pour échantillonner

1. OpenBUGS
2. STAN
3. MCMCpack
4. Bien d'autres possibilités

- Progressivement remplacé par STAN
- Langage pour l'écriture de modèles hiérarchiques
- Interface R (R2OpenBUGS package)
- Permet de traiter le cas de paramètres discrets

Mise en œuvre pratique

2. Stan : Avantages

- Pas besoin d'implémenter des algorithmes MCMC vous même
- Documentation très complète
- Open source
- Communauté de développeurs et d'utilisateurs très active, évolue rapidement

Mise en œuvre pratique

Stan : inconvénients

- Black box (on ne sait pas forcément toujours très bien ce qui se passe)
- Permet d'attaquer un grand nombre de problèmes, mais pas tous
  - Tous les paramètres doivent être des valeurs continues (il existe des moyens de contourner ce problème pour les utilisateurs *très* avancés)
  - La loi a posteriori doit être continue et dérivable
  - On aura potentiellement des soucis avec les lois multimodales (mais il n'y a pas de solutions simples à ce problème)

Mise en œuvre pratique

Comment ça marche : l'échantillonneur

- L'échantillonneur est un black box à qui on donne en entrée une fonction $\mathcal{L}(\theta)$ et sa dérivée $\frac{\partial}{\partial\theta}\mathcal{L}$, et qui génère (asymptotiquement) des échantillons de la loi

$$\pi(\theta) = \exp(L(\theta))$$

- L'algorithme utilise les infos du gradient pour accélérer l'échantillonnage
- Monte Carlo Hamiltonien avec ajustement automatique (NUTS, No U-Turn Sampler)

Mise en œuvre pratique

Comment ça marche : la définition du modèle

- Vous écrivez un programme qui décrit le modèle statistique
  - Quelles sont les données ?
  - Quels sont les paramètres ?
  - Quels sont les a priori sur les paramètres ? (i.e., $p(\theta)$)
  - Quelle est la vraisemblance ? ($p(\mathbf{y}|\theta)$)

Mise en œuvre pratique

Comment ça marche : interprétation, compilation

- Interprétation : stan lit le modèle, le transforme en fonctions C++ qui calculent :
  - Le log de la loi a posteriori $\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) + \log p(\theta)$
  - $\frac{\partial}{\partial \theta}\mathcal{L}$, sa dérivée (par différentiation automatique)
- Stan compile les fonctions C++ (cette étape prend du temps)
- On peut maintenant utiliser l'échantillonneur

Mise en œuvre pratique

3. MCMCpack : Avantages

- Package R
- Plus de liberté dans la définition du modèle
- Open source
- Fonctions prédéfinies pour certains modèles

Mise en œuvre pratique

Stan Exemples : syntaxe

```
data {
real<lower=0> sd_prior; //Contrôle l'écart type du prior

}
parameters {
real beta; //Coefficient
}
model {
beta ~ normal(0,sd_prior); //A priori sur beta
```

```
data {
real<lower=0> sd_prior;
real y;

}
parameters {
real beta;
}
model {
beta ~ normal(0,sd_prior);
y ~ normal(beta,1); //Vraissemblance: les données sont Gaussiennes centrées sur beta

}
```

Mise en œuvre pratique

Rajouter *des* observations

```
data {
real<lower=0> sd_prior;
int<lower=1> n; //Nombre d'observations

vector[n] y; //Observations

}
parameters {
real beta; //Coefficient
}
model {
beta ~ normal(0,sd_prior);
y ~ normal(beta,1); //Vraisemblance y|beta (vectorisée)

}
```

Mise en œuvre pratique

Régression bayésienne

- On va regarder un modèle de régression simple :
- Modèle :

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ \epsilon &\sim \mathbb{N}\left(0, \sigma^2\right) \\ \beta_i &\sim \mathbb{N}\left(0, \tau^2\right) \\ \sigma &\sim \Gamma(2,2) \end{aligned}$$

- La matrice $\mathbf{X}_{n\times m}$ contient les covariées (design matrix).
- On a m coefficients et n observations $\mathbf{y}$

Mise en œuvre pratique

Régression bayésienne (programme Stan)

```
data {
int<lower=1> n; //Nombre d'observations
int<lower=1> m; //Nombres de covariées
matrix[n,m] X; //Matrice de régression
real<lower=0> sd_prior; //Ecart type a priori des coeffs. de régression

vector[n] y; //Observations
}
parameters {
vector[m] beta; //Coefficients
real<lower=0> sd_noise; //Ecart-type du bruit
}
model {
sd_noise ~ gamma(2,2); //Loi a priori sur l'écart type
beta ~ normal(0,sd_prior); //Loi a priori sur les coefficients
y ~ normal(X*beta,sd_noise); //Vraissemblance
}
```

Mise en œuvre pratique

Qu'est-ce qui sort de Stan ?

- Par défaut, Stan lance quatre chaînes MCMC initialisées à des endroits différents, avec 1000 itérations "burn-in" et 1000 itérations conservées
- Si les chaînes se comportent bien, les résultats devraient dans toutes les chaînes être semblables
- Toujours vérifier que les chaînes ne sont pas trop autocorrélées !

Mise en œuvre pratique

MCMCpack : syntaxe

- Fonction : *MCMCmetrop1R* :

```
MCMCmetrop1R(fun, theta.init, burnin = 500, mcmc = 20000, thin = 1,....)
#fun est definie par l'utilisateur c'est la loi cible
# theta.init le point in initial
```

Chap.3

Méthodes d'estimation

Rappel de théorie de la décision

Comparaison des estimateurs Bayésiens et fréquentistes

Régions de crédibilité

Rappel de théorie de la décision

Perte et Risque I

Lorsqu'on dispose d'un estimateur $\hat{\theta}$, on peut chercher à savoir s'il est *bon*.

Définition (Perte)

On appelle fonction de perte toute fonction $L : \Theta \times \Theta \to \mathbb{R}^+$ telle que $L(\theta, \theta) = 0$

Pour θ le vrais paramètre ayant généré les données, la qualité de l'estimateur est mesurée par $L(\theta, \hat{\theta})$.

Rappel de théorie de la décision

Perte et Risque II

Comme $\hat{\theta} = \hat{\theta}(\mathbf{X}^n)$ on peut considérer la perte moyenne pour un estimateur

Définition (Risque)

La fonction de risque d'un estimateur $\hat{\theta}$ est l'espérance sous θ de la perte

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbb{E}_\theta^n L(\theta, \hat{\theta}(\mathbf{X}^n)).$$

- On peut définir une relation d'ordre partiel sur les estimateurs
- On dira qu'un estimateur $\hat{\theta}$ est *inadmissible* si il existe un estimateur $\tilde{\theta}$ tel que $\forall \theta$, $\mathcal{R}(\theta, \tilde{\theta}) \leq \mathcal{R}(\theta, \hat{\theta})$ et $\exists \theta_0$ tel que $\mathcal{R}(\theta_0, \tilde{\theta}) < \mathcal{R}(\theta_0, \hat{\theta})$

Rappel de théorie de la décision

Minimisation du risque I

On va chercher à minimiser le risque, *mais le risque dépend de θ...*

Approche minimax Trouver l'estimateur qui a le plus petit risque pour le pire θ

Approche Bayésienne On dispose d'une mesure de probabilité de chaque θ (la loi a priori), on va chercher à minimiser le risque moyen pour cette loi sur θ.

On va donc chercher à minimiser

$$\min_{\hat{\theta}} \mathbb{E}^{\pi}(\mathcal{R}(\theta, \hat{\theta})) = \min_{\hat{\theta}} \int_{\Theta} \mathcal{R}(\theta, \hat{\theta}) \pi(\theta) d\theta$$

Rappel de théorie de la décision

Minimisation du risque II

Mais $\mathcal{R}(\theta, \hat{\theta})$ est aussi une espérance !

$$\begin{aligned}
\mathbb{E}^{\pi}(\mathcal{R}(\theta, \hat{\theta})) &= \int_{\Theta} \mathcal{R}(\theta, \hat{\theta}) \pi(\theta) d\theta \\
&= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \hat{\theta}(X)) f_{\theta}(\mathbf{X}^n) d\mathbf{X}^n \pi(\theta) d\theta \\
&= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \hat{\theta}(\mathbf{X}^n)) \underbrace{\pi(\theta|\mathbf{X}^n)}_{\text{Posterior}} d\theta m(\mathbf{X}^n) d\mathbf{X}^n
\end{aligned}$$

Il nous suffit de minimiser $\int_{\Theta} L(\theta, \hat{\theta}(\mathbf{X}^n)) \pi(\theta|\mathbf{X}^n) d\theta$ pour tout $\mathbf{X}^n$ et on a notre estimateur Bayésien

- $L(\theta, \mu) = (\theta - \mu)^2$ L'estimateur bayésien est la moyenne a posteriori.
- $L(\theta, \mu) = |\theta - \mu|$ l'estimateur bayésien associé est la médiane a posteriori
- $L(\theta, \mu) = \mathbb{I}_{\theta=\mu}$ l'estimateur bayésien associé est le mode a posteriori

Comparaison des estimateurs Bayésiens et fréquentistes

En général, pour les modèles simples, les estimateurs Bayésiens et fréquentistes seront proches. Pour des modèles réguliers on a même le résultat suivant

Théorème (Berstein-von Mises)

Sous des conditions idoines

$$\|\Pi(\cdot|\mathbf{X}^n) - \mathcal{N}(\hat{\theta}, I_n^{-1}(\theta))\|_{TV} \to 0$$

Comparaison des estimateurs Bayésiens et fréquentistes

En général, pour les modèles simples, les estimateurs Bayésiens et fréquentistes seront proches. Pour des modèles réguliers on a même le résultat suivant

Théorème (Berstein-von Mises)

Sous des conditions idoines

$$\|\Pi(\cdot|\mathbf{X}^n) - \mathcal{N}(\hat{\theta}, I_n^{-1}(\theta))\|_{TV} \to 0$$

Ce résultat dit qu'asymptotiquement les approches fréquentistes et bayésiennes donnent les mêmes résultats.

Régions de crédibilité

Région de confiance

Définition

On appelle région de confiance de niveau α un ensemble C tel que

$$\mathbb{P}_\theta(C \ni \theta) \geq 1 - \alpha$$

Régions de crédibilité

Région de confiance

Définition

On appelle région de confiance de niveau α un ensemble C tel que

$$\mathbb{P}_\theta(C \ni \theta) \geq 1 - \alpha$$

- $\mathbb{P}_\theta$ est une probabilité sur $\mathbf{X}^n$ et non sur θ
- On ne va s'intéresser qu'aux ensemble C petits

Régions de crédibilité

Région α-Crédible

Lorsque l'on dispose d'une loi sur θ on peut chercher l'ensemble de paramètres $\mathcal{C}$ tels que

$$\Pi(\theta \in \mathcal{C} | \mathbf{X}^n) \geq 1 - \alpha$$

On parle alors de région α crédible.

Régions de crédibilité

Région α-Crédible

Lorsque l'on dispose d'une loi sur θ on peut chercher l'ensemble de paramètres $\mathcal{C}$ tels que

$$\Pi(\theta \in \mathcal{C} | \mathbf{X}^n) \geq 1 - \alpha$$

On parle alors de région α crédible.

- C'est bien une probabilité sur θ cette fois
- Il existe plein de méthodes pour construire de telles régions
- On peut les approcher par des méthodes de Monte-Carlo

Méthode simple pour construire des régions α-crédibles – Credible ball

Si l'on dispose d'un estimateur Bayésien $\hat{\theta}$ et d'une distance d sur Θ on peut chercher la plus petite boule $B(\hat{\theta}, c)$ centrée en $\hat{\theta}$ et de rayon c telle que $\Pi(B(\hat{\theta}, c)) \geq 1 - \alpha$.

Régions de crédibilité

HPD Région

Si la méthode précédente est facile pour créer des région de crédibilité, celles-ci ne seront souvent pas les *meilleures possibles*.

Régions de crédibilité

HPD Région

Si la méthode précédente est facile pour créer des région de crédibilité, celles-ci ne seront souvent pas les *meilleures possibles*.

- On peut chercher la région la plus petite contenant une masse d'au moins $1 - \alpha$.

Régions de crédibilité

HPD Région

Si la méthode précédente est facile pour créer des région de crédibilité, celles-ci ne seront souvent pas les *meilleures possibles*.

- On peut chercher la région la plus petite contenant une masse d'au moins $1-\alpha$.

La solution à ce problème donne

$$\mathcal{C}_\alpha = \{\theta, \pi(\theta|\mathbf{X}^n) > k_\alpha\}, \ k_\alpha \text{ tel que } \pi(\mathcal{C}_\alpha|\mathbf{X}^n) = 1-\alpha$$

Régions de crédibilité
HPD Région

Si la méthode précédente est facile pour créer des région de crédibilité, celles-ci ne seront souvent pas les *meilleures possibles*.

- On peut chercher la région la plus petite contenant une masse d'au moins $1-\alpha$.

La solution à ce problème donne

$$\mathcal{C}_\alpha = \{\theta, \pi(\theta|\mathbf{X}^n) > k_\alpha\}, \ k_\alpha \text{ tel que } \pi(\mathcal{C}_\alpha|\mathbf{X}^n) = 1-\alpha$$

- Trouver k_α peut être très compliqué en pratique, surtout lorsque l'on s'écarte des modèles conjugués.

Chap.4

Modèles de régression

Régression linéaire

Modèles linéaires généralisés

Soient $\mathbf{X}$ la matrice des co-variables et Y le vecteur des outputs, on rappelle le modèle de régression

$$Y = \mathbf{X}\beta + \sigma\epsilon, \ \epsilon \sim \mathcal{N}(0, I_n)$$

Pour chaque ligne on a donc $Y_i = \sum_{j=1}^{p} x_i^j \beta_j + \sigma\epsilon_i$. On va chercher à estimer β (et σ en fonction des cas).

Régression linéaire

Modèle Bayésien

Pour avoir un modèle bayésien, il faut choisir une loi a priori sur les paramètres β et σ.

- $\beta \in \mathbb{R}^p$, sauf indication contraire, on va chercher une loi sur continue sur $\mathbb{R}^p$
- $\sigma \in \mathbb{R}^+$
- On fait en général l'hypothèse que σ et β sont indépendants

Régression linéaire

A priori conjugué

Si σ est connu, on a simplement un modèle Gaussien, on peut donc choisir simplement un a priori Gaussien qui devrait être conjugué.

A priori Gaussien sur β

Si $\beta \sim \mathcal{N}(\mu, \sigma^2 S)$ alors a posteriori

$$\beta|\mathbf{X}^n \sim \mathcal{N}\left(\hat{\mu}, \sigma^2 \hat{S}\right),$$

$$\hat{\mu} = \left(X'X + S^{-1}\right)^{-1}\left(X'Y + S^{-1}\mu\right)$$

$$\hat{S} = \left(X'X + S^{-1}\right)^{-1}$$

Régression linéaire

A priori conjugué

Si σ est connu, on a simplement un modèle Gaussien, on peut donc choisir simplement un a priori Gaussien qui devrait être conjugué.

A priori Gaussien sur β

Si $\beta \sim \mathcal{N}(\mu, \sigma^2 S)$ alors a posteriori

$$\beta|\mathbf{X}^n \sim \mathcal{N}\left(\hat{\mu}, \sigma^2 \hat{S}\right),$$
$$\hat{\mu} = \left(X'X + S^{-1}\right)^{-1}(X'Y + S^{-1}\mu)$$
$$\hat{S} = \left(X'X + S^{-1}\right)^{-1}$$

Si σ est inconnu on peut vérifier qu'un a priori *Inverse Gamma* $IG(a, b)$ *sur* σ^2 est conjugué pour ce modèle, et l'a posteriori est une loi $IG(a_n, b_n)$ avec

$$a_n = a + n/2, \;\; b_n = b + \frac{1}{2}(Y'Y + \mu' S^{-1}\mu - \hat{\mu}'\hat{S}^{-1}\hat{\mu})$$

L'a priori conjugué n'est pas forcément simple à utiliser

- Difficile d'avoir de l'information a priori sur la structure de covariance S
- Posterior assez sensible au choix d'hyper-paramètres (notamment S)

Zellner Prior

L'idée est de prendre un choix par défaut $S = g(X'X)^{-1}$, l'a posteriori est donc

$$\beta|\sigma^2, \mathbf{X}^n \sim \mathcal{N}\left(\frac{g}{g+1}(\hat{\beta}^{OLS} + \mu/g), \frac{\sigma^2 g}{g+1}(X'X)^{-1}\right)$$

Régression linéaire

Zellner G-Prior

L'a priori conjugué n'est pas forcément simple à utiliser

- Difficile d'avoir de l'information a priori sur la structure de covariance S
- Posterior assez sensible au choix d'hyper-paramètres (notamment S)

Zellner Prior

L'idée est de prendre un choix par défaut $S = g(X'X)^{-1}$, l'a posteriori est donc

$$\beta|\sigma^2, \mathbf{X}^n \sim \mathcal{N}\left(\frac{g}{g+1}(\hat{\beta}^{OLS} + \mu/g), \frac{\sigma^2 g}{g+1}(X'X)^{-1}\right)$$

L'a priori de Zellner est un a priori impropre sur la variance $\pi(\sigma^2) \propto \sigma^{-2}$

Modèles linéaires généralisés

Retour sur les GLM

L'idée du modèle de régression est la suivante, on a $Y_i \sim f(X_i)$, et on modélise en général $\mathbb{E}(Y|X_i) = g(\theta' X_i)$.

Modèles linéaires généralisés

Retour sur les GLM

L'idée du modèle de régression est la suivante, on a $Y_i \sim f(X_i)$, et on modélise en général $\mathbb{E}(Y|X_i) = g(\theta' X_i)$.

- g est appelé la fonction de lien

Modèles linéaires généralisés

Retour sur les GLM

L'idée du modèle de régression est la suivante, on a $Y_i \sim f(X_i)$, et on modélise en général $\mathbb{E}(Y|X_i) = g(\theta' X_i)$.

- g est appelé la fonction de lien
- Elle sert principalement à prendre en compte les contraintes sur les paramètres de la loi f.

Modèles linéaires généralisés

Retour sur les GLM

L'idée du modèle de régression est la suivante, on a $Y_i \sim f(X_i)$, et on modélise en général $\mathbb{E}(Y|X_i) = g(\theta' X_i)$.

- g est appelé la fonction de lien
- Elle sert principalement à prendre en compte les contraintes sur les paramètres de la loi f.

Modèles linéaires généralisés

Retour sur les GLM

L'idée du modèle de régression est la suivante, on a $Y_i \sim f(X_i)$, et on modélise en général $\mathbb{E}(Y|X_i) = g(\theta' X_i)$.

- g est appelé la fonction de lien
- Elle sert principalement à prendre en compte les contraintes sur les paramètres de la loi f.

Exemple (Régression logistique)

$Y_i \sim \mathcal{B}(1, g(\theta' X_i))$ où $g(x) = e^x/(1+e^x)$ la fonction logistique.

Modèles linéaires généralisés

Retour sur les GLM

L'idée du modèle de régression est la suivante, on a $Y_i \sim f(X_i)$, et on modélise en général $\mathbb{E}(Y|X_i) = g(\theta' X_i)$.

- g est appelé la fonction de lien
- Elle sert principalement à prendre en compte les contraintes sur les paramètres de la loi f.

Exemple (Régression logistique)

$Y_i \sim \mathcal{B}(1, g(\theta' X_i))$ où $g(x) = e^x/(1+e^x)$ la fonction logistique.

Exemple (Régression Poisson)

$Y_i \sim \mathcal{P}(g(\theta' X_i))$ et $g(x) = e^x$

Chap.5

Régularisation et Conparaison de modèles

Comparaison de modèles

Régularisation pour les modèles de régression

Comparaison de modèles

Retour sur les tests

On défini un problème de test par 2 hypothèses/modèles en compétition :

- Une hypothèse nulle H_0
- Une hypothèse alternative H_1

Le but va être de *déterminer quelle hypothèse est la plus compatible avec les données*

Comparaison de modèles

Retour sur les tests

On défini un problème de test par 2 hypothèses/modèles en compétition :

- Une hypothèse nulle H_0
- Une hypothèse alternative H_1

Le but va être de *déterminer quelle hypothèse est la plus compatible avec les données*

Exemple

On dispose d'un n échantillon $X_1, \ldots, X_n$ iid de loi de Bernoulli de paramètre p, on veut déterminer si

$$H_0 : p \leq 1/2, \text{ versus } H_1 : p > 1/2$$

Comparaison de modèles

Test Bayésien

Dans un cadre Bayésien, comparer des hypothèses revient à comparer leur probabilités à posteriori.

Comparaison de modèles

Test Bayésien

Dans un cadre Bayésien, comparer des hypothèses revient à comparer leur probabilités à posteriori.

Fonction de perte 0-1

On considère le test générique suivant $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ et la fonction de perte

$$L(\theta, \delta) = \begin{cases} 1 & \text{si } \mathbb{I}_{\Theta_1}(\theta) \neq \delta \\ 0 & \text{sinon} \end{cases}$$

L'estimateur Bayésien de δ pour ce problème est

$$\delta^\pi(\mathbf{X}^n) = 1 \iff \Pi(\Theta_1|\mathbf{X}^n) > \Pi(\Theta_0|\mathbf{X}^n)$$

On remarque que l'a priori doit mettre un poids non nulle sur chacune des hypothèses...

Cas d'une hypothèse simple $\theta = \theta_0$

Dans le cas d'une hypothèse simple $\Theta_0 = \{\theta_0\}$, on va considérer une loi a priori de la forme

$$\Pi(\theta) = \alpha\delta_{\theta_0} + (1-\alpha)\Pi_1(\theta)\mathbb{I}_{\theta\neq\theta_0}.$$

Π_1 est n'importe quelle loi a priori sur Θ_1.

On remarque que l'a priori doit mettre un poids non nulle sur chacune des hypothèses...

Cas d'une hypothèse simple $\theta = \theta_0$

Dans le cas d'une hypothèse simple $\Theta_0 = \{\theta_0\}$, on va considérer une loi a priori de la forme

$$\Pi(\theta) = \alpha\delta_{\theta_0} + (1-\alpha)\Pi_1(\theta)\mathbb{I}_{\theta\neq\theta_0}.$$

Π_1 est n'importe quelle loi a priori sur Θ_1.

Remarque

Cela revient à considérer l'index du modèle comme un paramètre à estimer. Dans le cas précédent, on a $\{\mathfrak{M} = 0\} = \{\theta = \theta_0\}$ *et* $\{\mathfrak{M} = 1\} = \{\theta \neq 0\}$ *avec les probabilités a priori*

$$\Pi(\mathfrak{M} = 0) = \alpha = 1 - \Pi(\mathfrak{M} = 1)$$

Comparaison de modèles

Facteur de Bayes

Pour comparer de modèles, on peut comparer les *odds ratio*

$$B_{0,1} = \frac{\Pi(\mathfrak{M} = 0|\mathbf{X}^n)/\Pi(\mathfrak{M} = 0)}{\Pi(\mathfrak{M} = 1|\mathbf{X}^n)/\Pi(\mathfrak{M} = 1)}.$$

Comparaison de modèles

Facteur de Bayes

Pour comparer de modèles, on peut comparer les *odds ratio*

$$B_{0,1} = \frac{\Pi(\mathfrak{M} = 0|\mathbf{X}^n)/\Pi(\mathfrak{M} = 0)}{\Pi(\mathfrak{M} = 1|\mathbf{X}^n)/\Pi(\mathfrak{M} = 1)}.$$

On appelle cette quantité le Facteur de Bayes qui se comporte comme un rapport de vraisemblance. C'est une mesure de la confiance que qu'on peut avoir dans le choix de modèle.

Comparaison de modèles

Facteur de Bayes

Pour comparer de modèles, on peut comparer les *odds ratio*

$$B_{0,1} = \frac{\Pi(\mathfrak{M} = 0|\mathbf{X}^n)/\Pi(\mathfrak{M} = 0)}{\Pi(\mathfrak{M} = 1|\mathbf{X}^n)/\Pi(\mathfrak{M} = 1)}.$$

On appelle cette quantité le Facteur de Bayes qui se comporte comme un rapport de vraisemblance. C'est une mesure de la confiance que qu'on peut avoir dans le choix de modèle.

Échelle de Jeffrey

- $\log(B_{0,1}) \in [0, 0.5]$ Une confiance faible en faveur du modèle 0
- $\log(B_{0,1}) \in [0.5, 1]$ Une confiance substantielle en faveur du modèle 0
- $\log(B_{0,1}) \in [1, 2]$ Une confiance forte en faveur du modèle 0
- $\log(B_{0,1}) > 2$ Une confiance décisive en faveur du modèle 0

Régularisation pour les modèles de régression

Choix de variables

Un problème classique de choix de modèle est la sélection de co-variables pour le modèle de régression. Dans ce cas, on ne va pas comparer deux modèles mais potentiellement 2^p modèles !

Approche

Soit $\gamma \in \{0,1\}^p$ un index de modèle (par exemple $\gamma = (1,0,0,\ldots,0,1,0)$). On peut définir β^γ comme le sous vecteur de β avec uniquement les composantes actives de γ, idem pour $\mathbf{X}^\gamma$.

Régularisation pour les modèles de régression

Choix de variables

Un problème classique de choix de modèle est la sélection de co-variables pour le modèle de régression. Dans ce cas, on ne va pas comparer deux modèles mais potentiellement 2^p modèles !

Approche

Soit $\gamma \in \{0,1\}^p$ un index de modèle (par exemple $\gamma = (1,0,0,\ldots,0,1,0)$). On peut définir β^γ comme le sous vecteur de β avec uniquement les composantes actives de γ, idem pour $\mathbf{X}^\gamma$.
On va traiter γ comme un paramètre et trouver le modèle le plus probable.

Régularisation pour les modèles de régression

A priori pour γ

On a 2^p modèles à comparer, on va donc utiliser un a priori conjugué pour faciliter les calculs. On choisi l'a priori de Zellner

$$\begin{aligned} \gamma &\sim \pi_\gamma \\ \pi_{\alpha,\sigma^2}(\alpha,\sigma^2) &\propto \sigma^{-2} \\ \beta^\gamma|\sigma^2,\gamma &\sim \mathcal{N}\left(\bar{\beta}^\gamma, g\sigma^2\left(\mathbf{X}^{\gamma\prime}\mathbf{X}^\gamma\right)^{-1}\right) \end{aligned}$$

où $\bar{\beta}^\gamma = \left(\mathbf{X}^{\gamma\prime}\mathbf{X}\right)^{-1}\mathbf{X}^{\gamma\prime}\mathbf{y}$.

Régularisation pour les modèles de régression

A priori pour γ

On a 2^p modèles à comparer, on va donc utiliser un a priori conjugué pour faciliter les calculs. On choisi l'a priori de Zellner

$$\begin{aligned}\gamma &\sim \pi_\gamma \\ \pi_{\alpha,\sigma^2}(\alpha,\sigma^2) &\propto \sigma^{-2} \\ \beta^\gamma|\sigma^2,\gamma &\sim \mathcal{N}\left(\bar{\beta}^\gamma, g\sigma^2\left(\mathbf{X}^{\gamma\prime}\mathbf{X}^\gamma\right)^{-1}\right)\end{aligned}$$

où $\bar{\beta}^\gamma = \left(\mathbf{X}^{\gamma\prime}\mathbf{X}\right)^{-1}\mathbf{X}^{\gamma\prime}\mathbf{y}$. Pour π_γ on peut choisir un a priori uniforme $\pi(\gamma) = 2^{-p}$ ou un a priori qui pénalise la complexité. On a

$$\pi(\gamma|\mathbf{X}^n) \propto$$

$$(g+1)^{-(p_\gamma+1)/2}\left[\mathbf{y}'\mathbf{y} - \frac{g}{g+1}\mathbf{y}'\mathbf{X}^{\gamma\prime}\left(\mathbf{X}^{\gamma\prime}\mathbf{X}^\gamma\right)^{-1}\mathbf{X}^{\gamma\prime}\mathbf{y} - \frac{1}{g+1}\bar{\beta}^{\gamma\prime}\mathbf{X}^{\gamma\prime}\mathbf{X}^\gamma\bar{\beta}^\gamma\right]^{(n-1)/2}\pi(\gamma)$$

avec $p_\gamma = \sum_{i=1}^p \gamma_i$.

Le calcul précédent est possible uniquement quand p est de taille raisonnable (disons ¡ 15). Dans d'autre cas on peut essayer de simuler sous la loi a posteriori des γ, en particulier pour l'a priori de Zellner on peut utiliser un algorithme de Gibbs. Pour l'échantillonneur de Gibbs, on remarquera que $\pi(\gamma_i|\mathbf{X}^n, \gamma_{-i}) \propto \pi(\gamma|\mathbf{X}^n)$.

Le calcul précédent est possible uniquement quand p est de taille raisonnable (disons ¡ 15). Dans d'autre cas on peut essayer de simuler sous la loi a posteriori des γ, en particulier pour l'a priori de Zellner on peut utiliser un algorithme de Gibbs. Pour l'échantillonneur de Gibbs, on remarquera que $\pi(\gamma_i|\mathbf{X}^n, \gamma_{-i}) \propto \pi(\gamma|\mathbf{X}^n)$.

Remarque

Une fois l'échantillon de γ obtenu, on peut approximer la probabilité d'inclusion de chaque variable par

$$\Pi(\gamma_i = 1|\mathbf{X}^n) \approx \frac{1}{N}\sum_{t=1}^{N} \gamma_i^{(t)}$$