MA201 - Séance 2
Estimation Bayésienne

H. Piet-Lahanier - L. Meyer

Estimateur

Estimateurs

Approche fréquentiste : Hypothèses

- $(X_1, ...X_n)$ un n-échantillon suit une loi *connue* de paramètre $\theta \in \Theta \subset \mathbb{R}^p$ *inconnu*.
- Le paramètre θ est une valeur *déterministe*
- L'estimateur fréquentiste fournit un paramètre estimé $\widehat{\theta}$ qui est une variable aléatoire
- Caractéristiques dépendent de l'estimateur (Maximum de Vraisemblance, Moments, Risque quadratique)

Estimateur

Estimateurs

Approche Bayésienne : Hypothèses

- $(X_1, ...X_n)$ un n-échantillon suit une loi *connue* de paramètre $\theta \in \Theta \subset \mathbb{R}^p$ *inconnu*.
- Le paramètre θ est également une variable aléatoire associée à une distribution *a priori* $\pi(\theta)$
- L'estimateur Bayésien fournit une variable estimée $\widehat{\theta}$ qui est également une variable aléatoire
- Caractéristiques dépendent de l'estimateur et de la distribution a priori

Estimateur Bayésien

Déterministe vs bayésienne

- Non Bayésien : Besoin d'un rapport minimal entre nombre des observations et dimension du vecteur estimé
- Bayésien : dimension du vecteur estimé peut dépasser le nombre d'observations

Compensation de la déficience des observations par des hypothèses restrictives sur les valeurs à estimer
Représentation du vecteur θ comme un vecteur aléatoire de loi connue *a priori*

Hypothèses

Loi *a priori*, loi jointe, loi a posteriori

La valeur recherchée θ est un vecteur aléatoire de loi *a priori* $\pi(\boldsymbol{\theta})$.
Étant donné l'échantillon de données $\boldsymbol{z}$, on cherche à caractériser la loi *a posteriori* de θ, qui est la loi de θ conditionnelle à $\boldsymbol{z}$ qui s'obtient par le *Théorème de Bayes*.

Théorème de Bayes

Soient deux événements A et B. La probabilité conditionnelle $P(A|B)$ est obtenue par

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

où $P(A|B)$ désigne la probabilité conditionnelle de A sachant B.

Expression de la loi *a posteriori*

Expression de la loi *a posteriori*

On suppose que les lois *a priori* de θ et de $\boldsymbol{z}$ sont des lois à densité.

La densité conditionnelle s'exprime par Le théorème de Bayes sous la forme :

$$f_{\boldsymbol{\Theta}|Z}(\theta|\boldsymbol{z}) = \frac{f_{\boldsymbol{\Theta}Z}(\theta, \boldsymbol{z})}{f_Z(\boldsymbol{z})} = \frac{f_{Z|\boldsymbol{\Theta}}(\boldsymbol{z}|\theta) f_{\boldsymbol{\Theta}}(\theta)}{f_Z(\boldsymbol{z})}.$$

Lois *a priori*, marginale, *a posteriori*

Loi *a priori* et marginale

Loi *a priori* $\pi(\theta)$: hypothèse à faire, quel choix ?

Loi *Marginale* $\mathbb{P}(Z)$: c'est la loi des observations
$\mathbb{P}(Z) = \int f(Z|\theta)\pi(\theta)d\theta$
En général difficile à calculer

Loi *a posteriori*

Loi *a posteriori* $\mathbb{P}(\theta|Z)$: loi de $\boldsymbol{\theta}$ conditionnellement aux observations collectées Z
$\mathbb{P}(\theta|Z) = \frac{\mathbb{P}(Z;\theta)}{\mathbb{P}(Z)}$ par application du théorème de Bayes.
$\mathbb{P}(Z;\theta)$ se développe en $\mathbb{P}(\mathrm{Z}|\theta)\pi(\theta)$.

$$\mathbb{P}(\theta|Z) = \frac{\mathbb{P}(\mathrm{Z}|\theta)\pi(\theta)}{\mathbb{P}(Z)}$$

Lois *a posteriori*

Loi *a posteriori*

Quatre cas sont possibles :

- La loi de Z et la loi *a priori* sont discrètes :
  Loi *a posteriori* discrète $\mathbb{P}(\theta = \theta_i | Z = z) = \frac{\mathbb{P}(Z=z|\theta=\theta_i)\mathbb{P}(\theta=\theta_i)}{\mathbb{P}(Z=z)}$
- La loi de Z est discrète et la loi *a priori* est continue :
  Loi *a posteriori* continue $\pi(\theta | Z = z) = \frac{\mathbb{P}(Z=z|\theta)\pi(\theta)}{\int \mathbb{P}(Z=z|t)\pi(t)dt}$
- La loi de Z est continue et la loi *a priori* est discrète :
  Loi *a posteriori* discrète $\mathbb{P}(\theta = \theta_i | Z) = \frac{f(Z|\theta=\theta_i)\mathbb{P}(\theta=\theta_i)}{\sum_k f(Z|\theta=\theta_k)\mathbb{P}(\theta=\theta_k)}$
- La loi de Z et la loi *a priori* sont continues :
  Loi *a posteriori* continue $\pi(\theta | Z) = \frac{f(Z|\theta)\pi(\theta)}{\int f(Z|t)\pi(t)dt}$

Choix de la loi

Loi *a priori* Exemple

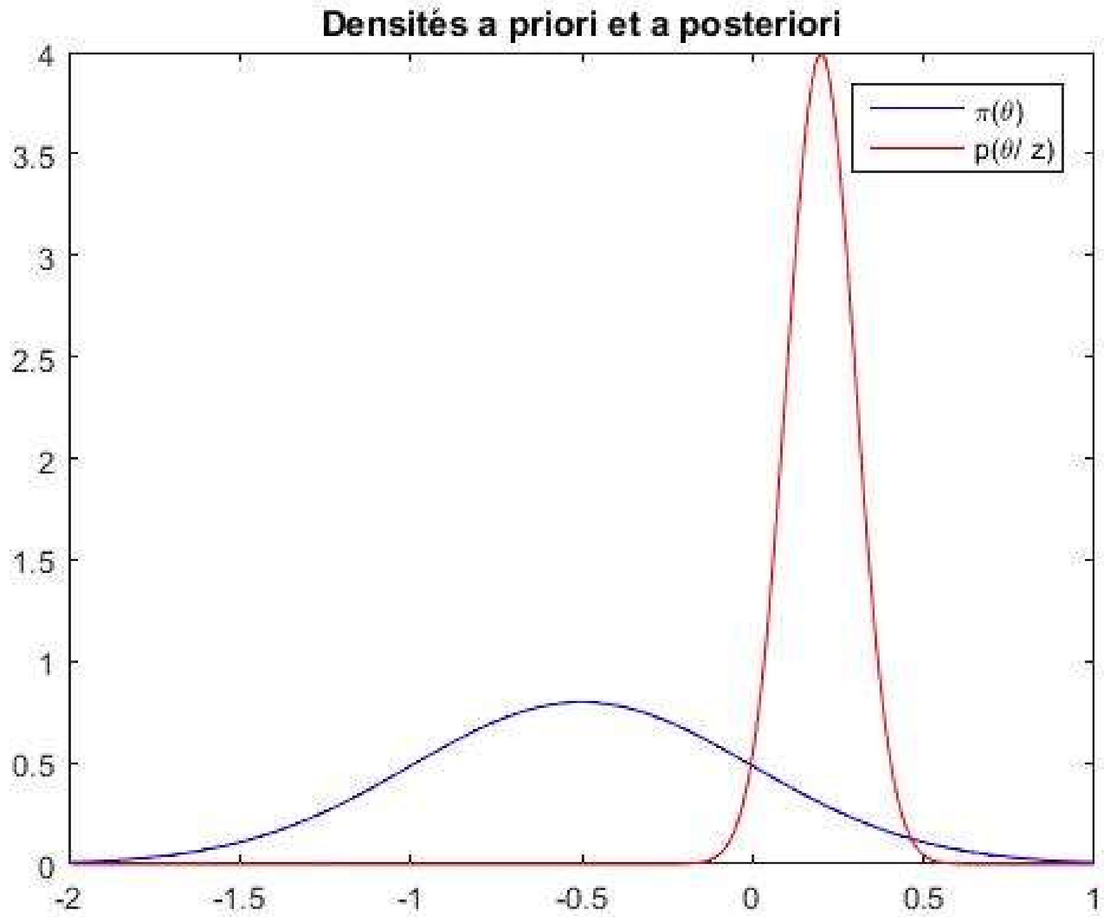
L'estimation Bayésienne fait intervenir la forme de la loi *a priori*
Cette loi reflète nos connaissances sur l'origine de la variabilité des données
Exemple :

- Mesure d'une position de cible $\boldsymbol{z} = \varphi(\theta) + \boldsymbol{b}$
- Variations : issues du bruit de capteur et de la variation de position possible de la cible
- Valeur du bruit de capteur : loi gaussienne de moyenne $\boldsymbol{m}_{\boldsymbol{B}}$ et de matrice de covariance centrée $\mathrm{R}_{\boldsymbol{B}}$.
- Position *a priori* de la cible $\varphi(\theta)$ avec θ : loi gaussienne de moyenne $\boldsymbol{m}_{\Theta}$ et de matrice de covariance centrée R_{Θ}

Choix de la loi

Exemple

La loi conditionnelle $\theta|z$ est une loi gaussienne dont les caractéristiques se déduisent des caractéristiques de la loi marginale et de la loi *a priori*.



Choix de la loi

Influence de la loi a priori

- Forme de la densité de probabilité a posteriori
- Caractéristiques

$\Rightarrow$ Choix à effectuer en tenant compte des connaissances disponibles

- Utilisation de lois non informatives
- Lois conjuguées

Choix de la loi : Loi *a priori* non informative

Définition

L'estimation Bayésienne fait intervenir la forme de la loi *a priori*
Si on ne dispose que de peu d'informations sur la forme de θ
La loi a priori à sélectionner est *peu ou non informative*
$\Rightarrow$ Intervient de façon la plus faible possible dans la loi *a posteriori*

Exemple de lois non informatives

- Lois invariantes : par translation $\pi(\theta) = \pi(\theta + \theta_0) \forall \theta_0$
  $\Rightarrow \pi$ constante : loi uniforme
- Lois invariantes : par facteur d'échelle $\pi(\theta) = \lambda \pi(\lambda\theta) \forall \lambda$
  $\Rightarrow \pi(\theta) = c/\theta$ c constant
- Loi *a priori* de Jeffreys

Loi *a priori* non informative

Loi *a priori* de Jeffreys

Loi basée sur la matrice d'information de Fisher
$I_{i,j}(\theta) = -\mathrm{E}[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln f(Z|\theta)]$
La loi de Jeffreys est définie par $\pi(\theta) \propto \sqrt{\det(I(\theta))}$

Intérêt de la loi

- $I(\theta)$ indique la quantité d'information apportée par $f(\boldsymbol{z}\,|\theta)$.
- Valeurs de θ pour lesquelles $I(\theta)$ est grande doivent être plus probables a priori
- Donne plus d'importance à l'information des données
- La loi est invariante par reparamétrisation $\phi = h(\theta)$

Lois conjuguées

Lois conjuguées

Calcul potentiellement difficile de la loi *a posteriori*
Facilité quand la loi *a priori* et la loi *a posteriori* ont la même forme $\Rightarrow$ Lois conjuguées

Exemple lois conjuguées

Soit $Z = (Z_1, \ldots, Z_n)$ un vecteur de n variables aléatoires
de loi Bernouilli P_θ
On suppose que θ a une loi *a priori* π_θ

$$f(\theta|\boldsymbol{z}) = \frac{\pi(\theta)\theta^{S_n(z)}(1-\theta)^{n-S_n(z)}}{\int_0^1 \pi(t)t^{S_n(z)}(1-t)^{n-S_n(z)}}$$

où $S_n(z) = \sum_{i=1}^n z_i$.
La loi *a posteriori* ne dépend des observations que par S_n

Exemple loi conjuguée

Exemple Bernouilli

θ doit varier entre 0 et 1.
Pour la loi *a priori* : Choix de lois Bêta : densité $f_{\alpha,\beta} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $0 < x < 1$
$B(\alpha,\beta)$ est la fonction définie par $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
Par substitution

$$f(\theta|y) = \frac{\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}}{B(y+\alpha, n-y+\beta)}$$

La loi *a posteriori* est donc aussi une loi Bêta de paramètres $y+\alpha, n-y+\beta$

Structure identique pour la loi *a priori* et la loi *a posteriori* : lois conjuguées

NB : Fonction Gamma : $\Gamma(a) = \int_0^{+\infty} x^{a-1}\exp(-x)dx$

Lois conjuguées

Exemples lois conjuguées

$f(z\|\theta)$	$f(\theta)$	$f(\theta\|z)$
$\mathcal{N}(\theta,\sigma^2)$	$\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}((\tau^2 z+\mu\sigma^2)/(\sigma^2+\tau^2), 1/(1/\sigma^2+1/\tau^2))$
$\mathcal{G}(n,\theta)$	$\mathcal{G}(a,b)$	$\mathcal{G}(n+a,b+z)$
$\mathcal{B}(n,\theta)$	$\mathcal{B}eta(a,b)$	$\mathcal{B}eta(n+a,b+z)$
$\mathcal{P}oisson(\mu)$	$\mathcal{G}(a,b)$	$\mathcal{G}(a+z,b+1)$

Loi a posteriori

Exemples lois conjuguées

$f(z\|\theta)$	$f(\theta)$	$E(\theta\|z)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$(\tau^2 z + \mu\sigma^2)/(\sigma^2 + \tau^2)$
$\mathcal{G}(n, \theta)$	$\mathcal{G}(a, b)$	$(a + n)/(b + z)$
$\mathcal{B}(n, \theta)$	$\mathcal{B}eta(a, b)$	$(a + n)/(a + n + b + z)$
$\mathcal{P}oisson(\mu)$	$\mathcal{G}(a, b)$	$(a + z)/(b + 1)$

Loi conjuguée

Propriétés particulières des lois exponentielles

Loi conjuguée pour lois exponentielles
Soit $f(x|\theta)$ famille exponentielle
Famille de lois conjuguées *a priori* : $\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda A(\theta))$
Loi *a posteriori* : $\pi(\theta|z) \propto \exp((\mu + z)\theta - (\lambda + 1)A(\theta))$

Remarques

- La proposition précédente est formelle et peut déboucher sur des lois *a priori* de θ non intégrables
- De nombreuses lois classiques sont sous forme exponentielle : Bernouilli, Poisson, Normale, Gamma ,...
- Les lois dont le support dépend de θ ne sont jamais exponentielles (ex loi uniforme)

Informations

Objectif de l'estimation Bayésienne

Règle de Bayes : fusionner deux sources d'information

- l'information en provenance des observations $\mathbb{P}(\mathrm{Z}|\theta)$
- l'information *a priori* représentée par $\pi(\boldsymbol{\theta})$.

Vraisemblance a posteriori

La fusion d'informations est représentée par la *vraisemblance a posteriori*.

- Cas Non Bayésien : vraisemblance $f_{\boldsymbol{Z}\,|\Theta}(\boldsymbol{z}\,|\theta)$
- Cas Bayésien : Vraisemblance a posteriori
  qui correspond à la loi *a posteriori* $\mathbb{P}(\theta|Z)$

Notion de risque

Définition d'un estimateur

Un estimateur Bayésien est une statistique T qui associe un vecteur de paramètres $\widehat{\theta}$ à un ensemble d'observations $\boldsymbol{z}$ et à un paramètre caractéristique θ ayant une distribution *a priori* connue .
Plusieurs estimateurs peuvent être déduits de la loi a posteriori : moyenne, médiane, modes ..

- Lequel choisir ,
- Quels sont les critères de choix ?
- existe-t-il des critères optimaux ?

Fonction de perte

Définition d'une fonction de coût/perte

Une fonction de perte ℓ est une fonction de $\Theta \times \Theta \Rightarrow \mathcal{R}^+$, mesurable qui vérifie : $\ell\,(\theta - \theta') = 0 \iff \theta = \theta', \forall\theta,\ \theta' \in \Theta$

Exemples

- Perte quadratique : $\ell\,(\theta,\ \theta') = (\theta - \theta')^2$
- Perte valeur absolue : $\ell\,(\theta,\ \theta') = |\theta - \theta'|^2$
- Perte de Hellinger : $\ell\,(\theta,\ \theta') = h(\theta,\ \theta')$ avec $h(P, Q)$ distance de Hellinger entre les lois $P = p\ d\mu$ et $Q = q\ d\mu$, définie par $h(P, Q)^2 = \int(\sqrt{p} - \sqrt{q})^2 d\mu$

Notion de risque

Risque

La fonction risque d'un estimateur T pour une fonction de perte ℓ est l'application

$$R \; : \Theta \; \Rightarrow \mathcal{R}^+ \tag{1}$$

$$\theta \; \Rightarrow R(\theta, T) = \mathrm{E}_\theta[\ell(\theta, T(\boldsymbol{z}))] = \int \ell(\theta, T(z)) dP_\theta(z) \tag{2}$$

Admissibilité

Un estimateur T est dit *inadmissible* s'il existe un estimateur T_1 tel que

$$\forall \theta \in \Theta, R(\theta, T1) \leq R(\theta, T) \tag{3}$$

$$\exists \theta_1, R(\theta_1, T1) < R(\theta_1, T) \tag{4}$$

Un estimateur est *admissible* s'il n'est pas inadmissible

Risque bayésien

Risque moyen

Cas non Bayésien : le risque est l'espérance sur Z du coût ou de la perte associé à un estimateur T.
Cas Bayésien : le risque doit faire intervenir l'a priori sur θ.
Le risque Bayésien est alors l'espérance sur Θ de l'espérance sur $\boldsymbol{Z}$ du coût pour T.

- $R_f(T) = \int_\Theta R(\theta, T) df(\theta) = \int_\Theta R(\theta, T)\pi(\theta) d(\theta)$
- $R_f(T) = \int_\Theta \int_{\boldsymbol{Z}} C(\theta, T(z)) f(z|\theta) dz \pi(\theta) d(\theta)$
- $R_f(T) = \int_\Theta \int_{\boldsymbol{Z}} C(\theta, T(z)) \pi(\theta|z) f(z) dz d(\theta)$

T est un estimateur *Bayésien* si il minimise le risque
$R_f(T) = \int_\Theta R(\theta, T) df(\theta)$

Risque bayésien

Estimateur de Bayes

L'estimateur défini par :
$\widehat{\theta} = T^{\pi}(\boldsymbol{z}) = arg \min_{T} \mathrm{E}^{\pi}[C(\theta, T)|z]$
où C est une fonction de coût
est *l'estimateur de Bayes* de θ associé à la loi *a priori* π

- Il est *admissible* au sens du risque de Bayes
- Il n'est pas nécessairement unique
- Il est convergent en probabilité (sous hypothèse de régularité)

Exemples

Risque en valeur absolue

- Fonction coût : $C(\theta, T) = |\theta - T|$
- L'estimateur de Bayes est la *médiane* de la loi *a priori*

Risque en moyenne quadratique

Risque quadratique moyen

Soient $\boldsymbol{Z}$ et Θ variables aléatoires du second ordre.
Le *risque quadratique moyen* est défini par

$$\mathrm{E}\left((T(\boldsymbol{Z}) - \Theta)^t \, (T(\boldsymbol{Z}) - \Theta)\right)$$

Estimateur associé

Soit $\mathrm{E}(\theta|\boldsymbol{Z} = \boldsymbol{z})$: *moyenne a posteriori* (Posterior Mean).
On a $\mathrm{E}(\theta|\boldsymbol{Z} = \boldsymbol{z}) = \int \theta f(\theta|\boldsymbol{Z} = \boldsymbol{z}) d\theta$

- Estimateur PM minimise le risque quadratique moyen (Estimateur MMSE)
- On a donc équivalence entre Estimateur MMSE et Estimateur Moyenne a posteriori

Estimateur moyenne quadratique minimale

Biais

Biais de l'estimateur MMSE (Minimum Mean Square Error) :
$\mathrm{E}_{\boldsymbol{Z},\theta}\left(\widehat{\theta}^{MMSE}-\theta\right)=\mathrm{E}_{\boldsymbol{Z}}\left(\mathrm{E}_{\theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\widehat{\theta}^{MMSE}-\theta\right)\right)=\mathrm{E}_{\boldsymbol{Z}}\,(0)=0.$
L'estimateur est donc *non biaisé*.

Covariance

La matrice de covariance de l'estimateur est
$\mathrm{E}_{\boldsymbol{Z},\theta}\left[(\widehat{\theta}^{MMSE}-\theta)(\widehat{\theta}^{MMSE}-\theta)^t\right]=\mathrm{E}_{\boldsymbol{Z}}\left[C_{\theta|\boldsymbol{Z}=\boldsymbol{z}}\right].$

où $C_{\theta|\boldsymbol{Z}=\boldsymbol{z}}=\mathrm{E}_{\theta|\boldsymbol{Z}=\boldsymbol{z}}\left[(\theta-\mathrm{E}_{\theta|\boldsymbol{Z}=\boldsymbol{z}}(\theta))(\theta-\mathrm{E}_{\theta|\boldsymbol{Z}=\boldsymbol{z}}(\theta))^t\right]$ est la matrice de covariance *a posteriori* de θ

Estimateur moyenne quadratique minimale

Borne Bayésienne

Soit $I_B(\theta)$ la matrice d'information moyenne

$$I_B(\theta) = \mathrm{E}_{\boldsymbol{Z},\theta}\left(-\frac{\partial^2 \ln f_{\boldsymbol{Z},\Theta}}{\partial\theta\partial\theta^T}\right)$$

$$I_B(\theta) = \mathrm{E}_\theta[\mathrm{E}_{\boldsymbol{Z}|\theta}\left(-\frac{\partial^2 \ln f_{\boldsymbol{Z}|\Theta}(\boldsymbol{Z}|\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 \ln f_\Theta(\theta)}{\partial\theta\partial\theta^T}\right)]$$

Soit finalement $I_B(\theta) = \mathrm{E}_\theta[I(\theta)] + \mathrm{E}_\theta\left[-\frac{\partial^2 \ln f_\Theta(\theta)}{\partial\theta\partial\theta^T}\right]$

La borne minimale de la variance (iid Cramer-Rao) est alors $I_B(\theta)^{-1}$

Maximum a posteriori

Estimateur

L'estimation au sens du maximum *a posteriori* (MAP) consiste à rechercher la valeur *la plus probable* du vecteur des paramètres compte tenu des observations en maximisant la vraisemblance *a posteriori* . (obtenue à partir de l'expression de la loi *a posteriori* de θ : $f_{\Theta|Z}$
Par application de la loi de Bayes
Le critère correspondant à maximiser est $J(\theta) = \ln f_{\boldsymbol{Z}\,|\Theta}(\boldsymbol{z}\,|\theta) + \ln f_{\Theta}(\theta)$
car $f_{\boldsymbol{Z}}\,(\boldsymbol{z}\,)$ ne dépend pas de θ.

Maximum a posteriori

Estimateur

Analyse des termes :

- $\ln f_{\boldsymbol{Z}\,|\Theta}(\boldsymbol{z}\,|\theta)$ est la log-vraisemblance de l'estimation non Bayésienne
- $\ln f_{\Theta}(\theta)$ est l'information *a priori* sur les variations des paramètres.

Estimateur du maximum de vraisemblance : cas limite d'estimation au sens du maximum *a posteriori* où $f_{\Theta}(\theta)$ est une loi uniforme dont le support tend vers le domaine de définition du vecteur des paramètres.

Exemple dans le cas Gaussien

Estimateur

Cas simple d'un bruit additionnel : $\boldsymbol{z} = \varphi(\theta) + \boldsymbol{b}$
Vecteur de bruit $\boldsymbol{b}$: gaussien de moyenne $\boldsymbol{m}_{\boldsymbol{B}}$ et de matrice de covariance centrée $\mathrm{R}_{\boldsymbol{B}}$.
Vecteur des paramètres θ : gaussien de moyenne $\boldsymbol{m}_{\Theta}$ et de matrice de covariance centrée R_{Θ}
Alors $J(\theta)$ s'écrit à une constante près

$$J(\theta) \# -\frac{1}{2}(\boldsymbol{z} - \varphi(\theta) - \boldsymbol{m}_{\boldsymbol{B}})^{\mathrm{t}} \mathrm{R}_{\boldsymbol{B}}^{-1}(\boldsymbol{z} - \varphi(\theta) - \boldsymbol{m}_{\boldsymbol{B}}) - \frac{1}{2}(\theta - \boldsymbol{m}_{\Theta})^{\mathrm{t}} \mathrm{R}_{\Theta}^{-1}(\theta - \boldsymbol{m}_{\Theta}),$$

Variante des moindres carrés, comprenant un terme de *rappel* de θ vers sa moyenne *a priori* $\boldsymbol{m}_{\Theta}$: on parle alors de *moindres carrés régularisés*.

Ce qu'il faut retenir

- L'estimation Bayésienne utilise la loi de θ conditionnelle aux observations pour estimer la valeur de ce (ces) paramètre (s).
- Dans le cadre Bayésien, le vecteur à estimer est un vecteur aléatoire auquel est associé une loi de densité *a priori*
- La loi conditionnelle est obtenue par application de la formule de Bayes et est fonction de la loi des observations conditionnelle à θ, de la loi a priori et de la loi des observations qui constitue une facteur de régularisation.
- Le calcul de ces lois conditionnelles est facilité par le choix de loi *a priori* conjuguée avec la loi $f(\boldsymbol{Z}\,|\theta)$.
- Les estimateurs bayésiens sont définis à partir de risques moyens tels le risque moyen quadratique ou le risque moyen en valeur absolue.
- L'estimateur du maximum de vraisemblance peut être vu comme une limite de l'estimateur du maximum de vraisemblance *a posteriori* lorsque la loi *a priori* de θ n'est pas informative.