NOTIONS D'ESTIMATION

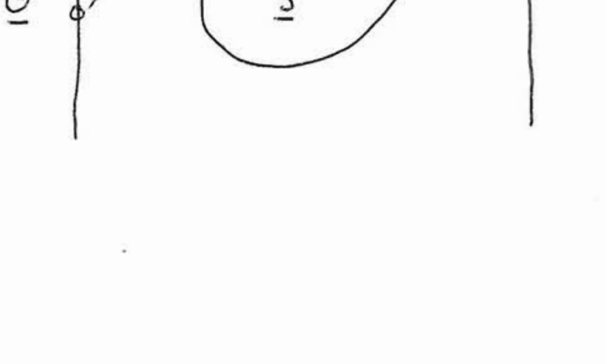
Introduction

À partir d'un ensemble de mesures, $\underline{y}$, il s'agit d'inférer la valeur d'un paramètre $\underline{\theta}$, relié aux mesures via une équation d'observation et/ou des informations statistiques.

Le paramètre $\underline{\theta}$ pourra être certain ou (considéré) comme aléatoire. On notera $\hat{\underline{\theta}}$ la valeur *estimée*, et la règle d'estimation, $\hat{\underline{\theta}}(\underline{y})$ est appelée *estimateur*.

La loi conditionnelle $p_{\underline{Y}|\underline{\Theta}}(\underline{y}|\underline{\theta})$ est appelée la *vraisemblance*, où le conditionnement indiquera simplement une "dépendance" dans le cas où $\underline{\theta}$ est déterministe (certain). Les connaissance ou les contraintes sur $\underline{\theta}$ peuvent être formulées sous la forme d'une *loi a priori* $p_{\underline{\theta}}(\underline{\theta})$, et la règle de Bayes permet alors d'obtenir la loi a posteriori $p_{\underline{\Theta}|\underline{Y}}(\underline{\theta}|\underline{y})$ selon

$$p_{\underline{\Theta}|\underline{Y}}(\underline{\theta}|\underline{y}) = \frac{p_{\underline{Y}|\underline{\Theta}}(\underline{y}|\underline{\theta})\, p_{\underline{\Theta}}(\underline{\theta})}{p_{\underline{Y}}(\underline{y})}.$$



Le problème est de se donner des règles pour construire $\hat{\underline{\theta}}(\underline{y})$ et déterminer ses performances.

Pour évaluer la qualité de l'estimation, on pourra considérer le biais, la variance (ou la matrice de covariance), une borne d'estimation, un intervalle de confiance, ...

Moyenne :

$$E[\hat{\underline{\theta}}] = \int \hat{\underline{\theta}}(\underline{y})\, P_{\underline{Y}|\underline{\theta}}(\underline{y})\, d\underline{y} \qquad \underline{\theta} \text{ certain}$$

$$E[\hat{\underline{\theta}}] = \int \underset{\underline{Y}|\underline{\theta}}{E}[\hat{\underline{\theta}}|\underline{\theta}]\, P_{\underline{\theta}}(\theta)\, d\underline{\theta} \qquad \underline{\theta} \text{ aléatoire}$$

Biais

$$B(\hat{\underline{\theta}}) = \underline{\theta} - E[\hat{\underline{\theta}}] \qquad \text{cas certain}$$

$$B(\hat{\underline{\theta}}) = E[\theta - E[\hat{\underline{\theta}}]] \qquad \text{cas aléatoire}$$

Covariance

$$\underline{\underline{R}}_{\hat{\underline{\theta}}} = E\left[(\hat{\underline{\theta}} - E[\hat{\underline{\theta}}])(\hat{\underline{\theta}} - E[\hat{\underline{\theta}}])^{+}\right]$$

Exemples :

- $y(n) = \sum a_i\, y(n-i) + b(n)$

  $\underline{\theta} = [a_1 \cdots a_p, \sigma_b^2]$

- $y(n) = \sum h(\ell)\, x(n-\ell) + b(n)$

  $\underline{\theta} = [h(0) \cdots h(N)]$ pb d'identification

  $\underline{\theta} = [x(n) \cdots x(n-N+1)]$ pb d'inversion (déconvolution)

- $y(n) = A \sin(2\pi f_0 t + \phi) + b(n)$

  $\underline{\theta} = [A, f_0, \phi]$ ou $\underline{\theta} = [f_0]$

- $y(n) = A\, s(n-r) + b(n)$

  $\underline{\theta} = [A, r]$ radar, sonar, échographie ...

Pour juger et construire un estimateur, il faut

- utiliser toutes les données,
- donner des résultats acceptables
  - être consistant : converger en probabilité

cas certain | $\lim_{N \to \infty} \text{Proba}\left(| \hat{\theta}_N - \theta | > \varepsilon \right) = 0$
où N est le nb de données,
i.e. il est non biaisé, au moins asymptotiquement

  - être efficace (atteint la borne de Cramer-Rao définie plus loin) où sa variance est minimale
Il pourra être asymptotiquement efficace.

Intervalle de confiance

$$\text{Proba}\left[\underline{\hat{\theta}} - \underline{\ell_1} \leq \underline{\theta} \leq \underline{\hat{\theta}} + \underline{\ell_2} \right] = 1 - \alpha$$

Quel est l'intervalle $[\ell_1, \ell_2]$ pour que θ soit compris dans $[\underline{\hat{\theta}} - \underline{\ell_1}, \underline{\hat{\theta}} + \underline{\ell_2}]$ avec une proba $(1-\alpha)$?

- Lorsque $\{\underline{\theta}\}$ n'est pas probabilisé, on utilisera un critère d'optimisation bati à partir d'une mesure d'erreur minimale $\| \hat{\theta}(y) - \theta \|$ ou d'une mesure d'observation minimale $\| \underline{y} - h(\underline{\hat{\theta}_y}) \|$ ou d'une construction à maximum de vraisemblance

- Lorsque $\{\underline{\theta}\}$ est probabilisé, on se donne une fonction de coût ou de perte $C(\underline{\hat{\theta}}, \underline{\theta})$, que l'on cherche à minimiser en moyenne. On parle de minimisation du risque bayésien.

Notons que dans ce cadre, c'est la loi a posteriori qui engrange toute l'information, et que c'est à partir de cette loi a posteriori que l'on définit un estimateur ponctuel $\cancel{p(\underline{\theta}}$ $\hat{\underline{\theta}}$.

Estimation bayésienne

On choisit une fonction de coût $c(\hat{\underline{\theta}},\underline{\theta})$, avec

- $c(\hat{\underline{\theta}},\underline{\theta}^{*}) \geq 0$,
- $c(\hat{\underline{\theta}},\hat{\underline{\theta}}) = 0$,
- non décroissante.

Le risque est $\mathcal{R} = E_{\underline{\theta},\hat{\underline{\theta}}}\left[c(\hat{\underline{\theta}},\underline{\theta})\right]$, soit

$$\mathcal{R} = \iint_{\underline{\theta}\,\underline{y}} p(\underline{y},\underline{\theta})\, c(\hat{\underline{\theta}},\hat{\theta})\, d\underline{\theta}\, d\underline{y}$$

Or $p(\underline{y},\underline{\theta}) = p(\underline{\theta}|\underline{y})\, p(\underline{y})$, ce qui conduit à

$$\mathcal{R} = \int_{\underline{y}} p(\underline{y}) \int_{\underline{\theta}} p(\underline{\theta}|\underline{y})\, c(\hat{\underline{\theta}},\underline{\theta})\, d\underline{\theta}$$

Minimiser $\mathcal{R}$ revient alors à minimiser l'intégrale sur $\underline{\theta}$, à $\underline{y}$ fixé.

a) Coût quadratique [MMSE – Minimum Mean Square Error]

Il faut minimiser

$$\begin{aligned}
f(\hat{\underline{\theta}}) &= \int_{\underline{\theta}} p(\underline{\theta}|\underline{y})\, c(\hat{\underline{\theta}},\underline{\theta})\, d\underline{\theta} \\
&= E_{\underline{\theta}|y}\left[c(\hat{\underline{\theta}},\underline{\theta})\right] \\
&= E_{\underline{\theta}|y}\left[(\hat{\underline{\theta}}-\underline{\theta})^T(\hat{\underline{\theta}}-\underline{\theta})\right] \\
&= \hat{\underline{\theta}}^T\hat{\underline{\theta}} - \hat{\underline{\theta}}^T E_{\underline{\theta}|y}[\underline{\theta}] - E_{\underline{\theta}|y}[\underline{\theta}]^T\underline{\theta} + E_{\underline{\theta}|y}[\underline{\theta}^T\underline{\theta}]
\end{aligned}$$

et $$\frac{df(\hat{\underline{\theta}})}{d\hat{\underline{\theta}}} = 2\hat{\underline{\theta}} - 2E_{\underline{\theta}|y}[\underline{\theta}]$$

Ce qui fournit :

$$\underline{\hat{\theta}}_{MMSE} \stackrel{\Delta}{=} E_{\underline{\theta}|\underline{y}}[\theta],$$

c'est-à-dire la <u>moyenne a posteriori</u>.

b) <u>Estimation en valeur absolue.</u>

$$f(\underline{\hat{\theta}}) = \int p(\underline{\theta}|\underline{y}) \, |\underline{\hat{\theta}} - \underline{\hat{\theta}}|^T d\underline{\theta} \qquad |\hat{\theta} - \theta|$$

$$= \int_{-\infty}^{\underline{\hat{\theta}}} p(\underline{\theta}|\underline{y})(\underline{\theta} - \underline{\hat{\theta}}) d\underline{\theta} + \int_{\underline{\hat{\theta}}}^{+\infty} p(\underline{\theta}|\underline{y})(\underline{\hat{\theta}} - \underline{\theta}) d\underline{\theta}$$

En intégrant par parties,

$$\int_{-\infty}^{\underline{\hat{\theta}}} (\underline{\hat{\theta}} - \underline{\theta}) \, p(\underline{\theta}|\underline{y}) \, d\underline{\theta} = \int_{-\infty}^{\underline{\hat{\theta}}} F(\underline{\theta}|\underline{y}) \, d\underline{\theta}$$

avec $F(\underline{\theta}|\underline{y}) = \int_{-\infty}^{\underline{\theta}} p(\underline{x}|\underline{y}) d\underline{x} = \text{Proba}\{\underline{x} < \underline{\theta} \,|\, \underline{y}\}$

et

$$\int_{\underline{\hat{\theta}}}^{+\infty} (\underline{\hat{\theta}} - \underline{\theta}) \, p(\underline{\theta}|\underline{y}) \, d\underline{\theta} = \int_{\underline{\hat{\theta}}}^{+\infty} -1 + F(\underline{\theta}|\underline{y}) \, d\underline{\theta}$$

$$= \int_{\underline{\hat{\theta}}}^{+\infty} \text{Proba}\{\underline{x} > \underline{\theta} \,|\, \underline{y}\} \, d\underline{\theta}$$

On obtient alors

$$f(\underline{\hat{\theta}}) = \int_{-\infty}^{\underline{\hat{\theta}}} \text{Proba}\{\underline{x} < \underline{\theta} \,|\, \underline{y}\} \, d\underline{\theta} + \int_{\underline{\hat{\theta}}}^{+\infty} \text{Proba}\{\underline{x} > \underline{\theta} \,|\, \underline{y}\} \, d\underline{\theta}$$

$$= \int_{-\infty}^{\underline{\hat{\theta}}} \text{Proba}\{\underline{x} < \underline{\theta} \,|\, \underline{y}\} - \text{Proba}\{\underline{x} > \underline{\theta} \,|\, \underline{y}\} \, d\underline{\theta}$$

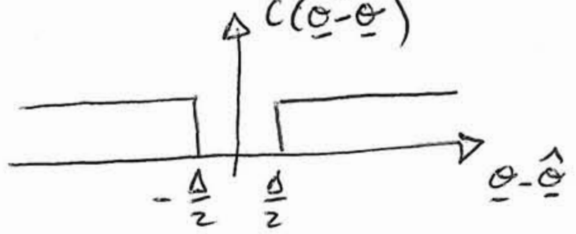
La dérivée par rapport à $\underline{\hat{\theta}}$ fournit alors

$$\left.\frac{df(\hat{\theta})}{d\hat{\theta}}\right|_{\hat{\theta} = \stackrel{\Delta}{\hat{\theta}}} = 0 \Rightarrow \text{Proba}\{\underline{x} < \underline{\hat{\theta}} \,|\, \underline{y}\} = \text{Proba}\{\underline{x} > \underline{\hat{\theta}} \,|\, \underline{y}\}$$

Et il faut donc choisir $\underline{\hat{\theta}}$ comme la <u>médiane a posteriori</u>. Cet estimateur sera confondu avec la moyenne a posteriori si $p(\underline{\theta}|\underline{y})$ est paire autour de sa moyenne.

c- Estimation avec coût uniforme

Le coût est uniforme, sauf sur un intervalle de largeur Δ où il est nul. Ce coût pénalise donc uniformément les écarts $\underline{\theta}-\underline{\hat{\theta}}$ dès que $\underline{\theta}-\underline{\hat{\theta}} > \frac{1}{2}\Delta$



représenté dans le cas scalaire

On a alors

$$f(\underline{\hat{\theta}}) = \int p_{\underline{\theta}|\underline{y}}(\underline{\theta}|\underline{y})\, C(\underline{\hat{\theta}}-\underline{\theta})\, d\underline{\theta}$$

$$= 1 - \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} p(\underline{\theta}|\underline{y})\, d\underline{\theta}$$

Bien sûr, $f(\underline{\hat{\theta}})$ est minimale lorsque $\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} p(\underline{\theta}|\underline{y})\,d\underline{\theta}$ est maximale. En faisant $\Delta \to 0$, on obtient ainsi le <u>maximum</u> de la loi <u>a posteriori</u> (MAP)

Deux résultats généraux :

<u>Théo</u>. Si la fonction de coût est paire, convexe, et la densité a posteriori paire autour de sa moyenne, alors tous les coûts mènent au même estimateur.

<u>Théo</u> - Si la fonction de coût est paire, la densité a posteriori paire autour de sa moyenne et unimodale, tous les estimateurs sont confondus.

Maximum de vraisemblance.

Dans la méthode du maximum de vraisemblance, le paramètre $\underline{\theta}$ est certain, et l'on recherche sa valeur afin de maximiser la probabilité des données :

$$\underline{\hat{\theta}}_{MV} = \underset{\underline{\theta}}{\mathrm{Max}} \; P(\underline{y}|\underline{\theta})$$

Souvent on travaille avec le logarithme de la fonction de vraisemblance (pour simplifier les exponentielles en particulier). La solution est inchangée (le max ne bouge pas lorsque l'on prend une fonction monotone)

$$\underline{\hat{\theta}}_{MV} = \mathrm{Arg}\,\underset{\theta}{\mathrm{Max}} \; \log P_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})$$

Il est à noter que le MV et le MAP sont <u>confondus</u> lorsque l'on prend $P_{\underline{\theta}}(\underline{\theta})$ uniforme. En effet,

$$\underbrace{\log P_{\underline{\theta}|\underline{Y}}(\theta|y)}_{\text{a posteriori}} = \underbrace{\log P_{\underline{Y}|\underline{\theta}}(y|\theta)}_{\text{vraisemblance}} + \underbrace{\log P_{\underline{\theta}}(\underline{\theta})}_{\text{constant si } P_{\underline{\theta}} \text{ uniforme}} - \underbrace{\log P_{Y}(y)}_{\text{ne dépend pas de } \underline{\theta}}$$

et la maximisation par rapport à $\underline{\theta}$ fournit alors le même résultat.

Exemple idiot :

$y = \theta + b$, cas scalaire, b gaussien

alors $P_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta}) = P_B(\underline{b}) = P_B(y-\theta)$

$P_B(x)$ est maximale pour $x = 0$ (si gaussienne centrée), et

la solution du max de vraisemblance est $\boxed{\theta_{MV} = y}$

Estimation selon les moindres carrés

Le paramètre $\underline{\theta}$ est déterministe. Seule la moyenne $E_{\underline{Y}|\underline{\theta}}[\underline{y}|\underline{\theta}]$ est connue et l'on utilise une distance à l'observation de la forme $d(\underline{y}, E_{\underline{Y}|\underline{\theta}}[\underline{y}|\underline{\theta}])$ ce qui conduit à l'estimateur

$$\underline{\hat{\theta}}(\underline{y}) = \operatorname{Arg}\min_{\underline{\theta}} d(\underline{y}, E_{\underline{Y}|\underline{\theta}}[\underline{y}|\underline{\theta}]).$$

Lorsqu'on choisit une distance associé à la norme euclidienne, éventuellement pondérée, on parle de moindres carrés. On obtient:

$$\underline{\hat{\theta}}_{MC}(\underline{y}) = \operatorname{Arg}\min_{\underline{\theta}} \left\{ (\underline{y} - E_{\underline{Y}|\underline{\theta}}[\underline{y}|\underline{\theta}])^T \underline{\underline{M}} (\underline{y} - E_{\underline{Y}|\underline{\theta}}[\underline{y}|\underline{\theta}]) \right\}$$

avec $\underline{\underline{M}}$ une matrice symétrique définie positive.

Modèle

Le modèle d'observation général est de la forme $\underline{y} = \underline{h}(\underline{\theta}, \underline{b})$ où h est une relation déterministe connue et $\underline{b}$ représente la part d'aléatoire.
Souvent le modèle est additif, et

$$\underline{y} = \underline{h}(\underline{\theta}) + \underline{b}.$$

Dans ce cas, $E_{\underline{Y}|\underline{\theta}}[\underline{y}|\underline{\theta}] = \underline{h}(\underline{\theta}) + E_B[\underline{B}]$, et le critère des MC devient

$$(\underline{y} - \underline{h}(\underline{\theta}) - E[\underline{B}])^T \underline{\underline{M}} (\underline{y} - \underline{h}(\underline{\theta}) - E[\underline{B}])$$

La solution est obtenue lorsque le gradient du critère est nul, soit

$$\left[\frac{\partial \underline{h}(\underline{\theta})}{\partial \underline{\theta}}\right]^T \underline{\underline{M}} (\underline{y} - \underline{h}(\underline{\theta}) - E_B[\underline{B}]) = 0$$

Lorsque le modèle est <u>linéaire</u> en $\underline{\theta}$, on a

$$\underline{y} = \underline{\underline{H}}\,\underline{\theta} + \underline{b},$$

et le gradient devient

$$\underline{\underline{H}}^t \underline{\underline{M}} \left[\underline{y} - \underline{\underline{H}}\,\underline{\theta} - E_B[B] \right]$$

Ce qui fournit finalement

$$\underline{\hat{\theta}}_{MC}(\underline{y}) = \left(\underline{\underline{H}}^T \underline{\underline{M}}\,\underline{\underline{H}}\right)^{-1} \underline{\underline{H}}^T \underline{\underline{M}} \left(\underline{y} - E[B] \right)$$

si la matrice $\underline{\underline{H}}^T \underline{\underline{M}}\,\underline{\underline{H}}$ a le bon goût d'être inversible.

La matrice $\underline{\underline{M}}$ sert à donner plus ou moins d'importance aux différentes composantes de l'observation $\underline{y}$. On la choisit souvent diagonale et on parle de <u>moindres carrés pondérés</u>. Lorsque $\underline{\underline{M}} = \underline{\underline{1}}$, on obtient les <u>moindres carrés ordinaires</u>.

<u>Caractéristiques</u> :

- <u>Biais</u>

$$E_{\underline{y}|\underline{\theta}}\left[\hat{\theta}_{MC}\right] = (H^T M H)^{-1} H^T M \left(E_{\underline{y}|\underline{\theta}}[Y] - E_B[B]\right)$$

or $\underline{y} = \underline{\underline{H}}\,\underline{\theta} + \underline{B}$, et $E_{Y|\theta}\left[Y|\underline{\theta}\right] = \underline{\underline{H}}\,\underline{\theta} + E_B[\underline{B}]$,

et il reste $E_{\underline{Y}|\underline{\theta}}\left[\underline{\hat{\theta}}_{MC}\right] = (\underline{\underline{H}}^T \underline{\underline{M}}\,\underline{\underline{H}})^{-1} (\underline{\underline{H}}^T \underline{\underline{M}}\,\underline{\underline{H}}) \theta$

$$= \underline{\theta}$$

L'estimateur est donc <u>non biaisé</u>.

- <u>Covariance</u>

Il faut disposer de la matrice de covariance du bruit. Notons là $\underline{\underline{R}}_B$.

$$\underline{\underline{R}}_B = E_B\left[\left(\underline{B} - E_{\underline{B}}[\underline{B}]\right)\left(\underline{B} - E_B[\underline{B}]\right)\right]$$

Dans ce cas, la matrice de covariance de l'estimateur est

$$\underline{\underline{R}}_{\underline{\hat{\theta}}} = E_{\underline{Y}|\underline{\theta}}\left[\left(\underline{\hat{\theta}} - E[\underline{\hat{\theta}}]\right)\left(\underline{\hat{\theta}} - E[\underline{\hat{\theta}}]\right)\right]$$

En remplaçant $\hat{\theta}_{MC}$ par son expression :

$$\underline{\underline{R}}_{\hat{\underline{\theta}}} = E_{\underline{Y}|\underline{\theta}}\left[\left\{(H^TMH)^{-1}H^TM\left(\underline{\underline{H}}\underline{\theta}+\underline{B}-E[\underline{B}]\right)-\underline{\theta}\right\}\left\{\quad\right\}^T\right]$$

$$= (H^TMH)^{-1}H^TME\left[(B-E[B])(\underline{B}-E[B])\right]MH(H^TMH)^{-1}$$

$$= (H^TMH)^{-1}H^TM R_B MH(H^TMH)^{-1}$$

Si l'on prend $\underline{\underline{M}} = \underline{\underline{R}}_B^{-1}$, cela se simplifie en

$$\underline{\underline{R}}_{\hat{\underline{\theta}}} = \left(\underline{\underline{H}}^T\underline{\underline{R}}_B^{-1}\underline{\underline{H}}\right)^{-1}.$$

Lien avec l'estimation à maximum de vraisemblance.

Si l'on dispose de la loi sur les observations $P_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})$, on peut chercher à utiliser toute cette information en utilisant comme critère le maximum de vraisemblance, qui maximise la probabilité de l'observation :

$$\hat{\underline{\theta}}_{MV} = \operatorname*{Arg\,max}_{\underline{\theta}} P_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})$$

Toujours dans le cadre linéaire, supposons que le bruit additif soit gaussien, de moyenne $\underline{m}_B = E_B[\underline{B}]$ et de covariance $\underline{\underline{R}}_B$. Alors

$$P_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta}) = P_B(\underline{y}-\underline{\underline{H}}\underline{\theta})$$

$$= \frac{1}{\left[(2\pi)^N|\underline{\underline{R}}_B|\right]^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\underline{y}-\underline{\underline{H}}\underline{\theta}-\underline{m}_B\right)^T\underline{\underline{R}}_B^{-1}\left(\underline{y}-\underline{\underline{H}}\underline{\theta}-\underline{m}_B\right)\right\}$$

et comme le coefficient de normalisation ne dépend pas de $\underline{\theta}$, maximiser $P_{\underline{Y}|\underline{\theta}}$ revient à minimiser

$$\left(\underline{y}-\underline{\underline{H}}\underline{\theta}-\underline{m}_B\right)\underline{\underline{R}}_B^{-1}\left(\underline{y}-\underline{\underline{H}}\underline{\theta}-\underline{m}_B\right),$$

c'est-à-dire le critère des moindres carrés pondérés avec $\underline{\underline{M}} = \underline{\underline{R}}_B^{-1}$. On a équivalence, dans le cas <u>linéaire</u> gaussien, entre MC et MV.

Lien avec l'estimation (sans biais) à variance minimale.

Pour un paramètre déterministe, une autre technique très prisée consiste à rechercher un estimateur non biaisé et à variance minimale (mais notons qu'on peut obtenir des estimateurs dont l'erreur quadratique moyenne est bien plus faible en se permettant un peu de biais — $E[|\hat{\underline{\theta}} - E[\hat{\underline{\theta}}]|^2] = Var[\hat{\underline{\theta}}] + B[\hat{\underline{\theta}}]^2$).

On choisit comme critère la Trace de la matrice de covariance de l'erreur d'estimation $(\hat{\underline{\theta}} - \underline{\theta})$.

Dans le cas d'un modèle linéaire, $\underline{Y} = \underline{\underline{H}}\,\underline{\theta} + \underline{B}$, et on note $\underline{\underline{R}}_B = E[(\underline{B} - \underline{m}_B)(\underline{B} - \underline{m}_B)^T]$ (matrice $p \times p$)

On s'impose un estimateur linéaire de la forme

$\hat{\underline{\theta}} = \underline{\underline{A}}\,\underline{Y} + \underline{c}$ avec $\underline{\underline{A}}$ une matrice $n \times p$

On recherche $\underline{\underline{A}}$ et $\underline{c}$ en (1) annulant le biais, (2) minimisant la trace de la matrice de covariance de $\hat{\underline{\theta}}$.

Biais

$$E_{\underline{Y}|\underline{\theta}}[\hat{\underline{\theta}} - \underline{\theta}] = \underline{\underline{A}}\, E_{\underline{Y}|\underline{\theta}}[\underline{Y}] + \underline{c} - \underline{\theta}$$
$$= \underline{\underline{A}}\,\underline{\underline{H}}\,\underline{\theta} + \underline{\underline{A}}\,\underline{m}_B + \underline{c} - \underline{\theta}$$

Le biais sera donc nul si $(\underline{\underline{A}}\,\underline{\underline{H}} - \underline{\underline{1}}) = \underline{0}$ et $\underline{\underline{A}}\,\underline{m}_b + \underline{c} = \underline{0}$

Comme $\underline{\underline{A}}$ et $\underline{\underline{H}}$ sont rectangulaires, la condition $\underline{\underline{A}}\,\underline{\underline{H}} = \underline{\underline{1}}$ ne suffit pas et l'on examine la covariance.

Covariance

$$\underline{\underline{R}}_{\hat{\underline{\theta}}} = E_{\underline{Y}|\underline{\theta}}\left[(\hat{\underline{\theta}} - \underline{\theta})(\hat{\underline{\theta}} - \underline{\theta})^T\right]$$
$$= E_{\underline{Y}|\underline{\theta}}\left[(\underline{\underline{A}}\,\underline{Y} + \underline{c} - \underline{\theta})(\underline{\underline{A}}\,\underline{Y} + \underline{c} - \underline{\theta})^T\right]$$
$$= E_{\underline{Y}|\underline{\theta}}\left[(\underline{\underline{A}}\,\underline{\underline{H}}\,\underline{\theta} + \underline{\underline{A}}\,\underline{B} + \underline{c} - \underline{\theta})(\text{idem})^T\right]$$
$$= E_{\underline{Y}|\underline{\theta}}\left[\underline{\underline{A}}(\underline{B} - \underline{m}_B)(\underline{B} - \underline{m}_B)^T \underline{\underline{A}}^T\right] \quad \text{car } \underline{\underline{A}}\,\underline{\underline{H}}\,\underline{\theta} - \underline{\theta} = 0, \ \underline{c} = -\underline{\underline{A}}\,\underline{m}_B$$

c'est-à-dire $\quad \underline{\underline{R}}_{\underline{\hat{\theta}}} = \underline{\underline{A}}\,\underline{\underline{R}}_B\,\underline{\underline{A}}^T$

On montre que la matrice $\underline{\underline{A}}$ qui minimise la trace de $\underline{\underline{R}}_{\underline{\hat{\theta}}}$ sous la contrainte $\underline{\underline{A}}\,\underline{\underline{H}} = \underline{\underline{1}}$ est donnée par

$$\underline{\underline{A}} = \left(\underline{\underline{H}}^T \underline{\underline{R}}_B^{-1} \underline{\underline{H}}\right)^{-1} \underline{\underline{H}}^T \underline{\underline{R}}_B .$$

L'estimateur linéaire à variance minimale s'écrit alors

$$\underline{\hat{\theta}}_{LVM} = (H^T R_B^{-1} H)^{-1} H^T R_B^{-1} (\underline{\theta} - \underline{m}_B)$$

qui est identique à l'estimateur des moindres carrés, avec $\underline{\underline{M}} = \underline{\underline{R}}_B^{-1}$.

Estimation au sens de l'Erreur Quadratique Moyenne Minimale

Dans le cas d'un paramètre aléatoire $\underline{\theta}$, on a vu que la minimisation du risque bayésien avec un coût quadratique mène à la moyenne a posteriori (ou moyenne conditionnelle)

$$\underline{\hat{\theta}}_{MQ} = E_{\underline{\theta}|\underline{Y}}[\underline{\theta}] .$$

MQ : moyenne quadratique
ou EQM : Erreur Quadratique Moyenne

Mais cet estimateur ne peut être obtenu que si l'on dispose complètement des différentes lois mises en jeu. Lorsque ce n'est pas le cas, et que l'information est réduite aux deux premiers moments, on impose à l'estimateur d'être une fonction linéaire (affine en fait) des observations et on poursuit avec un critère EQM.

On note $\underline{m}_Y$, $\underline{m}_{\underline{\theta}}$, $\underline{\underline{R}}_Y$ et $\underline{\underline{R}}_{\underline{\theta}\underline{Y}}$ les moyennes et matrices de covariance (centrées).

On impose à l'estimateur d'être de la forme

$$\underline{\hat{\theta}} = \underline{\underline{A}}\,\underline{Y} + \underline{c} \qquad A\,(p\times n) \quad c\,(p\times 1)$$

la condition d'annulation de biais fournit

$$\begin{cases} E[\underline{\hat{\theta}}] = \underline{\underline{A}}\,m_Y + \underline{c} = \underline{m}_\theta \\ \text{d'où} \quad \underline{c} = \underline{m}_\theta - \underline{\underline{A}}\,\underline{m}_Y \end{cases}$$

On recherche ensuite $\underline{\underline{A}}$ de sorte à minimiser la trace de la matrice de covariance. On travaillera avec les variables <u>centrées</u> pour soulager l'écriture et le rédacteur.

$$\begin{cases} \underline{\hat{\theta}}_c = \underline{\hat{\theta}} - E[\hat{\theta}] = \hat{\theta} - \underline{m}_\theta \\ \underline{Y}_c = \underline{Y} - E[\underline{Y}] \\ \underline{\theta}_c = \underline{\theta} - \underline{m}_\theta \end{cases}$$

On a alors
$$\underline{\underline{R}}_{\hat{\underline{\theta}}_c} = \underset{\underline{Y}\underline{\theta}}{E}\left[(\underline{\hat{\theta}}_c - \underline{\theta}_c)(\underline{\hat{\theta}}_c - \underline{\theta}_c)^T\right]$$
$$= E\left[(\underline{\underline{A}}\,\underline{Y}_c - \underline{\theta}_c)(\underline{\underline{A}}\,\underline{Y}_c - \underline{\theta}_c)^T\right]$$
$$= \underline{\underline{A}}\,E[\underline{Y}_c\underline{Y}_c^T]A^T - A\,E[\underline{Y}\,\underline{\theta}_c^T] - E[\underline{\theta}_c\underline{Y}^T]A^T + E[\underline{\theta}_c\underline{\theta}_c^T]$$
$$= \underline{\underline{A}}\,\underline{\underline{R}}_{YY}\,\underline{\underline{A}}^T - \underline{\underline{A}}\,\underline{\underline{R}}_{Y\theta} - \underline{\underline{R}}_{\theta Y}\,\underline{\underline{A}}^T + \underline{\underline{R}}_\theta$$

Ce qui se met également sous la forme

$$\underline{\underline{R}}_{\hat{\theta}_c} = \left(\underline{\underline{A}} - \underline{\underline{R}}_{\theta Y}\underline{\underline{R}}_{YY}^{-1}\right)\underline{\underline{R}}_{YY}\left(\underline{\underline{A}} - \underline{\underline{R}}_{\theta Y}\underline{\underline{R}}_{YY}^{-1}\right)^T + \underline{\underline{R}}_{\theta\theta} - \underline{\underline{R}}_{\theta Y}\underline{\underline{R}}_{Y}^{-1}\underline{\underline{R}}_{Y\theta}$$

Comme le premier terme est une forme quadratique (ses termes diagonaux sont tous positifs), on voit que la trace de $\underline{\underline{R}}_{\hat{\theta}}$ sera minimale pour $\boxed{\underline{\underline{A}} = \underline{\underline{R}}_{\theta Y}\underline{\underline{R}}_{YY}^{-1}}$, ce qui conduit finalement à

$$\underline{\hat{\theta}}_{ELMQ} = E_\theta[\underline{\theta}] + \underline{\underline{R}}_{\theta Y}\underline{\underline{R}}_{YY}^{-1}\left(\underline{Y} - E_Y[\underline{Y}]\right).$$

Il est important de noter que ce résultat a été obtenu de manière générale, sans s'appuyer sur une équation d'observation.

Cas linéaire (gaussien)

On reprend tous ces résultats dans le cas d'un modèle linéaire, avec une hypothèse gaussienne lorsque nécessaire :

$$\underline{y} = \underline{\underline{H}}\,\underline{\theta} + \underline{b}$$

exemple : identification de filtre, prédiction linéaire, etc.

Estimateurs :

1) MC - on a $g(\underline{\theta}) = \underline{\underline{H}}\,\underline{\theta}$ $\qquad \dfrac{dg(\underline{\theta})}{d\underline{\theta}} = \underline{\underline{H}}$

ce qui conduit à $\boxed{(\underline{\underline{H}}^T\underline{\underline{M}}\,\underline{\underline{H}})\,\hat{\underline{\theta}} = \underline{\underline{H}}^T\underline{\underline{M}}\,\underline{y}}$

(estimateur sans biais)

2) MV - on suppose $\underline{b}$ gaussien, centré, de cov. $\underline{\underline{R}}_B$

$$p(\underline{y}|\underline{\theta}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\underline{\underline{R}}_b|^{\frac{1}{2}}} \exp -\frac{1}{2}\left\{(\underline{y} - \underline{\underline{H}}\,\underline{\theta})^T \underline{\underline{R}}_b^{-1}(\underline{y} - \underline{\underline{H}}\,\underline{\theta})\right\}$$

et le MV est obtenu par la minimisation de l'opposé de l'argument de l'exponentielle

MCP avec $\underline{\underline{M}} = \underline{\underline{R}}_b^{-1}$ $\qquad \boxed{(\underline{\underline{H}}^T\underline{\underline{R}}_b^{-1}\underline{\underline{H}})\,\hat{\underline{\theta}} = \underline{\underline{H}}^T\underline{\underline{R}}_b^{-1}\underline{y}}$

3) MAP

$$p(\underline{\theta}|\underline{y}) \propto p_{\theta}(\underline{y}|\underline{\theta}) \cdot p_{\theta}(\underline{\theta}) \qquad p_{\theta} = \mathcal{N}(\underline{0}, \underline{\underline{R}}_{\theta})$$

$$\propto \exp\left[-\frac{1}{2}(\underline{y} - \underline{\underline{H}}\,\underline{\theta})^T\underline{\underline{R}}_b^{-1}(\underline{y} - \underline{\underline{H}}\,\underline{\theta}) - \frac{1}{2}\underline{\theta}^T\underline{\underline{R}}_{\theta}^{-1}\underline{\theta}\right]$$

ce qui fournit

$$\boxed{(\underline{\underline{H}}^T\underline{\underline{R}}_b^{-1}\underline{\underline{H}} + \underline{\underline{R}}_{\theta}^{-1})\,\hat{\underline{\theta}} = \underline{\underline{H}}^T\underline{\underline{R}}_b^{-1}\underline{y}}$$

4) EQMM

On a vu que $\underline{\hat{\theta}} = \underline{\underline{R}}_{\theta Y} \underline{\underline{R}}_{YY}^{-1} \underline{y}$

Ici $\begin{cases} \underline{\underline{R}}_{YY} = \underline{\underline{H}}\,\underline{\underline{R}}_\theta \underline{\underline{H}}^T + \underline{\underline{R}}_B \\ \underline{\underline{R}}_{\theta Y} = \underline{\underline{R}}_\theta H^T \end{cases}$

$$\underline{\hat{\theta}} = \underline{\underline{R}}_\theta \underline{\underline{H}}^T (H R_\theta H^T + R_B)^{-1} \underline{y}$$

En utilisant le lemme d'inversion matricielle

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1} DA^{-1},$$

il vient

$$\underline{\hat{\theta}}_{EQMM} = (\underline{\underline{H}}^T \underline{\underline{R}}_B^{-1} \underline{\underline{H}} + \underline{\underline{R}}_\theta^{-1})^{-1} \underline{\underline{H}}^T \underline{\underline{R}}_B\, y$$

Qui est identique à l'estimateur MAP.

Tous ces estimateurs présentent donc la même structure - NB ceci est « fortuit » et vrai uniquement dans le cas linéaire gaussien. Dans le cas général, les estimateurs peuvent être très différents.

Cas linéaire gaussien - MAP et MP.

Reprenons maintenant plus en détail le cas linéaire gaussien. On a ainsi

$$\underline{y} = \underline{\underline{H}}\,\underline{\theta} + \underline{b},$$

avec $\underline{b}\ \mathcal{N}(0, \underline{\underline{R}}_B)$

$\underline{\theta}\ \mathcal{N}(\underline{m}_\theta, \underline{\underline{R}}_\theta)$

On a alors $\quad p_{\underline{\theta}|Y}(\underline{\theta}|\underline{y}) \propto p_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})\, p_{\underline{\theta}}(\underline{\theta})$

avec $\quad p_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta}) = p_B(\underline{y} - \underline{\underline{H}}\,\underline{\theta})$

Il vient donc

$$p_{\underline{\theta}|\underline{Y}}(\underline{\theta}|\underline{y}) \propto \exp\left\{ -\frac{1}{2} (\underline{y} - \underline{\underline{H}}\,\underline{\theta})^T \underline{\underline{R}}_B^{-1} (\underline{y} - \underline{\underline{H}}\,\underline{\theta}) - \frac{1}{2} \underbrace{\underline{\theta}^T \underline{\underline{R}}_\theta \underline{\theta}}_{(\underline{\theta}-\underline{m}_\theta)^T R_B^{-1} (\underline{\theta}-\underline{m}_\theta)} \right\}$$

L'argument de l'exponentielle est une forme quadratique en $\underline{\theta}$ qui peut donc se mettre sous la forme

$$p_{\Theta|Y}(\theta|y) \propto \exp\left\{-\frac{1}{2}(\theta - m_{\theta|Y})^T R_{\theta|Y}^{-1}(\theta - m_{\theta|Y}) + K\right\}$$

En identifiant, on trouve

$$\begin{cases} m_{\theta|Y} = (H^T R_B^{-1} H + R_\theta)^{-1}(H^T R_B^{-1} y + R_\theta^{-1} m_\theta) \\ R_{\theta|Y}^{-1} = (H^T R_B^{-1} H + R_\theta^{-1}) \end{cases}$$

En utilisant le lemme d'inversion matricielle,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

les relations précédentes deviennent

$$\begin{cases} m_{\theta|Y} = m_\theta + R_\theta H^T(H R_\theta H^T + R_B)^{-1}\underbrace{(y - H m_\theta)}_{\text{erreur a priori}} \\ R_{\theta|Y} = R_\theta - R_\theta H^T(H R_\theta H^T + R_B)^{-1} H R_\theta \end{cases}$$

Comme la loi a posteriori est gaussienne, la moyenne a posteriori et le MAP sont confondus, et on a

$$\hat{\theta} = \hat{\theta}_{MAP} = \hat{\theta}_{MP} = m_{\theta|Y}$$

La procédure met donc à jour un a priori de moyenne m_θ et de covariance R_θ en une loi a posteriori de moyenne $\hat{\theta}$ et de covariance $R_{\hat{\theta}} = R_{\theta|Y}$.

On peut dès lors imaginer une procédure itérative $y^{(n)} \to y^{(n+1)}$: l'a priori est constitué par une loi normale $\mathcal{N}(\hat{\theta}_n, R_{\hat{\theta}}^{(n)})$ qui est utilisé pour fournir $\mathcal{N}(\hat{\theta}_{n+1}, R_{\hat{\theta}}^{(n+1)})$

$$\begin{cases} \hat{\theta}_{n+1} = \hat{\theta}_n + R^{(n)} H^T(H R^{n} H^T + R_B)^{-1}(y^{(n+1)} - H\hat{\theta}_n) \\ R^{(n+1)} = R^{(n)} - R^{(n)} H^T(H R^{(n)} H^T + R_B)^{-1} H R^{(n)} \end{cases}$$

On obtient un algorithme adaptatif analogue au filtrage de Kalman / Moindres carrés récursifs.

Borne de Cramér-Rao

La moyenne quadratique de l'erreur d'estimation pour tout estimateur possède une borne inférieure, la borne de Cramér-Rao, qui définit ainsi la qualité de n'importe (toute) procédure d'estimation.

On a $E\left[(\hat{\underline{\theta}}-\underline{\theta})(\hat{\underline{\theta}}-\underline{\theta})^T\right] = \mathrm{Cov}\{\hat{\underline{\theta}}\} + \underline{B}(\hat{\underline{\theta}})\, \underline{B}(\hat{\underline{\theta}})^T$

Pour simplifier, limitons nous ici au cas scalaire, et considérons un paramètre aléatoire θ et un estimateur $\hat{\theta}$ éventuellement biaisé, avec $B(\hat{\theta}) = E_{\underline{y}\theta}\left[\hat{\theta}-\theta\right]$.

On notera $B_{\underline{y}|\theta}(\hat{\theta}) = E_{\underline{y}|\theta}\left[\hat{\theta}-\theta\right]$ le biais conditionnel, avec bien sûr $B[\hat{\theta}] = E_{\theta}\left[B_{\underline{y}|\theta}[\hat{\theta}]\right]$.

Résultats préalables

On note que $\dfrac{dp_{y|\theta}}{d\theta} = p_{y|\theta}\dfrac{d\log p_{y|\theta}}{d\theta}$

On sait que $\int p_{y|\theta}(y|\theta)\,dy = 1$. Ainsi en dérivant par rapport à θ,

$$\int \frac{dp_{y|\theta}}{d\theta}\,dy = \int p_{y|\theta}\frac{d\log p_{y|\theta}}{d\theta}\,dy = 0$$

En dérivant à nouveau, on a

$$\int \underbrace{\frac{dp_{y|\theta}}{d\theta}}_{= p_{y|\theta}\frac{d\log p_{y|\theta}}{d\theta}}\frac{d\log p_{y|\theta}}{d\theta}\,dy + \int p_{y|\theta}\cdot\frac{d^2\log p_{y|\theta}}{d\theta^2}\,dy = 0$$

Par conséquent, $E\left[\left(\dfrac{d\log p_{y|\theta}}{d\theta}\right)^2\right] = -E\left[\dfrac{d^2\log p_{y|\theta}}{d\theta^2}\right]$.

Ces deux résultats sont également valables pour la loi conjointe $p_{y,\theta}$.

Finalement, il est instructif de considérer deux cas : θ certain et θ aléatoire. Débutons par le cas certain

θ certain

Le biais est défini par $B(\hat{\theta}) = E_{Y|\theta}[\hat{\theta} - \theta] = \int (\hat{\theta} - \theta) p_{Y|\theta}(y|\theta)\, dy$

Considérons la dérivée du biais :

$$\frac{dB(\hat{\theta})}{d\theta} = -\underbrace{\int p_{Y|\theta}(y|\theta)\, dy}_{1} + \int \underbrace{\frac{dp_{Y|\theta}}{d\theta}}_{p_{Y|\theta} \frac{d \log p_{Y|\theta}}{d\theta}} (\hat{\theta} - \theta)\, dy$$

Alors, $$\frac{dB(\hat{\theta})}{d\theta} = -1 + \int p_{Y|\theta}(y|\theta) \cdot (\hat{\theta} - \theta) \frac{d \log p_{Y|\theta}(y|\theta)}{d\theta}\, dy$$

Appliquons l'inégalité de Schwartz

$$|\langle X, Y \rangle|^2 \leq \langle X, X \rangle \langle Y, Y \rangle \quad \text{avec égalité si } Y = kX$$

avec $\begin{cases} X = \sqrt{p_{Y|\theta}}\, (\theta - \hat{\theta}) \\ Y = \sqrt{p_{Y|\theta}}\, \dfrac{d \log p_{Y|\theta}}{d\theta} \end{cases}$

Celà fournit : $$\left(1 + \frac{dB(\hat{\theta})}{d\theta}\right)^2 \leq E[(\hat{\theta} - \theta)^2] \cdot E\left[\left(\frac{d \log p_{Y|\theta}}{d\theta}\right)^2\right]$$

et finalement

$$\boxed{E[(\hat{\theta} - \theta)^2] \geqslant \frac{\left(1 + \frac{dB(\hat{\theta})}{d\theta}\right)^2}{E\left[\left(\frac{d \log p_{Y|\theta}}{d\theta}\right)^2\right]} = \frac{\left(1 + \frac{dB(\hat{\theta})}{d\theta}\right)^2}{-E\left[\frac{d^2 \log p_{Y|\theta}}{d\theta}\right]}}$$

Le dénominateur,

$$E\left[\left(\frac{d \log p_{Y|\theta}}{d\theta}\right)^2\right] = -E\left[\frac{d^2 \log p_{Y|\theta}}{d\theta^2}\right] = J_\theta$$

est appelé <u>information de Fisher</u>.

Lorsque l'estimateur est non biaisé, il reste simplement

$$\mathrm{Var}[\hat{\theta}] \geqslant \frac{1}{J_\theta}.$$

Efficacité

Dans le cas sans biais, un estimateur est dit <u>efficace</u> s'il atteint la borne de Cramér-Rao.

Dans l'inégalité de Cauchy Schwartz, l'égalité est atteinte pour $X = kY$ (vecteurs proportionnels), c'est à dire

$$\frac{d \log p_{y|\theta}}{d\theta} = k(\theta)(\hat{\theta} - \theta) \quad \text{pour tout } \underset{\approx}{y} \text{ et } \theta.$$

Lorsque $p_{y|\theta}$ ne peut se mettre sous la forme précédente (en ajustant $\hat{\theta}$), il n'existe pas d'estimateur efficient.

Lorsqu'on choisit l'estimateur du maximum de vraisemblance, on a

$$\left.\frac{d \log p_{y|\theta}}{d\theta}\right|_{\theta = \hat{\theta}_{MV}} = 0 = k(\theta)(\hat{\theta} - \theta)\Big|_{\theta = \hat{\theta}_{MV}}$$

et on trouve donc $\hat{\theta} = \hat{\theta}_{MV}$ si l'estimateur est efficient.

Si un estimateur efficient existe, c'est l'estimateur du maximum de vraisemblance.

Les propriétés asymptotiques du MV sont les suivantes :

1. le MV est non biaisé asymptotiquement (ou non biaisé tout court)
2. le MV est consistant (converge vers la vraie valeur)
3. le MV est asymptotiquement efficient
4. $\hat{\theta}_{MV}$ est asymptotiquement gaussien.

De belles propriétés, mais asymptotiques. Il peut être possible de trouver d'autres estimateurs qui présenteraient une EQM, ou une variance plus faible, à nombre de données fini.

<u>cas θ aléatoire</u>

Le biais conditionnel est $B_{Y|\theta}(\hat{\theta}) = \int (\hat{\theta}-\theta)\, p_{Y|\theta}\, dy$

On considère $p(\theta) B_{Y|\theta}(\hat{\theta})$ et on fait l'hypothèse que

$\lim_{|\theta|\to\infty} p(\theta) B_{Y|\theta}(\hat{\theta}) = 0$. Alors

$$\frac{d\, p(\theta) B_{Y|\theta}(\hat{\theta})}{d\theta} = -\int p_{Y,\theta}(y,\theta)\, dy + \int \frac{d\, p_{Y,\theta}(y,\theta)}{d\theta} (\hat{\theta}-\theta)\, dy$$

Puis on intègre par rapport à θ :

$$\underbrace{\Big[\, p(\theta) B_{Y|\theta}(\hat{\theta}) \Big]_{-\infty}^{+\infty}}_{0} = \underbrace{-\iint p_{Y,\theta}\, dy\, d\theta}_{-1} + \iint p_{Y,\theta} \frac{d \log p_{Y,\theta}}{d\theta} (\hat{\theta}-\theta) \times dy\, d\theta$$

Il vient donc $\quad 1 = \iint p_{Y,\theta} \dfrac{d \log p_{Y,\theta}}{d\theta} (\theta-\hat{\theta})\, dy\, d\theta$

et on applique l'inégalité de Schwartz comme précédemment, ce qui mène à

$$\boxed{E_{Y,\theta}\left[(\hat{\theta}-\theta)^2\right] \geqslant \frac{1}{E_{Y,\theta}\left[\left(\dfrac{d \log p_{Y,\theta}}{d\theta}\right)^2\right]}}$$

où l'on note que cette fois ci ce sont les lois conjointes et non plus conditionnelles qui interviennent dans la borne.

L'égalité est atteinte lorsque $\quad \dfrac{d \log p_{Y,\theta}}{d\theta} = k\,(\hat{\theta}-\theta)$

où la constante k ne dépend pas de θ (à cause de l'intégration sur θ).

De manière équivalente, $\quad \dfrac{d \log p_{\theta|Y}}{d\theta} = k(\hat{\theta}-\theta)$,

car $p_{Y,\theta} = p_{\theta|Y}\, p_Y$ et $\dfrac{d \log p_Y}{d\theta} = 0$

En dérivant à nouveau par rapport à θ on a

$$\frac{d^2 \log p_{\theta|Y}}{d\theta^2} = -k$$

En intégrant 2 fois et en prenant l'exponentielle, il vient alors

$$P_{\theta|y}(\theta|y) = \exp\left\{ -k\theta^2 + C_1\theta + C_2 \right\},$$

c'est-à-dire une loi gaussienne.

Pour qu'un estimateur atteigne la borne, il faut donc <u>nécessairement</u> que <u>la loi a posteriori soit gaussienne</u>.

À partir de la condition $\dfrac{d \log P_{\theta|y}}{d\theta} = k(\hat{\theta} - \theta)$,

on constate que $\hat{\theta}_{MAP}$ atteint la borne, puisque $\left.\dfrac{d \log P_{\theta|y}}{d\theta}\right|_{\theta = \hat{\theta}_{MAP}} = k(\hat{\theta} - \hat{\theta}_{MAP}) = 0$

En d'autres termes, si un estimateur atteint la borne, c'est le MAP. Et évidemment, comme le MMSE est l'estimateur présentant l'erreur quadratique minimale, on a alors $\hat{\theta}_{MMSE} = \hat{\theta}_{MAP} = \theta_{MP}$.