❑ Méthode des moments : un exemple simple

Soit une variable aléatoire Y distribuée suivant une loi Gamma(θ, k)

On souhaite estimer θ et k à partir de N réalisations indép. $\{y_n\}_{n=1\cdots N}$ de Y.

① Loi de la variable $Y|\boldsymbol{\theta}$:

Loi Gamma de densité de probabilité $f_{Y|\theta,k}(y|\theta,k) = \dfrac{y^{k-1}e^{-y/\theta}}{\theta^k\Gamma(k)}, \forall y > 0$

② Choix des moments : on sait (`wikipedia`) que pour une loi Gamma

$$m_Y = \mathrm{E}_{Y|\theta,k}\{Y|\theta,k\} = k\cdot\theta \quad \text{et} \quad \sigma_Y^2 = \mathrm{E}_{Y|\theta,k}\{(Y-m_Y)^2|\theta,k\} = k\cdot\theta^2$$

③ Calcul des estimateurs empiriques des moments à partir des données $\underline{\boldsymbol{y}}$

$$\text{Moyenne } \overset{o}{m}_Y = \frac{1}{N}\sum_{n=1}^{N} y_n \text{ et variance } \overset{o}{\sigma}{}_Y^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \overset{o}{m}_Y)^2$$

④ Résolution du système d'équations (non linéaire en $\overset{o}{\theta}$ et $\overset{o}{k}$) :

$$\left\{\begin{array}{l} \overset{o}{m}_Y = \overset{o}{k}\cdot\overset{o}{\theta} \\ \overset{o}{\sigma}{}_Y^2 = \overset{o}{k}\cdot\overset{o}{\theta}{}^2 \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} \overset{o}{\theta} = \dfrac{\overset{o}{\sigma}{}_Y^2}{\overset{o}{m}_Y} \\ \overset{o}{k} = \dfrac{\overset{o}{m}{}_Y^2}{\overset{o}{\sigma}{}_Y^2} \end{array}\right.$$

❑ Critères quadratiques de fidélité aux données

- Modèle de données dépendant des paramètres $\underline{\boldsymbol{\theta}}$: $y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}})$

$\Rightarrow$ Calculer $\underline{\boldsymbol{\theta}}$ tel que les $y_{\mathcal{M}}(n,\underline{\boldsymbol{\theta}})$ soient « le plus fidèle aux données ! »

- Estimateur $\overset{o}{\underline{\boldsymbol{\theta}}}$ minimisant la distance quadratique :

$$\overset{o}{\underline{\boldsymbol{\theta}}} = \arg\min_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) \text{ avec } J(\underline{\boldsymbol{\theta}}) = \sum_{n=1}^{N}\left(y[n] - y_{\mathcal{M}}(n,\underline{\boldsymbol{\theta}})\right)^2 = ||\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})||^2$$

- Forme pondérée (poid w_n associé à la confiance en la donnée $y[n]$) :

$$J(\underline{\boldsymbol{\theta}}) = \sum_{n=1}^{N} w_n\left(y[n] - y_{\mathcal{M}}(n,\underline{\boldsymbol{\theta}})\right)^2$$

- Forme matricielle :

$$J(\underline{\boldsymbol{\theta}}) = \left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right)^T \mathbf{Q}\left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right) = ||\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})||^2_{\mathbf{Q}}$$

Matrice de pondération $\mathbf{Q}$, symétrique, définie positive *i.e.* $\forall\underline{\boldsymbol{u}} \neq 0,\ \underline{\boldsymbol{u}}^H\mathbf{Q}\underline{\boldsymbol{u}} > 0$

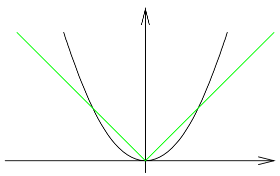
❑ Critères robustes (aux données aberrantes)

- Estimateur $\overset{o}{\underline{\boldsymbol{\theta}}}$ minimisant la distance non quadratique :

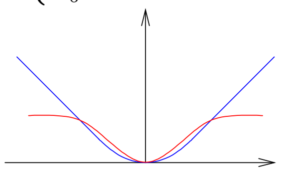
$$\overset{o}{\underline{\boldsymbol{\theta}}} = \arg\min_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) \text{ avec } J(\underline{\boldsymbol{\theta}}) = \sum_{n=1}^{N}\phi\left(y[n] - y_{\mathcal{M}}(n,\underline{\boldsymbol{\theta}})\right)$$

$\phi(x) \nearrow$ moins vite que $x^2 \Rightarrow$ moins sensible aux données aberrantes

- Critère en valeur absolue (L_1) : $\phi(x) = |x|$
- Critères L_2-L_1 : $\phi(x) = \sqrt{s^2+x^2}$, $\phi(x) = \begin{cases} \frac{1}{2}x^2 & \text{si } |x| < s \\ s|x| - \frac{1}{2}s^2 & \text{sinon.}\end{cases}$
- Critères L_2-L_0 : $\phi(x) = \frac{x^2}{s^2+x^2}$, $\phi(x) = \begin{cases} \frac{1}{2}(x^2 - \frac{x^4}{s^2} + \frac{x^6}{3s^4}) & \text{si } |x| < s \\ \frac{s^2}{6} & \text{sinon.}\end{cases}$



L_2 et L_1



L_2-L_1 et L_2-L_0

III.2 Estimateur du maximum de vraisemblance

Estimation d'après la loi des observations $\underline{\boldsymbol{y}}$ sachant $\underline{\boldsymbol{\theta}}$

❑ Définition

- Fonction de vraisemblance (*likelihood*) : loi des observations $\underline{\boldsymbol{y}}$ fonction de $\underline{\boldsymbol{\theta}}$

$$L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = \begin{cases} f_{Y|\theta}(\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}) \text{ si } \underline{\boldsymbol{y}} \text{ à valeurs continues } (\underline{\boldsymbol{y}} \in \mathbb{R}^N) \\ P_{Y|\theta}[\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}] \text{ si } \underline{\boldsymbol{y}} \text{ à valeurs discrètes } (\underline{\boldsymbol{y}} \in \mathbb{Z}^N)\end{cases}$$

- Anti-log-vraisemblance (*neg-log-likelihood*) : $NLL(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = -\ln L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}})$
- Estimateur du maximum de vraisemblance (MV) :

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{MV}} = \arg\max_{\underline{\boldsymbol{\theta}}} L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = \arg\min_{\underline{\boldsymbol{\theta}}} NLL(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}})$$

❑ Démarche pour le calcul de $\overset{o}{\underline{\boldsymbol{\theta}}}_{MV}$:

① Déduire des hypothèses la loi des observations $\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}} \Rightarrow$ Vraisemblance $L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}})$

② Calculer l'anti-log-vraisemblance $NLL(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = -\ln L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}})$

③ Calculer la valeur des paramètres minimisant l'anti-log-vraisemblance

❑ Cas des perturbations additives indépendantes gaussiennes centrées

$$\text{Perturbations } \epsilon[n] \begin{cases} \text{additives : } y[n] = y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}) + \epsilon[n] \\ \text{indépendantes : } f_E(\underline{\boldsymbol{\epsilon}}) = \prod_{n=1}^N f_{E_n}(\epsilon[n]) \\ \text{gaussiennes centrées : } \epsilon[n] \sim \mathcal{N}(0, \sigma_n^2) \end{cases}$$

① Vraisemblance

→ Les $y[n]$ (sachant $\underline{\boldsymbol{\theta}}$) sont indépendants ⇒ $f_{Y|\theta}(\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}) = \prod_{n=1}^N f_{Y_n}(y[n]|\underline{\boldsymbol{\theta}})$

→ D'après la formule de changement de variable page 16

$$f_{Y_n}(y[n]|\underline{\boldsymbol{\theta}}) = f_{E_n}(y[n] - y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}))$$

→ Les $\epsilon[n]$ sont gaussiens centrés : $f_{E_n}(\epsilon[n]) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp(-\frac{1}{2\sigma_n^2}\epsilon[n]^2)$

$$\Rightarrow L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma_n} \exp(-\frac{1}{2\sigma_n^2}(y[n] - y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}))^2)$$

② Anti-log-vraisemblance

$$\Rightarrow NLL(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = \frac{N}{2}\log(2\pi) + \sum_{n=1}^N \log(\sigma_n) + \sum_{n=1}^N \frac{1}{2\sigma_n^2}(y[n] - y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}))^2$$

③ Calculer $\underline{\boldsymbol{\theta}}$ minimisant la NLL : ici distance quadratique pondérée

$$J(\underline{\boldsymbol{\theta}}) = \sum_{n=1}^N \frac{1}{2\sigma_n^2}(y[n] - y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}))^2 = ||\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})||^2_{\mathbf{Q}} \text{ avec } \mathbf{Q} = \mathrm{diag}\{\sigma_1^2, \ldots \sigma_N^2\}$$

❑ Cas des perturbations additives non indépendantes

- Cas de modèle de perturbation corrélées de type filtrage connu :

  ⇒ Erreur de prédiction indépendante !

- Cas de loi de probabilité prenant en compte la corrélation :

  → Cas additif gaussien centré $f_E(\underline{\boldsymbol{\epsilon}}) \propto \mathcal{N}(\underline{\mathbf{0}}, \mathbf{\Gamma}_E)$: $\underline{\boldsymbol{y}} = \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) + \underline{\boldsymbol{\epsilon}}$

① Vraisemblance : d'après la formule de changement de variable page 23

$$L(\underline{\boldsymbol{\theta}}\,;\underline{\boldsymbol{y}}) = f(\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{N/2}(\det \mathbf{\Gamma}_E)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Big(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\Big)^T \mathbf{\Gamma}_E^{-1}\Big(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\Big)\right)$$

$$③\ \overset{o}{\underline{\boldsymbol{\theta}}}_{MV} = \arg\min_{\underline{\boldsymbol{\theta}}} \Big(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\Big)^T \mathbf{\Gamma}_E^{-1}\Big(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\Big) = \arg\min_{\underline{\boldsymbol{\theta}}} ||\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})||_{\mathbf{\Gamma}_E^{-1}}$$

❑ Propriétés de l'estimateur du maximum de vraisemblance

- Invariant par reparamétrisation : soit $\underline{\boldsymbol{\beta}} = \boldsymbol{g}(\underline{\boldsymbol{\theta}})$ alors $\overset{o}{\underline{\boldsymbol{\beta}}}_{MV} = \boldsymbol{g}(\overset{o}{\underline{\boldsymbol{\theta}}}_{MV})$
- Asymptotiquement ($N \to \infty$) :

  $\overset{o}{\underline{\boldsymbol{\theta}}}_{MV}$ est distribué suivant la loi $\mathcal{N}(\underline{\boldsymbol{\theta}}, \mathbf{F}(\underline{\boldsymbol{\theta}})^{-1})$: $\begin{cases} \text{non biaisé} \\ \text{variance minimale} \end{cases}$

III.3 Estimateurs bayésiens

Prise en compte d'information *a priori* sur les paramètres

❑ Règle de Bayes

$f_\theta(\underline{\boldsymbol{\theta}})$: loi *a priori* sur $\underline{\boldsymbol{\theta}}$, résume les informations disponibles sur les paramètres *avant* la mesure

$$\overbrace{f_{\theta|Y}(\underline{\boldsymbol{\theta}}\,|\,\underline{\boldsymbol{y}})}^{a\ posteriori} = \frac{f_{Y|\theta}(\underline{\boldsymbol{y}}\,|\,\underline{\boldsymbol{\theta}}) f_\theta(\underline{\boldsymbol{\theta}})}{f_Y(\underline{\boldsymbol{y}})} = \underbrace{f_{Y|\theta}(\underline{\boldsymbol{y}}\,|\,\underline{\boldsymbol{\theta}})}_{\text{vraisemblance}} \overbrace{f_\theta(\underline{\boldsymbol{\theta}})}^{a\ priori} \Big/ \underbrace{f_Y(\underline{\boldsymbol{y}})}_{\text{cte de normalisation}}$$

$f_{\theta|Y}(\underline{\boldsymbol{\theta}}\,|\,\underline{\boldsymbol{y}})$: loi *a posteriori* sur $\underline{\boldsymbol{\theta}}$, résume les informations disponibles sur les paramètres *après* la mesure

❑ Décision $f_{\theta|Y}(\underline{\boldsymbol{\theta}}\,|\,\underline{\boldsymbol{y}}) \overset{?}{\rightsquigarrow} \overset{o}{\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{y}})$ fonction de coût $\mathcal{J}(\overset{o}{\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{y}})\,|\,\underline{\boldsymbol{\theta}})$

⇒ minimisation d'un coût moyen sur $\underline{\boldsymbol{\theta}}$ et $\underline{\boldsymbol{y}}$

❑ Minimisation d'EQM et principe d'orthogonalité

Soit $\mathcal{H}(\{y[n]\}_{n=1\ldots N})$ l'espace vectoriel engendré par toutes les fonctions des variables aléatoires $\{y[n]\}_{n=1\ldots N}$.

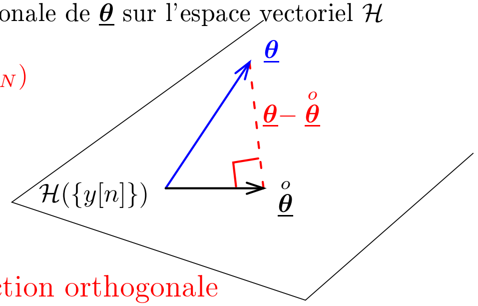
Soit le produit scalaire sur $\mathcal{H}$ défini par $<\underline{\boldsymbol{u}}, \underline{\boldsymbol{v}}> = \mathrm{E}\{\underline{\boldsymbol{u}}^T\underline{\boldsymbol{v}}\}$ et la norme $||\underline{\boldsymbol{u}}||^2 = <\underline{\boldsymbol{u}}, \underline{\boldsymbol{u}}>$ associée.

L'élément $\overset{o}{\underline{\boldsymbol{\theta}}}$ de $\mathcal{H}$ minimisant l'Erreur Quadratique Moyenne (EQM) :

$$\overset{o}{\underline{\boldsymbol{\theta}}} = \arg\min_{\underline{\boldsymbol{\beta}} \in \mathcal{H}} ||\underline{\boldsymbol{\beta}} - \underline{\boldsymbol{\theta}}||^2$$

correspond à la projection orthogonale de $\underline{\boldsymbol{\theta}}$ sur l'espace vectoriel $\mathcal{H}$

On a donc $\underline{\boldsymbol{\theta}} - \overset{o}{\underline{\boldsymbol{\theta}}} \perp \mathcal{H}(\{y[n]\}_{n=1\ldots N})$



Estimation EQM ⇔ projection orthogonale

❏ Estimateur en moyenne quadratique : minimise le *coût quadratique moyen*

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{EMQ}} = \arg\min_{\overset{o}{\underline{\boldsymbol{\theta}}}} E_{Y,\theta}\{(\overset{o}{\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{y}}) - \underline{\boldsymbol{\theta}})^T(\overset{o}{\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{y}}) - \underline{\boldsymbol{\theta}})\}$$

$\Rightarrow$ Moyenne *a posteriori* : $\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{EMQ}} = \overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{PM}} = E_{\theta|Y}\{\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}}\}$

(*posterior mean* PM)

Calcul d'une espérance $\Rightarrow$ Intégration

- C'est la projection orthogonale de $\underline{\boldsymbol{\theta}}$ sur l'*e.v.* des fonctions de $\underline{\boldsymbol{y}}$
- Non biaisée en moyenne sur $\underline{\boldsymbol{\theta}}$

❏ Estimateur linéaire en moyenne quadratique

$\overset{o}{\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{y}}) = \mathbf{A}\underline{\boldsymbol{y}} + \underline{\boldsymbol{b}}$ qui minimise le coût quadratique moyen

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{ELMQ}} = \mathbf{\Gamma}_{\underline{\boldsymbol{\theta}},\underline{\boldsymbol{y}}}\mathbf{\Gamma}_{\underline{\boldsymbol{y}}}^{-1}(\underline{\boldsymbol{y}} - \underline{\boldsymbol{m}}_{\underline{\boldsymbol{y}}}) + \underline{\boldsymbol{m}}_{\underline{\boldsymbol{\theta}}}$$

$\Rightarrow$ Estimation des moyennes et corrélations

❏ Maximum *a posteriori* : minimise un *coût moyen* de type tout ou rien

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{MAP}} = \arg\max_{\underline{\boldsymbol{\theta}}} f(\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}}) = \arg\min_{\underline{\boldsymbol{\theta}}} -\ln(f(\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}}))$$

$\Rightarrow$ Maximisation de la loi a posteriori

❏ Lien avec le maximum de vraisemblance :

- $\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{ML}} = \overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{MAP}}$ pour une loi *a priori* uniforme sur un intervalle

  Si $\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{ML}}$ appartient à l'intervalle de définition !

- $\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{MAP}} = \arg\max_{\underline{\boldsymbol{\theta}}} f(\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}}) = \arg\min_{\underline{\boldsymbol{\theta}}}(NLL(\underline{\boldsymbol{y}} \,|\, \underline{\boldsymbol{\theta}}) - \ln f(\underline{\boldsymbol{\theta}}))$

  $\Rightarrow$ Minimisation de l'anti-log-vraisemblance pénalisée

❏ Maximum *a posteriori* Marginal :

$$\overset{o}{\theta}_{m\mathrm{MAPM}} = \arg\max_{\theta_m} f(\theta_m \,|\, \underline{\boldsymbol{y}}), \ 1 \leqslant m \leqslant M$$

❏ Démarche pour calculer un estimateur bayésien de $\overset{o}{\underline{\boldsymbol{\theta}}}$:

1. Déduire des hypothèses la loi des observations $f(\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}})$ (vraisemblance)
2. Déduire des hypothèses la loi a priori sur les paramètres $f(\underline{\boldsymbol{\theta}})$
3. Par la règle de Bayes, en déduire la loi a posteriori $f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}})$

   $\Rightarrow$ Retrouve-t-on la forme d'une loi « classique » ?
4. Calculer l'estimateur

   $\rightarrow$ Si $f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}})$ est une loi « classique » : connaît-on $\left\{\begin{array}{l}\text{sa moyenne } \overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{PM}} \,? \\ \text{son maximum } \overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{MAP}} \,?\end{array}\right.$

   $\rightarrow$ Estimateur du Maximum a posteriori $\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{MAP}}$

   $\Rightarrow$ Maximiser $f(\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}})$ ou minimiser $-\ln(f(\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}}))$

   $\rightarrow$ Estimateur de la moyenne a posteriori $\overset{o}{\underline{\boldsymbol{\theta}}}_{\mathrm{PM}}$

   $\Rightarrow$ Calculer $E_{\theta|Y}\{\underline{\boldsymbol{\theta}} \,|\, \underline{\boldsymbol{y}}\}$ (intégrale)

❏ Cas perturbations additives gaussiennes + *a priori* gaussien :

$\underline{\boldsymbol{\epsilon}}$ additif gaussien : $\mathcal{N}(\underline{\mathbf{0}}, \mathbf{\Gamma}_E)$; $\underline{\boldsymbol{\theta}}$ gaussien : $\mathcal{N}(\underline{\boldsymbol{\theta}}_0, \mathbf{\Gamma}_0)$

1. Vraisemblance :

$$f(\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{N/2}(\det \mathbf{\Gamma}_E)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right)^T \mathbf{\Gamma}_E^{-1}\left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right)\right)$$

2. Loi a priori :

$$f(\underline{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{M/2}(\det \mathbf{\Gamma}_0)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}_0)^T \mathbf{\Gamma}_0^{-1}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}_0)\right)$$

3. Loi a posteriori $f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}})$

$$f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}}) \propto \exp\left(-\frac{1}{2}\left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right)^T \mathbf{\Gamma}_E^{-1}\left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right) - \frac{1}{2}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}_0)^T \mathbf{\Gamma}_0^{-1}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}_0)\right)$$

4. Estimateur : $f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}})$ n'est pas une loi classique

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{MAP} = \arg\min_{\underline{\boldsymbol{\theta}}} \left(\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right)^T \mathbf{\Gamma}_E^{-1}\left((\underline{\boldsymbol{y}} - \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}})\right) + (\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}_0)^T \mathbf{\Gamma}_0^{-1}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}_0)$$

$\overset{o}{\underline{\boldsymbol{\theta}}}_{PM} = ?$

❑ Même cas + Modèle linéaire en les paramètres :

$\underline{\boldsymbol{\epsilon}}$ additif gaussien : $\mathcal{N}(\underline{\mathbf{0}},\mathbf{\Gamma}_E)$; $\underline{\boldsymbol{\theta}}$ gaussien : $\mathcal{N}(\underline{\boldsymbol{\theta}}_0,\mathbf{\Gamma}_0)$

Modèle linéaire en les paramètres : $\underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}\underline{\boldsymbol{\theta}}$

3 Loi a posteriori $f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}})$

$$f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}}) \propto \exp\left(\underbrace{-\frac{1}{2}\left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right)^T\mathbf{\Gamma}_E^{-1}\left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right)-\frac{1}{2}\left(\underline{\boldsymbol{\theta}}-\underline{\boldsymbol{\theta}}_0\right)^T\mathbf{\Gamma}_0^{-1}\left(\underline{\boldsymbol{\theta}}-\underline{\boldsymbol{\theta}}_0\right)}_{\text{Terme quadratique en } \underline{\boldsymbol{\theta}}}\right)$$

On peut montrer (grace à la propriété 1. de l'exercice 1 du TD4) que :

$$f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}}) \propto \exp\left(\overbrace{-\frac{1}{2}\left(\underline{\boldsymbol{\theta}}-\underline{\boldsymbol{m}}_\theta\right)^T\mathbf{\Gamma}_\theta^{-1}\left(\underline{\boldsymbol{\theta}}-\underline{\boldsymbol{m}}_\theta\right)}^{\text{Terme quadratique en } \underline{\boldsymbol{\theta}}}\right) \quad \text{avec} \quad \begin{cases} \mathbf{\Gamma}_\theta = \left(\mathbf{R}^T\mathbf{\Gamma}_E^{-1}\mathbf{R}+\mathbf{\Gamma}_0^{-1}\right)^{-1} \\ \underline{\boldsymbol{m}}_\theta = \mathbf{\Gamma}_\theta\left(\mathbf{R}^T\mathbf{\Gamma}_E^{-1}\underline{\boldsymbol{y}}+\mathbf{\Gamma}_0\underline{\boldsymbol{\theta}}_0\right) \end{cases}$$

4 Estimateur : $f(\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{y}})$ est une loi Gaussienne $\mathcal{N}(\underline{\boldsymbol{m}}_\theta,\mathbf{\Gamma}_\theta)$

$\overset{o}{\underline{\boldsymbol{\theta}}}_{PM} = \underline{\boldsymbol{m}}_\theta$ (moyenne *a posteriori* $\underline{\boldsymbol{m}}_\theta$)

$\overset{o}{\underline{\boldsymbol{\theta}}}_{MAP} = \underline{\boldsymbol{m}}_\theta$ (Gaussienne maximale en sa moyenne)

CHAP IV CALCUL DES ESTIMATEURS

1/ Introduction à l'optimisation

2/ Modèles LP et critères quadratiques ; les Moindres Carrés

3/ Méthodes dérivées des Moindres Carrés

IV.1 INTRODUCTION À L'OPTIMISATION

❑ Dérivation matricielle :

→ Dérivée de $J(\underline{\boldsymbol{\theta}})$ par rapport à $\underline{\boldsymbol{\theta}} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}$: **gradient** $\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = \begin{bmatrix} \frac{\partial J}{\partial \theta_1}(\underline{\boldsymbol{\theta}}) \\ \vdots \\ \frac{\partial J}{\partial \theta_M}(\underline{\boldsymbol{\theta}}) \end{bmatrix}$

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}^T}(\underline{\boldsymbol{\theta}}) = \left[\frac{\partial J}{\partial \theta_1}(\underline{\boldsymbol{\theta}}), \frac{\partial J}{\partial \theta_2}(\underline{\boldsymbol{\theta}}), \ \dots \ \frac{\partial J}{\partial \theta_M}(\underline{\boldsymbol{\theta}})\right] \quad \text{(vecteurs de dimension } M\text{)}$$

→ Dérivée seconde de $J(\underline{\boldsymbol{\theta}})$ par rapport à $\underline{\boldsymbol{\theta}}$: **Hessien**

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}\mathrm{d}\underline{\boldsymbol{\theta}}^T} = \begin{bmatrix} \frac{\partial^2 J}{\partial^2 \theta_1}(\underline{\boldsymbol{\theta}}) & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_2}(\underline{\boldsymbol{\theta}}) & \cdots & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_M}(\underline{\boldsymbol{\theta}}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 J}{\partial \theta_M \partial \theta_1}(\underline{\boldsymbol{\theta}}) & \frac{\partial^2 J}{\partial \theta_M \partial \theta_2}(\underline{\boldsymbol{\theta}}) & \cdots & \frac{\partial^2 J}{\partial^2 \theta_M}(\underline{\boldsymbol{\theta}}) \end{bmatrix} \quad \text{(Matrice } M \times M\text{)}$$

❑ Exemple : critère quadratique $J(\underline{\boldsymbol{\theta}}) = \left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right)^T\mathbf{Q}\left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right)$

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = -2\mathbf{R}^T\mathbf{Q}\left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right) \quad \text{et} \quad \frac{\mathrm{d}^2J}{\mathrm{d}\underline{\boldsymbol{\theta}}\mathrm{d}\underline{\boldsymbol{\theta}}^T}(\underline{\boldsymbol{\theta}}) = 2\mathbf{R}^T\mathbf{Q}\mathbf{R}$$

❑ Conditions de minimalité

Condition nécessaire de minimalité du premier ordre :

- Si $\underline{\boldsymbol{\theta}} \in \mathcal{O}$ est un minimum local de J et si J est dérivable en $\underline{\boldsymbol{\theta}}$, alors :

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = \underline{\mathbf{0}} \quad \text{(gradient nul)}$$

Conditions de minimalité du second ordre :

- Si $\underline{\boldsymbol{\theta}} \in \mathcal{O}$ est un min. local de J et J est deux fois dérivable en $\underline{\boldsymbol{\theta}}$, alors :

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = \underline{\mathbf{0}} \quad \text{et} \quad \frac{\mathrm{d}^2J}{\mathrm{d}\underline{\boldsymbol{\theta}}\mathrm{d}\underline{\boldsymbol{\theta}}^T}(\underline{\boldsymbol{\theta}}) \text{ est définie positive (Hessien)}$$

(Matrice $\mathbf{A}$ définie positive si $\forall \underline{\boldsymbol{u}},\ \underline{\boldsymbol{u}}^T\mathbf{A}\underline{\boldsymbol{u}} > 0$)

- Si $\underline{\boldsymbol{\theta}} \in \mathcal{O}$ est un point où J est deux fois différentiable et si $\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = \underline{\mathbf{0}}$ et $\frac{\mathrm{d}^2J}{\mathrm{d}\underline{\boldsymbol{\theta}}\mathrm{d}\underline{\boldsymbol{\theta}}^T}(\underline{\boldsymbol{\theta}})$ est définie positive, alors $\underline{\boldsymbol{\theta}}$ est un minimum local strict de J

❑ Exemple : critère quadratique $J(\underline{\boldsymbol{\theta}}) = \left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right)^T\mathbf{Q}\left(\underline{\boldsymbol{y}}-\mathbf{R}\underline{\boldsymbol{\theta}}\right)$

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = \underline{\mathbf{0}} \quad \Leftrightarrow \quad \mathbf{R}^T\mathbf{Q}\underline{\boldsymbol{y}} = \mathbf{R}^T\mathbf{Q}\mathbf{R}\underline{\boldsymbol{\theta}} \quad \Leftrightarrow \quad \underline{\boldsymbol{\theta}} = (\mathbf{R}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}\underline{\boldsymbol{y}}$$

❏ Difficulté des problèmes d'optimisation :

→ Critères quadratique en les paramètres : $J(\underline{\boldsymbol{\theta}}) = \underline{\boldsymbol{\theta}}^T \mathbf{A}\, \underline{\boldsymbol{\theta}} - 2\underline{\boldsymbol{b}}^T \underline{\boldsymbol{\theta}} + c$

Pas un problème d'optimisation ⇒ résoudre un système linéaire !

$$\frac{\mathrm{d}J}{\mathrm{d}\underline{\boldsymbol{\theta}}}(\underline{\boldsymbol{\theta}}) = \underline{\mathbf{0}} \quad \Leftrightarrow \quad \mathbf{A}\underline{\boldsymbol{\theta}} = \underline{\boldsymbol{b}} \quad \Leftrightarrow \quad \underline{\boldsymbol{\theta}} = \mathbf{A}^{-1}\underline{\boldsymbol{b}} \quad \text{(Voir TD4, ex.1)}$$

→ Critères convexes : pas de minima locaux !

⇒ Il suffit d'annuler le gradient ! Le plus simple…

(problème éventuel pour les fonctions non différentiables en tout point !)

→ Optimisation sous contraintes

⇒ Pénalisation intérieure, Pénalisation extérieure, Techniques de Lagrangien

→ Présence de minima locaux : optimisation globale

$\Rightarrow \left\{ \begin{array}{l} \text{Algorithmes stochastiques : recuit simulé, algorithmes génétiques…} \\ \text{Algorithmes déterministes : encadrement par intervalles…} \end{array} \right.$

Pour l'optimisation : voir cours de M2 !

IV.2 Modèles LP et critères quadratiques : *a)* les Moindres Carrés

❏ Modèle linéaire en les paramètres (LP) : $\underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}\underline{\boldsymbol{\theta}}$

→ $y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}})$ modèle idéal (sans bruit) pour la *n*ième donnée, $n \in \{1, 2, \ldots, N\}$

Modèle LP si $y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}})$ est une combinaison linéaire des paramètres :

$$y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}) = r_{n,1}\theta_1 + r_{n,2}\theta_2 + \ldots r_{n,M}\theta_M = \sum_{m=1}^{M} r_{n,m}\theta_m = \underline{\boldsymbol{r}}_n^T \underline{\boldsymbol{\theta}}$$

avec $\underline{\boldsymbol{r}}_n^T = [r_{n,1}, r_{n,2}, \ldots, r_{n,M}]$ vecteur de régression (connu) à l'instant n

→ Modèle $\underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}\underline{\boldsymbol{\theta}}$ avec $\left\{ \begin{array}{l} \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) \text{ vecteur de dimension } N \\ \mathbf{R} \text{ matrice de dimension } N \times M \end{array} \right.$

$$\underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \begin{bmatrix} y_{\mathcal{M}}(1, \underline{\boldsymbol{\theta}}) \\ y_{\mathcal{M}}(2, \underline{\boldsymbol{\theta}}) \\ \vdots \\ y_{\mathcal{M}}(N, \underline{\boldsymbol{\theta}}) \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \underline{\boldsymbol{r}}_1^T \\ \underline{\boldsymbol{r}}_2^T \\ \vdots \\ \underline{\boldsymbol{r}}_N^T \end{bmatrix} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,M} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,M} \\ \vdots & \vdots & \vdots & \\ r_{N,1} & r_{N,2} & \cdots & r_{N,M} \end{bmatrix} \quad \underline{\boldsymbol{\theta}} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}$$

→ Modèle $\underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}\underline{\boldsymbol{\theta}}$ peut aussi s'interpréter comme :

$$\underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}(:,1)\theta_1 + \mathbf{R}(:,2)\theta_2 + \cdots + \mathbf{R}(:,M)\theta_M = \sum_{m=1}^{M} \mathbf{R}(:,m)\theta_m$$

avec $\mathbf{R}(:,m)$ la *m*ième colonne de la matrice $\mathbf{R}$

❏ Critère quadratique : $\overset{o}{\underline{\boldsymbol{\theta}}} = \arg\min_{\underline{\boldsymbol{\theta}}} (\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}})^T \mathbf{Q} (\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}})$

avec $\underline{\boldsymbol{y}} = \begin{bmatrix} y[1] \\ y[2] \\ \vdots \\ y[N] \end{bmatrix}$ et $\mathbf{Q}$ la matrice $(N \times N)$ de pondération (définie positive)

❏ Moindres carrés : (critère convexe ⇒ annuler son gradient)

$$\overset{o}{\underline{\boldsymbol{\theta}}} = \arg\min_{\underline{\boldsymbol{\theta}}} (\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}})^T \mathbf{Q} (\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}}) \Rightarrow \left\{ \begin{array}{l} \text{Équations normales : } \mathbf{R}^T\mathbf{Q}\mathbf{R}\, \overset{o}{\underline{\boldsymbol{\theta}}} = \mathbf{R}^T\mathbf{Q}\underline{\boldsymbol{y}} \\ \text{Solution : } \overset{o}{\underline{\boldsymbol{\theta}}} = (\mathbf{R}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}\underline{\boldsymbol{y}} \end{array} \right.$$

⇒ Système linéaire de M équations à M inconnues (Analyse numérique)

→ $\overset{o}{\underline{\boldsymbol{\theta}}}$ linéaire par rapport aux données $\underline{\boldsymbol{y}}$:

$$\overset{o}{\underline{\boldsymbol{\theta}}} = \mathbf{M}\underline{\boldsymbol{y}} \text{ avec } \mathbf{M} = (\mathbf{R}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}$$

❏ Propriétés statistiques :

→ Pour des perturbations **additives centrées** :

Si $\underline{\boldsymbol{y}} = \mathbf{R}\underline{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}}$ avec $E\{\underline{\boldsymbol{\epsilon}}\} = 0$ alors $E_{Y|\theta}\{(\mathbf{R}^T\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{Q}\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}\} = \underline{\boldsymbol{\theta}}$

⇒ Estimateur $\overset{o}{\underline{\boldsymbol{\theta}}}$ non biaisé

→ Pour des perturbations **additives centrées de matrice de covariance** $\mathbf{\Gamma}_E$

Si $\mathbf{Q} = \mathbf{\Gamma}_E^{-1}$ avec $\mathbf{\Gamma}_E$ la matrice de covariance des perturbations

alors $V_{\overset{o}{\underline{\boldsymbol{\theta}}}}(\underline{\boldsymbol{\theta}}) = (\mathbf{R}^T\mathbf{\Gamma}_E^{-1}\mathbf{R})^{-1}$

→ Pour des perturbations **centrées additives Gaussiennes** $\mathcal{N}(\underline{\mathbf{0}}, \mathbf{\Gamma}_E)$

$$\left. \begin{array}{l} \underline{\boldsymbol{\epsilon}} \text{ Gaussien} \\ \underline{\boldsymbol{y}} \text{ affine en } \underline{\boldsymbol{\epsilon}} \end{array} \right\} \Rightarrow \left. \begin{array}{l} \underline{\boldsymbol{y}} \text{ Gaussien} \\ \overset{o}{\underline{\boldsymbol{\theta}}} \text{ affine en } \underline{\boldsymbol{y}} \end{array} \right\} \Rightarrow \overset{o}{\underline{\boldsymbol{\theta}}} \text{ Gaussien : } \mathcal{N}(\underline{\boldsymbol{\theta}}, \overbrace{(\mathbf{R}^T\mathbf{\Gamma}_E^{-1}\mathbf{R})^{-1}}^{\text{Pour } \mathbf{Q} = \mathbf{\Gamma}_E^{-1}})$$

On a l'égalité $V_{\overset{o}{\underline{\boldsymbol{\theta}}}}(\underline{\boldsymbol{\theta}}) = \boldsymbol{F}^{-1}(\underline{\boldsymbol{\theta}})$ matrice d'information de Fisher

Inégalité de Cramer-Rao ⇒ Estimateur non biaisé à variance minimale !

Interprétation statistique :

$$\begin{cases} \text{Modèle LP : } \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}\underline{\boldsymbol{\theta}} \\ \text{Perturbations additives gaussiennes : } \underline{\boldsymbol{y}} = \mathbf{R}\underline{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}} \text{ avec } \underline{\boldsymbol{\epsilon}} \sim \mathcal{N}(\underline{\mathbf{0}}, \mathbf{\Gamma}_E) \\ \text{Estimateur du maximum de vraisemblance } \overset{o}{\underline{\boldsymbol{\theta}}}_{\text{MV}} = \overset{o}{\underline{\boldsymbol{\theta}}}_{\text{MC}} \text{ avec } \mathbf{Q} = \mathbf{\Gamma}_E^{-1} \end{cases}$$

① Vraisemblance :

Pour $\underline{\boldsymbol{\theta}}$ connu, $\underline{\boldsymbol{y}}$ est une transfo. affine d'un vect. gaussien $\underline{\boldsymbol{\epsilon}}$: $\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}} \sim \mathcal{N}(\mathbf{R}\underline{\boldsymbol{\theta}}, \mathbf{\Gamma}_E)$

$$L(\underline{\boldsymbol{\theta}}; \underline{\boldsymbol{y}}) = f_{Y|\theta}(\underline{\boldsymbol{y}}|\underline{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{\frac{N}{2}} (\det(\mathbf{\Gamma}_E))^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}})^T \mathbf{\Gamma}_E^{-1} (\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}})}$$

② Anti-log-vraisemblance :

$$\Rightarrow NLL(\underline{\boldsymbol{\theta}}; \underline{\boldsymbol{y}}) = -\log L(\underline{\boldsymbol{\theta}}; \underline{\boldsymbol{y}}) = \frac{1}{2}(\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}})^T \mathbf{\Gamma}_E^{-1} (\underline{\boldsymbol{y}} - \mathbf{R}\underline{\boldsymbol{\theta}}) + \text{cte.}$$

③ Estimateur du maximum de vraisemblance :

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{\text{MV}} = \arg\min_{\underline{\boldsymbol{\theta}}} NLL(\underline{\boldsymbol{\theta}}; \underline{\boldsymbol{y}}) = (\mathbf{R}^T \mathbf{\Gamma}_E^{-1} \mathbf{R})^{-1} \mathbf{R}^T \mathbf{\Gamma}_E^{-1} \underline{\boldsymbol{y}}$$

IV.2 *b)* L'algorithme des Moindres Carrés Récursifs

Traitement en ligne : soit $\overset{o}{\underline{\boldsymbol{\theta}}}_N$ estimés à partir des N données $\underline{\boldsymbol{y}}_N$

Mettre à jour l'estimé $\overset{o}{\underline{\boldsymbol{\theta}}}_N$ à l'arrivée de la $N+1^{\text{ème}}$ donnée $y[N+1]$

$$\text{Soit : } \underline{\boldsymbol{y}}_{N+1} = \mathbf{R}_{N+1}\underline{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}}_{N+1} \text{ avec } \underline{\boldsymbol{y}}_{N+1} = \begin{bmatrix} y[1] \\ \vdots \\ y[N+1] \end{bmatrix} \text{ et } \mathbf{R}_{N+1} = \begin{bmatrix} \underline{\boldsymbol{r}}_1^T \\ \vdots \\ \underline{\boldsymbol{r}}_{N+1}^T \end{bmatrix}$$

$\Rightarrow$ Éviter de stocker $\mathbf{R}_{N+1}$ et de calculer et inverser $\mathbf{R}_{N+1}^T \mathbf{Q}_{N+1} \mathbf{R}_{N+1}$!

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \overset{o}{\underline{\boldsymbol{\theta}}}_N + \underbrace{\underline{\boldsymbol{k}}_{N+1}}_{\text{Gain}} \cdot \underbrace{(y[N+1] - \underline{\boldsymbol{r}}_{N+1}^T \overset{o}{\underline{\boldsymbol{\theta}}}_N)}_{\text{Erreur de prédiction}}$$

Algorithme :

$$\begin{aligned} \underline{\boldsymbol{k}}_{N+1} &= \frac{\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}{\frac{1}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}} \\ \mathbf{P}_{N+1} &= \mathbf{P}_N - \underline{\boldsymbol{k}}_{N+1} \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \qquad \left(\mathbf{P}_N \triangleq (\mathbf{R}_N^T \mathbf{Q}_N \mathbf{R}_N)^{-1}\right) \\ \overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} &= \overset{o}{\underline{\boldsymbol{\theta}}}_N + \underline{\boldsymbol{k}}_{N+1}(y[N+1] - \underline{\boldsymbol{r}}_{N+1}^T \overset{o}{\underline{\boldsymbol{\theta}}}_N). \end{aligned}$$

Principe : (pour une matrice diagonale $\mathbf{Q}_N = \text{diag}\{q_1, q_2, \ldots q_N\}$)

$\rightarrow$ Décomposition par bloc de l'équation normale :

$$\mathbf{R}_{N+1}^T \mathbf{Q}_{N+1} \mathbf{R}_{N+1} \overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \mathbf{R}_{N+1}^T \mathbf{Q}_{N+1} \underline{\boldsymbol{y}}_{N+1}$$

$$\begin{bmatrix} \mathbf{R}_N^T & \underline{\boldsymbol{r}}_{N+1} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_N & \underline{\mathbf{0}} \\ \underline{\mathbf{0}}^T & q_{N+1} \end{bmatrix} \begin{bmatrix} \mathbf{R}_N \\ \underline{\boldsymbol{r}}_{N+1}^T \end{bmatrix} \overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \begin{bmatrix} \mathbf{R}_N^T & \underline{\boldsymbol{r}}_{N+1} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_N & \underline{\mathbf{0}} \\ \underline{\mathbf{0}}^T & q_{N+1} \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{y}}_N \\ y[N+1] \end{bmatrix}$$

Soit en développant :

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \overbrace{(\underbrace{\mathbf{R}_N^T \mathbf{Q}_N \mathbf{R}_N}_{\mathbf{P}_N^{-1}} + q_{N+1} \underline{\boldsymbol{r}}_{N+1} \underline{\boldsymbol{r}}_{N+1}^T)^{-1}}^{\mathbf{P}_{N+1}} (\mathbf{R}_N^T \mathbf{Q}_N \underline{\boldsymbol{y}}_N + q_{N+1} \underline{\boldsymbol{r}}_{N+1} y[N+1])$$

$\rightarrow$ Lemme d'inversion matriciel : $(\mathbf{M} + \alpha \underline{u}\, \underline{v}^T)^{-1} = \mathbf{M}^{-1} - \dfrac{\mathbf{M}^{-1} \underline{u}\, \underline{v}^T \mathbf{M}^{-1}}{\frac{1}{\alpha} + \underline{v}^T \mathbf{M}^{-1} \underline{u}}$

en prenant $\mathbf{M} = \mathbf{P}_N^{-1}$, $\alpha = q_{N+1}$ et $\underline{u} = \underline{v} = \underline{\boldsymbol{r}}_{N+1}$.

$\rightarrow$ Expression de $\overset{o}{\underline{\boldsymbol{\theta}}}_{N+1}$:

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \left(\mathbf{P}_N - \frac{\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1} \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N}{\frac{1}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}\right)(\mathbf{R}_N^T \mathbf{Q}_N \underline{\boldsymbol{y}}_N + q_{N+1} \underline{\boldsymbol{r}}_{N+1} y[N+1])$$

Soit en développant :

$$\begin{aligned} \overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \; & \underbrace{\mathbf{P}_N \mathbf{R}_N^T \mathbf{Q}_N \underline{\boldsymbol{y}}_N}_{\overset{o}{\underline{\boldsymbol{\theta}}}_N} - \frac{\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}{\frac{1}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}} \underline{\boldsymbol{r}}_{N+1}^T \underbrace{\mathbf{P}_N \mathbf{R}_N^T \mathbf{Q}_N \underline{\boldsymbol{y}}_N}_{\overset{o}{\underline{\boldsymbol{\theta}}}_N} \\ & + \underbrace{\left(\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1} - \frac{\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1} \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}{\frac{1}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}\right)}_{\frac{1}{q_{N+1}} \frac{\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}{\frac{1}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}} q_{N+1} y[N+1] \end{aligned}$$

Soit

$$\overset{o}{\underline{\boldsymbol{\theta}}}_{N+1} = \overset{o}{\underline{\boldsymbol{\theta}}}_N + \underbrace{\frac{\mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}{\frac{1}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T \mathbf{P}_N \underline{\boldsymbol{r}}_{N+1}}}_{\text{Gain } \underline{\boldsymbol{k}}_{N+1}} \underbrace{(y[N+1] - \underline{\boldsymbol{r}}_{N+1}^T \overset{o}{\underline{\boldsymbol{\theta}}}_N)}_{\text{Erreur de prédiction}}$$

❑ Cadre Bayésien :

Perturbation additive gaussiennes $\mathcal{N}(\underline{\mathbf{0}}, \mathbf{\Gamma}_E)$ et loi *a priori* gaussienne $\mathcal{N}(\underline{\boldsymbol{m}}_0, \mathbf{\Gamma}_0)$

$$\text{Critère : } J_N(\underline{\boldsymbol{\theta}}) = (\underline{\boldsymbol{y}}_N - \mathbf{R}_N\underline{\boldsymbol{\theta}})^T \mathbf{\Gamma}_E^{-1}(\underline{\boldsymbol{y}}_N - \mathbf{R}_N\underline{\boldsymbol{\theta}}) + (\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{m}}_0)^T \mathbf{\Gamma}_0^{-1}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{m}}_0)$$

$$\text{Équations normales : } \underbrace{(\mathbf{R}_N^T\mathbf{\Gamma}_E^{-1}\mathbf{R}_N + \mathbf{\Gamma}_0^{-1})}_{\mathbf{P}_N^{-1}}(\mathring{\underline{\boldsymbol{\theta}}}_N - \underline{\boldsymbol{m}}_0) = \mathbf{R}_N^T\mathbf{\Gamma}_E^{-1}(\underline{\boldsymbol{y}}_N - \mathbf{R}_N\underline{\boldsymbol{m}}_0)$$

$\Rightarrow$ Même type d'algorithme que MCR

❑ Initialisation :

$\rightarrow$ Cadre Bayésien : $\mathbf{P}_0 = \mathbf{\Gamma}_0$ et $\mathring{\underline{\boldsymbol{\theta}}}_0 = \underline{\boldsymbol{m}}_0$

$\rightarrow$ Cas non Bayésien :

En théorie :
- attendre au moins M données ;
- calculer $\mathring{\underline{\boldsymbol{\theta}}}_M$ et $\mathbf{P}_M$ par MC ;
- poursuivre par MCR.

En pratique : $\mathring{\underline{\boldsymbol{\theta}}}_0 = \underline{\mathbf{0}}$ et $\mathbf{P}_0 = \alpha I$ avec α grand.

❑ Variantes adaptatives :

$\Rightarrow$ Autoriser les paramètres à évoluer au cours du temps

$\rightarrow$ Version à fenêtre glissante : $\begin{cases} \bullet \text{ ajout } y[N+1] \\ \bullet \text{ suppression } y[N-L+2] \end{cases}$

Ne prendre en compte à chaque instant que L données

$\rightarrow$ Version à oubli exponentiel :

Pondérer par un poid λ^n la donnée de l'instant $N-n$ ($\lambda < 1$ proche de 1)

$\Rightarrow$ Algorithme très légèrement modifié :

$$\begin{aligned}
\underline{\boldsymbol{k}}_{N+1} &= \frac{\mathbf{P}_N\underline{\boldsymbol{r}}_{N+1}}{\frac{\lambda}{q_{N+1}} + \underline{\boldsymbol{r}}_{N+1}^T\mathbf{P}_N\underline{\boldsymbol{r}}_{N+1}} \\
\mathbf{P}_{N+1} &= \tfrac{1}{\lambda}(\mathbf{P}_N - \underline{\boldsymbol{k}}_{N+1}\underline{\boldsymbol{r}}_{N+1}^T\mathbf{P}_N) \\
\mathring{\underline{\boldsymbol{\theta}}}_{N+1} &= \mathring{\underline{\boldsymbol{\theta}}}_N + \underline{\boldsymbol{k}}_{N+1}(y[N+1] - \underline{\boldsymbol{r}}_{N+1}^T\mathring{\underline{\boldsymbol{\theta}}}_N)
\end{aligned}$$

❑ Filtre de Kalman :

$\Rightarrow$ Modèle d'état pour l'évolution des paramètres

IV.3 Méthodes dérivées des Moindres Carrés

❑ Régression pseudo-linéaire, Moindres carrés étendus

$\rightarrow$ Modèle pseudo-LP : $y_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}) = \underline{\boldsymbol{r}}_n(\underline{\boldsymbol{\theta}})^T\underline{\boldsymbol{\theta}}$ (régresseur dépend de $\underline{\boldsymbol{\theta}}$)

$\rightarrow$ Soit $\mathring{\underline{\boldsymbol{\theta}}}^{(\ell)}$ estimé à l'itération ℓ, modèle approx. LP : $\widetilde{y}_{\mathcal{M}}(n, \underline{\boldsymbol{\theta}}) = \underline{\boldsymbol{r}}_n(\mathring{\underline{\boldsymbol{\theta}}}^{(\ell)})^T\underline{\boldsymbol{\theta}}$

$\Rightarrow$ $\mathring{\underline{\boldsymbol{\theta}}}^{(\ell+1)}$ se calcule par MC

$\rightarrow$ Problèmes d'initialisation, convergence non garantie. . .

❑ Régression multi-linéaire, MC Généralisés, MV Généralisé

$\rightarrow$ Partition $\underline{\boldsymbol{\theta}} = [\underline{\boldsymbol{\theta}}_1, \underline{\boldsymbol{\theta}}_2, \ldots \underline{\boldsymbol{\theta}}_K]^T$ telle que :

$J(\underline{\boldsymbol{\theta}}_1, \underline{\boldsymbol{\theta}}_2, \ldots \underline{\boldsymbol{\theta}}_K)$ est quadratique en $\underline{\boldsymbol{\theta}}_k$ quand les autres sont fixés $\forall k = 1 \ldots K$

$\rightarrow$ Explorer cycliquement pour $k = 1 \ldots K$ et itérer. . .

$$\mathring{\underline{\boldsymbol{\theta}}}_k^{(\ell+1)} = \arg\min_{\underline{\boldsymbol{\theta}}_k} J(\mathring{\underline{\boldsymbol{\theta}}}_1^{(\ell)}, \mathring{\underline{\boldsymbol{\theta}}}_2^{(\ell)}, \ldots, \underline{\boldsymbol{\theta}}_k, \ldots, \mathring{\underline{\boldsymbol{\theta}}}_K^{(\ell)}) \text{ par MC}$$

$\rightarrow$ A chaque itération $J \searrow$, convergence garantie mais pas vers un min. local. . .

❑ Critère quadratique et modèle partiellement LP

$\rightarrow$ Modèle partiellement LP : $\underline{\boldsymbol{\theta}} = [\underline{\boldsymbol{\theta}}_{\mathrm{L}}, \underline{\boldsymbol{\theta}}_{\mathrm{NL}}]^T$ modèle LP vis-à-vis de $\underline{\boldsymbol{\theta}}_{\mathrm{L}}$

$$y_{\mathcal{M}}(k, \underline{\boldsymbol{\theta}}) = \underline{\boldsymbol{r}}_k(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})^T\underline{\boldsymbol{\theta}}_{\mathrm{L}} \quad \Leftrightarrow \quad \underline{\boldsymbol{y}}_{\mathcal{M}}(\underline{\boldsymbol{\theta}}) = \mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})\underline{\boldsymbol{\theta}}_{\mathrm{L}}$$

$$\Rightarrow J(\underline{\boldsymbol{\theta}}_{\mathrm{L}}, \underline{\boldsymbol{\theta}}_{\mathrm{NL}}) = (\underline{\boldsymbol{y}} - \mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})\underline{\boldsymbol{\theta}}_{\mathrm{L}})^T\mathbf{Q}(\underline{\boldsymbol{y}} - \mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})\underline{\boldsymbol{\theta}}_{\mathrm{L}}) \text{ quadratique en } \underline{\boldsymbol{\theta}}_{\mathrm{L}}$$

$$\text{Par MC} \quad \mathring{\underline{\boldsymbol{\theta}}}_{\mathrm{L}}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}}) = \left(\mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})^T\mathbf{Q}\mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})\right)^{-1}\mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})^{\mathbf{T}}\mathbf{Q}\underline{\boldsymbol{y}}$$

$$\mathcal{J}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}}) = J(\mathring{\underline{\boldsymbol{\theta}}}_{\mathrm{L}}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}}), \underline{\boldsymbol{\theta}}_{\mathrm{NL}}) = (\underline{\boldsymbol{y}} - \mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})\mathring{\underline{\boldsymbol{\theta}}}_{\mathrm{L}}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}}))^T\mathbf{Q}(\underline{\boldsymbol{y}} - \mathbf{R}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}})\mathring{\underline{\boldsymbol{\theta}}}_{\mathrm{L}}(\underline{\boldsymbol{\theta}}_{\mathrm{NL}}))$$

$\Rightarrow$ Critère à minimiser en $\underline{\boldsymbol{\theta}}_{\mathrm{NL}}$ de dimension $< M$

Intéressant par exemple si θ_{NL} est scalaire : Modèle LP sauf en un paramètre !

❏ Méthode de type Majorize-Minimization (MM)

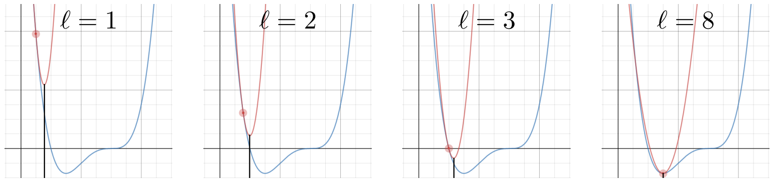
A chaque itération ℓ pour le point courant $\underline{\boldsymbol{\theta}}^{(\ell)}$:

→ Construction d'une approximation majorante du critère $J(\underline{\boldsymbol{\theta}})$:

$$K(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}^{(\ell)}) \text{ telle que } \begin{cases} \forall \underline{\boldsymbol{\theta}},\ K(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}^{(\ell)}) \geqslant J(\underline{\boldsymbol{\theta}}) \\ K(\underline{\boldsymbol{\theta}}^{(\ell)}, \underline{\boldsymbol{\theta}}^{(\ell)}) = J(\underline{\boldsymbol{\theta}}^{(\ell)}) \end{cases}$$

→ Minimisation du critère $K(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}^{(\ell)})$

$$\underline{\boldsymbol{\theta}}^{(\ell+1)} = \arg\min_{\underline{\boldsymbol{\theta}}} K(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}^{(\ell)})$$



→ Pour une approximation majorante quadratique :

$$\text{Minimisation de } K(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}^{(\ell)}) = \underline{\boldsymbol{\theta}}^T \mathbf{A}^{(\ell)} \underline{\boldsymbol{\theta}} - 2(\underline{\boldsymbol{b}}^{(\ell)})^T \underline{\boldsymbol{\theta}} \text{ de type MC}$$