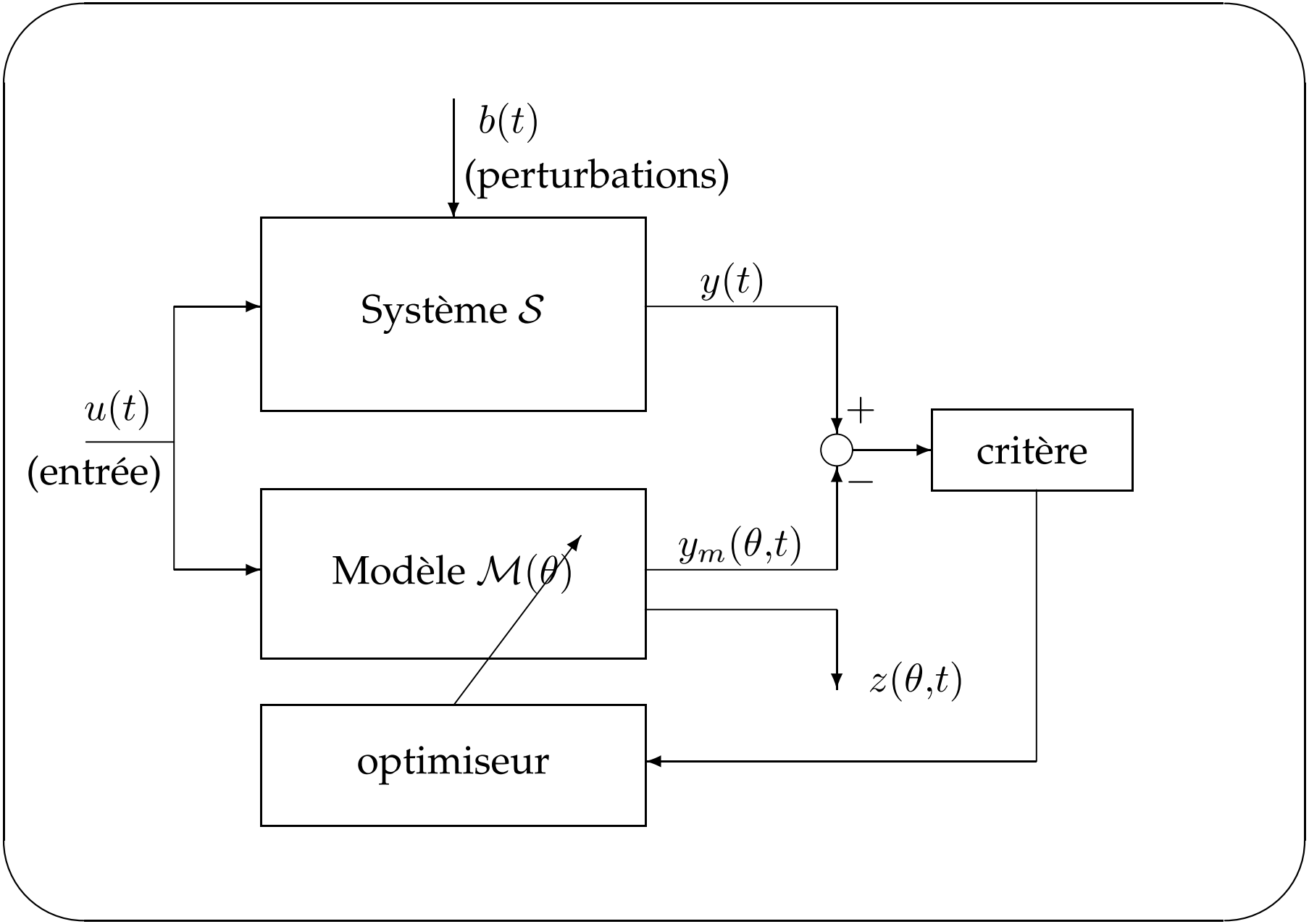
Université Nice SOPHIA ANTIPOLIS

I3S

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

ESTIMATION DE PARAMÈTRES

Luc Pronzato, 2004


$b(t)$
(perturbations)
Système $\mathcal{S}$
$y(t)$
$u(t)$
(entrée)
+
−
critère
Modèle $\mathcal{M}(\theta)$
$y_m(\theta,t)$
$z(\theta,t)$
optimiseur

0) Généralités

$\mathbf{y}$ vecteur de toutes les observations sur $\mathcal{S}$

$\mathbf{y}_m(\theta)$ vecteur des réponses de modèle associées pour des paramètres θ

Choisir θ pour que $\mathbf{y}_m(\theta)$ «ressemble» à $\mathbf{y}$

$\rightarrow$ minimiser $j(\theta)$, «distance» entre $\mathbf{y}_m(\theta)$ et $\mathbf{y}$ $\rightarrow$ $\hat{\theta}(\mathbf{y})$

Rq1 : $\mathbf{y}$ aléatoire, donc $\hat{\theta}(\mathbf{y})$ aléatoire aussi (biais, variance, etc.)

Rq2 : Comment optimiser $j(\cdot)$? — voir plus loin

1) Moindres carrés

Distance quadratique (norme L_2)

$$j_{MC}(\theta) = \|\mathbf{y} - \mathbf{y}_m(\theta)\|^2 = \sum_{i=1}^{N} [y(i) - y_m(\theta,i)]^2$$

i est le numéro de la mesure, peut correspondre au temps t_i où on observe, avec $t_i = Ti$ si système échantillonné avec une période T.
Plus généralement, i correspond à des conditions expérimentales ξ_i

1.1) Moindres carrés pondérés :

$$j_{MCP}(\theta) = [\mathbf{y} - \mathbf{y}_m(\theta)]^\top \mathbf{W} [\mathbf{y} - \mathbf{y}_m(\theta)] = \sum_{i=1}^{N} w_i [y(i) - y_m(\theta,i)]^2$$

avec $\mathbf{W} = \text{diag}(w_i, i = 1, \ldots, N)$ matrice de pondération $(w_i \geq 0, \forall i)$.
Choix des w_i ?

1.2) Répétition de mesures : n_i mesures $y_j(i)$, $j = 1, \ldots, n_i$, pour les mêmes conditions expérimentales ξ_i (au même instant par exemple).

Soit m le nb. de conditions expérimentales différentes ($\sum_{i=1}^{m} n_i = N$). Posons

$$y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_j(i)$$

$$
\begin{aligned}
j_{MCP}(\theta) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i[y_j(i) - y_m(\theta,i)]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i[y_j(i) - y_i + y_i - y_m(\theta,i)]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i[y_j(i) - y_i]^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i[y_i - y_m(\theta,i)]^2 \\
&\quad +2\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i[y_j(i) - y_i] \times [y_i - y_m(\theta,i)] \\
&= \sum_{i=1}^{m} n_i w_i[y_i - y_m(\theta,i)]^2 + \text{ terme ind. de } \theta
\end{aligned}
$$

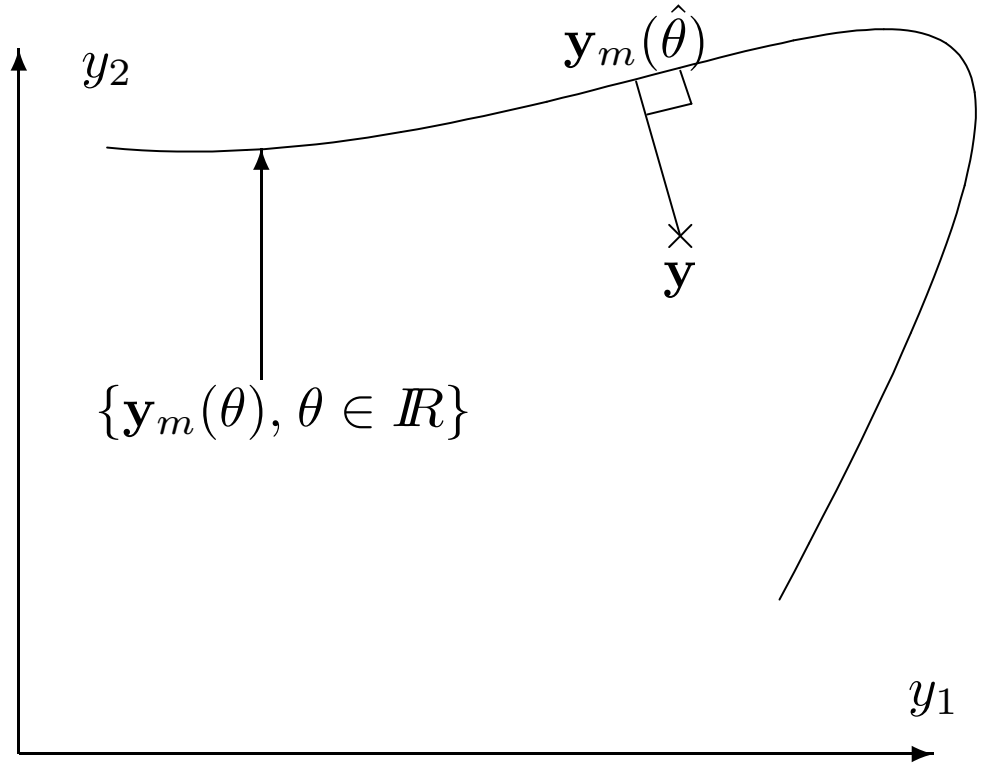
$\rightarrow$ moindres carrés pondérés pour les m observations y_i, pondérations $n_i w_i$

1.3) Minimas locaux :

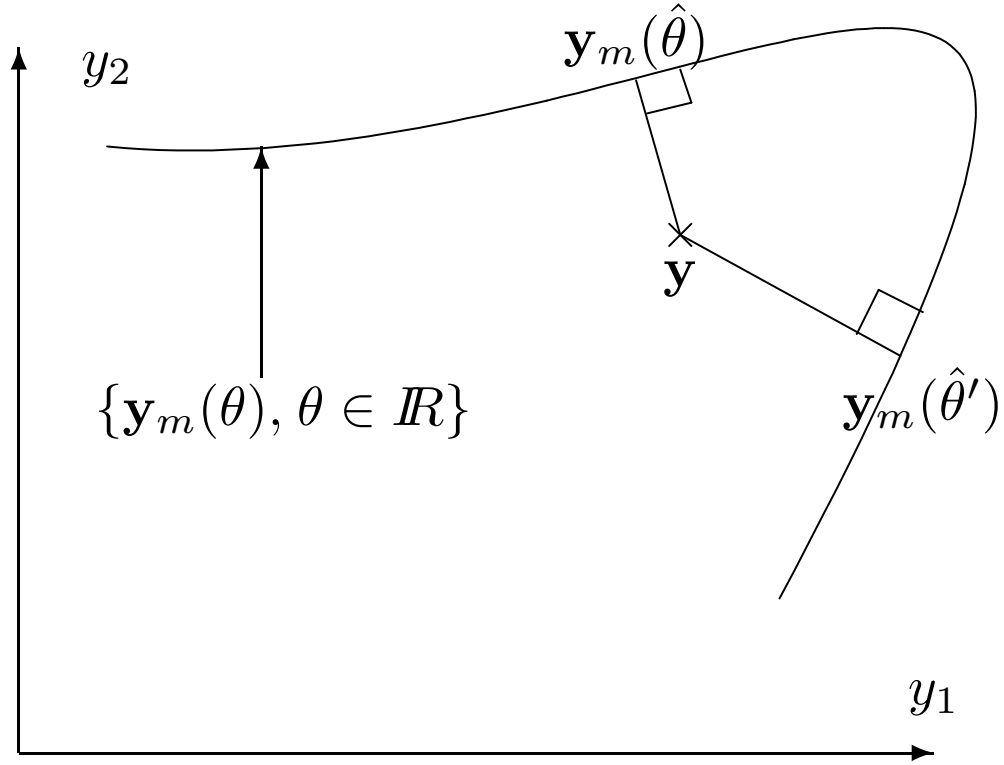
$\hat{\theta}$ = estimateur des moindres carrés

$\mathbf{y}_m(\hat{\theta})$ = projection $\perp$ de $\mathbf{y}$ sur la surface des réponses de modèle

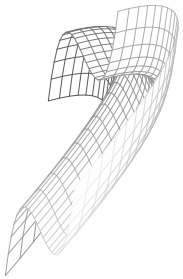
$\mathcal{S}_r = \{\mathbf{y}_m(\theta), \theta \in I\!R^p\}$.



Il peut y avoir des minimas locaux si $\mathcal{S}_r$ est courbe !

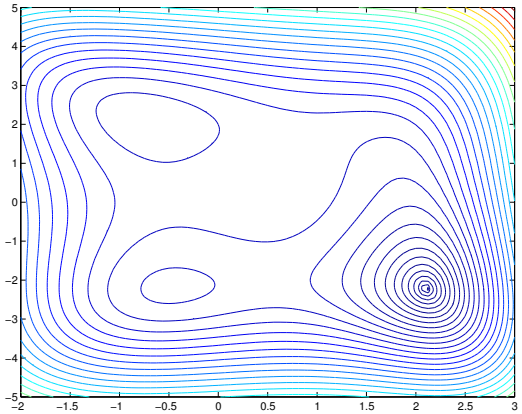


$y_m(\theta,\mathbf{x}) = \theta_1 x_1 + \theta_2 x_2 +$
$\theta_1^3(1-x_1) + \theta_2^2(1-x_2)$

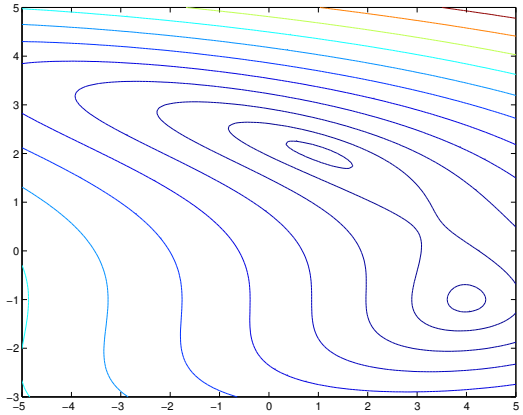


3 observations, en $\mathbf{x}^1 = (1,0)^\top$,
$\mathbf{x}^2 = (1,1)^\top$ et $\mathbf{x}^3 = (0,1)^\top$

$\mathbf{y}_m(\theta)$, $\theta_1 \in [-5,5]$, $\theta_2 \in [-2,5]$



isocritères dans le plan
(θ_1,θ_2), $\mathbf{y} = (5, -10,8)^\top$



répétition de mesures :
$\mathbf{x}^1 = (1, 0)^\top$, $\mathbf{x}^2 = \mathbf{x}^3 = (1, 1)^\top$
$\mathbf{y} = (5,2,4)^\top$

Répétition de mesures $\rightarrow$ pertes d'identifiabilité globale :
$\hat{\theta}^1 = (\hat{\theta}_1,\hat{\theta}_2)^\top$ et $\hat{\theta}^2 = (\hat{\theta}_1 + 2\hat{\theta}_2 - 1,1 - \hat{\theta}_2)^\top$ donnent le même $\mathbf{y}_m(\theta)$

$\rightarrow$ Supprimer la courbure de $\mathcal{S}_r$

Plusieurs observations pour les mêmes conditions expérimentales :

$m = p = \dim(\theta)$

$\Rightarrow \mathcal{S}_r$ est plane

$\Rightarrow$ plus de minimas locaux !

$\rightarrow$ On résout en θ : $y_m(\theta,i) = y(i)$, $i = 1,\ldots,p$.

Parfois plusieurs solutions (identifiabilité locale) :

- ☞ 1) chercher toutes les solutions
- ☞ 2) utiliser + de conditions expérimentales
- ☞ 3) $\exists$ peut être des minimas locaux : initialiser la recherche aux solutions trouvées en 1.

2) Moindres valeurs absolues (norme L_1)

$$j_{MVA}(\theta) = \|\mathbf{y} - \mathbf{y}_m(\theta)\|_1 = \sum_{i=1}^{N} |y(i) - y_m(\theta,i)|$$

$$j_{MVAP}(\theta) = \|\mathbf{W}[\mathbf{y} - \mathbf{y}_m(\theta)]\|_1 = \sum_{i=1}^{N} w_i |y(i) - y_m(\theta,i)|$$

Pas différentiable partout !

Optimum pas forcément unique (même si le modèle est s.g.i.)

Exemple : $y_m(\theta,1) = \theta_1 + \theta_2^2$, $y_m(\theta,2) = y_m(\theta,3) = \theta_1 + \theta_2$, $\mathbf{y} = (5,2,4)$.

$j_{MVA}(\theta) = 2\ \forall\theta$ tel que $\theta_1 + \theta_2^2 = 5$, $2 \leq \theta_1 + \theta_2 \leq 4$

Pas recommandé tel quel —voir + loin

Modèles LP : $y_m(\theta,i) = \mathbf{r}_i^\top \theta + c_i$

Le problème revient à minimiser $\sum_{i=1}^N \alpha_i$, sous les contraintes

$$y(i) - \mathbf{r}_i^\top \theta - c_i \leq \alpha_i \text{ et } y(i) - \mathbf{r}_i^\top \theta - c_i \geq -\alpha_i \, , \, i = 1, \ldots, N$$

$N + p$ variables (α et θ), $2N$ contraintes.

Les contraintes et la fonction sont linéaires en les variables $\rightarrow$ programmation linéaire

Répétition de mesures : Pas toujours comme en 1.2, mais sous certaines conditions ($m = p$) $\rightarrow$ moindres valeurs absolues pondérées, pour les observations

$$y_i = \text{med}\{y_j(i), \, j = 1, \ldots, n_i\}$$

avec les pondérations $n_i w_i$.

3) Minimax (norme L_∞)

$$j_{MM}(\theta) = \|\mathbf{y} - \mathbf{y}_m(\theta)\|_\infty = \max_{i=1,\ldots,N} |y(i) - y_m(\theta,i)|$$

$$j_{MMP}(\theta) = \|\mathbf{W}[\mathbf{y} - \mathbf{y}_m(\theta)]\|_\infty = \max_{i=1,\ldots,N} w_i|y(i) - y_m(\theta,i)|$$

Modèles LP : $y_m(\theta,i) = \mathbf{r}_i^\top \theta + c_i$

Minimiser α, sous les contraintes

$$y(i) - \mathbf{r}_i^\top \theta - c_i \leq \alpha \text{ et } y(i) - \mathbf{r}_i^\top \theta - c_i \geq -\alpha\,,\ i = 1,\ldots,N$$

$1 + p$ variables (α et θ), $2N$ contraintes.

$\rightarrow$ programmation linéaire

Répétition de mesures : Presque comme en 1.2, minimax pondéré, pour les observations

$$y_i = \frac{\max_{j=1,\ldots,n_i}\{y_j(i)\} + \min_{j=1,\ldots,n_i}\{y_j(i)\}}{2}$$

avec les pondérations $n_i w_i$

(La raison :

$$\max_{j=1,\ldots,n_i} |y_j(i) - y_m(\theta,i)| \quad = \quad |y(i) - y_m(\theta,i)| + \frac{\max_{j=1,\ldots,n_i}\{y_j(i)\} - \min_{j=1,\ldots,n_i}\{y_j(i)\}}{2})$$

4) Maximum de vraisemblance

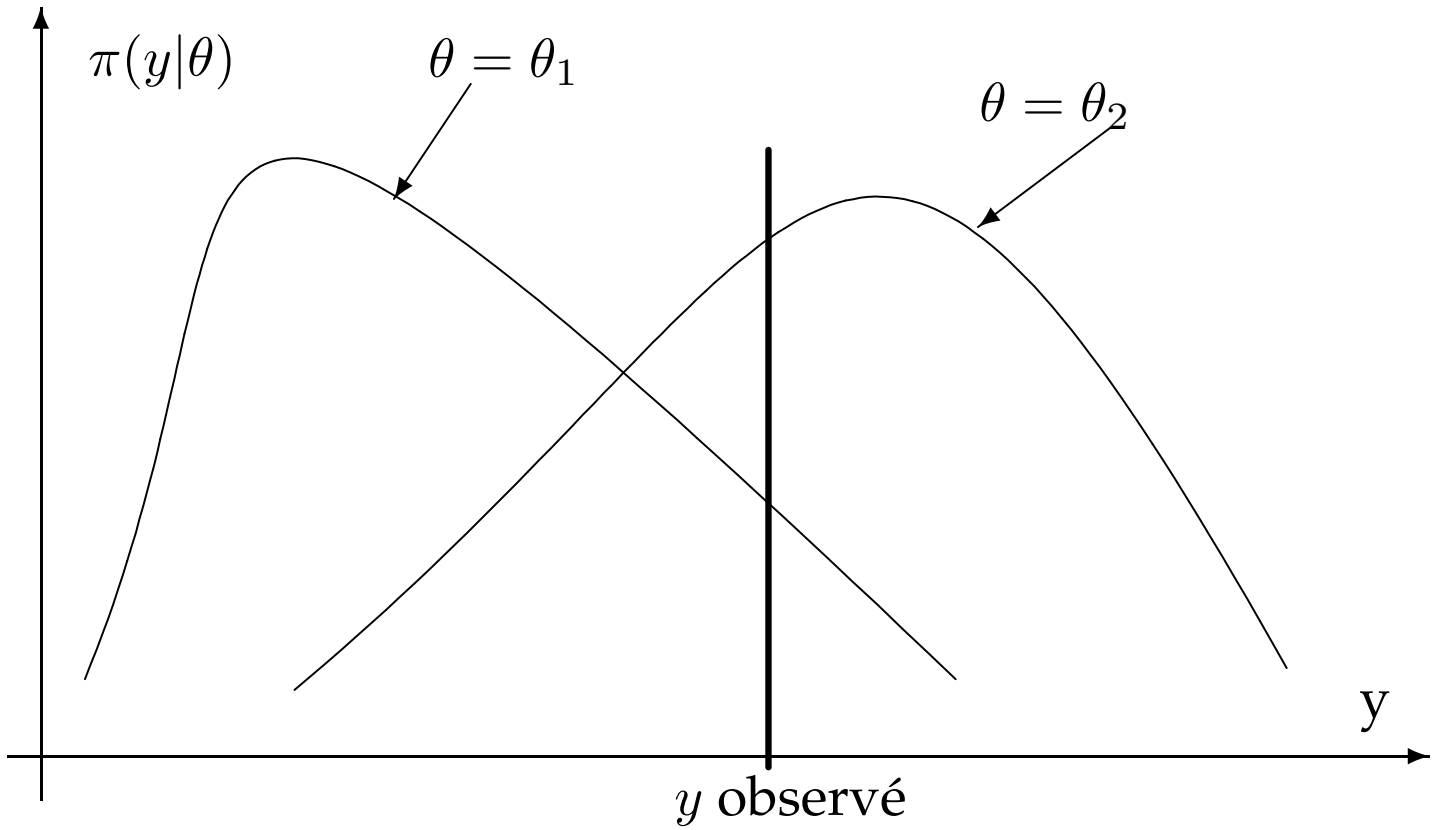
4.1) Généralités

$\mathbf{y}$ variable aléatoire, générée par un modèle déterministe, de paramètres $\bar{\theta}$ (inconnus), perturbé par $(\epsilon_k)_k$ (suite de v.a.i., éventuellement filtrée par un partie du système)

vraisemblance de $\mathbf{y}$: $\pi(\mathbf{y}|\theta)$

Ici, $\mathbf{y}$ est fixé (observations) : $\hat{\theta}_{MV}$ maximise $\pi(\mathbf{y}|\theta)$
(en pratique on maximise la log-vraisemblance $\log \pi(\mathbf{y}|\theta)$ — pourquoi ?)

Suivant la loi des perturbations $\rightarrow$ divers critères d'estimation


$\pi(y|\theta)$
$\theta = \theta_1$
$\theta = \theta_2$
y
y observé

4.2) Exemples (on fait toujours N observations, $i = 1, \ldots, N$)

Ex1 : $y(i) = \epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

$\hat{\theta}_{MV}$?

Biais ($\mathtt{E}\{\hat{\theta}(\mathbf{y})|\theta\} - \theta$)?

Variance ($\mathtt{E}\{[\hat{\theta}(\mathbf{y}) - \mathtt{E}\{\hat{\theta}(\mathbf{y})|\theta\}][\hat{\theta}(\mathbf{y}) - \mathtt{E}\{\hat{\theta}(\mathbf{y})|\theta\}]^\top|\theta\}$)?

Ex2 : $y(i) = y_m(\theta, i) + \epsilon_i$, $(\epsilon_i)_i$ ind. $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$

a) σ_i^2 connus $\rightarrow \hat{\theta}_{MV}$?

b) $\forall i$, $\sigma_i^2 = \sigma^2$ (inconnu), $\hat{\theta}_{MV}$ et $\hat{\sigma}^2_{MV}$?

c) σ_i^2 inconnus, comment faire ?

Ex3 : Sortie vectorielle

$\mathbf{y}(i) = \mathbf{y}_m(\theta, i) + \epsilon_i$, $\mathbf{y}_i \in I\!R^n$, $(\epsilon_i)_i$ ind., $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$

$\hat{\theta}_{MV}$ et $\hat{\Sigma}_{MV}$?

Ex4 : Bruit de Laplace, moyenne 0, variance s^2

$y(i) = y_m(\theta,i) + \epsilon_i$, $(\epsilon_i)_i$ ind.,

$$\pi(\epsilon_i) = \frac{1}{\sqrt{2}\sigma_i} \exp\left(-\frac{\sqrt{2}|\epsilon_i|}{\sigma_i}\right)$$

a) σ_i connus $\rightarrow \hat{\theta}_{MV}$?

b) $\forall i$, $\sigma_i = \sigma$ (inconnu), $\hat{\theta}_{MV}$ et $\hat{\sigma}_{MV}$?

Ex5 : Bruit uniforme

$y(i) = y_m(\theta,i) + \epsilon_i$, $(\epsilon_i)_i$ ind.,

$$\pi(\epsilon_i) = \begin{cases} \frac{1}{2A} & \text{si } |\epsilon_i| \leq A \\ 0 & \text{sinon} \end{cases}$$

a) A connu, $\hat{\theta}_{MV}$?

b) A inconnu, $\hat{\theta}_{MV}$, $\hat{A}_{MV}$?

4.3) Complexité

Plusieurs structures de modèles $\mathcal{M}_i$ en compétition (par exemple, avec un nb. croissant de paramètres p_i)

$\rightarrow$ estimer en même temps quelle est la bonne structure ?
$\rightarrow$ quelle est la valeur de ses paramètres ?

Akaike Information Criterion : $j_{AIC}(\theta,i) = \frac{1}{N}\left[-\log \pi(\mathbf{y}|\theta) + p_i\right]$

(à minimiser)
Max de vraisemblance pour chaque structure ($\rightarrow \hat{\theta}_i$), choisir $\mathcal{M}_{\hat{i}}$ avec $j_{AIC}(\hat{\theta}_i,i)$ minimum

Autres critères : toujours la même idée $\rightarrow$ pénaliser les structures complexes

4.4) Erreur de prédiction

Pour écrire $\pi(\mathbf{y}|\theta)$, on utilise l'indépendance ($\rightarrow$ produit de lois individuelles, $\log \pi(\mathbf{y}|\theta) = \sum$)

$\rightarrow$ toujours revenir à $(\epsilon_i)_i$

$\rightarrow$ construire des variables $e(\theta,i)$ (erreurs de prédiction) qui correspondent à ϵ_i sous l'hypothèse de données générées pas $\mathcal{M}(\theta)$.

Ex : Box & Jenkins

$$y = F(\theta,q)u + G(\theta,q)\epsilon \;\rightarrow\; e(\theta,i) = G^{-1}(\theta,q)[y(i) - F(\theta,q)u(i)]$$

Rq 1 : Système LI $\rightarrow$ on peut toujours propager l'erreur pour la rendre additive en sortie

Rq 2 : Th. de factorisation spectrale $\rightarrow \forall$ spectre de bruit sur y, on peut toujours le considérer comme un bruit blanc (suite de v.a.i.) filtré pas $G(\theta,q)$, rationnel en q, stable et d'inverse stable.

4.5) Propriétés

Pourquoi le maximum de vraisemblance ?

Hypothèses :

- ✓ données générées par $\mathcal{M}(\bar{\theta})$
- ✓ perturbations $(\epsilon_i)_i$ i.i.d. (quelque part)
- ✓ $\{\mathbf{y}/\pi(\mathbf{y}|\theta) > 0\}$ indépendant de θ
- ✓ $\frac{\partial^2 \log \pi(\mathbf{y}|\theta)}{\partial\theta\partial\theta^\top}$ existe, continue en θ uniformément en $\mathbf{y}$
- ✓ $\mathtt{E}\left\{\left|\frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial\theta_i}\right| \ |\theta\right\} < \infty, \forall i = 1,\ldots,p, \forall\theta$
- ✓ $\mathtt{E}\left\{\left|\frac{\partial^2 \log \pi(\mathbf{y}|\theta)}{\partial\theta_i\partial\theta_j}\right| \ |\theta\right\} < \infty, \forall (i,j) = 1,\ldots,p, \forall\theta$

...alors

$$\sqrt{N}(\hat{\theta}^N_{ML} - \bar{\theta}) \sim_{N\to\infty} \mathcal{N}(\mathbf{0},\mathbf{M}^{-1}(\bar{\theta},\xi))$$

avec $\mathbf{M}(\bar{\theta},\xi)$ la matrice d'information de Fisher (moyenne, par observation)

$$\begin{array}{rcl} \mathbf{M}(\theta,\xi) & = & \mathtt{E}\left\{ \frac{1}{N} \frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial \theta} \frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial \theta^\top} \; |\theta \right\} \\ & = & -\mathtt{E}\left\{ \frac{1}{N} \frac{\partial^2 \log \pi(\mathbf{y}|\theta)}{\partial \theta \partial \theta^\top} \; |\theta \right\} \end{array}$$

(ξ caractérise les conditions expérimentales utilisées)

asymptotiquement sans biais

asymptotiquement efficace (retour à l'exemple 1 de 4.2)

Pour $y(i) = y_m(\theta,i) + \epsilon_i$ avec $(\epsilon_i)_i$ i.i.d. (si ce n'est pas le cas, on se ramène à des erreurs de prédiction, et on considère $\pi[e(\theta,i)|\theta]$) :

$$\frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta}$$

et dans $\mathbf{M}(\theta,\xi)$,

$$\begin{aligned}
\mathtt{E}\left\{ \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta} \frac{\partial \log \pi[y(j)|\theta]}{\partial \theta^\top} |\theta \right\} &= \mathtt{E}\left\{ \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta} |\theta \right\} \\
&\quad \times \mathtt{E}\left\{ \frac{\partial \log \pi[y(j)|\theta]}{\partial \theta^\top} |\theta \right\} \\
\text{De plus, } \mathtt{E}\left\{ \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta} |\theta \right\} &= \int \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta} \pi[y(i)|\theta] dy(i) \\
&= \int \frac{\partial \pi[y(i)|\theta]}{\partial \theta} dy(i) = 0
\end{aligned}$$

et $\boxed{\mathbf{M}(\theta,\xi) = \mathtt{E}\left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta} \frac{\partial \log \pi[y(i)|\theta]}{\partial \theta^\top} |\theta \right\}}$

Inégalité de Cramér-Rao :

Pour tout estimateur $\hat{\theta}$ non-biaisé ($\mathtt{E}\{\hat{\theta}(\mathbf{y})|\theta\} = \theta$)

$$\mathrm{Cov}(\hat{\theta}) \succeq \mathbf{M}^{-1}(\theta,\xi)/N = \left[\mathtt{E}\left\{\frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial \theta}\frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial \theta^\top}\ |\theta\right\}\right]^{-1}$$

($\mathbf{A} \succeq \mathbf{B} \Leftrightarrow \mathbf{A} - \mathbf{B}$ définie non-négative)

Démonstration :

✍ Calculer $\mathbf{A} = \mathtt{E}\{\mathbf{v}(\theta,\mathbf{y})\mathbf{v}^\top(\theta,\mathbf{y})|\theta\}$ avec $\mathbf{v}(\theta,\mathbf{y}) = \begin{pmatrix} \hat{\theta}(\mathbf{y}) - \theta \\ \frac{\partial \log \pi(\mathbf{y}|\theta)}{\partial \theta} \end{pmatrix}$

$$\rightarrow \mathbf{A} = \begin{pmatrix} \mathrm{Cov}(\hat{\theta}) & \mathbf{I}_p \\ \mathbf{I}_p & N\mathbf{M}(\theta,\xi) \end{pmatrix} \succeq \mathbf{O}$$

✍ Calculer $\begin{pmatrix} \mathbf{I}_p & -\mathbf{M}^{-1}(\theta,\xi)/N \end{pmatrix} \mathbf{A} \begin{pmatrix} \mathbf{I}_p \\ -\mathbf{M}^{-1}(\theta,\xi)/N \end{pmatrix} \succeq \mathbf{O}$

On utilise la dérivation sous le signe $\int \ldots$ rappel :

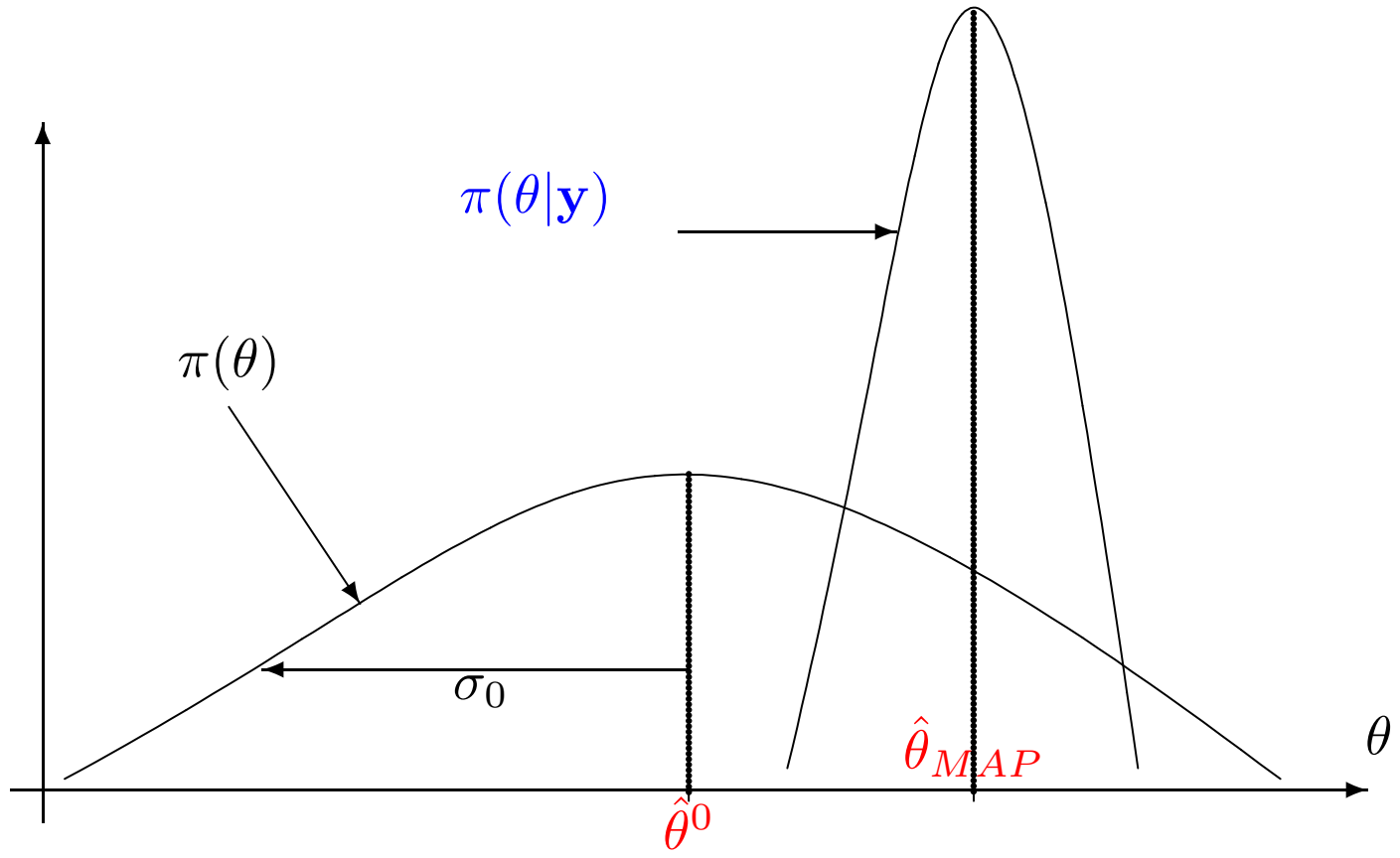
$$\text{Si} \begin{cases} f : [a,b) \times \mathcal{S} \to I\!R\,, \quad -\infty < a < b \leq \infty\,, \mathcal{S} \subseteq I\!R \\ \frac{\partial f(x,y)}{\partial y}\ \exists\,, \forall x \in [a,b)\,, \forall y \in \mathcal{S} \\ \forall y \in \mathcal{S}\,, f(x,y) \text{ et } \frac{\partial f(x,y)}{\partial y} \text{ sont intégrables sur } [a,b) \\ \forall x \in [a,b)\,, \forall y \in \mathcal{S}\,, \left|\frac{\partial f(x,y)}{\partial y}\right| \leq g(x)\,, \text{avec } g(x) \text{ intégrable} \end{cases}$$

alors

$$\frac{\partial \int_a^b f(x,y)\,dx}{\partial y} = \int_a^b \frac{\partial f(x,y)}{\partial y}\,dx$$

5) Estimation bayesienne

θ : v.a., densité connue $\pi(\theta)$, après observations $\mathbf{y} \rightarrow \pi(\theta|\mathbf{y})$ (loi *a posteriori*)



Règle de Bayes :

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta,\mathbf{y})}{\pi(\mathbf{y})} = \frac{\pi(\mathbf{y}|\theta)\pi(\theta)}{\pi(\mathbf{y})}$$

Estimateur du maximum *a posteriori* : $\hat{\theta}_{MAP}$ maximise $\pi(\theta|\mathbf{y})$

$\rightarrow$ maximise $\underbrace{\log \pi(\mathbf{y}|\theta)}_{\text{log-vraisemblance}} + \underbrace{\log \pi(\theta)}_{\textit{a priori}}$

Mêmes propriétés asymptotiques que max. de vraisemblance

Si loi *a priori* normale $\mathcal{N}(\hat{\theta}^0,\Omega)$: $\hat{\theta}_{MAP}$ maximise

$$\log \pi(\mathbf{y}|\theta) - \frac{1}{2}(\theta - \hat{\theta}^0)^\top \Omega^{-1}(\theta - \hat{\theta}^0)$$

Si en plus, modèle LP $\mathbf{y}(\theta) = \mathbf{R}\theta + \epsilon$, avec $\epsilon \sim \mathcal{N}(\mathbf{0},\Sigma)$,
$\hat{\theta}_{MAP} = (\mathbf{R}^\top \Sigma^{-1}\mathbf{R} + \Omega^{-1})^{-1}(\mathbf{R}^\top \Sigma^{-1}\mathbf{y} + \Omega^{-1}\hat{\theta}^0)$

Estimateur du risque minimum :

Coût $j(\hat{\theta}|\bar{\theta})$ d'estimer θ par $\hat{\theta}$ si vraie valeur $\bar{\theta}$ (commande optimale...)

$$
\begin{aligned}
j_{RM}(\theta) &= \int j(\theta|\bar{\theta})\pi(\bar{\theta}|\mathbf{y})d\bar{\theta} \\
&= \frac{1}{\pi(\mathbf{y})}\int j(\theta|\bar{\theta})\pi(\mathbf{y}|\bar{\theta})\pi(\bar{\theta})d\bar{\theta} \\
&= \text{const.}\int \underbrace{j(\theta|\bar{\theta})}_{\text{coût}} \underbrace{\pi(\mathbf{y}|\bar{\theta})}_{\text{perturbations}} \underbrace{\pi(\bar{\theta})}_{\textit{a priori}} d\bar{\theta}
\end{aligned}
$$

Si $j(\theta|\bar{\theta})$ quadratique en θ

$$j(\theta|\bar{\theta}) = (\theta - \bar{\theta})^\top \mathbf{Q}(\theta - \bar{\theta}) \, , \, \mathbf{Q} \text{ sym. } \succ \mathbf{O}$$

Stationnarité du critère $j_{RM}(\theta)$:

$$\frac{\partial j_{RM}(\theta)}{\partial \theta}_{|\hat{\theta}_{RM}} = \mathbf{0} = 2\mathbf{Q} \int (\hat{\theta}_{RM} - \bar{\theta}) \pi(\bar{\theta}|\mathbf{y}) d\bar{\theta}$$

$\Rightarrow \hat{\theta}_{RM} = \int \bar{\theta} \pi(\bar{\theta}|\mathbf{y}) d\bar{\theta} = \mathtt{E}\{\theta|\mathbf{y}\}$ moyenne *a posteriori* de θ

...et si la loi *a posteriori* est symétrique, $\hat{\theta}_{RM} = \mathtt{E}\{\theta|\mathbf{y}\} = \hat{\theta}_{MAP}$

6) Contraintes

Égalités ($\mathbf{c}_e(\theta) = \mathbf{0}$) ou inégalités ($\mathbf{c}_i(\theta) \leq \mathbf{0}$)

On les prend en compte soit par le critère, soit par l'algorithme d'optimisation

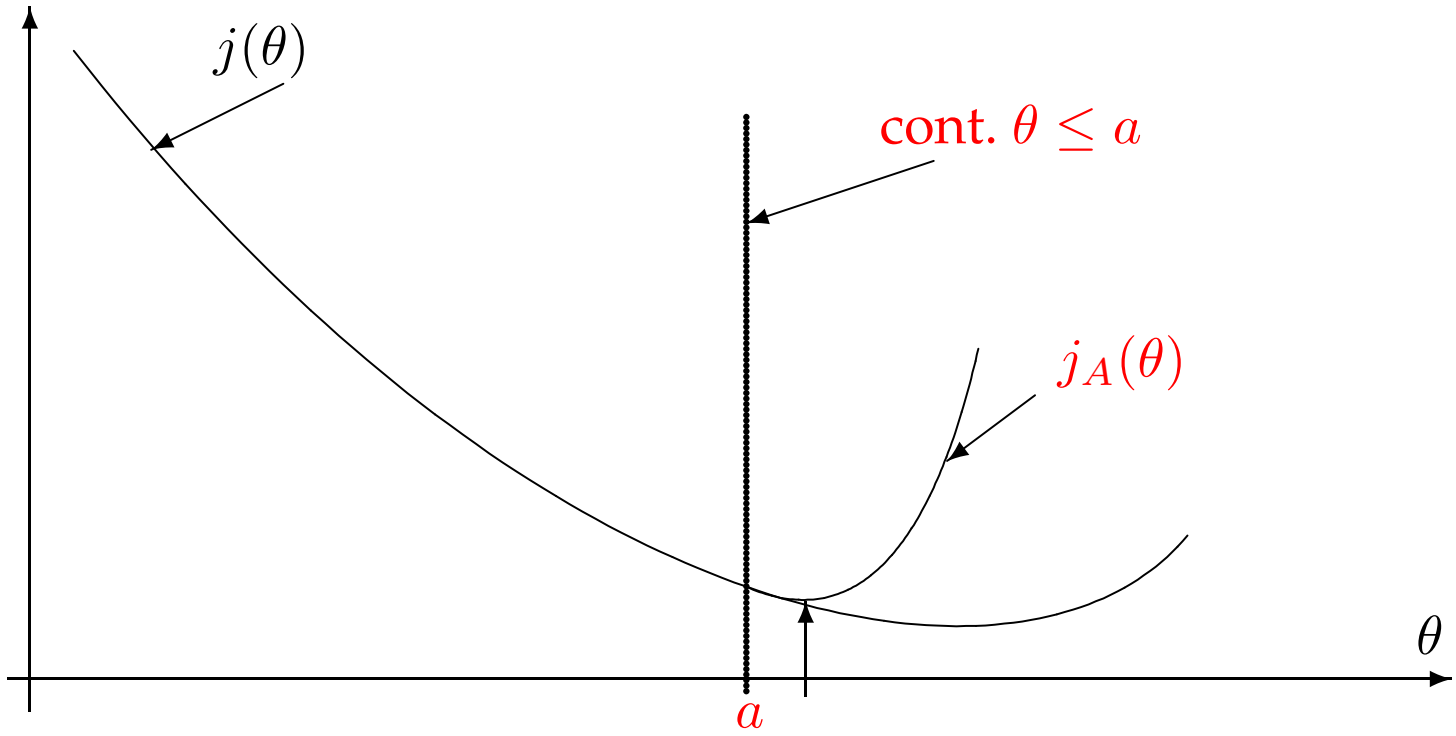
Par le critère : pénalisation (on suppose $j(\theta)$ à minimiser)

6.1) Inégalités

On définit $\mathbf{c}_i^+(\theta) = \max\{\mathbf{c}_i(\theta), 0\}$

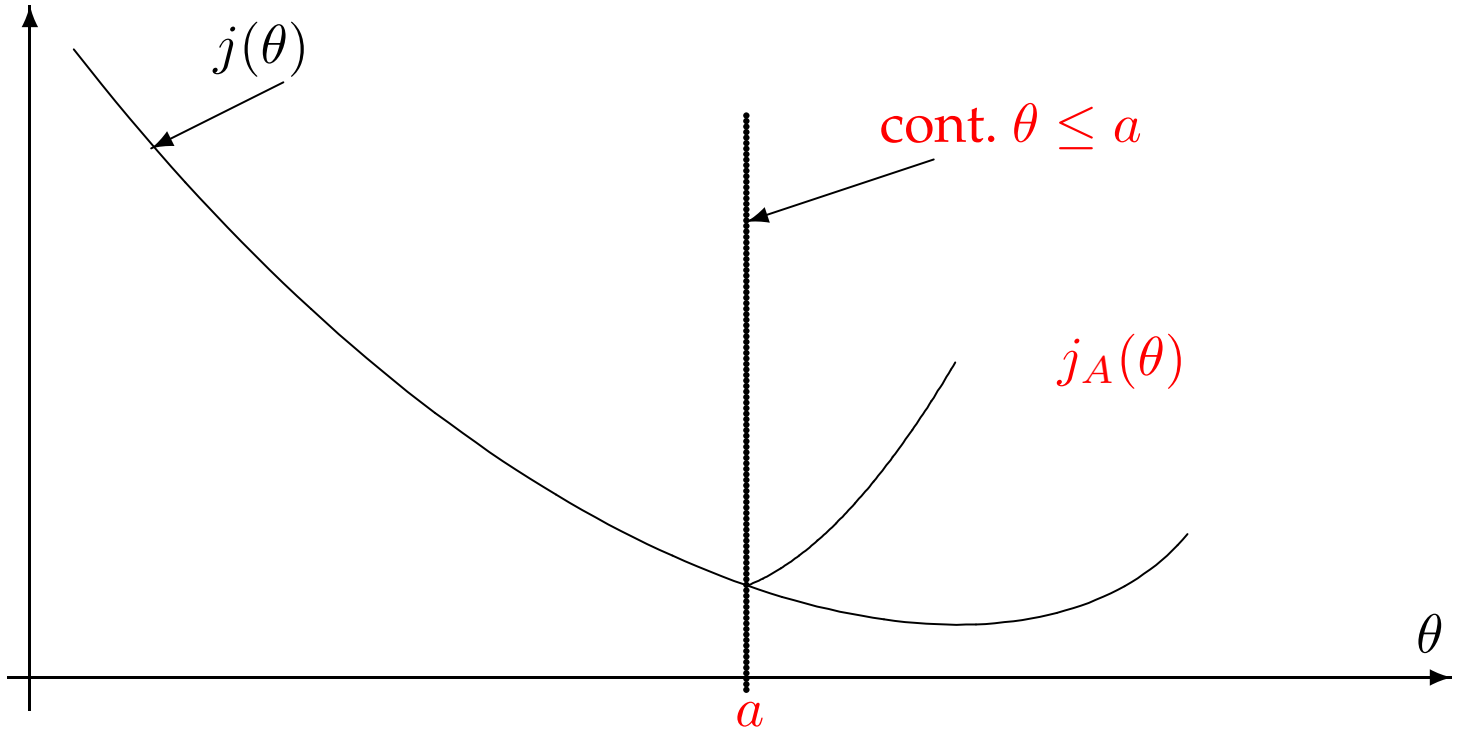
Pénalisation extérieure :

$j_A(\theta) = j(\theta) + (A/2)\|\mathbf{c}_i^+(\theta)\|^2$: contrainte violée à l'optimum, $\forall A$



Problème : la tangente de la pénalisation $(A/2)\|\mathbf{c}_i^+(\theta)\|^2$ en $\theta = a$ est nulle

$\rightarrow$ utiliser une pénalisation exacte : $j_A(\theta) = j(\theta) + A\|\mathbf{c}_i^+(\theta)\|$



Pénalisation intérieure : reparamétrer

Ex : $\theta \leq a \rightarrow \theta = a - \exp(\theta')$ ou $\theta = a - \theta'^2$ ...

6.1) Égalités

Pénalisation : comme pour contraintes inégalités
$j_A(\theta) = j(\theta) + (A/2)\|\mathbf{c}_e(\theta)\|^2$
ou $j_A(\theta) = j(\theta) + A\|\mathbf{c}_e(\theta)\|$ (pénalisation exacte)

Lagangien : $\mathcal{L}(\theta,\mathbf{d}) = j(\theta) + \mathbf{d}^\top \mathbf{c}_e(\theta)$

$\rightarrow$ minimum sur θ maximum sur $\mathbf{d}$

Lagangien augmenté : $\mathcal{L}(\theta,\mathbf{d}) = j(\theta) + \mathbf{d}^\top \mathbf{c}_e(\theta) + (A/2)\|\mathbf{c}_e(\theta)\|^2$

Alterner
des minimisations sur θ : $\hat{\theta}^k = \arg\min_\theta \mathcal{L}(\theta,\mathbf{d}^{k-1})$
des maximisations sur $\mathbf{d}$: $\mathbf{d}^k = \mathbf{d}^{k-1} + A\mathbf{c}_e(\hat{\theta}^k)$ (gradient à pas fixe A)

7) Estimation robuste

7.1) Robustesse par rapport à la loi des erreurs : variance

Si $y(i) = y_m(\theta,i) + \epsilon_i$, avec $(\epsilon_i)_i$ i.i.d. $\pi(\epsilon)$, matrice d'inf. de Fisher :

$$\begin{aligned}
\mathbf{M}(\theta,\xi) &= \mathtt{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \log \pi[y(i)|\theta]}{\partial\theta} \frac{\partial \log \pi[y(i)|\theta]}{\partial\theta^\top} \;|\theta \right\} \\
&= \frac{1}{N}\sum_{i=1}^{N} \left[\int \left(\frac{d \log \pi(e)}{de} \right)^2 \pi(e) de \right] \frac{\partial y_m(\theta,i)}{\partial\theta} \frac{\partial y_m(\theta,i)}{\partial\theta^\top} \\
&= \mathcal{I}(\pi) \frac{1}{N}\sum_{i=1}^{N} \frac{\partial y_m(\theta,i)}{\partial\theta} \frac{\partial y_m(\theta,i)}{\partial\theta^\top}
\end{aligned}$$

avec $\mathcal{I}(\pi)$ l'information de Fisher de la loi π : $\mathcal{I}(\pi) = \int \left(\frac{d\pi(e)}{de} \right)^2 \frac{1}{\pi(e)} de$

Propriétés asymptotiques de $\hat{\theta}_{MV}$: grande précision $\rightarrow$ «grande» matrice $\mathbf{M}(\theta,\xi)$

Approche minimax : MV pour la pire distribution dans une classe donnée Π

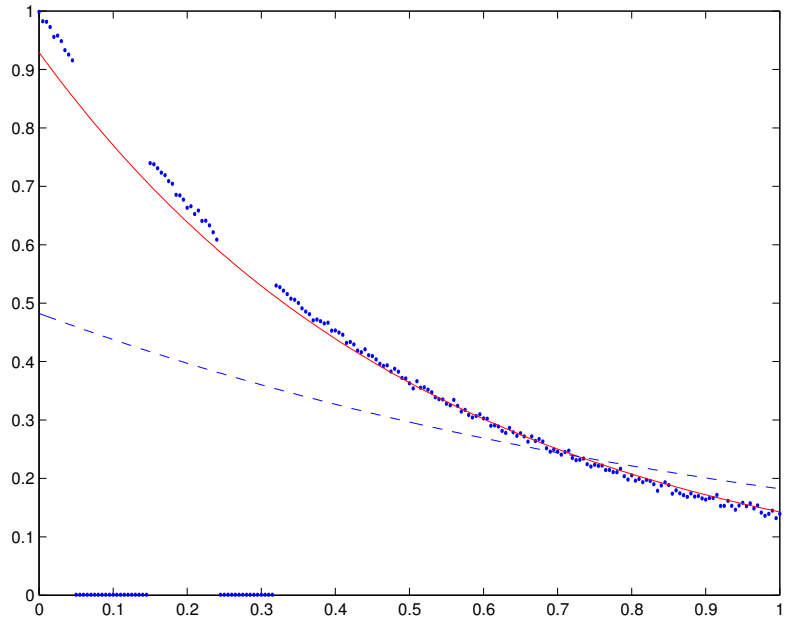
$$\hat{\pi} = \arg\min_{\pi\in\Pi} \mathcal{I}(\pi)$$

Ex1 : $\Pi = \{\pi / \mathcal{I}(\pi)$ existe et $\pi(0) \geq 1/(s\sqrt{2}) > 0\}$

$\rightarrow \hat{\pi}$ = loi de Laplace, de moyenne 0 et variance s^2

Ex2 : $\Pi = \{\pi / \mathcal{I}(\pi)$ existe et $\text{var}(\pi) \leq \sigma^2\}$

$\rightarrow \hat{\pi}$ = loi normale, de moyenne 0 et variance σ^2



$y_m(\hat{\theta}_{MC},i)$ (- - -), $y_m(\hat{\theta}_{MVA},i)$ (—)

Observations

$y(i) = \bar{\theta}_1 \exp(-\bar{\theta}_2 t_i) + \epsilon_i$,

$\bar{\theta} = (1,2)^\top$,

$(\epsilon_i)_i$ i.i.d. $\mathcal{N}(0,\sigma^2)$,

$\sigma = 5\ 10^{-3}$

t_i : 200 points dans $[0,1]$

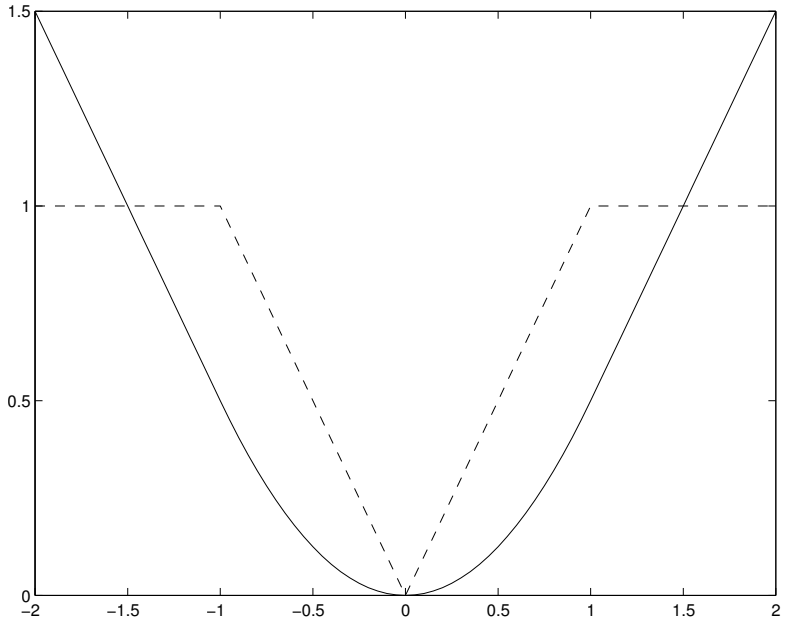
Données aberrantes : $y(i) = 0$ de $i = 11$ à $i = 30$, puis de $i = 50$ à $i = 64$

7.2) M-estimateurs

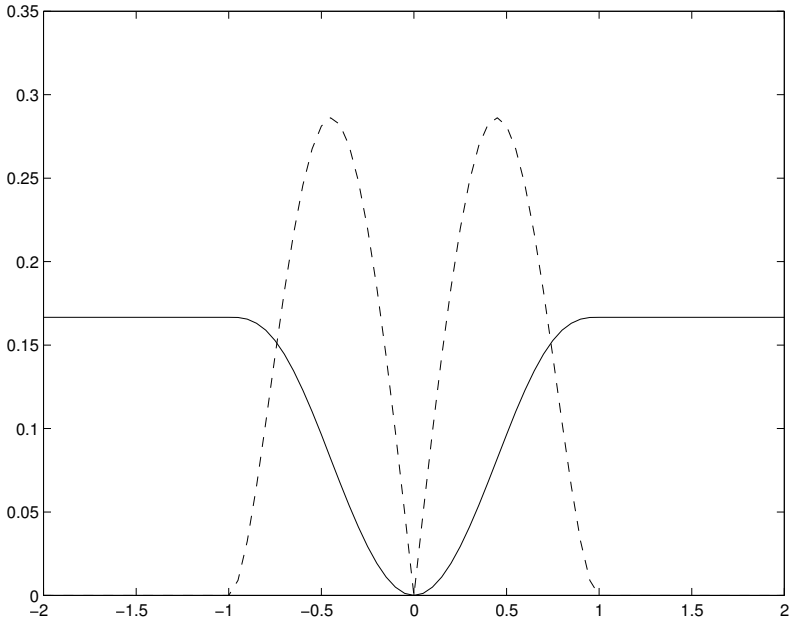
Moindres carrés, moindre valeurs absolues : cas particulier de

$$j(\theta) = \sum_{i=1}^{N} \rho[e(\theta,i)]$$

(avec $e(\theta,i) = y(i) - y_m(\theta,i)$ si erreur de sortie)



$\rho_1(e)$ (—), $|d\rho_1(e)/de|$ (- - -)



$\rho_2(e)$ (—), $|d\rho_2(e)/de|$ (- - -)

Estimateur non-redescendant : Huber

$$\rho_1(e) = \begin{cases} e^2/2 & \text{si } |e| \leq \delta \\ \delta|e| - \delta^2/2 & \text{sinon} \end{cases}$$

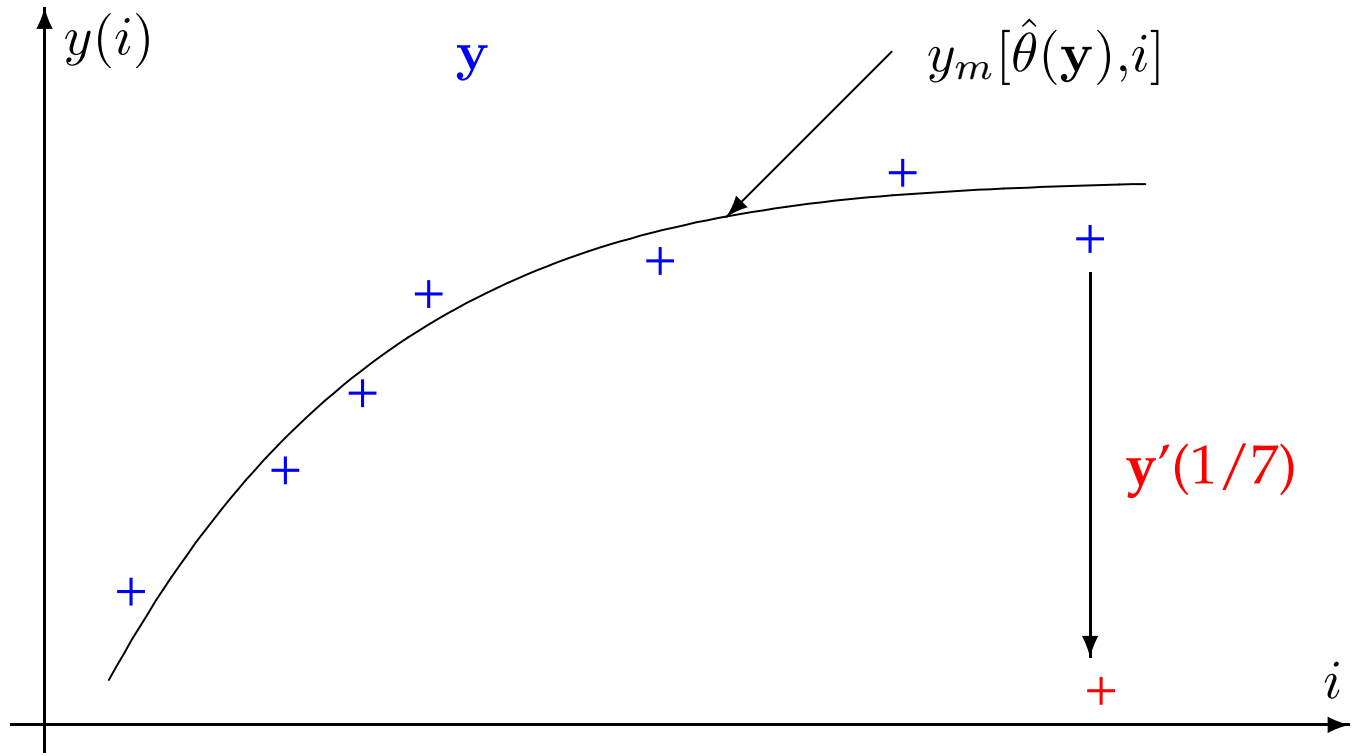
Estimateur redescendant : Tukey

$$\rho_2(e) = \begin{cases} \frac{1}{2}\left(e^2 - \frac{e^4}{\delta} + \frac{e^6}{3\delta^4}\right) & \text{si } |e| \leq \delta \\ \delta^2/6 & \text{sinon} \end{cases}$$

7.3) Point de rupture : données aberrantes $\rightarrow$ biais

$\mathbf{y}$ données correctes $\rightarrow \mathbf{y}'(\alpha)$ avec $\alpha\%$ de données aberrantes

$$\text{biais } = \mathbf{b}[\mathbf{y},\alpha,\hat{\theta}(\cdot)] = \max_{\mathbf{y}'(\alpha)} \|\hat{\theta}(\mathbf{y}) - \hat{\theta}[\mathbf{y}'(\alpha)]\|$$

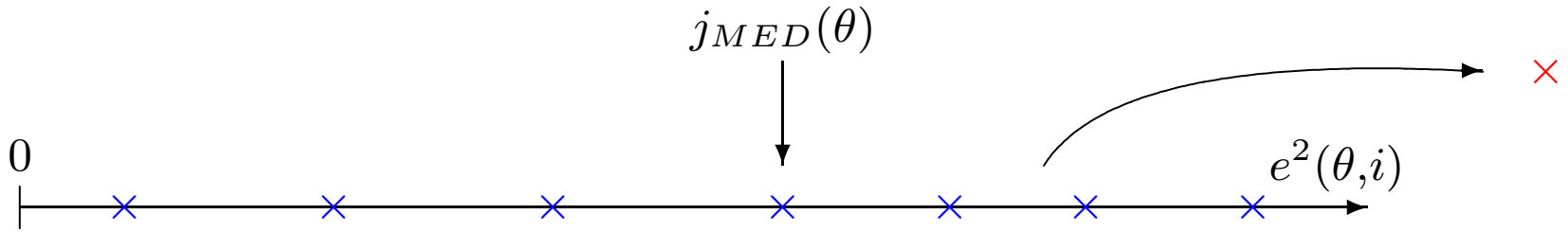


Point de rupture : $\mathrm{PR}[\mathbf{y},\hat{\theta}(\cdot)] = \min\{\alpha / \mathbf{b}[\mathbf{y},\alpha,\hat{\theta}(\cdot)] = \infty\}$

$\mathrm{PR}[\mathbf{y},\hat{\theta}_{MC}(\cdot)] = \mathrm{PR}[\mathbf{y},\hat{\theta}_{MVA}(\cdot)] = 1/N \to 0$ quand $N \to \infty$!

Médiane du carré des erreurs :

$j_{MED}(\theta) = \text{med}\{e^2(\theta,i), i = 1,\ldots,N\}$



ou somme des h + petits carrés d'erreurs :

$e_c^2(\theta,1) \leq e_c^2(\theta,2) \leq \cdots \leq e_c^2(\theta,N)$, puis

$j_{Sh}(\theta) = \sum_{i=1}^{h} e_c^2(\theta,i)$

$\text{PR}[\mathbf{y},\hat{\theta}_{MED}(\cdot)] = ([N/2] - p + 2)/N$

$h = [N/2] + 1 \rightarrow \text{PR}[\mathbf{y},\hat{\theta}_{Sh}(\cdot)] = ([N/2] - p + 2)/N$

$h = [N/2] + [(p+1)/2] \rightarrow \text{PR}[\mathbf{y},\hat{\theta}_{Sh}(\cdot)] = ([(N-p)/2] + 1)/N$ =Borne sup!

et $\lim_{N\rightarrow\infty} \text{PR}[\mathbf{y},\hat{\theta}_{MED}(\cdot)] = \lim_{N\rightarrow\infty} \text{PR}[\mathbf{y},\hat{\theta}_{Sh}(\cdot)] = 1/2$

7.4) Application en traitement d'images

T_θ : Image $A \rightarrow$ Image $T_\theta(A)$, avec θ paramètres de translation, rotation, facteur d'échelle, changement de niveaux de gris...

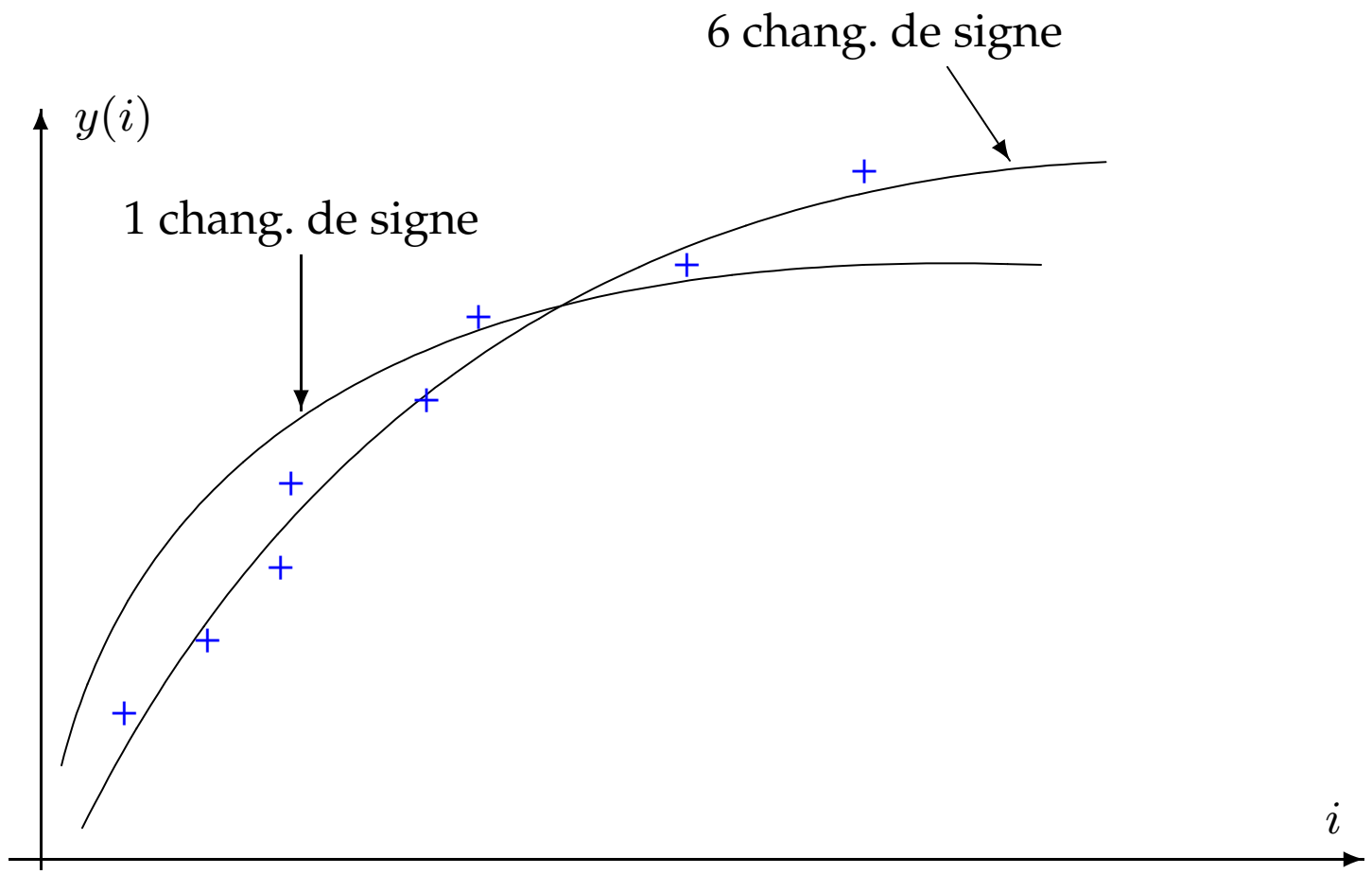
2 images A et B, calibration automatique : transformer B pour qu'elle ressemble le + possible à A (imagerie médicale)

Considérer B comme $T_\theta(A)+$ bruit, et estimer θ qui maximise un critère de ressemblance entre A et B

Problème : $\exists$ différences «significatives», non explicables par T_θ !
$\Leftrightarrow$ données aberrantes !

Transformer A et B en un signal à une dimension (par ex., balayage horizontal),

$$\hat{\theta} = \arg\max_{\theta} \text{Nb. de changements de signe dans } [T_\theta(A) - B]$$



Rapide à calculer, robuste vis-à-vis de données aberrantes