Estimation par mélange de Gibbs

Christophe Giraud

Université de Nice-Sophia Antipolis

INRA Jouy-en-Josas, 12 février 2007

Un exemple illustratif

Observations

60 observations bruitées d'un signal

$$Y_i = f(t_i) + \sigma\varepsilon_i, \qquad i = 1, \dots, 60$$

avec

- $t_i = i/60$, pour $i = 1, \dots, 60$,
- les ε_i i.i.d. $\mathcal{N}(0, 1)$,
- le niveau de bruit σ inconnu.

Observations

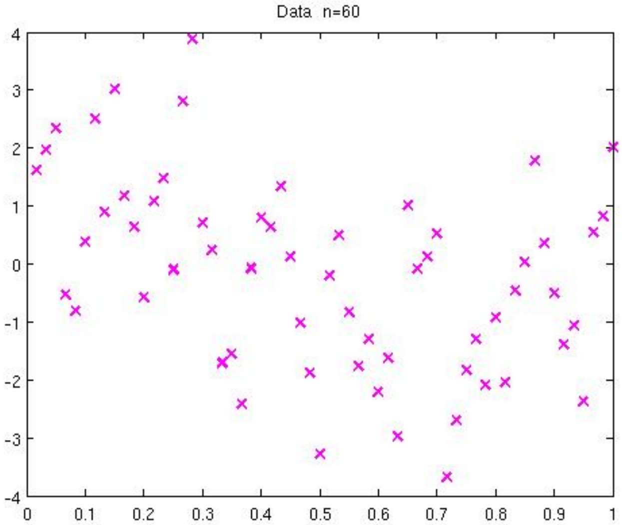
60 observations bruitées d'un signal

$$Y_i = f(t_i) + \sigma\varepsilon_i, \qquad i = 1, \dots, 60$$

avec

- $t_i = i/60$, pour $i = 1, \dots, 60$,
- les ε_i i.i.d. $\mathcal{N}(0, 1)$,
- le niveau de bruit σ inconnu.

Observations



Les moindres carrés

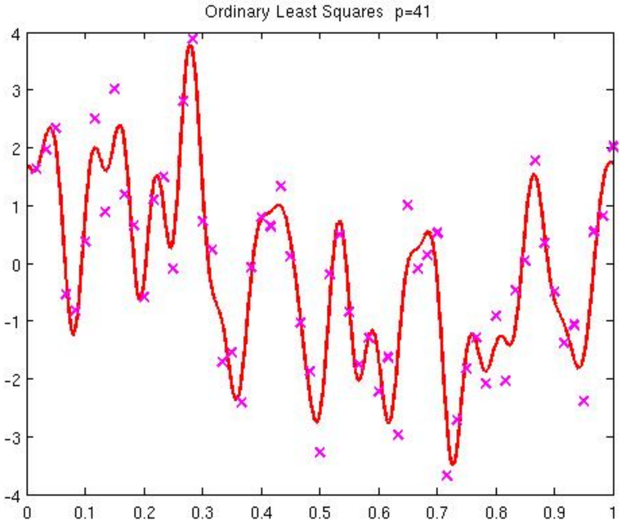
Estimateur "naïf "

$$\hat{f}(t) = \hat{a}_0 + \sum_{j=1}^{20} \hat{a}_j \cos(2\pi j t) + \sum_{j=1}^{20} \hat{b}_j \sin(2\pi j t)$$

avec $(\hat{a}_j, \hat{b}_j)$ minimisant les moindres carrés

$$\sum_{i=1}^{n} (Y_i - \hat{f}(t_i))^2$$

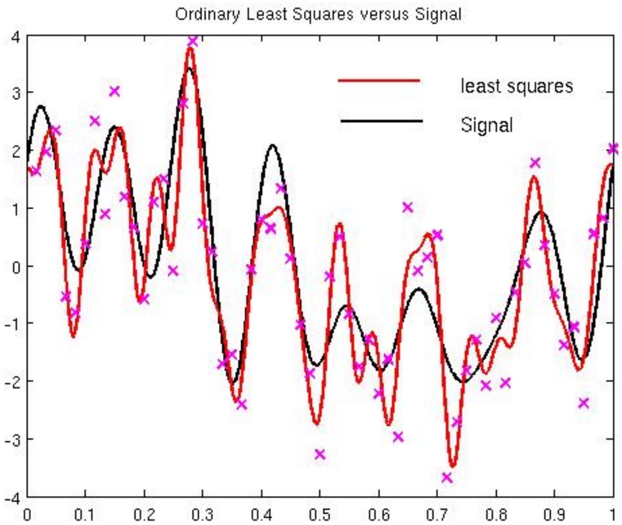
Les moindres carrés



comparaison au signal...

$$f(t) = 0.7\cos(t) + \cos(7t) + 1.2\sin(t) + 0.8\sin(5t) + 0.9\sin(8t)$$

Moindres carrés versus Signal



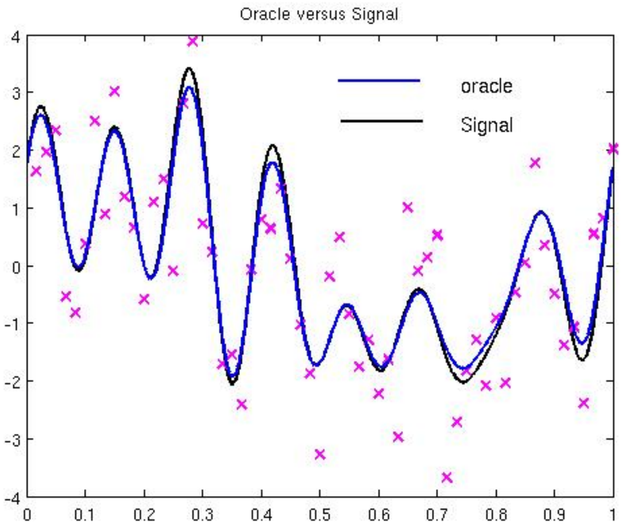
Avec l'aide d'un oracle...

Un oracle sugère de chercher $\hat{f}$ sous la forme

$$\hat{f}(t) = \hat{a}_1 \cos(t) + \hat{a}_7 \cos(7t) + \hat{b}_1 \sin(t) + \hat{b}_5 \sin(5t) + \hat{b}_8 \sin(8t)$$

avec $(\hat{a}_1, \hat{a}_7, \hat{b}_1, \hat{b}_5, \hat{b}_8)$ minimisant les moindres carrés...

Oracle versus Signal



Et sans oracle?

Collection d'estimateurs

$$\hat{f}_m(t) = \sum_{j \in m^+} \hat{a}_j \cos(2\pi j t) + \sum_{j \in m^-} \hat{b}_{|j|} \sin(2\pi |j| t)$$

indexée par $m = m^+ \cup m^- \in \mathcal{P}\big(\{-20, \ldots, 20\}\big)$,
avec $(\hat{a}_j, \hat{b}_j)$ minimisant les moindres carrés.

Quel estimateur $\hat{f}$ choisir?

Et sans oracle?

Collection d'estimateurs

$$\hat{f}_m(t) = \sum_{j \in m^+} \hat{a}_j \cos(2\pi j t) + \sum_{j \in m^-} \hat{b}_{|j|} \sin(2\pi |j| t)$$

indexée par $m = m^+ \cup m^- \in \mathcal{P}\big(\{-20, \ldots, 20\}\big)$,
avec $(\hat{a}_j, \hat{b}_j)$ minimisant les moindres carrés.

Quel estimateur $\hat{f}$ choisir?

Première possibilité: Sélection d'un modèle

Choix de $\hat{f} = \hat{f}_{\hat{m}}$ avec $\hat{m}$ obtenu par un critère de sélection

Exemple. choix de $\hat{m}$ minimisant

$$\mathrm{crit}(m) = (1 + \mathrm{pen}(m)) \sum_{i=1}^{n} (Y_i - \hat{f}_m(t_i))^2$$

avec $\mathrm{pen}(m) = 2|m|/41$, $2|m|\log(41)/41$, etc...

Première possibilité: Sélection d'un modèle

Choix de $\hat{f} = \hat{f}_{\hat{m}}$ avec $\hat{m}$ obtenu par un critère de sélection

Exemple. choix de $\hat{m}$ minimisant

$$\mathrm{crit}(m) = (1 + \mathrm{pen}(m)) \sum_{i=1}^{n} (Y_i - \hat{f}_m(t_i))^2$$

avec $\mathrm{pen}(m) = 2|m|/41$, $2|m|\log(41)/41$, etc...

Une alternative: mélange des estimateurs

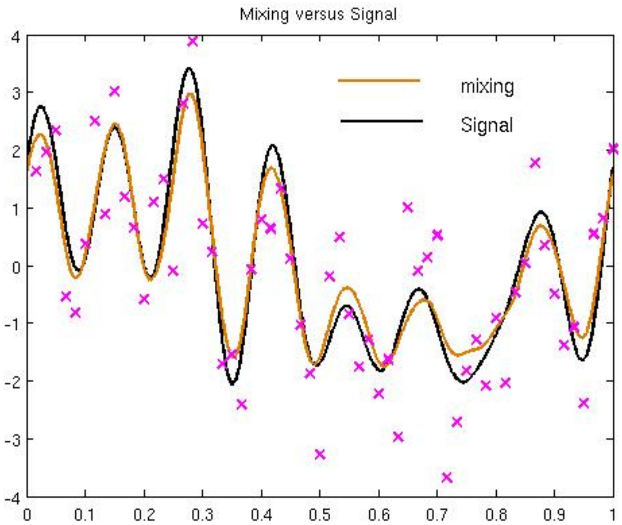
Choix d'une combinaison convexe (ou linéaire) des $\hat{f}_m$

$$\hat{f} = \sum_{m \in \mathcal{M}} w_m \hat{f}_m$$

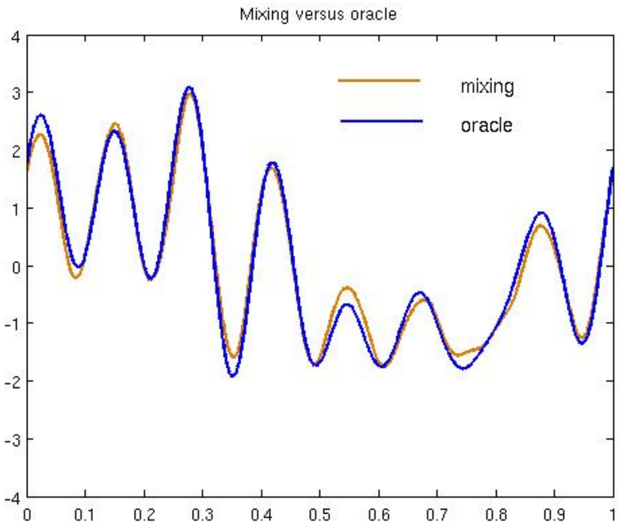
avec $\mathcal{M} = \mathcal{P}\left(\{-20, \ldots, 20\}\right)$ et les poids w_m fonctions de Y seulement.

Quels w_m?

Mélange de Gibbs versus Signal



Mélange de Gibbs versus Oracle



Mélange de Gibbs

Cadre statistique

Observations: $Y_i = \mu_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$

avec

- $\mu = (\mu_1, \ldots, \mu_n)' \in \mathbb{R}^n$ et $\sigma > 0$ inconnus
- $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. de loi $\mathcal{N}(0, 1)$.

Collection de modèles

Collection de modèles / estimateurs

- une collection finie $\{S_m,\ m \in \mathcal{M}\}$ de sous-espace vectoriels de $\mathbb{R}^n$ (modèles)
- $\hat{\mu}_m = \Pi_{S_m} Y$.

Risque L^2: $\mathbb{E}\left(\|\mu - \hat{\mu}_m\|^2\right) = \|\mu - \Pi_{S_m}\mu\|^2 + \dim(S_m)\sigma^2$

Estimateur: $\hat{\mu} = \sum_{m\in\mathcal{M}} w_m \hat{\mu}_m$. Quels w_m?

Un estimateur (pseudo-)bayésien

Loi a priori sur μ

- Tirage d'un modèle S_m selon la probabilité π_m
- Tirage de μ "uniformément" sur S_m

Estimateur Bayésien de la forme $\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}_m$ avec

$$w_m = \frac{\pi_m}{\mathcal{Z}} \exp\left(-\frac{1}{2}\left[\frac{\|Y - \hat{\mu}_m\|^2}{\sigma^2} + 2\dim(S_m)\right]\right)$$

Un estimateur (pseudo-)bayésien

Loi a priori sur μ

- Tirage d'un modèle S_m selon la probabilité π_m
- Tirage de μ "uniformément" sur S_m

Estimateur Bayésien de la forme $\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}_m$ avec

$$w_m = \frac{\pi_m}{\mathcal{Z}} \exp\left(-\frac{1}{2}\left[\frac{\|Y - \hat{\mu}_m\|^2}{\sigma^2} + 2\dim(S_m)\right]\right)$$

Mélange de Gibbs

- Une mesure de Gibbs

$$w_m = \frac{\pi_m}{\mathcal{Z}} \exp(-\beta \hat{U}_m), \quad m \in \mathcal{M},\ \beta > 0$$

minimise "l'énergie libre"

$$\hat{F}_\beta(w) = \sum_{m\in\mathcal{M}} w_m \hat{U}_m + \frac{1}{\beta}\mathcal{D}(w|\pi)$$

où $\mathcal{D}(w|\pi) = \sum_{m\in\mathcal{M}} w_m \log\left(w_m/\pi_m\right)$.

- **Idée:** prendre
  - $\hat{U}_m$ un estimateur du risque de $\hat{\mu}_m$
  - π_m poids fonction de la complexité de S_m

Mélange de Gibbs

- Une mesure de Gibbs

$$w_m = \frac{\pi_m}{\mathcal{Z}} \exp(-\beta \hat{U}_m), \quad m \in \mathcal{M}, \ \beta > 0$$

minimise "l'énergie libre"

$$\hat{F}_\beta(w) = \sum_{m \in \mathcal{M}} w_m \hat{U}_m + \frac{1}{\beta} \mathcal{D}(w|\pi)$$

où $\mathcal{D}(w|\pi) = \sum_{m \in \mathcal{M}} w_m \log\left(w_m/\pi_m\right)$.

- **Idée:** prendre
  - $\hat{U}_m$ un estimateur du risque de $\hat{\mu}_m$
  - π_m poids fonction de la complexité de S_m

Choix de $\hat{U}_m$

Hypothèse: $S_m \subset S_*, \ \forall\, m \in \mathcal{M}$, avec $\dim(S_*) < n$

Estimation variance: $\hat{\sigma}^2 = \|Y - \Pi_{S_*} Y\|^2 / N_*$ où $N_* = n - \dim(S_*)$

Choix de $\hat{U}_m$:

$$\hat{U}_m = \frac{\|\Pi_{S_*} Y - \hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{1}{\beta} L_m \qquad \text{où } L_m \geq \dim(S_m)/2.$$

Théorème

Sous les conditions

$$n \geq 3, \quad \beta < 1/4 \quad \text{and} \quad N_* \geq 2 + \frac{\log n}{\phi(4\beta)},$$

avec $\phi(x) = (x - 1 - \log x)/2$, on a

$$\begin{aligned}
&\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right) \\
&\leq \; -(1+\varepsilon_n)\,\frac{\bar{\sigma}^2}{\beta}\,\log\left[\sum_{m\in\mathcal{M}} \pi_m e^{-\beta\|\mu-\Pi_{S_m}\mu\|^2/\bar{\sigma}^2 - L_m}\right] + \frac{\sigma^2}{2\log n} \\
&\leq \; (1+\varepsilon_n) \inf_{m\in\mathcal{M}} \left\{\|\mu - \Pi_{S_m}\mu\|^2 + \frac{\bar{\sigma}^2}{\beta}\left(L_m - \log \pi_m\right)\right\} + \frac{\sigma^2}{2\log n}
\end{aligned}$$

où $\varepsilon_n = (2n\log n)^{-1}$ et $\bar{\sigma}^2 = \sigma^2 + \|\mu - \Pi_{S_*}\mu\|^2/N_*$.

Théorème

Sous les conditions

$$n \geq 3, \quad \beta < 1/4 \quad \text{and} \quad N_* \geq 2 + \frac{\log n}{\phi(4\beta)},$$

avec $\phi(x) = (x - 1 - \log x)/2$, on a

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right)$$

$$\leq \quad -(1+\varepsilon_n)\,\frac{\bar{\sigma}^2}{\beta}\,\log\left[\sum_{m\in\mathcal{M}} \pi_m e^{-\beta\|\mu-\Pi_{S_m}\mu\|^2/\bar{\sigma}^2 - L_m}\right] + \frac{\sigma^2}{2\log n}$$

$$\leq \quad (1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu - \Pi_{S_m}\mu\|^2 + \frac{\bar{\sigma}^2}{\beta}\left(L_m - \log \pi_m\right)\right\} + \frac{\sigma^2}{2\log n}$$

où $\varepsilon_n = (2n\log n)^{-1}$ et $\bar{\sigma}^2 = \sigma^2 + \|\mu - \Pi_{S_*}\mu\|^2/N_*$.

Théorème

Sous les conditions

$$n \geq 3, \quad \beta < 1/4 \quad \text{and} \quad N_* \geq 2 + \frac{\log n}{\phi(4\beta)},$$

avec $\phi(x) = (x - 1 - \log x)/2$, on a

$$\begin{aligned}
&\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right) \\
&\leq \quad -(1+\varepsilon_n)\,\frac{\bar{\sigma}^2}{\beta}\,\log\left[\sum_{m\in\mathcal{M}} \pi_m e^{-\beta\|\mu-\Pi_{S_m}\mu\|^2/\bar{\sigma}^2 - L_m}\right] + \frac{\sigma^2}{2\log n} \\
&\leq \quad (1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu - \Pi_{S_m}\mu\|^2 + \frac{\bar{\sigma}^2}{\beta}\left(L_m - \log \pi_m\right)\right\} + \frac{\sigma^2}{2\log n}
\end{aligned}$$

où $\varepsilon_n = (2n\log n)^{-1}$ et $\bar{\sigma}^2 = \sigma^2 + \|\mu - \Pi_{S_*}\mu\|^2/N_*$.

Esquisse de preuve

Faits:

1. $\mathbb{E}\left[\frac{\|\mu-\hat{\mu}\|^2}{\sigma^2}\right] \leq \frac{\|\mu-\Pi_{S_*}\mu\|^2}{\sigma^2} + \mathbb{E}\left[(1+\hat{\varepsilon}_n)\hat{F}_\beta(w)\right] - \dim(S_*)$
2. $\hat{F}_\beta(w) \leq \hat{F}_\beta(\alpha)$ pour toute probabilité α sur $\mathcal{M}$.

d'où

$$\mathbb{E}\left[\|\mu-\hat{\mu}\|^2\right] \leq$$

$$(1+\varepsilon_n)\bar{\sigma}^2\left[\sum_{m\in\mathcal{M}}\alpha_m\left[\frac{\|\mu-\Pi_{S_m}\mu\|^2}{\bar{\sigma}^2}+\frac{L_m}{\beta}\right]+\frac{\mathcal{D}(\alpha|\pi)}{\beta}\right]+n\varepsilon_n\sigma^2$$

$+$ **optimisation en** α**.**

Choix des paramètres de l'exemple

- Famille de modèles indexée par $\mathcal{M} = \mathcal{P}\left(\{-20, \ldots, 20\}\right)$

- Distribution $\pi_m = (1 + 1/p)^{-p} p^{-|m|}$ avec $p = 41$, et

$$w_m \propto p^{-|m|} \exp\left(\beta \|\hat{\mu}_m\|^2 / \hat{\sigma}^2 - |m|\right), \quad \text{avec } \beta = 1/3$$

- Problème: $|\mathcal{M}| = 2^{41} \approx 2000$ milliards...

Choix des paramètres de l'exemple

- Famille de modèles indexée par $\mathcal{M} = \mathcal{P}\left(\{-20, \ldots, 20\}\right)$

- Distribution $\pi_m = (1 + 1/p)^{-p} p^{-|m|}$ avec $p = 41$, et

$$w_m \propto p^{-|m|} \exp\left(\beta \|\hat{\mu}_m\|^2 / \hat{\sigma}^2 - |m|\right), \quad \text{avec } \beta = 1/3$$

- Problème: $|\mathcal{M}| = 2^{41} \approx 2000$ milliards…

Un seuillage mou

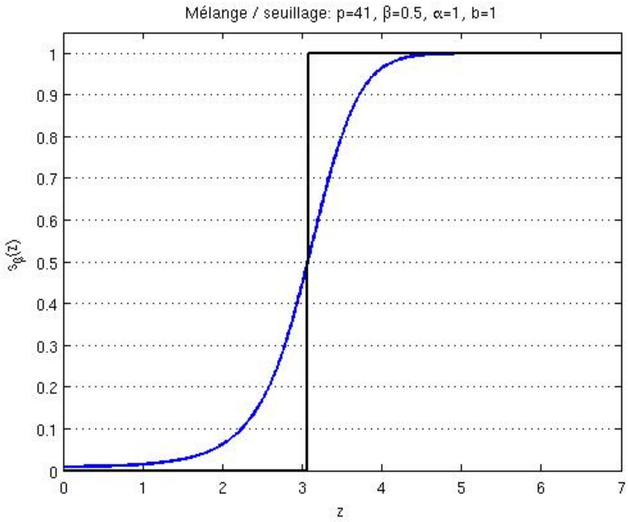
Dans le cas particulier où:

1. $\mathcal{M} = \mathcal{P}(\{1, \ldots, p\})$
2. $S_m = \mathrm{span}\left\{\overrightarrow{v_j},\ j \in m\right\}$ avec $\left\{\overrightarrow{v_1}, \ldots, \overrightarrow{v_p}\right\}$ famille o.n. de $\mathbb{R}^n$
3. $\pi_m = (1 + p^{-\alpha})^{-p}\, p^{-\alpha|m|}$, pour $m \in \mathcal{M}$, $\alpha > 0$
4. $L_m = b|m|$, avec $b \geq 0$

on a $\hat{\mu} = \sum_{j=1}^{p} s_\beta(Z_j/\hat{\sigma}) Z_j\, \overrightarrow{v_j}$ où

$$Z_j = < Y, \overrightarrow{v_j} > \quad \text{et} \quad s_\beta(z) = \frac{e^{\beta z^2}}{p^\alpha e^b + e^{\beta z^2}}\,.$$

Coefficient de "shrinkage" $s_\beta(z)$, pour $p = 41$



Mélange versus Sélection

Mélange versus Sélection

Points forts	Points faibles
• constante $1 + \varepsilon_n$ • combinaison convexe (mélange de modèles)	• ne sélectionne pas un modèle • Hypothèse: $S_m \subset S_*$, $\forall\, m \in \mathcal{M}$

Problème numérique: $|\mathcal{M}|$ potentiellement très grand...

Seuillage mou versus Seuillage: rapport des risques 1-D