Gaussian Model Selection with Unknown Variance

Y. Baraud, C. Giraud and S. Huet

Université de Nice - Sophia Antipolis,
INRA Jouy en Josas

Luminy, 13-17 novembre 2006
The statistical setting

The statistical model

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

- $\mu = (\mu_1, \ldots, \mu_n)' \in \mathbb{R}^n$ and $\sigma > 0$ are unknown

- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d standard Gaussian

Collection of models / estimators

- $S = \{S_m, m \in M\}$ a countable collection of linear subspaces of $\mathbb{R}^n$ (models)

- $\hat{\mu}_m =$ least-squares estimator of $\mu$ on $S_m$
Example: change-points detection

- $\mu_i = f(x_i)$ with $f : [0, 1] \mapsto \mathbb{R}$, piecewise constant.

- $\mathcal{M}$ is the set of increasing sequences $m = (t_0, \ldots, t_q)$ with $q \in \{1, \ldots, p\}$, $t_0 = 0$, $t_q = 1$, and $\{t_1, \ldots, t_{q-1}\} \subset \{x_1, \ldots, x_n\}$.

- models:

$$S_m = \{(g(x_1), \ldots, g(x_n))', \ g \in \mathcal{F}_m\},$$

where

$$\mathcal{F}_{(t_0, \ldots, t_q)} = \left\{ g = \sum_{j=1}^{q} a_j 1_{[t_{j-1}, t_j[} \text{ with } (a_1, \ldots, a_q) \in \mathbb{R}^q \right\}.$$

- No residual squares to estimate the variance.
Risk on a single model

Euclidean risk on $S_m$:

$$
\mathbb{E} \left[ \| \mu - \hat{\mu}_m \|^2 \right] = \underbrace{\| \mu - \mu_m \|^2}_{\text{bias}} + \underbrace{D_m \sigma^2}_{\text{variance}}
$$

Ideal: estimate $\mu$ with $\hat{\mu}_{m^*}$, where $m^*$ minimizes $m \mapsto \mathbb{E} \left[ \| \mu - \hat{\mu}_m \|^2 \right]$ ...
Model selection

Selection rule: we set $D_m = \dim(S_m)$ and select $\hat{m}$ minimizing

$$\text{Crit}_L(m) = \| Y - \hat{\mu}_m \|_2^2 \left( 1 + \frac{\text{pen}(m)}{n - D_m} \right)$$

(1)

or

$$\text{Crit}_K(m) = \frac{n}{2} \log \left( \frac{\| Y - \hat{\mu}_m \|_2^2}{n} \right) + \frac{1}{2} \text{pen}'(m).$$

(2)

Some classical penalties:

<table>
<thead>
<tr>
<th></th>
<th>FPE</th>
<th>AIC</th>
<th>BIC</th>
<th>AMDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>pen$(m)$</td>
<td>$2D_m$</td>
<td>$2D_m$</td>
<td>$D_m \log n$</td>
<td>$3D_m \log n$</td>
</tr>
<tr>
<td>pen'(m)</td>
<td>$2D_m$</td>
<td>$2D_m$</td>
<td>$D_m \log n$</td>
<td>$3D_m \log n$</td>
</tr>
</tbody>
</table>
Model selection

Selection rule: we select $\hat{m}$ minimizing

$$\text{Crit}_L(m) = \|Y - \hat{\mu}_m\|^2 \left(1 + \frac{\text{pen}(m)}{n - D_m}\right)$$

or

$$\text{Crit}_K(m) = \frac{n}{2} \log \left(\frac{\|Y - \hat{\mu}_m\|^2}{n}\right) + \frac{1}{2} \text{pen}'(m).$$

Criteria (1) and (2) are equivalent with

$$\text{pen}'(m) = n \log \left(1 + \frac{\text{pen}(m)}{n - D_m}\right).$$
Objectives

- for classical criteria: to analyze the Euclidean risk of $\hat{\mu}_{\hat{m}}$ with regard to the complexity of the family of model $S$, and compare this risk to

$$\inf_{m \in M} \mathbb{E} [\| \mu - \hat{\mu}_m \|^2].$$

- to propose penalties versatile enough to take into account the complexity of $S$ and the sample size.

Complexity:

We say that $S$ has an index of complexity $(M, a)$ if for all $D \geq 1$

$$\text{card} \{ m \in \mathcal{M}, \ D_m = D \} \leq M e^{aD}.$$
Theorem 1: Performances of classical penalties

Let $K > 1$ and $S$ with complexity $(M, a) \in \mathbb{R}^2_+$. If for all $m \in \mathcal{M}$,

$$D_m \leq D_{\text{max}}(K, M, a) \quad \text{(explicit)}$$

and

$$\text{pen}(m) \geq K^2 \phi^{-1}(a) D_m,$$

with $\phi(x) = (x - 1 - \log x)/2$ for $x \geq 1$, then

$$\mathbb{E} \left[ \| \mu - \hat{\mu}_m \|^2 \right] \leq \frac{K}{K - 1} \inf_{m \in \mathcal{M}} \left[ \| \mu - \mu_m \|^2 \left( 1 + \frac{\text{pen}(m)}{n - D_m} \right) + \text{pen}(m) \sigma^2 \right] + R$$

where

$$R = \frac{K \sigma^2}{K - 1} \left[ K^2 \phi^{-1}(a) + 2K + \frac{8KM e^{-a}}{\left( e^{\phi(K)/2} - 1 \right)^2} \right].$$
Performances of $\hat{\mu}_m$

- under the above hypotheses if $\text{pen}(m) = K \phi^{-1}(a) D_m$ with $K > 1$

$$
\mathbb{E} \left[ \|\mu - \hat{\mu}_m\|^2 \right] \leq c(K, M) \phi^{-1}(a) \left[ \inf_{m \in M} \mathbb{E} \left[ \|\mu - \hat{\mu}_m\|^2 \right] + \sigma^2 \right]
$$

- The condition $\text{pen}(m) \geq K^2 \phi^{-1}(a) D_m$ with $K > 1$" is sharp (at least when $a = 0$ and $a = \log n$).

Roughly, for large values of $n$ this imposes the restrictions:

<table>
<thead>
<tr>
<th>Criteria</th>
<th>FPE</th>
<th>AIC</th>
<th>BIC</th>
<th>AMDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>$a &lt; 0.15$</td>
<td>$a &lt; 0.15$</td>
<td>$a &lt; \frac{1}{2} \log(n)$</td>
<td>$a &lt; \frac{3}{2} \log(n)$</td>
</tr>
</tbody>
</table>
For \( x \geq 0 \), we define

\[
D\text{kh}
\left[D, N, x\right] = \frac{1}{\mathbb{E}(X_D)} \times \mathbb{E}\left[\left(X_D - x \frac{X_N}{N}\right)_+\right] \in [0, 1].
\]

where \( X_D \) and \( X_N \) are two independent \( \chi^2(D) \) and \( \chi^2(N) \).

**Computation:** \( x \mapsto D\text{kh}[D, N, x] \) is decreasing and

\[
D\text{kh}[D, N, x] = \mathbb{P}\left(F_{D+2, N} \geq \frac{x}{D + 2}\right) - \frac{x}{D} \mathbb{P}\left(F_{D, N+2} \geq \frac{(N + 2)x}{DN}\right),
\]

where \( F_{D, N} \) is a Fischer random variable with \( D \) and \( N \) degrees of freedom.
Theorem 2: a general risk bound

Let $\text{pen}$ be an arbitrary non-negative penalty function and assume that $N_m = n - D_m \geq 2$ for all $m \in M$. If $\hat{m}$ exists a.s., then for any $K > 1$

$$
\mathbb{E} \left[ \| \mu - \hat{\mu}_{\hat{m}} \|^2 \right] \leq \frac{K}{K - 1} \inf_{m \in M} \left[ \| \mu - \mu_m \|^2 \left( 1 + \frac{\text{pen}(m)}{N_m} \right) + \text{pen}(m) \sigma^2 \right] + \Sigma
$$

where

$$
\Sigma = \frac{K^2 \sigma^2}{K - 1} \sum_{m \in M} (D_m + 1) \text{Dkhi} \left[ D_m + 1, N_m - 1, \frac{N_m - 1}{KN_m} \text{pen}(m) \right].
$$
Minimal penalties

• Choose \( K > 1 \) and \( \mathcal{L} = \{L_m, m \in \mathcal{M}\} \) non-negative numbers (weights) such that

\[
\Sigma' = \sum_{m \in \mathcal{M}} (D_m + 1) e^{-L_m} < +\infty.
\]

• For any \( m \in \mathcal{M} \) set

\[
\text{pen}^L_{K, \mathcal{L}}(m) = K \frac{N_m}{N_m - 1} \text{Dkhi}^{-1} [D_m + 1, N_m - 1, e^{-L_m}]
\]

• When \( L_m \vee D_m \leq \kappa n \) with \( \kappa < 1 \):

\[
\text{pen}^L_{K, \mathcal{L}}(m) \leq C(K, \kappa) (L_m \vee D_m).
\]
How to choose the $L_m$?

- When $S$ has a complexity $(M, a)$: a possible choice is $L_m = aD_m + 3 \log(D_{m+1})$. Then

  $$\Sigma' = \sum_{m \in M} (D_m + 1)e^{-L_m} \leq M \sum_{D \geq 1} D^{-2}$$

- For change-point detection: We choose $L_m = L(|m|) = \log \left[ \binom{n}{|m|-2} \right] + 2 \log(|m|)$, for which

  $$\Sigma' = \sum_{D=2}^{p+1} \binom{n}{D-2} D e^{-L(D)} = \sum_{D=2}^{p+1} \frac{1}{D} \leq \log(p + 1).$$
rouge : Dstar= 4 , vert :
DhatK= 5 , bleu : DAMDL= 5
rouge : $D_{star} = 15$, vert : $D_{hatK} = 13$, bleu : $D_{AMDL} = 8$