Gaussian model selection with an unknown variance

Yannick Baraud

Laboratoire J.A. Dieudonné
Université de Nice Sophia Antipolis
baraud@unice.fr

Joint work with C. Giraud and S. Huet

The statistical framework

We observe

$$Y \sim \mathcal{N}\left(\mu, \sigma^2 I_n\right)$$

where both parameters $\mu \in \mathbb{R}^n$ and $\sigma > 0$ are unknown.

Our aim: Estimate μ from the observation of Y.

Example : Variable selection

$$Y \sim \mathcal{N}\left(\mu, \sigma^2 I_n\right) \quad \text{with} \quad \mu = \sum_{j=1}^{p} \theta_j X_j.$$

and p possibly larger than n but expect that

$$\left|\{j,\ \theta_j \neq 0\}\right| \ll n$$

Our aim: Estimate μ and $\{j,\ \theta_j \neq 0\}$.

The estimation strategy: model selection

We start with a collection $\{S_m,\ m \in \mathcal{M}\}$ of linear subspaces (models) of $\mathbb{R}^n$.

$$S_m \longrightarrow \hat{\mu}_m = \Pi_{S_m} Y$$

Our aim : select $\hat{m} = \hat{m}(Y)$ among $\mathcal{M}$ in such a way

$$\mathbb{E}\left[|\mu - \hat{\mu}_{\hat{m}}|^2\right] \text{ close to } \inf_{m \in \mathcal{M}} \mathbb{E}\left[|\mu - \hat{\mu}_m|^2\right].$$

Variable selection (continued)

$$Y \sim \mathcal{N}\left(\mu, \sigma^2 I_n\right) \quad \text{with} \quad \mu = \sum_{j=1}^{p} \theta_j X_j$$

For $m \subset \{1, \ldots, p\}$, such that $|m| \leq D_{\max} < n$ we set

$$S_m = \mathrm{Span}\left\{X_j,\ j \in m\right\}.$$

- Ordered variable selection. Take

$$\mathcal{M}_o = \{\{1, \ldots, D\},\ D \leq D_{\max}\} \cup \{\varnothing\}$$

- (Almost) complete variable selection. Take

$$\mathcal{M}_c = \{m \subset \mathcal{P}(\{1, \ldots, p\}),\ |m| \leq D_{\max}\}$$

Some selection criteria

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \left(|Y - \hat{\mu}_m|^2 + \mathrm{pen}(m) \right)$$

- Mallows' C_p (1973): $\mathrm{pen}(m) = 2D_m\sigma^2$ where $D_m = \dim(S_m)$.
- Birgé & Massart (2001): $\mathrm{pen}(m) = \mathrm{pen}(m, \sigma^2)$.

- Advantages :
  - Non-asymptotic theory
  - Variable selection: no assumption on the predictors X_j.
  - Bayesian flavor : allows (into some extent) to take into account knowlege/intuition
- Drawbacks :
  - The computation of $\hat{m}$ may not feasible if $\mathcal{M}$ is too large

For the problem of variable selection :

- Tibshirani(1996) Lasso :

$$\hat{\theta}^{\lambda} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left\{ \left| Y - \sum_{j=1}^{p} \theta_j X_j \right|^2 + \lambda \, |\theta|_1 \right\}.$$

- Candès & Tao (2007) Dantzig selector:

$$\hat{\theta}^{\lambda} = \text{argmin} \left\{ |\theta|_1 \, , \; \max_{j=1,\dots,p} \left| \langle X_j, Y - \sum_{j'=1}^{p} \theta_{j'} X_{j'} \rangle \right| \leq \lambda \right\}$$

$$\longrightarrow \quad \hat{m}^{\lambda} = \left\{ j, \; \hat{\theta}_j^{\lambda} \neq 0 \right\} \; \text{ and } \; \hat{\mu}_{\hat{m}^{\lambda}} = \sum_{j \in \hat{m}^{\lambda}} \hat{\theta}_j^{\lambda} X_j$$

- Advantages :
  - The computation is feasible even if p is very large
  - Non-asymptotic theory
- Drawbacks :
  - The procedure work under suitable assumptions on the predictors X_j
  - There is no way to check these assumptions if p is very large
  - Blind to knowledge/intuition

For all these procedures, remains the problem of estimating σ^2 or choosing λ

- These parameters depends on the data distribution and must be estimated
- In general, there is no natural estimator of σ^2 (complete variable selection with $p > n$)
- Cross-validation...
- The performance of the procedure crucially depends upon these parameters.

Other selection criteria

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

or

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

Both criteria are the same if one takes

$$\mathrm{pen}'(m) = n \log\left(1 + \frac{\mathrm{pen}(m)}{n - D_m}\right) \approx \mathrm{pen}(m)$$

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

or

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

- Akaike(1969) FPE : $\mathrm{pen}(m) = 2D_m$
- Akaike(1973) AIC : $\mathrm{pen}'(m) = 2D_m$
- Schwarz/Akaike (1978) BIC/SIC : $\mathrm{pen}'(m) = D_m \log(n)$
- Saito(1994) AMDL : $\mathrm{pen}'(m) = 3D_m \log(n)$

Two questions

1. What can be said about these selection criteria from a non-asymptotic point of view?

2. Is it possible to propose other penalties that would take into account the complexity of the collection $\{S_m,\ m \in \mathcal{M}\}$?

What do we mean by complexity?

We shall say that that the collection $\{S_m,\ m \in \mathcal{M}\}$ is a-complex (with $a \geq 0$) if

$$|\{m \in \mathcal{M},\ D_m = D\}| \leq e^{aD} \quad \forall D \geq 1.$$

- For the collection $\{S_m,\ m \in \mathcal{M}_o\}$

$$|\{m \in \mathcal{M},\ D_m = D\}| \leq 1 \quad \Longrightarrow \quad a = 0$$

- For the collection $\{S_m,\ m \in \mathcal{M}_c\}$

$$|\{m \in \mathcal{M},\ D_m = D\}| \leq \binom{p}{D} \leq p^D \quad \Longrightarrow \quad a = \log(p)$$

Penalty choice with regard to complexity

Let $\phi(x) = (x - 1 - \log(x))/2$ for $x \geq 1$.

Consider a a-complex collection $\{S_m,\ m \in \mathcal{M}\}$. If for some $K, K' > 1$

$$K \leq \frac{\mathrm{pen}(m)}{\phi^{-1}(a) D_m} \leq K', \ \ \forall m \in \mathcal{M}^*$$

and select

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} |Y - \hat{\mu}_m|^2 \left(1 + \frac{\mathrm{pen}(m)}{n - D_m}\right)$$

then

$$\frac{\mathbb{E}\left[\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2}\right]}{\inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right] \vee 1} \ \leq \ C(K)K' \ \ \phi^{-1}(a)$$

Case of ordered variable selection

$a = 0$, $\phi^{-1}(a) = 1$. For all $m \in \mathcal{M}$ such that $D_m \neq 0$

$$1 < K \leq \frac{\mathrm{pen}(m)}{D_m} \leq K'$$

one has

$$\frac{\mathbb{E}\left[\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2}\right]}{\inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right] \vee 1} \leq C(K)K'$$

$\longrightarrow$ FPE and AIC (for n large enough)

Case of the complete variable selection with $p = n$

$a = \log(n)$, $\phi^{-1}(a) \approx 2\log(n)$. If for all $m \in \mathcal{M}$ such that $D_m \neq 0$

$$1 < K \leq \frac{\mathrm{pen}(m)}{2 D_m \log(n)} \leq K'$$

then

$$\frac{\mathbb{E}\left[\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2}\right]}{\inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right] \vee 1} \leq C(K)K' \log(n)$$

$\longrightarrow$ AMDL (but not AIC, FPE, BIC)

New penalties

Definition

Let $X_D \sim \chi^2(D)$, $X_N \sim \chi^2(N)$, be two independent χ^2. Define

$$\mathcal{H}_{D,N}(x) = \frac{1}{\mathbb{E}(X_D)} \times \mathbb{E}\left[\left(X_D - x\frac{X_N}{N}\right)_+\right], \quad x \geq 0$$

Definition

To each S_m with $D_m < n-1$, we associate a weight $L_m \geq 0$ and the penalty

$$\mathrm{pen}(m) \quad = \quad \frac{1.1 N_m}{N_m - 1} \mathcal{H}^{-1}_{D_m+1, N_m-1}\left(e^{-L_m}\right) \text{ where } N_m = n - D_m.$$

Theorem

Let $\{S_m,\ m \in \mathcal{M}\}$ *be a collection of models and* $\{L_m,\ m \in \mathcal{M}\}$ *a family of weights. Assume that* $N_m \geq 7$ *and* $D_m \vee L_m \leq n/2$ *for all* $m \in \mathcal{M}$*. Define*

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} |Y - \hat{\mu}_m|^2 \left(1 + \frac{\mathrm{pen}(m)}{n - D_m}\right)$$

The estimator $\hat{\mu}_{\hat{m}}$ *satisfies*

$$\square \times \mathbb{E}\left(\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2}\right) \leq \inf_{m \in \mathcal{M}} \left[\mathbb{E}\left(\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right) + L_m\right] + \sum_{m \in \mathcal{M}} (D_m + 1)e^{-L_m}.$$

Ordered variable selection

For $m \in \mathcal{M}_o$, $m = \{1, \ldots, D\}$,

$$L_m = |m|$$
$$\longrightarrow \quad \sum_{m \in \mathcal{M}} (D_m + 1)\, e^{-L_m} \leq 2.51$$

If $|m| \leq D_{\max} \leq [n/2] \wedge p$,

$$\mathbb{E}\left(\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2} \right) \leq \Box \inf_{m \in \mathcal{M}} \left[\mathbb{E}\left(\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2} \right) \vee 1 \right].$$

Complete Variable selection

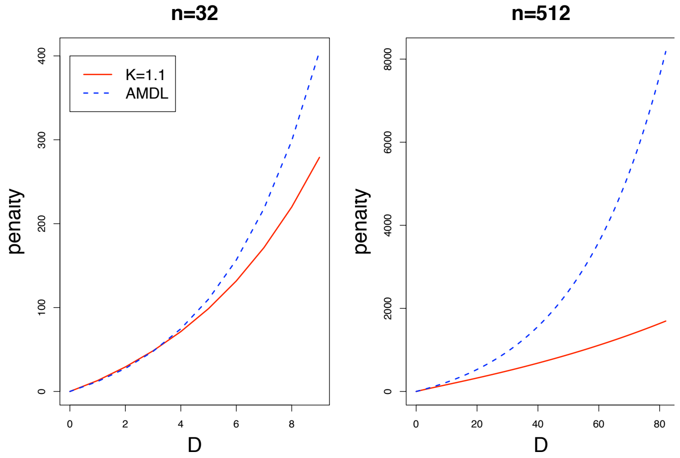
For $m \in \mathcal{M}_c$,

$$L_m = \log\left[\binom{p}{|m|}\right] + 2\log(|m|+1)$$
$$\longrightarrow \quad \sum_{m\in\mathcal{M}} (D_m+1)\, e^{-L_m} \leq \log(p).$$

If $|m| \leq D_{\max} \leq [n/(2\log(p))] \wedge p$,

$$\mathbb{E}\left(\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2}\right) \leq \square \log(p) \inf_{m\in\mathcal{M}} \left[\mathbb{E}\left(\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right) \vee 1\right].$$

Complete Variable selection: order of magnitude of the penalty



Comparison with Lasso/Adaptive Lasso

The ”Adaptive Lasso” Proposed by Zou(2006).

$$\hat{\theta}^{\lambda} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \left| Y - \sum_{j=1}^{p} \theta_j X_j \right|^2 + \lambda \sum_{j=1}^{p} \frac{1}{\left|\tilde{\theta}_j\right|^{\gamma}} \times |\theta_j| \right\}.$$

$\longrightarrow$ λ, γ obtained by cross-validation

Simulation 1

Consider the predictors $X_1, \ldots, X_8 \in \mathbb{R}^{20}$ such that for all $i = 1, \ldots, 20$

$$X_i^T = (X_{1,i}, \ldots, X_{8,i}) \text{ are i.i.d. } \mathcal{N}(0, \Gamma) \text{ with } \Gamma_{j,k} = 0.5^{|j-k|}.$$

and

$$\mu = 3X_1 + 1.5X_2 + 2X_5$$

	$\sigma = 1$			
	r	$\mathbb{E}(\lvert\widehat{m}\rvert)$	$\%\{\widehat{m} = m_0\}$	$\%\{\widehat{m} \supseteq m_0\}$
Our procedure	1.57	3.34	72%	97.8%
Lasso	2.09	5.21	10.8%	100%
A. Lasso	1.99	4.56	16.8%	99%

	$\sigma = 3$			
	r	$\mathbb{E}(\lvert\widehat{m}\rvert)$	$\%\{\widehat{m} = m_0\}$	$\%\{\widehat{m} \supseteq m_0\}$
Our procedure	3.08	2.01	10.3%	15.7
Lasso	2.06	4.56	10.5%	100%
A. Lasso	2.44	3.81	13.2	52%

Simulation 2

Let X_1, X_2, X_3 be three vectors of $\mathbb{R}^n$ defined by

$$\begin{array}{lcllllll} X_1 & = & (& 1, & -1, & 0, & \ldots, & 0) \ /\sqrt{2} \\ X_2 & = & (& -1, & 1.001, & 0, & \ldots, & 0) \ /\sqrt{1+1.001^2} \\ X_3 & = & (& 1/\sqrt{2}, & 1/\sqrt{2}, & 1/n, & \ldots, & 1/n) \ /\sqrt{1+(n-2)/n^2} \end{array}$$

and $X_j = e_j$ for all $j = 4, \ldots, n$.

We take $p = n = 20$, $D_{\max} = 8$ and

$$\mu = (n, n, 0, \ldots, 0) \in \mathrm{Span}\,\{X_1, X_2\}.$$

$\longrightarrow$ μ almost $\perp$ X_1, X_2 and very correlated to X_3.

The result

	r	$\mathbb{E}(\|\widehat{m}\|)$	$\%\{\widehat{m} = m_0\}$	$\%\{\widehat{m} \supseteq m_0\}$
Our procedure	2.24	2.19	83.4%	96.2%
Lasso	285	6	0%	30%
A. Lasso	298	5	0%	25%

Mixed strategy

Let $m \in \mathcal{M}_c$.

$$\begin{aligned} L_m &= |m| \text{ if } m \in \mathcal{M}_o \\ &= \log\left[\binom{p}{|m|}\right] + \log(p(|m|+1)) \quad \text{if } m \in \mathcal{M}_c \setminus \mathcal{M}_o \end{aligned}$$

$$\longrightarrow \quad \sum_{m \in \mathcal{M}} (D_m + 1) e^{-L_m} \leq 3.51$$

$$\square \mathbb{E}\left(\frac{|\mu - \hat{\mu}_{\hat{m}}|^2}{\sigma^2}\right) \leq$$

$$\left\{\inf_{m \in \mathcal{M}_o} \mathbb{E}\left(\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right) \vee 1\right\} \wedge \left\{\log(p) \inf_{m \in \mathcal{M}_c} \mathbb{E}\left(\frac{|\mu - \hat{\mu}_m|^2}{\sigma^2}\right) \vee 1\right\}.$$