Model selection

<u>objective</u>: to adapt to unknown hidden structures.

(1) Regression model

- <u>goal</u>: predict $y \in \mathbb{R}$ from covariates $x \in \mathbb{R}^p$
- <u>Regression model</u>: $y = \underset{\uparrow \text{ unknown}}{f(x)} + \underset{\nwarrow\ \mathbb{E}[\varepsilon]=0}{\varepsilon}$

why? $\quad Y = \underbrace{\mathbb{E}[Y|X]}_{=f(X)} + \underbrace{(Y - \mathbb{E}[Y|X])}_{=\varepsilon \text{ with } \mathbb{E}[\varepsilon]=0}$

- <u>Linear model</u>: $f(x) = \langle \beta, x \rangle$
  $\beta \in \mathbb{R}^p$, unknown

- <u>Examples</u>
  - <u>linear approximation</u>: for x close to $\bar{x} \in \mathbb{R}^p$ we have
  $$f(x) \simeq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle$$
  - <u>Frame / basis expansion</u>: we can expand f on a Fourier basis, on wavelets, on cubic splines, etc...
  $$f(x) = \sum_j \beta_j \varphi_j(x)$$
  $$= \langle \beta, \Phi(x) \rangle \quad \text{where } \Phi(x) = (\varphi_j(x))_j$$
  Typically, the expansion over a wavelet or spline basis is sparse which means that only a few β_j are (significantly) different from 0.

- **Additive model:**

$$f(x) = \sum_{k=1}^{p} f_k(x_k)$$

$$= \sum_{k=1}^{p} \sum_{j \in J_k} \beta_{j,k} \varphi_j(x_k) \qquad \text{(basis expansion)}$$

$$= \langle \beta, \Phi(x) \rangle \quad \text{where} \quad \beta = (\beta_{j,k})_{\substack{j \in J_k \\ k=1,\dots,p}}$$

$$\Phi(x) = (\varphi_j(x_k))_{\substack{j \in J_k \\ k=1,\dots,p}}$$

If only a few x_k are influencial, then only a few f_k are non zero i.e only a few "group" $(\beta_{j,k})_{j \in J_k}$ are non zero.

- **Observations:** we have n observations

  - $y_i = f(x^{(i)}) + \varepsilon_i, \quad i=1,\dots,n$

$$\Downarrow$$

$$Y = f^* + \varepsilon \in \mathbb{R}^n \quad \text{with } f^*_i = f(x^{(i)})$$

  - In the following, we assume that

    → $f(x) = \langle \beta^*, x \rangle$ so that

$$f^* = X\beta^* \quad \text{where} \quad X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(n)})^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

    → $(\varepsilon_i)_{i=1,\dots,n}$ iid $\mathcal{N}(0, \sigma^2)$

    so that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

- Hidden structures

  - coordinate sparsity: $|\beta^*|_0 := \text{card}\{j : \beta^*_j \neq 0\}$ small

    unknown? $\quad \text{supp}(\beta^*) = \{j : \beta^*_j \neq 0\}$

  - group sparsity: $\underbrace{\{1,\dots,p\} = \bigsqcup_{k=1}^{n} G_k}_{\text{known}}$ and $\text{card}\{k : \beta_{G_k} \neq 0\}$ small

    unknown? $\quad \{k : \beta_{G_k} \neq 0\}$

2 Models and oracle

a/ Known structure

- for example, if we know $m^* := \text{supp}(\beta^*)$, then we can fit the model

$$y_i = \sum_{j \in m^*} \beta^*_j x^{(i)}_j + \varepsilon_i, \quad i = 1,\dots,n.$$

- more generally, if we know that $f^* \in S^*$, with $S^* \subset \mathbb{R}^n$ a linear span, we can maximize the likelihood, with the constraint that $\hat{f} \in S^*$:

$$\hat{f} \in \underset{f \in S^*}{\text{argmin}} -\log L(f), \quad \text{where } -\log L(f) = \frac{\|Y - f\|^2}{2\sigma^2} + \frac{n}{2}\log(2\pi\sigma^2)$$

The solution is $\hat{f}_{S^*} = \text{Proj}_{S^*} Y$.

b/ Collection of models

. Problem: S^* is unknown in practice.

→ Take a collection $\{S_m, m \in \mathcal{M}\}$ of linear spans (called models), corresponding to the different possible structure

→ Use the best of the estimator in the collection $\{\hat{f}_m, m \in \mathcal{M}\}$, where $\hat{f}_m := \mathrm{Proj}_{S_m} Y$

. Examples:

- coordinate sparse:
  - → $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$
  - → $S_m = \mathrm{span}\{X_j : j \in m\}$ where $X_j = X[\cdot, j]$ for $m \in \mathcal{M}$
- group sparse:
  - → $\mathcal{M} = \mathcal{P}(\{1, \dots, M\})$
  - → $S_m = \mathrm{span}\{X_j : j \in \bigcup_{k \in m} G_k\}$, for $m \in \mathcal{M}$.

. Best estimator?

- risk: $R(\hat{f}) = \mathbb{E}[\|\hat{f} - f^*\|^2]$
- oracle: $\hat{f}_{m_o}$ where $m_o \in \underset{m \in \mathcal{M}}{\mathrm{argmin}}\, R(\hat{f}_m)$

③ Selecting a model

a/ Risk of $\hat{f}_m$

- Since $Y = f^* + \varepsilon$ and $\hat{f}_m = \mathrm{Proj}_{S_m} Y$, we have

$$R(\hat{f}_m) = \mathbb{E}\left[\| \mathrm{Proj}_{S_m}(f^* + \varepsilon) - f^* \|^2\right]$$

$$\underset{\text{Pythagore}}{=} \mathbb{E}\left[\| \mathrm{Proj}_{S_m}(\varepsilon)\|^2\right] + \mathbb{E}\left[\| f^* - \mathrm{Proj}_{S_m} f^* \|^2\right]$$

$$= \sigma^2 \, \mathrm{Tr}(\mathrm{Proj}_{S_m}) + \| f^* - \mathrm{Proj}_{S_m} f^* \|^2$$

$$= \underbrace{\sigma^2 \dim(S_m)}_{\substack{\text{variance} \\ \nearrow \text{ with } S_m}} + \underbrace{\| f^* - \mathrm{Proj}_{S_m} f^* \|^2}_{\substack{\text{bias} \\ \searrow \text{ with } S_m}}$$

- The oracle m_o minimizes

$$m_o \in \underset{m \in \mathcal{M}}{\mathrm{argmin}} \left\{ \underbrace{\| f^* - \mathrm{Proj}_{S_m} f^* \|^2}_{\uparrow \text{ unknown!}} + \sigma^2 \dim(S_m) \right\}$$

1. Estimate $R(\hat{f}_m)$ by some $\hat{R}(\hat{f}_m)$
2. Take $\hat{f}_{\hat{m}}$ with $\hat{m} \in \underset{m \in \mathcal{M}}{\mathrm{argmin}} \, \hat{R}(\hat{f}_m)$

Questions:

- which $\hat{R}(\hat{f}_m)$?
- which performance?

b/ History

Naive : take $\hat{f}_m$ with the best fit on the data

$$\hat{m}_{naive} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \|Y - \hat{f}_m\|^2$$

Since $\hat{f}_m = \mathrm{Proj}_{S_m} Y \rightsquigarrow \hat{m}_{naive} = \text{largest model}$:(

AIC : unbiased estimation of the risk

$$\mathbb{E}\left[\|Y - \hat{f}_m\|^2\right] = \mathbb{E}\left[\|f^* + \varepsilon - \mathrm{Proj}_{S_m}(f^* + \varepsilon)\|^2\right]$$

$$= \|f^* - \mathrm{Proj}_{S_m} f^*\|^2 + \underbrace{\mathbb{E}\left[\|\varepsilon - \mathrm{Proj}_{S_m}\varepsilon\|^2\right]}_{(n - \dim(S_m))\sigma^2} + 2\,\underbrace{\mathbb{E}\left[\langle f^* - \mathrm{Proj}_{S_m} f^*, \varepsilon - \mathrm{Proj}_{S_m}\varepsilon\rangle\right]}_{=0}$$

$$= R(\hat{f}_m) + (n - 2\dim(S_m))\sigma^2$$

So $\hat{R}(\hat{f}_m) := \|Y - \hat{f}_m\|^2 + 2\sigma^2 \dim(S_m) - n\sigma^2$ is an unbiased estimate of $R(\hat{f}_m)$

$$\hat{m}_{AIC} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}_m\|^2 + 2\dim(S_m)\sigma^2 \right\}$$

⚠

- It does not work when $\mathcal{M}$ is very large (with an exponential number of models per dimension)
- To do: Exercise 2.8.1 parts A] and B]

BIC : bayesian approximation.

bayesian model: $\to$ m^* is sampled according to $\pi = (\pi_m)_{m \in \mathcal{M}}$

$\to$ f^* is sampled according to a diffuse distribution on S_{m^*}

$\to$ $Y = f^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

Then, we can prove that when $n \to \infty$

$$-\log \mathbb{P}[m^* = m \mid Y] \underset{n\to\infty}{\approx} \frac{\|Y - \hat{f}_m\|^2}{2\sigma^2} + \frac{\dim(S_m)}{2} \log(n) - \log(\pi_m) + \underbrace{\text{remaining terms}}_{\text{smaller or independent of } m}$$

So for $(\pi_m)_{m \in \mathcal{M}}$ = uniform distribution on $\mathcal{M}$, we get

$$\hat{m}_{BIC} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}_m\|^2 + \dim(S_m) \log(n) \sigma^2 \right\}$$

⚠

- asymptotic justification (p fixed and $n \to \infty$)
- it does not work when $\mathcal{M}$ is very large
- look again at exercise 2.8.1 (A) and B)).

? and you, what would you do?

c/ Analytic design of selection criterion

- We observe that AIC and BIC are of the form

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}_m\|^2 + \mathrm{pen}(m)\,\sigma^2 \right\}$$

- Which pen(m) ?

- for a given pen(m), analyse $R(\hat{f}_{\hat{m}})$
- design pen(m) in order to have a good $R(\hat{f}_{\hat{m}})$

- Ideal: $$R(\hat{f}_{\hat{m}}) \leq \underbrace{(\text{constante})}_{\text{hopefully close to } 1.} \times \underbrace{\min_{m \in \mathcal{M}} R(\hat{f}_m)}_{\text{oracle risk}} + \underbrace{\text{remaining term}}_{\text{hopefully small}}$$

Such an inequality is called "oracle inequality"

Theorem 2.2

Set $B(d,\alpha) := \mathbb{E}\left[\left((\sqrt{d} + \sqrt{2\xi})^2 - \alpha\right)_+\right]$, where $\xi \sim \mathcal{E}xp(1)$

Then, for any $a > 1$

$$\frac{a-1}{a} R(\hat{f}_{\hat{m}}) \leq \min_{m \in \mathcal{M}} \left\{ R(\hat{f}_m) + \mathrm{pen}(m)\,\sigma^2 \right\} + a\sigma^2 \rho(\mathcal{M})$$

where $\rho(\mathcal{M}) = 1 + \sum_{m \in \mathcal{M}} B\left(\dim(S_m), \frac{1}{a}\,\mathrm{pen}(m)\right)$

Sketch of proof :

- useful lemma : $\forall a>0$

$$\boxed{\begin{array}{ll}(i) & 2\langle x,y\rangle \leq a\|x\|^2+\frac{1}{a}\|y\|^2\\ (ii) & \|x+y\|^2 \leq (1+a)\|x\|^2+(1+\frac{1}{a})\|y\|^2\end{array}}$$

proof :

$$\left\|\sqrt{a}\,x+\frac{1}{\sqrt{a}}\,y\right\|^2 \geq 0$$

$\square$

- Starting inequality : $\|Y-\hat{f}_{\hat{m}}\|^2+\text{pen}(\hat{m})\sigma^2 \leq \|Y-\hat{f}_m\|^2+\text{pen}(m)\sigma^2, \quad \forall m \in \mathcal{M}$

$$\overset{Y=f^*+\varepsilon}{\Longrightarrow} \|f^*-\hat{f}_{\hat{m}}\|^2 \leq \underbrace{\|f^*-\hat{f}_m\|^2+\text{pen}(m)\sigma^2}_{\text{OK}}+\underbrace{2\langle\varepsilon,f^*-\hat{f}_m\rangle}_{(\text{I})}+\underbrace{2\langle\varepsilon,\hat{f}_{\hat{m}}-f^*\rangle-\text{pen}(\hat{m})\sigma^2}_{(\text{II})}$$

(I): $\mathbb{E}[\langle\varepsilon,f^*-\hat{f}_m\rangle] = -\mathbb{E}[\|\text{Proj}_{S_m}\varepsilon\|^2] \leq 0$

(II): Let us set $\bar{S}_m = S_m+\langle f^*\rangle = \langle f^*\rangle \perp \tilde{S}_m$

$$2\langle\varepsilon,\hat{f}_{\hat{m}}-f^*\rangle = 2\langle \text{Proj}_{\bar{S}_{\hat{m}}}\varepsilon,\hat{f}_{\hat{m}}-f^*\rangle$$

$$\leq a\underbrace{\|\text{Proj}_{\bar{S}_{\hat{m}}}\varepsilon\|^2}_{=\|\text{Proj}_{\langle f^*\rangle}\varepsilon\|^2+\|\text{Proj}_{\tilde{S}_{\hat{m}}}\varepsilon\|^2}+\frac{1}{a}\|\hat{f}_{\hat{m}}-f^*\|^2$$

- So, for any $m\in\mathcal{M}$

$$\left(1-\frac{1}{a}\right)R(\hat{f}_{\hat{m}}) \leq R(\hat{f}_m)+\text{pen}(m)\sigma^2+a\underbrace{\mathbb{E}\left[\|\text{Proj}_{\langle f^*\rangle}\varepsilon\|^2\right]}_{=\sigma^2}+a\,\underbrace{\mathbb{E}\left[\left(\|\text{Proj}_{\tilde{S}_{\hat{m}}}\varepsilon\|^2-\frac{\text{pen}(\hat{m})\sigma^2}{a}\right)\right]}_{?}$$

- $\mathbb{E}\left[\left(\|\text{Proj}_{\tilde{S}_{\hat{m}}}\varepsilon\|^2-\frac{\text{pen}(\hat{m})\sigma^2}{a}\right)\right] \leq \mathbb{E}\left[\sup_{m\in\mathcal{M}}\left(\|\text{Proj}_{\tilde{S}_m}\varepsilon\|^2-\frac{\text{pen}(m)\sigma^2}{a}\right)\right]$

$$\leq \sum_{m\in\mathcal{M}}\mathbb{E}\left[\left(\|\text{Proj}_{\tilde{S}_m}\varepsilon\|^2-\frac{\sigma^2}{a}\text{pen}(m)\right)_+\right]$$

From lecture 1 (Gaussian concentration inequality), there exists $\xi_m \sim \text{Exp}(1)$ such that

$$\|\text{Proj}_{\tilde{S}_m}\varepsilon\|^2 \leq \left(\sqrt{\mathbb{E}[\|\text{Proj}_{\tilde{S}_m}\varepsilon\|^2]}+\sigma\sqrt{2\xi_m}\right)^2 = \left(\sigma\sqrt{\dim(\tilde{S}_m)}+\sigma\sqrt{2\xi_m}\right)^2$$

$$\leq \left(\sqrt{\dim(S_m)}+\sqrt{2\xi_m}\right)^2\sigma^2$$

$\square$

which pen(m)?

- We have $B(d,\alpha) \asymp \exp\left(-\frac{1}{2}(\sqrt{\alpha}-\sqrt{d})^2\right)$

  and we want $\sum_{m\in\mathcal{M}} B\left(\dim(S_m), \frac{1}{a}\,\mathrm{pen}(m)\right) \asymp 1$

- So we take pen(m) such that $B\left(\dim(S_m), \frac{1}{a}\,\mathrm{pen}(m)\right) \asymp \pi_m$,

  with $\sum_{m\in\mathcal{M}} \pi_m = 1$.

- solving $\exp\left(-\frac{1}{2}\left(\sqrt{\frac{\mathrm{pen}(m)}{a}} - \sqrt{\dim(S_m)}\right)^2\right) = \pi_m$, we find

$$\mathrm{pen}_{BM}(m) = a\left(\sqrt{\dim(S_m)} + \sqrt{2\log\frac{1}{\pi_m}}\right)^2, \text{ with } a>1$$

Corollary:

For the choice $\mathrm{pen}_{BM}(m)$, there exists a constant $C_a>1$ such that

$$R(\hat{f}_{\hat{m}}) \leq C_a \min_{m\in\mathcal{M}} \left\{ R(\hat{f}_m) + \left(1+\log\frac{1}{\pi_m}\right)\sigma^2 \right\}$$

Proof: see proof of Theorem 2.2 in the lecture notes □

Questions:

1/ can we choose π_m to get an oracle inequality?

2/ optimality of $\hat{f}_{\hat{m}}$ → see next lecture

3/ can we choose pen(m) smaller? No, see again the exercise 2.8.1.

Choice of π_m :

→ we choose π_m in order to have $\min_{m\in\mathcal{M}} \left\{ R(\hat{f}_m) + \sigma^2 \log \frac{1}{\pi_m} \right\}$ as small as possible

→ we would like to have an oracle inequality :

since $R(\hat{f}_m) = \| \mathrm{Proj}_{S_m} f^* - f^* \|^2 + \dim(S_m)\sigma^2 \geq \dim(S_m)\sigma^2$

we want to have $\log \frac{1}{\pi_m} \leq c \dim(S_m)$

not always possible when $\mathcal{M}$ is very large (see below)

Examples :

- coordinate sparse: $\mathcal{M} = \mathcal{P}(\{1,\dots,p\})$ and $\dim(S_m) \leq |m|$

taking $\pi_m \propto e^{-s|m|}$, we get

$$\sum_{m\in\mathcal{M}} e^{-s|m|} = \sum_{d=0}^{p} C_p^d e^{-sd} = (1+e^{-s})^p \quad \text{and hence}$$

$$\pi_m = \frac{e^{-s|m|}}{(1+e^{-s})^p} \quad \text{and} \quad \log\frac{1}{\pi_m} = s|m| + \underbrace{p\log(1+e^{-s})}_{\text{requires } s \simeq \log p}$$

<u>choice 1</u>: . $\pi_m = (1+\frac{1}{p})^{-p} p^{-|m|}$

. $\log\frac{1}{\pi_m} \leq 1 + |m| \log p$

<u>choice 2</u>: . $\pi_m = \frac{1}{C_p^{|m|}} \times \frac{e-1}{e-e^{-p}} \times e^{-|m|}$

. since $\log C_p^d \leq d \log\left(\frac{ep}{d}\right)$ (Lemma 2.1)

$$\log\frac{1}{\pi_m} \leq \log\frac{e}{e-1} + |m| \log\left(\frac{e^2 p}{|m|}\right)$$

Remark: Set $m^* = \mathrm{supp}(\beta^*)$. For choice 2

$$R(\hat{f}_{\hat{m}}) \leq C_a \inf_{m \in \mathcal{M}} \left\{ R(\hat{f}_m) + \left(1 + \log \frac{1}{\pi_m}\right)\sigma^2 \right\}$$

$$\leq C_a \left(\underbrace{R(\hat{f}_{m^*})}_{= |m^*|\sigma^2} + \left(1 + \log \frac{1}{\pi_{m^*}}\right)\sigma^2 \right)$$

$$\leq C_a' \, |m^*| \left(1 + \log \frac{p}{|m^*|}\right)\sigma^2$$

- group sparse: $\mathcal{M} = \mathcal{P}(\{1, .., M\})$
  so idem with $p \leftarrow M$

4 Numerical illustration

Section 2.5, in the sparse basis expansion setting
$$f(x) = \sum_j \beta_j \varphi_j(x) \quad \text{with } (\varphi_j)_{j \geq 1} \text{ Fourier basis.}$$

5 Computational issues

- In principle solving $\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \{ \|Y - \hat{f}_m\|^2 + \mathrm{pen}(m)\sigma^2 \}$
  requires $|\mathcal{M}|$ evaluations. It is prohibitive for large $\mathcal{M}$, as in the coordinate sparse setting where $|\mathcal{M}| = 2^p$.
- Cases where it is possible though:
  → when the columns of X are orthogonal (hard thresholding)
  see (again!) exercise 2.8.1
  → when $f(x)$ is piecewise constant (with dynamic programming)
  Do exercise 2.8.4

otherwise?

→ convexification (Lecture 4, chap. 5)

→ greedy algorithms:

- forward - backward (p. 43)
- Iterative Hard Thresholding (Lecture 5, chap. 6)

⑥ Take home message

- Model selection is a powerful theory for conceptualizing estimation in high dimensional setting
- it gives optimal estimators (see next lecture)
- prohibitive computational complexity (but in a few cases)
- good baseline for deriving practical procedures
  - → convex criterion (Lecture 4)
  - → greedy algorithm (Lecture 5)