March 23 Recitation

- BIC Score

$$BIC = \ell(X;\hat{\theta}) - \frac{k}{2}\log(n) \quad \begin{cases} k: & \text{num. of params.} \\ n: & \text{num of samples} \\ \hat{\theta}: & \text{maximum likelyhood estimates of params} \end{cases}$$

Regression tree example:

Model 1 : expr is not factor specific

$$\Rightarrow \ell(X;\theta) = \log \prod_{t=1}^{m}\prod_{i=1}^{n} N(X_{it};\mu,\sigma^2) = \sum_{t=1}^{m}\sum_{i=1}^{n}\left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(X_{it}-\mu)^2\right]$$

MLE : find $\hat{\theta} = \underset{\mu,\sigma}{\operatorname{argmax}}\ \ell(X;\theta) = \left\{(\mu,\sigma) \mid \frac{\partial \ell(X;\theta)}{\partial \mu} = \frac{\partial \ell(X;\theta)}{\partial \sigma} = 0\right\}$

(maximum likelyhood estimation)

$$\Rightarrow \hat{\mu} = \frac{1}{nm}\sum_{t=1}^{m}\sum_{i=1}^{n} X_{it}^2$$

$$\hat{\sigma} = \left(\frac{1}{nm}\sum_{t=1}^{m}\sum_{i=1}^{n}(X_{it}-\hat{\mu})^2\right)^{\frac{1}{2}}$$

$$\Rightarrow \ell(X;\hat{\theta}) = nm\left(-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma}^2)\right)$$

$$\Rightarrow BIC = nm\left(-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma}^2)\right) - \frac{2}{2}\log(nm)$$

Model 2: Expr. depends on one factor w/ one threshold

Assume it's factor j. For a given threshold f_j^* $\begin{cases} I_1 = \{t: f_{jt} < f_j^*\} \\ I_2 = \{t: f_{jt} > f_j^*\} \end{cases}$

$$\Rightarrow \ell(X;\theta,\hat{f_j}) = \log\left[\prod_{t\in I_1}\prod_{i=1}^{n} N(X_{it};\mu_1,\sigma_1)\prod_{t\in I_2}\prod_{i=1}^{n} N(X_{it};\mu_2,\sigma_2)\right]$$

Find $\hat{\mu_1}, \hat{\sigma_1}, \hat{\mu_2}, \hat{\sigma_2}$ in a similar way

Note: also need to optimize over all possible thresholds f_i^*

$$\Rightarrow BIC = n|I_1|\left(-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma_1}^2)\right) + n|I_2|\left(-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma_2}^2)\right) - \frac{5}{2}\log(nm)$$

- Bayesian model selection
  - ⇒ Compare model 1 and model 2 in general, rather than best (model 1) vs. best (model 2)

    ↓ defined by MLE
  - ⇒ Take the distribution of parameter into account
  - ⇒ Avoids overfitting (Bayesian Occam's Razor)

$$P(M|D) \sim P(D|M)\,P(M)$$

assume $p(M) \sim$ uniform

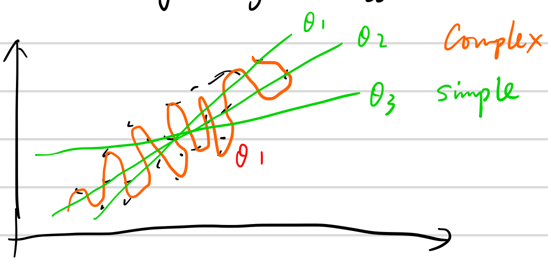
$$\sim p(D|M) \quad \text{marginal likelyhood / evidence}$$

$$P(D|M) = \int_\theta \underbrace{P(D|M,\theta)}_{\text{likelyhood}}\ \underbrace{P(\theta|M)}_{\substack{\text{prior on } \theta \\ (\text{assuming model } M)}}\, d\theta$$

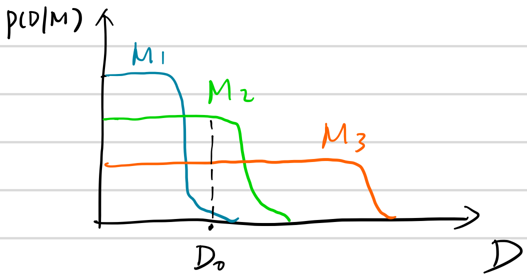
Complex model might not always have the best $p(D|M)$

- Complex model can get very good $p(D|M,\theta)$ but only for very few θ

  Simple model might get decent $p(D|M,\theta)$ for a large range of θ ⇒ Intergral might be bigger



- conservation of prob. mass $\sum_{D'} p(D'|M) = 1$



D_0 : actual data ⇒ M_2 fits best

Complexity : $M_1 < M_2 < M_3$

(more complex model can model a larger range of D)

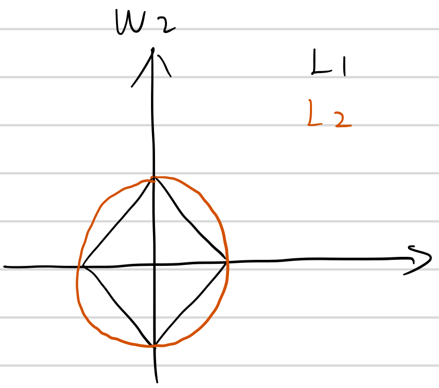
but resulting in "thiner" distribution

Note : • BIC is an approximation of BMS

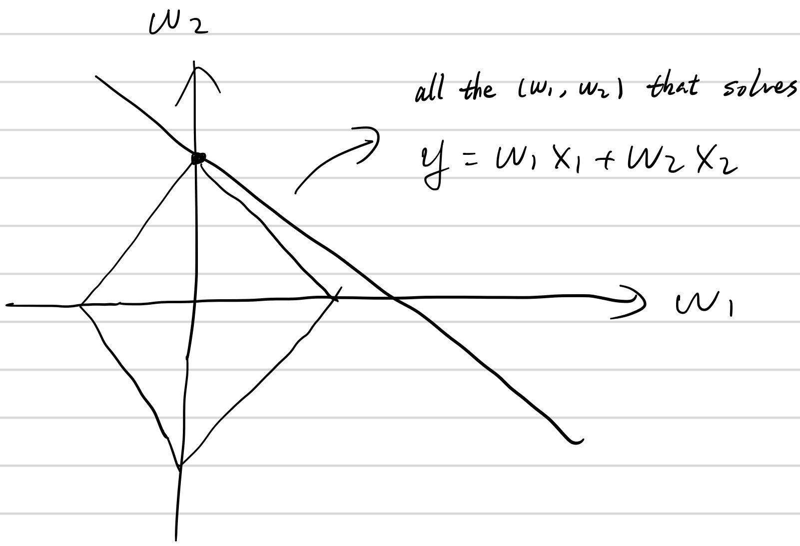
• BIC also penalizes complex model (by $\frac{k}{2}\log n$)

- L_1 / L_2 penalty

$$L = \sum_i (y_i - wx_i)^2 + R \qquad R = \begin{cases} \|W\|_1 = \sum_{k=1}^{d} w_k \\ \frac{1}{2}\|w\|_2^2 = \frac{1}{2}\sum_{k=1}^{d} w_k^2 \end{cases}$$



All the (w_1, w_2) with a fixed R



Graudually shrink the diamond until only one point crosses with $y = w_1 x_1 + w_2 x_2$ $\Rightarrow$ end up at a vertex

$\Rightarrow$ sparsity ($w_1 = 0$)

OLS $\Rightarrow$ $(X^T X)^{-1} X^T y$

OLS w/ L_2 $\Rightarrow$ $(X^T X + \alpha I)^{-1} X^T y$