Model Selection

We used the LRT to compare models where a set of parameters were restricted. However, the LRT cannot be used to compare nonnested models, models with different dimensionality or family.

We consider models f_j, $j=1,\ldots,J$ where

$$\{f_j = f_{\theta_j},\ \theta_j \in \Theta_j\}, \quad p_j = \dim(\theta_j)$$

We cannot use $l_{f_j}(\hat{\theta}_j) = \log L_f(\hat{\theta}_j | \underset{\sim}{x}) = \sum_{i=1}^{n} \log f_j(x_i | \hat{\theta}_j)$

since model with max p_j 'wins' by construction

Focus on <u>generalization</u> - does the model, and the MLE under this model, work on future data?

Our criterion is

Choose f_j to maximize

$$\mathbb{E}_g \left[\log L_{f_j}(\theta_{f_j} | X) \right]$$

where we assume $X \sim^d g$ (unknown, true distribution)

Too many unknowns (don't know θ_{f_j}) $\Rightarrow$ replace with

$$\mathbb{E}_g \left[\log L_{f_j}(\hat{\theta}_{f_j}(\underset{\sim}{X}) | X) \right] \equiv Q_{f_j}$$

- $\mathbb{E}_g$: expectation over both uncertainty in $\underset{\sim}{X}$ and X
- $\hat{\theta}_{f_j}(\underset{\sim}{X})$: MLE under f_j using $\underset{\sim}{X} = X_1, \ldots \sim^d g$ (unknown)
- X: generic, new future observation $X \sim^d g$

How to get a good and computable estimate of Q_{f_j}? (2)

We observe $\tilde{x}$, can estimate $\hat{\theta}_{f_j}(\tilde{x})$ and the maximized likelihood $\frac{1}{n} l_{f_j}(\hat{\theta}_{f_j}(\tilde{x})) = \frac{1}{n} \sum_{i=1}^{n} \log f_j(x_i | \hat{\theta}_{f_j})$

- Note, $\frac{1}{n} l_{f_j}(\hat{\theta}_{f_j})$ is <u>not</u> a good estimate of Q_{f_j}

  Since we use $\tilde{x}$ both for training (to obtain $\hat{\theta}$) and testing (to compute the likelihood).

  $\frac{1}{n} l_{f_j}(\hat{\theta}_{f_j})$ is an <u>overestimate</u> of the Q_{f_j}.

<u>Properties & notations</u> (drop f_j for now and work with one f)

- $\hat{\theta}_f = \underset{\theta_f}{\operatorname{argmax}}\ l_f(\theta_f | \tilde{x})$
- $\frac{1}{n} \sum_{i=1}^{n} \log f(x_i | \theta_f) \longrightarrow \mathbb{E}_g [\log f(X | \theta_f)]$ by WLLN

  and max of LH $\rightarrow$ max of RH
- Define θ_0 as maximizer of RH $\mathbb{E}_g [\log f(X | \theta_f)]$

  $\hat{\theta}_f \rightarrow \theta_0$

- Note, θ_0 is not "the true" parameter unless $f = g$

  Instead, θ_0 is the best choice of θ_f to match f as closely as possible to g — best f-approximation of g

- $\hat{\theta}_f = \underset{\theta_f}{\text{argmax}}\ l_f(\theta_f|\tilde{x})$ or zero-crossing of the

Score $S_f(\theta_f) = \frac{\partial}{\partial \theta_f} \sum_{i=1}^{n} \log f(x_i|\theta_f) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_f} \log f(x_i|\theta_f)$

- $S_f(\theta_f) \simeq S_f(\hat{\theta}_f) + \frac{\partial^2}{\partial \theta_f \partial \theta_f^T} \log L_f(\theta_f|\tilde{x})\Big|_{\theta^*} (\theta_f - \hat{\theta}_f)$

  - $S_f(\theta_f)$: approx near $\hat{\theta}_f$
  - $S_f(\hat{\theta}_f)$: $= 0$ by def.
  - $\frac{\partial^2}{\partial \theta_f \partial \theta_f^T} \log L_f(\theta_f|\tilde{x})$: 2nd order derivative
  - θ^*: value between θ_f and $\hat{\theta}_f$

(For simplicity of notation, drop subscript f on θ below)

So $S_f(\theta) \simeq \frac{\partial^2}{\partial \theta \partial \theta^T} \log L_f(\theta|\tilde{x})\Big|_{\theta^*} (\theta - \hat{\theta})$

$= \sum_{i=1}^{\hat{n}} \underbrace{\frac{\partial}{\partial \theta} \log f(x_i|\theta)}_{y_i}$

where $E(y_i) = E_g\left[\frac{\partial}{\partial \theta} \log f(x|\theta)\right] = \frac{\partial}{\partial \theta} E_g[\log f_\theta(x)]$

$= 0$ at θ_0 since θ_0 maximizer of $\rightarrow$

At θ_0, $V(y_i) = E(y_i^2) = E_g\left[\left(\frac{\partial}{\partial \theta} \log f_\theta(x)\right)\left(\frac{\partial}{\partial \theta} \log f_\theta(x)\right)^T\right]$

$\equiv J(\theta_0)$

- By CLT $\frac{1}{\sqrt{n}} S_f(\theta_0) \xrightarrow{d} N(0, J(\theta_0))$

We have

$$\frac{1}{\sqrt{n}} S_f(\theta_0) = \frac{1}{\sqrt{n}} \frac{\partial^2}{\partial\theta\partial\theta^T} \log L_f(\theta|\tilde{x})\Big|_{\theta^*} (\theta_0 - \hat{\theta})$$

$$\to^d N(0, J(\theta_0))$$

we want distribution of this.

- $\frac{\partial^2}{\partial\theta\partial\theta^T} \log L(\theta|\tilde{x})\Big|_{\theta^*} \approx \frac{\partial^2}{\partial\theta\partial\theta^T} \log L_f(\theta|\tilde{x})\Big|_{\theta_0}$ for large n

- $\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log L_f(\theta|\tilde{x})\Big|_{\theta_0} \to E_g\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log f_\theta(X)\Big|_{\theta_0}\right]$ by WLLN

$$\equiv -FI(\theta_0)$$

All together we have

$$\frac{1}{\sqrt{n}} S_f(\theta_0) \approx \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log f_\theta(X_i)\Big|_{\theta_0}\right] \sqrt{n}(\theta_0 - \hat{\theta})$$

$$\to^d N(0, J(\theta_0)) \qquad \to^p -FI(\theta_0)$$

$\Rightarrow$ Slutsky's Thm

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, FI(\theta_0)^{-1} J(\theta_0) FI(\theta_0)^{-1})$$

Key result: CLT of MLE under wrong model

If $f = g$ $FI(\theta_0)^{-1} = J(\theta_0)$ $\Rightarrow$ $\hat{\theta} \to^d N(\theta_0, (nFI(\theta_0))^{-1})$

Comparing models

Back to original problem: $f_j, j = 1, \dots, J$

- $\hat{\theta}_j$ = maximizer of $\sum_{i=1}^{n} \log f_{\theta_j}(x_i) = l_{f_j}(\theta_j)$
- θ_{j0} = maximizer of $E_G[\log f_{\theta_j}(X)]$

Want to choose f_j to maximize

$$Q = E_{X \text{ and } \tilde{X}}\left[\log f_j(X \mid \hat{\theta}_j(\tilde{X}))\right]$$

- We know $\sqrt{n}(\hat{\theta}_j - \theta_{j0}) \to_d N(0, I_j^{-1}(\theta_{j0}) J_j(\theta_{j0}) I_j^{-1}(\theta_{j0}))$

We start by examining the estimation variance
(dealing with randomness of $\tilde{x} \Rightarrow$ impact on $\hat{\theta}_j$)

We have $l_{f_j}(\theta_j) \approx l_{f_j}(\hat{\theta}_j) + \frac{\partial}{\partial \theta_j} l_{f_j}(\theta_j)\Big|_{\hat{\theta}_j} (\theta_j - \hat{\theta}_j)$

$= 0$

$$+ \frac{1}{2}(\theta_j - \hat{\theta}_j)^T \left(\frac{\partial^2}{\partial \theta_j \partial \theta_j^T} l_{f_j}(\theta_j)\Big|_{\theta_j^*}\right)(\theta_j - \hat{\theta}_j)$$

$$= l_{f_j}(\hat{\theta}_j) - \frac{1}{2}(\theta_j - \hat{\theta}_j)^T \boxed{**} (\theta_j - \hat{\theta}_j)$$

where $\boxed{**} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_j \partial \theta_j^T} \log f_j(x_i \mid \theta_j)\Big|_{\theta_j^*} \to E_G\left[-\frac{\partial^2}{\partial \theta_j \partial \theta_j^T} \log f_j(X \mid \theta_j)\Big|_{\theta_{j0}}\right]$

$= I_j(\theta_{j0})$

So we have

$$l_{f_j}(\theta_j) \approx l_{f_j}(\hat{\theta}_j) - \frac{n}{2}(\theta_j - \hat{\theta}_j)^T \mathcal{FI}_j(\theta_{j0})(\theta_j - \hat{\theta}_j) \quad (***)$$

At θ_{j0} we have

$$n(\hat{\theta}_j - \theta_{j0})^T \mathcal{FI}_j(\theta_{j0})(\hat{\theta}_j - \theta_{j0})$$

$$= \left((n\mathcal{FI}_j(\theta_{j0}))^{1/2}(\hat{\theta}_j - \theta_{j0})\right)^T \left((n\mathcal{FI}_j(\theta_{j0}))^{1/2}(\hat{\theta}_j - \theta_{j0})\right)$$

$$= \tilde{\theta}_j^T \tilde{\theta}_j \qquad \tilde{\theta}_j \text{ = standardized version of } \hat{\theta}_j$$

$$= \sum_{p=1}^{p_j} \tilde{\theta}_{jp}^2 = \text{sum of diagonal elements of } \tilde{\theta}_j \tilde{\theta}_j^T$$

$$= \text{Trace}\, \tilde{\theta}_j \tilde{\theta}_j^T$$

We are interested in $E[\tilde{\theta}_j^T \hat{\theta}_j] = \text{Trace}\left(E[\tilde{\theta}_j \tilde{\theta}_j^T]\right)$

$$= \text{Trace}\left(E\left[n\mathcal{FI}_j(\theta_{j0})(\hat{\theta}_j - \theta_{j0})(\hat{\theta}_j - \theta_{j0})^T\right]\right)$$

$$= \text{Trace}\left(n\mathcal{FI}_j(\theta_{j0})\, E\left[(\hat{\theta}_j - \theta_{j0})(\hat{\theta}_j - \theta_{j0})^T\right]\right)$$

$$Cov(\hat{\theta}_j) = \mathcal{FI}_j(\theta_{j0})^{-1} \frac{\mathcal{J}_j(\theta_{j0})}{n} \mathcal{FI}_j(\theta_{j0})^{-1}$$

$\Rightarrow$ So $E\left[n(\hat{\theta}_j - \theta_{j0})^T \mathcal{FI}_j(\theta_{j0})(\hat{\theta}_j - \theta_{j0})\right]$

$$= \text{Trace}\left(n\mathcal{FI}_j(\theta_{j0})\left(\mathcal{FI}_j(\theta_{j0})^{-1} \frac{\mathcal{J}_j(\theta_{j0})}{n} \mathcal{FI}_j(\theta_{j0})^{-1}\right)\right) = \text{Trace}\left(\mathcal{J}_j(\theta_{j0})\mathcal{FI}_j(\theta_{j0})^{-1}\right)$$

Back to (eq 4)

$$l_{f_j}(\theta_j) \approx l_{f_j}(\hat{\theta}_j) - \frac{n}{2}(\hat{\theta}_j - \theta_{j0})^T F I_j(\theta_{j0})(\hat{\theta}_j - \theta_{j0})$$

and @ θ_{j0}

$$l_{f_j}(\theta_{j0}) \approx l_{f_j}(\hat{\theta}_j) - \frac{n}{2}(\hat{\theta}_j - \theta_{j0}) F I_j(\theta_{j0})(\hat{\theta}_j - \theta_{j0})$$

$$\approx l_{f_j}(\hat{\theta}_j) - \frac{1}{2}\text{Trace}\left(J_j(\theta_{j0}) F I_j(\theta_{j0})\right)$$

[We used: • Taylor expansion of l_{f_j}
• CLT for $\hat{\theta}_j$ under f_j but g true model
• replacing 2nd order derivatives in approx with their expectations]

Take it all together:

$$l_{f_j}(\theta_{j0}) = \sum_{i=1}^{n} \log f_j(x_i|\theta_{j0}) \approx \sum_{i=1}^{n} \log f_j(x_i|\hat{\theta}_j) - \frac{1}{2}\text{Trace}\left(J_j(\theta_{j0}) F I_j(\theta_{j0})^{-1}\right)$$

$$\Downarrow E_{g,\tilde{x}} \Rightarrow n E[\log f_j(X|\theta_{j0})] = E\left[\sum_{i=1}^{n} \log f_j(x_i|\theta_{j0})\right]$$

$$\approx E[l_{f_j}(\hat{\theta}_j)] - \frac{1}{2}\text{Trace}\left(J_j(\theta_{j0}) F I_j(\theta_{j0})^{-1}\right)$$

Take a step back; our goal is to maximize (wrt f_j)

$$\theta_j = E_{X,\tilde{X}}\left[\log f_j(X \mid \hat{\theta}_j(\tilde{X}))\right]$$

testing X, training $\tilde{X}$

We have so far done the training expectation, $E_{\tilde{X}}$.

To do testing, we expand

$$E_X[\log f_j(X \mid \theta_j)] \text{ near } \theta_{j0}$$

$$E_X[\log f_j(X \mid \theta_j)] = E[\log f_j(X \mid \theta_{j0})]$$

$$+ \frac{\partial}{\partial \theta_j} E[\log f_j(X \mid \theta_j)]\Big|_{\theta_{j0}} (\theta_j - \theta_{j0})$$

$$= 0$$

$$+ \frac{1}{2}(\theta_j - \theta_{j0})^T \frac{\partial^2}{\partial \theta_j \partial \theta_j^T} E[\log f_j(X \mid \theta_j)]\Big|_{\theta_{j*}} (\theta_j - \theta_{j0})$$

$$\approx E[\log f_j(X \mid \theta_{j0})] - \frac{1}{2}(\theta_j - \theta_{j0})^T E\left[\frac{-\partial^2}{\partial \theta_j \partial \theta_j^T} \log f_j(X \mid \theta_{j0})\Big|_{\theta_{j0}}\right](\theta_j - \theta_{j0})$$

$$= I_j(\theta_{j0})$$

$$= E[\log f_j(X \mid \theta_{j0})] - \frac{1}{2n}\left[n(\theta_j - \theta_{j0})^T I_j(\theta_{j0})(\theta_j - \theta_{j0})\right]$$

Consider $\theta_j = \hat{\theta}_j$ especially;

$$E_X[\log f_j(X|\hat{\theta}_j)] \approx E[\log f_j(X|\theta_{j0})]$$

$$-\frac{1}{2n}\left[n(\hat{\theta}_j - \theta_{j0})^T FI_j(\theta_{j0})(\hat{\theta}_j - \theta_{j0})\right]$$

Now take $E_{\tilde{X}}$ also $\Rightarrow$

$$E_{\tilde{X}}\left[E_X[\log f_j(X|\hat{\theta}_j)]\right] \approx E_X[\log f_j(X|\theta_{j0})] \quad (A)$$

$$\| \quad Q_j \qquad -\frac{1}{2n}\text{Trace}\{J_j(\theta_{j0})FI_j(\theta_{j0})^{-1}\}$$

from training error on page (6)

From page (6) we have

$$n E_X[\log f_j(X|\theta_{j0})] \approx E_{\tilde{X}}[\ell_{f_j}(\hat{\theta}_j)] - \frac{1}{2}\text{Trace}\{J_j(\theta_{j0})FI_j(\theta_{j0})^{-1}\}$$

$$\| \quad E_{\tilde{X}}\left[\sum_{i=1}^{n} \log f_j(X_i|\hat{\theta}_j(\tilde{X}))\right] \quad (B)$$

Combine (A) and (B) to see

$$n Q_j \approx E_{\tilde{X}}\left[\sum_{i=1}^{n} \log f_j(X_i|\hat{\theta}_j(\tilde{X}))\right] - \text{Trace}\{J_j(\theta_{j0})FI_j(\theta_{j0})^{-1}\}$$

We replace the average (expected) training error $E_{\tilde{x}}\left[\sum_{i=1}^{\hat{n}} \log f(x_i | \hat{\theta}_j(\tilde{x}))\right]$ by the observed training error $\sum_{i=1}^{n} \log f_j(x_i | \hat{\theta}_j(\tilde{x})) = l_{f_j}(\hat{\theta}_j)$

$$\Rightarrow n Q_j \approx l_{f_j}(\hat{\theta}_j) - \text{Trace}\{J_j(\theta_{j0}) I_j(\theta_{j0})^{-1}\}$$

This depends on the unknown g so cannot be computed.

BUT, assume f_j a good approximation of g at θ_{j0}, then

$$J_j(\theta_{j0}) \approx I_j(\theta_{j0})$$

$$\Rightarrow n Q_j \approx l_{f_j}(\hat{\theta}_j) - \text{Trace}\{I_{p_j \times p_j}\}$$

$$= l_{f_j}(\hat{\theta}_j) - p_j$$

Usually we see this expressed as

$$AIC_j = -2 n Q_j = \underbrace{-2 l_{f_j}(\hat{\theta}_j)}_{\text{Deviance under model } f_j} + \underbrace{2 p_j}_{\text{penalty on model } f_j = 2 * \# \text{ parameters}}$$

Strategy: • find $\hat{\theta}_j$: maximizer of l_{f_j}, for all f_j

• evaluate $AIC_j = -2 l_{f_j}(\hat{\theta}_j) + 2p_j$

• choose f_j that minimizes AIC

Alternatives • create a training and test set from original data $\tilde{x}$

• LOOCV, let each x_i take turns as X and $x_{i' \neq i} = \tilde{x}$ training

Leave-one-out cross-validation

• $\hat{Q}_j = \frac{1}{n} \sum_{i=1}^{n} \log f_j(x_i \mid \hat{\theta}_j(x_{i' \neq i}))$

$= LOOCV_j$

• One can show $AIC_j \approx -2n\, LOOCV_j$

• Can also try leave-k-out cross-validation