Notes on Model Order Selection

Umut Orguner

This document is adapted from Professor Erik G. Larsson's lecture notes for the course Detection and Estimation Theory.

We first start with some definitions of the quantities we use in the following.

- $\boldsymbol{r}$ denotes the data.
- We suppose there are M alternative models out of which we need to decide which one would represent the data $\boldsymbol{r}$ best. This is in general an M-ary hypothesis testing problem where

  $$H_i \text{ is the hypothesis that the } i\text{th model is true.}$$

- Typically each model comes with its own unknown parameters denoted as $\boldsymbol{\theta}_i$.
- Model parameters has prior distribution denoted as $p(\boldsymbol{\theta}_i|H_i)$.
- Examples (model selection):
  - Regression-like polynomial fit with unknown polynomial order
  - FIR (tapped delay line) coefficient identification — unknown length
  - Determine rank of sample covariance matrix
  - Estimate of number of targets in radar

In the binary hypothesis case i.e., $M = 2$, the generalized likelihood ratio test (GLRT) is applicable. In the following, we consider the different approaches for the general case $M > 2$. Note that these approaches are equally applicable for the case $M = 2$.

1 Deterministic approach

- As a simple suggestion, suppose one simply tried

  $$\max_i \left(\max_{\boldsymbol{\theta}_i} p(\boldsymbol{r}|H_i, \boldsymbol{\theta}_i) \right) \tag{1}$$

  Unfortunately, this decision rule usually cannot do the job. If models are inclusive: $\boldsymbol{\theta}_i = [\boldsymbol{\theta}_{i-1}, \bar{\boldsymbol{\theta}}_i]$ it will always choose H_i over H_{i-1}

  $$\max_{\boldsymbol{\theta}_i} p(\boldsymbol{r}|\boldsymbol{\theta}_i) = \max_{\bar{\boldsymbol{\theta}}_i} \max_{\boldsymbol{\theta}_{i-1}} p(\boldsymbol{r}|\boldsymbol{\theta}_{i-1}, \bar{\boldsymbol{\theta}}_i) \geq \max_{\boldsymbol{\theta}_{i-1}} p(\boldsymbol{r}|\boldsymbol{\theta}_{i-1}, \bar{\boldsymbol{\theta}}_i = \text{fixed})$$

- Fundamental difficulty: A more flexible model fits the data better.
- Many methods rely on *penalization* of models with many parameters:

  $$\max_i \left(\log \left(\max_{\boldsymbol{\theta}_i} p(\boldsymbol{r}|H_i, \boldsymbol{\theta}_i) \right) - \psi_i \right)$$

  where ψ_i depends on the number of elements in $\boldsymbol{\theta}_i$ and in $\boldsymbol{r}$. Typical examples are

– Akaike information criterion (AIC):

$$\psi_i \sim \dim(\boldsymbol{\theta}_i) \tag{2}$$

– Generalized cross validatory Kullback-Leibler approach (GIC):

$$\psi_i \sim \frac{\dim(\boldsymbol{r})}{\dim(\boldsymbol{r}) - \dim(\boldsymbol{\theta}_i) - 1}\dim(\boldsymbol{\theta}_i)$$

– Minimum description length (MDL) – Bayesian information criterion:

$$\psi_i \sim \frac{1}{2}\dim(\boldsymbol{\theta}_i) \cdot \log(\dim(\boldsymbol{r})). \tag{3}$$

- What we want is "Occam's razor": Among two explanations for the same data, choose the "simplest".
- The deterministic approach naturally reduces to a GLRT when $M = 2$.
- The deterministic approach needs "ad hoc" regularization. We are going to show how a Bayesian approach would consider this problem in the next section.

2 Bayesian approach

- Eliminate $\boldsymbol{\theta}_i$ by marginalization

$$p(\boldsymbol{r}|H_i) = \int p(\boldsymbol{r}|H_i, \boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|H_i)d\boldsymbol{\theta}_i$$

and use standard Bayesian-cost criterion

With the minimum probability of error criterion, just compute

$$\boxed{\arg\max_i P(H_i|\boldsymbol{r}) = \arg\max_i \left(P(H_i)\int p(\boldsymbol{r}|\boldsymbol{\theta}_i, H_i)p(\boldsymbol{\theta}_i|H_i)d\boldsymbol{\theta}_i\right).} \tag{4}$$

- This integral can be hard to find in closed form!
- A tool we need: saddle-point (Laplace) approximation to a function $q(\boldsymbol{x})$ (think of $q(\boldsymbol{x})$ as "un-normalized pdf"). Suppose that $q(\cdot)$ is maximized at x_0. Then a Taylor series approximation for $\log q(\cdot)$ around x_0 is given as

$$\begin{aligned}\log q(\boldsymbol{x}) &= \log q(\boldsymbol{x}_0) + \underbrace{\nabla_{\boldsymbol{x}}^T \log q(\boldsymbol{x}_0)}_{=0}(\boldsymbol{x} - \boldsymbol{x}_0) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0) + \text{H.O.T.} \\ &= \log q(\boldsymbol{x}_0) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0) + \text{H.O.T.}\end{aligned} \tag{5}$$

where $\nabla_{\boldsymbol{x}}$ denote the gradient operator with respect to $\boldsymbol{x}$ and

$$\boldsymbol{A} \triangleq -\Delta_{\boldsymbol{x}}^{\boldsymbol{x}} \log q(\boldsymbol{x}_0) \tag{6}$$

where $\Delta_{\boldsymbol{x}}^{\boldsymbol{x}}$ denote the Hessian operation with respect to $\boldsymbol{x}$. Neglecting now the higher order terms and taking the exponential of both sides, we get

$$q(\boldsymbol{x}) \approx q(\boldsymbol{x}_0)\exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0)\right) \tag{7}$$

Then $\int q(x)dx \approx q(x_0) \cdot \sqrt{\left|2\pi\boldsymbol{A}^{-1}\right|}$.

- Interpretation: Gaussian approximation

$$\frac{q(\boldsymbol{x})}{\int q(x)dx} \approx \frac{1}{\sqrt{\left|2\pi \boldsymbol{A}^{-1}\right|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_0)} = N(\boldsymbol{x}_0, \boldsymbol{A}^{-1}) \tag{8}$$

- Apply this to the function $q(\boldsymbol{\theta}_i) = p(\boldsymbol{r}|\boldsymbol{\theta}_i, H_i)p(\boldsymbol{\theta}_i|H_i)$ in the integral of (4) at the maximizing point

$$\hat{\boldsymbol{\theta}}_i = \underset{\boldsymbol{\theta}_i}{\operatorname{argmax}} \left[p(\boldsymbol{r}|\boldsymbol{\theta}_i, H_i)p(\boldsymbol{\theta}_i|H_i)\right] = \underset{\boldsymbol{\theta}_i}{\operatorname{argmax}}\, P(\boldsymbol{\theta}_i|\boldsymbol{r}, H_i). \tag{9}$$

This corresponds to the approximation $P(\boldsymbol{\theta}_i|\boldsymbol{r}, H_i) \sim N(\hat{\boldsymbol{\theta}}_i, \boldsymbol{Q}_i^{-1})$ whose accuracy increases with size of data record. The matrix $\boldsymbol{Q}_i$ is given as

$$\begin{aligned} \boldsymbol{Q}_i &= -\Delta_{\boldsymbol{\theta}_i}^{\boldsymbol{\theta}_i} \log p(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i) - \Delta_{\boldsymbol{\theta}_i}^{\boldsymbol{\theta}_i} \log p(\hat{\boldsymbol{\theta}}_i|H_i) && (10)\\ &\approx -\Delta_{\boldsymbol{\theta}_i}^{\boldsymbol{\theta}_i} \log p(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i) && (11) \end{aligned}$$

in the large sample limit since the prior $p(\boldsymbol{\theta}_i|H_i)$ is independent of data length and the first term on the right hand side of (10) grows with data length. It is known that

$$\frac{1}{\dim(\boldsymbol{r})}\boldsymbol{Q}_i \approx \mathcal{O}(1) \tag{12}$$

in the large sample limit. Moreover, as the amount of data increases $\hat{\boldsymbol{\theta}}_i$ converges to the ML estimate.

- Using the approximation above

$$\begin{aligned} P(H_i|\boldsymbol{r}) &\propto P(\boldsymbol{r}|H_i)P(H_i) = P(H_i) \int \underbrace{P(\boldsymbol{r}|\boldsymbol{\theta}_i, H_i)P(\boldsymbol{\theta}_i|H_i)}_{\propto P(\boldsymbol{\theta}_i|\boldsymbol{r}, H_i)} d\boldsymbol{\theta}_i \\ &\approx P(H_i) \quad \underbrace{P(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i)P(\hat{\boldsymbol{\theta}}_i|H_i)}_{\text{peak of integrand (at } \boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i)} \cdot \underbrace{\sqrt{\left|2\pi \boldsymbol{Q}_i^{-1}\right|}}_{\text{Laplace factor}} \\ &= \underbrace{P(H_i)}_{\text{a priori}} \cdot \underbrace{P(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i)}_{\text{best fit likelihood}} \cdot \underbrace{P(\hat{\boldsymbol{\theta}}_i|H_i) \cdot \sqrt{\left|2\pi \boldsymbol{Q}_i^{-1}\right|}}_{\text{Occam factor}} \end{aligned}$$

Taking the logarithm of both sides

$$\begin{aligned} \log P(H_i|\boldsymbol{r}) &= \log P(H_i) + \log P(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i) + \log P(\hat{\boldsymbol{\theta}}_i|H_i) + \log \sqrt{\left|2\pi \boldsymbol{Q}_i^{-1}\right|} && (13)\\ &= \log P(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i) - \frac{1}{2}\log |\boldsymbol{Q}_i| + \mathcal{O}(1) && (14)\\ &= \log P(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i) - \frac{1}{2}\log \Big|\dim(\boldsymbol{r}) \underbrace{\frac{1}{\dim(\boldsymbol{r})}\boldsymbol{Q}_i}_{\approx \mathcal{O}(1)}\Big| + \mathcal{O}(1) && (15)\\ &= \log P(\boldsymbol{r}|\hat{\boldsymbol{\theta}}_i, H_i) - \frac{\dim(\boldsymbol{\theta}_i)}{2}\log(\dim(\boldsymbol{r})) + \mathcal{O}(1) && (16) \end{aligned}$$

which is the maximized function for BIC after neglecting the constants independent of $\dim(\boldsymbol{r})$.

- The "Occam factor" will automatically favor the less flexible model.

  No need to penalize flexibility nor to bias $P(H_i)$ towards simpler models.