THE ROCKEFELLER UNIVERSITY
Science for the benefit of humanity

Bayesian Model Selection

Luc Demortier

Statistics Miniworkshop: What Have We Learnt from the LHC Higgs Search?

CERN, February 13–14, 2013

Outline

Bayesian Foundations - General

(From J. M. Bernardo & A. F. M. Smith, "Bayesian Theory," Wiley, 2000)
A decision problem is defined by the following elements:

1. An algebra of events $\mathcal{E}$;
2. A set of possible consequences $\mathcal{C}$;
3. A set of options $\mathcal{A}$, or potential acts, consisting of functions that map events into consequences;
4. A preference ordering $\leq$ among some elements of $\mathcal{A}$.

In order to analyze a given decision problem $\{\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq\}$ in accordance with the axioms of quantitative coherence, we must represent degrees of belief about uncertain events in the form of a finite probability measure over $\mathcal{E}$ and values for consequences in the form of a utility function over $\mathcal{C}$. Options are then to be compared on the basis of expected utility.

Bayesian Foundations - Inferences

The problem of reporting inferences is essentially a special case of a decision problem. This can be seen with the following correspondences:

- Each event E_j in $\mathcal{E}$ has the interpretation $E_j \equiv$ "the hypothesis H_j is true";
- The actions available to an individual, or options, are the various inference statements that might be made about the E_j; for coherence reasons these inference statements must take the form of a conditional probability distribution over the E_j, given observed data D and initial information;
- Corresponding to each inference statement and each E_j, there will be a consequence, namely the record of the statement together with what actually turned out to be the case.

There is nothing in this framework that says an individual has to report the probability distribution corresponding to his or her personal beliefs. There still needs to be provided a utility function $u(q, E_j)$ describing the value of reporting the probability distribution q as the final inferential summary of the investigation, if E_j is the true "state of nature". This special type of utility function is often called a score function. The probability distribution to report is then the one that maximizes expected utility:

$$\sum_{j \in J} u(q, E_j)\, P(E_j \mid D).$$

Model Selection Perspectives

Suppose we have a set of models $\mathcal{M} = \{M_i, i \in I\}$ under consideration for observations x. Before we examine methods for comparing these models or choosing among them, it is important to distinguish three ways in which the set of models can be viewed:

1. The $\mathcal{M}$-closed view
   In this view we believe that $\mathcal{M}$ actually contains the true model for the observed data. We just don't know which one it is. We have prior weights $P(M_i)$ for each model, so that the overall belief for x is of the form
   $$p(x) = \sum_{i \in I} P(M_i)\, p(x \mid M_i).$$
2. The $\mathcal{M}$-completed view
   Here we still have an overall belief $p(x)$ for x, but it may be intractable from the point of view of computation or implementation or communication of results, so instead we work with a range of models that serve as a proxy. There is no sense in assigning prior weights $P(M_i)$. However the models will still have to be evaluated and compared in the light of the actual beliefs $p(x)$.
3. The $\mathcal{M}$-open view
   Again the M_i constitute a range of models available for comparison, but now there isn't even an overall belief specification $p(x)$.

Model Selection as a Decision Problem (1)

There are several types of decision problems involving model selection:

1. Just choosing a model
   The utility function has the form $u(m_i, \omega)$ with ω some unknown of interest. The optimal model choice $m^\star$ is given by:

$$\bar{u}(m^\star \mid x) = \sup_{i \in I} \bar{u}(m_i \mid x), \quad \text{where} \quad \bar{u}(m_i \mid x) = \int u(m_i, \omega)\, p(\omega \mid x)\, d\omega,$$

   and $p(\omega \mid x)$ represents actual beliefs about ω having observed x.

2. Choosing a model M_i, and then given that model, obtaining an answer a_i about an unknown ω of interest
   Here the optimal model choice is

$$\bar{u}(m^\star \mid x) = \sup_{i \in I} \bar{u}(m_i \mid x), \quad \text{where} \quad \bar{u}(m_i \mid x) = \int u(m_i, a^\star, \omega)\, p(\omega \mid x)\, d\omega,$$

   and $a^\star$ is obtained by maximizing:

$$\int u(m_i, a_j, \omega)\, p(\omega \mid x)\, d\omega.$$

Model Selection as a Decision Problem (2)

3. Directly obtaining an answer a to an inference problem
   Here we omit the model choice step and the optimal answer $a^\star$ is given by:

$$\bar{u}(a^\star \mid x) = \sup_a \bar{u}(a \mid x), \quad \text{where} \quad \bar{u}(a \mid x) = \int u(a, \omega)\, p(\omega \mid x)\, d\omega.$$

   A model comparison is still being done implicitly.

Zero-One Utilities and Bayes Factors

In the $\mathcal{M}$-closed perspective, the problem of choosing the "true" model can be solved with a zero-one form for the utility function:

$$\begin{aligned} u(m_i \mid \omega) &= 1 \text{ if } \omega = M_i, \\ &= 0 \text{ if } \omega \neq M_i. \end{aligned}$$

The expected utility of choosing M_i given x is therefore:

$$\bar{u}(m_i \mid x) = \int u(m_i, \omega)\, p(\omega \mid x)\, d\omega = p(M_i \mid x).$$

The optimal decision is therefore to choose the model with the highest posterior probability. Pairwise model comparisons can then be done using the posterior-odds ratio:

$$\frac{p(M_i \mid x)}{p(M_j \mid x)} = \frac{p(x \mid M_i)}{p(x \mid M_j)} \times \frac{p(M_i)}{p(M_j)},$$

where for example

$$p(x \mid M_i) = \int p_i(x \mid \theta_i)\, p_i(\theta_i)\, d\theta_i.$$

The Jeffreys-Lindley Paradox in the M-Closed Perspective (1)

Suppose that for some data $x = (x_1, \ldots, x_n)$ we have two alternative models:

$$M_1: \quad p_1(x) = \prod_{i=1}^{n} G(x_i \mid \mu_0, \sigma^2),$$

$$M_2: \quad p_2(x) = \int \prod_{i=1}^{n} G(x_i \mid \mu, \sigma^2) G(\mu \mid \mu_1, \sigma_1^2)\, d\mu,$$

where μ_0, μ_1, σ, and σ_1 are all known.

This is equivalent to testing $H_1 : \mu = \mu_0$ versus $H_2 : \mu \neq \mu_0$. The Bayes factor in favor of H_1 only depends on the mean $\bar{x}$ and is:

$$B_{12}(x) = \frac{G(\bar{x} \mid \mu_0, \sigma^2/n)}{\int G(\bar{x} \mid \mu, \sigma^2/n)\, G(\mu \mid \mu_1, \sigma_1^2)\, d\mu} = \sqrt{\frac{\sigma_1^2 + \sigma^2/n}{\sigma^2/n}} \frac{\exp\left[-\frac{1}{2}\frac{(\bar{x}-\mu_0)^2}{\sigma^2/n}\right]}{\exp\left[-\frac{1}{2}\frac{(\bar{x}-\mu_1)^2}{\sigma_1^2+\sigma^2/n}\right]}.$$

For any fixed $\bar{x}$, regardless of how many σ's $\bar{x}$ is from μ_0, $B_{12}(x) \to \infty$ as $\sigma_1 \to \infty$, i.e. the evidence in favor of M_1 becomes overwhelming as the prior precision on M_2 vanishes.

The Jeffreys-Lindley Paradox in the M-Closed Perspective (2)

An alternative approach is possible for Lindley's paradox by viewing it as model rejection problem, where one model (M_1) is a parametric restriction of the other (M_2). In this case one can look at the difference in utilities in favor of the larger model, and reject the restricted model if

$$t(x) = \int [u(m_2, \mu) - u(m_1, \mu)]\ p(\mu \mid x)\ d\mu > c(x),$$

where $c(x)$ is the utility premium attached to keeping the simpler model M_1.

Using a logarithmic score function for the utilities yields

$$\delta(\mu) = u(m_2, \mu) - u(m_1, \mu) = \int G(x \mid \mu, \sigma^2) \log \frac{G(x \mid \mu, \sigma^2)}{G(x \mid \mu_0, \sigma^2)}\ dx = \frac{1}{2} \frac{(\mu - \mu_0)^2}{\sigma^2/n}.$$

Within an M_2-closed perspective we can calculate a reference posterior for μ and obtain the corresponding *expected* difference in utilities:

$$t(x) = \int \frac{1}{2} \frac{(\mu - \mu_0)^2}{\sigma^2/n}\ G(\mu \mid \bar{x}, \sigma^2/n)\ d\mu = \frac{1}{2}[1 + z^2],$$

where $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$ is the standard significance test statistic for a normal location null hypothesis.

Note on Nested Hypothesis Testing

Since the publication of his book with A. F. M. Smith, Bernardo has written more extensively about nested hypothesis testing, see for example J. M. Bernardo, "Nested hypothesis testing: The Bayesian reference criterion," in *Bayesian Statistics 6*, pp. 101-130, Oxford University Press, 1999.

In Lindley's paradox the distributions are all Gaussians and there is, *for all sample sizes,* a one-to-one correspondence between thresholds on $t(x)$ and frequentist significance levels. However this exact correspondence does not carry over to other distributions.

For testing the parameter θ of an exponential distribution for example,

$$p(x \mid \theta) = \theta^n \ \exp(-n\bar{x}\theta),$$

using n data points with average $\bar{x}$, Bernardo's procedure calls for rejecting a null value θ_0 whenever the posterior expected intrinsic discrepancy is greater than 5, which implies Type-I error probabilities of 0.0572, 0.0051, 0.0029, and 0.0027 when the sample size is, respectively, 2, 10, 100, and 1000. Hence Bernardo's procedure corresponds to a decreasing significance level as the sample size increases.

The Jeffreys-Lindley Paradox in the M-Open Perspective

In the $\mathcal{M}$-open view, we do not have any prior beliefs about M_1 or M_2. How can we proceed? One possibility is to do a type of cross-validation analysis: "how well does M_1 or M_2 perform when it attempts to predict a left-out subset of the data from the remaining subset of the data?" In other words, under which model do the data achieve the highest level of "internal consistency"?

This can be given a formulation in terms of quadratic score functions. For Lindley's paradox, the result is that model M_2 will be preferred if on average the posterior mean does better as a predictor than μ_0:

$$\frac{1}{n}\sum_{j=1}^{n}\left[\{(1-w_{n-1})\mu_1 + w_{n-1}\bar{x}_{n-1}(j)\} - x_j\right]^2 > \frac{1}{n}\sum_{j=1}^{n}(\mu_0 - x_j)^2,$$

where $w_{n-1} = \sigma_1^2/[\sigma_1^2 + \sigma^2/(n-1)]$ and $\bar{x}_{n-1}(j)$ is the mean of the sample x with x_j omitted. In the limit of $\sigma_1 \to \infty$, the above inequality is approximately equivalent to:

$$\frac{n(\mu_0 - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2/(n-1)} > 2$$

(The left-hand side is a random Snedecor $F_{1,n-1}$ quantity.)

8. Concluding Remarks

- The decision-theory approach with utility functions allows for a wide variety of model selection and model comparison techniques.
- The problem of inference after testing can be solved by an appropriate choice of utility function.
- The resolution of Lindley's paradox depends on one's view of the available models and on the information available.
- One of the lessons of Lindley's paradox is that Bayes factors are sensitive to prior information.