Seminar über Statistik FS2008:

Model Selection

Alessia Fenaroli, Ghazale Jazayeri

Monday, April 21, 2008

1 Introduction

Model Choice deals with the comparison of models and the selection of a model. It can be considered as a special case of testing where the sampling distribution is depending on possibly infinitely many unknown parameters:

$$M_i : x \sim f_i(x \mid \theta_i)$$

where M_i are the models in comparison, $\theta_i \in \Theta_i$ the parameter space and $i \in I$ possibly infinite.
From another point of view model choice is closer to estimation than to testing. It involves many possibilities M_1,...,M_p and selecting a model is equivalent to estimating the index of this model.
The purpose is estimating the true sampling distribution f, which is unknown, by using posterior probabilities and Bayes factors.

2 Two approaches to Bayesian testing

Testing hypothesis $H_0 : \theta \in \Theta_0 \; vs \; H_A : \theta \in \Theta_1$ in Bayesian fashion is interesting because the notion of probability of a hypothesis can only be defined trough this way:

$$X \mid \theta \sim f(x \mid \theta)$$

where θ has prior distribution π and $\theta \in \Theta = \Theta_0 \cup \Theta_1$ for Θ_0 and Θ_1 disjoint.
Two approaches to Bayesian testing are considered: decision theoretic approach and Bayes factor.

2.1 Decision theoretic approach

Consider the action space $\{0, 1\}$ and the test procedure φ.
If we use the generalized $0-1$ loss

$$L(\theta, \varphi) = \begin{cases} 0, & \varphi = I\{\theta \in \Theta_0\} \\ K_0, & \theta \in \Theta_0 \text{ and } \varphi = 0 \\ K_1, & \theta \notin \Theta_0 \text{ and } \varphi = 1 \end{cases}$$

then the Bayesian decision (which minimizes the posterior expected loss) is

$$\varphi^\pi(x) = \begin{cases} 1 & \text{if } P_\pi[\theta \in \Theta_0 \mid x] > \frac{K_1}{K_0+K_1} = \frac{1}{\frac{K_0}{K_1}+1} \\ 0 & \text{otherwise} \end{cases}$$

The null hypothesis is rejected when the posterior probability is too small (smaller than the acceptance level $K_1/(K_0 + K_1)$). The Bayesian decision only depends on the ratio K_0/K_1: the larger it is the smaller the posterior probability of H_0 needs to be (for H_0 to be accepted).
Here the difficulty is the choice of K_0 and K_1, which usually are selected automatically rather than from utility considerations.

2.2 Bayes factor

Definition 1. *The Bayes factor in favor of H_0 is defined by*

$$B_{01}^{\pi}(x) = \frac{P[\theta \in \Theta_0 \mid x]/P[\theta \in \Theta_1 \mid x]}{\pi(\theta \in \Theta_0)/\pi(\theta \in \Theta_1)}$$

This factor evaluates the modification of the odds of Θ_0 against Θ_1 due to the observation and measures the relative change in prior odds once the evidence is collected.
Jeffreys developed a scale to judge the evidence in favor or against H_0 brought by the data:

- if $\log_{10}(B_{01}^{\pi})$ varies between 0 and 0.5, the evidence against H_0 is *poor*,
- if it is between 0.5 and 1, it is *sustantial*,
- if it is between 1 and 2, it is *strong*, and
- if it is above 2, it is *decisive*

The precise bounds of this scaling separating one strength from another are a matter of convention and they can be arbitrarily changed.
In the particular case where Θ_0=$\{\theta_0\}$, Θ_1=$\{\theta_1\}$ and the prior probabilities are $\pi(\theta = \theta_0) = \pi_0$ and $\pi(\theta = \theta_1) = \pi_1 = 1 - \pi_0$, the Bayes factor simplifies to the usual *likelihood ratio*:

$$B_{01}^{\pi}(x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)}$$

If the prior is a mixture of two priors, $\xi_0(\theta)$ under H_0 and $\xi_1(\theta)$ under H_A with weights π_0 and π_1, i.e. $\pi(\theta) = \pi_0\xi_0(\theta) + \pi_1\xi_1(\theta)$, the bayes factor is:

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x \mid \theta_0)\pi_0\xi_0(\theta)\mathrm{d}\theta / \int_{\Theta_1} f(x \mid \theta_1)\pi_1\xi_1(\theta)\mathrm{d}\theta}{\pi_0/\pi_1} = \frac{m_0(x)}{m_1(x)}$$

where m_0 and m_1 are the marginals under both hypotheses.

3 Some criticisms and remarks

3.1 Continuos prior distribution

In general model choice is incompatible with absolutely continuos (w.r.t. the Lebesgue measure) prior distributions.
The motivation is that testing a point null hypothesis H_0: $\theta = \theta_0$ vs. H_A: $\theta \neq \theta_0$ with a continuos prior distribution implies $\pi(\theta \in \Theta_0) = 0$, but the Bayes factor is only defined when $\pi(\theta \in \Theta_0) \neq 0$.
The solution of this problem requires a modification of the prior:

$$\pi(\theta) = \pi_0\delta_{\theta_0}(\theta) + \pi_1\xi(\theta)$$

where $\pi_0 = 1 - \pi_1$ prior weights, $\delta_{\theta_0}(\theta)$ Dirac mass and $\xi(\theta)$ the spread distribution under H_A.
Then we get the following marginal density for $X \mid \theta \sim f(x \mid \theta)$:

$$m(x) = \int_{\Theta} (\pi_0\delta_{\theta_0}(\theta) + \pi_1\xi(\theta))f(x \mid \theta)\mathrm{d}\theta = \pi_0 f(x \mid \theta_0) + \pi_1 m_1(x)$$

where $m_1(x) = \int f(x \mid \theta)\xi(\theta)\mathrm{d}\theta$ is the marginal density under H_A.

3.2 Relation between Bayes factor and posterior probability

Under the above situation holds:

$$\pi(\Theta_0 \mid x) = \Big[1 + \frac{1 - \pi_0}{\pi_0}\frac{1}{B_{01}^{\pi}(x)}\Big]^{-1}$$

3.3 Improper priors and Pseudo-Bayes factors

Improper priors should not be used at all in tests, since they are incompatible with most tests of point-null hypothesis. It doesn't feel right to use improper priors because they seem to lead to too much arbitrariness (many competing solutions contradict the Likelihood principle). A solution to overcome the difficulties related to improper priors can be represented by Pseudo-Bayes factors, which we are going to introduce.

Definition 2. *Given the improper prior π, a sample $(x_1, ..., x_n)$ is a* training sample *if $\pi(. \mid x_1, ..., x_n)$ is proper.*
It is a minimal training sample *if no subsample is a training sample.*

The idea is to use minimal training sample $x_{(\ell)}$ (where ℓ is its lenght) to "properize" π into $\pi(. \mid x_{(\ell)})$ and then use this posterior as if it were a regular proper prior for the remainder sample $x_{(-\ell)}$.

Definition 3. *Consider the null hypothesis H_0 with prior π_0 and the alternative hypothesis H_A with π_1. Let $x_{(\ell)}$ be the minimal training sample s. t. $\pi_0(. \mid x_{(\ell)})$ is also proper, then the pseudo-Bayes factor is:*

$$B_{10}^{(\ell)} = \frac{\int_{\Theta_1} f_1(x_{(-\ell)} \mid \theta_1)\pi_1(\theta_1 \mid x_{(\ell)})\mathrm{d}\theta_1}{\int_{\Theta_0} f_0(x_{(-\ell)} \mid \theta_0)\pi_0(\theta_0 \mid x_{(\ell)})\mathrm{d}\theta_0}$$

There are still some difficulties with the pseudo-Bayes factor. Moreover, there is no obvious choice for $x_{(\ell)}$, while this choice of the training sample influences the resulting value of $B_{10}^{(\ell)}$.
A way to remove this dependence on the training sample is to average the different pseudo-Bayes factors over all the possible training samples $x_{(\ell)}$.

4 Model Choice

4.1 Introduction

In contrast with the other sections, we are now dealing with models, rather than with parameters, and the sampling distribution f is unknown rather than simply depending on an unknown (finite dimensional) parameter.

4.1.1 Choice between models

Model choice seems to elude the Bayesian paradigm in that the sampling distribution f is itself uncertain, making it difficult to condition on the observation x.
We consider the more restricted setting where several (parametric) models are in competition,

$$M_i : x \sim f_i(x \mid \theta_i), \qquad \theta_i \in \Theta_i, \qquad i \in I$$

the index set I being possibly infinite.
A prior distribution can be constructed for each model M_i, as if it were the only and true model under consideration.
Remarks:

- In the simpliest case: the choice is between a small number of models that have been chosen for convenience, historical or more motivated reasons.
- In more complex cases: the number of potential models is large because the available information is too limited to eliminate most of them. We are then closer to the nonparametric perspective.

- In case of regression models: the variety of models stems from the large number of combinations of covariates (explanatory variables) that could be included in the model.
- There is often a high degree of arbitrariness involved in the selection of the models to choose from.
- While no model is true, *several* models may be appropriate.
- We have also to consider that some models are submodels of others.

From a modeling point of view, the larger model should be preferred, while from a statistical point of view, this is not so clear, given that more parameters need to be estimated from the same sample! The model choice criterion must thus include parts that weight the fit, as well as parts that incorporate the estimation error.

4.1.2 Model choice: motives and uses

We can identify the choice of a model as:

- a first step in *model construction*, where few models come to mind and we want to decide which one fits best the data at hand. There is no reason to believe that one of these models is correct.
- a last step of *model checking*. A model or a family of models has been selected for various theoretical and practical reasons, and we want to know whether the data agrees with this type of model.
- a call for *model improvement*. The goal is to introduce possible refinements of a given model to improve the fit or to create an *embedding* of the existing model in a class of models to check whether the current model is good enough.
- the reverse need of *model pruning*, where the current model is too complicated to be of practical use and where simpler submodels are examined to see whether they fit the data well enough. We want to reduce the whole range of covariates to a few important covariates.
- a *model comparison*, when a few models are proposed because they fitted correctly other samples and we wonder which of these models best fits the current sample.
- a purpose of *hypothesis testing*, where several models are built from theoretical considerations and then tested through specially designed experiments.
- a requirement of *prediction efficiency*, as in finance. Here we are only interested in the prediction preformances of different models.

4.2 Standard framework

4.2.1 Prior modeling for model choice

We can write the parameter space associated with the set of models as

$$\Theta = \bigcup_{i \in I} (\{i\} \times \Theta_i),$$

the model indicator $\mu \in I$ being part of the parameters.
So, if we can assign probabilities p_i to the models M_i $(i \in I)$ and then define

priors $\pi_i(\theta_i)$ on the parameter subspaces Θ_i, we get, by Bayes formula, the posterior model probability given the data:

$$\begin{aligned} p(M_i \mid x) &= P(\mu = i \mid x) = \frac{p(x \mid M_i)p_i}{\sum_j p(x \mid M_j)p_j} \\ &= \frac{p_i \int_{\Theta_i} f_i(x \mid \theta_i)\pi_i(\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x \mid \theta_j)\pi_j(\theta_j)d\theta_j}, \end{aligned}$$

where $p(x \mid M_i) = \int_{\Theta_i} f_i(x \mid \theta_i)\pi_i(\theta_i)d\theta_i$.
A common solution, based on this prior modeling, for model selection is simply to determine the model with the largest $p(M_i \mid x)$.
Difficulties:

- the solution based on the representation of the collection of models requires the construction of a prior distribution (π_i,p_i) for each $i \in I$, which is delicate when I is infinite. Moreover, these priors π_i must all be proper because there is no unique scaling for improper priors.

- if some models are embedded into others, that is, if $M_{i0} \subset M_{i1}$, there should be some coherence in the choice of π_{i0} given π_{i1} and maybe also in the choice of p_{i0} given p_{i1}.
  For instance, if $M_1 = M_2 \cup M_3$, one could argue that $p(M_1) = p(M_2) + p(M_3)$ or at least $p(M_1) \geq p(M_2) + p(M_3)$. Similarly, if two models M_{i0} and M_{i1} are not embedded in one another, there is the possibility of a third model M_{i2} embedding both M_{i0} and M_{i1}.

Another type of difficulty is associated with the computation of predictives, marginals and other quantities related to the model choice procedures.
An important point is that parameters common to several models must be treated as separate entities.

4.2.2 Bayes factors

Once the modeling representation is accepted we get a generic testing problem. The solution proposed is to call for Bayes factors:

$$B_{12} = \frac{P(M_1 \mid x)}{P(M_2 \mid x)} \Big/ \frac{P(M_1)}{P(M_2)} = \frac{\int_{\Theta_1} f_1(x \mid \theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int_{\Theta_2} f_2(x \mid \theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}$$

for the comparison of models M_1 and M_2, where $p(M_i \mid x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}$ and $p_i = P(M_i)$.
Bayes factors are independent of the priors to the models p_i and measure the relative strength of evidence on the data of model M_1 over model M_2.
Note that the comparison of models based on Bayes factors can proceed one pair (M_i, M_j) at a time.
Unfortunately, while Bayes factors are rather intuitive, as a practical matter, they are often quite difficult to calculate. The main problem is given by the evaluation of the integral $\int_{\Theta_i} f_i(x \mid \theta_i)\pi_i(\theta_i)d\theta_i$.
Improper priors cannot be used (they create additional difficulties because then the integral doesn't exist) and vague priors, that is, proper priors with a large variance, do not solve the difficulty.
The solution to this fundamental difficulty with improper priors is to use approximative Bayesian solutions, calling for minimal training samples. Intrinsic and fractional Bayes factors can be proposed as evaluation of the models under improper priors.

The Jeffrey's scale of evidence for Bayes factor is then so defined:

Bayes factor	Interpretation
$B_{12} < \frac{1}{10}$	strong evidence for M_2
$\frac{1}{10} < B_{12} < \frac{1}{3}$	moderate evidence for M_2
$\frac{1}{3} < B_{12} < 1$	weak evidence for M_2
$1 < B_{12} < 3$	weak evidence for M_1
$3 < B_{12} < 10$	moderate evidence for M_1
$B_{12} > 10$	strong evidence for M_1

4.2.3 Schwarz's criterion

There are many methods for chooosing between competing models. We have discussed the Bayesian approach.
The two most common alternatives to the Bayes factor are Schwarz's criterion (also called BIC) and Akaike's Information Criterion (AIC). We will concentrate our attention on the first one.
We consider asymptotic approximations to Bayes factors.
Applying the *Laplace expansion*, which is an approximation, to the Bayes factor we get:

$$\log(B_{12}) \simeq \log(\lambda_n) + \frac{p_2 - p_1}{2}\log(n) + K(\hat{\theta}_{1,n}, \hat{\theta}_{2,n}),$$

where p_1 and p_2 are the dimensions of Θ_1 and Θ_2, λ_n is the standard likelihood ratio for the comparison of M_1 with M_2,

$$\lambda_n = \frac{L_{1,n}(\hat{\theta}_{1,n})}{L_{2,n}(\hat{\theta}_{2,n})},$$

$L_{1,n}$ and $L_{2,n}$ are the likelihood functions based on n observations, and $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ are the maxima of L_1 and L_2, respectively. $K(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ denotes the remainder term.
This approximation leads to *Schwarz's criterion*, also called BIC (Bayes Information Criterion):

$$S = -\log(\lambda_n) - \frac{p_2 - p_1}{2}\log(n)$$

when $M_1 \subset M_2$, if the remainder term $K(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ is negligible compared with both other terms.
For regular models, when $M_1 \subset M_2$, the likelihood ratio is approximately distributed as a $\chi^2_{p_2-p_1}$ distribution, $-2\log(\lambda_n) \approx \chi^2_{p_2-p_1}$, if M_1 is the true model. Since $P(M_2 \quad \text{chosen} \mid M_1) = P(\lambda_n < c \mid M_1) \simeq P(\chi^2_{p_2-p_1} > -2\log(c)) > 0$, it follows that a criterion based only on the likelihood ratio does not converge to a sure answer under M_1.
This is why penalization factors have been added to the (log) likelihood ratio, starting with Akaike's Information criterion:

$$-2\log\lambda_n - \alpha(p_2 - p_1)$$

and then also Bayes Information Criterion.
Remarks:

- The "best" model is the one with maximum BIC.
- With BIC we take into account both the statistical goodness of fit and the number of parameters that have to be estimated to achieve this particular degree of fit, by imposing a penality for increasing the number of parameters.
  Hence upper BIC implies either fewer parameters, better fit, or both.

Schwarz's criterion provides an approximation to the Bayes factor, but this criterion is not relevant in a Bayesian setting, since

- the dependence on the prior assumption disappears
- the approximation only works for regular models.

Moreover it requires the derivation of the maximum likelihood estimates for all models!!

Comparing AIC and BIC we see that:

- BIC penalizes models with more parameters even more severely than AIC.
- BIC is consistent if the data are generated by one model with fixed dimension, whereas AIC tends to overestimate the dimension.
- AIC tends to do better than BIC for prediction.

4.2.4 Bayesian deviance

Deviance Information Criterion (DIC) is a Bayesian alternative, based on the *deviance*, to both AIC and BIC.
This criterion is more satisfactory because it takes into account the prior information and gives a natural penalization factor to the log-likelihood. It also allows for improper priors.
Like BIC and AIC it's an asymptotic approximation as the sample size becomes large.

$$\text{DIC} = \mathbb{E}[D(\theta) \mid x] + p_D = \mathbb{E}[D(\theta) \mid x] + \{\mathbb{E}[D(\theta) \mid x] - D(\mathbb{E}[\theta \mid x])\},$$

where for a model $f(x \mid \theta)$ associated with a prior distribution $\pi(\theta)$, the Deviance is $D(\theta) = -2\log(f(x \mid \theta))$.
$\mathbb{E}[D(\theta) \mid x]$ can be interpreted as a measure of fit (how well the model fits the data). The larger this is the worse the fit.
p_D is a measure of complexity, also called the *effective number of parameters*.
Models with smaller DIC should be preferred to models with larger DIC.

References

[1] Robert: *The Bayesian Choice*, Springer NY 2007

[2] Wasserman: *Bayesian Model Selection and Model Averaging*

[3] Kuensch: *"Model selection: An overview "*

[4] www2.isye.gatech.edu/∼ brani/isyebayes/handouts.html *Handouts 7 and 15*