1 Model comparison

The choice of an appropriate structure to represent features observed in data is an inherent part of data analysis. In statistical modeling this corresponds to the choice of an appropriate probability model and necessitates formulation of the model structure and often also estimation of its dimension.

The question of *how* one should compare probability models is to a large extent a philosophical issue. In broad terms a widely accepted Occam's razor principle says that unnecessary parts should be eliminated from a scientific theory, *i.e.* the parts which cannot be empirically verified. In the current context it is quite natural to restrict ourselves to the empirical verification, although other forms, such as the use of logic, can be considered in general. We shall see that the Occam's razor principle is automatically built in the Bayesian approach.

In reality, in general, there always exists a discrepancy between models and observations. Such a discrepancy might already arise from the consideration of the accuracy of some measurements that are made and give rise to our data. Second, since all data analysis is performed by computers where any numbers are represented by finite precision binary digits, it is natural to view models involving densities for continuous variables as approximations for the data. Nevertheless, to give us guidance in the construction of tools for model comparison, it is sometimes useful to imagine a "computer game scenario", where the observations are generated from a model which belongs to a class of models known to us. The uncertainty in this situation arises from the fact that we don't know which of the models can be seen as responsible for the data. This imaginary construct enables us to distinguish good ideas (model comparison strategies) from the less good ones in the ideal world. What one really hopes then, is that the solutions found to be good in the computer game scenario, would continue to be good in the real world if our models are sufficiently good descriptions of the regularities involved in the phenomenon under investigation. Conversely, if a model comparison strategy turns out to be a poor one in the ideal world, we expect it to be a poor one even in the real world.

From the Bayesian point of view, one could easily say that models exist in our heads, representing subjective beliefs about some phenomenon. They are abstract constructs aimed to give perceivable structures to complicated real-world phenomena that can be communicated with others. The typical statistical interpretation of models as data generating machines is not very realistic or elaborate in this respect. Quoting Rissanen (1987, p.223), "As in Bayesian theory the class of models is not intended to include any "true" distribution for the data, but rather is only regarded as a language in which the properties of the data are to be expressed. This is a minimum requirement for any kind of learning, for how can we find regular features in the data unless we can describe them."

2 A decision theoretic framework

We now consider the model comparison issue in the Bayesian framework. To proceed, assume that all elements in our finite class of belief models $\mathcal{M}$ are such that the joint density of the observations may be described in terms of a finite-dimensional parameter. In this section we use I as the set of values for an index variable, which corresponds to a unique mapping of the elements of $\mathcal{M}$. Thus, the predictive distributions for the alternative models are described by

$$p_i(\mathbf{x}) = p(\mathbf{x}|M_i) = \int p_i(\mathbf{x}|\boldsymbol{\theta}_i)p_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i,\ i \in I \tag{1}$$

Notice, that each element in I constitutes according to our beliefs a possible predictive model for the data. In order to proceed coherently in the Bayesian framework we then have to build an overall belief model for $\mathbf{x}$ by assigning weights to the different alternatives, so that our subjective beliefs are represented probabilistically. The overall model takes the form

$$p(\mathbf{x}) = \sum_{i \in I} P(M_i)p(\mathbf{x}|M_i) \tag{2}$$

where $P(M_i)$ are the weights of the individual belief models such that $\sum_{i\in I} P(M_i) = 1$. In the literature there has been a considerable amount of discussion about the interpretation of the weights $P(M_i)$. On one hand, they may be considered as *a priori* probabilities that the corresponding models are "true". On the other hand, they can simply be regarded as means of representing the degree of dominance (or perhaps functions of odds) of a subjective belief over another.

We first investigate the case where the only action to be taken is a choice of a model $M_i, i \in I$. Let ω be an unknown of interest, such that the utility function for our decision problem has the form $u(M_i, \omega)$. Using the decision theoretic approach we know that the optimal decision is to choose the model M^* which maximizes the expected utility according to

$$\bar{u}(M^*|\mathbf{x}) = \sup_{i \in I} \bar{u}(M_i|\mathbf{x}) \tag{3}$$

where

$$\bar{u}(M_i|\mathbf{x}) = \int u(M_i, \omega)p(\omega|\mathbf{x})d\omega,\ i \in I \tag{4}$$

where $p(\omega|\mathbf{x})$ represents the beliefs about ω having observed $\mathbf{x}$. These have further the form

$$p(\omega|\mathbf{x}) = \sum_{i \in I} p_i(\omega|M_i, \mathbf{x})P(M_i|\mathbf{x}) \tag{5}$$

where

$$P(M_i|\mathbf{x}) = \frac{P(M_i)p(\mathbf{x}|M_i)}{\sum_{i\in I} P(M_i)p(\mathbf{x}|M_i)} \tag{6}$$

is the posterior predictive weight or subjective posterior probability of the individual model being the "true" model. Notice that there is certainly nothing wrong in the latter definition if one restricts the interpretation to an observer's narrow (or perhaps naive) perception of the world.

If we let the above unknown of interest ω be simply the "true" model among those in I, the decision problem is concretized as follows. A natural utility function takes the form (0-1 loss)

$$u(M_i, \omega) = \begin{cases} 1 \text{ if } \omega = M_i \\ 0 \text{ if } \omega \neq M_i \end{cases} \tag{7}$$

It then follows that

$$p_i(\omega|M_i, \mathbf{x}) = \begin{cases} 1 \text{ if } \omega = M_i \\ 0 \text{ if } \omega \neq M_i \end{cases} \tag{8}$$

and

$$p(\omega|\mathbf{x}) = \begin{cases} P(M_i|\mathbf{x}), & \text{if } \omega = M_i \\ 0, & \text{if } \omega \neq M_i \end{cases} \tag{9}$$

The expected utility of the choice M_i is

$$\begin{aligned} \bar{u}(M_i|\mathbf{x}) &= \int u(M_i, \omega) p(\omega|\mathbf{x}) d\omega \\ &= P(M_i|\mathbf{x}) \end{aligned} \tag{10}$$

As might be intuitively expected we see that the optimal decision in this case is to choose the model with the highest posterior probability. It can be shown that, under the "computer game scenario" mentioned earlier, $P(M_i|\mathbf{x}) \rightarrow 1$ for the "true" model as $n \rightarrow \infty$, meaning that the Bayes procedure is *consistent*.

In the case where only two models (say M_1 and M_2) are available for comparison, a measure of plausibility is the *Bayes factor* specified below.

Definition 1 ***Bayes factor***. *Given two models M_1 and M_2 for data $\mathbf{x}$, the Bayes factor in favor of M_1 (and against M_2) is given as the posterior to prior odds ratio*

$$B_{12} = \frac{p(\mathbf{x}|M_1)}{p(\mathbf{x}|M_2)} = \frac{P(M_1|\mathbf{x})}{P(M_2|\mathbf{x})} \Big/ \frac{P(M_1)}{P(M_2)} \tag{11}$$

Intuitively, the Bayes factor says whether the data have increased ($B_{12} > 1$) or decreased ($B_{12} < 1$) the odds on M_1. Clearly, if the prior weights are uniform, the Bayes factor is simply a ratio of the posterior weights. A thorough discussion about the properties and guidelines for interpretation of the Bayes factor can be found in Kass and Raftery (1995).

The above described approach as such is not the most sensible solution in all situations, especially if we are aiming to produce some kind of statements about observables using our models (*e.g.* prediction of future values). An optimal strategy under such circumstances does not even necessitate a choice of a model, which can be formalized using a different utility structure.

Let a be an answer relating to the unknown of interest ω. The answer can for instance be a value of a future observation or an estimate of parameter common to all models in I. With the utility function $u(a,\omega)$, the expected utility of an answer a^* becomes

$$\bar{u}(a^*|\mathbf{x}) = \int u(a^*,\omega)p(\omega|\mathbf{x})d\omega \tag{12}$$

and this is indeed the optimal answer if

$$\bar{u}(a^*|\mathbf{x}) = \sup_a \bar{u}(a|\mathbf{x}) \tag{13}$$

Note that $p(\omega|\mathbf{x})$ still has the posterior weighted mixture form. This type of a strategy is often called Bayesian model averaging, and its sensibility for the problem at hand is dependent on whether the chosen utility function reflects the relevant issues (which ultimately need to be considered by the modeler).

An important question yet to be stated is that: What happens if all models in I are poor descriptions of $\mathbf{x}$**?** It should be clear from the above that we cannot use the formalism to directly detect this. Notice that the crucial question in such a situation is not the comparison of models, but the criticism of a model *without suggesting anything to replace it.* If one had, for instance, a more general model in mind than those included in I, the problem would be easily resolved by taking that model also into consideration and proceeding as before. Indeed, in many situations a generalization of the finite-dimensional parametric models $p(\mathbf{x}|\boldsymbol{\theta})$ could be obtained by considering models involving directly a probability measure on the space of distribution functions (these are typically called non-parametric models in the Bayesian framework), as was discussed in the previous section. Including such models to the class I and then performing the formal analysis, can be seen as one promising strategy to check formally the plausibility of the various parametric assumptions.

3 Asymptotic behavior of statistical model comparison: Part I

Here we investigate some central asymptotic concepts of frequentist and Bayesian statistical analysis in the context of model comparison.

The classical Neyman-Pearson theory for testing models requires pairwise processing of the elements of a model class $\mathcal{M}$, and therefore, let us concentrate for a moment on the situation where $\mathcal{M}$ contains only two models: M_1 and M_2. A generally accepted device for comparing models' appropriateness for a particular data set $\mathbf{x}$ is the likelihood ratio

$$\frac{p(\mathbf{x}|M_1)}{p(\mathbf{x}|M_2)}, \tag{14}$$

which is identical to the Bayes factor (Kass and Raftery, 1995) in the case of *completely specified models* (no parameters are estimated). Typically, however, models contain unknown parameters and the frequentist comparison procedure differs from the Bayes factor.

Using Neyman-Pearson theory we formulate the null hypothesis H_1 : the observations have arisen from the model M_1, and the alternative H_2 : the observations have arisen from the model M_2. To be able to formulate a regular likelihood ratio test (see *e.g.* Cox and Hinkley, 1974) of H_1 against H_2, assume the "nested hypothesis" case where M_2 is the full model and M_1 a reduced version of M_2 where some parameter(s) have been given fixed values. Let $d(\boldsymbol{\theta}_i)$ generally denote the number of unrestricted parameters in $M_i, i = 1, 2$.

The likelihood ratio test is formulated as: reject H_1 if

$$\lambda_n = \frac{L(\hat{\boldsymbol{\theta}}_1|\mathbf{x})}{L(\hat{\boldsymbol{\theta}}_2|\mathbf{x})} < c < 1 \tag{15}$$

where c is *a priori* specified threshold and $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimate of $\boldsymbol{\theta}_i, i = 1, 2$. We notice the difference with the Bayes factor where the uncertainty about parameters is accounted for by integrating them out with respect to the prior distribution, instead of maximization.

Under general regularity conditions on $L(\boldsymbol{\theta}_i|\mathbf{x})$ (*e.g.* $d(\boldsymbol{\theta}_i)$ remains fixed as $n \to \infty$), $-2\log\lambda_n$ is approximately chi-square distributed with $d(\boldsymbol{\theta}_2) - d(\boldsymbol{\theta}_1)$ degrees of freedom (denoted by $\chi^2_{d(\boldsymbol{\theta}_2)-d(\boldsymbol{\theta}_1)}$). As illustrated in Gelfand and Dey (1994), an inconsistency of this procedure is evident, since

$$\begin{aligned} \lim_{n\to\infty} \{P(\text{choose } M_2|M_1 \text{ true})\} &= \lim_{n\to\infty} \{P(\lambda_n < c|M_1 \text{ true})\} \qquad (16) \\ &= \lim_{n\to\infty} \{P(-2\log\lambda_n > -2\log c)\} \\ &= P(\chi^2_{d(\boldsymbol{\theta}_2)-d(\boldsymbol{\theta}_1)} > -2\log c) > 0 \end{aligned}$$

Thereby, even with unlimited amounts of data the procedure is not guaranteed to pick out the correct model. A more severe problem associated with the above testing scenario is that it provides no general yardstick for comparison of a range of different models. For instance, when the evidence against each of the models in $\mathcal{M}$ is measured by the p-value according to (15) where the unrestricted model M_2 is the most general model in $\mathcal{M}$, it follows that the p-value is a decreasing function of the number of restrictions imposed on $\boldsymbol{\theta}$. Thereby, the largest possible model is by definition associated with a p-value equal to unity, while the remaining models attain p-values smaller than or equal to unity depending on their degree of fit to data with respect to the full model. Generally, this framework makes especially the comparison of non-nested models difficult.

Hypothesis tests are designed to detect *any* discrepancies between a model and reality. Since models are virtually never exact descriptions of reality, we know by definition that for large enough samples the discrepancies will be detected by (15) and lead to a rejection of M_1 even if it is a good model for the purpose at hand. The point is that rejection of M_1 does not necessarily mean that M_2 offers a better description of the data, and hence, one should *compare* the two models instead of simply looking at the discrepancy between M_1 and the data. In this respect, a fundamental flaw of the hypothesis test scenario is that it cannot provide directly evidence *for* a model but only *against* it.

Even some of the advocates of the frequentist approach to statistical inference have clearly pointed out that such framework is unfortunate in the context of model selection and suggested that other approaches should preferably be followed (*e.g.* see Lindsey, 1996).

4 Asymptotic behavior of statistical model comparison: Part II

Consider the parametric case with a model labeled by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for an exchangeable sequence of observations. We then have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \\ &\propto \exp\{\log p(\boldsymbol{\theta}) + \log p(\mathbf{x}|\boldsymbol{\theta})\} \end{aligned} \tag{17}$$

Let $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_n$ denote the respective maxima of the two logarithmic terms in (17), *i.e.* the prior mode and the maximum likelihood estimate, respectively. These are determined by setting $\nabla \log p(\boldsymbol{\theta}) = 0$ and $\nabla \log p(\mathbf{x}|\boldsymbol{\theta}) = 0$, respectively. By expanding both logarithmic terms about their respective maxima we obtain

$$\begin{aligned} \log p(\boldsymbol{\theta}) &= \log p(\hat{\boldsymbol{\theta}}_0) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)' H(\hat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) + R_0 \\ \log p(\mathbf{x}|\boldsymbol{\theta}) &= \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' H(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + R_n \end{aligned} \tag{18}$$

where R_0, R_n denote remainder terms and

$$\begin{aligned} H(\hat{\boldsymbol{\theta}}_0) &= \left(-\frac{\partial^2 \log p(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0} \\ H(\hat{\boldsymbol{\theta}}_n) &= \left(-\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \end{aligned} \tag{19}$$

are the Hessian matrices. Under regularity conditions which ensure that the remainder terms R_0, R_n are small for large n, we get the result

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)' H(\hat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' H(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\right\} \tag{20}$$

The Hessian matrix $H(\hat{\boldsymbol{\theta}}_n)$ measures the local curvature of the log-likelihood function at it maximum $\hat{\boldsymbol{\theta}}_n$ and is typically called the *observed information matrix.* Further, by ignoring the prior terms (which are swamped by the data as n grows) we see that the posterior can be approximated by the multivariate normal distribution with mean $\hat{\boldsymbol{\theta}}_n$ and covariance matrix $\hat{\Sigma}_n = H(\hat{\boldsymbol{\theta}}_n)^{-1}$. However, asymptotics also reveal that

$$\lim_{n\to\infty}\left\{\frac{1}{n}\left(-\frac{\partial^2\log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{l=1}^{n}\left(-\frac{\partial^2\log p(x_l|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\} \tag{21}$$
$$= \int p(x|\boldsymbol{\theta})\left(-\frac{\partial^2\log p(x|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx$$

so that $H(\hat{\boldsymbol{\theta}}_n) \to n\mathbf{I}(\hat{\boldsymbol{\theta}}_n)$, where $\mathbf{I}(\boldsymbol{\theta})$ is (again) the *Fisher information matrix*, defined as

$$(\mathbf{I}(\boldsymbol{\theta}))_{ij} = \int p(x|\boldsymbol{\theta})\left(-\frac{\partial^2\log p(x|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx \tag{22}$$

The above results can be utilized in the model comparison framework through an approximation to the key quantity $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, the marginal likelihood. An important assumption concerning the validity of the asymptotic approximation is that the dimension $d(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ remains fixed as $n \to \infty$. Using the properties of the multivariate normal distribution (*i.e.* the form of its normalizing constant), an approximation to the marginal likelihood can be written as

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &\approx (2\pi)^{d(\hat{\boldsymbol{\theta}}_n)}|\hat{\Sigma}_n|^{1/2}p(\hat{\boldsymbol{\theta}}_0)p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) \end{aligned} \tag{23}$$

Using this approximation the posterior weights of the different models in $\mathcal{M}$ can be calculated. Under the assumption that the prior is continuous in $\boldsymbol{\Theta}$ and bounded at $\hat{\boldsymbol{\theta}}_0$, an approximate Bayes solution to the model comparison problem under the 0-1 loss function is to choose the model which maximizes

$$\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) + \frac{1}{2}\log|\hat{\Sigma}_n| + d(\hat{\boldsymbol{\theta}}_n)\log(2\pi) \tag{24}$$

This result is valid under a rather general setting (see Kim, 1998) and defines a consistent model selection procedure. However, a yet simpler and *still consistent* model comparison criterion is obtained, when terms not depending on n are ignored, and an asymptotic expansion of $\log|\hat{\Sigma}_n|$ is used. Under certain conditions (see Kim, 1998) the log-determinant can be written as

$$\log|\hat{\Sigma}_n| = -2\log\left(\prod_{l=1}^{d(\hat{\boldsymbol{\theta}}_n)} s_l(n)\right) + R_0 \tag{25}$$

where the remainder is bounded in n and the terms $s_l(n)$ are the *rates of convergence* of the maximum likelihood estimate $\hat{\theta}_{l(n)}$ to the true value of the (l)th component θ_l of $\boldsymbol{\theta}$. Under regular $\sqrt{n}$-convergence we are led to the criterion

$$\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \log\left(\prod_{l=1}^{d(\hat{\boldsymbol{\theta}}_n)} n^{1/2}\right) = \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \frac{d(\hat{\boldsymbol{\theta}}_n)}{2}\log n \tag{26}$$

This is precisely the widely-known criterion derived by Schwarz (1978), often called BIC or SBC (sometimes the above is multiplied by two). In the two model case, we can more concretely write

$$\log B_{12} \approx \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{1(n)}) - \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{2(n)}) - \frac{d(\hat{\boldsymbol{\theta}}_{1(n)}) - d(\hat{\boldsymbol{\theta}}_{2(n)})}{2} \log n \qquad (27)$$

Although (26) is a rather rough approximation, it can generally be considered as guideline for model comparison in a situation where the prior information is vague and difficult to specify precisely. Notice that also from (26) one can derive approximate posterior weights for the elements of $\mathcal{M}$. Generally, the criterion (26) has in various simulation studies shown to be conservative, such that for small n it may underestimate the true model dimension.

Since the introduction of the model comparison criterion AIC by Akaike (1974), a considerable interest has been attained in the statistical literature to criteria of the *penalized maximum likelihood* type

$$\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - c \cdot d(\hat{\boldsymbol{\theta}}_n) \log g(n) \qquad (28)$$

where different choices of c and $g(n)$ lead to different suggested criteria. For instance, $c = 1$ and $g(n) = e^1$ give rise to the AIC, $c = 1$ and $g(n) = \log n$ to the criterion of Hannan and Quinn (1979), and $c = 1/2$ and $g(n) = n$ to (26). It can be shown that for problems where $d(\hat{\boldsymbol{\theta}}_n)$ is *not increasing* with n, any choice of $g(n)$ equal to a constant, will lead to an inconsistent criterion. In particular, AIC not consistent, and it typically leads to a gross overestimation of the true dimension of $\boldsymbol{\theta}$ when n is large.

References

[1] Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Auto. Control,* **19**, 716-723.

[2] Cox, D. R. and Hinkley, D. V. (1974). Theoretical statistics. London: Chapman&Hall.

[3] Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc.*, **B 56**, 501-514.

[4] Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc.*, **B 41**, 190-195.

[5] Kass, R. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Stat. Assoc.* **90**, 773-795.

[6] Kim, J-Y. (1998). Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models. *Econometrica*, **66**, 359-380.

[7] Lindsey, J. (1996). Parametric statistical inference. Oxford: Oxford University Press.

[8] Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc.*, **B 49**, 223-239.

[9] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461-464.