Hacettepe Journal of Mathematics and Statistics
Volume 39 (1) (2010), 131 – 135

A MODEL SELECTION APPROACH IN STATISTICAL MODELING

Turhan Menteş*

Received 14 : 10 : 2009 : Accepted 12 : 12 : 2009

Abstract

It is argued that quantitative results from statistical surveys and experiments should be communicated as inferences of the model maximising the log Bayes factor against a reference model penalised by a subjectively chosen constant times the difference in model complexity. Model complexity is measured by the degrees of freedom. In this study, an efficient algorithm is proposed to select a model from among a large set of models with unit penalties in some interval. The algorithm utilizes the penalised log Bayes factor with only the likelihood ratio statistic, model dimensions and a constant. This approach seems to be a more realistic screening device than related criteria similar to the Bayesian information criterion.

Keywords: Model selection, Bayes factor, Penalty function, Utility function.

2000 AMS Classification: 62 C 10.

1. Introduction

A probabilistic model is a simplified description of the phenomenon under study, facilitating the reflection of parts of the reality associated with the phenomenon. Throughout this paper, a model is understood as a likelihood function or similar specification of model structure. A model should be seen as a device through which to view the world, as in Gelfand & Ghosh [2]. This view of the likelihood function is just a reformulation of the original reason for constructing the Bayesian predictive distribution from the prior and likelihood [8].

Space limitations necessitate that only one or just a few models be singled out for reporting results from a given scientific experiment or survey, in terms of parameter estimates and associated inferences. This paper deals with the decision-theoretic problem of choosing one or more models from a finite set of candidates.

*Hacettepe University, Department of Statistics, 06800 Beytepe, Ankara, Turkey. E-mail: mentes@hacettepe.edu.tr

Box's [1] dictum "all models are wrong, but some are useful" is taken as a premise here. However, some standard statistical methods for model selection do assume that a true model exists. Model selection methods yield sensible results most of the time because of lack of data. In practise, with enough data, there is a tendency for the larger models to be selected. These techniques are unsound and counterproductive. Model selection should be based on a criterion maximising a utility depending on more than just the amount of data.

The alternative to basing inference on one or just a few models is to employ a mixture of models. This approach reflects model uncertainty and improves predictive performance, see e.g. Kass & Raftery [7]. In the scientific literature, only the most important parameter estimates are given and correlations between parameter estimates are not reported, suggesting that accurate prediction is not a primary concern and that simplicity is at a premium.

In the Bayesian paradigm, likelihood and prior distributions are used together to update the person's belief through the Bayes formula. This is a completely objective entity, but the subjectively chosen priors play a role in the selection of the model.

The methodology developed for describing the relevant utilities is given in Section 2. The criteria for choosing the unit penalty is presented in Section 3. A particular approximation of the utilities used in automated model screening is elaborated in Section 4 of this paper.

2. Defining the utility

Subjective probability measures the degree of belief in a proposition [8]. Therefore it is the most natural measure of a model's compatibility with the data at hand, as it is the probability of the observed data conditional on the model considered.

The dimensions of the parameter space seems to be a natural measure of complexity of a fixed-effects model without smoothers. It is the measure which will be used for this model class in Section 4. Here the parameter sample spaces are assumed to be continuous. In recent years, measuring the dimension, especially of hierarchical and other random effects models, has received much attention. This effort is aimed at reconciling intuition with an objective measure of dimension, see for example Spiegelhalter *et al.* [9], Gelman, Carlin and Stern [4], Hodges and Sargent [6], and references therein. The suggestions are not conclusive, and how to count the degrees of freedom in various types of model is left to the modeler.

Let U_i be the utility of model M_i on data D. Let $f(D \mid M_i)$ denote the predictive density function of data D under model M_i, and let $C \geq 0$ denote the complexity penalty constant. Then we may define

$$U_1/U_0 = f(D \mid M_1)/f(D \mid M_0).$$

When M_0 and M_1 have the same dimension $m_0 = m_1$ this implies $U_i = f(D \mid M_i)h(Cm_i)$. Here it is assumed that $h(x)$ is continuous, strictly positive and strictly decreasing for $x \geq 0$. It is natural to have $h(0) = 1$ and $h''(x) > 0$, since the smaller the model the larger the consequence of a unit increase in dimensionality. For given C, $h(x)$ is convex so that M_1 with parameter space dimension $m_1 = m_0 + 1$ is preferred to M_0 depending on whether $f(D \mid M_1)/f(D \mid M_0) > \alpha_C > 1$ or not, regardless of the value of m_0. These properties are possessed by $h(x) = \exp(-x)$. The extreme variation in $f(D \mid M_i)$ with sample size and probability space is scaled away by looking at the equation

$$\log(U_i/U_0) = \log\{f(D \mid M_1)/f(D \mid M_0)\} - C(m_i - m_0).$$

This is a rank preserving mapping enabling the identification of the optimal model as a function of C. The quantity $\mathrm{LSU}\,(i, C)$ denotes the log scaled utility, and $\mathrm{LSU}\,(C)$ stands for $\max_i \mathrm{LSU}\,(i, C)$. For convenience, all models satisfy $m_0 \leq m_i$ in this development.

The model maximising the log scaled utility is selected for a given C, rather than averaging over models. Therefore, a prior distribution of models is not necessary, and there no notion of a "true" model conditional on the universe of models considered.

Having specified the proposed criteria for model selection as just support for the model in data, and the dimension of the parameter space, it is implicitly understood that the models being compared are equally plausible *a priori* in all other respects. Otherwise it would be necessary to group models into model strata and perform the comparisons within each group. Choosing between preferred models from different strata would then need to be based on other criteria.

3. Choosing the unit penalty

The log scaled utility $LSU(i, C)$ for model i is of the form $a(i) - Cb(i)$. It is assumed in this study that the models considered here are sufficiently regular to ensure that $a(i)$ is finite. By definition, $C \geq 0$ and $0 = b(0) \leq b(i) < \infty$. Assuming no ties among the $a(i)$'s, the minimal model maximising $LSU(C)$ is unique for any C. Any $LSU(i, C)$ equal to $LSU(C)$ can occur in only one interval. Defining

$$m\dim(C) = \min\{b(i) \mid \mathrm{LSU}\,(i, C) = \mathrm{LSU}\,(C)\}.$$

we see that $m\dim(C)$ is non-increasing and attains its minimum for

$$C \geq C_{\max} = \min\{C \mid m\dim(C) = 0\} < \infty.$$

Thus a plot of $m\dim(C)$ versus C marks each time a new model generates $m\dim(C)$. This plot provides a shorthand description of the model, and an overview of how the model selection depends on C.

The plot of $m\dim(C)$ versus C could be used to say something about the robustness of the inference, e.g. that only x out of y models maximised the utility, and that model z was the maximising function for a very broad range of C. Of course, the same could be communicated in a tabulation of the models generating $m\dim(C)$, and the left end point of the corresponding interval.

Basically, C should be chosen as the investigator sees fit. It is the minimum improvement in fit as measured by the log Bayes factor which is worth an extra parameter. Tentatively, C should be constrained to an interval defined by $\delta_L \leq m\dim(C) \leq \delta_U$. One often finds nothing of significance in small studies. Even documenting the negative result requires a model of some minimal dimensionality δ_L. In large studies an upper bound on the complexity of the models is important, and this is provided for by δ_U. If C is at or near the upper bound, it should be checked whether vastly different models eventually becomes available, or could be avoided by changing C. If this interval is too wide, it is necessary to find C by a thought experiment. For instance two hypothetical density functions $f_1(D, \theta_1)$ and $f_0(D, \theta_0)$ could be considered for the data D, with $\theta_1 = (\theta_0, \alpha)$, α being a one-dimensional parameter of a magnitude exactly large enough to bother about, and θ_0 is fixed. Then C can be chosen as the expected value of the log Bayes factor which (upon considering $f_1(D, \theta_1)$ as the true density) is the Kullback-Leibler distance between the two densities.

4. Priors for model screening

So far we have described how to choose a model, once the utility $U_i(C)$ for each model is known. Clearly, $U_i(C)$ depends on the i^{th} prior. There may be a logistic problem in eliciting and assigning a large number of high-dimensional priors to obtain the LSU (C) using approximations.

The proposed utilities are maximised by delta function priors placed at an MLE; i.e. the effect of the prior is to penalise the maximum available utility, $L(\psi)$. In the following it is assumed that all the models possess a MSE. The models are also sufficiently continuous to ensure that the likelihood function is a mapping of the parameter space, Θ, into an interval, i.e. $L(\Theta) = (0, L(\psi)]$. Under these circumstances any desired penalising is possible.

When considering a set of nested models, it is natural to expect the penalties caused by the choice of priors to increase monotonically with model dimension. This may not always be so in practise, but it is a good approximation. The simplest way to obtain this result is to have the following utility for the i^{th} model,

$$U_i = L_i(\psi_i)h(-\kappa m_i)\exp(-Cm_i).$$

Here m_i the dimension of the i^{th} model, h is a strictly positive and increasing function, and $\kappa > 0$.

As a side effect, this implies that for models i and j with $m_i \geq m_j$ and $L_i(\psi_i) < L_j(\psi_j)$, model j is preferred for any $C \geq 0$. Thus this procedure picks the model with the largest maximum likelihood for each model dimension, and may be viewed as the model with the largest potential. For reasons of symmetry the natural choice of h is exponential. Then the log scaled utility for model i is defined as follows:

$$LSU(i, C) = \log(\lambda_i) - \frac{1}{2}\pi[2\kappa + 2C],$$

where $\lambda_i = L_i(\psi_i)/L_0(\psi_0)$, p_i is the difference in model dimensions between models i and 0, and $\kappa > 0$.

When the set of candidate models forms a binary tree the function $m\dim(C)$ may be obtained by evaluating a small fraction of these models. This is done by sweeping the model universe in such a way that a sub tree is discarded if a hypothetical model combining the fit of the largest model in the sub tree with the complexity of the smallest model in the sub tree does not change the current estimate of the function $m\dim(C)$. Here the fit is the maximum log-likelihood.

5. Conclusion

The proposed approach fits into the modern tradition of explorative rather than confirmatory data analysis. The novelty is the flexibility in choosing the trade-off between compatibility with data and penalising model complexity. The deliberate aim of penalising model complexity in itself, rather than just as a quest for consistent model selection and/or good predictive performance sets it apart from other model selection criteria in the literature. Most model selection criteria in the literature utilize a log likelihood-ratio penalised by some function $Cf(p, n)$, where p is the difference in model dimensions, n the sample size, and C some fixed positive constant, see Gelfand & Day [3] for a review. These criteria are not meant to answer the important question "What is the most useful representation of the information embedded in the data at hand?". Only the study of Goutis and Robert [5], which provides the same order of flexibility in penalising complexity, is an alternative approach to the one proposed here.

In this study, model criticism is hardly mentioned. One may devise some initial model selection based on an evaluation of some particularly unsatisfactory model feature(s). However, this is rarely done. The main use for model criticism is as an aid in the creative process of inventing better models for consideration. Nonetheless, if some model criticism criterion could be formalised it could be entered into the model selection process proposed here, by screening and eventually dropping models which maximise $\mathrm{LSU}(C)$ for a given C.

References

[1] Box, G. E. P. *Robustness in the strategy of scientific model building*, In Launer, R. L. & Wilkinson, G. N. (eds.) *Robustness in Statistics* (Academic Press, New York, 1979).

[2] Gelfand, A. E. and Ghosh, S. K. *Model choice: A minimum posterior predictive loss approach*, Biometrika **85**, 1–11, 1998.

[3] Gelfand, A. E. and Dey, D. K. *Bayesian Model Choice: Asymptotics and Exact Calculations*, J. R. Statist. Soc. B **56**, 501-514, 1994.

[4] Gelman, A., Carlin, J. and Stern, A. *Bayesian Data Analysis* (Chapman & Hall, 2003).

[5] Goutis, C. and Robert, C. P. *Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections*, Biometrika **85**, 29–37, 1998.

[6] Hodges, J. S. and Sargent D. J. *Counting degrees of freedom in hierarchical and other richly-parameterised models*, Biometrika **88**, 367–379, 2001.

[7] Kass, R. E. and Raftery, A. E. *Bayes Factors*, J. Am. Statist. Assoc. **90**, 773–95, 1995.

[8] O'Hagan, A. *Kendall's Advanced Theory of Statistics. Volume 2B. Bayesian Inference* (Arnold, London, 1994)

[9] Spiegelhalter, D., Thomas, D., Best, N. and Carlin, B. P. *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*, J. R. Statist. Soc. B., 35–42, 1998.