ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 46 — 1997

Choice Among Hypotheses using Estimation Criteria

Constantinos GOUTIS, Christian P. ROBERT*

ABSTRACT. – A rule for choosing among nested models is presented, taking into account that the usual model selection procedure is a sequence of tests, followed by estimation of the parameters that remain in the model. We take a decision theoretical approach and formulate the loss functions and the rules from a Bayesian point of view. The rule resembles a Neyman-Pearson type test but it takes into consideration the estimation loss across different candidate models and reports an estimate together with an associated loss, taking into account the uncertainty in the selection. The method is compared with other existing procedures and illustrated by examples.

Choix de modèles à partir de critères d'estimation

RÉSUMÉ. – Nous développons une méthode de sélection de modèles pour des modèles emboités. L'apport de cette méthode est prendre en compte l'usage courant en sélection de modèle, qui est d'appliquer une suite de tests et d'effectuer ensuite l'estimation des paramètres restants. Nous reprenons cette démarche d'un point de vue décisionnel, en introduisant des fonctions de coût adéquates et en définissant les estimateurs de Bayes associés. La règle de décision est similaire à celle de Neyman-Pearson mais elle prend en compte les coûts respectifs des différents modèles et propose un estimateur en sus d'un modèle, en intégrant l'incertitude liée au choix du modèle. Nous comparons cette nouvelle méthode avec les techniques existantes et l'illustrons sur des exemples standards.

* C. GOUTIS; Ch. P. ROBERT: CREST-INSEE. Constantinos Goutis died tragically in a scubadiving accident on July 21, 1996, near Seattle. He was then 33 and a visiting professor at University Carlos III, Madrid. He will be remembered for his original contributions to Mathematical Statistics as well as for his indomitable personality.

1 Introduction

It is common among practitioners to implement a testing procedure as a means of reducing a complicated parametric model to a more parsimonious one, usually before undertaking the estimation step. The main goal is to reduce the number of parameters as much as possible without losing the essential features of the phenomenon under scrutiny. Usually one decides to include a parameter in the model or to set it equal to a fixed value on the basis of a significance test at an α level.

Once a satisfactory model is obtained, the remaining unknown parameters are estimated by some procedure. It is desirable that the parameter estimates are optimal in some sense, as for example unbiasedness, consistency or admissibility. A basic feature of the optimal estimators is that they are deemed to be close to a (real or conceptual) true value of the parameter. This property reflects the desire to have a model that describes reality sufficiently well to help us understand underlying structures of the data and, possibly, to be usable for some practical goals.

Typically choosing and fitting a model are viewed as different processes. On the one hand the form of the model depends on Neyman-Pearson type tests and on the other hand parameters are determined through estimation, implying a kind of discrepancy between the goals of the two procedures. This paper deals with this discrepancy and proposes a procedure of selecting a model based on the subsequent estimation of parameters and on the aim of the construction of the model, which takes into account both imperatives of parsimony and of good description of the analysed phenomenon.

Our procedure is decision theoretical in spirit and, more specifically, we take a Bayesian approach. Therefore, the optimality properties of the estimators we select encompass both steps of analysis instead of focusing on testing or estimating, thus reflecting more closely the overall goal of practitioners. Indeed, on the one hand, the separation between testing and estimation in usual model choice procedures forces the use of traditional ill-adapted loss functions, or the reference to standard acceptance levels which do not correspond to real life imperatives. On the other hand, practitioners always request model choice steps in the derivation of an acceptable model, both for parsimony and for practical reasons.

The paper is organised as follows: Section 2 examines critically the testing and estimation procedures, from a decision theoretical point of view. In Section 3 we propose a model selection rule which avoids the pitfalls of existing procedures and present some simple examples, whereas in Section 4 we generalise the rule to more complicated setups and examples. We conclude with a discussion.

2 Decision Theoretical Comparison

Usual testing starts with a pair of hypotheses of the form

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad vs. \quad H_1 : \boldsymbol{\theta} \in \Theta_1 \tag{1}$$

where θ is a possibly vector valued parameter, Θ_0 and Θ_1 are subsets of the parameter space. The goal is to decide between H_0 and H_1, on the basis of data $\mathbf{y}$ having a distribution $f(\mathbf{y}|\boldsymbol{\theta})$. We develop our arguments in the context of nested hypotheses but in Section 4 we discuss extensions to non-nested ones. We consider the case for which, possibly after reparameterisation, we have

$$\begin{array}{l} H_0 : (\theta_1, \theta_2, \ldots, \theta_q) \in \Re^q, \quad \theta_{q+1} = \theta_{q+2} = \ldots = \theta_p = 0 \\ H_1 : (\theta_1, \theta_2, \ldots, \theta_p) \in \Re^p, \end{array} \tag{2}$$

where $q < p$.

The classical approach to hypothesis testing produces a $0-1$ answer. This is an optimum solution to the decision problem of estimating $I(\boldsymbol{\theta} \in \Theta_0)$ under the familiar $0-1$, or more general $0-K_i$ loss function (see e.g. BERGER 1985, p. 355). Typically, in hypothesis tests one reports the decision (accept or reject), but practitioners prefer to report some highly data dependent measures of precision or accuracy of the procedure, such as a p-value or the posterior probability of the null hypothesis. The problem of a report of accuracy in testing, though important, has not been thoroughly studied [see KIEFER [1977], HWANG *et al.* [1992] or CASELLA and GOUTIS [1992] for some approaches and discussions).

Contrariwise, estimation procedures consider the parameter space as fixed. It is assumed that $\boldsymbol{\theta} \in \Re^p$ (or a subset of $\Re^p$) and the purpose is to determine a $\widehat{\boldsymbol{\theta}}$ as close as possible to the parameter value $\boldsymbol{\theta}$. A decision theoretical approach measures the distance between the values of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, or, more generally, the consequences of replacing $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$, by the squared error loss

$$L_2(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = ||\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}||^2, \tag{3}$$

or other measures of distance or utility functions. Once an estimate is derived, it is often reported together with an estimate of its precision. The use of the variance is implicit in (3), but other losses also allow for estimates of the attained loss or risk of the estimation rule. It seems that, for point estimation, the need of an accuracy report is more prevalent than for hypothesis testing.

In practice, a typical model selection procedure is a combination of hypothesis testing and point estimation. The testing component often consists of a sequence of tests, where a result determines which of the subsequent tests, if any, should be undertaken. In such cases the frequentist probabilities of error differ widely from the nominal α levels and a

calculation of the true rejection or acceptance probabilities involves some non-trivial conditioning. Furthermore, a report on the accuracy of estimators does not reflect the model uncertainty or the fact that they were derived after a stepwise model selection process. This has long been known, extensively studied (see, for example, DRAPER *et al.* [1971], POPE and WEBSTER [1972], JUDGE and BOCK [1978], RUUD [1984], BREIMAN [1992], CHATFIELD [1995]) and routinely ignored.

Similarly, a Bayesian approach to the problem requires that the prior is determined by the sequence of the tests. In particular, the prior must put some positive mass on a set Θ_0 which has (p-dimensional Lebesgue) measure equal to zero. This contradicts the very nature of the prior, since, theoretically the prior represents one's beliefs before data are collected and the beliefs should not depend on the inference of interest. Since the result depends critically on such "spike" masses, there is no consensus on a standard procedure in this case (see BERGER and DELAMPADY [1987], CASELLA and BERGER [1987] and BERGER and SELLKE [1987] and the discussions therein for thorough treatments in some simple cases). Replacing a point null hypothesis by an interval one seems reasonable but is not free of counter-intuitive behaviour (BASU [1992]). A sequence of "spike" masses embedded on subsets is even more counterintuitive as the prior distribution required for the subsets at each step is influenced by the results of earlier tests. A related difficulty in Bayesian testing is that diffuse priors do not yield any result in this setting. Indeed, the Bayes factor involves an arbitrary constant in the limit and the conclusions are indeterminate. The problem has been addressed by SMITH and SPIEGELHALTER [1980], SPIEGELHALTER and SMITH [1982], ATKINSON [1978] and WOLPERT [1995].

From a decision theoretical point of view, the underlying loss is far from obvious because for the actual sequence of inferential actions, the function of the parameter to be estimated (model label or parameter) and the loss change in a data dependent way. A simple estimation loss is unreasonable, since the best action would then be to retain the most complicated model. The loss $L_2(\theta, \widehat{\theta})$ at the final stage is not an appropriate report as it fails to account for the earlier steps of the sequence and the additional error resulting from the earlier decisions. On the other hand, it is not informative enough to know which parameters are non-zero, as we do in a testing only procedure. Hence, regular decision theoretical tools are not of much help, as they cannot incorporate the sequential nature in a single loss function. It is necessary to account for the vagueness of the information on the structure.

A different solution to the problem of estimation under alternative models is to avoid the problem of choice and consider averaging of all models (MOULTON [1991], OSIEWALSKI and STEEL [1993], RAFTERY [1993], RAFTERY *et al.* [1994], DRAPER [1995]). Although an attractive possibility in some cases, model averaging opposes the general scientific principle of parsimony. This may not be important for predictive purposes, but parsinomious models are indispensable if the goal is to understand the modelled phenomenon. Furthermore, such a method cannot be useful in, say, pilot studies, where various possibly irrelevant variables are included and the goal is to eliminate those that have little or not effect or are too costly to measure in the future.

It is more than rarely that the resulting mixed models are intractable for practitioners.

Estimation procedures can be modified towards parsimony so that the estimates may belong to the set Θ_0. This problem has been addressed via loss functions, by various authors, including LINDLEY [1968] and DICKEY [1974] who add "parsimony rewards" to an estimation loss. However, such losses combine heterogeneous quantities measured in non comparable scales. There is a vast literature on non-decision theoretical approaches, specially in a multiple regression setup (see MILLER [1990] and references therein). A popular approach based on information criteria and likelihood methods was introduced by AKAIKE [1974], who combined the likelihood with a "non-parsimony penalty". For other methods based on similar considerations see SAWA [1978], SCHWARZ [1978], GOUTIS and ROBERT [1994]. The main difference between all these methods and the approach in this paper is that we introduce explicitly comparisons between priors and decision theoretical loss functions across models.

Loss comparison is a straightforward method of choosing estimation rules in a single model but presents difficulties when performed across models. This is because each model is associated with a prior and the priors are supported in different sets. Typically, priors for the reduced models involve less uncertainty which in turn implies that the posterior expected loss will be smaller for reduced models. This is the Bayesian version of the well known phenomenon of greater accuracy of parameter estimates for simple models and prevents a direct comparison of posterior losses.

Though direct comparison does not yield sensible answers, a modification can produce a choice among models based directly on estimation criteria. In the following section we derive a procedure that addresses this issue, as well as earlier ones and effectively amalgamates hypothesis testing and parameter estimation. We first illustrate the concepts through an application to squared error loss and then describe extensions to some other cases.

3 A Model Selection Rule

Suppose that the null and alternative hypotheses are as in (2) and let π_1 be a possibly generalised prior supported on Θ_1. Suppose that the loss is $L_2(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ so the Bayes rule is the posterior mean $\widehat{\boldsymbol{\theta}}_1 = E_1(\boldsymbol{\theta}|\mathbf{y})$ whereas the posterior variance is $V_1 = E_1[(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_1)(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_1)^T|\mathbf{y}|$, where these expectations are taken with respect to the posterior corresponding to π_1. On the other hand, let π_0 be a possibly generalised prior supported on Θ_0 and let $E_0(\boldsymbol{\theta}|\mathbf{y})$ and $V_0 = E_0[(\boldsymbol{\theta} - E_0(\boldsymbol{\theta}|\mathbf{y}))(\boldsymbol{\theta} - E_0(\boldsymbol{\theta}|\mathbf{y}))^T|\mathbf{y}]$ be the mean and variance of the corresponding posterior distribution. Note that the expectations are taken over different sets so the vector $E_0(\boldsymbol{\theta}|\mathbf{y})$ is q-dimensional whereas $E_1(\boldsymbol{\theta}|\mathbf{y})$ is p-dimensional.

Under the most general model the Bayes rule is $\widehat{\boldsymbol{\theta}}_1$, whereas under the reduced model, the estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}_0 = (E_0(\boldsymbol{\theta}|\mathbf{y})^T, 0, 0, \ldots, 0)^T$. The standard decision theoretical Bayesian approach is to select the rule that minimises the posterior expected loss, hence choosing between models is equivalent to determining which procedure will produce a minimum loss. The *estimation* expected losses of $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}_1$ with respect to their corresponding priors are the sums of the variances of the parameter components, that is, $\mathrm{tr}(V_0)$ and $\mathrm{tr}(V_1)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

Typically a direct comparison yields $\mathrm{tr}(V_0) < \mathrm{tr}(V_1)$ and a way to avoid the bias towards the reduced model is to adjust the estimation loss to take into account the loss due to the change of the parameter values. That leads us to add, under H_0, some measure of error due to a wrong model selection. This additional loss term must reflect the effects of model choice. Since by choosing the reduced model we set the parameter estimate equal to $\widehat{\boldsymbol{\theta}}_0$ instead of $\widehat{\boldsymbol{\theta}}_1$ and we consider the quadratic loss appropriate, the model choice loss is measured by the distance of the two models in the quadratic metric, that is, the Euclidian distance of $\widehat{\boldsymbol{\theta}}_0$ from $\widehat{\boldsymbol{\theta}}_1$. This is really an additional error due to the choice of the "wrong" model. In fact, similarly to VARDEMAN's [1987] requirements, if H_0 is selected but H_1 is "true", it still makes sense to select H_0 if the resulting errors (the one due to estimation of $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}_0$ plus the one by replacing $\widehat{\boldsymbol{\theta}}_1$ by $\widehat{\boldsymbol{\theta}}_0$) is smaller that the error associated with the use of $\widehat{\boldsymbol{\theta}}_1$. Hence, a measure of the posterior expected loss associated with $\widehat{\boldsymbol{\theta}}_0$ is the sum of the pure estimation loss $\mathrm{tr}(V_0)$ and of the model selection loss $||\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1||^2$. A Bayesian rule for choosing between the hypotheses, and, consequently, parameter estimates, can be derived by comparing $\mathrm{tr}(V_0) + ||\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1||^2$ with $\mathrm{tr}(V_1)$ and accept the model with the smaller loss. Formally we have

$$\widehat{\boldsymbol{\theta}} = \begin{cases} \widehat{\boldsymbol{\theta}}_0 & \text{if} \quad \mathrm{tr}(V_0) + ||\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1||^2 \leq \mathrm{tr}(V_1) \\ \widehat{\boldsymbol{\theta}}_1 & \text{if} \quad \mathrm{tr}(V_0) + ||\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1||^2 > \mathrm{tr}(V_1). \end{cases} \tag{4}$$

In each case the reported loss should be the minimum attained loss, that is, if $\widehat{\boldsymbol{\theta}}_0$ is the estimate then the associated loss is $\mathrm{tr}(V_0) + ||\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1||^2$ whereas for $\widehat{\boldsymbol{\theta}}_1$ the loss is $\mathrm{tr}(V_1)$. If $\widehat{\boldsymbol{\theta}}_0$ is selected, we are indeed trying to prevent against the worst case, namely when H_1 is "true".

Loss comparison as in (4) can also be considered as an unbiasedness criterion for choosing between the priors π_0 and π_1, hence between the corresponding hypotheses: we have the alternatives to use $\widehat{\boldsymbol{\theta}}_0$ or $\widehat{\boldsymbol{\theta}}_1$, whether H_0 is acceptable or not. Therefore, we are making one of two errors when using $\widehat{\boldsymbol{\theta}}_1$: under H_1 the error is $\mathrm{tr}(V_1) = E_1[(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1)^T(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1)|y]$, whereas under H_0 the error is

$$E_0[(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1)^T(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1)|\mathbf{y}] = \frac{\int_{\Theta_1}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1)^T(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1) f(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Theta_1} f(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{5}$$

where in the above integral the measure of the prior π_0 is concentrated in Θ_0. The choice of the retained hypothesis is then dictated by the comparison of

errors. Actually, (5) is equal to $\mathrm{tr}(V_0)+\|\widehat{\boldsymbol{\theta}}_0-\widehat{\boldsymbol{\theta}}_1\|^2$, so both ways of thinking are equivalent. This may be thought as being specific to squared error loss, but in the next section we show a similar equivalence for other losses.

Whichever way we may think about the above procedure, it effectively works as a testing rule using an estimation loss. Furthermore, the report is a point estimate and an honest report of the obtained loss. Before discussing the procedure further, we illustrate its behavior for a simple case.

Example 1. Suppose that $Y_i|\theta_i \sim N(\theta_i, 1)$, $i = 1, 2, \ldots, p$, independently, the null hypothesis is $H_0 : \theta_1 = \theta_2 = \ldots = \theta_p$ and the alternative $H_1 : (\theta_1, \theta_2, \ldots, \theta_p) \in \Re^p$. Suppose that under the alternative, π_1 is a $N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ distribution, where $\boldsymbol{\mu}$ is a p-dimensional vector of prior means and $\mathbf{I}$ denotes the identity matrix. Let the prior under the null be the distribution of $\boldsymbol{\theta}$ under the alternative, conditionally on $\theta_1 - \theta_2 = \theta_2 - \theta_3 = \ldots = \theta_{p-1} - \theta_p = 0$. After some straightforward algebra, it follows that

$$\widehat{\boldsymbol{\theta}}_0 = ((\overline{\mu} + \sigma^2\overline{y})/(1+\sigma^2), \ldots, (\overline{\mu} + \sigma^2\overline{y})/(1+\sigma^2)) \tag{6}$$

and

$$\widehat{\boldsymbol{\theta}}_1 = ((\mu_1 + \sigma^2 y_1)/(1+\sigma^2), \ldots, (\mu_p + \sigma^2 y_p)/(1+\sigma^2)). \tag{7}$$

The estimates are given by

$$\widehat{\boldsymbol{\theta}} = \begin{cases} \widehat{\boldsymbol{\theta}}_0 & \text{if} \quad \displaystyle\sum_{i=1}^{p}(y_i-\overline{y})^2 + \frac{\displaystyle\sum_{i=1}^{p}(\mu_i-\overline{\mu})^2}{\sigma^4} + \frac{2\displaystyle\sum_{i=1}^{p}(y_i-\overline{y})(\mu_i-\overline{\mu})}{\sigma^2} \\ & \qquad\qquad \leq \left(p-\frac{1}{p}\right)\left(1+\frac{1}{\sigma^2}\right) \\ \widehat{\boldsymbol{\theta}}_1 & \text{if} \quad \displaystyle\sum_{i=1}^{p}(y_i-\overline{y})^2 + \frac{\displaystyle\sum_{i=1}^{p}(\mu_i-\overline{\mu})^2}{\sigma^4} + \frac{2\displaystyle\sum_{i=1}^{p}(y_i-\overline{y})(\mu_i-\overline{\mu})}{\sigma^2} \\ & \qquad\qquad > \left(p-\frac{1}{p}\right)\left(1+\frac{1}{\sigma^2}\right) \end{cases} \tag{8}$$

and the report of the loss is

$$\min\left\{\left\| \frac{\sigma^2(y_i-\overline{y}) - (\mu_i-\overline{\mu})}{\sigma^2+1} \right\|^2 + \frac{\sigma^2}{p(\sigma^2+1)}, \frac{p\sigma^2}{\sigma^2+1}\right\}. \tag{9}$$

From (8) one can see that the reduced model is preferred if the observations are close to each other, if the prior means are close to each other or if the correlation coefficient of the prior means and the observations is negative, indicating conflicting information about the truth of the null. The denominators at the left-hand sides of the inequalities of (8) scale the distances of the observations and the prior means by the variances in the usual way. The rule is well defined in the "noninformative" case, *i.e.* when

$\sigma^2 \to \infty$, and then the null hypothesis is preferred if $s^2 = \sum_{i=1}^{p}(y_i - \overline{y})^2$ is small. More precisely, equality of the means is accepted when $s^2 \leq p - \frac{1}{p}$. Note that, under the usual test, $s^2 \sim \chi^2_{p-1}$ and that the median of a χ^2_{p-1} distribution is between $p-1$ and $p-2$.

In the above example, the rule (4) resembles to Neyman-Pearson type tests, but the rationale is completely different. In particular, we do not fix a level of the test and we never consider probabilities of selecting the "correct model", as our losses are estimation losses. Hence, our focus is not on the question of estimating a model label, for which an approach via posterior probabilities or Bayes factors (e.g. ZELLNER [1984], KASS and RAFTERY [1995]) may be more appropriate. In that case, the loss is (either implicitly or explicitly) a $0 - K_i$ for $i = 1, 2$ loss rather than squared error.

A better look at the asymptotic behaviour of (4) might yield some further insight. Suppose that there exists a data generating process described by the density $f(\mathbf{y}|\boldsymbol{\theta})$ and the sample size goes to infinity. Then, as long as the prior π_1 gives positive mass to all open sets of Θ_1 and under suitable regularity conditions on the likelihood, the posterior mean $\widehat{\boldsymbol{\theta}}_1$ will be a consistent estimator of $\boldsymbol{\theta}$ and the posterior variance V_1 will tend to a zero matrix. We also suppose that the prior π_0 gives positive mass to all open sets of Θ_0 and note that, by definition, it does not give any mass to subsets of Θ_0^C. Hence, under suitable regularity conditions on the likelihood $p\lim \widehat{\boldsymbol{\theta}}_0 \in \Theta_0$ and V_0 tends to a zero matrix. The value $p\lim \widehat{\boldsymbol{\theta}}_0$ is some kind of *asymptotic pseudo true value* (see SAWA [1978], GOURIÉROUX *et al.* [1983], or FLORENS and MOUCHART [1989], [1993] for a Bayesian version) but it is derived as a limit of posterior means rather than as a limit of maximum likelihood estimates under a misspecified model. Under H_1, we have

$$(10) \qquad \|p\lim \widehat{\boldsymbol{\theta}}_0 - p\lim \widehat{\boldsymbol{\theta}}_1\|^2 = \|p\lim \widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}\|^2 > 0$$

and the estimator $\widehat{\boldsymbol{\theta}}$ given by (4) is consistent and asymptotically unbiased because, for sufficiently large sample size it is equal to $\widehat{\boldsymbol{\theta}}_1$. In that sense, if we want to view (4) as a testing procedure, the asymptotic power is equal to one. Under H_0,

$$(11) \qquad p\lim \widehat{\boldsymbol{\theta}}_0 = p\lim \widehat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}$$

so $\widehat{\boldsymbol{\theta}}$ is also consistent and asymptotically unbiased. Whether it is equal to $\widehat{\boldsymbol{\theta}}_0$ or to $\widehat{\boldsymbol{\theta}}_1$ depends, similarly to any other Bayesian method, on the prior distributions and the data but asymptotically it does not matter because of (11).

The form of (4), namely comparing the distance of the two estimators under the two hypothesis with their variances is similar to the specification test introduced by HAUSMAN [1978], which evaluates the asymptotic sampling distribution of $\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1$ under H_0. Specification tests later received a large attention in the econometric literature (see e.g. HOLLY [1982], HAUSMAN and TAYLOR [1982], HOLLY and MONFORT [1986], GOURIÉROUX and MONFORT [1989], or the review by RUUD [1984]). However, there is a major difference with our approach. We arrive at (4) by using

comparisons of Bayesian posterior expected losses, rather than through the sampling distribution of estimates of the parameter. Unlike Hausman type specification tests, where the asymptotic distributions of the test statistics involves second moments, the appearance of variances is incidental and follows from the default choice of $L_2(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$. However, squared error loss is just one of the possible choices, albeit a widely used one in general statistical contexts. In the next section we describe the application our method using other loss functions and we show that it is a straightforward generalisation of (4).

Furthermore, our method does not appeal at any asymptotic results and can be used as long as the ingredients of the decision problem (priors and loss functions) have been specified. Hausman's specification tests rely on the asymptotic distribution of parameter estimates and are typically applicable in conjunction with maximum likelihood (HOLLY [1982], NEWE [1985]) or other likelihood methods (HOLLY and MONFORT [1986]). In our approach the likelihood enters the problem only in the calculation of the posterior via Bayes theorem. As in all Bayesian methods, the sampling distribution of our parameter estimates is irrelevant because we consider the distribution of the parameters given the data rather than the distribution of the data given the parameters.

It is worth noting that $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}_1$ are defined as Bayes estimators in their own right. In that sense, they are distinct from the pseudo-true values or pseudo-maximum likelihood estimators (GOURIÉROUX *et al.* [1983], GOURIÉROUX and MONFORT [1993], FLORENS and MOUCHART [1989, 1993]). The latter arrive from considerations of a restricted model which is "close" in the Kullback-Leibler sense to another one. In all these methods, as well as in the Bayesian specification tests (FLORENS and MOUCHART [1989, 1993]) distances of models and "projections" of one model onto another play a crucial role. Our approach is based on decision theoretic loss comparisons of two Bayes estimators. There is however some vague connection, in that we consider the effects of replacing a model with another.

4 Generalisations and Examples

Though we developed our arguments in the context of the hypotheses (2), the method can be easily extended to non-nested hypotheses. In particular, it can be used when the parameter of interest has a similar meaning under both models so that it is sensible to use its (possibly zero) estimate under any model and the estimation loss under H_0 is smaller than under H_1. This typically follows from the fact that a more parsimonious use of the parameters raises the accuracy of the estimates, but is not restricted to hypotheses (2). Note that even in that case, we do not follow the encompassing principle (MIZON and RICHARD [1986], HENDRY and RICHARD [1990], FLORENS and MOUCHART [1993], GOURIÉROUX and MONFORT [1994]), in the sense that both models are considered separately.

The rule (4) can be easily modified to accommodate other losses. In particular, an easy modification solves the important problem of handling nuisance parameters of interest in different ways. The loss (3) can be restricted only to the norm of the vector of the parameters of interest. Extensions to losses other than squared error can be more relevant for model selection settings as a loss function can be derived from reasons underlying the selection purpose (e.g. prediction of future observations or deeper understanding of causality relationships). Then the appropriateness of a model should be judged according to how well it serves its purposes and different criteria may lead to different models (LINDLEY [1968]).

In a general setup, suppose that $\widehat{\boldsymbol{\theta}}$ estimates $\boldsymbol{\theta}$ and the incurred loss is $L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$, a real valued function of the parameter and the estimate. Let $\widehat{\boldsymbol{\theta}}_1$ be the value of $\widehat{\boldsymbol{\theta}}$ minimising

$$E_1(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})|\mathbf{y}) = \frac{\int_{\Theta_1} L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) f(\mathbf{y}|\boldsymbol{\theta}) \pi_1(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta_1} f(\mathbf{y}|\boldsymbol{\theta}) \pi_1(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{12}$$

and $\widehat{\boldsymbol{\theta}}_0$ be the value minimising

$$E_0(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})|\mathbf{y}) = \frac{\int_{\Theta_1} L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) f(\mathbf{y}|\boldsymbol{\theta}) \pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta_1} f(\mathbf{y}|\boldsymbol{\theta}) \pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{13}$$

where the first argument of the loss $L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ is constrained to $\theta_{q+1} = \theta_{q+2} = \ldots = \theta_p = 0$. Then the estimate of $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}} = \begin{cases} \widehat{\boldsymbol{\theta}}_0 & \text{if} \quad E_0(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_0)|\mathbf{y}) + L_e(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1) \leq E_1(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_1)|\mathbf{y}) \\ \widehat{\boldsymbol{\theta}}_1 & \text{if} \quad E_0(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_0)|\mathbf{y}) + L_e(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1) > E_1(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_1)|\mathbf{y}), \end{cases} \tag{14}$$

and the associated loss report is

$$\min\{E_0(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_0)|\mathbf{y}) + L_e(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1), E_1(L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_1)|\mathbf{y})\}. \tag{15}$$

The above relations are similar in spirit to the squared error loss case. One compares the posterior expected loss under the full model with the sum of the posterior expected loss of the reduced model and the model selection loss, measured by the loss incurred by using $\widehat{\boldsymbol{\theta}}_1$ if the parameter were $\widehat{\boldsymbol{\theta}}_0$. If $L_e(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1)$ is not symmetric in $\widehat{\boldsymbol{\theta}}_0$ $\widehat{\boldsymbol{\theta}}_1$, in (14) we would still use $L_e(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1)$ rather than $L_e(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_0)$, because H_1 is in some sense the *a priori* model that we want to reduce.

We illustrate the general procedure with two examples, one involving normal linear regression and the other a contingency table. Both cases are important, since they represent typical situations in which both a parsimonious model and parameter estimates are desired and the usual derivation goes through a sequence of tests and point estimations.

Example 2. Suppose that $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{I})$, where $\mathbf{X}$ is a known $n \times p$ matrix, β is a $p \times 1$ vector of parameters, and the hypotheses of interest are $H_0 : \beta \in \Re^q \times \{0, 0, \ldots, 0\}$ *vs.* $H_1 : \beta \in \Re^p$. If we want to understand

the underlying structure of the dependence relations then we would use the quadratic loss (3) and the resulting estimates (4). If, however, the goal is to control a future observation at some fixed value of the independent variables, say $\mathbf{x}_0$, then a reasonable loss is $(\mathbf{x}_0^T\beta - \mathbf{x}_0^T\widehat{\beta})^2$. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be the posterior means and V_0 and V_1 the posterior variances of β under π_0 and π_1 respectively. Here we take $\widehat{\beta}_0$ to be a $p \times 1$ vector but the last $p-q$ elements are equal to zero, whereas V_0 is the $q \times q$ variance-covariance matrix of the non-zero coordinates of β under the null hypothesis. Then if $\mathbf{x}_{00}$ is the vector consisting of the first q values of $\mathbf{x}_0$, rule (14) becomes

$$\widehat{\beta} = \begin{cases} \widehat{\beta}_0 & \text{if} \quad \mathbf{x}_{00}^T V_0 \mathbf{x}_{00} + \mathbf{x}_0^T(\widehat{\beta}_0 - \widehat{\beta}_1)(\widehat{\beta}_0 - \widehat{\beta}_1)^T\mathbf{x}_0 \leq \mathbf{x}_0^T V_1 \mathbf{x}_0 \\ \widehat{\beta}_1 & \text{if} \quad \mathbf{x}_{00}^T V_0 \mathbf{x}_{00} + \mathbf{x}_0^T(\widehat{\beta}_0 - \widehat{\beta}_1)(\widehat{\beta}_0 - \widehat{\beta}_1)^T\mathbf{x}_0 > \mathbf{x}_0^T V_1 \mathbf{x}_0. \end{cases} \tag{16}$$

If, instead, we use the model in order to predict future observations for unspecified $\mathbf{x}$, then we want to minimise $(\mathbf{x}^T\beta - \mathbf{x}^T\widehat{\beta})^2$ for all $\mathbf{x}$. Since this loss is quadratic in $\mathbf{x}$, performing well for all $\mathbf{x}$ simultaneously can be thought of as minimising the volume contained in the various quadratic forms in (16) (for a fixed length of $\mathbf{x}$), that is, minimising the product of the non-zero eigenvalues of the involved matrices. Hence, the rule becomes

$$\widehat{\beta} = \begin{cases} \widehat{\beta}_0 & \text{if} \quad \det(V_0) + (\widehat{\beta}_0 - \widehat{\beta}_1)^T(\widehat{\beta}_0 - \widehat{\beta}_1) \leq \det(V_1) \\ \widehat{\beta}_1 & \text{if} \quad \det(V_0) + (\widehat{\beta}_0 - \widehat{\beta}_1)^T(\widehat{\beta}_0 - \widehat{\beta}_1) > \det(V_1). \end{cases} \tag{17}$$

We study the behaviour of above rules in a numerical example, the data set given by HALD [1952] and reproduced in DRAPER and SMITH [1981]. The data consist of four explanatory and one response variables. The explanatory variables represent the amounts of four cement ingredients: tricalcium aluminate, tricalcium silicate, tetracalcium alumino ferrite and didalcium silicate, and the response variable is the heat evolved during hardening, in calories per gram of cement. The explanatory variables are measured as percentages of weight, so they are higly collinear. Apart from the collinearity arising from the fact that they add to (almost) one hundred, there are some strong correlations being between the first and the third and between the second and the fourth variables.

The various tested submodels have the form $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, where the design matrix $\mathbf{X}$ depends on the model, and we used *g-priors* (ZELLNER [1971, 1986]) for the mean and a noninformative prior for the variance (a nuisance parameter here), that is,

$$\beta|\sigma^2 \sim N(\boldsymbol{\mu}, \frac{\sigma^2}{n_0}(\mathbf{X}^T\mathbf{X})^{-1}) \tag{18}$$

$$\sigma^2 \sim \pi(\sigma^2)\alpha\frac{1}{\sigma^2}. \tag{19}$$

To avoid unnecessary collinearities, we centered all but the first column of $\mathbf{X}$ by substracting their respective means. The vector $\boldsymbol{\mu}$ was taken to

have elements equal to zero, except the first element which was taken, in a somewhat empirical Bayes way, to be equal to the average $\overline{\mathbf{Y}}$.

The prior for β represents knowledge from a similar (imaginary) experiment, with a relative precision of the two experiments equal to n_0, that is, if $n_0 = 1$, the prior information comes from an experiment of the same size as the one in hand. This is a sensible prior in this case since it takes into account the collinearity of the predictors and, furthermore, we have apparently an observational study. Of course an engineer or a chemist might have a better prior which should be used instead, but our point here is mainly to illustrate the method. For the priors (18), (19) one can easily compute the marginal posteriors of β which are t distributions with appropriate parameters.

Figures 1-4 show the estimates of each component of β (except the intercept) as a function of the prior precision, for the null hypothesis $H_0 : \beta_3 = \beta_4 = 0$. The estimates were computed using the three different rules: the "Difference" rule (4) where the structure of the model is of interest, the "Control" rule (16) to control the evolved heat at a fixed value of the composition and the "Prediction" rule (17) where we are interested in prediction in some arbitrary composition. The figures also show the Bayes estimates derived by the rule "Use the reduced model if the Bayes factor is more than $10^{-0.5}$" and the maximum likelihood estimates under the null and the alternative hypotheses (for these data, a significance test would not reject the null hypothesis). Note that the estimates from (4), (16), (17) do not change abruptly as the ones arising from the use of Bayes factor as a testing rule. Figure 5 gives the losses of the various rules as functions of n_0. The losses are rescaled so that direct comparisons can be made and the sums of the sampling variances of the least squares estimates are also shown on the graph. As the prior precision incrases, all Bayes losses increase, as a result of stronger prior information conflicting with the data.

Example 3. Consider $\mathbf{y} = (y_{ij})$, $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, representing frequencies arranged in a contingency table and let $\mathbf{p} = (p_{ij})$ be the corresponding cell probabilities. A plus ('+') in the subscript will denote summation over the respective values and we will consider $n = y_{++}$ fixed. We treat the common independence test and we use a natural estimation loss in this setup, the *entropy* loss or Kullback-Leibler distance:

$$L_{\mathcal{E}}(\mathbf{p}, \widehat{\mathbf{p}}) = n \sum_{ij} p_{ij} \log \frac{p_{ij}}{\widehat{p}_{ij}}. \tag{20}$$

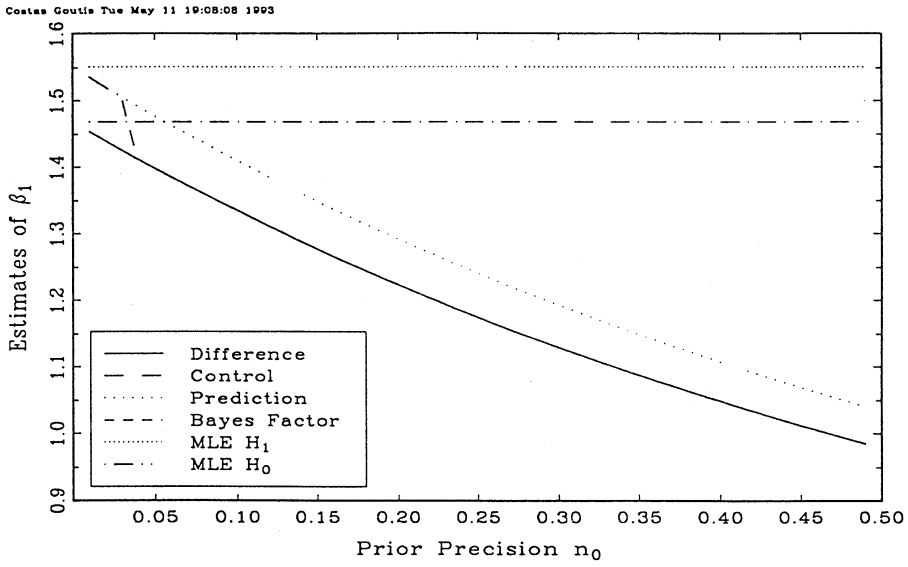
This loss gives a global measure of the distance of the density using the estimated parameters from the true density.

Under the dependency model and using a conjugate Dirichlet $\mathcal{D}(\alpha_{ij})$ prior, the posterior is also Dirichlet $\mathcal{D}(\alpha_{ij} + y_{ij})$, so the Bayes estimator $\widehat{\mathbf{p}}_1$, corresponding to $L_{\mathcal{E}}(\mathbf{p}, \widehat{\mathbf{p}})$, has components

$$\widehat{p}_{ij}^{\,1} = \frac{\alpha_{ij} + y_{ij}}{\alpha_{++} + n}, \tag{21}$$

FIGURE 1

Estimates of β_1 as functions of the prior precision n_0 for the hypothesis $\beta_3 = \beta_4 = 0$ under various rules. The "Difference" rule corresponds to (4), the "Control" rule to (16) and the "Prediction" rule to (17). A Bayes estimates using Bayes factor for model choice and the maximum likehood estimates under the null and the alternative hypotheses are also shown. The "Difference" and the "Bayes factor" rules are identical, whereas the "Control" rule overlaps with them for most values of n_0.



whereas the posterior expected loss is

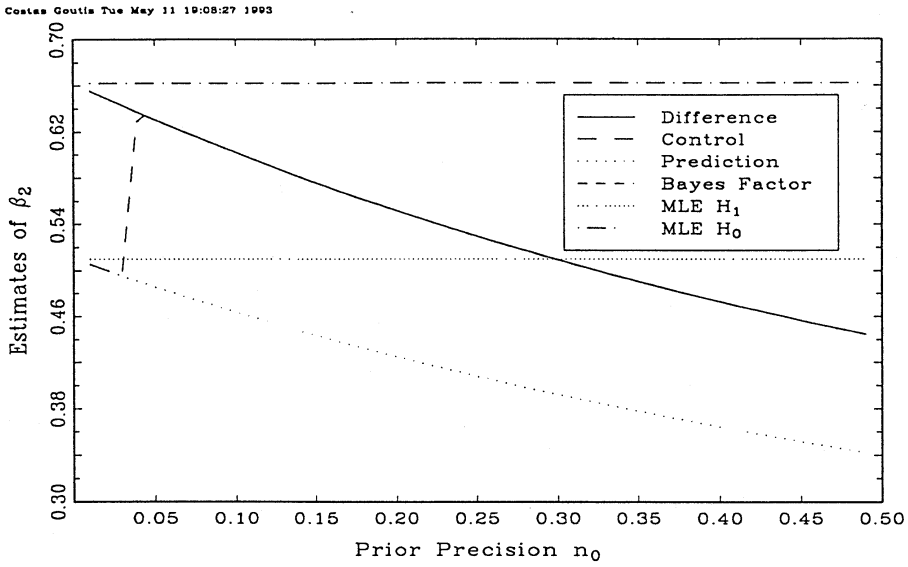
$$
\begin{aligned}
E_1[L_{\mathcal{E}}(\mathbf{p}, \widehat{\mathbf{p}}_1)|\mathbf{y}] &= n \sum_{ij} E_1(p_{ij} \log p_{ij} - p_{ij} \log \widehat{p}^1_{ij}|\mathbf{y}) \\
&= n \sum_{ij} E_1(p_{ij} \log p_{ij}|\mathbf{y}) - \widehat{p}^1_{ij} \log \widehat{p}^1_{ij}.
\end{aligned} \tag{22}
$$

Under independence, p_{ij} can be written as $p_{ij} = p_{i+}p_{+j}$ and we take the projected $\mathcal{D}(\alpha_{i+})$ and $\mathcal{D}(\alpha_{+j})$ priors as marginal priors on the probabilities p_{i+} and p_{+j} respectively. Obviously, this is by no means the only way to express independence (see e.g. ALBERT and GUPTA [1982] or GOUTIS [1993]). The restricted Bayes estimator $\widehat{\mathbf{p}}_0$ has components

$$
\widehat{p}^0_{ij} = \frac{(\alpha_{i+} + y_{i+})(\alpha_{+j} + y_{+j})}{(\alpha_{++} + n)^2} \tag{23}
$$

FIGURE 2

Estimates of β_2 for the same hypothesis and presented the same way as in Figure 1. The "Difference" and the "Bayes factor" rules are identical, whereas the "Control" rule overlaps with them for most values of n_0.



and some algebra shows

(24)
$$E_0[L_{\mathcal{E}}(\mathbf{p},\widehat{\mathbf{p}}_0)|\mathbf{y}] = n\sum_{ij} E_0[p_i + p_{+j}\log(p_{i+}p_{+j}) - p_{i+}p_{+j}\log\widehat{p}^0_{ij}|\mathbf{y}]$$
$$= n\{\sum_i E_0[p_{i+}\log p_{i+}|\mathbf{y}) - \widehat{p}^0_{i+}\log\widehat{p}^0_{i+}$$
$$+ \sum_j E_0(p_{+j}\log p_{+j}|\mathbf{y}) - \widehat{p}^0_{+j}\log\widehat{p}^0_{+j}\}$$

and

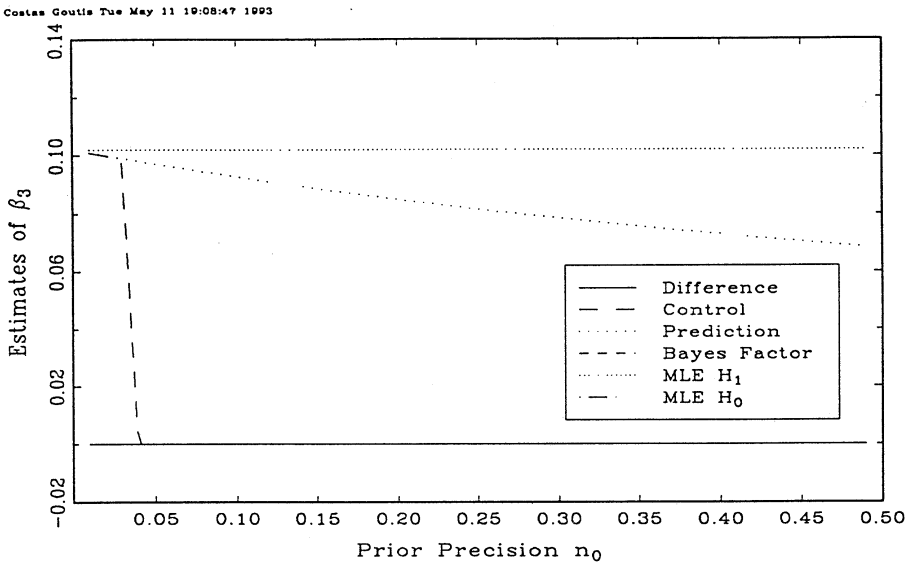
(25) $$L_{\mathcal{E}}(\widehat{\mathbf{p}}_0,\widehat{\mathbf{p}}_1) = n\{\sum_i \widehat{p}^0_{i+}\log\widehat{p}^0_{i+} + \sum_j \widehat{p}^0_{+j}\log\widehat{p}^0_{+j} - \sum_{ij}\widehat{p}^0_{ij}\log\widehat{p}^1_{ij}\}.$$

In order to find the rule (14), we first derive $E_1(p_{ij}\log p_{ij}|\mathbf{y})$. Since, under $\pi_1(\mathbf{p}|\mathbf{y})$, the cell probabilities p_{ij} have marginally a $\mathcal{B}e(\alpha_{ij}+y_{ij},\ \alpha_{++}+n-\alpha_{ij}-y_{ij})$ distribution, we have

(26) $$E_1(p_{ij}\log p_{ij}|\mathbf{y}) = \frac{\int_0^1 t^{\alpha_{ij}+y_{ij}}(1-t)^{\alpha_{++}+n-\alpha_{ij}-y_{ij}-1}\log(t)dt}{B(\alpha_{ij}+y_{ij},\ \alpha_{++}+n-\alpha_{ij}-y_{ij})}$$
$$= -\frac{\alpha_{ij}+y_{ij}}{\alpha_{++}+n}\sum_{m=0}^{\alpha_{++}+n-\alpha_{ij}+y_{ij}-1}\frac{1}{\alpha_{ij}+y_{ij}+1+m}.$$

FIGURE 3

Estimates of β_3 for the same hypothesis and presented the same way as in Figure 1. The MLE under H_0, the "Difference" and the "Bayes factor" rules are identical, whereas the "Control" rule overlaps with them for most values of n_0.



Using similar expressions for $E_0(p_{i+}\log p_{i+}|\mathbf{y})$ and $E_0(p_{+j}\log p_{+j}|\mathbf{y})$, the reduced model is preferred if

$$(27)\qquad -n\Bigg\{\frac{1}{\alpha_{++}+n}\sum_i\sum_{m=0}^{\alpha_{++}+n-\alpha_{i+}-y_{i+}-1}\frac{\alpha_{i+}+y_{i+}}{\alpha_{i+}+y_{i+}+1+m} + \frac{1}{\alpha_{++}+n}\sum_{m=0}^{\alpha_{++}+n-\alpha_{+j}-y_{+j}-1}\frac{\alpha_{+j}+y_{+j}}{\alpha_{+j}+y_{+j}+1+m}+\sum_{ij}\widehat{p}^{\,0}_{ij}\log\widehat{p}^{\,1}_{ij}\Bigg\}$$

is smaller than

$$(28)\qquad -n\Bigg\{\frac{1}{\alpha_{++}+n}\sum_{ij}\sum_{m=0}^{\alpha_{++}+n-\alpha_{ij}-y_{ij}-1}\frac{\alpha_{ij}+y_{ij}}{\alpha_{ij}+y_{ij}+1+m}+\sum_{ij}\widehat{p}^{\,1}_{ij}\log\widehat{p}^{\,1}_{ij}\Bigg\}$$

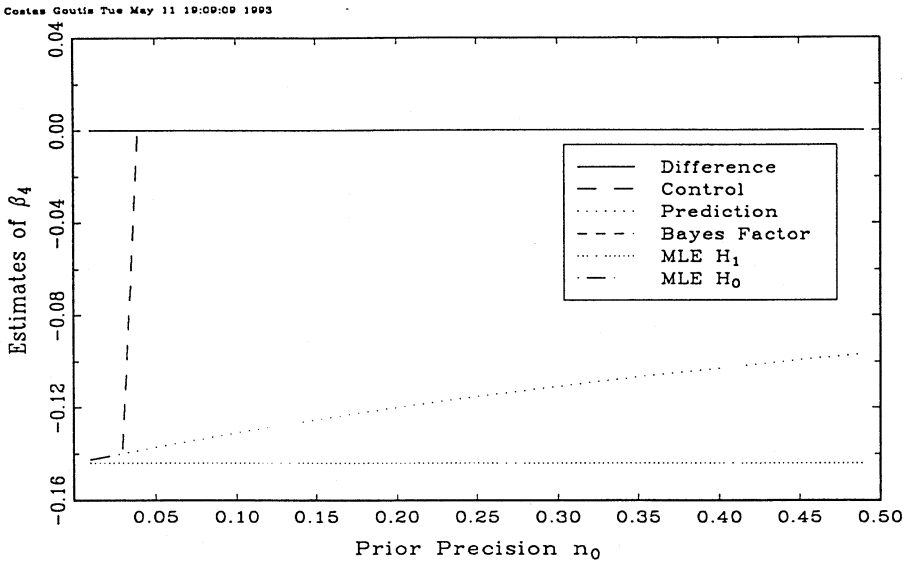
whereas the minimum of (27) and (28) is the reported loss.

It is interesting to note that in the above example expression (27) is equal to the posterior expected loss of $\widehat{\mathbf{p}}_1$ under the posterior corresponding to the prior π_0, that is,

$$(29)\qquad E_0[L_{\mathcal{E}}(\mathbf{p},\widehat{\mathbf{p}}_0)|\mathbf{y})+L_{\mathcal{E}}(\widehat{\mathbf{p}}_0,\widehat{\mathbf{p}}_1)=E_0[L_{\mathcal{E}}(\mathbf{p},\widehat{\mathbf{p}}_1)|\mathbf{y}].$$

FIGURE 4

Estimates of β_4 for the same hypothesis and presented the same way as in Figure 1. The MLE under H_0, the "Difference" and the "Bayes factor" rules are identical, whereas the "Control" rule overlaps with them for most values of n_0.



As mentioned in Section 3, this equality also holds for the squared error loss, giving an additional interpretation to the selection criterion as an unbiasedness criterion between hypotheses. It turns out that (29) holds for the entropy loss under general conditions, namely, if the sampling density $f(y|\theta)$ belongs to the exponential family, that is, if

$$f(\mathbf{y}|\theta) = h(\mathbf{y})\exp\{\boldsymbol{\theta}\cdot\mathbf{y} - \psi(\boldsymbol{\theta})\}. \tag{30}$$

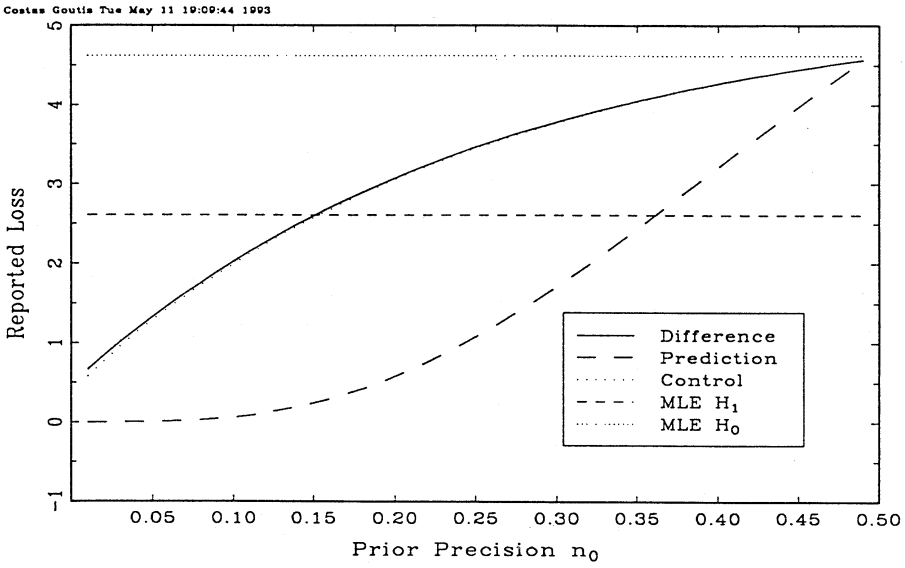
Then, using the results summarised in BROWN [1986, Chapter 6],

(31)

$$\begin{aligned}
&E_0(L_{\mathcal{E}}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_1)|\mathbf{y}) \\
&\quad= \int \pi_0(\boldsymbol{\theta}|\mathbf{y}) \int f(\mathbf{t}|\boldsymbol{\theta}) \log \frac{f(\mathbf{t}|\boldsymbol{\theta})}{f(\mathbf{t}|\widehat{\boldsymbol{\theta}}_1)}\, d\mathbf{t}\, d\boldsymbol{\theta} \\
&\quad= \int \pi_0(\boldsymbol{\theta}|\mathbf{y}) \Bigg\{ \int f(\mathbf{t}|\boldsymbol{\theta}) \log \frac{f(\mathbf{t}|\boldsymbol{\theta})}{f(\mathbf{t}|\widehat{\boldsymbol{\theta}}_1)}\, dt \\
&\qquad - \int f(\mathbf{t}|\boldsymbol{\theta}) \log \frac{f(\mathbf{t}|\boldsymbol{\theta})}{f(\mathbf{t}|\widehat{\boldsymbol{\theta}}_0)}\, d\mathbf{t} \Bigg\} d\theta + E_0(L_{\mathcal{E}}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_0)|\mathbf{y}) \\
&\quad= \int \pi_0(\boldsymbol{\theta}|\mathbf{y})\{(\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1)\cdot\nabla\psi(\boldsymbol{\theta}) - (\psi(\widehat{\boldsymbol{\theta}}_0) - \psi(\widehat{\boldsymbol{\theta}}_1))\} d\boldsymbol{\theta} + E_0(L_{\mathcal{E}}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_0)|\mathbf{y}) \\
&\quad= (\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_1)\cdot \int \pi_0(\boldsymbol{\theta}|\mathbf{y})\nabla\psi(\boldsymbol{\theta}) d\boldsymbol{\theta} - (\psi(\widehat{\boldsymbol{\theta}}_0)) + E_0(L_{\mathcal{E}}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_0)|\mathbf{y}).
\end{aligned}$$

FIGURE 5

Reported losses of the various rules as functions of n_0. The "Prediction" and "Control" reported losses have been rescaled. The sums of the sampling variances of the least squares estimates are also shown.



The Bayes rule under entropy loss minimises

$$\int \pi_0(\boldsymbol{\theta}|\mathbf{y})\{(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\nabla\psi(\boldsymbol{\theta})-(\psi(\boldsymbol{\theta})-\psi(\widehat{\boldsymbol{\theta}}))\}d\boldsymbol{\theta} \tag{32}$$

hence

$$\int \pi_0(\boldsymbol{\theta}|\mathbf{y})\nabla\psi(\boldsymbol{\theta})d\boldsymbol{\theta}=\nabla\psi(\widehat{\boldsymbol{\theta}}_0) \tag{33}$$

and substituting in (31) we obtain the result.

One can extend the model selection procedure to more than two candidate models. We will describe it in the case of three nested hypotheses, further extensions being straightforward.

Suppose that, instead of (2), we have

$$\begin{aligned}
&H_0:(\theta_1,\theta_2,\ldots,\theta_r)\in\Re^r,\quad \theta_{r+1}=\theta_{r+2}=\ldots\theta_q=\theta_{q+1}=\ldots=\theta_p=0,\\
&H_1:(\theta_1,\theta_2,\ldots,\theta_r)\in\Re^q,\quad \theta_{q+1}=\theta_{q+2}=\ldots\theta_p=0,\\
&H_2:(\theta_1,\theta_2,\ldots,\theta_r)\in\Re^p,
\end{aligned} \tag{34}$$

where $r<q<p$. The choice between any pair of hypotheses is similar to the earlier case but the reported losses differ. If one compares H_1 and

H_2 and chosses H_1 the estimate is $\widehat{\boldsymbol{\theta}}_1$ and the estimation loss is (with an obvious notation)

$$E_1[L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_1)|\mathbf{y}] + L_e(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2). \tag{35}$$

If subsequently one compares H_0 to H_1 the estimating rule is the same as if H_0 and H_1 were directly compared, but, if ones accepts H_1 the loss is (35) whereas otherwise, it is

$$E_0[L_e(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_0)|\mathbf{y}] + L_e(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1) + L_e(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2). \tag{36}$$

Hence, if one chooses the parameter estimate associated with the simplest hypothesis H_0 after two tests, the reported loss is always larger than after one test. This reflects, in some sense, the fact that, if a parsimonious model is selected after having tried several models and reducing them, a report of a loss associated with the parameter estimates only is overoptimistic, and the reported loss should depend on the number of the models that were tried in the process.

5 Discussion

There are several problems, described in Section 2, that the above procedure is addressing. The first one is to avoid any artificial mixing of the null and alternative hypotheses with the difficulty of getting the "proper" weights of a modified prior distribution. In order to implement the rule, we may need prior distributions for the null and the alternative hypothesis but not a "spike" on the set Θ_0. This spike seems to be a serious problem with no satisfactory answer (see VARDERMAN [1987], ROBERT and CASELLA [1994] or ROBERT and CARON [1996] for discussions). Here the probability of the null hypothesis has no meaning, since a null hypothesis is chosen not because it is more probable but because the restricted estimates are better in some precise loss sense. Furthermore, unlike standard Bayesian tests of hypotheses, the rule is determined even if $\pi_0(\boldsymbol{\theta})$ and $\pi_1(\boldsymbol{\theta})$ are improper, as long as the posterior expected loss is finite. Although the use of non-informative or improper priors is an object of debate, it is always useful to be able to extend the Bayesian paradigm. One can use as $\pi_0(\boldsymbol{\theta})$ the distribution derived from $\pi_1(\boldsymbol{\theta})$ by conditioning on the event $\boldsymbol{\theta} \in \Theta_0$ (as it was done in Example 1) but the derived distribution is not parametrisation invariant.

It is worthwhile to note that the rule derived by the use of squared error loss is invariant to orthogonal transformations of the parameters, since both the traces of the variances and the distance are rotation invariant. Under the entropy loss there is a stronger invariance, namely the procedure does not depend on the parametrisation of the problem (see also GUTTIÉRREZ-PEÑA [1992]).

This is very appealing when one is willing to use a global measure of distance of distribution as a loss function. Of course, in general, invariance does not hold since the rule is essentially an estimation rule and, except when derived by maximum likelihood methods, estimates are typically not invariant to reparameterisation.

However, our main effort was to amalgamate testing and estimation procedures so that the resulting methodology takes into account what is actually done in practice. Alghough there is a vast amount of research on the effects of testing before estimation and on "testimators", it is rarely from a decision theoretical point of view. Decision theory has traditionally viewed testing and estimation as two separate problems, requiring different solutions, but we consider this separation rather misleading.

Furthermore, we avoid the artificiality of the α level tests and replace the whole problem in its true perspective, that is to produce models approximating reality as well as possible. A way to judge the adequacy of the models is by estimating the errors resulting from the choice of each model. This is exactly what the criterion is doing. After all, forcing a zero-one answer to the testing problem and using this answer to build an estimation procedure which is by nature imprecise, does not seem sensible. The imprecision of the answer must be taken into account from the very beginning of the process.

• References

AKAIKE, H. (1974). – "A New Look at the Statistical Model Identification", *IEEE Trans. Auto. Control.*, 19, pp. 716-723.

ALBERT, J. H., GUPTA, A. K. (1982). – "Mixtures of Dirichlet Distributions and Estimation in Contingency Tables", *Ann. Statist.*, 10, pp. 1261-1268.

ATKINSON, A. C. (1978). – "Posterior Probabilities for Choosing a Regression Model", *Biometrika*, 65, pp. 39-48.

BASU, S. (1992). – "A New Look at Bayesian Point Null Hypothesis Testing: HPD Sets, Volume Minimising Sets and Robust Bayes", *Technical Report, Department of Statistics and Applied Probability, University of California,* Santa Barbara.

BERGER, J. O. (1985). – *Statistical Decision Theory and Bayesian Analysis*, Second Edition. Springer-Verlag, New York.

BERGER, J. O., DELAMPADY, M. (1987). – Testing Precise Hypotheses (with discussion)", *Statist. Sci.*, 2, pp. 317-352.

BERGER, J. O., SELLKE, T. (1987). – "Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence (with discussion)", *J. Amer. Statist. Assoc.*, 82, pp. 112-122.

BREIMAN, L. (1992). – "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: the X-Fixed Prediction Error", *J. Amer. Statist. Assoc.*, 87, p. 738-754.

BROWN, L. D. (1986). – *Fundamentals of Statistical Exponential Families*. IMS Monograph Series, Institute of Mathematical Statistics, Hayward, CA.

CASELLA, G., BERGER, R. L. (1987). – "Reconciling Evidence in the One-Sided Testing Problem (with discussion)", *J. Amer. Statis. Assoc.*, 82, pp. 106-111.

CASELLA, G., GOUTIS, C. (1992). – "Relationships between Post-Data Accuracy Measures", *Biometrics Unit Technical Report BU-1047-M*, Cornell University.

CHATFIELD, C. (1995). – "Model Uncertainty, Data Mining and Statistical Inference (with discussion)", *J. R. Statist. Soc. Ser. A*, 158.

DICKEY, J. (1974). – Bayesian Alternatives to the F-Test and Least Squares Estimate in the Normal Linear Model, In: *Studies in Bayesian Econometrics and Statistics* (S. E. Fienberg and A. Zellner, eds.), North-Holland, Amsterdam.

DRAPER, D. (1995). – "Assessment and Propagation of Model Uncertainty (with discussion)", *J. R. Statist. Soc. Ser. B*, 57, pp. 45-97.

DRAPER, N. R., GUTTMAN, I., KANEMASU, H. (1971). – "The Distribution of Certain Regression Statistics", *Biometrika*, 58, pp. 295-298.

DRAPER, N. R., SMITH, H. (1981). – *Applied Regression Analysis*, 2nd Edition. John Wiley & Sons, New York.

FLORENS, J. P., MOUCHART, M. (1989). – "Bayesian Specification Tests", In: *Contributions to Operations Research and Economics* (B. CORNET and H. TULKENS eds.), *MIT Press*, Cambridge MA.

FLORENS, J. P., MOUCHART, M. (1993). – "Bayesian Testing and Testing Bayesians", In: *Handbook of Statistics*, vol. 11, (MADDALA G. S., RAO C. R. and VINOD H. D. eds.), Elsevier Science Publishers, Amsterdam.

GOURIÉROUX, C., MONFORT, A. (1989). – *Statistique et Modèles Econométriques*, 2 volumes, Economica, Paris.

GOURIÉROUX, C., MONFORT, A. (1993). – "Pseudo-Likelihood Methods", In: *Handbook of Statistics*, vol. 11, (MADDALA G. S., RAO C. R., VINOD H. D. eds.), Elsevier Science Publishers, Amsterdam.

GOURIÉROUX, C., MONFORT, A. (1994). – "Testing Non-Nested or Non-Nested Hypotheses", *J. Econometrics*, IV, (ENGLE R. F., MCFADDEN D. L. eds.), Elsevier Science Publishers, Amsterdam.

GOURIÉROUX, C., MONFORT, A., TROGNON, A. (1983). – "Testing Nested or Non-Nested Hypotheses", *J. Econometrics*, 21, pp. 83-115.

GOUTIS, C. (1993). – "Bayesian Estimation Methods for Contingency Tables", *J. Ital. Statist. Soc.*, 2, pp. 35-54.

GOUTIS, C., ROBERT, C. P. (1994). – "Model Choice in Generalised Linear Models: a Bayesian Approach via Kullback-Leibler Projections", *Research report No. 137, Department of Statistical Science*, University College London.

GUTIÉREZ-PEÑA, E. (1992). – "Expected Logarithmic Divergence for Exponential Families", In: *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith, eds.), *Oxford University Press*, New York.

HALD, A. (1952). – *Statistical Theory and Engineering Applications*, John Wiley & Sons, New York.

HAUSMAN, J. A. (1978). – "Specification Tests in Econometrics", *Econometrica*, 46, pp. 1251-1271.

HAUSMAN, J. A., TAYLOR, W. E. (1982). – "A Generalized Specification Test", *Economics Letters*, 8, pp. 239-245.

HENDRY, D. F., RICHARD, J. P. (1989). – "Recent Developments in the Theory of Encompassing", In: *Contributions to Operations Research and Economics*, (B. Cornet and H. Tulkens, eds.), *MIT Press*, Cambridge MA.

HOLLY, A. (1982). – "A Remark on Hausman's Specification Test", *Econometrica*, 50, pp. 749-759.

HOLLY, A, MONFORT, A. (1986). – "Some Useful Equivalence Properties of Hausman's Test", *Economics Letters*, 20, pp. 39-43.

HWANG, J. T., CASELLA, G., ROBERT, C., WELLS, M. T., FARRELL, R. H. (1992). – "Estimation of Accuracy in Testing", *Ann. Statis.*, 20, pp. 490-509.

JUDGE, G. G., BOCK, M. E. (1978). – *The Statistical Implications of Pre-Test and Stein-Rule Estimation in Econometrics*, North-Holland, Amsterdam.

KASS, R. E., RAFTERY, A. E. (1995). – "Bayes Factors and Model Uncertainty", *J. Amer. Statist. Assoc.*, 90, pp. 773-795.

KIEFER, J. (1977). – "Conditional Confidence Statements and Confidence Estimators (with discussion)", *J. Amer. Statist. Assoc.*, 72, pp. 789-827.

LINDLEY, D. V. (1968). – "The Choice of Variables in Multiple Regression (with discussion)", *J. R. Statist. Soc. Ser. B*, 30, pp. 31-66.

MILLER, A. J. (1990). – *Subset Selection in Regression*, Chapman & Hall, London.

MIZON, G. E., RICHARD, J. P. (1986). – "The Encompassing Principle and its Application to Testing Non-Nested Hypotheses", *Econometrica*, 54, pp. 657-678.

MOULTON, B. R. (1991). – "A Bayesian Approach to Regression Selection and Estimation with Application to a Price Index for Ratio Services", *J. Econometrics*, 49, pp. 169-193.

NEWEY, W. K. (1985). – "Maximum Likehood Specification Testing and Conditional Moment Tests", *Econometrica*, 53, pp. 1047-1070.

OSIEWALSKI, J., STEEL, M. F. J. (1993). – "Regression Models under Competing Covariance Structures: A Bayesian Perspective", *Annales d'Économie et de Statistique*, 32, pp. 65-79.

POPE, P. T., WEBSTER, J. T. (1972). – "The Use of an F-Statistic in Stepwise Regression Procedures", *Technometrics*, 14, pp. 327-340.

RAFTERY, A. E. (1993). – "Bayesian Model Selection in Structural Equation Models", In: *Testing Structural Equation Models*, (K. A. Bollen and J. S. Long eds.), Sage, Beverly Hills.

RAFTERY, A. E., MADIGAN, D., VOLINSKY, C. T. (1994). – "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance", In *Bayesian Statistics* 5, (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith eds.).

ROBERT, C. P., CARON, N. (1996). – "Noninformative Bayesian Testing and Neutral Bayes Factors", *Test*, 5, pp. 411-437.

ROBERT, C. P., CASELLA, G. (1994). – "Distance Weighted Losses for Testing and Confidence Set Evaluation", *Test*, 3, pp. 165-184.

RUUD, P. A. (1984). – "Tests of Specification in Econometrics", *Econometric Reviews*, 3, pp. 211-242.

SAWA, T. (1978). – "Information Criteria for Discriminating Among Alternative Regression Models", *Econometrica*, 46, pp. 1273-1291.

SCHWARZ, G. (1978). – "Estimating the Dimension of a Model", *Ann. Statit.*, 6, pp. 461-464.

SMITH, A. F. M., SPIEGELHALTER, D. J. (1980). – "Bayes Factors and Choice Criteria for Linear Models", *J. R. Statist. Soc. Ser. B*, 42, pp. 213-220.

SPIEGELHALTER, D. J., SMITH, A. F. M. (1982). – "Bayes Factors for Linear and Log-Linear Models with Vagues Prior Information", *J. R. Statist. Soc. Ser. B*, 44, pp. 377-387.

VARDEMAN, S. B. (1987). – "Comments on Testing a Point Null Hypothesis", *J. Amer. Statis. Assoc.*, 82, pp. 130-131.

WOLPERT, R. L. (1995). – "Comments on Inference from a Deterministic Population Dynamics Model for Bowheaded Whales", *J. Amer. Statis. Assoc.*, 90, p. 426.

ZELLNER, A. (1971). – *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons, New York.

ZELLNER, A. (1984). – "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results", In: *Basic Issues in Econometrics*, University of Chicago Press, Chicago IL.

ZELLNER, A. (1986). – "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions", In: *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.), North-Holland, Amsterdam.