A fair comparison between regression models of different dimension

Vos, A.F.

1993

document version
Publisher's PDF, also known as Version of record

Faculteit der Economische Wetenschappen en Econometrie
Vakgroep Econometrie

Serie Research Memoranda

A fair comparison between regression models of different dimension

Aart F. de Vos

Research Memorandum 1993-78

December 1993

vrije Universiteit *amsterdam*

A fair comparison between regression models of different dimension.

Aart F. de Vos[1],

october 1993

ABSTRACT

This paper shows new ways to discriminate between regression models. The most appealing aspect is that the criteria come from all possible prediction exercises one can imagine, but may be computed directly from the residual sums of squares. The computational clue is the Binet-Cauchy theorem from linear algebra, which points the way to new possibilities, encompassing several recent proposals. The context is Bayesian with noninformative priors, but the basic results are also of interest from a Classical viewpoint. The use of predictions allows to derive Bayes Factors circumventing the problem that noninformative priors lead to probability statements of incomparable dimensions. These Bayes Factors perform well in simulation studies, but a number of questions remains.

[1]Free University
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

1. An informal summary.

This paper shows new ways to discriminate between regression models. The most appealing aspect is that the criteria come from all possible prediction exercises one can imagine, but may be computed directly from the residual sums of squares. The computational clue is the Binet-Cauchy theorem from linear algebra, which points the way to new possibilities, encompassing several recent proposals. The context is Bayesian with noninformative priors, but the basic results are also of interest from a Classical viewpoint. The use of predictions allows to derive Bayes Factors circumventing the problem that noninformative priors lead to probability statements of incomparable dimensions. These Bayes Factors perform well in simulation studies, but a number of questions remains.

That real prediction errors should provide the clue in discriminating between models is a widely held belief. An often used classical procedure is to discard part of the data in estimation, and to use these data for a forecast exercise. In Bayesian inference with improper priors, predictive likelihoods are the main clue in providing Bayes Factors, needed to update prior beliefs in models. The drawback of all these methods is that they depend strongly on the choice of starting values; they only use one possibility and thus are inefficient. This paper shows the way to combine all $\binom{n}{m}$ possibilities to predict (n-m) values from m starting values into one criterion. Moreover, it appears to be possible to combine these in such a way that the only thing one has to do (with an approximation that depends on the X matrices, but works very well in most cases) is to perform the two regressions. The result is invariably of the format of a likelihood ratio

$$\frac{p(y|M_0)}{p(y|M_1)} = c(n,k,r,m).\left(\frac{s_0^2}{s_1^2}\right)^{-(n-k)/2}$$

with the s_1^2 the "unbiased" estimate of the residual variance in both models, k the number of parameters in the largest model (M_1) and k-r in the smaller model (M_0). Whether the models are nested is irrelevant.

The constant c comes from the prediction errors, combined using the Binet-Cauchy theorem. A plausible though somewhat heuristic procedure is used to evaluate sets of forecast errors based on $m \geq k$ starting values; for each value of m, it becomes apparent that the complete set of forecast errors contains all relevant information, but the resulting values for c differ. This reveals some indeterminacy in the constant, which serves like a penalty for the use of parameters like in the criteria of Schwarz and Akaike. Using the smallest possible value for m, m=k, gives a criterion close to Schwarz's,

using large values results in lower penalties for the use of parameters. Using the likelihood ratio as Bayes factor to update prior beliefs, this indeterminacy is enlarged by the difficulty to define prior odds for the models (though unit odds seem fair).

Despite these indeterminacies, a simulation study mimicking realistic situations with both models having equal prior probabilities, shows that interpretation of the likelihood ratio as a Bayes factor yields good results. The choice between nonnested models of equal dimension (then c=1), is in most cases easy, the criterion favoring one of the models strongly. Nonnested models of different dimension also give straight answers in most cases. For nested models, the risk of an error of the second kind (rejecting the large model while it is right) is in many cases very small; by using high penalties (m=k) errors of both kinds become small. But in cases where the large model is very close to the small model, lower penalties (m=n-1) give better results by keeping the error of the second kind low. Thus the possibility to use several values for m brings back the familiar uncertainty of model choice for nested models with unspecified priors for the parameters extending the small model. However, new arguments for the choice of the proper criterion are involved in the derivations, and it may well be that the simulation study does not reflect the proper arguments.

Anyhow, despite the questions that remain, a useful tool arises from the algebra in this paper. New, computationally simple criteria with new interpretations are the result, and neither priors on parameters nor asymptotic arguments are involved.

The setup is as follows: section 2 treats model choice with a fixed set of starting values in general, section 3 the case of the regression model. Section 4 considers the fact that predictive likelihoods are contained in the degenerate distribution of regression residuals. Some consequences of the Binet-Cauchy theorem are derived in section 5, and used in section 6 to combine predictive densities for different starting values into one criterion. In section 7 the outcomes for minimal training samples (sets of starting values, size m=k) are derived, in section 8 training samples of m>k, and in section 9 the largest possible size (m=n-1). Section 10 shows that consistent updating formulae result. Section 11 gives the results of the simulation studies. Their are no separate conclusions apart from those mentioned above.

2. Model choice the Bayesian way and predictive likelihoods.

Let be

$M_1 \mid \theta_1$	the models being considered,
$\pi(M_1)$; $\pi(\theta_1 \mid M_1)$	the prior probabilities,
y	the data.

For two models M_0 and M_1

$$\frac{P(M_0 \mid y)}{P(M_1 \mid y)} = \frac{\pi(M_0)}{\pi(M_1)} \frac{p(y \mid M_0)}{p(y \mid M_1)} = \frac{\pi(M_0)}{\pi(M_1)} \frac{\int p(y \mid \theta_0, M_0)\pi(\theta_0 \mid M_0)}{\int p(y \mid \theta_1, M_1)\pi(\theta_1 \mid M_1)} , \qquad (1)$$

or: posterior odds is prior odds times Bayes factor.

With proper priors, the posterior odds give an elegant answer to a well defined question. And by analyzing these answers for different priors, it becomes apparent that the Bayes factor has many attractive features. The direct interpretation as relative evidence for both hypotheses avoids classical misunderstandings as does the robustness with respect to the meaning of sharp null hypotheses (see Berger and Delampady(1987) for a good summary of these arguments).

One may wonder, however, whether this is the question one wants to answer (Leamer 1978). If one really thinks of the models as two possible "true" descriptions of reality, and has prior ideas about both chances, it is the valid answer. But, once one has proper prior ideas, Bayesian inference allows to formulate, depending on the nature of the research,a decision function (one may wish simplification, plausibility, an argument for further research, or a real decision), all giving different answers. The standard problems of model choice and hypothesis testing, for which Bayesians advocate Bayes Factors are typically activities performed in the absence of sufficient prior information. Leamer (1992) has pointed the way to simplify the prior information needed by deriving elicitation diagnostics, and this may provide a solution in some cases, but it remains in most econometric models an almost impossible task and most researchers will feel more confident in using Jeffreys' (1961) noninformative priors.

However, this concept leads to serious trouble in model comparison, see, e.g. Leamer (1978, section 4.5), who shows that totally different answers arise for different concepts of noninformativeness. The standard answer - fitting nicely in the predictive context we will use is simple: the model with more parameters simply gets a posterior probability of zero. The essence of this result is that noninformative priors for models with a different number of

parameters lead to probability statements of different dimensions.

Leamer suggests a solution, based on a stationarity assumption for explanatory variables and asymptotic expansions, but the result is not very convincing. Some further reflection (see O'Hagan(1993) for an extensive discussion) shows that there can be no solution within the standard setup. The Bayes factor is an essentially non-robust concept, in the sense that the outcome of the analysis depends heavily on the prior assumptions, and that this dependency does not vanish asymptotically.
In the likelihood context, the most popular Bayesian criterion comes from Schwarz(1978): a penalty of ln(n)/2 per parameter in the loglikelihood. This result rests even more heavily on assumptions and asymptotic expansions.

All these problems seem to vanish when looking at predictive densities. A "training set" of m observations is used to get rid of the indeterminacy of the probability statements and the comparison is based on predictions. This may be done recursively, like in the Kalman filter. Phillips and Ploberger (1992) consider this as the essential mechanism in obtaining Bayes factors.
Starting with the first k observations, one may evaluate

$$p(y_{k+1}|y_1..y_k).p(y_{k+2}|y_1..y_{k+1}).....p(y_n|y_1...y_{n-1}) \qquad (2)$$

for both models. This leads to an intuitively attractive procedure. Each observation provides an update of the Bayes factor, which is multiplied by the ratio of the probabilities both models did assign to the realization. Typically one starts with unit odds and the better predicting model wins. The dilemma of comparing models of different dimension is directly clear in this setup: if one takes for k the number of parameters in the small model, this model makes a well defined prediction for observation k+1, while the larger model can't, leading to a zero score for this model.
The problem in using (2) is the dependency on the choice of the set k. Three solutions are proposed. Berger and Pericchi (1993) suppose to use all possible starting sets and averaging the score in some way. This is essentially the same solution as proposed in section 7 of this paper, though our approach involves a rather special type of averaging. O'Hagan (1993) supposes to use a larger training set than the minimal one. This is less efficient, but in section 8 we show that combination with the idea to use all possible starting sets is possible. Phillips and Ploberger (1992) do not use the idea of equal sets of starting values directly; they use the complete likelihood in both models and use the Radon Nykodym derivative which appears to have properties

that avoid the problem of different dimension. The high level of abstraction in their derivation makes this seem magic, but our more down to earth section 4 suggests that such a solution may be justified. For the time being, however, our analysis of the standard regression model suggests that more clarity is needed about the question we want to answer.

3. The regression model with σ known

For the standard regression model with σ known, the result of (2) may be calculated directly using Bayes' rule. This takes care of the switch between parameters and data in an elegant way (notation: y_{n-k} is the vector of y values excluding k):

$$p(y_{n-k}|y_k) = \int p(y_{n-k}|y_k,\beta)p(\beta|y_k)d\beta = \int p(y_{n-k}|\beta)p(\beta|y_k)d\beta \qquad (3)$$

by definition this is equal to (2). As

$$\beta|y_k \overset{d}{=} N(b_k,\sigma^2[X_k'X_k]^{-1}) \qquad (4)$$

with $b_k=[X_k'X_k]^{-1}X_k'y_k$,

$$p(\beta|y_k)= |X_k'X_k|^{0.5}(2\pi)^{-(k-r)/2}\sigma^{-(k-r)}\exp(-0.5[(\beta-b_k)'[X_k'X_k](\beta-b_k)]/\sigma^2) \qquad (4a)$$

This is a well defined probability distribution, provided $X_k'X_k$ is not singular (as is the case for k smaller than the number of regression parameters). Working out (3) gives (we use $\alpha=\sqrt{2\pi}$)

$$p(y_{n-k}|y_k,\sigma) = |X'X|^{-0.5}|X_k'X_k|^{0.5}(\alpha\sigma)^{-(n-k)}\exp\{-0.5(C- B'A^{-1}B)/\sigma^2\} \qquad (6)$$

with

$$C - B'A^{-1}B = y_{n-k}'y_{n-k} + b_k'[X_k'X_k]b_k \; y'X[X'X]^{-1}X'y \qquad (6a)$$

We first consider the term (6a). If X_k is of full rank, the second term is $y_k'y_k$, and (6a) becomes the familiar expression for the residual sum of squares: y'My, with $M = I - X[X'X]^{-1}X'$. Note that this expression is symmetric in the observations, despite the asymmetric probability statement.

If X_k is k*p, with p the number of parameters (k>p), (6a) is not a standard residual sum of squares; it may be written as

$$C- B'A^{-1}B = y'My - (y_k'y_k - b_k'[X_k'X_k]b_k)$$

$$= y'My - y_k'M_k y_k \qquad (7)$$
$$= RSS - RSS_k$$

where RSS is the total residual sum of squares and RSS_k the residual sum of squares obtained from the regression with the p regressors on the first k observations (so $RSS_k=0$ if k=p).

This suggests a simple direct comparison of models. One takes for k the number of parameters in the large model, and compares the total RSS in the model with parameters k with the adjusted sum of squares RSS - RSS_k in the smaller model. Evaluating the probability statements at the resulting maximum likelihood estimates of the σ's (for large n a good approximation of the final purely Bayesian result, which we will treat later on)

$$\hat{\sigma}_1^2 = (RSS_1 - RSS_{1k})/(n-k) \qquad (8)$$

i referring to model i, we get the result:

$$\frac{p(y_{n-k}|y_k,M_0)}{p(y_{n-k}|y_k,M_1)} \cong \frac{p(y_{n-k}|y_k,M_0\,\hat{\sigma}_0)}{p(y_{n-k}|y_k,M_1,\hat{\sigma}_1)} = \frac{|X_0'X_0|^{-0.5}|X_{0k}'X_{0k}|^{0.5}\hat{\sigma}_0^{-(n-k)}}{|X_1'X_1|^{-0.5}|X_{1k}'X_{1k}|^{0.5}\hat{\sigma}_1^{-(n-k)}} \qquad (9)$$

Unfortunately the outcome depends heavily upon the choice of starting values.
An extreme example is when one of the $X_k'X_k$ matrices happens to be singular: the corresponding model gets zero probability. Moreover the choice of k is rather arbitrary. So the next task is to get rid of the choice of starting values and to decide on the optimal value of k. But first we have a closer look at the concept of predictive densities.

4. The direct link between predictive likelihoods and residuals

The preoccupation with predictive densities is put in another perspective if we look at the distribution of the residual vector. This is an n-p dimensional singular normal distribution, proportional to

$$(\alpha\sigma)^{-(n-p)}\, e^{-0.5\{y'My/\sigma^2\}}$$

Writing

$$M = \begin{bmatrix} M_{11} & M_{21} \\ M_{12} & M_{22} \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad (10)$$

the indices 1 and 2 referring to partitions of size m and (n-m) respectively, it is easily deduced (see appendix A) that

$$y'My = y'My = y_1'M_{11}y_1 + 2y_1'M_{12}y_2 + y_2'M_{22}y_2 \qquad (11)$$

and thus that (10) corresponds with a distribution of $y_2|y_1$ which is n-m dimensional normal, with variance $M_{22}^{-1}$and expectation - $M_{22}^{-1}M_{21}y_1$, multiplied by a factor

$$\exp(-0.5(y_1'[M_{11} - M_{12}M_{22}^{-1}M_2]y_1/\sigma^2)) = \exp(-0.5(RSS_1/\sigma^2)). \qquad (12)$$

As moreover (see again appendix A)

$$|X_m'X_m|/|X'X| = |M_{22}| \qquad (13)$$

we conclude that

$$p_2(y) = p_2(y_2|y_1) \qquad (14)$$

in other words:

$$p_{n-m}(y) = c_m p(y_{n-m}|y_m). \qquad (14a)$$

This shows that *all* possible predictive likelihoods may directly be derived from the density of the vector of residuals. This leads to two points.
First, the different probability statements arising for different starting values appear to be only different ways to look at the same distribution. Note that even in making a probability statement on one model, the statements based on m starting values, which all depend on y by y'My, depend on the choice of starting values by the $X_m'X_m$ matrices. Different predictive probabilities thus may be essentially equal. In other words: to compare predictive probabilities, one must convert them to the same base.
Second, it must be possible to combine predictive probabilities to get some unique number. Every set of predictive probabilities containing all information should then lead to the same probability statement if converted to the same base.
In the next section we show how such a combination may be performed, such that the choice of starting set drops out in a plausible way.

5. The Binet-Cauchy theorem

The theorem of Binet-Cauchy says:

if A and B are m*n and n*m matrices, respectively, m≤n, then

$|AB| = \sum_{(k)} |A_k B_k|$, with A_k and B_k corresponding m*m submatrices of A and B,

and the summation being over all $\binom{n}{m}$ possible sets {k}.

We only need symmetrical expressions of the form $|A'A| = \sum_{(k)} |A_k' A_k|$. For A=X, the theorem says

$$|X'X| = \sum_{(k \in n)} |X_k' X_k|$$

For the selections of size m, when the rank of X is p, we may apply the theorem again in the form

$$|X_m' X_m| = \sum_{(p \in m)} |X_p' X_p|$$

implying that all $X_p' X_p$ matrices are selected $\binom{n}{m}\binom{m}{p}/\binom{n}{p} = \binom{n-p}{m-p}$ times, so

$$\sum_{(m \in n)} |X_m' X_m| = \binom{n-p}{m-p} \sum_{(p \in n)} |X_p' X_p| = \binom{n-p}{m-p} |X'X|. \qquad (15)$$

For A = [X y], Binet-Cauchy is applicable to selections of size m+1 and

$$|X'X| y'My = \sum_{(m+1 \in n)} |X_{m+1}' X_{m+1}| y_{m+1}' M_{m+1} y_{m+1}'.$$

For X with rank p<m, we must convert the theorem to p+1 dimensional matrices. With a similar argument as before we get for selections of size m

$$\sum_{(m)} |X_m' X_m| y_m' M_m y_m' = \binom{n-p-1}{m-p-1} |X'X| y'My. \qquad (16)$$

6. The combination of predictions from all possible starting sets

Suppose we consider for a model with p parameters training samples of size m≥p. We then have $\binom{n}{m}$ possibilities of choosing a starting set. The predictive densities are

$$p(y_{n-m} | y_m, \sigma) = (\alpha\sigma)^{-(n-m)} \{|X'X| / |X_m' X_m|\}^{0.5} \exp\{-0.5[y'My - y_m' M_m y_m]/\sigma^2\} \qquad (17)$$

The Binet-Cauchy theorem provides clues for the combination of these densities. In fact we think that the theorem is so powerful and suited for the

problem that it *must* show the way to combine the densities. However, despite intensive trials, we only partially succeeded in finding the clue.
The only way that leads to comprehensive results is a linear combination of the exponential term. First consider weights

$$c_m = |X'_m X_m| \binom{n-p}{m-p}^{-1} |X'X|^{-1} \tag{18}$$

(16) implying that $\sum_{(m)} c_m = 1$. The consequence is

$$\sum_{(m)} c_m [y'My - y'_m M_m y'_m] = y'My - \binom{n-p}{m-p}^{-1} \binom{n-p-1}{m-p-1} y'My = \frac{n-m}{n-p} y'My \tag{19}$$

so, instead of different residual sums of squares, each time adjusted for the residual sum of squares in the starting values, this weighted sum is independent from the starting set and adjusts the overall n-p dimensional RSS according to the difference in dimension using m starting values.
Substituting (19) into (17) learns

$$\sum_{(m)} c_m \ln p(y_{n-m} | y_m, \sigma) = - (n-m)\ln(\alpha\sigma) + 0.5\sum_{(m)} c_m \ln(|X'_m X_m| / |X'X|) - 0.5[\frac{n-m}{n-p} y'My]/\sigma^2 \tag{20}$$

The first interesting thing about this expression is that it provides a valid n-m dimensional expression for p(y), symmetrical in y. "Valid" in the view of the "mixing lemma":

$$p(M|y,R) = \sum_S p(M|S,y,[R])p(S|y,R) \tag{21}$$

where R is a dimension reduction procedure, to obtain a proper probability statement, S is the set of values chosen by R and $p(M|S,y,[R]) = p(M|y,S)$ is the probability statement we get by taking the starting values set S ((1..m) in our example).

So, to get a proper statement we must define $p(S|y,R)$, in other words a procedure which selects S, possibly depending on y, and possibly taking several sets of starting values with probabilities (or weights) summing to one.
As our weighted geometric average between the possible predictive probability statements, with weights summing to unity, always lies between the lowest and the highest one, it is equal to some arithmetic average of the probabilities, and thus equal to the expected probability in some process selecting sets of starting values with probabilities summing to one.

The second interesting outcome is the maximum likelihood estimator for σ^2 based on the "mean predictive likelihood" (20):

$$\hat{\sigma}^2 = y'My/(n-p) = s^2 \qquad (22)$$

the classic unbiased estimator, independent of the size of m.
However, it remains to be seen whether all information from the forecasts is used efficiently by taking a weighted geometric average. The different (vector) forecasts contain to some degree independent information, and independent information leads to multiplication of probability statements. Completely independent log-probability statements, - the extreme case - may simply be added, so the weights sum to n. In our case, we know that n-p independent probability statements may be obtained from y'My. This dimension may be obtained by using weights

$$d_m = \frac{n-p}{n-m} c_m = |X'_m X_m| \binom{n-p-1}{m-p-1}^{-1} |X'X|^{-1}$$

which gives

$$\sum_{(m)} d_m \ln p(y_{n-m} | y_m, \sigma) = c - (n-p)\ln\sigma + 0.5\sum_{(m)} d_m \ln(|X'_m X_m|/|X'X|) - 0.5[y'My]/\sigma^2 \qquad (23)$$

which proves that, for any size $m \geq p$ of the training sample, it is possible to reproduce, up to a factor depending on X, the distribution of the residual sum of squares. We conclude from this fact, in view of section 4, that this is the proper way to combine predictions. The only problem is the rather artificial base for the predictive density.
Note that now for inference on σ^2, the sets of forecasts always provide the same posterior, which is also equal to the posterior based on the concentrated likelihood. This once more confirms that our way to combine predictions is the optimal one.

7. Bayes factors for minimal training samples of equal size

First we explore the weighted geometric mean of predictive probability statements of maximal dimension (23). This is interesting as it provides useful insights and as it is a variant upon a proposal by Berger and Pericchi (1993), who suggests to average all possible Bayes factors.
With k we denote the higher number of regression parameters, with k-r the lower number, such that in the case of nested models r is the number of

restrictions. The "mixture" (19) then becomes

$$\sum_{(k)} c_k [y'My - y_k' M_k y_k'] = \frac{n-k}{n-(k-r)}\ y'My. \qquad (24)$$

Substituting (24) into (17) learns

$$\sum_{(k)} c_k \ln p(y_{n-k}|y_k,\sigma) = -(n-k)\ln\alpha + \\ - (n-k)\ln\sigma + 0.5\sum_{(k)} c_k \ln(|X_k'X_k|/|X'X|) - 0.5[\frac{n-k}{n-(k-r)}\ y'My]/\sigma^2\ ) \qquad (25)$$

evaluating this for the ML estimate for σ from (20) we get

$$\ln(p_{n-k}(y)) \cong -(n-k)\ln\alpha - (n-k)\ln\hat{\sigma} + 0.5\sum_{(k)} c_k \ln(|X_k'X_k|/|X'X|) - 0.5(n-k) \qquad (26)$$

In appendix B it is explained that

$$\sum_{(k)} c_k \ln(|X_k'X_k|/|X'X|) = \ln\binom{n-k+r}{r} + \sum_{(k)} c_k \ln c_k \qquad (27)$$

approximately equals

$$\ln\binom{n-k+r}{r} - \ln\binom{n}{k} + c \qquad (27a)$$

The last expression, with c being a small constant with slight variation among models is explained in appendix B. Assuming that the approximation holds, we get from (26) a simple formula for what we may call the mean predictive density:

$$\ln(p_{n-k}(y)) \cong c^* - (n-k)\ln\hat{\sigma} + 0.5\ln\left(\frac{(n-k+r)!}{r}\,\frac{k!}{n!}\right) \qquad (28)$$

with $c^* = -(n-k)\ln\alpha - (n-k)/2 + c$ about equal among models.
The corresponding Bayes factor becomes

$$\frac{\hat{p}(y_{n-k}|H_0)}{\hat{p}(y_{n-k}|H_1)} \cong \binom{n-k+r}{r}^{1/2}\left(\frac{s_0^2}{s_1^2}\right)^{-(n-k)/2} \qquad (29)$$

where we use $\hat{p}$ to denote the aspect that maximum likelihood estimates for σ are used.

The main part looks like a residual variance criterion. Theil (1971, p543) shows that the expectation of s^2 is lowest for the correct model. In terms of the ratio of residual sums of squares the criterion may be written

$$\frac{\hat{p}(y_{n-k}|H_0)}{\hat{p}(y_{n-k}|H_1)} \cong \binom{n-k+r}{r}^{1/2} \left(\frac{n-k}{n-k+r}\right)^{-(n-k)/2} \left(\frac{RSS_0}{RSS_1}\right)^{-(n-k)/2} \qquad (30)$$

note that in asymptotic criteria like Schwarz's the RSS ratio occurs without adjustment for degrees of freedom and to the power n instead of n-k. In both respect our criterion seems more plausible.
Noting that our "dimension penalty" for the loglikelihood is $0.5 \ln\binom{n-k+r}{r} \cong$ $r/2 \ln(n-k) - 1/2 \ln(r!)$, we see another difference with Schwarz' criterion: apart from using n-k instead of n again, the factor $- 1/2 \ln(r!)$ means higher penalties for more restrictions. Putting all differences together, and using the approximation

$$\left(\frac{n-k}{n-k+r}\right)^{-(n-k)/2} \cong e^{r/2}$$

our criterion may be written as

$$\ln \frac{\hat{p}(y_{n-k}|H_0)}{\hat{p}(y_{n-k}|H_1)} \cong r/2 \ln(n-k) - 1/2\ln(r!)+r/2 - (n-k)/2\ln \left(\frac{RSS_0}{RSS_1}\right) \qquad (31)$$

the difference with Schwarz is, apart from using n-k (which has obvious appeal), the factor $r/2 - 1/2\ln(r!)$. Numerical values for this factor are

1	2	3	4	5	6	7	8	9	10
0.50	.65	.60	.41	.11	-.29	-.76	-1.30	-1.90	-2.55
11	12	13	14	15	16	17	18	19	20
-3.25	-3.99	-4.78	-5.60	-6.45	-7.34	-8.25	-9.20	-10.17	-11.17

So we see that the penalty for the larger dimensional model quite close to Schwarz's for r<8. For large r our penalty is lower.

8. Optimizing the size of the training sample(s)

The reason to use minimal training sets is to use as much of the data as possible for evaluation. There is a good reason however to consider larger training sets: small training samples put an unfair large penalty upon high dimensional models, the uncertainty about parameter estimates being huge. For this reason O'Hagan (1993) suggests "moderate" training sets (without saying how moderate). This argument seems to lose much of its value once one uses all possible starting sets by the Binet-Cauchy mixing, but the dimension of the criterion still remains that of a probability statement on n-k values

conditional upon k values. And, as will become clear, there are possibilities to use larger training sets leading to lower penalties for large models.

Consider training samples of size m, and two models M_0 and M_1 with k-r (=p_0) and k (=p_1) parameters respectively, the Bayes factor (29) changes into

$$\frac{\hat{p}(y_{n-m}|H_0)}{\hat{p}(y_{n-m}|H_1)} \cong \binom{n-p_0}{m-p_0}^{1/2} \binom{n-p_1}{m-p_1}^{-1/2} \left(\frac{s_0^2}{s_1^2}\right)^{-(n-m)/2} \qquad (32)$$

which shows that, in increasing m, we lose power in the part representing the fit (the ratio s_0^2/s_1^2), but decrease the penalty function for the dimension. All as expected. Considering that it is unfair to the larger model to compare on the base of small training samples (the relevant training sample has size n), this gives rise to another Bayes factor, less unfair than that based on training sample k. We simply minimize (32) with respect to m, so choose the Bayes factor as favorable as possible for the larger model. In this procedure the ratio s_0^2/s_1^2 is fixed, but this is valid in view of the mixing lemma of section 5.

This procedure makes sense for the ratio s_0^2/s_1^2 larger than 1, so the large model having the better fit. For values near 1, the result is m=n-1, and the Bayes factor is about 1. The larger the ratio, the smaller values for m result, but always "reasonable" values (like O'Hagan proposes), much larger than k. For moderate ratio's the increase in the Bayes factor is spectacular, as the following table shows:

Table 1 n = 50, p_1 = 7

s_1^2/s_0^2	p_0 = 0	1	2	3	4	5	6
0.80	35	31	27	23	19	15	11
0.85	46	41	35	29	29	18	12
0.90	49	49	49	43	34	25	16
0.95	49	49	49	49	49	44	26
1.00	49	49	49	49	49	49	49

for other values of k and r similar results arise (r, the difference in dimension, being the decisive factor, different values for k hardly matter.

We also computed the table for the Bayes factor, integrating σ^2 out for both models with respect to the noninformative prior $\pi(\sigma^2) \propto \sigma^{-2}$ (assuming the proportionality constants equal, which seems the only fair option). Then, (32) changes into

$$\frac{\hat{p}(y_{n-m}|H_0)}{\hat{p}(y_{n-m}|H_1)} \cong \binom{n-p_0}{m-p_0}^{1/2} \binom{n-p_1}{m-p_1}^{-1/2} \left(\frac{s_0^2}{s_1^2}\right)^{-(n-m-1)/2} \tag{33}$$

which only differs slightly from (32). The outcomes for the criterion become

Table 2 $n = 50$, $p_1 = 7$

s_1^2/s_0^2	$p_0 = 0$	1	2	3	4	5	6
0.80	35	31	26	23	19	15	11
0.85	46	40	35	29	24	18	12
0.90	49	49	49	43	34	25	16
0.95	49	49	49	49	49	45	25
1.00	49	49	49	49	49	49	49

indeed only slightly different from table 1

Obviously we are making progress towards a fair comparison, as far as we believe in the arguments to do so. Still, in this view the Bayes factor for better fitting large models must be to low, as the relevant "training set" is n, much larger than m. Moreover the methodology is unsatisfactory for the case the smaller model fits better: the argument then leads to m=n-1, throwing away almost all the evidence. What one actually wants is a procedure that uses large training samples, without throwing the evidence from the fit away. Such a procedure exists, as the sequel will show.

9. Probability statements of maximal equal dimensions

In section 6 we showed that using Binet-Cauchy weights for the log- predictive probabilities with weights summing to more than unity, may reproduce all available information. However, combining of all information available in the forecasts for both models leads to the problem we started with: probability statements of incomparable dimension. This problem may be met by using weights

$$e_m = \frac{n-k}{n-m} c_m = |X'_m X_m| \binom{n-p-1}{m-p-1}^{-1} |X'X|^{-1} \tag{34}$$

and thus

$$\sum_{(m)} e_m \ln p(y_{n-m}|y_m,\sigma) = -(n-k)\ln\alpha + \\ - (n-k)\ln\sigma + 0.5\sum_{(m)} e_m \ln(|X'_m X_m|/|X'X|) - 0.5\,\frac{n-k}{n-p}[y'My]/\sigma^2 \tag{35}$$

which looks like a probability statement of dimension n-k (in the next section we will show that it may be considered as such). So, we may construct

probability statements of maximal equal dimension. The choice of m in this context only reflects the choice of the base, and we may do this to maximize the fairness with respect to the largest model by putting m=n-1, i.e.

$$e_m = (n-k)\, c_m = |X'_m X_m| \binom{n-p-1}{m-p-1}^{-1} |X'X|^{-1}, \text{ and} \tag{36}$$

$$\sum_{(m)} e_m \ln p(y_{n-m}|y_m,\sigma) = -(n-k)\ln\alpha +$$

$$- (n-k)\ln\sigma + 0.5\sum_{(m)} e_m \ln(|X'_m X_m|/|X'X|) - 0.5\,\frac{n-k}{n-p}[y'My]/\sigma^2 \tag{37}$$

resulting in the "Fair Bayes Factor"

$$\frac{\hat{p}(y_{n-m}|H_0)}{\hat{p}(y_{n-m}|H_1)} \cong \left(\frac{n-p_0}{n-p_1}\right)^{(n-k)/2} \left(\frac{s_0^2}{s_1^2}\right)^{-(n-k)/2} \tag{38}$$

10. Consistent updating

The question is now whether we may treat (34) as a (n-k) dimensional probability statement on y. We will prove that, in a sense, we may.

The crucial aspect of the kind probability statement we need, is that it leads to proper updates of evidence for the models when adding new data.

Bayes' formula requires that

$$p(\vec{y}_n|M)=p(y_n|\vec{y}_{n-1},M)p(\vec{y}_{n-1}|M) \tag{39}$$

as

$$p(y_n|\vec{y}_{n-1},M)=(\alpha\sigma)^{-1}\cdot\left(|X'_{n-1}X_{n-1}|/|X'_n X_n|\right)^{1/2}\exp(-1/2(y'My-y'_{n-1}M_{n-1}y_{n-1})/\sigma^2) \tag{39a}$$

this obviously cannot be true in the normal sense. This is not amazing as, if (39a) would hold for all recursions from k, formula (1) would result with the dependency on the starting values.

The solution is to consider (39) backwards. We require that a probability statement on y_1 .....y_n is a proper combination of all forecast statements for the n elements:

$$p(\vec{y}_n|M)=\Pi_{(m)}\{p(y_{n-m}|\vec{y}_m,M)p(\vec{y}_m|M)\}^{c_m} \tag{40}$$

So, as m=n-1, a mixture of the n forecasts to be made by deleting all y's once. Not surprising, we will base the proper mixture on the Binet-Cauchy

theorem.

In the exponential term, we get (divided by σ^2) for the right hand side

$$\Sigma_{(m)} c_m \left(\frac{n-k-1}{n-p-1} \, y_m' M_m y_m + y'My - y_m' M_m y_m \right) = \left(\frac{p-k}{n-p-1} \cdot \frac{n-p-1}{n-p} + 1 \right) y'My = \frac{n-k}{n-p} \, y'My$$

so (40) holds exactly in this respect.

The power of σ obviously fulfills its requirement. The remaining term of interest is

$$\prod_{(n-1 \in n)} \left(|X'_{n-1} X_{n-1}| / |X'X| \right)^{c_{n-1}} \left(\prod_{(n-2 \in n-1)} \left(|X'_{n-2} X_{n-2}| / |X'_{n-1} X_{n-1}| \right)^{c_{n-2}} \right) \qquad (41)$$

with

$$c_{n-1} = |X'_{n-1} X_{n-1}| / |X'X| \; (n-p)^{-1} \text{ and } c_{n-2} = |X'_{n-2} X_{n-2}| / |X'_{n-1} X_{n-1}|.$$

Substituting the approximation

$$\Sigma_{(n-1 \in n)} c_{n-1} \ln(c_{n-1}) \cong -\ln(n) \qquad (42)$$

in (41) and using (27) we derive the following "identity"

$$\exp(\ln(n) + \ln(n-p) + (n-k-1)(\ln(n-1) + \ln(n-1-p))). \qquad (43)$$

Fortunately (43) is (for large n) approximately equal to

$$\exp((n-k)(\ln(n)+\ln(n-p)) \qquad (44)$$

which gives us our desired result (40).

11. Some philosophical questions and simulations

What does our "Fair Bayes Factor" mean? We hope of course that it is a Bayes factor in the traditional sense, to be interpreted as posterior odds for unit prior odds. The argument that it can't be that because this would contradict Bayes Factors based on informative priors is invalid: all probability statements depend on information, and the idea is that we excluded a part of the prior information on parameters. Not all prior information, as "prior odds being unity" implies some prior ideas, as will become clear in the sequel.

If only some information is deleted, our Bayes Factor should give valid, though not fully efficient answers if there is an informative prior. And the least one can hope is that the results are better than those of other criteria

using the same information.

To check this, we performed some simulation studies. It is not simple, however, to define a simulation experiment that shows what question is answered. The central message of the criteria seems to be the update message: if a priori both models are equally likely, adding the forecasts updates the probabilities. The problem is to define a situation where "equally likely" has a specific meaning. To solve this we used is a presample:

We generate drawings from two different models, using proper priors. We use a small dataset (size ns) and compute for any case s_0^2 and s_1^2. Next we select pairs of cases coming from different models with equal ratios s_0^2/s_1^2. Thus we obtain a prior situation with equal probabilities that the models are correct, in a way that is the same for all criteria depending on this ratio, that is for almost all criteria including the F test.

Next we generate new data for all selected models, and obtain new ratios. The question is now whether criteria computed from these new ratio are suited to calculate posterior probabilities for the right model. For each criterion we calculate the loglikelihood score

$$\sum_{M_1}(\ln p(M_1))+\sum_{M_2}(\ln p(M_2))$$

The criteria we compared are Schwarz, Akaike, p(F) which is the classical "p value for which the F test just rejects the null, and the two extremes of our prediction criteria: Pred(k) based on training samples of minimal size and Pred(n-1) based on training samples of maximal size. Also we looked at the optimal discrimination function of the format

$$\frac{p(y|M_0)}{p(y|M_1)} = c.\left(\frac{s_0^2}{s_1^2}\right)^{-b}$$

for varying c and b.

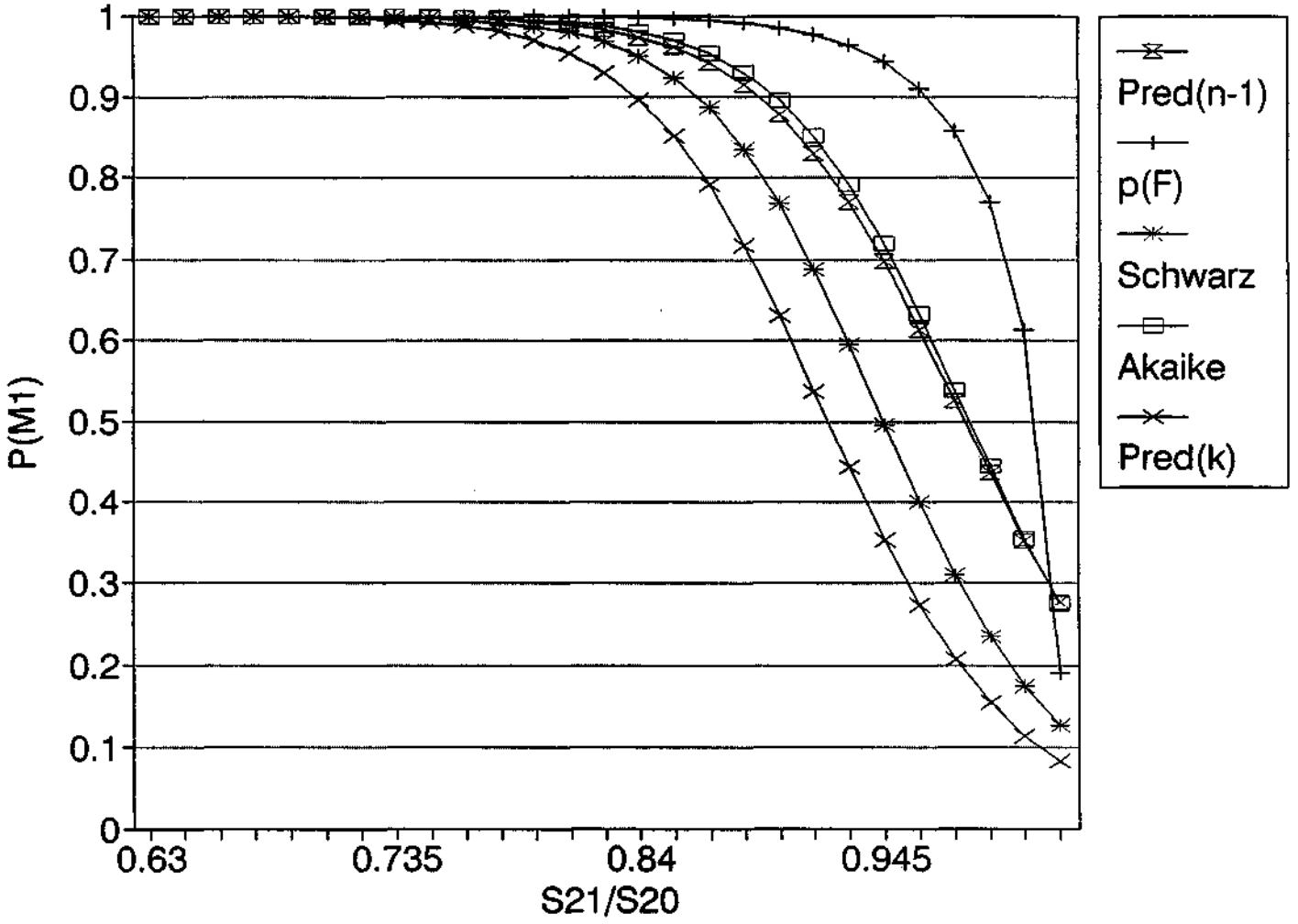
The answers appears to be a mixed success. Only the hypothesis b=(n-k)/2 is confirmed. The optimal value for c and the relative performance of the criteria depends on the priors, on the size of the starting set and (less so) on the size of the sample. One can distinguish three situations:

A. Nested models

In nested models, M_0 being the small model and M_1 the model with more parameters, s_1^2/s_0^2 is close to unity if M_0 is true. The distribution of the ratio if M_1 is true depends on the situation. One can get them close to unity (around 0.9, say) either by using large presamples (the ratio under M_0 is than

Fig 1. P(M1) as dependent on the ratio S_1^2/S_0^2 according to five criteria

n=50, k0=2, k1=3



so strongly concentrated that only the small percentage of values near 1 are selected), or by specifying a prior with a low signal-to-noise ratio. With the latter we mean, rewriting the model as

$y = X_0\beta + X_1\gamma + u$ with $X_0'X_1 = 0$ and $X_1'X_1$ diagonal,

that $[\Sigma_j \gamma_j^2 \mathrm{Var}(x_{1j})]/(n-k)$ is small compared to σ^2. In our simulations we used small presamples and manipulated $\mathrm{Var}(x_{1j})$ to obtain different situations.

Fig 1 makes clear (for one specific situation) what happens. The criteria simply differ with respect to the penalties they impose on the large model. In this case (and other cases differ little) in decreasing order of penalties Pred(k), Schwarz, Pred(n-1), Akaike, p(F).

If the ratio under M_1 tends to be small, high penalties tend to win. In the first part of table 3 (corresponding with fig 1) this is the case. Schwarz and Pred(k) win, Pred(k) just losing by a bad score if M_1 is true (for even lower ratios, Pred(k) would win). The looser is p(F), corresponding with the well known effect that the interpretation of the classical error of the first kind as evidence against the null (which is formally not justified, but quite natural) grossly tends to overestimate this evidence.

The second part of table 3 shows that for n=100 the results are similar, but the scores get better (as to the interpretation of the scores: the -22.07 for Pred(k) corresponds with an expected score for 100 probability statements having a probability of 94% of being right and saying so).

In the third part of table 3, the mean ratio under M_1 is manipulated to be much larger. Now the discrimination is much harder (the scores are worse) and low penalties win. That is to say: Pred(n-1) and Akaike do, the p(F) criterion still looses by a bad score if M_0 is true, and its overall score is worse than a fifty-fifty gamble (giving a certain score of $100.\ln(0.5) = -69.31$). But the latter is also true for Pred(k) by its horrible score if M_1 is true, caused by underestimating the probability of this being the case.

Minimax considerations seem to advocate Pred(n-1), on the other hand it is clear that choosing the wrong model is considerably worse when the models differ a lot, so with other criteria Schwarz and Pred(k) might be the winners.

We also performed studies with k1 much larger, but the conclusions remained the same.

Table 4 concerns a much easier situation: the choice between two (correlated) explanatory variables. Here the difference in penalties is irrelevant. Discrimination between the models is good (-36.26 corresponds to a 88% score), the difference between the criteria is irrelevant, but the prediction criteria saying that one may simply raise the ratio of residual sums of squares to the

power (n-k)/2 to obtain odds is an elegant outcome. That even the F test scores well is of little relevance: it is not applicable.
For nonnested models with different degrees of freedom, the scores also tend to be good for all criteria, the winner depending on the circumstances.
So, firm conclusions are difficult to draw, especially for non-nested models. The main conclusions are that everything is better than the F-test, and that further thinking is required about the questions we want to answer.

Appendix A. The distribution of the residuals.

If X_1 is an arbitrary training sample of size m of X and X_2 the remainder of X, we can write

$$M = I - X(X'X)^{-1}X' = I - \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}(X'X)^{-1}\begin{bmatrix} X_1 X_2 \end{bmatrix}$$

$$= \begin{bmatrix} I - X_1(X'X)^{-1}X_1'X_1(X'X)^{-1}X_2' \\ -X_2(X'X)^{-1}X_1' \quad I - X_2(X'X)^{-1}X_2' \end{bmatrix}$$

$$= \begin{bmatrix} M_{11} M_{12} \\ M_{21} M_{22} \end{bmatrix}.$$

Equation (6) written (explicitly) as a normal distribution in y_2 ($y_{>m}$) gives

$$y_2 \overset{d}{=} N(-M_{22}^{-1} M_{21} y_1, \sigma^2 M_{22}^{-1}). \qquad (1)$$

As (1) and (6) are essentially the same, we derive the following identity

$$|M_{22}| \exp(-y_1'M_{12}M_{22}^{-1}M_{21}y_1/\sigma^2) = |X_1'X_1|/|X'X| \exp((y_1'M_1y_1 - y_1'M_{11}y_1)/\sigma^2). \qquad (2)$$

with $M_1 = I - X_1(X_1'X_1)^{-1}X_2'$. A simple matrix theorem (see e.g. Judge...),
if A, D are singular matrices of order m, n and B, C are (m x n) and (m x n) matrices, respectively, then

$$|A| \ |D + CA^{-1}B| = |D| \ |A + BD^{-1}C|, \qquad (3)$$

applied with $A = X'X$, $B = X_2'$, $C = X_2$ and $D = I$ learns

$$|M_{22}| = |X_1'X_1|/|X'X|. \qquad (4)$$

And thus

$$y_1'M_1y_1 = y_1'M_{11}y_1 - y_1'M_{12}M_{22}M_{21}y_1. \quad (5)$$

Our (n-p) dimensional probability statement would look something like

$$(\alpha\sigma)^{-(n-p)/2} \exp(-y'My/2\sigma^2). \quad (6)$$

Now y'My can be written as

$$y'My = y_1'M_{11}y_1 + y_2'M_{21}y_1 + y_1'M_{12}y_2 + y_2'M_{22}y_2 \quad (7)$$

$$= (y_2 + M_{22}^{-1}M_{21}y_1)'M_{22}(y_2 + M_{22}^{-1}M_{21}y_1) + y_1'(M_{11} - M_{12}M_{22}^{-1}M_{21})y_1$$

$$= (y_2 + M_{22}^{-1}M_{21}y_1)'M_{22}(y_2 + M_{22}^{-1}M_{21}y_1) + y_1'M_{11}y_1.$$

So (6) corresponds with a distribution of $y_2|y_1$ which is (n-m) dimensional normal with variance M_{22}^{-1} and expectation $-M_{22}^{-1}M_{21}y_1$ multiplied by a factor $\exp(-y_1'M_{11}y_1/2\sigma^2)$. In other words

$$p_{n-k}(y) = c_1\, p(y_2|y_1). \quad (8)$$

Appendix B. Heuristic proof that the the factor $\sum_{(k)} c_k \ln c_k$ differs little among models

In both models the c_k are $N = \binom{n}{k}$ very small positive weights summing to 1 with, in most cases, only moderate variation.

One possibility to get some further understanding of the size of $\sum_{(k)} c_k \ln c_k$ is the following:

Imagine the c_k ordered by size (descending) on intervals numbered 1...N. Approximate the resulting discrete density by its continuous counterpart, assuming a truncated exponential distribution:

$$c_k \cong \lambda e^{-\lambda k}/(1-e^{-\lambda N}) \qquad k=1...N$$

so the facts that they are descending and summing to unity have been incorporated. Variation among distributions of c_k can be represented now as variations in $N\lambda$. Let α be the percentage of c_k covering 90% of the probability. A rather extreme case is α=0.1: the largest 10% of the c_k sum to 0.9. The other reasonable extreme is α=0.8, meaning very little variation among the c_k.

As is easily shown, with each value of α there corresponds a value of λN, while

$$E(\ln c_k) = \sum_{(k)} c_k \ln c_k = -\ln(1-e^{-\lambda N}) + \ln\lambda - (1-(1+\lambda N)e^{-\lambda N})/(1-e^{-\lambda N})$$

leading to the following table:

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
λN	23	11.5	7.6	5.7	4.4	3.4	2.5	1.51
$\sum_{(k)} c_k \ln c_k + \ln N$	2.13	1.44	1.04	.76	.55	.38	.23	.09

The last line shows that for extreme differences between the conditioning of the X matrices a difference of 2 points in $\sum_{(k)} c_k \ln(c_k)$, so of 1 point in the loglikelihood (26), may occur (the value of N is equal among models). In most cases only some decimals difference will occur. Note that one can estimate the necessary adjustment by taking samples of size k from the X matrices.

References

Berger, J.O. and M. Delampady (1987) "Testing precise hypotheses". *Statistical Science*, 3, 317-352.

Berger, J.O. and L.R. Pericchi (1993) "*The Intrinsic Bayes Factor for model selection and prediction*" Purdue University Technical Report 93-43C.

Jeffreys, H (1961) *Theory of Probability*. Oxford University Press, London.

Leamer, E.E. (1978) "*Specification searches: Ad Hoc Inference with Nonexperimental Data*". Wiley, New York.

Leamer (1992) "Bayesian Elicitation Diagnostics" *Econometrica* Vol 60, pp 919-942.

O'Hagan, A.(1993) "*Partial Bayes Factors for Model Comparison*" Working paper, University of Nottingham, presented at the Fist Riverboat Conference on Bayesian Econometrics and Statistics.

Phillips, P.C.B. and W. Ploberger(1992) "*Posterior odds resting for a unit root with data-based model selection*" Research Paper, Cowles Foundation.

Schwarz, G (1978) "Estimating the dimension of a model." *Annals of Statistics*. 6, 461-464.

Theil, H.(1970) "*Principles of Econometrics*". North Holland, Amsterdam.

1992-1	R.J. Boucherie N.M. van Dijk	Local Balance in Queueing Networks with Positive and Negative Customers
1992-2	R. van Zijp H. Visser	Mathematical Formalization and the Analysis of Cantillon Effects
1992-3	H.L.M. Kox	Towards International Instruments for Sustainable Development
1992-4	M. Boogaard R.J. Veldwijk	Automatic Relational Database Restructuring
1992-5	J.M. de Graaff R.J. Veldwijk M. Boogaard	Why Views Do Not Provide Logical Data Independence
1992-6	R.J. Veldwijk M. Boogaard E.R.K. Spoor	Assessing the Software Crisis: Why Information Systems are Beyond Control
1992-7	R.L.M. Peeters	Identification on a Manifold of Systems
1992-8	M. Miyazawa H.C. Tijms	Comparison of Two Approximations for the Loss Probability in Finite-Buffer Queues
1992-9	H. Houba	Non-Cooperative Bargaining in Infinitely Repeated Games with Binding Contracts
1992-10	J.C. van Ours G. Ridder	Job Competition by Educational Level
1992-11	L. Broersma P.H. Franses	A model for quarterly unemployment in Canada
1992-12	A.A.M. Boons F.A. Roozen	Symptoms of Dysfunctional Cost Information Systems
1992-13	S.J. Fischer	A Control Perspective on Information Technology
1992-14	J.A. Vijlbrief	Equity and Efficiency in Unemployment Insurance
1992-15	C.P.M. Wilderom J.B. Miner A. Pastor	Organizational Typology: Superficial Foursome of Organization Science?
1992-16	J.C. van Ours G. Ridder	Vacancy Durations: Search or Selection?
1992-17	K. Dzhaparidze P. Spreij	Spectral Characterization of the Optional Quadratic Variation Process
1992-18	J.A. Vijlbrief	Unemployment Insurance in the Netherlands, Sweden, The United Kingdom and Germany