NOTES AND CORRESPONDENCE

Significance Tests for Regression Model Hierarchies

T. P. BARNETT, R. W. PREISENDORFER, L. M. GOLDSTEIN AND K. HASSELMANN

Climate Research Group, Scripps Institution of Oceanography, La Jolla, CA 92093

29 May 1981

ABSTRACT

Methods of estimating the significance of optimal regression models selected from a model hierarchy proposed by Barnett and Hasselmann (1979) are reexamined allowing for the multiple-candidate nature of the selection criteria. It is found that the single-candidate models' significance value previously used can over- or underestimate the true multiple-candidate significance level of the selected model depending on the selection criteria used. A number of possible selection strategies to remove these problems are discussed and evaluated both theoretically and by Monte Carlo simulations.

1. Introduction

In a recent paper Barnett and Hasselmann (1979, hereafter called BH) discussed methods for constructing linear-regression models that balanced the competing requirements of skill and statistical significance. The methods are designed to select an optimum model from a model hierarchy of increasing order. The selection decision was based on the individual statistical significance of each member of the hierarchy. However, the statistical significance of the model actually selected will in general differ from the individual significance value computed for the model since the selection procedure represents a multivariate test of statistically dependent candidate models, whereas the individual significance values correspond to a single model test. The actual significance level of the selected model can be greater or smaller than its single-model significance measure depending on the selection criteria used. This fact was ignored by BH and the purpose of the present note is to elucidate the relation between the actual multicandidate significance level of the selected model (hereafter referred to as c') and its single-model significance measure (c) for a number of possible model selection criteria.

The significance of a set of estimated regression-model coefficients (a_i) is evaluated by BH with respect to the null hypothesis of a true model with zero coefficients ($a_i^0 \equiv 0$). The covariance matrix $M_{ij} = \langle \delta a_i \delta a_j \rangle$ of the differences, $\delta a_i = a_i - a_i^0$, between the coefficients a_i estimated from a finite realization and the true model coefficients a_i^0 is assumed known (although in practice it must be estimated). The distribution of errors δa_i is assumed Gaussian.

The appropriate test statistic characterizing the significance of the model is given by

$$\rho_n^2 = \sum_{i,j=1}^{n} M_{ij}^{-1} \delta a_i \delta a_j. \tag{1}$$

The variable ρ^2 presents a χ^2 variable with n degrees of freedom. ρ^2 is invariant with respect to linear transformations, and rotation of (1) into a diagonal form yields the sum of n squares. The model of order n was therefore regarded as statistically significant at confidence level c if ρ_n^2 exceeds the appropriate critical value $\chi_{n,c}^2$. We note that the statistic ρ^2 represents the optimal quadratic form for testing significance in the sense that it defines the smallest volume in a_i space for a given probability volume. Also, the surfaces ρ^2 = constant represent surfaces of constant probability density.

To arrive at a suitable balance between the competing requirements of achieving both model skill and significance, the significance test was applied by BH not to a single model, but to a nested hierarchy of models. The model selection criterion was proposed in which the model with the largest order n was chosen for which model n and all lower order models (m) exceeded at prescribed confidence level c: $\rho_m^2 > \chi_{m,c}^2$ for $1 \leq m \leq n$, $\rho_{n+1}^2 \leq \chi_{n+1,c}^2$. This criterion, in general, is rather stringent and requires particularly good *a priori* insight in choosing the ordering of the lowest predictors. Theoretical analysis and Monte Carlo tests show that the multicandidate significance level (c') of the selected highest-order model is normally considerably higher than the single-model significance level (c) (cf. Rule B below and Table 1).

In practice, however, BH tacitly relaxed the re-

0022-3670/81/081150-05$05.25

TABLE 1. Multi-candidate significance (c') vs model order (N') for a single-candidate significance of 0.90.

N'	c'
1	0.89987
2	0.94540
3	0.96507
5	0.97831
7	0.98328
10	0.98787

straint $\rho_m^2 \geq \chi_{m,c}^2$ for all $1 \leq m \leq n$ and accepted models in which the first few members of the heirarchy failed to satisfy the significance criteria. Theoretical analysis and Monte Carlo tests have shown that this imprecise procedure generally yields significance levels c' that are smaller than c and depend on the number of members in the hierarchy that are allowed to fall below the critical χ^2 value. We are indebted to Russ Davis for pointing this out to us.

2. Linear model hierarchies and optimal model-selection criteria

BH first orthogonalized the predictors and then ordered them with respect to variance. Next, all predictors were discarded that could not be distinguished from white noise, using the method of Preisendorfer and Barnett (1977) (see also Preisendorfer *et al.*, 1981). This yielded a maximum number of N predictors, $z_1, z_2, \ldots, z_N$, and hence the *maximum model order*. The model hierarchy was defined as a set of N models constructed from the first predictor, the first two predictors, . . . , all N predictors. This, or any other, means of constructing a model hierarchy is largely subjective, but must be made *a priori* in order to carry out meaningful significance tests.

The set of N values, ρ_n^2, characterizing the individual significance of each model of the hierarchy was then calculated and formed the basis for selecting the "optimal" model. The *a priori* selection criterion is again essentially arbitrary, reflecting the modelers' individual choice of tradeoffs between significance and skill. However, once the selection criterion is specified, the statistical significance relative to the null hypothesis of a zero prediction model ($a_i^0 = 0$) can be evaluated by Monte Carlo simulations, or in special cases, analytically. It is a straightforward matter to generalize the procedure to models which have $a_i^0 \neq 0$ (cf. Preisendorfer, 1979). Examples where this may be useful are given later. Some examples of how the ρ^2 might evolve with respect to the $\chi_{n,c}^2$ are shown in Fig. 1 and will be useful in a later discussion.

In the Monte Carlo simulations to be discussed below the covariance matrix was assumed to be diagonalized so that the statistic ρ_n^2 reduces to

$$\rho_n^2 = \sum_{i=1}^{n} a_i^2 \lambda_i^{-1}, \tag{2}$$

where the λ_i are eigenvalues of M_{ij}. The members of the ρ^2 set for a given data realization were then

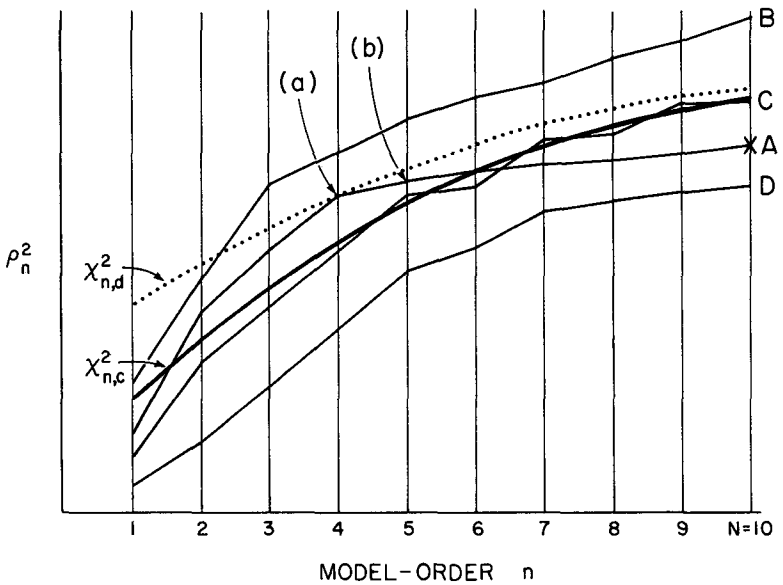

FIG. 1. Sample model paths in the space of multi-candidate model paths for $N = 10$. Two χ^2-curves for significance levels of c and d also are shown. See Section 3 for a discussion of this figure.

TABLE 2. Single-candidate (c) versus multi-candidate (c') confidence levels for various model order (N).

Single candidate confidence	Model order									
	1	2	3	4	5	6	7	8	9	10
0.9000	0.90053	0.85490	0.82616	0.80298	0.78630	0.77143	0.75858	0.74719	0.73789	0.72947
0.9100	0.90961	0.86720	0.83874	0.81842	0.80324	0.78807	0.77638	0.76694	0.75765	0.74965
0.9200	0.92202	0.88342	0.85893	0.84003	0.82517	0.81332	0.80332	0.79418	0.78563	0.77805
0.9300	0.93013	0.89631	0.87333	0.85575	0.84262	0.83237	0.82250	0.81348	0.80583	0.79885
0.9400	0.94010	0.91054	0.89054	0.87570	0.86321	0.85305	0.84388	0.83596	0.82906	0.82244
0.9500	0.95064	0.92482	0.90838	0.89505	0.88454	0.87544	0.86827	0.86122	0.85491	0.84975
0.9600	0.96024	0.93939	0.92503	0.91384	0.90566	0.89802	0.89136	0.88572	0.88072	0.87571
0.9700	0.97100	0.95584	0.94472	0.93592	0.92873	0.92261	0.91746	0.91281	0.90879	0.90457
0.9750	0.97482	0.96151	0.95213	0.94464	0.93882	0.93362	0.92936	0.92511	0.92136	0.91831
0.9800	0.97997	0.96893	0.96096	0.95441	0.94902	0.94467	0.94099	0.93763	0.93482	0.93205
0.9850	0.98601	0.97686	0.97110	0.96655	0.96228	0.95868	0.95555	0.95285	0.95065	0.94884
0.9900	0.98993	0.98418	0.98013	0.97677	0.97410	0.97164	0.96931	0.96756	0.96605	0.96435
0.9950	0.99525	0.99240	0.99004	0.98809	0.98661	0.98849	0.98434	0.98327	0.98243	0.98144
0.9975	0.99761	0.99613	0.99494	0.99385	0.99295	0.99214	0.99151	0.99102	0.99055	0.99002
0.9999	0.99941	0.99842	0.99806	0.99768	0.99738	0.99700	0.99677	0.99664	0.99650	0.99628

obtained by drawing n independent normal random numbers for each coefficient a_i, squaring the coefficients and summing. Because of the invariance of ρ_n^2 to linear transformation, the resulting significance tables (e.g., Table 2) generated by the Monte Carlo simulations apply generally for any covariance matrix M_{ij}.

We ignore here complications arising from the fact that the error-covariance matrix is not known precisely but must also be estimated from the finite data realization. The estimation uncertainties in the λ_i are generally small, if N is small compared with the effective number of independent data samples N_s. This was the case for the examples considered by BH. More careful consideration is required when N becomes of order N_s.

The following selection criteria have been considered.

a. Rule A: Unordered predictor set

In this case the optimal model is taken as the maximum-order model N. This approach considers only the maximum number of predictors, paying no attention to their ordering, and places substantial importance on the objective rules used to determine N. After filtering out predictors that cannot be distinguished from white noise, one simply constructs the model with the remaining predictors. The significance of the selected model in this case is identical to the significance measure inferred from the single-model test statistic ρ_N^2, i.e., $c' = c$.

The inherent danger of using this technique is that for large N the model will be normally insignificant. However, lower-order significant models of the hierarchy may have been constructed if a physically plausible, *a priori* hypothesis had been introduced to define the more important predictors of the set. This potential problem is illustrated in Fig. 1 by curve A. For the maximum model order N ($=10$) the value of ρ^2, shown by the "$\times$", is below $\chi^2_{N,c}$ and model would be rejected at the c significance level. However, for $2 \leq n \leq 5$, ρ_n^2 is above $\chi^2_{n,c}$ indicating that a lower order model may have been accepted. Methods for determining this situation are given below (cf. Rule C).

b. Rule B: Sequential fixed-significance-level selection criterion

In this case the optimal model is defined as the highest order model N', satisfying $\rho_m^2 \geq \chi^2_{m,c}$, for all m in the interval $1 \leq m \leq n$; $n \leq N$. This corresponds to the stringent form of the selection criterion proposed by BH and corresponds to curve B in Fig. 1. For orthogonal errors δa_i the probability of the null model satisfying this condition at the order n or greater is given by

$$c' = \int_{b_1}^{\infty}\int_{b_2}^{\infty}\cdots\int_{b_n}^{\infty} f(u_1)f(u_2)\cdots f(u_n)du_n du_{n-1}\cdots du_1, \quad (3)$$

where

$$b_n = d_n - \sum_{j=1}^{n-1} u_j, \quad b_1 = d_1, \quad (4)$$

and where

$$f(u) = [2^{1/2}\Gamma(\tfrac{1}{2})]^{-1}u^{-1/2}e^{-u/2}. \quad (5)$$

is the probability density of a χ^2-variate with one degree of freedom.

For large n the computation of c' according to Eq. (3) becomes rather tedious and it is simpler to determine by Monte Carlo simulations. Table 1 shows an example of the dependence c' on c and n.

A selection criterion requiring $\rho_m^2 \geq \chi_{m,c}^2$ for all m in the range $1 \leq m \leq n$, $n \leq N$ is appropriate only if the modeler has high confidence in his ordering of predictors. In this sense, the criterion may be regarded as the antithesis of Rule A, which is appropriate if no ordering can be devised. In most practical applications selection criteria intermediate between these two limits will be most useful. Two examples of such selection criteria are given below.

c. Rule C: Nonsequential fixed-significance-level selection criterion

The optimal model is taken to be the highest-order model n satisfying $\rho_n^2 \geq \chi_{n,c}^2$ in the range $1 \leq n \leq N$. This criterion corresponds approximately to that actually applied by BH. In this case the model associated with curve A (Fig. 1) would now be accepted as significant. The last ρ^2 value that exceeds χ_c^2 occurs at $n = 5$ [point (b)] and so the model order $N' \equiv 5$. Similarly, model C would be accepted with order 9. Model D which never exceeds the critical curve would be rejected.

The probability c' of finding a significant model for given single-candidate confidence level c in a hierarchy of N models is given by

$$c' = 1 - \int_0^{b_1} \cdots \int_0^{b_N} f(u_1) \cdots f(u_N) du_N \cdots du_1. \quad (6)$$

The relation between c' and the parameter c and N are shown in Table 2 and Fig. 2. For $n \geq 4$ the Monte Carlo computations again prove more efficient than the direct integration of (6). A comparison between the theoretical and Monte Carlo estimates of c' (Table 3) shows excellent agreement for the two sets of calculations.

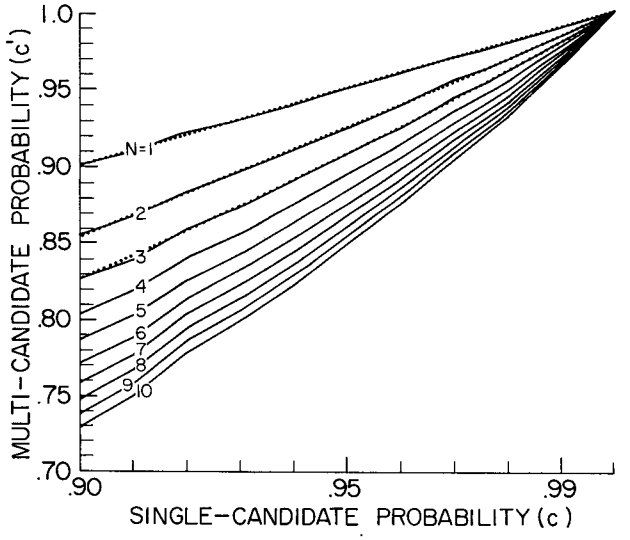

FIG. 2. Single-candidate versus multi-candidate confidence levels for various model orders (N). The solid line is from the Monte Carlo simulation while the dotted line came from exact analytical solutions of Eq. (6).

TABLE 3. Comparison of theoretical (Th) and Monte Carlo (MC) estimates of multi-candidate confidence for regression models up to order 3.

Confidence	Model order 1		2		3	
	MC	Th	MC	Th	MC	Th
0.90	0.9005	0.9000	0.8549	0.8451	0.8262	0.8242
0.95	0.9506	0.9500	0.9248	0.9249	0.9084	0.9077
0.99	0.9899	0.9900	0.9842	0.9843	0.9801	0.9801

As noted earlier, the above test, like Rules A and B and D below, can be generalized to include models assumed to have $a_i^0 \neq 0$ and to be additively perturbed by Gaussian noise. The subsequent values of ρ^2 may now be compared against a noncentral χ^2 distribution (cf. Preisendorfer, 1979) and so the test procedure, evaluation of c', etc., go through as above. Applications of these features of the test might include the significance testing of a given theoretical model against a data set or in estimating the degree of similarity between two (theoretical) models competing to represent a given data base (cf. BH, Section 7).

d. Rule D: Maximal significance model

The optimal model in this case is defined as the model that is associated with the highest single-candidate significance level d as determined by the equality $\rho_n^2 = \chi_{n,d}^2$. This situation is illustrated on Fig. 1 by the point (a) on curve A. This choice represents a tradeoff which emphasizes significance relative to skill, whereas the previous choice (Rule C) yields higher hindcast-skill models at smaller significance levels, e.g., point (b), Fig. 1. However, the relation between c' and c is the same for both selection criteria. In both cases, for given fixed single-candidate confidence level c, a given data realization will yield an accepted model if at least one of the set of variables $\rho_1^2, \ldots, \rho_N^2$ exceeds its associated significance level $\chi_{1,c}^2, \chi_{2,c}^2, \ldots, \chi_{N,c}^2$; the two selection criteria differing only in the choice of model order. The differences in the statistical distribution of optimal models selected by the two criteria appear only if the two-dimensional probability densities with respect to both c and selected model order n are considered.

e. Other possible rules

A number of further selection criteria intermediate to the option for maximal single-candidate significance or maximal skill for given single-candidate significance can be obtained by consider-

ing the incremental changes in significance associated with the addition of individual predictors. Here again it is important that the selection criteria and the ordering of predictors be defined *a priori* since an *a posteriori* screening of the full set of N predictors automatically ensures that the significance of the selected model is of the same order as the significance of the full N-predictor model (cf. Rule A). Provided the selection rules require, as before, that at least one of the set of variables $\rho_1^2, \ldots, \rho_N^2$ exceeds its associated single-candidate significance value at a given confidence level c', then the same relations between c and c' apply as in Rules C and D.

3. Summary

The problem of estimating the significance and order of linear regression models has been reexamined. Several new significance tests are proposed to replace those originally suggested by Barnett and Hasselmann (1979) since the latter tests either underestimated or overestimated model significance. The newer tests and the tables needed to apply them, given here in detail, offer the potential user alternative approaches to the construction of regression-model hierarchies.

Acknowledgments. Support for this work was made available by National Science Foundation ATM-7918206 (TB, LG), NOAA-PMEL (RP) and the Max Planck Society through Max Planck Institute for Meteorology (KH). The authors are indebted to Russ Davis for pointing out the overestimation property of the original BH significance test.

REFERENCES

Barnett, T. P., and K. Hasselmann, 1979: Techniques of linear prediction, with application to oceanic and atmospheric fields in the tropical Pacific. *Rev. Geophys. Space Phys.* **17**, 949–968.

Preisendorfer, R. W., 1979: Model skill and model significance in linear regression hindcasts. SIO Ref. Ser. 79-12, Scripps Institution of Oceanography, 75 pp.

——, and T. P. Barnett, 1977: Significance tests for empirical orthogonal functions. *Preprints Fifth Conf. Probability and Statistics*, Las Vegas, Amer. Meteor. Soc., 169–172.

——, R. W. Zwiers and T. P. Barnett, 1981: Foundations of principal component selection rules. SIO Ref. Ser. 81-4, Scripps Institution of Oceanography (in press).