A Hybrid Test for Non-Nested Models

Paul Wilson[1]

[1] Department of Mathematics, National University of Ireland, Galway

Abstract: Compared to tests for nested models, little attention has been given to methods that test the log-likelihood ratios of non-nested models. We outline two such methods, Cox's and Vuong's, highlighting the advantages and disadvantages of both. We propose a hybrid test that combines the advantages both methods, without their disadvantages.

Keywords: Non-nested models; Cox's test; Vuong's test; model discrimination.

1 Introduction

Standard statistical theory provides us with a range of tools for choosing between nested models. However, in many practical data analysis problems we wish to choose between non–nested models, i.e. models where neither model is a special case of the other. The problem of choosing between non-nested models arises in many areas of scientific research. Recent examples include, in environmental science, Dobbie and Welsh (2001); in agricultural science, Allcroft and Glasby (2003); and in political science, Smith (1999).
This problem was first considered by Cox (1961,1962), who developed an analytic test; a further analytic test was later considered by Vuong (1989). Williams (1970) and Hinde (1992) have proposed simulation based alternatives to Cox's approach.

2 Cox's Method

Let M_f and M_g be non-nested models for observations Y_t, $t = 1, 2, \ldots, n$, conditional on covariates X_t and Z_t, and with parameters θ and γ, respectively. Let $\hat{\theta}$ and $\hat{\gamma}$ be the maximum likelihood estimators of θ and γ, and let $\hat{\gamma}_\theta$ be the maximum likelihood estimator of γ if, in fact, M_f is the correct model. Let $LR_n(\hat{\theta}_n, \hat{\gamma}_n)$ be the log–likelihood ratio: $\sum_{t=1}^{n} \log \frac{f(Y_t|X_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}$.
Cox's test is based on the statistic:

$$T_f = \left\{ LR_n(\hat{\theta}_n, \hat{\gamma}_n) - E_f\Big(LR_n(\hat{\theta}_n, \hat{\gamma}_{\theta_n}) \Big) \right\} \tag{1}$$

i.e. the difference between the observed and the expected values of the log-likelihood ratio of the data where the null hypothesis, H_f, is that M_f is the true model, as opposed to M_g.

Cox (1962) shows that under H_f, T_f is asymptotically normally distributed with approximate mean zero and variance

$$n\left\{V_f\left(\log\frac{f\left(Y_t|X_t;\hat{\theta}_n\right)}{g\left(Y_t|Z_t;\hat{\gamma}_{\theta_n}\right)}\right)-\frac{C_f^2\left(\log\frac{f\left(Y_t|X_t;\hat{\theta}_n\right)}{g\left(Y_t|Z_t;\hat{\gamma}_\theta\right)},\frac{\partial}{\partial\theta}\log f\left(Y_t|X_t;\hat{\theta}_n\right)\right)}{V_f\left(\frac{\partial}{\partial\theta}\log f\left(Y_t|X_t;\hat{\theta}_n\right)\right)}\right\} \tag{2}$$

where V_f is the expected value of the variance under H_f, C_f is the expected value of the covariance under H_f, and $\frac{\partial}{\partial\theta}\log f\left(Y_t|X_t;\hat{\theta}_n\right)$ is the score statistic under H_f. We may therefore calculate the p-value associated with a given value of T_f, "small" p-values indicating rejection of T_f.
Similarly, reversing the roles of f and g and θ and γ in the above we may calculate the p-value associated with a given value of T_g. Combining these two results we obtain the range of conclusions summarised by Table 1.

TABLE 1. Possible outcomes of Cox's test

		$H_0: M_f$ is the true model		
	p-value	*small*	*medium*	*large*
	small	Neither	M_f	Neither
$\underline{H_0: M_g}$	*medium*	M_g	Both	M_g
	large	Neither	M_f	–

Cox's method may be performed both analytically or by simulation. Analytic evaluation can be complicated, while simulation requires the refitting of the model for each resample and this can be very time-consuming.

3 Vuong's Test

Vuong's test considers the null hypothesis:

$$H_0: E\left[LR_n\left(\hat{\theta}_n,\hat{\gamma}_n\right)\right]=0 \tag{3}$$

i.e. that the expected value of the log–likelihood ratio under H_0 is zero, (and hence models M_f and M_g are equivalent). The alternative hypotheses are thus that M_f is "better" than M_g and vice-versa. The variance of LR_n can be estimated by the empirical variance:

$$\omega_n^2\equiv\frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f\left(Y_t|X_t;\hat{\theta}_n\right)}{g\left(Y_t|Z_t;\hat{\gamma}_n\right)}\right]^2-\left[\frac{1}{n}\sum_{t=1}^{n}\log\frac{f\left(Y_t|X_t;\hat{\theta}_n\right)}{g\left(Y_t|Z_t;\hat{\gamma}_n\right)}\right]^2 \tag{4}$$

Vuong shows that, under fairly general conditions, $\frac{LR_n\left(\hat{\theta}_n,\hat{\gamma}_n\right)}{\omega_n\sqrt{n}} \xrightarrow{\text{D}} N(0,1)$ under the null hypothesis, and to $\pm\infty$ otherwise.
Vuong's test is undoubtedly quick and simple to execute. It is however very conservative, (see, for example, Table 2), and in those cases where the null hypothesis is rejected, this only indicates that, say, M_f is preferable to M_g, not necessarily that M_f is suitable. (See Tables 3 and 4).

4 The Hybrid Test

Clearly it would be advantageous to develop a hybrid of Cox's and Vuong's test that combines the ease of use of the latter with the accuracy of the former. We do this by replacing the single null hypothesis of Vuong with the double null hypotheses of Cox, and adjusting equation (4) to be the expected value of the variance under each null hypothesis respectively. This is equivalent to applying the analytic version of Cox's test with the right-hand term of (2) omitted when calculating the variance, or applying a simulation–based Cox's test where the model parameters are *not* refitted at each resample. Given that the requirement to refit the parameters of each resample is by far the most time consuming aspect of simulation-based versions of Cox's test, the practical benefits of not having to do so are enormous. Such a "resampling without refitting" approach has been used by Allcroft and Glasby(2003). Given that the variance of the hybrid test is greater than that of Cox's test, it is necessarily more conservative than Cox's test. Clearly the larger the right-hand term of equation (2) relative to the left-hand term, the greater the conservatism of the hybrid test when compared to Cox. Note that the ratio of the right and left hand terms of (2) is the ratio of the expected value of the covariance of the log-likelihood ratio and the score statistic under H_f to the product of the expected values of their variances. We denote the value of this correlation coefficient type statistic by r_f^2 for $H_0 : M_f$ and r_g^2 for $H_0 : M_g$. In general, these two values are different, indicating that the conservatism of the hybrid test is not symmetrical. Note that for the example of Table 4 below r_g^2 (H_0:*geometric*) is more than five times greater than r_p^2 (H_0:*Poisson*), and hence the hybrid test is more conservative when rejecting a geometric distribution than it is when rejecting a Poisson distribution, in relation to Cox's test. Tables 2 to 4 each show the results obtained when the three tests were each used to classify 1,000 samples of size 50 taken from data that followed a geometric distribution with mean 0.8, a binary distribution consisting of equal numbers of zeros and ones, and a geometric distribution with mean 6, respectively. For example, Cox's test classified 35 of the $1,000$ samples taken from geometric(0.8) data as Poisson, 764 as geometric, 158 as possibly both, and 43 as neither. We see that the hybrid test performs well in relation to Cox's test, and considerably better than Vuong's test.

5 Examples

We further illustrate the usefulness of the hybrid test by comparing its performance with that of the simulation-based version of Cox's test, (1,000 resamples), and Vuong's test, when applied to two "real life" data sets. Please note that the purpose of these

TABLE 2. Classification of 1,000 samples drawn from Geometric(0.8)

Test	Pois	Geom	Both	Neither
Vuong	5	223	768	–
Cox	35	764	158	43
Hybrid	24	755	181	40
mean value of $r_p^2 = 0.105$				
mean value of $r_g^2 = 0.147$				

TABLE 3. Classification of 1,000 samples drawn from binary distribution

Test	Pois	Geom	Both	Neither
Vuong	1000	0	0	–
Cox	4	0	0	996
Hybrid	4	0	0	996
mean value of $r_p^2 = 0.092$				
mean value of $r_g^2 = 0.109$				

TABLE 4. Classification of 1,000 samples drawn from Geometric(6)

Test	Pois	Geom	Both	Neither
Vuong	0	995	5	–
Cox	0	907	0	93
Hybrid	0	962	0	38
mean value of $r_p^2 = 0.061$				
mean value of $r_g^2 = 0.309$				

examples is to emphasise the merits of the hybrid test for comparing models, not to determine a suitable model.

Firstly, we consider data from Leroux and Puterman (1992) that gives the number of movements made by a fetal lamb in each of 240 consecutive 5-second intervals (Table 5). Clearly this data is overdispersed, hence two possible "candidate" models are the negative-binomial (Poisson-Gamma) and the Neyman-A (Poisson-Poisson). Due to the presence of infinite sums in its probability density function, algorithms for fitting Neyman-A models are slow, even in the absence of covariates. As shown in Table 5, the (elapsed) time taken to complete the hybrid test is over 150 times less than that of Cox's test, both tests concluding that, at $\alpha = 0.05$, we may reject the Neyman-A model in favour of the negative binomial. Note that whilst Vuong's test is practically instantaneous, it fails to distinguish between the models.

Secondly, we look at data from Ridout, Demétrio, and Hinde (1998) describing the number of roots produced by 270 micropropagated shoots of the apple cultivar *Trajan*. Two covariates were present. *Period*, at 2 levels, and *Hormone* at 4. Ridout et al. show

TABLE 5. p-values for Fetal Lamb Data

Vuong (< 1 second)		Cox (143 minutes)		Hybrid (52 seconds)	
Neyman–A	Neg–Bin	Neyman–A	Neg–Bin	Neyman–A	Neg–Bin
0.183	0.817	0.019	0.863	0.025	0.828
$r^2_{NA} = 0.101$			$r^2_{NB} = 0.082$		

that *hormone* has little effect. In general, the presence of covariates increases the time taken for model-fitting. This is not of major consequence if the number of covariates in the model is small and efficient algorithms for fitting the model in question exist. For example, a comparison of two zero-inflated negative-binomial models:
roots $\sim$ *period* and *roots* $\sim$ *hormone*, (where both the mean and the over–inflation parameter are fitted by the given covariate), where both models were fitted by the R-package *Zicounts*, took approximately 12 minutes to complete for Cox's test, as opposed to approximately 11 seconds using the hybrid test. If many covariates are present, or efficient model-fitting algorithms do not exist, then Cox's test may prove impractical. An example is that of Table 6 which illustrates the results obtained when the three tests were used to compare Neyman-A and zero-inflated Poisson models of the form *roots* $\sim$ *period*, (all parameters varying over *period*). Vuong's test failed to reject either model. Cox's test proved impractical: less than a quarter of the 1,000 resamples had occurred after two days, and the test was abandoned, whereas the hybrid test completed in under 2 minutes, rejecting both models.

TABLE 6. p-values for Trajan Apple Data, (roots$\sim$period)

Vuong (< 1 second)		Cox (estimate: 9 days)		Hybrid (117 seconds)	
Neyman–A	ZIP	Neyman–A	ZIP	Neyman–A	ZIP
0.506	0.494	—	—	0.018	0.000
$r^2_{NA} = 0.101$			$r^2_{ZIP} = 0.082$		

6 Conclusion

The hybrid test is a suitable alternative to Cox's test and Vuong's test, being much quicker than the former, and more decisive than the latter. The results above are dependent upon models being non–nested, and the extension of the hybrid test to models where this is not the case is a possible area of future research.

Acknowledgement

The Author wishes to thank Prof. John Hinde for the generous support and assistance he provided with the preparation of this paper.