The Dragnet Test: A New Approach to Choosing Between Models

Paul Wilson[1]

[1] Department of Mathematics, National University of Ireland, Galway
Paul.Wilson@nuigalway.ie

Abstract: Traditional log-likelihood based methods for choosing between models, be they nested or non-nested, all concentrate on log-likelihoods evaluated at the maximum likelihood estimates of the model parameters. The true model parameters may in fact differ considerably from their maximum likelihood estimates. We propose a method that examines the relative fits of the models over a cross-section of likely parameter values; this method is based upon testing simple hypotheses, and hence avoids pitfalls associated with compound null hypotheses such as biased estimation of p-values.

Keywords: Model Discrimination, Hypothesis Testing, Cox's test, nested models, non-nested models, hybrid test.

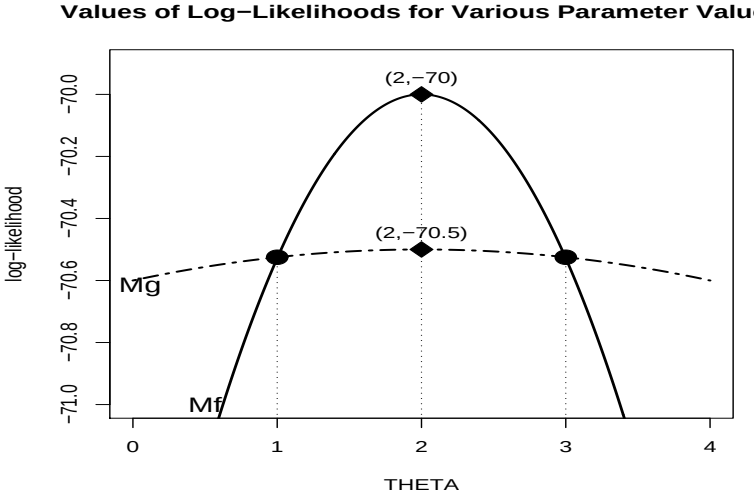
1 Introduction

In statistical analysis the substitution of a maximum likelihood estimator, $\hat{\tau}$, for the true, but unknown, parameter τ_0 of a model for given data, is so commonplace that possible consequences of the fact that the values of τ_0 and $\hat{\tau}$ may differ considerably are often overlooked.

Consider the situation illustrated in Figure 1, which plots, for some (fictional) data, the log likelihoods of two models M_f and M_g, where both models are functions of the same single-valued parameter, θ. We see that, for both models, the maximum-likelihood estimate of θ occurs at $\hat{\theta} = 2$, thus as $\ell_{M_f}(2) > \ell_{M_g}(2)$ M_f is to be preferred over M_g *assuming that* $\hat{\theta}$ *is the "true" value of* θ. Say the "true" value of θ is 0.9, we see from Figure 1 that $\ell_{M_g}(0.9) > \ell_{M_f}(0.9)$, and hence, for $\theta = 0.9$, M_g is to be preferred to M_f. Thus, whilst M_f is clearly the "better" model if it is highly likely that the true parameter value lies between 1 and 3, it is far from clear which model is "better" if there is a reasonable chance that the true parameter value lies outside of this interval.

Here, when we say that we "prefer", say, M_f to M_g at $\theta = \theta_\star$, we simply mean that the log likelihood of the former, evaluated at $\theta_\star$ is greater than that of the latter, (also evaluated at $\theta_\star$). For any test to be of practical use we need to determine criteria that determine whether we may reject:

FIGURE 1. Possible log likelihood Values



$$H_0 : M_g(\theta_\star) \text{ is a suitable model for the data} \tag{1}$$

against:

$$H_1 : M_f(\theta_\star) \text{ is a suitable model for the data} \tag{2}$$

The criteria we adopt for the dragnet test are basically those of the standard Cox test for non-nested models, (Cox (1962)): we reject the null hypothesis if the observed log likelihood ratio is inconsistent with what would be expected if the null hypothesis were true, rejection being possible both towards and away from the alternative hypothesis. We then reverse the hypotheses, and repeat the procedure. This results in two p-values, one for each null hypothesis, from which we may classify the models as illustrated in Table 1. Thus, unlike conventional log-likelihood based methods, or score tests, which merely determine if one model is significantly better than another, not whether it is suitable, the Cox test determines whether one or other, either, or both of the two models under consideration are appropriate. This desirable property is also incorporated into the dragnet test.

The test proposed in Cox (1962) was analytic. Following Williams (1970) and Hinde (1992), who proposed simulation based analogues of Cox's test, the various p-values of the dragnet test are estimated by bootstrap methods. Whereas the Cox test only evaluates "inconsistency" at the maximum likelihood estimate of the parameters of the models concerned, the dragnet test

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
$H_0 : M_g$	*medium*	M_g	Both	M_g
	large	Neither	M_f	–

evaluates it at S possible *fixed* parameter values determined by sampling from the parameter spaces of both models, thus obtaining a weighted cross section of possible parameter values. With regard to the testing of models at fixed parameters, the dragnet test may be viewed as an extension of the hybrid test proposed in Wilson (2007), which could be regarded as a dragnet test where the dragnet consists solely of the maximum likelihood estimate of the model parameters. Hence, as the hypotheses of the dragnet test are simple, i.e. they specify the parameters of M_f and M_g, problems with bias estimation of p-values are avoided. (See Wilson (2008)). This fixing of parameters also enables Cox's method to be extended to nested or overlapping models.

2 Example: Zero-Inflated Poisson versus Poisson Models

The dragnet test, with and $S = 1,000$ was used to analyse the random sample of data summarised in Table 2. This sample was drawn from $ZIP(0.1, 2)$ data.

TABLE 2. $ZIP(0.1, 2)$ data

Value	0	1	2	3	4	5	6	≥ 7	Total
Count	9	8	13	9	9	0	2	0	50

When a zero-inflated Poisson model is fitted to these data, parameter estimates $\hat{\gamma} = 0.100$ and $\hat{\lambda} = 2.423$ are obtained. A score test returns a p-value of 0.078, not enabling the rejection of H_0 :*Poisson* at $\alpha = 0.05$. Table 3 describes the overall classification, at $\alpha = 0.05$, by a ZIP dragnet test (i.e. where the cross-section of parameter values used to determine the dragnet assumes a ZIP distribution), and a Poisson dragnet test.
We see that the ZIP dragnet favours the ZIP model to the Poisson at 0.780 of likely parameter values, and indicates that if a zero modifiedinflated

TABLE 3. Classification of the Table 2 data, $\alpha = 0.05$

$S = 1,000$	ZIP	Poisson	Both	Neither
ZIP dragnet	0.780	0.018	0.110	0.092
Poisson Dragnet	0.188	0.070	0.025	0.217

Poisson distribution is not suitable, then probably neither is a Poisson distribution. If we examine the classification of the Poisson dragnet not only is there is little support for the Poisson model, but there is no particular support for any classification. Overall, the evidence appears to support the ZIP model, but is not conclusive.

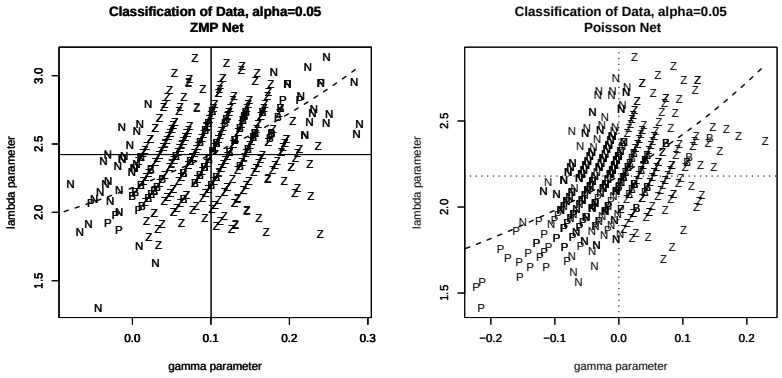
2.1 Dragnet Classification Diagrams

The *dragnet classification diagrams* of Figure 2 illustrates the classification, at $\alpha = 0.05$, of the data at the various parameter values. Letter Z's indicate a "ZIP" classification, P's a "Poisson" classification, B's a classification of "both", N's a "neither" classification. The "intermingling" of "ZIP" and 'both" classifications in some regions, and of "Poisson" and 'neither" classifications in others indicates that these classifications are borderline in these regions. The solid black lines correspond to the maximum likelihood estimators for the ZIP model, and the dotted black lines to those of the Poisson. We see that in the vicinity of the ZIP maximum likelihood estimator for the ZIP model the data is classified as ZIP/both, (indicating that $H_0 : ZIP$ is not rejected in this vicinity, but that for H_0 :Poisson, $p \approx 0.05$ and hence H_0 :Poisson is borderline accepted or rejected) whereas in the vicinity of the maximum likelihood estimators of the Poisson model all four possible classifications occur, indicating that *both* H_0:*Poisson* and $H_0 : ZIP$ are borderline accepted/rejected in the immediate vicinity of the Poisson maximum likelihood estimator. Given that if a model is true one would expect stability in the vicinity of the maximum likelihood estimator for the model, this lends support to the ZIP model. Also, along the dashed line representing the locus of the data mean both Poisson and ZIP classification is supported, except at extreme points. As one would expect support for Poisson classification to be strongest along this line, this is further evidence in favour of the ZIP model.

3 A Zero-Inflated Negative Binomial versus a Zero-Inflated Generalised Poisson models

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*

FIGURE 2. ZIP versus Poisson Classification of the Table 2 data for both dragnets at $\alpha = 0.05$



at 4. Ridout et al. fit various standard and zero-inflated Poisson and negative binomial models to the Trajan data, and show that a zero-inflated negative binomial model where both the mean and the zero-inflation parameters are modelled by *period* fits the data well, with a BIC of $1,271.9$, compared to a BIC of $1,283.7$ for the zero-inflated Poisson model.
An alternative, not considered by Ridout at al., is a *zero-inflated generalised Poisson model* based upon the generalised Poisson distribution:

$$f_Y(y;\mu,\phi) = \frac{\mu(\mu+(\phi-1)y)^{y-1}}{y!}\phi^{-y}\exp(-\frac{1}{\phi}(\mu+(\phi-1)y) \qquad (3)$$

Such a model, (fitted using the R package *ZIGP*, Erhardt (2007)), has a BIC of 1270.0, indicating a slightly better fit than the ZINB model. Table 4 presents the results of zero-inflated generalised Poisson and zero-inflated negative binomial dragnet tests.

TABLE 4. ZIGP *versus* ZINB Classification of Trajan data, ZIGP net, $\alpha = 0.05$.

$S = 100$	ZIGP	ZINB	Both	Neither
ZIGP dragnet	0.56	0.01	0.39	0.04
ZINB dragnet	0.19	0	0.81	0

We see that the ZIGP dragnet tends to prefer a "ZIGP" to a "Both" classification, but not overwhelmingly so, whereas the ZINB dragnet strongly favours a "Both" classification, and interesting, favours a "ZIGP" otherwise. This indicates that if a ZINB model is suitable, then so also is a

ZIGP model, but not necessarily vice-versa. We may conclude that, except possibly at some outlying parameters, the ZIGP model is to be preferred.

4 Conclusion

The dragnet test is an exciting new approach to choosing between models. Unlike score-tests or standard (log) likelihood based methods it may be applied to nested, non-nested or overlapping models, and it is not dependent upon the maximum likelihood estimates of the model parameters being close approximations to the true parameters. Unlike analytic or simulation-based Cox tests it is free of bias, but it retains the desirable property of being able to accept or reject both models, as opposed to determining the relative merits of one model in relation to the other.

Acknowledgement

The author wishes to thank Prof. John Hinde for his constructive criticism and feedback concerning the development of the dragnet test.

References

Cox DR. (1962). Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society. Series B* **24**, 406–423.

Erhardt V. (2007), ZIGP: Zero Inflated Generalized Poisson regression models.
www.m4.ma.tum.de/Papers/Czado/Czado-Erhardt-Min-Wagner.pdf

Hinde JP. (1992). Choosing Between Non-nested Models: a Simulation Approach. In Fahrmeir L et al. eds. *Advances in Glim and Statistical Modelling: Proceedings of the Glim92 Conference and 7th International Workshop on Statistical Modelling*. New York: Springer.

Williams DA. (1970), Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures, *Biometrics*, **28**, 23–32.

Wilson P. (2007). A Hybrid Test for Non-Nested Models. In *Proceedings of the 22rd International Workshop on Statistical Modelling*, Barcelona:Universitat Autònoma de Barcelona.

Wilson P. (2008).Bias estimation of p-values in analytic and simulated Cox Tests for non-nested models. In *Proceedings of the 23rd International Workshop on Statistical Modelling*, Utrect: Universiteit Utrecht.