Proceedings of the World Congress on Engineering 2007 Vol II
WCE 2007, July 2 - 4, 2007, London, U.K.

Distributions of Maximum Likelihood Estimators and Model Comparisons

Peter Hingley *

Abstract—**Experimental data need to be assessed for purposes of model identification, estimation of model parameters and consequences of misspecified model fits. Here the first and third factors are considered via analytic formulations for the distribution of the maximum likelihood estimates. When estimating this distribution with statistics, it is a tradition to invert the roles of population quantities and quantities that have been estimated from the observed sample. If the model is known, simulations, normal approximations and p*-formula methods can be used. However, exact analytic methods for describing the estimator density are recommended. One of the methods (TED) can be used when the data generating model differs from the estimation model, which allows for the estimation of common parameters across a suite of candidate models. Information criteria such as AIC can be used to pick a winning model. AIC is however approximate and generally only asymptotically correct. For fairly simple models, where expressions remain tractable, the exact estimator density under TED allows for comparisons between models. This is illustrated via a novel information criterion. Three linear models are compared and fitted to econometric data on patent filings.**

Keywords: AIC, likelihood, model comparison, patent filing, technique for estimator densities

1 Introduction

In the context of data analysis or data mining, a number of indicators are used for assessing data structure and for the comparison of explanatory models [7]. However, a danger is that such measures may be used without understanding the logic behind their construction. When confronted with data from an experiment that contain errors, some basic questions arise. These include, firstly, what (if any) model can be identified. Secondly, how can parameters from a candidate model be best estimated. Thirdly, what are the consequences for the parameter estimates if the chosen model is wrong.

There are practical considerations that should be considered in order to minimise these problems. For example, an experiment should be well designed in a statistical sense by including replication and the sample should span a useful region of the parameter space. It may well be that the model structure is already known from previous experience and that prior information can somehow be incorporated into the analysis. An experiment can also be strictly designed in advance to discriminate between a set of prespecified models via a formal hypothesis testing procedure.

But one may nevertheless encounter unexpected results The data could be unique (e.g. in astronomy) or expensive to reproduce (e.g. samples taken from a nuclear reactor), or broadly consistent with several emergent hypotheses (e.g. in social studies). A statistical toolbox is required to deal with these situations. In this paper, some developments will be described. However this toolbox is not yet complete.

It will be assumed that data are to be analysed by using maximum likelihood estimation under a specified model. The approach will involve studying the distributions of maximum likelihood estimates in terms of their probability density functions (estimator densities). The spectrum of methods that are available for determining these densities will be reviewed. In particular a method will be highlighted that determines the exact estimator density when a distinct model generates the data from the model that is used for estimation. Then there is a discussion of information criteria for comparing models on observed data and associated applications of the exact estimator density.

2 Definitions

Suppose that the members of a sample w_i, $(i = 1, ..., n)$ correspond to a model in the form of a probability distribution $g(w_i|\theta)$, where θ is a $px1$ vector of estimable parameters. The space of w is W, and the space of θ is Θ. $g(w_i|\theta)$ will be considered to be continuous, though simpler equivalents exist for the results when $g(w_i|\theta)$ is discrete. In vector notation, write the sample as $w_{(px1)}$, with likelihood given by the joint density $g(w, \theta)$. The MLE is the value $\hat{\theta}$ of θ that maximises $g(w, \theta)$.

It will be remembered that, in statistical modelling as well as in life, managing expectations is an important thing. $E[h(w)]$ indicates mathematical expectation of a function $h(w)$ of the data w, but possibly refers to the

*European Patent Office, Erhardtstrasse 27, D-80469 Munich, Germany. Email: phingley@epo.org.

data indirectly via the MLE or some other calculable statistic.

$\mathrm{E[h(w)]} = \int_W h(w)g(w)dw$

Useful quantities for inference are obtained from the log likelihood $l(\theta, w) = log(g(w, \theta))$. The first derivative wrt θ is termed the observed score $l'(\theta, w)_{(px1)}$, where $'$ indicates differentiation wrt θ. The MLE then satisfies the normal equation $(l'(\theta, w)|_{\theta=\hat{\theta}}) = 0_{(px1)}$. From the second derivative can be obtained a quantity called the observed information $j(\theta, w)_{(pxp)} = -l''(\theta, w)$. These quantities are *observed* because they depend upon the realisation of a particular sample w. But expected equivalents are also useful, in particular the expected information matrix $i(\theta, w)_{(pxp)} = E[j(\theta, w)]$. A vector of independent variables z can be introduced in the above expressions to represent covariates in the regression situation.

3 Estimator densities from a statistical model

Consider the probability density function of the MLE from a specified statistical model. Inferences about the parameters from sample data can be based on descriptors of this density. For example, the mean of the density gives an estimate of the parameter itself and the variance allows the construction of a confidence interval for the parameter. It is also relevant to describe the density when the data generating mechanism differs from the model that is assumed for estimation. This might happen in a regression setting, where the estimation model is one of a number of possible candidate models for a process. Special techniques can be applied when there is doubt about the form of the data generating model, including distribution free and robust approaches to estimation. But the use of a specific estimation model is usual when the data are presumed to be distributed in a certain way according to a scientific hypothesis. Nevertheless the modeller may accept that alternative models are possible.

In the remainder of this section and in Section 4, it will be assumed that the parameters for the formulae are already known before generating the estimator densities. This is of course not the case when confronted with a real set of data. Now the parameters must be estimated and presumptions made about their true values, so that estimator densities can be generated. This is usually done by a principle of inversion: assume that the parameter estimate is in fact the true value and generate an estimator density on this assumption. Typically confidence intervals for the true parameter value can be generated around the parameter estimates on the basis of the estimator density. However this approach becomes questionable if the density is highly asymmetric.

A simple example of an estimator density is that of the mean of a normally distributed variable w, distributed as $N(\theta_0, \sigma^2)$, with known variance σ^2. If the same model $N(\theta, \sigma^2)$ is fitted with the mean θ as the estimable parameter, then the MLE $\hat{\theta}$ is distributed as $\mathrm{N}(\theta_0, \sigma^2/n)$, where n is the sample size. This result is exact and can be easily proved (see Section 4.1). When the data are from a more intricate model, a distribution for the MLE $\hat{\theta}$ is given by a normal approximation as $\mathrm{N}(\theta_0, i(\theta, w)^{-1})$ [5]. However this expression is generally only an approximation to the actual distribution $g(\hat{\theta})$, to $O(n^{-1})$ accuracy [4]. This means that it is only correct in the asymptotic limit as $n \to \infty$, where almost everything that can be known about the population is present in the sample and $\hat{\theta} \to \theta_0$ due to consistency of the MLE.

Alternative techniques are available. One of these is the p*-formula [2] [9], which is based on ideas of the asymptotic sufficiency of the MLE and can be more accurate than the normal approximation. The p*-formula is exact for a range of common models, although it is in general only asymptotically correct.

$\mathrm{g}(\hat{\theta}|a) = c|j(\hat{\theta}, \hat{\theta}, a)|^{1/2} e^{\bar{l}(\theta_0, \hat{\theta}, a)}$,

where c is a normalising constant and $||$ indicates absolute value. In this formula, the log likelihood is expressed not in terms of the sample vector w, but via $(\hat{\theta}, a)$, where a is an ancillary statistic. This means a distribution-constant statistic which, together with the MLE, constitutes a sufficient statistic. Sufficient, in the classical sense, means that the conditional density for the data w does not depend on the true parameter value θ_0 [14]. $\bar{l}$ is the normed form of the likelihood given by the difference between the log likelihood at θ_0 and its value at the MLE $\hat{\theta}$, $\bar{l}(\theta_0, \hat{\theta}, a) = l(\theta_0, \hat{\theta}, a) - l(\hat{\theta}, \hat{\theta}, a)$. The p*-formula is generally again only an approximation to the actual distribution $g(\hat{\theta})$, to $O(n^{-1})$ accuracy. But it can be considerably better than the normal approximation, because it takes more of the model structure into account. It is exact for a number of commonly used distributions.

In case an exact distribution is required, the default option is to carry out a simulations based analysis [13]. But simulation is essentially a numerical technique that requires a lot of computation. Some progress has been made in the formulation of exact analytic expressions for estimator densities. The trade off with approximate analytic forms is that computations become more difficult, though less extensive than with simulations and not at all impractical where prewritten routines can be made available. The user should ensure that the initial conditions are fulfilled for the method that is selected.

An exact formula for $g(\hat{\theta})$ was given by Skovgaard [15]. This is specified for the general class of contrast estimators, but is restricted here to the case of maximum likelihood estimation.

$$g(\hat{\theta}) = E[|j(\theta,w)||_{l'(\theta,w)=0}].g_s(0,\theta_0), \quad (1)$$

where $g_s(l'(\theta,w),\theta_0)$ is the marginal density of the observed score.

Another formulation for $g(\hat{\theta})$ was proposed by Hillier and Armstrong [8]. This is an integral equation, that can also apply to more general statistics than the MLE. It also depends upon $l'(\theta,w)$ and $j(\theta,w)$.

$$g(\hat{\theta}) = \int_{W_{\hat{\theta}}} |j(\theta,w)|.|l'(\theta,w).[l'(\theta,w)]^T|^{-1/2} g(w|\theta_0)(dW_\theta), \quad (2)$$

where (dW_θ) denotes a volume element on the surface W_θ.

In both equations (1) and (2), only the data, the observed score and the observed information are involved. Thus it is not required that an analytic expression be available for $\hat{\theta}$ in terms of the underlying data, which is useful because the MLE can usually only be found by an iterative estimation routine. The formulae however differ from each other in that (2) eschews the use of an expectation term within the formula, rather leaving the whole formula as a kind of extended expectation. Both sets of authors go on to show that their exact techniques can be used to re-establish previously known approximate approaches.

For simple densities, approximate methods such as p*-formula are at least asymptotically correct. They can also be used to make approximate inferences for small samples where the asymptotic results do not hold. But a trade-off should be applied, with exact techniques to be preferred unless they are too difficult to apply in a particular situation. Approximate techniques are often sufficiently close to exactness to be acceptable when the estimation model is equivalent to the data generating model. However they are not usually appropriate when models differ, and neither expression (1) nor (2) can be used directly in this situation either. In the next section an alternative formula will be discussed that can operate when the data generating model differs from the estimation model.

4 An exact technique for estimator densities (TED)

In the following, statistical models will be specified in terms of the densities of data that are generated by them. Consider that $g_0(w)$ is the true density of w, and $g_1(w|\theta)$ is a presumed density for estimation of θ. The log likelihood corresponding to $g_1(w|\theta)$ is $l(\theta,w)$. g_0 and g_1 can also be the same, in which case $g_0(w) = g_1(w|\theta_0)$.

When models differ, the estimate obtained by minimising $l(\theta,w)$ is technically a quasi maximum likelihood estimate (QMLE) [17]. Nevertheless it will continue to be termed MLE ($\hat{\theta}$) here and is given by the same expression as that for the MLE in the usual situation, $(l'(\theta,w)|_{\theta=\hat{\theta}}) = 0$. The space of $\hat{\theta}$ is $\hat{\Theta}$, a subspace of Θ.

Consider a $(p\times 1)$ vector T.

$$T(\theta,\theta^*,w) = l'(\theta^*,w) - l'(\theta,w) \quad (3)$$

θ^* is fixed at an arbitrary value. Under a simple set of regularity conditions, the exact density for $\hat{\theta}$ is given as follows [10].

$$g(\hat{\theta}) = E_w[|j(\theta,w)||_{\theta=\hat{\theta}}] \cdot g_{[T(\hat{\theta},\theta^*=\hat{\theta},w)]}(0), \quad (4)$$

where $j(\theta,w) = -l''(\theta,w)$ is the observed information under the estimation model $g_1(w|\theta)$, and the second term represents the value of the density $g_{[T(\hat{\theta},\theta^*,w)]}(t)$, for which $\theta^* = \hat{\theta}$, and hence $t = 0$ by (3). This is to be derived as the density of a transform $T(\hat{\theta},\theta^*,w)$ of the data w on the data generating model $g_0(w)$. The term $E_w[|j(\theta,w)||_{\theta=\hat{\theta}}]$ describes a conditional expectation, that is conditional on $\theta = \hat{\theta}$ and is taken wrt w over $g_1(w|\theta)$.

$$\left[\frac{\int_{W_{\widehat{\theta(v)}}} |j(\theta,w(v))||_{\theta=\hat{\theta}} \cdot g_1(w(v)|_{\theta=\hat{\theta}}) \cdot ||w'(v)||\, dv}{\int_{W_{\widehat{\theta(v)}}} g_1(w(v)|_{\theta=\hat{\theta}}) \cdot ||w'(v)||\, dv}\right] \quad (5)$$

Here, integration is carried out on a manifold $W_{\widehat{\theta(v)}}$, which runs over an (n - p) dimensional subset of W. The term $||w'(v)||$ indicates the magnitude of the Jacobian from co-ordinates v that index the manifold to w. In practice $||w'(v)||\, dv$ is equivalent to the volume element dW_θ in equation (2). However equation (4) may be easier to use than equation (2), since the evaluation of expectation (5) involves taking conditional expectations over data sets, which can usually be done without evaluation of the integrals in the expression. This is because of the following plug-in principle. In practice terms in w can be replaced by $E_w[w|_{\theta=\hat{\theta}}]$, terms in w^2 by $E_w[w^2|_{\theta=\hat{\theta}}]$, etc. For example, if the model $g_1(w|\theta)$ was normal $N(r[\theta_1],\theta_2)$, terms proportional to w would be replaced by terms proportional to $r[\hat{\theta_1}]$, and terms proportional to w^2 would be replaced by terms proportional to $\hat{\theta_2} + [r(\hat{\theta_1}]^2$. Equation (4) reduces to equation (1) when $g_0(w) = g_1(w|\theta_0)$.

4.1 An example of the mean of a sample from a normal distribution

This is an example where the estimation model is equivalent to the data generating model. The application of

TED for determining the pdf of the MLE will now be shown for the simple case of estimating the mean μ of a sample from a normal distribution N(μ_0, σ^2), with variance σ^2 assumed to be known. The argument is given to illustrate how the technique can work in more complex cases, and there are certainly easier ways to establish this particular result, for example via the moment generating function [6].

The data generating and estimation models are written as follows, in vector form, based on the analytic expression for a normal distribution.

$$\mathrm{g}_0(w) = exp[-nlog(\sqrt{2\pi\sigma^2}) - (\tfrac{1}{2\sigma^2}) \; . \; [w^T w - 2\mu_0 w^T 1 + \mu_0 1^T 1 \mu_0] \quad (6)$$

$$\mathrm{g}_1(w|\theta) = exp[-nlog(\sqrt{2\pi\sigma^2}) - (\tfrac{1}{2\sigma^2}) \; . \; [w^T w - 2\mu w^T 1 + \mu 1^T 1 \mu], \quad (7)$$

where ^{T} indicates transposition and 1 is a (nx1) vector of 1s.

Here the parameter vector θ is just a scalar μ because $p = 1$. The log-likelihood is the logarithm of (7), and its derivative wrt μ is

$$\mathrm{l'}(\theta, w) = (\tfrac{1}{\sigma^2})[1^T \, . \, (w - \mu 1)]$$

By (3),

$$\mathrm{T}(\theta, \theta^*, w) = (\tfrac{1}{\sigma^2})[1^T \, . \, (w - \mu^* 1)] - (\tfrac{1}{\sigma^2})[1^T \, . \, (w - \mu 1)]$$

When $l'(\theta, w)|_{\theta=\hat{\theta}} = 0$, $\mu = \hat{\mu} = w^T 1$. Therefore

$$T(\hat{\theta}, \theta^*, w) = (\frac{1}{\sigma^2})[1^T \, . \, (w - \mu^* 1)] \quad (8)$$

The observed information is $j(\theta, w) = -l''(\theta, w) = \frac{n}{\sigma^2}$. This is a constant here, so the conditional expectation (5) is also $\frac{n}{\sigma^2}$.

In order to develop the density $g_{[T(\hat{\theta},\theta^*,w)]}(t)$, recall that this is derived as the density of $T(\hat{\theta}, \theta^*, w)$ under the data generating model $g_0(w)$. In this case, $g_0(w)$ is given by equation (6). Equation (8) indicates that $T(\hat{\theta}, \theta^*, w)$ is a linear transform of w, and the use of a Jacobian shows that $g_{[T(\hat{\theta},\theta^*,w)]}(t)$ is $N(\frac{n\mu_0}{\sigma^2} - \frac{n\mu^*}{\sigma^2}, \frac{n}{\sigma^2})$. Therefore

$$\mathrm{g}_{[T(\hat{\theta},\theta^*=\hat{\theta},w)]}(0) = \frac{1}{\sqrt{2\pi . \frac{n}{\sigma^2}}} \; . exp\left[\frac{-1}{\frac{2n}{\sigma^2}} . [0 - (\frac{n\mu_0}{\sigma^2} - \frac{n\hat{\mu}}{\sigma^2})]^2\right] =$$

$$\frac{1}{\sqrt{2\pi . \frac{n}{\sigma^2}}} \; . exp\left[\frac{-1}{\frac{2\sigma^2}{n}} . [(\hat{\mu} - \mu_0)^2]\right] \quad (9)$$

Applying (4), $g(\hat{\theta})$ is obtained as the product of $\frac{n}{\sigma^2}$ and (9), which is $N(\mu_0, \sigma^2/n)$, as required.

The step of calculating $g_{[T(\hat{\theta},\theta^*,w)]}(t)$ sets the limit of tractability for TED. For curved exponential families in general, $T(\hat{\theta}, \theta^*, w)$ is a linear function of w, which facilitates the required calculations.

4.2 Example of the mean of a sample from a negative exponential data generating model using a normal distribution estimation model

This is an example where the estimation model is not equivalent to data generating model. The data generating model is given as follows.

$$g_0(w) = \frac{1}{\mu_0} . exp[\frac{-1}{\mu_0} w^T 1], \quad w_i \geq 0, \quad i = 1, ..., n \quad (10)$$

The estimation model remains normal as in equation (7).

$T(\hat{\theta}, \theta^*, w)$ is still given by (8) and $E_w[|j(\theta, w)||_{\theta=\hat{\theta}}]$ is still $\frac{n}{\sigma^2}$. Use can be made of the fact that the sum of n i.i.d. negative exponential variables has a gamma distribution with shape parameter of n and mean μ_0 [12],

$$g(s) = \frac{s^{n-1} exp[\frac{-1}{\mu_0} s]}{\mu_0^n (n-1)!}, \quad s = w^T 1 \geq 0 \quad (11)$$

Equation (8) shows that $T(\hat{\theta}, \theta^*, w)$ is a linear transform of s, and $s = \sigma^2 T + n\mu^*$. Therefore, transformation of equation (11) gives

$$\mathrm{g}_{[T(\hat{\theta},\theta^*,w)]}(t) = \frac{\sigma^2 (\sigma^2 T + n\mu^*)^{n-1} exp[\frac{-1}{\mu_0}(\sigma^2 T + n\mu^*)]}{\mu_0^n (n-1)!}, \quad \sigma^2 T + n\mu^* \geq 0 \quad (12)$$

The density $g(\hat{\mu})$ is obtained by setting $T = 0$, $\mu^* = \hat{\mu}$ in equation (12) and multiplying by $E_w[|j(\theta, w)||_{\theta=\hat{\theta}}] = \frac{n}{\sigma^2}$.

$$\mathrm{g}(\hat{\mu}) = \frac{n^n \hat{\mu}^{n-1} exp[\frac{-n\hat{\mu}}{\mu_0}]}{\mu_0^n (n-1)!}, \quad \hat{\mu} \geq 0$$

This is a gamma distribution with a shape parameter of n and mean $\frac{\mu_0}{n}$.

In this case, the parameter estimate $\hat{\mu}$ is necessarily positive. The distribution of $\hat{\mu}$ is independent of σ^2, which is the assumed value for the variance of the normal estimation model. In fact $g(\hat{\mu})$ is the same density that is found when the negative exponential data generating model (10) is retained and an equivalent negative exponential distribution is used as the estimation model, instead of the normal distribution estimation model. This is no longer the case however if a negative exponential estimation model is used with a normal data generating model. The normal distribution therefore dominates the negative exponential as estimation model for the mean

with this pair of models, which is related to the fact that the negative exponential has only one fixable parameter while the normal distribution has two. Although the example is simplistic, it shows that it can be possible to pick dominant estimation models wrt the parameter of interest.

5 Methods for comparing models

While some analysts are happy to assess the robustness of estimation of common parameters throughout a set of candidate models, others want to discriminate between models in order to identify the best one. Akaike [1] suggested an information criterion that can be used for this.

AIC = -2[l$(\theta, w|_{\theta=\hat{\theta}})$] + 2$p$

The best model is considered to be the member of the candidate set that minimises AIC. In practice, the first term in the expression measures the goodness of fit of the model to data, while the second term is a penalty based on the number of parameters that have to be fitted. The latter term is a guard against over fitting.

The theoretical underpinning of this approach is that $AIC/2$ can be shown to be asymptotically equivalent to a constant minus the expected value of the Kullback-Leibler information $I(g_0, g_1)$.

$$I(g_0, g_1) = log\left(\frac{g_0(w)}{g_1(w|\theta)}\right) \tag{13}$$

$$E[I(g_0, g_1)] = \int_W \left[log\left(\frac{g_0(w)}{g_1(w|\theta)}\right)\right] g_0(w)dw \tag{14}$$

For models with the same number of parameters, maximisation of $E[I]$ at the MLE is asymptotically equivalent to minimisation of AIC. $I(g_0, g_1)$ itself is equivalent to the difference between the (Boltzmann) entropy of the models under consideration, $log(g_0(w))$ and $log(g_1(w|\theta))$ respectively.

In the case of a fixed data generating model $g_0(w)$, it can be shown that the QMLE $\hat{\theta}$ for a particular estimation model $g_1(w|\theta)$ minimises $I(g_0, g_1)$ [17].

AIC provides a benchmark for comparing adequacies of competing models on real data sets. In case the data generating model $g_0(w)$ is unknown, it can be argued that the true model need not be a member of the considered set of models [3]. However, AIC suffers to some extent from the same problem that was alluded to earlier for approximate forms of the estimator density, in that it is absolutely valid only in the asymptotic limit and it is also subject to problems of inversion. The first problem can be tackled by attempting a more exact approach, while the second problem is inescapable unless one is prepared to carry out an encompassing analysis on theoretical data sets.

Several other related information criteria for comparing models have been proposed as improvements to AIC. One of these is TIC [16].

TIC = -2[l$(\theta, w|_{\theta=\hat{\theta}})$] + 2[$trace(K(\hat{\theta})_{(pxp)}.[j(\hat{\theta}, w)]^{-1})$],

where $K(\hat{\theta}) = E_{g0}[l'(\theta, w).l'(\theta, w)^T|_{\theta=\hat{\theta}}]$.

$TIC/2$ is a closer approximation than AIC to a constant minus $E[I]$. For large samples, as $n \to \infty$, TIC and AIC converge as $trace(K(\hat{\theta})_{(pxp)}.[j(\hat{\theta}, w)]^{-1}) \to p$.

Both AIC and TIC provide rational ways to compare nested or non-nested models of various degrees of complexity - with a balance between the log likelihood and the number of parameters as a guard against over fitting. Otherwise the common practice is to compare models with different numbers of parameters via an analysis of variance based F test, using the sums of squares associated with the additional parameters. This is however only strictly valid for nested linear models. So the use of information criteria that are based on K-L information is attractive.

The TED estimator density (4) can be used for comparing models. One possible way to use the technique is via a robustness index (RI), when the parameters are common in the two models. Hingley [10] explains RI and gives an example of the comparison for a gamma data generating model and a negative exponential estimation model.

But TED can also be used directly to construct new information criteria. For example, since it gives exact densities, it can make an indicator from K-L information without requiring the use of either expectations or asymptotic approximation. However inversion is still required. A particularly simple form is obtained when comparing the effect of two alternative data generating models under a common estimation model, because then the term $E_w[|j(\theta, w)||_{\theta=\hat{\theta}}]$ in the expectation term (5) is the same for both models and cancels out when taking the ratio.

It is possible to calculate two densities for $\hat{\theta}$. For both densities assume that the estimation model is $g_1(w|\theta)$. For the first density, the data generating model is $g_1(w|\theta_1)$, and is written as $g(\hat{\theta}) = m_A(\hat{\theta}|\theta_1)$. For the second density, the data generating model is $g_0(w|\theta_0)$, and is written as $g(\hat{\theta}) = m_B(\hat{\theta}|\theta_0)$.

An information criterion can be constructed as follows.

$$\mathrm{H}(\mathrm{m}_A, m_B) = log\left(\frac{m_A(\hat{\theta}|\theta_0)}{m_B(\hat{\theta}|\theta_1)}\right) = log\left(\frac{g_{A[T(\hat{\theta},\theta^*=\hat{\theta},w)]}(0)}{g_{B[T(\hat{\theta},\theta^*=\hat{\theta},w)]}(0)}\right) \tag{15}$$

This takes a fairly simple form, and it is suggested that θ_0 should be set as $\hat{\theta}$ for the data under $g_0(w|\theta)$ as estimation model. For θ_1, it is possible to use the estimate $\hat{\theta}$ from $g_1(w|\theta)$ as estimation model. For the particular example that will be given below, it will turn out to be unnecessary to estimate θ_1.

As a demonstration, consider a pair of nested linear models.

$$\mathrm{g}_1(w|\theta_1) = MN_w(X_{1(nxq)}\theta_{1(qx1)}, \sigma^2 I_{(nxn)})$$

$$\mathrm{g}_0(w|\theta_0) = MN_w(X_{0(nxp)}\theta_{0(px1)}, \sigma^2 I_{(nxn)}),$$

where the error variance term σ^2 is assumed to be known. q and p can be unequal, and it is not prespecified which model is nested in the other. Pairwise comparisons are envisaged, but a larger set of candidate models can be used with comparisons between all pairs of models in the set.

Under m_A, both the data generating model and the estimation model are the same, and application of the TED formula gives $g(\hat{\theta})$ as $MN_W(\theta_1, (X_1^T X_1)^{-1}\sigma^2)$. However, under m_B, $g(\hat{\theta})$ is $MN_W((X_1^T X_1)^{-1}X_1^T X_0\theta_0, (X_1^T X_1)^{-1}\sigma^2)$. Then the information criterion (15) is as follows.

$$\mathrm{H}(\mathrm{m}_A, m_B) = \tfrac{-1}{2} log\left|X_1^T X_1\right| + \tfrac{1}{2} log\left|X_0^T X_0\right| + \tfrac{1}{2\sigma^2}\left[\hat{\theta} - (X_1^T X_1)^{-1}X_1^T X_0\theta_0\right]^T (X_1^T X_1)\left[\hat{\theta} - (X_1^T X_1)^{-1}X_1^T X_0\theta_0\right] - \tfrac{1}{2\sigma^2}\left[(\hat{\theta} - \theta_1)\right]^T (X_1^T X_1)\left[(\hat{\theta} - \theta_1)\right]$$

Now, θ_1 can be estimated from the data set under $g_1(w|\theta)$ and θ_0 can be estimated from the same data set under $g_0(w|\theta)$. These estimates are substituted into the above expression, using the principle of invariance, to make an estimated indicator $\hat{H}$. This simplifies matters by causing the final term to disappear. Hence, as was stated above, in this case θ_1 does not actually need to be estimated at all.

$$\hat{H(m_A}, m_B) = \tfrac{-1}{2} log\left|X_1^T X_1\right| + \tfrac{1}{2} log\left|X_0^T X_0\right| + \tfrac{1}{2\sigma^2}\left[\hat{\theta} - (X_1^T X_1)^{-1}X_1^T X_0\hat{\theta_0}\right]^T (X_1^T X_1)\left[\hat{\theta} - (X_1^T X_1)^{-1}X_1^T X_0\hat{\theta_0}\right]$$

As an example of a particular application, consider three competing linear models for data on the development of numbers of patent filings per year at the European Patent Office from Germany (DE), Japan (JP) and USA (US). The models are based on an econometric specification that is used to forecast future patent filing levels [11].

Model i: $\mathrm{Y}_{t,j} = a + b.Y_{t-1,j} + c.X_{t-5,j} + error_{t,j}$

Model ii: $\mathrm{Y}_{t,j} = a_j + b.Y_{t-1,j} + c.X_{t-5,j} + error_{t,j}$

Model iii: $\mathrm{Y}_{t,j} = a_j + b.Y_{t-1,j} + c_j.X_{t-5,j} + error_{t,j}$,

where $j = 1, 2, 3$ indicates DE, JP, US; $t = 1980, ..., 2005$. The main variables are first standardised between countries by means of the following transformations. $Y_{t,j}$ is the number of patent filings from country j in year t, divided by the number of workers in country j in year t. $X_{t,j}$ is the discounted research and development expenditures in country j in year t, again divided by the number of workers in country j in year t. Model i assumes common intercepts and regression parameters for the three countries. Model ii allows separate intercepts per country a_j. Model iii further allows separate slopes per country for the standardised research and development expenditures variable c_j. See Figure 1.

The information criteria are calculated pairwise between all combinations of Models i, ii and iii. The fixed value of σ^2 is estimated from the fit of the data generating model in each case. From the above expression for $H(\hat{m_A}, m_B)$, it can be seen that when the formulations of $g_0(w|\theta_0)$ and $g_1(w|\theta_1)$ are the same, $H(\hat{m_A}, m_B) = 0$.

	Model i estimate	Model ii estimate	Model iii estimate
Model i generate	0	1.30	3.54
Model ii generate	1.64	0	2.40
Model iii generate	1.25	-0.40	0

Table 1. Values of the information criterion $H(\hat{m_A}, m_B)$ for pairwise combinations of Models i, ii and iii as data generating model and estimation model.

Table 1 shows the 3x3 representation of $\hat{H}$ that was found. Minimisation of $\hat{H}$ is appropriate. In the case of data generated by Model iii, the criteria suggest that estimation by Model ii is slightly better than estimation by Model iii.

	Calculated SS	DF	Mean SS	F test	P value
Model i	35.5066	3			
Diff. i to ii	0.06581	2	0.0329	3.2	< 0.05
Model ii	35.5724	5			
Diff. ii to iii	0.0412	2	0.0206	2.0	> 0.05
Model iii	35.6136	7			
Resid. to iii	0.7199	70	0.0103		
Total	36.3335	77			

Table 2. Analysis of variance table to compare the significance of differences ('Diff.') between Model i (nested in) Model ii (nested in) Model iii. SS is Sums of Squares. DF is degrees of freedom. F tests are of Mean SS against the Residual ('Resid.') to Model iii. P values are against null hypotheses that the difference terms are not significant.

In a case like this, where nested linear models have been used, the information based analysis can be complemented by a conventional analysis of variance analysis, as shown in Table 2. The two F tests that are reported in the table suggest that Model ii gives a significantly better fit to the data than Model i does, but that no significant improvement in fit can be established when using model

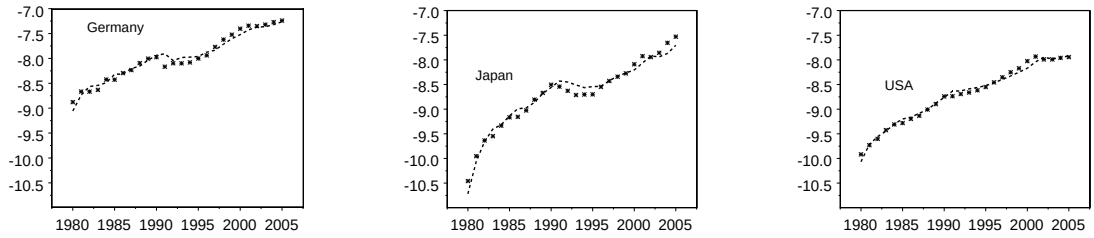

Figure 1: Fig. 1. Transformed patent filings at the European Patent Office. $Y_{t,j}$ vs. Year (t), filings from three countries (j). Fitted values (lines) by Model iii.

iii instead of Model ii. This is consistent with Table 1, in that it suggests that Model ii is adequate for the data.

The particular information criterion that has been suggested here could be extended easily to other types of models, particularly to nonlinear regression models. Other designs for information criteria can also be imagined.

6 Conclusion

Analysts should consider using exact densities for maximum likelihood estimates whenever it is practical to do so. The densities are particularly appropriate where alternative models are to be compared and assessed. The example in the previous section used exact densities to construct a trial information criterion. It may also be profitable to use exact methods to complement the range of existing approximate techniques with exact analytic equivalents. Although the exact techniques require more complex calculations, it can turn out that statistics can be created that have a simpler direct interpretation than the approximate ones. Further work is needed on these issues.

References

[1] Akaike, M., "Information theory as an extension of the maximum likelihood principle", Petrov, B., Csaki, C. (eds), *Second International Symposium on Information Theory*, Akademiai Kiado, 1973.

[2] Barndorff-Nielsen, O.E., Cox, D.R., *Inference and asymptotics*, Chapman and Hall, 1994.

[3] Burnham, K.P., Anderson, D.R., *Model selection and multimodel inference*, Second Edition, Springer, 2002.

[4] Cox, D.R., Hinkley, D.V., *Theoretical statistics*, Chapman and Hall, 1974.

[5] Davison, A.C., *Statistical models*, p. 118, Cambridge, 2003.

[6] DeGroot, M., *Probability and statistics, Second edition*, p. 271, Addison-Wesley, 1986.

[7] Giudici, P., *Applied data mining*, Wiley, 2003.

[8] Hillier, G., Armstrong, M., "The density of the maximum likelihood estimator", *Econometrica*, V67, pp. 1459-1470, 1999.

[9] Hingley, P.J., "p*-formula", Kotz, S., Read, C., Banks, D. (eds), *Encyclopedia of statistical sciences, Update volume 3*, Wiley, 1999.

[10] Hingley, P.J., "Analytic estimator densities for common parameters under misspecified models", *Statistics for Industry and Technology*, pp. 119-130, 2004.

[11] Park, W.G., "International patenting at the European Patent Office: aggregate, sectoral and family filings", Hingley, P., Nicolas, M. (eds), *Forecasting innovations*, Springer, 2006.

[12] Johnson, N.L., Kotz, S., Balakrishnan, N., *Continuous univariate distributions, Volume 1*, p. 340, Wiley, 1994.

[13] Morgan, B.J.T., *Elements of simulation*, Chapman and Hall, 1984.

[14] Schervish, M.J., *Theory of statistics*, Springer, p. 85, 1995.

[15] Skovgaard, I.M., "On the density of minimum contrast estimators", *Annals of Statistics*, V18, pp. 779-789, 1990.

[16] Takeuchi, K., "Distribution of informational statistics and a criterion of model fitting", *Suri-Kagaku*, V153, pp. 12-18, 1976.

[17] White, H., *Estimation, inference and specification analysis*, Cambridge, 1996.