# Testing Restrictions and Comparing Models*

## 1. The Problem

We consider here the problem of comparing two parametric models for the data $X$, defined by pdf's $p(X \mid \theta)$ and $q(X \mid \phi)$, with $\theta$ ranging over a parameter space $\Theta$ and $\phi$ ranging over $\Phi$, with the two parameter spaces being of possibly different dimension. A special case that we will discuss further is that in which the second "$q$" model is defined by restricting $\theta$ to a a submanifold of $\Theta$ defined by the vector of restrictions $R(\theta) = 0$. We aim at forming posterior probabilities of the two models, or in the case of the model defined by restrictions, for the truth of the restriction $R(\theta) = 0$, conditional on the data.

It is easy to prescribe how to handle this problem in general terms. The full parameter space is really $\Theta \cup \Phi$. We will have some prior over it, which we can think of as built from conditional pdf's over $\Theta$ and $\Phi$ and prior probabilities on the two parameter spaces. The posterior probability can then be calculated by the usual formulas. However, in contrast to the usual situation where the prior is continuous (i.e. has a pdf), the effects of the prior do not disappear in large samples in this situation. It is therefore useful to examine whether there remains anything useful to say about large-sample approximations for this problem.

## 2. Gaussian Approximation

In large samples, under rather general regularity conditions, the likelihood comes to dominate the prior if the parameter space is a subset of $\mathbb{R}^m$ with non-empty interior and the true value of the parameter is in the interior. In this case the use of a "flat prior" gives accurate conclusions. Furthermore, the likelihood takes on an approximately Gaussian shape in large samples, in the sense that a second-order Taylor expansion of the log likelihood in the neighborhood of its maximum gives accurate results if used to approximate the posterior pdf.[1] Even if the true model is not any of the pdf's $p(X \mid \theta)$ for $\theta \in \Theta$, under reasonable regularity conditions likelihood concentrates in the neighborhood of a "pseudo-true value" in $\Theta$ and the quadratic approximation to the log likelihood's shape becomes accurate in large samples.



[1] An informal sketch of a proof is in Gelman, Carlin, Stern, and Rubin (1995, Appendix B). A careful proof is in Schervish (1995, Chapter 7.4). Both of these references include further references to the literature.

For functions whose logs are quadratic (and which therefore are scaled versions of Gaussian pdf's) we can characterize their integrals as functions of their maxima and their second derivative matrices.[2] This leads to some easily computed approximations to posterior probabilities, based on the height of the likelihood function at its maxima in the two spaces and on the second derivative matrices of the log likelihood at these maxima — which are minus the usual classical asymptotic approximations to the covariance matrices of the MLE estimates.

If the log likelihood is exactly quadratic when restricted to $\Theta$, then the likelihood itself has the form within $\Theta$

$$\log(p(X \mid \theta)) = K - \tfrac{1}{2}(\theta - \hat{\theta})'\Omega^{-1}(\theta - \hat{\theta}) \,, \tag{1}$$

where $K = \log(p(X \mid \hat{\theta})$ is the value of log-likelihood at its peak and $\hat{\theta}$ is the value of $\theta$ that maximizes $p(X \mid \theta)$. We can then rewrite it as a constant plus the log of a Gaussian pdf,

$$\begin{aligned}\log(p(X \mid \theta)) = K + \frac{m}{2}\log(2\pi) + \tfrac{1}{2}\log|\Omega| \\ + \left[ -\left(\frac{m}{2}\log(2\pi) + \tfrac{1}{2}\log|\Omega|\right) - \tfrac{1}{2}(\theta - \hat{\theta})'\Omega^{-1}(\theta - \hat{\theta}) \right] \,,\end{aligned} \tag{2}$$

where $m$ is the dimension of the space $\Theta$. Obviously, then,

$$\int_{\Theta} p(X \mid \theta)\, d\theta = e^{K}(2\pi)^{\frac{m}{2}}\, |\Omega|^{\frac{1}{2}} \,. \tag{3}$$

If the likelihood is concentrated on small enough subsets of $\Phi$ and $\Theta$ that the prior pdf on each space is approximately constant where likelihood is non-trivially large, then we can apply (3) to approximate the integrals of the joint pdf of data and parameters over $\Theta$ and $\Phi$, respectively, with the data $X$ held fixed at the observed value, as

$$g(\hat{\theta})p(X \mid \hat{\theta})(2\pi)^{\frac{m}{2}}\, |\Sigma_\theta|^{\frac{1}{2}} \tag{4}$$

$$h(\hat{\phi})q(X \mid \hat{\phi})(2\pi)^{\frac{n}{2}}\, \left|\Sigma_\phi\right|^{\frac{1}{2}} \tag{5}$$

where $n$ is the dimension of $\Phi$, $\Sigma_\theta$ and $\Sigma_\phi$ are the usual asymptotically justified estimates of the covariance matrix of the MLE (minus the inverse of the second derivative of the log likelihood) within each model's parameter space, and $g$ and $h$ are the conditional prior pdf's on $\Theta$ and $\Phi$, respectively. Posterior odds on the two

[2] This type of approximation is known as the *method of Laplace* and is discussed in more general form in Schervish (1995, section 7.4.3)

models $\Theta$ and $\Phi$ are then approximated as the ratio of prior probabilities of the two models multiplied by the ratio of the expressions (4) and (5).[3]

It is clear, then, that use of the asymptotic normal approximation will not make the choice of prior asymptotically irrelevant in computing posterior probabilities. The odds ratio between the two models will contain the ratio of prior probabilities times the ratio of conditional prior densities at the maximum $g(X \mid \hat{\theta})/h(X \mid \hat{\phi})$ no matter how large the sample or how good the Gaussian approximation.

However, if we do not aim at getting an asymptotically accurate odds ratio, but instead only to get an asymptotically accurate decision — that is, to determine accurately which model the odds ratio favors — then we can avoid dependence on the prior under some additional regularity conditions.

In models of i.i.d. data, or of stationary time series data, it is usually true that

$$T\Sigma_\theta \xrightarrow[T\to\infty]{P} \Omega_\theta \,, \tag{6}$$

where $\Omega_\theta$ is a fixed matrix, with a similar result for $\Sigma_\phi$. Letting $\mu(\Theta)$ be the prior probability of $\Theta$, we can then write the approximation to the log of the posterior odds in favor of $\Theta$ as

$$\log\left(\frac{p(X \mid \hat{\theta})\mu(\Theta)g(\hat{\theta})}{q(X \mid \hat{\phi})(1-\mu(\Theta))h(\hat{\phi})}\right) + \frac{m-n}{2}\log(2\pi) \\ - \log\left(T^{\frac{m}{2}}\left|\Omega_\theta\right|^{-\frac{1}{2}}\right) + \log\left(T^{\frac{n}{2}}\left|\Omega_\phi\right|^{-\frac{1}{2}}\right) \,. \tag{7}$$

Also, for i.i.d. or time series models, $\log p(X \mid \theta)$ is a sum of similarly distributed random variables, so that under reasonable regularity conditions

$$\frac{1}{T}p(X \mid \hat{\theta}) \xrightarrow[T\to\infty]{P} \bar{p} \,, \tag{8}$$

again with a similar result holding for $q$. Every term in (7) is constant, as $T$ increases, except for $\log(p/q)$ and the two terms on the end involving $\Omega$'s. The odds ratio will, under usual regularity conditions, converge to infinity or zero as $T \to \infty$, so that eventually the terms that depend on $T$ dominate its behavior. Thus we can form an approximate "odds ratio" considering only the terms that vary with $T$, i.e.

$$\log(p(X \mid \hat{\theta})) - \log(q(X \mid \hat{\phi})) - \frac{m-n}{2}\log T \,. \tag{9}$$

[3]We could improve the order of approximation by taking account of the fact that $\partial g(\hat{\theta})/\partial\theta$ and $\partial h(\hat{\phi})/\partial\phi$ are not zero. This could be done by replacing $\hat{\theta}$ and $\hat{\phi}$ with the values of $\theta$ and $\phi$ that maximize the posterior pdf's, instead of those that maximize the likelihood, or else by computing $\partial g(\hat{\theta})/\partial\theta$ and $\partial h(\hat{\phi})/\partial\phi$ and using them to make a first-order correction to $\hat{\theta}$ and $\hat{\phi}$ as estimates of the posterior mode. This would in turn imply a second-order (in $\theta - \hat{\theta}$ and $\phi - \hat{\phi}$) correction to the heights of the posterior pdf's at their maxima.

This criterion will converge in probability to $+\infty$ if only $\Theta$ contains the true model and $-\infty$ if only $\Phi$ does. It is often called the Schwarz criterion, as Schwarz (1978) introduced it.[4]

The Schwarz criterion is fairly widely applied. Its popularity stems from the fact that it can be computed without any consideration of what a reasonable prior might be. But it is apparent from (7) that unless the Schwarz criterion is extremely large or small, it is likely to differ substantially from the odds ratio, even when the sample size is large enough to make the Gaussian approximation to log likelihood work well. Serious applied work ought to include in the reporting of results a consideration of what might be a reasonable specification of the prior, as well as a consideration of whether the normal approximation to the likelihood is accurate in the sample at hand.

Note that we could avoid having to think about the prior without dropping so many terms from (7). Dropping only those terms that depend on the prior, instead of all those terms that fail to grow with $T$, leads to

$$\log\left(\frac{p(X\,|\,\hat{\theta})}{q(X\,|\,\hat{\phi})}\right) + \frac{m-n}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{|\Sigma_\theta|}{|\Sigma_\phi|}\right). \tag{10}$$

This expression makes it clear that it is the precision of the estimates that produces the dependence on $T$ in the Schwarz Criterion. The expression is more robust — it converges to the same limit as the Schwarz Criterion when the Schwarz criterion is justified, but produces asymptotically correct decisions in certain situations where the Schwarz criterion does not. For example, in unit root time series models the asymptotic Gaussian approximation to the shape of the likelihood is accurate[5], but $T\Sigma_\theta$ fails to be bounded in probability. In this situation the Schwarz criterion may not lead to asymptotically correct decisions, whereas direct use of (10) does.

## 3. Infinitely Many Parameters

There are a number of classes of cases in which it is natural to think of a nested sequence of models, or a model with a countable infinity of parameters, as containing the truth. For example, AR or MA models with different numbers of lags, or models that fit a polynomial time trend or that use polynomials to approximate a nonlinear regression function. These situations can be treated as involving a Bayesian prior over the infinite sequence of models or over the infinite sequence of parameters. Obviously in the case of a sequence of models we have to have probabilities on

[4]However, it was implicit in earlier work on Laplace approximations, and Schwarz considered only its application to cases where $\Phi$ is a lower-dimensional subset of $\Theta$.

[5]This was shown by Kim (1994)

models that eventually get small as we go out the sequence of models or probabilities — otherwise we could not have probability one on the whole sequence of models. Correspondingly, when we have infinite sequences of parameters we usually have to assume that the parameters far out in the sequence are very likely to be unimportant.

Consider the case of a simple autoregression

$$y_t = \sum_{s=0}^{k} \beta_s y_{t-s} + \varepsilon_t \,. \tag{11}$$

Choice of $k$ here does involve choice among models, but it is clear that our priors on the sequence of $\beta$'s probably vary systematically with $k$. Not only are we likely to think that large values of $k$ are less likely than smaller values, at least for large enough $k$, but we are also likely to think that $\beta_s$ for large $s$, within a model with a given $k$, are probably smaller in absolute value than values of $\beta_s$ for smaller $s$. These beliefs might be justified by experience with fitting models to many economic time series. But they also can be justified by supposing that we put some substantial probability (maybe even 1) on the region of the parameter space that implies stationarity or unit-root behavior, and little or no probability on explosive models. Most $\beta$ sequences in these regions will have the properties we have suggested are natural.

In a distributed lag model

$$y_t = \sum_{s=0}^{k} \gamma_s x_{t-s} + \eta_t \,, \tag{12}$$

our prior is likely to be that the coefficients on $\gamma_s$ for large $s$ are smaller than the coefficients on small $s$, because otherwise our priors would imply that the variance of the explained part of $y_t$ grows without bound with increasing $k$.

We don't have time to cover these cases in any generality, but consider the case, for example, where we have a distributed lag model and within each model $k$, we assume exponential decay of the $\gamma$'s, i.e.

$$\begin{bmatrix} \gamma_0 \\ \vdots \\ \gamma_k \end{bmatrix} \sim N\left(0, \begin{bmatrix} \theta & 0 & 0 & \dots & 0 \\ 0 & \theta\lambda & 0 & \dots & 0 \\ \vdots & \dots & \ddots & \dots & \vdots \\ 0 & 0 & \cdots & \theta\lambda^{k-1} & 0 \\ 0 & 0 & \cdots & 0 & \theta\lambda^{k} \end{bmatrix}\right) . \tag{13}$$

It is not hard to see that this kind of prior pdf will produce systematic behavior in the $\log|\Omega|$ terms in (7), and in particular that if we look across $k$'s they will contribute a component that varies as $O(k^2 \log\lambda)$. Since it is the inverse of $\Omega$ that enters the posterior pdf and $\lambda < 1$, these terms increase with $k$. Since they do not vary with

$T$, they will be dominated by the SC terms as $T \to \infty$. But in a given sample, they will imply less of a penalty on large models than would be apparent from approximations to posterior odds (like the SC, or (10)) that ignore the contribution of priors. This may account for the reputation of the SC, that it tends to favor small models more strongly than is reasonable.

On the other hand, we also must recognize that the prior weights on models must decrease with $k$ if they are to sum to a finite number, and this will work in the opposite direction, disfavoring larger models. If the weights on models decline only as $O(e^{-k})$ or $O(k^{-p})$, however, the effect above from exponentially declining sizes of coefficients will dominate.

## 4. Comparison to Classical Methods

Classical asymptotic methods produce results only for the case where $\Phi$ is defined as $\{\theta \in \Theta \mid R(\theta) = 0\}$ for a function $R$ that takes values in $\mathbb{R}^{m-n}$. It is then conventional procedure to use the fact that, on the null hypothesis that $R(\theta) = 0$, the likelihood ratio (LR) statistic $2\big(\log(p(X \mid \hat{\theta})/q(X \mid \hat{\theta}))\big)$ is in large samples distributed approximately as $\chi^2(m-n)$. Usually some standard significance level, say $\alpha = .01$ or $\alpha = .05$, is chosen, and the null hypothesis is rejected if the LR exceeds $\chi^2_\alpha(m-n)$. Because both the Schwarz criterion and the standard likelihood ratio test compare the LR to a critical value, the Schwarz criterion is always equivalent, in any given sample, to a classical likelihood ratio test for some $\alpha$. However the value to which the Schwarz criterion compares the LR is increasing in $T$, while the classical test, if $\alpha$ is kept at a constant level, compares the LR to a fixed value.

In fact it is obvious by construction that the choice of $\Theta$ vs. $\Phi$ does not converge in probability to the truth as $T \to \infty$ if the choice is made with a conventional LR test with fixed $\alpha$. By definition, such a test in large samples has probability $\alpha$ of rejecting a true null hypothesis — i.e. of giving the wrong decision — even for very large samples. The Schwarz criterion (and some other variants on it that have been proposed) instead gives a probability of wrong decision that goes to zero as $T \to \infty$.

The Bayesian approach then suggests that significance levels for tests should generally be set more stringently (lower $\alpha$'s) in large samples. Though there is no classical statistical argument for doing so, applied workers do tend in fact to use lower $\alpha$'s in very large samples, or indeed, as (10) would suggest, in any context where estimated standard errors are very small.

## References

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995): *Bayesian Data Analysis*. Chapman and Hall, London.

KIM, J. Y. (1994): "Bayesian Asymptotic Theory in a Times Series Model with a Possible Nonstationary Process," *Econometric Theory*, 10(3), 764–773.

SCHERVISH, M. J. (1995): *Theory of Statistics*, Springer Series in Statistics. Springer, New York.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–64.