Model comparison

Christopher A. Sims
Princeton University
sims@princeton.edu

October 15, 2015

Model comparison as estimating a discrete parameter

- Data Y, models 1 and 2, parameter vectors θ_1, θ_2.
- Models are $p_1(Y \mid \theta_1)$ and $p(Y \mid \theta_2)$. m_1, m_2 are the dimensions of θ_1, θ_2.
- We can treat them as a single model, with pdf $q(Y \mid (\theta_1, \theta_2, \gamma))$, where the parameter space for θ_1, θ_2 is some subset (or the whole of) $\mathbb{R}^{m_1+m_2}$ and that for γ, the model number, is $\{1, 2\}$.
- Then the posterior probability over the two models is just the marginal distribution of the γ parameter — i.e., the posterior p.d.f. with θ_1, θ_2 integrated out.

Marginal data densities, Bayes factors

- The kernel of the probability distribution over model i is therefore

$$\int p(Y \mid \theta_1)\pi(\theta_1, \theta_2, i)\, d\theta_1\, d\theta_2) \; .$$

- The ratios of these kernel weights are sometimes called Bayes factors.

- The integral values themselves are sometimes called marginal data densities.

- The posterior probabilities are the ratios of the mdd's to their sum — i.e. the kernel weights normalized to sum to one.

Notes

- Often the prior over θ_1, θ_2 makes them independent, in which case only the prior pdf of θ_i enters the formula.

- The prior matters here, even asymptotically. With independent priors, assuming the likelihood concentrates asymptotically in the neighborhood of $\bar{\theta}_i$ for each model i, a change in $\pi(\theta_i)/\pi(\theta_j)$ always has a direct proportional effect on the posterior odds, even as sample size goes to infinity.

- It is still true in a sense that the prior doesn't matter: The posterior odds favoring the true model will go to infinity, regardless of the prior. But in any sample in which the odds ratio turns out to be of modest size, it can shift to favor one model or the other based on reasonable variations in the prior.

- This discussion generalizes directly to cases with any finite number of models.

Calculating mdd's

- This is just a question of how to evaluate an integral. For some cases, including reduced form VAR's with conjugate priors, this can be done analytically.

- Suppose, though, we can't evaluate the integral analytically, but can generate a sample from the posterior pdf.
  - Case 1: We can generate a sample from the whole posterior, over both the continuous and discrete components. Then posterior mdd is estimated just by counting relative frequencies of $\gamma = 1$ vs. $\gamma = 2$.
  - Case 2: We can generate a sample from either model's conditional posterior on θ_i, but not a sample from the joint $\theta_1, \theta_2, \gamma$ posterior.

Case 2

- We have a sample $\left\{\theta_i^j, j = 1, \ldots, N\right\}$ of draws from the posterior on θ_i conditional on model i being correct. We also assume here that the priors on θ_1 and θ_2 are independent.
- Form $\left\{k(Y, \theta_i^j)\right\} = \left\{p(Y \mid \theta_i^j)\pi(\theta_i^j)\right\}$. This is the kernel of the posterior conditional on model i being the truth.
- Pick a weighting function $g(\theta_i)$ that integrates to one over θ_i-space.
- Form

$$\frac{1}{N}\sum_{j=1}^{N}\frac{g(\theta_i^j)}{k(Y, \theta_i^j)}$$

- This converges to $(\int k(Y, \theta_i) d\theta_i)^{-1}$, because

$$E\left[\frac{g(\theta_i)}{k(Y, \theta_i)}\right]$$

$$= \int \frac{g(\theta_i)}{k(Y, \theta_i)} q(\theta_i \mid Y, i) d\theta_i$$

$$= \int \frac{g(\theta_i)}{k(Y, \theta_i)} \frac{k(Y, \theta_i)}{\int k(Y, \theta_i) d\theta_i} d\theta_i$$

Choosing g

- The method we have described is called the **modified harmonic mean** method.
- The original idea, called the harmonic mean method, was to use $\pi()$ as the weighting function.
- Since $k(Y, \theta) = p(Y \mid \theta_i)\pi(\theta_i)$, this amounted to simply taking the sample mean of $1/p_i(Y \mid \theta_i)$, i.e. of one over the likelihood.
- So long as g integrates to one, the expectation of $g/(p_i\pi_i)$, under the posterior density on θ_i, exists and is finite. We already proved this.

- But there is no guarantee that its second moment is finite. For example, if g declines slowly as $|\theta_i|$ goes to infinity, i.e. if g has "fat tails", while k does not, then g/k will be unbounded. It will have finite expectation, but only because the very large values of g/k that will eventually occur will occur rarely — because we are sampling from the thin-tailed k kernel.

- The rapidly declining k kernel offsets the k in the denominator of g/k. It is not enough to offset the k^2 in the denominator of g^2/k^2. So if $\lim_{\theta_i \to \infty} g^2/k > 0$, g/k has infinite second moment.

- This means usual measures of uncertainty, based on variances, do not apply, and that convergence of the sample mean to its expected value will be slow.

Geweke's suggested g

- Convergence will be more rapid the closer g/k is to a constant.

- Geweke suggested taking g to be the standard normal approximation to the posterior density at its peak $\hat{\theta}_i$— i.e. to be a

$$N\left(\hat{\theta}_i, -\left(\frac{\partial^2 \log(k(Y,\theta_i)}{\partial\theta_i \partial\theta_i}\right)^{-1}\right)$$

  density, but to truncate it at $(\theta_i - \hat{\theta}_i)' V^{-1} (\theta_i - \hat{\theta}) = \kappa$ for some κ, where V is the asymptotic covariance matrix. The integral of the normal density over this truncated region can be looked up in a chi-squared

distribution table, so the density can easily be rescaled to integrate to exactly one over this region.

- The idea is that g should be close to k near the peak, while the truncation avoids any possibility that g/k becomes unbounded in the tails.

Remaining problems with g choice

- Geweke's idea often works well, but not always.

- The tails are not the only place that g/k might become unbounded. If there is a point θ^* at which $k(\theta^*) = 0$, and if k is twice-differentiable at that point, while g is continuous at that point, then g/k has infinite variance under the posterior density. In high-dimensional models it may be difficult to know whether there are points with $k = 0$.

- With, say, 30 parameters, the typical draw of θ_i^j will have $(\theta_i^j - \hat{\theta}_i)' V^{-1} (\theta_i^j - \hat{\theta}_i) \doteq 30$ (because this quanitity has a chi-squared(30) distribution.) But this means that the density at $\theta_i = \hat{\theta}_i$ is e^{30} times larger than the density at a typical draw.

- Of course if the ratio of k at the peak and at a typical draw is also of this size, this causes no problem, but obviously there is tremendous room for g/k to vary, and hence for convergence of the estimate of the Bayes factor to be slow.

- There are ways to do better, if sampling g/k is not working well: bridge sampling, e.g.

- An important dumb idea: What about just taking sample averages of $p(Y \mid \theta_i^j)$ from the simulated posterior distribution? (Important because every year one or more students does this on an exam or course paper.)

The Schwarz Criterion

In large samples, under fairly general regularity conditions, the posterior density is well approximated as proportional to a $N(\hat{\theta}, V)$ density, where V is minus the inverse of the second derivative matrix for the log posterior density, evaluated at the MLE $\hat{\theta}$. We know the integral of a Normal density, so if the approximation is working well, we will find that the integral of the posterior density is approximately

$$(2\pi)^{k/2} \, |V|^{1/2} \, p(Y \mid \hat{\theta}) \pi(\hat{\theta}) \; .$$

V convergence

In a stationary model, $V \cdot T \xrightarrow[T\to\infty]{} \bar{V}$, where $\bar{V}$ is a constant matrix. This follows because

$$\log\big(p(Y \mid \theta)\pi(\theta)\big) = \sum_{t=1}^{T} \log\big(p(Y_t \mid \{Y_s, s < t\}, \theta)\big) \tag{1}$$

$$\therefore \frac{1}{T}\frac{\partial^2 k(Y \mid \theta)}{\partial\theta\partial\theta'} \xrightarrow{a.s.} E\left[\frac{\partial^2 \log\big(p(Y_t \mid \{Y_s, s < t\}, \theta)\big)}{\partial\theta\partial\theta'}\right] \tag{2}$$

and V is just minus the inverse of this second derivative matrix. Of course above we have invoked ergodicity of the Y process and also assumed that Y_t depends on Y_{t-s} for only finitely many lags s, so that after the first few observations all the terms in the sum that makes up the log likelihood are of the same form.

The Schwarz Criterion, II

Now we observe that since $\log\left\{\bar{V}/T\right\} = -n\log T + \log\left\{\bar{V}\right\}$ (where n is the length of θ), the log of the approximate value for the integrated density is

$$\frac{n}{2}\log(2\pi) - \frac{n}{2}\log T + \frac{1}{2}\log\left\{\bar{V}\right\} + \log(k(Y,\hat{\theta}))\ .$$

Dominating terms when taking difference across two models:

$$\log\big(k_1(Y,\theta_1)\big) - \log\big(k_2(Y,\theta_2)\big) - \frac{n_1 - n_2}{2}\log(T)$$

Comments on the SC

- The formula applies whether the models are "nested" or not.
- Frequentist model comparison, for the special case where one model is just a restriction of the other, treats the difference of log likelihoods, times 2, as chi-squared$(n_1 - n_2)$ and rejects the restricted model if this statistic exceeds the $1 - \alpha$ per cent level of the chi-squared distribution.
- The Schwarz criterion compares this statistic to a level that increases not just with $n_1 - n_2$, but also with $\log T$.
- In other words, Bayesian model comparison penalizes more richly parameterized models in large samples, and does so more stringently (relative to a frequentist likelihood ratio test) the larger the sample.

The Lindley Paradox

- Take a small collection of reasonable models, calculate Bayesian posterior odds. A typical result: One model has probability $1 - \varepsilon$, the others probability less than ε, with ε *very* small (say .00001).

- Frequentist comparisons (for the nested case where they work) seem not to give such drastic answers.

- One interpretation — Bayesian methods are more powerful.

- Another — Bayesian methods produce unreasonably sharp conclusions.

Interpreting and accounting for the Lindley Paradox

- The reason it's a problem is that these sharp results are often unstable: If, say, we introduce one more model, we may find that the new one is the one with $1 - \varepsilon$ probability, which makes our previous near certainty look foolish.

- In other words, the problem is that the result is conditional on our being certain that the set of models compared contains the truth, that there are no other possibilities waiting in the wings.

- Gelman, Carlin, Stern and Rubin suggest (almost) never doing model comparison: Their view is that it should (almost) always be reasonable, when comparing models, to make each model a special case of a larger model with a continuously varying parameter.

- In other words, when you think you want to use a $\{0, 1\}$-valued "model number" parameter, figure out how to replace this with a continuously varying parameter, say γ, such that the two models arise as $\gamma = 1$ and $\gamma = 0$.