BAYESIAN ECONOMETRICS

1. Outline

(I) The difference between Bayesian and non-Bayesian inference.
(II) Confidence sets and confidence intervals. Why?
(III) Bayesian interpretation of frequentist data analysis.
(IV) A case where Bayesian and frequentist approaches seem aligned: SNLM
(V) Cases where they conflict: Unit root time series models; breaks.
(VI) MCMC basics
(VII) Inference for DSGE's.

2. Bayesian Inference is a Way of Thinking, Not a Basket of "Methods"

- Frequentist inference makes only pre-sample probability assertions.
  - A 95% confidence interval contains the true parameter value with probability .95 only *before* one has seen the data. After the data has been seen, the probability is zero or one.
  - Yet confidence intervals are universally interpreted in practice as guides to *post*-sample uncertainty.
  - They often are reasonable guides, but only because they often are close to posterior probability intervals that would emerge from a Bayesian analysis.
- People want guides to uncertainty as an aid to decision-making. They want to characterize uncertainty *about parameter values*, given the sample that has actually been observed. That it aims to help with this is the distinguishing characteristic of Bayesian inference.

3. Is the difference that Bayesian methods are subjective?

- No.
- The objective aspect of Bayesian inference is the set of rules for transforming an initial distribution into an updated distribution conditional on observations.
- Bayesian thinking makes it clear that for decision-making, pre-sample beliefs are therefore in general important.
- But most of what econometricians do is not decision-making. It is reporting of data-analysis for an audience that is likely to have diverse initial beliefs.
- In such a situation, as was pointed out long ago by Hildreth (1963) and Savage (1977, p.14-15), the task is to present useful information about the shape of the likelihood.

Date: October 20, 2009.

4. How to characterize the likelihood

- Present its maximum.
- Present a local approximation to it based on a second-order Taylor expansion of its log. (Standard MLE asymptotics.)
- Plot it, if the dimension is low.
- If the dimension is high, present slices of it, marginalizations of it, and implied expected values of functions of the parameter. The functions you choose might be chosen in the light of possible decision applications.
- The marginalization is often more useful if a simple, transparent prior is used to downweight regions of the parameter space that are widely agreed to be uninteresting.

5. Confidence set simple example number 1

- $X \sim U(\beta + 1, \beta - 1)$
- β known to lie in (0,2)
- 95% confidence interval for β is $(X - .95, X + .95)$
- Can truncate to (0,2) interval or not — it's still a 95% interval.

6. Example 1, cont.

- Bayesian with $U(0, 2)$ prior has 95% posterior probability ("credibility") interval that is generally a subset of the intersection of the $X \pm 1$ interval with (0,2), with the subset 95% of the width of the interval.
- Frequentist intervals are simply the intersection of $X \pm .95$ with (0,2). They are wider than Bayesian intervals for X in $(0, 2)$, but narrower for X values outside that range, and in fact simply vanish for X values outside $(-.95, 2.95)$.

7. Are narrow confidence intervals good or bad?

- A 95% confidence interval is always a collection of all points that fail to be rejected in a 5% significance level test.
- A completely empty confidence interval, as in example 1 above, is therefore sometimes interpreted as implying "rejection of the model".
- As in example 1, an empty interval is approached as a continuous limit by very narrow intervals, yet very narrow intervals are usually interpreted as implying very precise inference.

8. Confidence set simple example number 2

- $X = \beta + \varepsilon$, $\varepsilon \sim e^{-\varepsilon}$ on $(0, \infty)$.
- likelihood: $e^{-X+\beta}$ for $\beta \in (-\infty, X)$
- Bayesian credible sets show reversed asymmetry
- contrast with naive (but commonly applied) bootstrap, which would produce an interval entirely concentrated on values of β *above* $\hat{\beta}_{MLE} = X$, which are impossible.
- (naive parametric bootstrap: Find an estimator $\hat{\beta}$. Simulate draws from $X \mid \hat{\beta}$, take the 2.5% tails of this distribution as the 95% confidence interval.)

9. Confidence set simple example number 3

- $Y \sim N(\beta, 1)$; we are interested in $g(\beta)$, where g is nonlinear and monotone, but unknown.
- We observe not Y, but $X = g(Y)$.
- We have a maximum likelihood estimator $\hat{g}$ for $g(\beta)$.
- If we could observe Y, a natural confidence and credible interval for β would be $Y \pm 1.96$. If we also knew g, we could then use $(g(Y - 1.96), g(Y + 1.96))$ as an excellent confidence and credible interval.
- Using the naive bootstrap here, if we base it on an estimator $\hat{g}$ asymptotically equivalent to the MLE, gives exactly the natural interval.
- So, without an analysis of likelihood, there is no answer as to whether the naive bootstrap gives reasonable results.

10. Bayesian interpretation of frequentist asymptotics

- A common case: $\sqrt{T}(\hat{\beta}_T - \beta) \mid \theta \xrightarrow[T\to\infty]{\mathcal{D}} N(0, \Sigma)$
- Under mild regularity conditions, this implies $\sqrt{T}(\beta - \hat{\beta}_T) \mid \hat{\beta}_T \xrightarrow[T\to\infty]{\mathcal{D}} N(0, \Sigma)$
- Kwan (1998) presents this standard case. Kim (2002) extended the results beyond the $\sqrt{T}$ normalization case and thereby covered time series regression with unit roots.

11. SNLM

The SNLM often denoted by the equation $Y = X\beta + \varepsilon$, asserts the following conditional pdf for the vector of Y data conditional on the matrix of X data and on the parameters β, σ^2:

$$p(\underset{T\times 1}{Y} \mid \underset{T\times k}{X}) = \varphi(Y - X\beta; \sigma^2 I) = (2\pi)^{-T/2}\sigma^{-T} \exp\left(\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}\right)$$

The most common framework for Bayesian analysis of this model asserts a prior that is flat in β and $\log\sigma$ or $\log\sigma^2$, i.e. $d\sigma/\sigma$ or $d\sigma^2/\sigma^2$. We will assume the prior has the form $d\sigma/\sigma^p$, then discuss how the results depend on p.

12. Marginal for σ^2

We let $u(\beta) = Y - X\beta$ and denote the least squares estimate as $\hat{\beta} = (X'X)^{-1}X'Y$. Also $\hat{u} = u(\hat{\beta})$. Then the posterior can be written, by multiplying the likelihood above by σ^{-p}

and rearranging, as proportional to

$$\sigma^{-T-p}\exp\left(-\frac{\hat{u}'\hat{u}+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2\sigma^2}\right)d\beta\,d\sigma$$
$$\propto\sigma^{-T-p+k}\left|X'X\right|^{-\frac{1}{2}}\exp\left(-\frac{\hat{u}'\hat{u}}{2\sigma^2}\right)\varphi\big(\beta-\hat{\beta};\sigma^2(X'X)^{-1}\big)\,d\beta\,d\sigma$$
$$\propto v^{(T+p-k)/2}\exp\left(\frac{-\hat{u}'\hat{u}}{2}v\right)\left|X'X\right|^{-\frac{1}{2}}\varphi\big(\beta-\hat{\beta};\sigma^2(X'X)^{-1}\big)\,d\beta\,\frac{dv}{v^{3/2}}\,,$$

where $v=1/\sigma^2$.

13.

Integrating this expression w.r.t. β and setting $\alpha=\hat{u}'\hat{u}/2$ gives us an expression proportional to

$$v^{(T+p-k-3)/2}\exp\left(-\frac{\hat{u}'\hat{u}}{2}v\right)dv\propto\alpha^{(T+p-k-1)/2}v^{(T+p-k-3)/2}e^{-\alpha v}dv\,,$$

which is a standard $\Gamma((T+p-k-1)/2,\alpha)$ pdf.

Because it is $v=1/\sigma^2$ that has the Γ distribution, we say that σ^2 itself has an **inverse-gamma** distribution. Since a $\Gamma(n/2,1)$ variable, multiplied by 2, is a $\chi^2(n)$ random variable, some prefer to say that $\hat{u}'\hat{u}/\sigma^2$ has a $\chi^2(T-k)$ distribution, and thus that σ^2 has an inverse-chi-squared distribution.

14. Marginal on β

Start with the same rearrangement of the likelihood in terms of v and β, and rewrite it as

$$v^{(T+p-3)/2}\exp\left(-\frac{1}{2}u(\beta)'u(\beta)v\right)dv\,d\beta\,.$$

As a function of v, this is proportional to a standard $\Gamma\big((T+p-1)/2,u(\beta)'u(\beta)/2\big)$ pdf, but here there is a missing normalization factor that depends on β. When we integrate with respect to v, therefore, we arrive at

$$\left(\frac{u(\beta)'u(\beta)}{2}\right)^{-(T+p-1)/2}d\beta\quad\propto\quad\left(1+\frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{\hat{u}'\hat{u}}\right)^{-(T+p-1)/2}d\beta\,.$$

This is proportional to what is known as a multivariate $t_n\big(0,(\hat{u}'\hat{u})/n\big)$ pdf, where $n=T+p-k-1$ is the degrees of freedom. It makes each element β_i of β an ordinary univariate $t_n(\hat{\beta}_i,s_\beta^2)$, where $s_\beta^2=s^2(X'X)^{-1}_{ii}$ and $s^2=\hat{u}'\hat{u}/n$. Thus the statistics computed from the data can be analyzed with the same tables of distributions from either a Bayesian or non-Bayesian perspective.

15. Unit roots

- Time series models with possible unit roots are an area of application where Bayesian and frequentist approaches lead to quite different calculations and reporting of evidence.
- In an autoregressive model with Gaussian disturbances, the likelihood, conditional on initial conditions, is Gaussian in shape regardless of whether the model is stationary or non-stationary.
- This suggests using the usual OLS estimates and standard errors in such models to summarize the likelihood shape.
- Frequentist theory shows that the asymptotic distribution of the OLS estimators changes discontinuously at the unit root boundary of the parameter space, suggesting that when unit roots may be present the usual OLS standard errors and test statistics cannot be used.

16. Unit roots, cont.

- Sims and Uhlig (1991) explains geometrically how the frequentist and Bayesian results can coexist.
- The Bayesian perspective suggests that *priors* might need to pay special attention to the unit root boundary in the parameter space. It also suggests that using the initial conditions for inference, rather than just the likelihood conditional on initial conditions, is important, and difficult when unit roots may be present.
- But these considerations are quite different from what makes possibly non-stationary models special from the frequentist perspective.

17. Breaks

- Suppose we have a model specifying Y_t is i.i.d. with a pdf $f(y_t; \mu_t)$, and with μ_t taking on just two values, $\mu_t = \underline{\mu}$ for $t < t^*$, $\mu_t = \bar{\mu}$ for $t \geq t^*$.
- The pdf of the full sample $\{Y_1, \ldots, Y_T\}$ therefore depends on the three parameters, $\underline{\mu}, \bar{\mu}, t^*$.
- If f has a form that is easily integrated over μ_t and we choose a prior $\pi(\underline{\mu}, \bar{\mu})$ that is conjugate to f (meaning it has the same form as the likelihood), then the posterior marginal pdf for t^* under a flat prior on t^* is easy to calculate: for each possible value of t^*, integrate the posterior over $\underline{\mu}, \bar{\mu}$.
- The plot of this integrated likelihood as a function of t^* gives an easily interpreted characterization of uncertainty about the break date.
- Frequentist inference about the break date has to be based on asymptotic theory, and has no intepretation for any observed complications (like multiple local peaks, or narrow peaks) in the global shape of the likelihood.

18. Posterior simulation

- For the situation where we know $\pi(\theta)p(Y \mid \theta) = f(y, \theta)$, in the sense that we can write an expression or program that gives its value for each possible value of θ, but we do not know how to draw randomly from the pdf in θ defined by $f(Y,\theta)/\int f(Y,\theta)d\theta$, because it defines no standard distribution.

- Two main approaches: Importance sampling and Markov chain Monte Carlo (MCMC).

19. Importance Sampling

- Suppose θ has pdf $p(\theta)$, defining a non-standard distribution. We would like to calculate the expectation of a function $g(\theta)$ under the distribution defined by p.
$$E_p[g(\theta)] = \int g(\theta)p(\theta)d\theta = \int \frac{g(\theta)p(\theta)}{q(\theta)}q(\theta)d\theta = E_q\left[\frac{g(\theta)p(\theta)}{q(\theta)}\right] .$$
for any pdf $q(\theta)$.
- So estimate the expectation of θ under p by drawing randomly from the distribution defined by q and weighting the draws of $g(\theta)$ by $p(\theta)/q(\theta)$.
- requires that $p(\theta) > 0 \Rightarrow q(\theta) > 0$. In fact there are problems even if p/q just becomes very large in parts of the parameter space, because this tends to make a few, rare draws completely dominate the weighted average.

20. MCMC: General principles

- Given a draw θ_j, one generates a new draw θ_{j+1} from a distribution that may depend on θ_j (but not on earlier draws). The draws are generally serially correlated across j (unlike the importance sampling draws), but eventually their sample distribution function converges to that of the target distribution.
- Need to have the target a fixed point. Often proving this can proceed by showing that, when θ_j is drawn from the target p_0 pdf, the transition mechanism implies that the joint pdf $p(\theta_{j+1},\theta_j)$ satisfies $p(\theta_{j+1},\theta_j) = p(\theta_j,\theta_{j+1})$.

21. More principles

- But then need also to insure that the algorithm will not get stuck. This will depend on the particular algorithm, on the shape of the boundaries of the parameter space, and on the nature of the target pdf.
- Can't even get the target to be a fixed point if the target is not integrable. Note that we do not need to know the target's scale in order to implement these algorithms, so failing to detect non-integrability is a real possibility.
- These methods really do require the Markov property. One can be tempted to systematically tune up the algorithm based on what has been learned about the target distribution from previous draws. If this is done systematically and repeatedly, it makes the algorithm deliver wrong answers.

22. Checking convergence and accuracy

- Accuracy: *assuming* current sample is representative, do we have enough accuracy for our purposes?
- Accuracy can be different for different functions of β in the same sample.
- Convergence: Can we treat this sample as "representative", i.e. as having visited all relevant regions and displayed all relevant modes of variation?

23. Effective sample size

- Assuming convergence, we can estimate a model (e.g. an AR or VAR), use it to generate the ACF $R_\theta(j)$ of the draws, and compute

$$\operatorname{Var}\left(\frac{1}{T}\sum\theta^j\right) = \frac{1}{T}\sum_{j=-T}^{T}\frac{|T-j|}{T}R_\theta(j) \doteq \frac{1}{T}\sum_{-\infty}^{\infty}R_\theta(j)$$

- This can be compared to what the variance of the mean θ_j would have been if the sample were i.i.d.:

$$\frac{1}{T}R_\theta(0)\,.$$

- Then we can use as a measure of "effective sample size"

$$\frac{TR_\theta(0)}{\sum R_\theta(j)}\,.$$

- It is not uncommon for MCMC to produce effective sample sizes of only a few hundred when the number of draws is in the hundreds of thousands. This implies such great dependence across j that even the effective sample size numbers, not to mention the mean of θ^j, is unreliable. It is likely to be a symptom of non-convergence.

24. Another approach to effective sample size

- Break the sample into k pieces indexed by i. Calculate sample average $\bar{g}_i$ of $g(\beta_j)$ for each piece. $1/\sqrt{k}$ times sample standard deviation of the $\bar{g}_i$'s is an estimate of the standard error of the sample mean from the overall sample.
- This is accurate to the extent that the pieces are long enough so that dependence between them is weak. Span of dependence among β_j draws must be considerably shorter than the length of the pieces.
- Variance from the whole sample, if had i.i.d. sample, would be N/k times the variance of sample means across pieces. So if s_k^2 is the sample variance of the means of the pieces and s_N^2 the sample variance from the whole sample, effective sample size is ks_N^2/s_k^2. This could in principle be larger than N, but in practice is usually much smaller than N.

25. Convergence checks based on subsample blocks

- That effective sample size is similar with different choices of k and is growing more or less linearly with N is a criterion for convergence.
- We are more suspicious of non-convergence in the beginning blocks of an MCMC chain. So some convergence checks (e.g. one suggested by Geweke) compare behavior of draws in the first part of the sample to behavior in the latter part.

26. Convergence checks based on multiple chains

- Start from different places.

- After one or two, start from a place that is fairly unlikely according to initial runs. Variation across runs from different starting points can be treated like variation across pieces of the sample.
- Often this leads to different conclusions about accuracy and convergence than working with pieces of a single run.

27. Is accuracy sufficient?

- If convergence is ok, the sample size may or not be big enough: That depends on whether the estimated accuracy of your estimate of $E[\beta]$ is within tolerances based on substantive considerations.
- Accuracy may be adequate for some g's and not others. Effective sample size may differ across g's. But if convergence looks bad for one g, it should not be very comforting that for other g's it looks ok.

28. Trimming

- Thinning. If effective sample size is running at about $N/10$, why not throw out all but every 10'th draw?
- This will make the result look more like an i.i.d. sample, but will not improve, and may harm, accuracy of estimates of $E[g(\beta_j)]$.
- However, it is often done, because the cost of moderate thinning (more like every 3rd or 4th draw, here) in reduced accuracy will be relatively small compared to the savings in disk space, if all the draws are being saved.

29. Trace plots

- These simply plot elements of θ^j, or functions $g(\theta^j)$, against j. They will show clearly trending behavior, or slow oscillations, or switches between regions with different characteristics. The `coda` package (there is an R version, but also an Octave/Matlab version) generates these with a single command.
- They are not foolproof. If there are large high-frequency oscillations, they may obscure trends and low-frequency oscillations. In large MCMC samples the plot may even look like a black smear.
- If effective sample size is small, but the trace plots are black, it may help to do trace plots of thinned samples.

30. Metropolis algorithm

Target kernel: $p(\theta)$
Proposal density: $q(\theta' \mid \theta)$
Procedure:

(1) Generate a proposed θ^*_{j+1} from θ_j from the $q(\theta^*_{j+1} \mid \theta_j)$ distribution.
(2) Calculate $\rho = p(Y \mid \theta^*_{j+1})/p(Y \mid \theta_j)$.
(3) Draw u from a $U(0,1)$ distribution.
(4) If $\rho \geq u$, set $\theta_{j+1} = \theta^*_{j+1}$. Otherwise $\theta_{j+1} = \theta_j$.

31. Proof of fixed point property for Metropolis

We want to show that if θ is drawn from the target density $p(\theta)$, and θ' is then drawn by the Metropolis algorithm conditioned on θ as the previous draw using a q that is symmetric, i.e. with $q(\theta' \mid \theta) = q(\theta \mid \theta')$, the resulting joint distribution is symmetric in θ and θ'. This implies that θ' has the same marginal distribution as θ, i.e. that given by the target density p.

The distribution consists of a certain amount of probability concentrated on the $\theta = \theta'$ subspace, plus a density over the rest of the parameter space. The part on the $\theta = \theta'$ subspace is obviously symmetric in θ and θ', so we can confine our attention to the density over the rest.

32. Symmetry of the density

Consider the region in which $p(\theta) < p(\theta')$. In this region, draws are never rejected. Since the draws are not rejected in this region, in this region the joint density is just $q(\theta' \mid \theta)\pi(\theta)$.

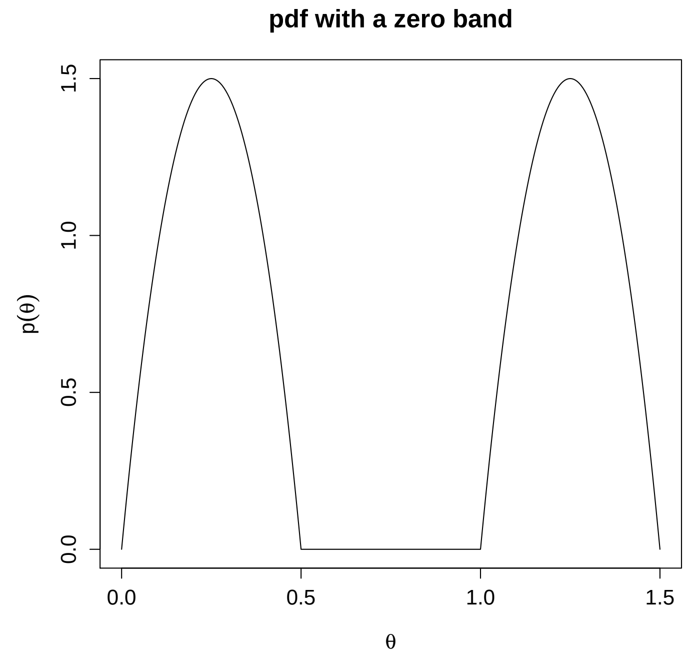
Now consider the part of the parameter space, in which $p(\theta') < p(\theta)$. In this region, a proposal draw θ' is rejected with probability $p(\theta')/p(\theta)$, so the joint density is

$$(p(\theta')/p(\theta))q(\theta' \mid \theta)p(\theta) = p(\theta')q(\theta' \mid \theta) = p(\theta')q(\theta \mid \theta')$$

where the last equality invokes the symmetry condition. Since $p(\theta') < p(\theta)$, this last expression is exactly the density at the symmetric opposite point — where θ' is the previous point and θ is the proposal. □

33. How to pick jump distributions, pitfalls

If the parameter space Θ has pieces separated by bands with zero probability under the target density, the proposal density must have large enough support to jump across the bands.



Obviously for the pictured target density, the jump density has to have support spanning an interval of length greater than .5.

34. More on jump distributions

- Even with a jump density with unbounded support, like a normal density, if the probability is concentrated mainly in a small region around 0, it may take an extremely long time before the Metropolis iterations actually start sampling the second lobe of the distribution.
- In a multivariate distribution that is approximately Gaussian in shape in its high-probability region, practice suggests that finding the local Gaussian approximation in the neighborhood of the peak $\hat{\theta}$, i.e. a $N(\hat{\theta}, \Sigma)$ with $\Sigma = -(\partial^2 \log(f(Y,\theta))/\partial\theta\partial\theta')^{-1}$, and then taking the jump distribution to be $N(0, k\Sigma)$, with k about .3, is a practical starting point.

35. Picking k

- By making k very small, we can make the acceptance rate for draws very high, but since the steps the algorithm takes will be very small, serial correlation will be high and convergence slow.
- By making k very large, we can make the steps *proposed* large, but this may make the rate at which proposals are rejected very high, which again will create serial correlation and slow convergence.
- And this strategy of using the normal approximation around the peak only makes sense if the target density is in fact somewhat Gaussian in shape.
- For a bimodal distribution, one might use a jump distribution with a $k\Sigma$ that varied with θ, matching the local expansion at one mode when the last draw was near that mode and matching the local expansion at the other node when near that one.

36. Metropolis-Hastings

Removes the symmetry requirement from the Metropolis algorithm.

(1) Draw θ^*_{j+1} from $q(\cdot \mid \theta_j)$.
(2) Construct

$$\rho = \frac{p(\theta_{j+1})q(\theta_j \mid \theta_{j+1})}{p(\theta_j)q(\theta_{j+1}/\theta_j)} .$$

(3) If $\rho > 1$, $\theta_{j+1} = \theta^*_{j+1}$.
(4) Otherwise $P[\theta_{j+1} = \theta^*_{j+1}] = \rho$; $P[\theta_{j+1} = \theta_j] = 1 - \rho$.

37. Gibbs sampling

- Suppose our parameter vector has two blocks, i.e. $\theta = (\theta_1, \theta_2)$, and that we use as our proposal distribution $q(\theta' \mid \theta) = p(\theta'_2 \mid \theta_1)$, holding $\theta'_1 = \theta_1$.

- Since this makes the ratio of the proposal density to the target density constant, it satisfies the conditions for an M-H algorithm if we accept every draw.

- If on the next draw we draw from $p(\theta'_1 \mid \theta_2)$ and then continue to alternate, the algorithm is likely (subject to regularity conditions) to converge.

- Obviously this is a useful idea only if, despite p being non-standard, the conditional densities for the blocks are standard. This happens more than you might expect.

38. Metropolis-within-Gibbs, etc.

- The fixed-point theorems for these algorithms concern a single step.
- Therefore the fixed point property holds even if we change the algorithm, say alternating Metropolis with Gibbs, with M-H, etc.
- Also Gibbs can obviously be done for more than two blocks.
- A common situation: Gibbs can be done for k of n blocks, but there is a small number $n - k$ of blocks for which the conditional distributions are non-standard.
- One can then do straight Gibbs for the k blocks, and for the other $n - k$ use Metropolis steps with the conditional densities as targets.

39. Model comparison: in principle

- When we have a collection of n models indexed by $j = 1, \dots, n$, all explaining the same vector y_t of time series, we can treat the model number j as just one additional parameter.
- This is formally no different from the "break date" model we discussed earlier.
- And the formal solution is the same: for each j, integrate prior times posterior over the other parameters; the result, normalized to sum to one, is the posterior probabilities of the models.

40. Model comparison: conceptual difficulties

- It is often found in practice that in a collection of models that all seem to fit fairly well by conventional measures of fit, one model nonetheless has posterior probability close to one.
- If the models were in fact a complete collection of all possible models, this would be a desirable outcome, of course.
- But more commonly a collection of models is just a sample from the space of possible models.
- When the models are complex and hard to solve, the collection is particularly likely to be a sparse subset of the all the possible models.
- In this case, the high probability attached to one of the models is misleading — There are likely to be other models that, if they were included in the collection, would also have high probability.
- Some Bayesian statisticians argue that it is hardly ever reasonable to form posterior probabilities over collections of models. Instead, some way should be found to parameterize the differences among the models continuously, which would avoid the misleading pileup of probability on one model.
- In practice, such a continuously parameterized super-model may not be feasible, but we should recognize then that the posterior probabilities from Bayesian odds calculations are not reliable decision-making posterior probabilities.
- This does not imply that there is some other, better, way to measure model fit. The cure for over-sharp conclusions from posterior probabilities is always to expand the collection of models, not to resort to some ad hoc measure of fit.

41. Model Comparison: how to do the computation

- Generally we don't know how to do the required integrations analytically.
- It is convenient if we can use the same set of MCMC draws we use to do inference on parameters within each model to find the posterior weights on the models.
- A method to do this: the "modified harmonic mean" (MHM).
- If $g(\cdot)$ is a pdf defined on the parameter space, then

$$\int \frac{g(\theta)}{\pi(\theta)p(Y \mid \theta)} \frac{\pi(\theta)p(Y \mid \theta)}{\int \pi(\theta')p(Y \mid \theta')\,d\theta'}\,d\theta$$

$$= \int \frac{g(\theta)}{\int \pi(\theta')p(Y \mid \theta')\,d\theta'}\,d\theta = \left(\int \pi(\theta')p(Y \mid \theta')\,d\theta'\right)^{-1}$$

- If we have a set of MCMC draws $\{\theta_j\}$ and values of the posterior weight $\pi(\theta_j)p(Y \mid \theta_j)$ for each draw, we can estimate the integral in this expression by taking sample averages of $g(\theta_k)/(\pi(\theta_j)p(Y \mid \theta_j)$. The result then estimates the inverse of the integrated posterior density.

42. MHM pitfalls

- If $\pi(\theta)p(Y \mid \theta)$ approaches zero smoothly at a value of θ where $g(\theta) > 0$, or if $\pi(\theta)p(Y \mid \theta)$ approaches zero in its tails faster than $g(\theta)$, then the random variable

$g(\theta_j)/(\pi(\theta_j)p(Y \mid \theta_j)$ is likely to have infinite variance. The sample averages are guaranteed to converge nonetheless, because the integral is by construction finite. But convergence may be very slow if the variance is infinite.

- This is why we use the "modified" harmonic mean. The original harmonic mean method took $g(\theta) = \pi(\theta)$, which seems initially appealing, but it is very likely that $\pi(\theta)$ is more spread out than the posterior density and thus has many large $g/(\pi p)$ values and likely infinite variance.
- Geweke suggested a widely applied g: Take the Gaussian local approximation to the posterior density around its peak, but truncate it at some reasonable choice of level curve of the multivariate Gaussian density. The integral of this truncated density can easily be normalized to one by looking up the integral in a χ-squared test table.
- Geweke's method takes care of the problem of g possibly having fatter tails than $p\pi$, but in highly multivariate models it can easily happen that $p\pi$ is zero at locations in the parameter space that are hard to characterize in advance. Then MHM may converge slowly.

43. NON-MHM APPROACHES

- One approach is "bridge sampling".
- Another is to use importance sampling, drawing a sample from some proposal density $q((\theta)$ and evaluating the sample average of $\pi(\theta_j)p(Y \mid \theta_j)/q(\theta_j)$ for that sample.
- Both these approaches, unlike MHM, require additional random sampling, beyond the MCMC draws from the posterior. Importance sampling has pitfalls of its own. Bridge sampling attempts to systematically work around the problem of $g/(\pi p)$ blowing up at some parameter values.

44. APPLICATION TO DSGE'S

MCMC Bayesian methods were first applied to macroeconomic DSGE's in a series of papers by Frank Smets and Raf Wouters. They have since been applied widely by other researchers. Handy computational tools for solving DSGE's and estimating them with MCMC methods are available in the DYNARE package, which is freely accessible on the internet.

45. DEL NEGRO/SCHORFHEIDE

- Marco Del Negro and Frank Schorfheide have proposed a new approach to dynamic macro policy modeling.
- They use a DSGE not directly as a model of the data, but instead as a device to generate a prior for a structural VAR (SVAR).

REFERENCES

HILDRETH, C. (1963): "Bayesian Statisticians and Remote Clients," *Econometrica*, 31(3), 422–438.

KIM, J.-Y. (2002): "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, 107, 175–193.

KWAN, Y. K. (1998): "Asymptotic Bayesian analysis based on a limited information estimator," *Journal of Econometrics*, 88, 99–121.

SAVAGE, L. J. (1977): "The Shifting Foundations of Statistics," in *Logic, Laws and Life*, ed. by R. Colodny, pp. 3–18. University of Pittsburgh Press.

SIMS, C. A., AND H. D. UHLIG (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59(6), 1591–1599.