PRINCIPLES OF BAYESIAN ANALYSIS

AAEC 6564
INSTRUCTOR: KLAUS MOELTNER

Textbooks: Koop (2003), Ch.1; Koop et al. (2007), Ch.1-2; Hoff (2009), Ch. 1
`Matlab` scripts: `mod1s1a, mod1s1apublish`

Bayesian vs. Classical Estimation

Bayesian data analysis is distinctly different from classical (or "frequentist") analysis in its treatment of probabilities, and in its resulting treatment of model parameters when compared to classical parametric analysis.[1]

Bayesian analysts formulate probabilistic statements about uncertain events before collecting any additional evidence (i.e. "data"). These ex-ante probabilities (or, more generally, probability distributions plus underlying parameters) are called *priors*. This notion of *subjective probabilities* is absent in classical estimation. In the classical world, all estimation and inference is based solely on observed data.

Both Bayesian and classical econometricians aim to learn more about a set of parameters, say $\boldsymbol{\theta}$. In the classical mindset, $\boldsymbol{\theta}$ contains fixed but unknown elements, usually associated with an underlying population of interest (e.g. the mean and variance for credit card debt amongst U.S. college students). Bayesians share with Classicals the interest in $\boldsymbol{\theta}$ and the definition of the population of interest. However, they assign ex ante a prior probability to $\boldsymbol{\theta}$, labeled $p(\boldsymbol{\theta})$, which usually takes the form of a probability distribution with "known" moments. For example, Bayesians might state that the above stated debt amount has a normal distribution with mean \$3000 and standard deviation of \$1500. This prior may be based on previous research, related findings in the published literature, or it may be completely arbitrary. In any case, it's an inherently subjective constructs.

Both schools then develop a theoretical framework that relates $\boldsymbol{\theta}$ to observed data, say a "dependent variable" $\mathbf{y}$, and a matrix of explanatory variables $\mathbf{X}$. This relationship is formalized via a likelihood function, say $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$ to stay with Bayesian notation. To stress, this likelihood function takes the exact same analytical form for both schools.

The Classical analyst then collects a sample of observations from the underlying population of interest and, combining these data with the formulated structural model, produces an estimate of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$. Any and all uncertainty surrounding the accuracy of this estimate is solely related to the notion that results are based on a sample, not data for the entire population. A different

[1] Throughout this course we will largely remain within the realm of parametric analysis. However, students should note that there also exists a variety of Bayesian methods for non- and semiparameteric modeling.

sample (of equal size) may produce slightly different estimates. Classicals express this uncertainty via "standard errors" assigned to each element of $\hat{\boldsymbol{\theta}}$. They also have a strong focus on the behavior of $\hat{\boldsymbol{\theta}}$ as the sample size increases. The behavior of estimators under increasing sample size falls under the heading of "Asymptotic Theory". The properties of most estimators in the Classical world can only be assessed "asymptotically", i.e. are only understood for the hypothetical case of an infinitely large sample. Also, virtually all specification tests used by Frequentists hinge on Asymptotic Theory.

Bayesians, in turn, combine prior and likelihood via Bayes' Rule to derive the *posterior distribution* of $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{p(\boldsymbol{\theta},\mathbf{y}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} = \frac{p(\boldsymbol{\theta})\,p(\mathbf{y}|\boldsymbol{\theta},\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} \propto p(\boldsymbol{\theta})\,p(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}) \tag{1}$$

Simply put, the posterior distribution is just an updated version of the prior. If the data have high informational content (i.e. allow for substantial learning about $\boldsymbol{\theta}$), the posterior will generally look very different from the prior. In most cases, it is much "tighter" (i.e. has a much smaller variance) than the prior. There is no room in Bayesian analysis for the Classical notions of "sampling uncertainty", and less a-priori focus on the "asymptotic behavior" of estimators.[2]

The term in the denominator of (1) is called "marginal likelihood", is not a function of $\boldsymbol{\theta}$, and can usually be ignored for most components of Bayesian analysis. Thus, we usually work only with the nominator (i.e. prior times likelihood) for inference about $\boldsymbol{\theta}$. From (1) we know that this expression is proportional ("$\propto$") to the actual posterior. However, the marginal likelihood is crucial for model comparison, so we'll learn a few methods to derive it as a by-product of or following the actual posterior analysis. For some choices of prior and likelihood there exist analytical solutions for this term.

In summary, Frequentists start with a "blank mind" regarding $\boldsymbol{\theta}$. They collect data to produce an estimate $\hat{\boldsymbol{\theta}}$. They formalize the characteristics and uncertainty of $\hat{\boldsymbol{\theta}}$ for a finite sample context (if possible) and a hypothetical large sample (asymptotic) case. Bayesians collect data to *update a prior*, i.e. a pre-conceived probabilistic notion regarding $\boldsymbol{\theta}$.

Practical Implications of Choosing a Classical or Bayesian Estimation Framework

If the sample size is large and the likelihood function "well-behaved" (which usually means a simple function with a clear maximum, plus a small dimension for $\boldsymbol{\theta}$), Classical and Bayesian analysis are essentially on the same footing and will produce virtually identical results. This is because the likelihood function and empirical data will dominate any prior assumptions in the Bayesian approach.

If the sample size is large but the dimensionality of $\boldsymbol{\theta}$ is high and the likelihood function is less tractable (which usually means highly nonlinear, with local maxima, flat spots, etc.), a Bayesian approach may be preferable purely from a computational standpoint. It can be very difficult to get

[2]However, at times Bayesian analysis does rest on asymptotic results - see Koop et al. (2007), Ch. 9.. Naturally, the general notion that a larger sample, i.e. more empirical information, is better than a small one also holds for Bayesian analysis.

good and reliable estimates via Maximum Likelihood (MLE) techniques, but it is usually straightforward to derive a posterior distribution for the parameters of interest using Bayesian estimation approaches, which usually operate via sequential draws from known distributions.

If the sample size is small, Bayesian analysis can have substantial advantages over a Classical approach. First, Bayesian results do not depend on asymptotic theory to hold for their interpretability. Second, the Bayesian approach combines the sparse data with subjective priors. If these priors are well informed, this can truly add value (i.e. gain in accuracy and efficiency) to the analysis. Conversely, of course, poorly chosen priors[3] can produce misleading posterior inference in this case. Thus, under small sample conditions, the choice between Bayesian and Classical estimation often boils down to a choice between trusting the asymptotic properties of estimators and trusting one's priors.

Model Comparison

The Bayesian setting also offers a very flexible framework for the comparison of competing models. The models don't have to be "nested" in the Classical sense. All that's required is that the competing specifications share the same dependent variable, i.e. $\mathbf{y}$. In contrast, model comparison can be quite tricky in the Classical setting when competing specifications are not nested.

Assume you're considering two models, say M_1 and M_2, each associated with a respective set of parameters, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. We want to know which model is more "probable", given the observed data. We start by re-writing (1) with explicit inclusion of model indexes (we'll drop $\mathbf{X}$ since the exact composition of explanatory data is implicitly covered by model index M_i):

$$p\left(\boldsymbol{\theta}_i|\mathbf{y}, M_i\right) = \frac{p\left(\boldsymbol{\theta}_i|M_i\right) p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right)}{p\left(\mathbf{y}|M_i\right)} \quad i = 1, 2 \tag{2}$$

This expression shows that differences across models can occur due to differing priors for $\boldsymbol{\theta}$ and / or differences in the likelihood function. The marginal likelihood in the denominator will usually also differ across models. We now re-apply Bayes' Rule to derive an expression for the *posterior model probability*

$$p\left(M_i|\mathbf{y}\right) = \frac{p\left(M_i\right) p\left(\mathbf{y}|M_i\right)}{p\left(\mathbf{y}\right)} \quad i = 1, 2 \tag{3}$$

where then numerator is the product of *prior model probability* (often set to equal values across models in absence of strong priors) and the marginal likelihood from (2). We can now construct the *posterior odds ratio* for the two models as

$$\frac{p\left(M_1|\mathbf{y}\right)}{p\left(M_2|\mathbf{y}\right)} = \frac{p\left(M_1\right) p\left(\mathbf{y}|M_1\right)}{p\left(M_2\right) p\left(\mathbf{y}|M_2\right)} \tag{4}$$

Under equal model priors (i.e. $p\left(M_1\right) = p\left(M_2\right)$) this reduces to the *Bayes Factor* for model 1 vs. 2, i.e.

$$BF_{1,2} = \frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_2\right)} \tag{5}$$

[3]For example, priors that place substantial probability mass on practically infeasible ranges of θ - this often happens inadvertently when parameter transformations are involved in the analysis.

which is simply the ratio of marginal likelihoods for the two models. Since Bayes Factors can become quite large, we usually prefer to work with its logged version

$$logBF_{1,2} = \log p(\mathbf{y}|M_1) - \log p(\mathbf{y}|M_2) \tag{6}$$

The derivation of BF's and thus model comparison is straightforward if expressions for marginal likelihoods are analytically known or can be easily derived. However, often this can be quite tricky, and we'll learn a few techniques to compute marginal likelihoods in this course.

On a final note, marginal likelihoods can also be used to derive *model weights* in *Bayesian Model Averaging (BMA)*. Simply put, the intuition behind BMA is that we're never fully convinced that a single model is the correct one for our analysis at hand. There are usually several (and often millions of) competing specifications. To explicitly incorporate this notion of "model uncertainty", one can estimate every model separately, compute relative probability weights for each model, and then generate model-averaged posterior distributions for the parameters (and predictions) of interest. This would be awkward to accomplish in a Classical framework, and thus constitutes another advantage of employing a Bayesian estimation approach.[4].

References

Hansen, B. (2007). Least Squares model averaging, *Econometrica* **75**: 1175–1189.

Hjort, N. and Claeskens, G. (2003). Frequentist model average estimators, *Journal of the American Statistical Association* **98**: 879–899.

Hoff, P. (2009). *A first course in Bayesian statistical methods*, Springer.

Koop, G. (2003). *Bayesian Econometrics*, Wiley.

Koop, G., Poirier, D. and Tobias, J. (2007). *Bayesian Econometric Methods*, Cambridge University Press.

[4]That said, there is now an emerging literature on "Frequentist Model Averaging". See, for example, Hjort and Claeskens (2003) and Hansen (2007).