Econ. 517, Econometric Theory I Fall 2000 Chris Sims

Bayesian Basics *

Reading that covers most of the ideas we will discuss is in Chapters 1 and 2 of *Bayesian Data Analysis* by Andre Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin, Chapman and Hall (London: 1995). This book should be on reserve and available at the PU Store under the Economics 513 heading.

1. Inference as application of Bayes' Rule

(1) 2×2 genetic testing example
- Likelihood
- Type I and Type II errors — equivalence to likelihood
- Dependence of decision making on prior
- Confidence regions, pre-sample and post-sample probability statements

	Defect present	Defect not present
Test +	99	1
Test -	1	99

(2) 3×3 specimen box example
- Likelihood
- Type I and Type II errors — *not* equivalent to likelihood
- Confidence regions, pre-sample and post-sample probability statements.
  - Confidence regions as collections of hypothesis tests
  - Confidence regions don't correspond to sharpness of information about location
  - Confidence regions are not unique.
  - Confidence statements depend on probabilities of things that didn't happen.

	Frog	Salamander	Earthworm
Red	40	5	55
Green	55	89	40
Blue	5	6	5

	Frog	Salamander	Earthworm
Red	40	5	45
Light Green	34	2	6
Medium Green	15	3	38
Dark Green	6	85	6
Blue	5	6	5

Bayes' Rule:

$$p(x \mid y) = \frac{p(x,y)}{\int p(x,y)dx} \tag{1}$$

or

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{\int p(y \mid x)p(x)dx} \tag{2}$$

(3) Scientific Reporting: Summarizing the Likelihood

2. Bayesian Asymptotics

(1) Posterior means converge with probability one if the prior variance of the parameter is finite,
- to the true value, if any consistent estimator exists,
- to something whose prior expectation is the true value, otherwise.
- This is the Martingale Convergence Theorem.
- Priors, if smooth, stop mattering for the shape of the posterior in large samples.

(2) Likelihood shape asymptotics

Type I: Log likelihood 2nd-order Taylor expansion in the neighborhood of $\hat{\theta}_{MLE}$ is asymptotically accurate; i.e. assuming

$$\theta \sim N\left(\hat{\theta}_{MLE}, -\left(\frac{\partial^2 \log p(Y_T \mid \hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right)$$

is a good approximation in large samples. Why? If $\log\big(p(y_t \mid \theta)\big)$ has three derivatives, in i.i.d. case

$$\log\big(p(Y_T \mid \theta\big) =$$

$$\log\big(p(Y_T \mid \hat{\theta}\big) + 0 + \frac{1}{2}T^{.5}(\theta-\hat{\theta})'\left(T^{-1}\frac{\partial^2 \log p(Y_T \mid \hat{\theta})}{\partial\theta\partial\theta'}\right)T^{.5}(\theta-\hat{\theta}) + T^{-.5}T^{-1}\frac{\partial^3 \log p}{\partial\theta^3}\,0\left(T^{3/2}\left\|\theta-\hat{\theta}\right\|^3\right).$$

Third order term dwindles in importance in any $0(T^{-.5})$ neighborhood of $\hat{\theta}_{MLE}$. Proving that this is the only relevant neighborhood is some work.

Type II: If a vector of statistics S_T is asymptotically $N(\mu, \Sigma)$ according to a classical CLT, then under regularity conditions $\mu \mid S_T$ converges in distribution to $N(S_T, \Sigma)$. So approximate Bayesian posteriors can be constructed from the classical asymptotics. Distributional assumptions are as weak as for the underlying classical results.

Comparison: These two types of asymptotic results are quite different. The former gives us guidance on how to approximate the shape of a likelihood function that we can evaluate at any point. It assumes we have a true likelihood, meaning we know the true model, up to a finite list of unknown parameters. The latter result is, in contrast, a robustness result. It states that we can make approximate probability statements conditional on a given list of statistics, even though we may not have a true parametric model for the data. The former, approximation-type result can be checked in a given sample just by exploring the actual shape of the likelihood to see if it matches the approximation. The latter, robustness-type result cannot be checked except by comparing it to results with the true likelihood — which requires an exact model and possibly more difficult computations, which we were trying to avoid by invoking the approximation.

3. Gaussian mechanics

- Normal linear regression with known variance:

$$\underset{T\times 1}{y} \sim N\left(\underset{k\times 1}{X\,\beta}, \sigma^2 I\right)$$

$$\Rightarrow p(y \mid X) = (2\pi)^{T/2}\sigma^{-T}\exp\left(-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}\right)$$

$$= (2\pi)^{T/2}\sigma^{-T}\exp\left(-\frac{\hat{u}'\hat{u}}{2\sigma^2} - \frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2\sigma^2}\right)$$

$$= (2\pi)^{(T-k)/2}\sigma^{-(T-k)}\exp\left(-\frac{\hat{u}'\hat{u}}{2\sigma^2}\right)|X'X|^{-1}\,\varphi\big(\beta-\hat{\beta}, \sigma^2(X'X)^{-1}\big)$$

  where $\hat{\beta}$ is the OLS estimator of β and $\hat{u} = y - X\hat{\beta}$.
- Normal prior, Normal likelihood, $\rightarrow$ Normal posterior.
  - posterior mean is a weighted average of prior mean and MLE.

$$p(\beta, y) = \varphi(\beta-\bar{\beta};\Sigma)\cdot\varphi(y-\beta;\Omega)$$
$$= \varphi\left(\Sigma(\Sigma+\Omega)^{-1}(y-\bar{\beta})+\bar{\beta};\Sigma(I-\Sigma^{-1}\Omega)^{-1}\right)$$
$$= \varphi\left((\Omega^{-1}+\Sigma^{-1})^{-1}\left(\Omega^{-1}y+\Sigma^{-1}\bar{\beta}\right);\left(\Sigma^{-1}+\Omega^{-1}\right)^{-1}\right)$$

  - When the model has a Gaussian likelihood, a Gaussian prior is "as if" we had additional observations ("dummy observations") on the original model.
  - In multivariate models, "weighted averages" need not stay "between" prior mean and posterior.
  - Geometric interpretation: Ellipsoids and contract curves. Experimenting with priors as a device for describing likelihood.

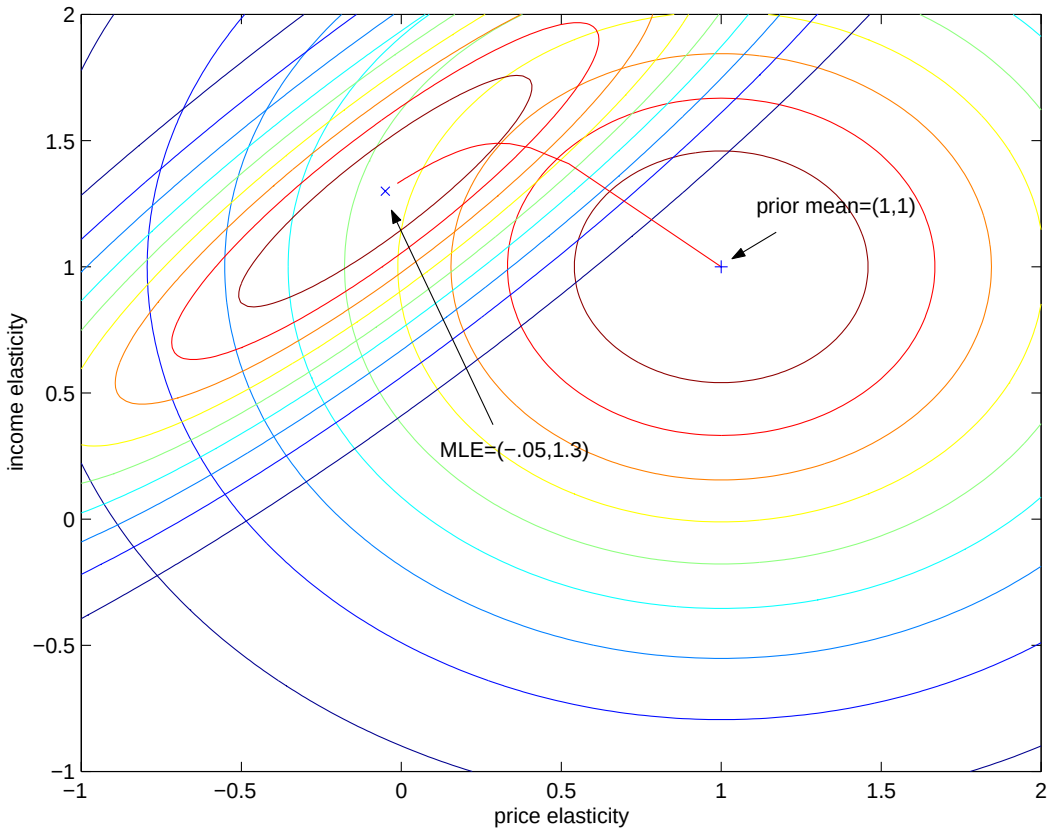

FIGURE 1. Posterior mean lies on "contract curve" between MLE and prior mean

4. Decision theory in a nutshell

We have a **loss function** $L(\delta(y), \beta)$ that depends on our **decision** δ and the "unknown parameter" β. We observe y, and we have a **model** $p(y \,|\, \beta)$ that characterizes the distribution of the observation y as a function of the parameter β. We choose δ, a function that maps observations y into decisions $\delta(y)$, from a set Δ of feasible decision rules, trying to keep L small.

The **risk function** is defined as

$$R(\delta, \beta) = E[L(\delta(y), \beta) \,|\, \beta] \,.$$

A **Bayesian decision rule** is a $\delta_0 \in \Delta$ such that for some pdf $\pi(\beta)$ over β,

$$E_\pi[R(\delta_0, \beta)] = \int R(\delta_0, \beta)\pi(\beta)\, d\beta = \min_{\delta \in \Delta} E_\pi[R(\delta, \beta)] \,.$$

An **inadmissible** decision rule is a $\delta_0 \in \Delta$ such that there is another $\delta^* \in \Delta$ with $R(\delta^*, \beta) \le R(\delta_0, \beta)$ for all β, and the inequality is strict for some β. An **admissible** δ is one that is not inadmissible.

Results:

(1) Under mild regularity conditions, Bayesian decision rules are admissible.
(2) (Complete Class Theorem) Under somewhat restrictive regularity conditions, all admissible decision rules are Bayesian or limits of Bayesian decision rules.
(3) OLS estimation of β in the normal linear regression model $y = X\beta + \varepsilon$ is not admissible if the β vector has 3 or more elements and the loss function is $\|\beta - \delta\|^2$, where $\|\cdot\|^2$, applied to a vector, returns its sum of squared elements. That the OLS estimator is dominated was shown by James and Stein.

Decision theory is discussed at much greater length, and more carefully, in chapter 3 of *Theory of Statistics* by Mark J. Schervish (Springer,1995), which should be on reserve and available at the PU store under Eco513. Schervish takes up a complete class theorem and the James-Stein estimator, among other topics.