

## Economics 696: Lecture Note 10

First, we will begin with a general discussion of the relationship between Bayesian estimation and inference to other commonly used methods. Then we will return to the MNP model, and discuss some non-Bayesian estimation methods that have been proposed.

# 1 Bayesian and Other Estimation Methods

## 1.1 Point Estimation

Recall that a point estimator of a parameter  $\theta \in \Theta$  is a function  $\hat{\theta}(z)$  mapping data into  $\Theta$ . A Bayes point estimator solves:

$$\hat{\theta}_B(z) = \arg \min_{a \in \Theta} \int L(\theta, a) p(\theta|z) d\theta = \arg \min_{a \in \Theta} E[L(\theta, a)|z].$$

where  $L(\theta, a)$  is a loss function. For example, when  $L(\theta, a) = (\theta - a)^2$ , the Bayes estimate is equal to the posterior expected mean.

Other commonly used estimators include the maximum likelihood estimator

$$\hat{\theta}_{ML}(z) = \max_{a \in \Theta} l(a),$$

where  $l(a) = \log f(z|a)$  is the log likelihood evaluated at  $\theta = a$ . Method of moments estimators are also commonly used.

In all of these cases, we have a function of the data that gives a single “guess” about the parameter. So we can compare different estimators from various perspectives.

First, if we take a subjectivist perspective, viewing the data analyst as a decision-maker satisfying the Savage axioms, then we know that they should act as if they have placed a prior on the parameter space and minimized posterior expected loss.

Next, we take a ex ante, repeated sampling perspective. Since the estimator is a function of the data  $z$ , and since the data have a distribution  $p(z|\theta)$ , we can view the point estimator as a random variable and look into its properties.

1. Unbiasedness: Recall that an estimator is unbiased if

$$E_{\theta}[\hat{\theta}(z)] = \theta.$$

The expectation on the left is with respect to the distribution of  $z$  under  $\theta$ :

$$E_{\theta}[\hat{\theta}(z)] = \int \hat{\theta}(z) f(z|\theta) dz.$$

Unbiasedness is sometimes considered a desirable property, but in many settings it is difficult to come up with an exactly unbiased estimator. In many models, both the Bayes estimator and the ML estimator are biased.

2. Admissibility: recall that an estimator is admissible if its risk function is not dominated by the risk function of another estimator. For a given estimator, the risk function is:

$$R(\theta, \hat{\theta}) = E_{\theta}[L(\theta, \hat{\theta}(z))] = \int L(\theta, \hat{\theta}(z)) f(z|\theta) dz.$$

Notice that risk is a frequentist concept: it takes the estimator as a random variable and looks at its behavior in repeated samples.

An estimator  $\hat{\theta}$  is dominated by another estimator  $\tilde{\theta}$  if

$$R(\theta, \hat{\theta}) \geq R(\theta, \tilde{\theta}) \quad \forall \theta \in \Theta$$

and

$$R(\theta, \hat{\theta}) > R(\theta, \tilde{\theta}) \quad \text{for some } \theta.$$

We have shown earlier that Bayes decision rules are admissible, so for point estimation problems, Bayes estimators are admissible. (Actually, our proof was only for the case where the parameter space  $\Theta$  was finite, and the prior was proper and strictly positive on the parameter space. It can happen that in other cases, Bayes estimators are not quite admissible, although they are usually close to being admissible.)

Other estimators such as ML and method of moments estimators are not guaranteed to be admissible. However, we will discuss below that ML often gives results similar to a Bayes estimate.

Third, we can take a frequentist perspective, but also use asymptotics to approximate the sampling distribution of the estimators. In “nice” cases, we know that method of moments and ML estimators are generally consistent:

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{as } n \rightarrow \infty.$$

Also, these estimators are typically asymptotically normally distributed:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, J),$$

for some variance matrix  $J$ . This implies that the estimator is approximately unbiased. The ML estimator is generally superior to the method of moments estimator, in that its variance  $J$  is smaller (in the matrix sense).

Bayes point estimators are generally asymptotically equivalent to the ML estimator (it is consistent and has the same asymptotic distribution). So the Bayes estimator has the same desirable properties as the ML estimator.

## 1.2 Inference

Classical inference (hypothesis testing and confidence intervals) is conceptually quite different from Bayesian inference, but it is easy to become confused about the distinction.

Suppose  $\theta$  is scalar and we want to test a hypothesis that  $\theta \leq h$  against the alternative that  $\theta > h$ . In a Bayesian analysis, we might simply report the posterior probability that  $\theta > h$ :

$$Pr(\theta > h|z) = \int_h^\infty p(\theta|z)d\theta.$$

This type of statement is now allowed in the Neyman-Pearson approach to testing, because  $\theta$  is never viewed as a random variable. Instead, we construct a test statistic  $T(z)$  and an associated critical region  $C$  such that

$$\sup_{\theta \leq h} Pr_\theta(T(z) \in C) \leq \alpha.$$

Intuitively, we want the probability of rejection to be high when  $\theta > h$ , and no greater than the significance level  $\alpha$  when  $\theta \leq h$ . Notice that the probabilities we are working with are

$$Pr_\theta(T(z) \in C) = \int_{z:T(z) \in C} f(z|\theta)dz.$$

For classical hypothesis testing, the sampling distribution of  $T(z)$  under a fixed  $\theta$  is key, not the posterior distribution for  $\theta$  given  $z$ . These two distributions are different.

**Example:** suppose that  $z = (z_1, \dots, z_n)$  where the  $z_i$  are IID Bernoulli with

$$f(z_i|\theta) = Pr(z_i = 1|\theta) = \theta.$$

If we want to test that  $\theta > 1/2$  in the Neyman-Pearson approach, we might base the test on the statistic

$$T(z) = \frac{1}{n} \sum_{i=1}^n z_i,$$

which is also the maximum likelihood estimator. Notice that for a given  $n$  and  $\theta$ , the distribution of  $z$  is discrete, and therefore the distribution of  $T(z)$  is also discrete. In particular,  $T(z)$  can only take on values in  $\{0, 1/n, 2/n, \dots, 1\}$ .

In a Bayesian approach, we might use a Beta prior for  $\theta$ . Then the posterior distribution (even for the limiting “noninformative” prior) will also be a Beta distribution, which is continuously supported on  $[0, 1]$ . So in this case, the sampling distribution of  $T(z)$  is discrete, whereas the posterior distribution of  $\theta$  given  $z$  is continuous.

□

A similar line of reasoning applies to confidence intervals. For classical confidence intervals, we seek statistics  $L(z), U(z)$  such that

$$\sup_{\theta \in \Theta} Pr_{\theta}(L(z) \leq \theta \leq U(z)) \geq 1 - \alpha.$$

Again, the probability in this expression is based on the probability distribution of  $L(z)$  and  $U(z)$  for a fixed  $\theta$ .

A Bayesian approach to interval estimation could look for numbers  $l, u$  such that

$$Pr(l \leq \theta \leq u|z) = \int_l^u p(\theta|z)d\theta = 1 - \alpha.$$

(There are different pairs  $l, u$  that would satisfy this equation. For example we could look for  $l, u$  such that each tail probability is  $\alpha/2$ , or we could look for the shortest interval satisfying the definition.)

The Bayesian interval estimator is based on the posterior  $p(\theta|z)$ , not on the sampling distribution of  $z$  given  $\theta$ . A Bayesian 95% interval need not have a 95% coverage property in the classical sense.

At this point, it should be clear that classical inference is different at a very fundamental level from Bayesian posterior inference. So we should not expect, in general, that the two approaches will lead to the same conclusions (for a test) or the same intervals (for an interval estimate).

However, there are some cases where the two approaches happen to give the same results.

**Example:** Suppose that  $z = (z_1, \dots, z_n)$ , and  $z_i \sim N(\theta, \sigma^2)$ , where  $\sigma^2$  is known. The estimator

$$\hat{\theta}(z) = \frac{1}{n} \sum_{i=1}^n z_i = \bar{z}_n$$

is the maximum likelihood estimator and also the Bayesian posterior mean estimator for a flat prior. The sampling distribution (for a given  $\theta$ ) of  $\hat{\theta}$  is

$$\hat{\theta}|\theta \sim N\left(\theta, \frac{\sigma^2}{n}\right).$$

Using this sampling distribution, we can construct a 95% confidence interval for  $\theta$  as  $[L(z), U(z)]$ , where

$$\begin{aligned} L(z) &= \hat{\theta} - 1.96\sqrt{\frac{\sigma^2}{n}} \\ U(z) &= \hat{\theta} + 1.96\sqrt{\frac{\sigma^2}{n}}, \end{aligned}$$

and it satisfies the coverage property:

$$Pr_{\theta}(L(z) \leq \theta \leq U(z)) = 0.95.$$

In a Bayesian analysis with a flat prior, the posterior distribution is

$$\theta|z \sim N\left(\hat{\theta}, \frac{\sigma^2}{n}\right).$$

Note that this mirrors the sampling distribution of  $\hat{\theta}$ . Then if we look for  $(l, u)$  such that

$$Pr(l \leq \theta \leq u|z) = \int_l^u p(\theta|z)d\theta = 0.95,$$

this will be satisfied by setting  $l = L(z)$  and  $u = U(z)$ . So the 95% Bayesian interval will coincide with the 95% confidence interval.

□

### 1.3 Bernstein-von Mises Theorem

Recall that for the ML estimator,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} N(0, I_{\theta}^{-1}),$$

where  $I_{\theta}$  is the Fisher information matrix. So, loosely speaking, the sampling distribution of  $\hat{\theta}_{ML}$  is approximately

$$\hat{\theta}|z \approx N(\theta, I_{\theta}^{-1}/n).$$

The Bernstein-von Mises Theorem states that, under roughly the same regularity conditions as those required for asymptotic normality of the ML estimator, the posterior distribution is approximately normal, centered at the MLE:

$$\theta|z \approx N(\hat{\theta}_{ML}, I_{\theta}^{-1}/n).$$

This is very similar to the IID normal example given in the previous subsection, and by the same reasoning, a 95% confidence interval will be (approximately) a 95% Bayesian interval and vice versa.

## 2 Simulation-Based Estimation of MNP Models

We’ve seen that in parametric models satisfying some regularity conditions, the Bayes estimator is asymptotically equivalent to the MLE, and that one can form approximate posterior distributions from the MLE and its variance matrix. In some models, it may be difficult to calculate the MLE, but a Bayesian analysis (perhaps using MCMC to simulate the posterior) may be feasible; in other models, it may be considerably easier to calculate the MLE and its standard errors.

Now, let’s return to the MNP model of the previous note, and consider other estimation methods. We generalize the model slightly from the previous note:

### 2.1 Multinomial Probit Model of Discrete Choice (McFadden, 1989)

Choices:  $c = 1, \dots, C$

Measured characteristics of choices:  $x_1, \dots, x_C$ , where each  $x_c$  is a  $k \times 1$  vector.

Examples of characteristics: prices, distance of choice to consumer.

Preferences are based on “weights”:  $\alpha \sim N_k(\beta, \Omega)$ .

Utilities:

$$\begin{aligned} u_1 &= x_1' \alpha \\ u_2 &= x_2' \alpha \\ &\vdots \\ u_C &= x_C' \alpha \end{aligned}$$

Choice rule: pick  $c$  if  $u_c \geq u_{c'} \quad \forall c' \in \{1, \dots, C\}$ .

(Assume no ties, and only one object chosen.)

Let

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_C \end{pmatrix}, \quad X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_C \end{pmatrix}.$$

Then  $u = X\alpha$ , so

$$u|X, \beta, \Omega \sim N_C(X\beta, X\Omega X').$$

Let  $d_c = 1$  if choice  $c$  is chosen. Then

$$\begin{aligned} E[d_c|X, \beta, \Omega] &= Pr[d_c = 1|X, \beta, \Omega] \\ &= Pr[u_c \geq u_1, \dots, u_c \geq u_C|X, \beta, \Omega] \\ &= \int \cdots \int \mathbf{1}(u_c \geq u_1, \dots, u_c \geq u_C) dN(u_1, \dots, u_C|X\beta, X\Omega X'), \end{aligned}$$

where  $dN(\cdot|\cdot, \cdot)$  means integration with respect to the density of the multivariate normal distribution.

## Sample Data

Individuals  $i = 1, \dots, n$ .

Weights:  $\alpha_i \stackrel{\text{iid}}{\sim} N(\beta, \Omega)$ .

Now assume that  $\theta$  denotes the distinct parameters of  $\beta, \Omega$ . For example, we might restrict  $\Omega$  to be diagonal. So we can write  $\beta = \beta(\theta)$  and  $\Omega = \Omega(\theta)$ , and

$$\alpha_i \stackrel{\text{iid}}{\sim} N(\beta(\theta), \Omega(\theta)).$$

Measured characteristics:  $x_{i1}, \dots, x_{iC}$ . (For example, individuals might face different prices for the different choices.)

$$X_i = \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{iC} \end{pmatrix}.$$

Choice indicators:

$$d_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{iC} \end{pmatrix}.$$

Likelihood for individual  $i$ :

$$\prod_{c=1}^C Pr(d_{ic} = 1|X_i, \theta)^{d_{ic}}.$$

Full-sample likelihood:

$$\prod_{i=1}^n \prod_{c=1}^C Pr(d_{ic} = 1|X_i, \theta)^{d_{ic}}.$$

Log likelihood:

$$L(\theta) = \sum_{i=1}^n \sum_{c=1}^C d_{ic} \log Pr(d_{ic} = 1|X_i, \theta).$$

ML Estimator:

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta).$$

First order conditions:

$$0 = \frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n \sum_{c=1}^C d_{ic} \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1|X_i, \theta).$$

Note that

$$\sum_{c=1}^C Pr(d_{ic} = 1|X_i, \theta) = 1.$$

So

$$\sum_{c=1}^C \frac{\partial}{\partial \theta} Pr(d_{ic} = 1|X_i, \theta) = 0.$$

Also,

$$\left[ \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1|X_i, \theta) \right] \cdot Pr(d_{ic} = 1|X_i, \theta) = \frac{\partial}{\partial \theta} Pr(d_{ic} = 1|X_i, \theta),$$

So

$$0 = \sum_{c=1}^C \frac{\partial}{\partial \theta} Pr(d_{ic} = 1|X_i, \theta) \cdot Pr(d_{ic} = 1|X_i, \theta),$$

and we can write the FOC alternatively as

$$0 = \sum_{i=1}^n \sum_{c=1}^C \left[ \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1|X_i, \theta) \right] \{d_{ic} - Pr(d_{ic} = 1|X_i, \theta)\}.$$

This is somewhat intuitive: the second term  $\{d_{ic} - Pr(d_{ic} = 1|X_i, \theta)\}$  is the outcome minus its conditional mean given  $X_i$ .

To solve for the MLE, we need to be able to calculate  $Pr(d_{ic} = 1|X_i, \theta)$  (and possibly its derivatives) for each  $i, c$  and different possible parameter values  $\theta$ .

However, the integral defining  $Pr(d_{ic} = 1|X_i, \theta)$  does not have a simple closed-form expression.

For relatively few choices ( $C \leq 4$ ), there exist deterministic (nonrandom) numerical integration routines that are efficient, but these cannot be used when the number of choices is large.

Lerman and Manski (1981) suggest a Monte Carlo procedure:

For  $s = 1, \dots, S$ :

- Draw  $u^s = (u_1^s, \dots, u_C^s)' \sim N(X_i\beta(\theta), X_i\Omega(\theta)X_i')$ .
- Form  $d^s = (d_1^s, \dots, d_C^s)'$ .

Then approximate  $Pr(d_{ic} = 1|X_i, \theta)$  by

$$\tilde{P}_S(d_{ic} = 1|X_i, \theta) = \frac{1}{S} \sum_{s=1}^S d_c^s.$$

By the LLN,

$$\tilde{P}_S(d_{ic} = 1|X_i, \theta) \xrightarrow{p} Pr(d_{ic} = 1|X_i, \theta) \quad \text{as } S \rightarrow \infty.$$

Simulation-based estimator:

$$\hat{\theta}_{LM} = \arg \max_{\theta} \sum_{i=1}^n \sum_{c=1}^C d_{ic} \cdot \log \tilde{P}_S(d_{ic} = 1|X_i, \theta).$$

For this to work well,  $S$  needs to be very large so that the simulation error  $\tilde{P}_S(d_{ic} = 1|X_i, \theta) - Pr(d_{ic} = 1|X_i, \theta)$  does not affect the inference. Unfortunately, this makes the procedure too computationally burdensome.

## 2.2 Method of Moments and Simulation

For the moment, suppose that  $Pr(d_{ic} = 1|X_i, \theta)$  is not hard to calculate, and consider an alternative estimation approach.

Recall that  $Pr(d_{ic} = 1|X_i, \theta)$  also had the interpretation as the conditional mean of  $d_{ic}$  given  $X_i$ . So

$$E[d_{ic} - Pr(d_{ic} = 1|X_i, \theta)|X_i] = 0.$$

This suggests moment conditions

$$E \begin{bmatrix} w_{i1}\{d_{i1} - Pr(d_{i1} = 1|X_i, \theta)\} \\ \vdots \\ w_{iC}\{d_{iC} - Pr(d_{iC} = 1|X_i, \theta)\} \end{bmatrix} = 0,$$

where the  $w_{ic}$  are vector-valued functions of  $X_i$ . (They could also be additional instrumental variables that are mean-independent of  $d_{ic} - Pr(d_{ic} = 1|X_i, \theta)$ .)

So we have a GMM moment function

$$g(w_i, X_i, d_i, \theta) = \begin{pmatrix} w_{i1}\{d_{i1} - Pr(d_{i1} = 1|X_i, \theta)\} \\ \vdots \\ w_{iC}\{d_{iC} - Pr(d_{iC} = 1|X_i, \theta)\} \end{pmatrix}.$$

Notice the similarity between this choice of moment function and the MLE score equation. Basically, the MLE score involves replacing  $w_{ic}$  by  $\frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1|X_i, \theta)$ , which also depends on  $\theta$ . This suggests that by choosing  $w_{ic}$  cleverly, we could achieve the same efficiency as MLE.

Assume that the instruments  $w$  are such that the dimension of  $g$  is greater than or equal to the dimension of  $\theta$ . Then we could estimate  $\theta$  by solving

$$\min_{\theta} \left[ \sum_{i=1}^n g(w_i, X_i, d_i, \theta) \right]' \left[ \sum_{i=1}^n g(w_i, X_i, d_i, \theta) \right].$$

Here, we are implicitly using an identity matrix as the GMM weighting matrix. Alternatively, we could have a different weighting matrix  $A$ , but then we could modify the instruments  $w$  to account for this weighting.

By stacking the observations, we can rewrite this estimator as

$$\hat{\theta}_{MM} = \arg \min_{\theta} [d - P(\theta)]' W' W [d - P(\theta)],$$

where

$$d = \begin{pmatrix} d_{11} \\ \vdots \\ d_{1C} \\ \vdots \\ d_{n1} \\ \vdots \\ d_{nC} \end{pmatrix}, \quad P(\theta) = \begin{pmatrix} Pr(d_{11} = 1|X_1, \theta) \\ \vdots \\ Pr(d_{1C} = 1|X_1, \theta) \\ \vdots \\ Pr(d_{n1} = 1|X_n, \theta) \\ \vdots \\ Pr(d_{nC} = 1|X_n, \theta) \end{pmatrix}.$$

Now, consider simulation of the choice probabilities. For a finite, possibly small value of  $S$ , the simulation error will not be negligible. However, the simulations are unbiased in the sense that

$$E[\tilde{P}_S(d_{ic} = 1|X_i, \theta) - Pr(d_{ic} = 1|X_i, \theta)|X_i] = 0.$$

Also  $\tilde{P}_S(d_{ic} = 1|X_i, \theta)$  is independent of the actual outcomes  $d_{ic}$  conditional on  $X_i$ . So it is true that

$$E[d_{ic} - \tilde{P}_S(d_{ic} = 1|X_i, \theta)|X_i] = 0.$$

The variance of  $d_{ic} - \tilde{P}_S(d_{ic} = 1|X_i, \theta)$  will differ from the variance of  $d_{ic} - Pr(d_{ic} = 1|X_i, \theta)$ , due to the simulation error, but the conditional means are both 0.

So, letting  $\tilde{P}(\theta)$  denote the stacked set of simulated response probabilities, we could use the simulated method of moments estimator

$$\hat{\theta}_{SMM} = \arg \min_{\theta} [d - \tilde{P}(\theta)]' W' W [d - \tilde{P}(\theta)],$$

This estimator does not require  $S \rightarrow \infty$ ; instead the simulation error is averaged out across individual observations. It will have a different asymptotic variance than the MM estimator, since the moment equation has a different variance.

## Implementation Notes

1. Notice that we need to construct the simulated probability vector  $\tilde{P}(\theta)$  for all  $\theta \in \Theta$  (or at least enough values so that we can maximize the criterion function). Exactly how we do this turns out to matter.

In particular, suppose we generate new random draws for each different value of  $\theta$ . Then it turns out that the SMM estimator is no longer consistent and asymptotically normal.

Instead, we have to use the same random draws to construct  $\tilde{P}(\theta)$  for different  $\theta$ . Here is one way to do this:

For  $s = 1, \dots, S$ , we want to draw  $u_i^s(\theta) \sim N(X_i\beta(\theta), X_i\Omega(\theta)X_i')$ . This can be done by drawing

$$\eta_{i1}^s, \dots, \eta_{iC}^s \stackrel{\text{iid}}{\sim} N(0, 1),$$

and forming

$$u_i^s(\theta) = X_i\beta(\theta) + \Gamma_i(\theta)'(\eta_{i1}^s, \dots, \eta_{iC}^s)',$$

where  $\Gamma_i(\theta)$  is the Cholesky factor of  $X_i\Omega(\theta)X_i'$ , an upper triangular matrix such that  $\Gamma_i(\theta)'\Gamma_i(\theta) = X_i\Omega(\theta)X_i'$ .

Then we form  $d_{ic}^s(\theta)$  by seeing which choice gives highest utility among the elements of  $u_i^s$ , and construct

$$\tilde{P}_S(d_{ic} = 1|X_i, \theta) = \frac{1}{S} \sum_{s=1}^S d_{ic}^s(\theta).$$

The key here is that  $\eta_{i1}^s, \dots, \eta_{iC}^s$  do not change with  $\theta$ . We draw them once, and only change  $\beta(\theta)$  and  $\Gamma_i(\theta)$  when calculating  $u_i^s(\theta)$ .

2. Another issue is that  $\tilde{P}(\theta)$  is not continuous in  $\theta$ . Take a typical element

$$\tilde{P}_S(d_{ic} = 1|X_i, \theta) = \frac{1}{S} \sum_{s=1}^S d_{ic}^s(\theta).$$

For fixed  $S$ , this will be a number in the set  $\{0, 1/S, 2/S, \dots, S/S\}$ . As  $\theta$  is varied,  $\tilde{P}_S(d_{ic} = 1|X_i, \theta)$  will jump between fractions of  $S$ .

As a consequence,  $\tilde{P}(\theta)$  is discontinuous in  $\theta$ , and so is the criterion function  $[d - \tilde{P}(\theta)]'W'W[d - \tilde{P}(\theta)]$ .

This has the practical problem that simple gradient-based numerical optimization routines will not work for solving the GMM minimization problem. We need to use some more sophisticated optimization routines, such as simulated annealing or simplex methods. (Amoeba is a simplex method.)

The discontinuity of the criterion function also means that our standard proofs of consistency and asymptotic normality of method of moments estimators do not work. More powerful asymptotic techniques are needed – see McFadden (1989) and Pakes and Pollard (1989).

3. Since the discontinuity of  $\tilde{P}(\theta)$  can make solving the estimator problematic, one alternative is to modify the simulated probabilities to make them smooth in  $\theta$ . McFadden discusses some different ways to do this. Some versions lead to simulated probabilities that are slightly biased. In this case, we need the bias to be small asymptotically.