- Please complete course evaluations online:

https://academicaffairs.ucsd.edu/Modules/Evals?e1120525

VII. Model selection

A. Marginal likelihood

Suppose we're trying to choose among a series of models:

Model 1: $p(\mathbf{y}|\boldsymbol{\theta}_1)$

$\vdots$

Model M: $p(\mathbf{y}|\boldsymbol{\theta}_M)$

where $\boldsymbol{\theta}_m$ are possibly of different dimension

The Bayesian might think in terms of an unobserved random variable:

$$s = 1 \text{ if Model 1 is true}$$

$$\vdots$$

$$s = M \text{ if Model } M \text{ is true}$$

and assign prior probabilities

$$\pi_1 = \Pr(s = 1)$$

$$\vdots$$

$$\pi_M = \Pr(s = M)$$

with associated priors on the parameters

$$p(\boldsymbol{\theta}_1 | s = 1)$$

$$\vdots$$

$$p(\boldsymbol{\theta}_M | s = M)$$

From such a perspective, the probability that Model m is true given the data is

$$p(s = m|\mathbf{y}) = \frac{\pi_m \int p(\mathbf{y}|\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|s = m)d\boldsymbol{\theta}_m}{\sum_{j=1}^{M} \pi_j \int p(\mathbf{y}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|s = j)d\boldsymbol{\theta}_j}$$

$$\equiv \frac{\pi_m p_m(\mathbf{y})}{\sum_{j=1}^{M} \pi_j p_j(\mathbf{y})}$$

The expression

$$p_m(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|s = m)d\boldsymbol{\theta}_m$$

is sometimes called the "marginal likelihood" of Model m

The Bayesian would say that the data favor the model for which $p(s = m|\mathbf{y})$ is biggest. With uniform priors $(\pi_m = 1/M)$, this is equivalent to choosing the model with the highest marginal likelihood.

VII. Model selection

A. Marginal likelihood

B. Schwarz criterion

First let's examine the behavior of

$$p_m(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|s = m)d\boldsymbol{\theta}_m$$

as the sample size T gets large

Suppose

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \log p(\mathbf{y}_t|\boldsymbol{\theta})$$

and let $\hat{\boldsymbol{\theta}}_T$ denote the MLE

$$\hat{\boldsymbol{\theta}}_T = \arg\max \log p(\mathbf{y}|\boldsymbol{\theta})$$

Recall Taylor's Theorem:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - \tfrac{1}{2}\sqrt{T}\,(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)' \times$$

$$T^{-1}\sum\nolimits_{t=1}^{T} \mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T)\,\sqrt{T}\,(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)$$

$$\tilde{\boldsymbol{\theta}}_T = \lambda_T \boldsymbol{\theta} + (1 - \lambda_T)\hat{\boldsymbol{\theta}}_T$$

$$\mathbf{H}_t(\boldsymbol{\theta}) \equiv -\frac{\partial^2 \log p(\mathbf{y}_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}$$

If λ_{2T} is smallest eigenvalue of $T^{-1}\sum_{t=1}^{T}\mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T)$ then

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - \tfrac{1}{2}\sqrt{T}\,(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)' \times$$
$$T^{-1}\sum_{t=1}^{T}\mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T)\,\sqrt{T}\,(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)$$
$$\leq \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - \tfrac{T\lambda_{2T}}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)$$

If also there exists B such that

$B \geq p(\boldsymbol{\theta})$, we should have

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$\leq \int \exp\left\{\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - \frac{T\lambda_{2T}}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)\right\}Bd\boldsymbol{\theta}$$

$$= Bp(\mathbf{y}|\hat{\boldsymbol{\theta}}_T)\int \exp\left\{-\frac{T\lambda_{2T}}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)\right\}d\boldsymbol{\theta}$$

But

$$\int \exp\left\{-\frac{T\lambda_{2T}}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)'(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)\right\}d\boldsymbol{\theta}$$

$$= \left[\frac{2\pi}{T\lambda_{2T}}\right]^{k/2} \times$$

$$\int \left[\frac{T\lambda_{2T}}{2\pi}\right]^{k/2} \exp\left\{-\frac{T\lambda_{2T}}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)'(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)\right\}d\boldsymbol{\theta}$$

$$= \left[\frac{2\pi}{T\lambda_{2T}}\right]^{k/2}$$

for k the dimension of $\boldsymbol{\theta}$

Conclusion:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$\leq Bp(\mathbf{y}|\hat{\boldsymbol{\theta}}_T)\left[\frac{2\pi}{T\lambda_{2T}}\right]^{k/2}$$

$$\log p(\mathbf{y}) \leq \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - (k/2)\log T + R_{2kT}$$

$$R_{2kT} = \log\left\{B\left[\frac{2\pi}{\lambda_{2T}}\right]^{k/2}\right\}$$

Also $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}^*$

$$T^{-1}\sum_{t=1}^{T}\mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T) = -T^{-1}\sum_{t=1}^{T}\frac{\partial^2 \log p(\mathbf{y}_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_T}$$

$$\xrightarrow{p} -E\left[\frac{\partial^2 \log p(\mathbf{y}_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right] = \boldsymbol{\Upsilon}^*$$

$$R_{2kT} = \log\left\{B\left[\frac{2\pi}{\lambda_{2T}}\right]^{k/2}\right\}$$

$$\xrightarrow{p} \log\left\{B\left[\frac{2\pi}{\lambda_2^*}\right]^{k/2}\right\}$$

for λ_2^* smallest eigenvalue of $\boldsymbol{\Upsilon}^*$

Similar argument reasons that

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - \tfrac{1}{2}\sqrt{T}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)' \times T^{-1}\sum_{t=1}^{T}\mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T)\sqrt{T}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)$$

$$\geq \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - \tfrac{T\lambda_{1T}}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)$$

for λ_{1T} biggest eigenvalue of

$$T^{-1}\sum_{t=1}^{T}\mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T)$$

$$\log p(\mathbf{y}) \geq \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - (k/2)\log T + R_{1kT}$$

$$\log p(\mathbf{y}) \leq \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - (k/2)\log T + R_{2kT}$$

Implication: for large T we have the approximation

$$\log p(\mathbf{y}) \simeq \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_T) - (k/2)\log T$$

for k the dimension of θ

Choosing the model m for which

$$\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{mT}) - (k_m/2)\log T$$

is biggest is known as the using the Schwarz Information Criterion (SIC) or Bayesian Information Criterion (BIC)

Since it is asymptotically a Bayesian decision rule, SIC inherits the properties of being asymptotically admissible and consistent

However, note that this result required the same regularity conditions needed to get asymptotic Normality of MLE

VII. Model selection

A. Marginal likelihood
B. Schwarz criterion
C. Calculating the marginal likelihood with the Gibbs sampler

Goal: calculate

$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$

Couldn't we get this by drawing

$\boldsymbol{\theta}^{(j)}\ j = 1, \ldots, J$ from $p(\boldsymbol{\theta})$ and then

$\hat{p}(\mathbf{y}) = J^{-1}\sum_{j=1}^{J} p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$?

Answer: no, this algorithm is

badly behaved numerically.

Chib's idea: think of evaluating at a point with a lot of mass (say the posterior mean $\boldsymbol{\theta}^*$). Note that for any $\boldsymbol{\theta}^*$ we have the identity

$$p(\boldsymbol{\theta}^*|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)$$

$$p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^*|\mathbf{y})$$

In many applications, we know $p(\mathbf{y}|\boldsymbol{\theta}^*)$ and $p(\boldsymbol{\theta}^*)$ analytically (evaluating the likelihood and prior at posterior mean, respectively), but couldn’t calculate $p(\boldsymbol{\theta}^*|\mathbf{y})$ explicitly

Suppose we've generated draws from a two-block Gibbs sampler:

$p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$

The object of interest is given by

$p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*|\mathbf{y}) = p(\boldsymbol{\theta}_1^*|\boldsymbol{\theta}_2^*, \mathbf{y})p(\boldsymbol{\theta}_2^*|\mathbf{y})$

where we may know $p(\boldsymbol{\theta}_1^*|\boldsymbol{\theta}_2^*, \mathbf{y})$ analytically.

We know that

$$p(\boldsymbol{\theta}_2^*|\mathbf{y}) = \int p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1, \mathbf{y}) p(\boldsymbol{\theta}_1|\mathbf{y}) d\boldsymbol{\theta}_1$$

and can therefore estimate

$$\hat{p}(\boldsymbol{\theta}_2^*|\mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^{(g)}, \mathbf{y})$$

that is, the average value of

$p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1, \mathbf{y})$ across Gibbs simulated

draws for $\boldsymbol{\theta}_1$

Conclusion: for a two-block Gibbs sampler, we can estimate the marginal likelihood from

$$\hat{p}(\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)p(\boldsymbol{\theta}_1^*)p(\boldsymbol{\theta}_2^*)}{p(\boldsymbol{\theta}_1^*|\boldsymbol{\theta}_2^*, \mathbf{y})G^{-1}\sum_{g=1}^{G} p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^{(g)}, \mathbf{y})}$$

How about 3 blocks? Now we want to estimate the denominator of

$$p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*|\mathbf{y})$$

$$p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*|\mathbf{y}) = p(\boldsymbol{\theta}_1^*|\mathbf{y})p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^*, \mathbf{y})p(\boldsymbol{\theta}_3^*|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{y})$$

$$p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^* | \mathbf{y})$$

$$= p(\boldsymbol{\theta}_1^* | \mathbf{y}) p(\boldsymbol{\theta}_2^* | \boldsymbol{\theta}_1^*, \mathbf{y}) p(\boldsymbol{\theta}_3^* | \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{y})$$

First term can be estimated as before:

$$\hat{p}(\boldsymbol{\theta}_1^* | \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(\boldsymbol{\theta}_1^* | \boldsymbol{\theta}_2^{(g)}, \boldsymbol{\theta}_3^{(g)}, \mathbf{y})$$

Third term $p(\boldsymbol{\theta}_3^* | \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{y})$ is known analytically

$$p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^* | \mathbf{y})$$

$$= p(\boldsymbol{\theta}_1^* | \mathbf{y}) p(\boldsymbol{\theta}_2^* | \boldsymbol{\theta}_1^*, \mathbf{y}) p(\boldsymbol{\theta}_3^* | \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{y})$$

Second term:

$$p(\boldsymbol{\theta}_2^* | \boldsymbol{\theta}_1^*, \mathbf{y}) = \int p(\boldsymbol{\theta}_2^* | \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3, \mathbf{y}) p(\boldsymbol{\theta}_3 | \boldsymbol{\theta}_1^*, \mathbf{y}) d\boldsymbol{\theta}_3$$

But how do we generate a sample from $p(\boldsymbol{\theta}_3 | \boldsymbol{\theta}_1^*, \mathbf{y})$?

Suppose we do a 2-block Gibbs sampler between $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ with $\boldsymbol{\theta}_1^*$ fixed throughout:

$p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3, \mathbf{y})$

$p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2, \mathbf{y})$

The ergodic distribution of $\boldsymbol{\theta}_3$ determined by this Markov chain ($q = 1, \ldots, Q$) is $p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1^*, \mathbf{y})$

$\hat{p}(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^*, \mathbf{y}) = Q^{-1} \sum_{q=1}^{Q} p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^{(q)}, \mathbf{y})$

So we estimate $p(\mathbf{y})$ from

$$\frac{p(\mathbf{y}|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*)p(\boldsymbol{\theta}_1^*)p(\boldsymbol{\theta}_2^*)p(\boldsymbol{\theta}_3^*)}{\hat{p}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*|\mathbf{y})}$$

$$\hat{p}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*|\mathbf{y}) =$$

$$G^{-1} \sum_{g=1}^{G} p(\boldsymbol{\theta}_1^*|\boldsymbol{\theta}_2^{(g)}, \boldsymbol{\theta}_3^{(g)}, \mathbf{y}) \times$$

$$Q^{-1} \sum_{q=1}^{Q} p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^{(q)}, \mathbf{y}) \times$$

$$p(\boldsymbol{\theta}_3^*|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{y})$$