# Bayesian Econometrics: Meet-up 5

Kevin Arthur Tittel

March 22, 2023

## 1 Recap Meet-up 4

In this Section we briefly review some main take-aways from Meet-up 4. We discussed an exercise in which the data $y$ followed an exponential distribution with parameter $1/\lambda$, $\lambda$ was attached a conjugate prior distribution in terms of an inverted Gamma distribution with parameters 2 and $\beta$, and subsequently $\beta$ itself was also attached uncertainty through a Gamma distribution with parameters $a$ and $b$.

First, the exercise allowed us to analytically derive the posterior distribution $\lambda|y$, which turned out to be an inverted Gamma distribution as well (i.e., the prior for $\lambda$ is conjugate). Second, the exercise challenged us to interpret the information added by the inverted Gamma prior in terms of adding data. Notably, we concluded that the shape parameter (i.e., $\tilde{\alpha} = N + 2$ in our case) informs us about the amount of data points added, whereas the scale parameter (i.e. $\tilde{\beta} = \beta + \sum_{i=1}^{N} y_i$ in our case) informs us about the size of the data. We can compute the amount of added data points by comparing $\tilde{\alpha}$ (when using a conjugate prior) to $\alpha$ (when using a diffuse prior). Since $\alpha = N-1$ under an uninformative prior, the amount of data points added is $\tilde{\alpha} - \alpha = (N + 2) - (N - 1) = 3$.

Third, the exercise allowed us to practice with Bayesian testing using the (logarithm of the) Bayes Factor. Importantly, since we observed that models A and B were nested in terms of parameter restrictions (i.e. $\lambda = 1 \subseteq \lambda \sim$ inverted Gamma(2,2)), we decided to compute an easier version of $BF_{A|B}$ by computing the Savage-Dickey Density Ratio. That is, we simply calculated values of the posterior density function $p_B(\lambda|y)$ and prior density function $p_B(\lambda)$ of model B only (note that the kernel does not suffice), while imputing

$\lambda = 1$ of the model A restriction. Last, we computed the distribution of $\beta|\lambda$, which turned out to follow a Gamma-distribution equal to $p(\beta)$.

# 2 Discussion Theory

In this Section we start off with a brief description of the theory given the slides of Lecture 3, which we will subdivide into a selection of key concepts. The overarching concept of the lecture is the fact that numerical methods need to be used in case it is not possible to obtain analytical results (e.g. posterior means/variances for logit and probit models). In order to solve the necessary integrals, one can choose from two standard approaches: i) numerical integration or ii) Monte Carlo integration.

**What is numerical integration?** When applying numerical integration, one writes the integral over the whole region as the sum of integrals over smaller regions, after which the latter integrals are approximated. An advantage of this method is its fast convergence with smaller approximation errors for low-dimensional integrals. In contrast, disadvantages of this method are i) decreasing convergence speed for higher dimensional integrals and ii) an exponentially increasing computational burden with the dimension of the integral.

**What is Monte Carlo integration?** When applying Monte Carlo integration, one approximates the integral of interest (e.g. $\mathbb{E}_{\theta|y}[h(\theta)] = \int_{-\infty}^{\infty} h(\theta)p(\theta|y)d\theta$) by Monte Carlo simulation (i.e. $\mathbb{E}_{\theta|y}[h(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})$). Trivially, Monte Carlo integration requires methods that are able to simulate $\theta^{(m)}$ from the posterior distribution $p(\theta|y)$, such as:

- Direct sampling;
- Importance sampling;
- Markov Chain Monte Carlo [MCMC] methods:
  - Gibbs sampling;
  - Metropolis-Hasting method.

Each of these simulation methods manages to provide a set of draws $\{\theta^{(m)}, m = 1, ..., M\}$ from the posterior distribution $p(\theta|y)$, which can be used to compute posterior results (e.g. posterior means, modes, or variances). Advantages of Monte Carlo integration are that i) the speed of convergence does not depend on the dimension of the integral, and ii) the

computation burden does not have to increase exponentially with the dimension of the integrals, such that iii) this method is suited for high-dimensional integrals. In contrast, a disadvantage is that the method converges at slower speed with larger approximation errors.

**What is importance sampling?** In case the posterior distribution is difficult to sample directly from, but an easier, so-called importance function $g(\theta)$ exists, then we sample $\theta^{(m)}$ from $g(\theta)$. The importance function must be a good approximation of the posterior density function $p(\theta|y)$, or the posterior kernel $k(\theta|y) = p(\theta)p(y|\theta)$. In addition, we compute importance weights $w(\theta^{(m)})$, which reflect how important the sampled draw is for approximating the integral. Given that we know either the posterior density or the posterior kernel function, we distinguish the following two cases:

- **If we know the exact posterior density function $p(\theta|y)$ with integrating constants**: Importance weights are defined as $w(\theta^{(m)}) = \frac{p(\theta^{(m)}|y)}{g(\theta^{(m)})}$. Hence, the original integral $\int h(\theta)p(\theta|y)d\theta$ changes to $\int h(\theta)\frac{p(\theta|y)}{g(\theta)}g(\theta)d(\theta)$, and the original Monte Carlo simulation expression changes from $\mathbb{E}_{\theta|y}[h(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})$ to $\mathbb{E}_{\theta|y}[h(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} w(\theta^{(m)})h(\theta^{(m)})$.

- **In case we only know the posterior kernel $p(\theta)p(y|\theta)$**: Importance weights are defined as $w(\theta^{(m)}) = \frac{p(\theta^{(m)})p(\theta^{(m)}|y)}{g(\theta^{(m)})}$. As an intermediate step, note that the original integral $\int h(\theta)p(\theta|y)d\theta$ can be rewritten as $\frac{\int h(\theta)p(\theta)p(y|\theta)d\theta}{\int p(\theta)p(y|\theta)d\theta}$. Then, the original integral changes to $\frac{\int h(\theta)\frac{p(\theta)p(y|\theta)}{g(\theta)}g(\theta)d\theta}{\int \frac{p(\theta)p(y|\theta)}{g(\theta)}g(\theta)d\theta}$, and the original Monte Carlo simulation expression changes from $\mathbb{E}_{\theta|y}[h(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})$ to $\mathbb{E}_{\theta|y}[h(\theta)] \approx \frac{\sum_{m=1}^{M} w(\theta^{(m)})h(\theta^{(m)})}{\sum_{m=1}^{M} w(\theta^{(m)})}$.

**What are Markov Chain Monte Carlo [MCMC] methods?** Whereas with importance sampling one samples from the same function $g(\theta)$ that is only an approximation of the posterior throughout the entire simulation process, with MCMC methods one iteratively samples from a Markov chain constructed to equal the posterior in the limit. Once converged, one can use draws from the limiting distribution to compute results of interest (e.g. posterior means, variances, or marginal densities).

**What is Gibbs sampling?** [Applications: sampling from a (truncated) bivariate normal distribution.] The Gibbs sampler is a type of MCMC sampler that divides the parameter of interest $\theta$ into $d$ blocks $(\theta_1, ..., \theta_d)$. Next, each of these blocks is associated with a

marginal distribution, that is, $p(\theta_1|\theta_2, ..., \theta_d, y)$, ..., $p(\theta_d|\theta_1, ..., \theta_{d-1}, y)$, respectively. Since the Gibbs sampler is an MCMC sampler, the Markov chain component appears through the marginal distributions: for the next iteration $(m+1)$, we simulate draws $\theta_1^{(m+1)}, ..., \theta_d^{(m+1)}$ from $p(\theta_1|\theta_2^{(m)}, ..., \theta_d^{(m)}, y)$, ..., $p(\theta_d|\theta_1^{(m)}, ..., \theta_{d-1}^{(m)}, y)$, respectively. After point of convergence $m = m^*$, all draws $\{\theta^m, m \geq m^*\}$ can be used as a sample from the joint posterior $p(\theta_1, ..., \theta_d|y)$.

**What is the simulation error?** The error margin of the simulation procedure can be computed for each posterior quantity of interest, given whether the draws are independent (e.g. importance sampling) or correlated (e.g. MCMC sampling). In the easy case of computing the posterior mean $\hat{\bar{\theta}} = \frac{1}{M}\sum_{m=1}^{M} \theta^{(m)}$, the variance of this estimator (i.e. the variance of the simulation error for computing the posterior mean, and NOT the posterior variance) is given by:

- **In case of independent draws**: $V[\hat{\bar{\theta}}] = \frac{\sigma_\theta^2}{M}$;
- **In case of correlated draws**: $\frac{\sigma_\theta^2}{M}(1+2\sum_{j=1}^{M} \rho_j) \geq \frac{\sigma_\theta^2}{M}$, where $\rho_j = \text{corr}(\theta^{(m)}, \theta^{(m-j)})$ can be estimated using a HAC type estimator.

A larger correlation causes slower convergence of the Markov chain, and requires more simulation iterations $M^*$ to get a more accurate posterior mean estimate (in fact, the effective sample size under $M$ correlated draws is only $\text{EES} = \frac{M}{1+2\sum_{j=1}^{M} \rho_j}$). Next, the concept of thinning (with thinning value $k$) can be applied to i) obtain an independent sample of draws where each $k$th draw is selected (and remaining draws deleted), ii) reduce storage space, or iii) reduce the number of draws when making graphs.

**How to check for convergence of MCMC sampler?** We can assess convergence of, for example, our Gibbs sampler by the following tools:

- Traceplots (informal): visualize consecutive draws to assess whether parameters fluctuate around a constant mean.
- Geweke test (formal test for single Markov chain): Equal mean test to test whether the means of the first 10% and last 50% of draws after convergence (or burn-in) are the same.
- Gelman-Rubin test (formal test to compare $L$ chains): Based on the chains' estimated within-variance $\hat{W}$ and between-variance $\hat{B}$ (which has to be small for con-

vergence!), the overall variance of $\theta$ is given by $\hat{V} = (1 - \frac{1}{M})\hat{W} + \frac{1}{M}\hat{B}$, where the Gelman-Rubin then equals $\hat{R} = \frac{\hat{V}}{\hat{W}}$ (which needs to converge to 1 if $\hat{B} \to 0$).

# 3 True or False Questions

1. When applying Monte Carlo integration, importance sampling can be used to simulate from the posterior distribution in case only the kernel function is known.

2. When the importance function $g(\theta|y)$ is a bad approximation of the posterior distribution, importance weights $w(\theta^{(m)})$ may be close to one such that the sampling method fails.

3. In case one uses a non-conjugate yet informative prior $p(\beta)$ leading to an unknown posterior kernel, one may choose the importance function $g(\theta|y)$ to equal the proper likelihood function such that importance weights simplify to $w(\theta^{(m)}) = p(\beta)$.

4. MCMC sampling uses the same function approximating the posterior throughout the entire simulation process, whereas importance sampling uses an iterative simulation procedure to simulate from a function that equals the posterior in the limit.

5. The lower the thinning value $k$ for MCMC sampling when thinning under an initial amount of $M$ independent draws, the more samples $M^*$ one should draw to obtain an effective sample size of EES $= M$.

6. The simulation error variance tends to be larger for sample draws from importance sampling than from Gibbs sampling.

7. The larger the correlation between subsequent sample draws from an MCMC sampler, the slower the convergence of the Markov chain.

8. For a large number of draws $M$, the Monte Carlo estimate of the mean equals $\mathbb{E}_{\theta|y}[h(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})p(\theta^{(m)}|y)$, where the sampling error decreases with $\sqrt{M}$ for correlated draws.