ECO545A: Bayesian Econometrics

February 19, 2025

Chapter 4: Prior Distributions

We will explore prior distributions with respect to the linear regression model, which is the workhorse of econometric and statistical modelling..

Regression Model

The linear regression model can be expressed by the following equation:

$$y_i = x_i'\beta + \epsilon_i,$$

where $x_i = (1, x_{i2}, ..., x_{ik})'$ and $\beta = (\beta_1, ..., \beta_k)'$ are column vectors of dimension $k \times 1$. Furthermore, we assume that $\epsilon_i | x_i \sim N(0, \sigma^2)$ where σ^2 is unknown. Consequently, $y_i \sim N(x_i'\beta, \sigma^2)$ for $i = 1, \cdots, n$.

The assumption $E(\epsilon_i|x_i) = 0$ implies that $E(\epsilon_i) = 0$ and $cov(x_i, \epsilon_i) = 0$. Under the assumption of joint normality of (ϵ_i, x_i), the assumption $cov(x_i, \epsilon_i) = 0$ (following from $E(\epsilon_i) = 0$) implies that each x_{ik} is independent of ϵ_i. Such covariates are said to be exogenous.

We minimize $\sum \epsilon_i^2$ with respect to β, to obtain the least-squares estimator. However, the estimators have a convenient form when the regression model is expressed in the matrix formulation.

Regression in Matrix Form

The regression model in the matrix formulation can be expressed as follows,

$$y = X\beta + \epsilon,$$

where $y = (y_1, ..., y_n)'$ is a column vector of dimension $n \times 1$, X is a matrix of dimension $n \times k$ defined as,

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

$\beta = (\beta_1, \cdots, \beta_k)'$ is a $k \times 1$ column vector of unknown parameters, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$ is $n \times 1$ vector of errors.

Inference in this model involves finding point estimates for β and σ^2, construct interval estimates for the parameters, compare models that contains different set of covariates and predicts a value of y_i for a given set of covariates values.

Here, β_1 is the intercept and $\beta_k = \frac{\partial E(y_i|x_i)}{\partial x_{ik}}$ gives the change on the expected value of y_i due to a small change in x_{ik}, if x_{ik} is continuous. When x_{ik} is discrete (indicator or dummy variable), β_k is the shift in the intercept associated with a change from $x_{ik} = 0$ to $x_{ik} = 1$.

If y is in log terms and x_d is a dummy variable, then $\frac{y_1}{y_0} - 1 \approx \beta_d$, for small β_d, where $y_1(y_0)$ is the value of y when $x_d = 1$ ($x_d = 0$).

Least Squares Estimator

To find the least-squares estimator, we do not need a distributional assumption on ϵ. Instead, we just need $E(\epsilon_i) = 0$, $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$ and $V(\epsilon_i) = \sigma^2$ for all $i = 1, \cdots, n$.

$$\begin{aligned} S = \arg\min_{\beta} \epsilon'\epsilon &= \arg\min_{\beta} (y - X\beta)'(y - X\beta) \\ &= \arg\min_{\beta} (y'y - \beta'X'y - y'X\beta + \beta'X'X\beta) \end{aligned}$$

Differentiating with respect β and equating to 0, yields

$$\begin{aligned} \frac{\partial S}{\partial \beta} &= -2X'y + 2X'X\beta = 0, \qquad \text{or,} \\ \hat{\beta} &= (X'X)^{-1}(X'Y) \end{aligned}$$

Please check that the SOC satisfies the minimization criterion.

Least squares estimator for σ^2 is $\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-k}$. Note that, unlike the maximum-likelihood estimator of σ^2, $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

According to the Gauss-Markov theorem, $\hat{\beta} = (X'X)^{-1}(X'Y)$ is the best linear unbiased estimator (BLUE) i.e., the OLS estimator has the smallest variance amongst the class of all linear and unbiased estimator.

MLE and Bayesian Method

Both maximum likelihood estimator (MLE) and Bayesian method assumes a distribution on the error term to get the likelihood.

Suppose, we assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ then the likelihood for our regression model has the

following expression,

$$\begin{aligned} L(\beta, \sigma^2; y) &= \prod_{i=1}^{n} f(y_i|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right] \\ &\propto (2\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right] \end{aligned}$$

While obtaining the ML estimator, we always maximize the log of the likelihood. Logarithm being a monotonic transformation yields the same estimator, but simplifies the calculation.

Prior Distributions

Improper priors: Prior distributions that are not integrable i.e. their integral is infinite.

Example 1: Consider a data generated from a normal distribution with unknown mean μ. A possible way to show prior ignorance is to assume a uniform distribution i.e. $\pi(\mu) \propto c$, $c > 0$, $-\infty \leq \mu \leq \infty$. This prior is improper, its integral is unbounded and it cannot be normalized to one.

Example 2: For a normal linear regression model, a uniform prior on β i.e. $\pi(\beta) \propto c$, $c > 0$ is improper. Similarly, the Jeffreys prior on σ is also improper, which is $\pi(\sigma) \propto \frac{1}{\sigma}$. This corresponds to a uniform prior on $\log \sigma$. To see this, let $u = \log(\sigma)$ and apply the usual rule for the change of variable: $f(u) = g(\sigma)\left|\frac{d\sigma}{du}\right| = \frac{1}{\sigma}\sigma = 1$.

Note: Jeffrey's prior is the square root of the determinant of the information matrix.

The ability to specify proper priors is crucial for the use of Bayes factors and posterior odds ratio in comparing models. Because an improper prior is not normalizable, $c\pi(\cdot)$, $c > 0$ is equivalent to the prior specified by $\pi(\cdot)$. This implies

$$f(y|M) = \int f(y|\theta, M) c\pi(\theta|M) d\theta.$$

So, the marginal likelihood can be set to any positive number by choice of 'c'. This has implications for Bayes factor.

If two models are compared with improper priors,

$$\begin{aligned} B_{12} &= \frac{\int f_1(y|\theta_1, M_1) c_1 \pi_1(\theta|M_1) d\theta_1}{\int f_2(y|\theta_2, M_2) c_2 \pi_2(\theta|M_2) d\theta_2} \\ &= \frac{c_1}{c_2} \frac{\int f_1(y|\theta_1, M_1) \pi_1(\theta|M_1) d\theta_1}{\int f_2(y|\theta_2, M_2) \pi_2(\theta|M_2) d\theta_2}. \end{aligned}$$

Since, both c_1 and c_2 are arbitrary, BF can take any value chosen by the researcher. This is true even if only one prior is improper.

Where the prior is proper, the value of marginal likelihood is well defined. We will (almost) always work with proper priors.

Conjugate Priors

For the normal linear regression model, a conjugate prior distribution is one for which the posterior distribution $\pi(\beta, \sigma^2|y)$ is in the same family of distributions as the prior $\pi(\beta, \sigma^2)$. Two different distributions are in the same family when they have the same form and different parameters.

For the normal linear regression model, the conjugate prior distribution is given by the normal-inverse Gamma distribution given by:

$$\pi(\beta, \sigma^2) = \pi(\beta|\sigma^2)\pi(\sigma^2) = N_k(\beta|\beta_0, \sigma^2 B_0)\ IG\bigg(\sigma^2\Big|\frac{\alpha_0}{2}, \frac{\delta_0}{2}\bigg)$$

where β_0, B_0, α_0 and δ_0 are the hyperparameters and known. Point to note is that the prior for β depends on σ^2.

Posterior distribution

Employing the Bayes' theorem, the joint posterior distribution can be obtained as the product of the likelihood and the prior distributions. This is done as follows:

$$\begin{aligned}
\pi(\beta, \sigma^2|y) &\propto f(y|\beta, \sigma^2)\pi(\beta|\sigma^2)\pi(\sigma^2) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\bigg[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\bigg]\left(\frac{1}{\sigma^2}\right)^{k/2} \\
&\quad \times \exp\bigg[-\frac{1}{2\sigma^2}(\beta - \beta_0)'B_0^{-1}(\beta - \beta_0)\bigg]\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0}{2}+1} \exp\bigg[-\frac{\delta_0}{2\sigma^2}\bigg] \\
&\propto \left(\frac{1}{\sigma^2}\right)^{(\alpha_0+n)/2+1}\left(\frac{1}{\sigma^2}\right)^{k/2} \exp\bigg[-\frac{1}{2\sigma^2}\Big\{(y - X\beta)'(y - X\beta) \\
&\quad + (\beta - \beta_0)'B_0^{-1}(\beta - \beta_0) + \delta_0\Big\}\bigg],
\end{aligned}$$

By expanding the expression under curly braces and completing the square in β, we have

$$\pi(\beta, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{k/2} \exp\bigg[-\frac{1}{2\sigma^2}(\beta - \bar{\beta})'B_1^{-1}(\beta - \bar{\beta})\bigg] \times \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_1}{2}+1} \exp\bigg[-\frac{\delta_1}{2\sigma^2}\bigg]$$

where the updated parameters are:

$$\begin{aligned}
B_1 &= (X'X + B_0^{-1})^{-1} \\
\bar{\beta} &= B_1(X'y + B_0^{-1}\beta_0) \\
\alpha_1 &= \alpha_0 + n \\
\delta_1 &= \delta_0 + y'y + \beta_0 B_0^{-1}\beta_0 - \bar{\beta}'B_1^{-1}\bar{\beta}
\end{aligned}$$

The above equation is recognized as the product of a normal and inverse gamma distribution i.e. $\beta|\sigma^2, y \sim N(\bar{\beta}, \sigma^2 B_1)$ and $\sigma^2|y \sim IG(\frac{\alpha_1}{2}, \frac{\delta_1}{2})$.

The conjugate prior allows us to find analytically, the marginal posterior distribution of β

and σ^2. We already found the marginal distribution of σ^2. The marginal distribution of β is derived below:

$$\pi(\beta|y) = \int \pi(\beta, \sigma^2|y)d\sigma^2 \propto \int \left(\frac{1}{\sigma^2}\right)^{\frac{k+\alpha_1}{2}+1} \exp\left[-\frac{Q}{2\sigma^2}\right]d\sigma^2$$

where $Q = \delta_1 + (\beta - \bar{\beta})' B_1^{-1}(\beta - \bar{\beta})$ and the integrand is in the form of an inverted-Gamma function. Therefore,

$$\begin{aligned}
\pi(\beta|y) &= \Gamma\left(\frac{k+\alpha_1}{2}\right)\left(\frac{Q}{2}\right)^{-\frac{(k+\alpha_1)}{2}} \\
&\propto Q^{-\frac{(k+\alpha_1)}{2}} \\
&\propto \left[\delta_1 + (\beta - \bar{\beta})' B_1^{-1}(\beta - \bar{\beta})\right]^{-\frac{(k+\alpha_1)}{2}} \\
&\propto \left[1 + \frac{1}{\alpha_1}(\beta - \bar{\beta})'\Big[(\delta_1/\alpha_1)B_1\Big]^{-1}(\beta - \bar{\beta})\right]^{-\frac{(k+\alpha_1)}{2}}.
\end{aligned}$$

This is recognized as the kernel of a multivariate -t distribution. So, $\pi(\beta|y) = t_k(\alpha_1, \bar{\beta}, (\delta_1/\alpha_1)B_1)$.

Exchangeability

The concept of exchangeability was proposed by De Finnetti. The random variables $(z_1, ..., z_n)$ are finitely exchangeable if $f(z_1, ..., z_n)$ is invariant to permutations in the indices $1, 2, ..., n$. For example, if n=3, then the random variables are exchangeable if:

$$f(z_1, z_2, z_3) = f(z_1, z_3, z_2) = f(z_2, z_1, z_3) = f(z_2, z_3, z_1) = f(z_3, z_1, z_2) = f(z_3, z_2, z_1)$$

Exchangeability generalizes the concept of independence. Identically distributed and mutually independent random variables are exchangeable but exchangeability does not imply independence.

As an example of exchangeability applied to prior distributions, consider a linear regression model with heteroscedasticity:

$$y_i = x_i'\beta + \epsilon_i \qquad \text{where} \qquad V(\epsilon_i) = \lambda_i^{-1}\sigma^2$$

So, the likelihood for the heteroscedastic linear regression model is given by:

$$f(y|\beta, \sigma^2, \lambda) = \prod_{i=1}^{n} f(y_i|\beta, \sigma^2, \lambda_i) = \prod_{i=1}^{n} N(x_i'\beta, \lambda_i^{-1}\sigma^2),$$

and the assumed prior distributions are,

$$\text{Priors}: \quad \beta \sim N(\beta_0, B_0), \;\; \sigma^2 \sim IG\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right), \;\; \lambda_i \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

The assumption about the distribution of λ_i embodies exchangeability: each i is associated

with a particular λ_i but knowing the value i gives no additional information because they are independently drawn from a Gamma distribution.

The heteroskedastic regression model is also important as an extension of the linear model. Note that the prior family is not conjugate and the posterior distribution does not permit analytical integration to obtain marginal posterior distributions of β and σ^2. However, the model has an interesting property. From $\epsilon_i|\lambda_i, \sigma^2 \sim N(0, \lambda_i^{-1}\sigma^2)$ and $\lambda_i \sim Ga\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big)$, we have,

$$\begin{aligned} f(\epsilon_i, \lambda_i|\sigma^2) &\propto \lambda_i^{1/2} \exp\left[-\frac{\lambda_i}{2\sigma^2}\,\epsilon_i^2\right] \lambda_i^{\frac{\nu}{2}-1}\, \exp\left[-\frac{\nu\lambda_i}{2}\right] \\ &= \lambda_i^{\frac{(\nu+1)}{2}-1} \exp\left[-\frac{\lambda_i\,(\epsilon_i^2+\nu\sigma^2)}{2\sigma^2}\right]. \end{aligned}$$

We see that $\lambda_i|\epsilon_i, \sigma^2 \sim Ga\left(\frac{(\nu+1)}{2}, \frac{(\epsilon_i^2+\nu\sigma^2)}{2\sigma^2}\right)$ and hence the marginal distribution of ϵ_i has the integrating constant $(\epsilon_i^2+\nu\sigma^2)^{-\frac{(\nu+1)}{2}}$ which yields $(\epsilon_i^2+\nu\sigma^2)^{-\frac{(\nu+1)}{2}} \propto \left(1+\frac{\epsilon_i^2}{\nu\sigma^2}\right)^{-\frac{(\nu+1)}{2}}$ and is recognized as the kernel of a student-t distribution $t(\nu, 0, \sigma^2)$.

So if $\epsilon_i|\lambda_i, \sigma^2 \sim N(\cdot)$ and $\lambda_i \sim Ga(\cdot)$, then $\epsilon_i|\sigma^2 \sim t$ distribution. We can say that the distribution of ϵ_i is a conditionally heteroscedastic because the variance of $\epsilon_i|\lambda_i, \sigma^2 = \lambda_i^{-1}\sigma^2$ but the distribution of $\epsilon_i|\sigma^2$ is homoscedastic.

Hierarchical Models

In hierarchical models, the hyperparameters themselves are given a prior distribution depending on another set of hyperparameters. Thus, we add one or more additional levels. For example,

$$y \sim f(y|\theta), \qquad \theta \sim \pi(\theta|\alpha_0), \qquad \text{and} \qquad \alpha_0 \sim \pi(\alpha_0|\alpha_{00})$$

where α_{00} is specified. Note the following points: (1) α_0 is not identified from the data because $f(y|\theta, \alpha_0) = f(y|\theta)$, and (2) α_0 can be eliminated from the model because

$$\pi(\theta|\alpha_{00}) = \int \pi(\theta|\alpha_0)\pi(\alpha_0|\alpha_{00})d\alpha_0.$$

Accordingly, α_0 is neither identified nor necessary for analyzing the model.

Example: Consider the heteroscedastic linear model. Here $\lambda_i \sim Ga(\frac{\nu}{2}, \frac{\nu}{2})$. The parameter $\nu > 0$ can be given a prior distribution such as Gamma or Poisson truncated to $\nu > 0$.

Conditionally Conjugate Priors

We assume, $\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2)$ instead of $\pi(\beta, \sigma^2) = \pi(\beta|\sigma^2) \times \pi(\sigma^2)$. So, we allow the priors to be independent, instead of them being dependent on each other. We continue to assume Normal and inverse-Gamma distribution for β and σ^2, respectively. Hence, we have

$$\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2) \propto \exp\left[-\frac{1}{2\sigma^2}(\beta-\beta_0)'B_0^{-1}(\beta-\beta_0)\right] \times \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0}{2}+1} \exp\left[-\frac{\delta_0}{2\sigma^2}\right]$$

The posterior distribution can be obtained as product of the likelihood times the prior distributions. However, we cannot separate them into product of two marginal distributions or of a marginal and conditional distribution. However, we can obtain the conditional posteriors $\pi(\beta|\sigma^2, y)$ and $\pi(\sigma^2|\beta, y)$.

For the linear regression model $y_i = x_i'\beta + \epsilon_i$ with independent priors, the conditional posterior distribution for β has the form,

$$\begin{aligned} \beta|\sigma^2, y &\sim N(\bar{\beta}, B_1), \\ \text{where} \quad B_1 &= \Big[\sigma^{-2}X'X + B_0^{-1}\Big]^{-1}, \\ \bar{\beta} &= B_1\Big[\sigma^{-2}X'y + B_0^{-1}\beta_0\Big], \end{aligned}$$

and the conditional posterior distribution for σ^2 has the form,

$$\begin{aligned} \sigma^2|\beta, y &\sim IG\left(\frac{\alpha_1}{2}, \frac{\delta_1}{2}\right), \\ \text{where} \quad \alpha_1 &= \alpha_0 + n, \\ \delta_1 &= \delta_0 + (y - X\beta)'(y - X\beta). \end{aligned}$$

When the conditional posterior distribution is in the same family as the prior, the prior is said to be conditionally conjugate or semi-conjugate.

Note: How to find out the conditional posterior distribution? Find out the joint posterior and then collect terms for the parameter for which you want to find out the distribution, holding others fixed. Then try to recognize the distribution.

As another example of conditionally conjugate prior, consider the heteroscedastic linear regression model.

$$y_i = x_i'\beta + \epsilon_i, \qquad \text{where} \qquad \epsilon_i|\lambda_i \sim N(0, \lambda_i^{-1}\sigma^2),$$

with the following prior distribution on the parameters,

$$\beta \sim N(\beta_0, B_0), \quad \sigma^2 \sim IG\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right), \quad \text{and} \quad \lambda_i \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

We can adopt an hierarchial approach by constructing ν as an additional parameter and specifying a prior distribution.

Prior: Assume that ν takes on J values $\nu_1, \ldots, \nu_J$ with the probabilities $p_{10}, \ldots, p_{J0}$ respectively. Then the joint posterior distribution can be shown to have the expression,

$$\begin{aligned} \pi(\beta, \sigma^2, \lambda, \nu|y) &\propto \pi(\beta)\pi(\sigma^2)\prod_{i=1}^{n}\left\{\left(\frac{\lambda_i}{\sigma^2}\right)^{1/2}\exp\Big[-\frac{\lambda_i}{2\sigma^2}(y_i - x_i'\beta)^2\Big]\right. \\ &\times \left.\lambda_i^{\frac{\nu}{2}-1}\frac{1}{\Gamma(\frac{\nu}{2}, \frac{\nu}{2})}\exp\left[-\frac{\nu\lambda_i}{2}\right]\right\}\sum_{j}^{J} p_{j0}1(\nu = \nu_j), \end{aligned}$$

where 1(A) is the indicator function. Note that only the last four terms involve ν and these

involve no parameters other than the λ_i. This reflects that ν is not identified by the data, although it depends on y through λ_i. We therefore have,

$$\pi(\nu_j|\lambda) \quad \propto \quad p_{j0}\left(\prod_{i=1}^{n}\lambda_i\right)^{\nu_j/2-1}\frac{1}{\Gamma(\frac{\nu_j}{2},\frac{\nu_j}{2})}\exp\left[-\frac{\nu_j\sum\lambda_i}{2}\right] \qquad \text{for} \quad j=1(1)J.$$

The requirement $\sum_j \pi(\nu_j|\lambda) = 1$ is enforced by dividing the individual terms in the above equation by their sum.

Chapter 5: Classical Simulation

Simulation has greatly expanded the scope of Bayesian Inference. We first review methods for generating independent samples from probability distributions.

Inverse CDF or Probability Integral Transformation Methhod

Let Z be a random variable and Z=z be a particular value, such that $F(Z) = z$ is invertible (i.e. you can take an inverse) and $F^{-1}(\cdot)$ is unique with probability 1. Draw $U \sim Unif(0,1)$ and set $F(z) = u$, then solve for z. The value of z is a draw from $F(Z)$

Algorithm 1 (Inverse CDF Method)

---

1. Draw u from Unif (0,1)
2. Return $y = F^{-1}(u)$ as a draw from $f(y)$

Example 1: Suppose you want to draw $Z \sim Unif(a,b)$, then the *cdf* is given by,

$$F(z) = \frac{(z-a)}{(b-a)} 1(a \leq z \leq b).$$

Draw $u \sim Unif(0,1)$ and solve $F(z) = u$ for z. The values $z = a + (b-a)u$, are drawn from $U(a,b)$.

Example 2: Suppose, we want to draw $Z \sim \mathcal{E}(\lambda)$ with *pdf* and *cdf* as given by,

$$\begin{aligned} f(z) &= \lambda \exp(-\lambda z) \\ F(z) &= 1 - \exp(-\lambda z). \end{aligned}$$

Draw $u \sim Unif(0,1)$ and solve $u = 1 - \exp(-\lambda z)$, which gives $z = -\log(1-u)/\lambda$.

Example 3: The inverse CDF method is routinely used to draw random numbers from truncated distributions. Suppose, the *cdf* of X is given by $F(X)$ and $c_1 \leq X \leq c_2$. Then the *cdf* of the truncated values is

$$\frac{[F(X) - F(c_1)]}{[F(c_2) - F(c_1)]} \qquad for\, c_1 \leq X \leq c_2.$$

Generate $u \sim U(0,1)$ and set

$$u = \frac{F(x) - F(c_1)}{F(c_2) - F(c_1)},$$

which implies that,

$$x = F^{-1}\Big[F(c_1) + u\big(F(c_2) - F(c_1)\big)\Big].$$

Method of Composition

This method uses the relationship

$$f(x) = \int g(x|y)h(y)dy$$

where f, g, and h are all densities. The method is useful when we know how to sample y from $h(y)$. By drawing a y from $h(y)$ and then drawing x from $g(x|y)$, we are drawing x from $f(x)$.

Example 4: Heteroskedastic linear regression model. By drawing $\lambda_i \sim Ga(\frac{\nu}{2}, \frac{\nu}{2})$, and then drawing $\epsilon_i|\lambda_i \sim N(0, \lambda_i^{-1}\sigma^2)$, we are actually drawing $\epsilon_i \sim t(v, 0, \sigma^2)$.

The method of composition can be thought of as a mixture distribution where the density of interest can be written as the marginal distribution of a joint distribution, $f(x) = \int g(x, y)dy$. In this form $g(\cdot)$ is not written as the product of a conditional and marginal. The mixture distribution idea can be used when it is convenient to sample a joint distribution. The expression implies that the values of x are a sample from its marginal distribution when a sample (x, y) is generated from their joint distribution.

Accept-Reject Algorithm

The AR algorithm can be used to simulate values from a density function $f(\cdot)$ if it is possible to simulate values from a density $g(\cdot)$ and if a number 'c' can be found such that $f(Y) \leq cg(Y)$, $c \geq 1$ for all Y in the support of $f(\cdot)$. Here, $f(Y)$ is the target density, and $g(Y)$ is the proposal density.

Note: The target density must be dominated over the entire support of Y.

Algorithm 2 (Accept-Reject Algorithm)

---

1. Generate a value of y from $g(\cdot)$
2. Draw a value u from Unif(0,1)
3. Return y as a draw from $f(\cdot)$ if $u \leq \frac{f(y)}{cg(y)}$, else return to Step 1. The effect is to accept y with probability $\frac{f(y)}{cg(y)}$

A simple proof is below which shows that the method works. The distribution of the accepted values of y is $h\left[y|u \leq \dfrac{f(y)}{cg(y)}\right]$. By Bayes Theorem,

$$h\left[y|u \leq \frac{f(y)}{cg(y)}\right] = \frac{P\left[u \leq \frac{f(y)}{cg(y)}|y\right] g(y)}{\int P\left[u \leq \frac{f(y)}{cg(y)}|y\right] g(y)dy} = \frac{\left[\frac{f(y)}{cg(y)}\right] g(y)}{\frac{1}{c}\int f(y)dy} = f(y)$$

Note that $1/c$ is the unconditional probability that a generated value of y is accepted.

Example 5: Sampling from Beta$(3,3)$ with uniform distribution as the proposal density.

Maximum of Beta(3,3) occurs at $y = 1/2$ and $f(y) = 1.8750$. So set $c = 1.8750$. In this context, the uniform density is not a good proposal because it generates values uniformly along the horizontal axis and so values near 0 and 1 are over-sampled.

For a value 0.15 generated by the proposal, it is accepted with probability $0.4877/1.875 = 0.2601$

Algorithm 3 (Sampling from Beta(3,3) with Uniform as Proposal)

---

1. Draw u_1 and u_2 from $Unif(0,1)$
2. If $u_2 \leq \left[u_1^2(1-u_1)^2\right] / \left[B(3,3)1.875\right]$, return $y = u_1$, otherwise reject and return to step 1.

Example 6: We want to sample from N(0,1) using the proposal density $g(y) = \frac{1}{2}e^{-|y|}$.

Generate y from an exponential distribution. If the value of y is accepted, assign a positive value with probability 0.5 and a negative value with probability 0.5.

Determination of c: The ratio $f(y)/e^{-y} = (2\pi)^{-1/2}[e^{-y^2/2}/e^{-y}]$ is maximized at $y = 1$, implying $c = \sqrt{e/2\pi}$ and probability of acceptance equals $1/c = 0.6577$.

Algorithm 4 (Sampling from N(0,1) using the proposal density $g(y) = \frac{1}{2}e^{-|y|}$)

---

1. Generate u_1, u_2 and $u_3 \sim Unif(0,1)$
2. Sample x from exp(1) using the inverse CDF method.
   Setting $u_1 = G(x) = 1 - e^{-x}$ and solving for x yields $x = -\log(1-u_1) = -\log(u_1)$ because u_1 and $1-u_1$ have the same distribution.
3. If the condition,

$$\begin{aligned} u_2 &\leq (2\pi)^{-1/2} e^{-x^2/2} / c e^{-x}, \\ u_2 &\leq (2\pi)^{-1/2} \frac{e^{-x^2/2}}{\sqrt{\frac{e}{2\pi}} e^{-x}}, \\ u_2 &\leq e^{x - \frac{x^2}{2} - \frac{1}{2}}, \end{aligned}$$

   holds, then y $= \begin{cases} x, & \text{if } u_3 \leq 1/2 \\ -x, & \text{otherwise} \end{cases}$

   else return to step (1)

Importance Sampling

Importance sampling is a type of Monte Carlo Integration. Suppose $X \sim f(X)$ and we wish to estimate

$$E[g(X)] = \int g(x)f(x)dx$$

but the integral is not computable analytically and the method of composition is not available because we cannot sample from $f(x)$. We then use importance sampling.

Let $h(X)$ be a distribution from which we can sample, so,

$$E[g(X)] = \int \frac{g(x)f(x)}{h(x)}\, h(x)dx$$

We draw values from $h(x)$ and evaluate

$$E[g(X)] \equiv \frac{1}{G}\sum_{g=1}^{G} g(X^{(g)})\frac{f(X^{(g)})}{h(X^{(g)})},$$

to estimate the quantity of interest. The weight $w(X^{(g)}) = \frac{f(X^{(g)})}{h(X^{(g)})}$ is called the importance weight or importance ratio.

Note: Importance sampling is not a useful method if the importance ratios vary substantially. Worst scenario: importance ratios are small with high probability and very large with small probability. This happens when $g.f$ has thick tails compared to h. So, you want the proposal density to dominate the targets in the tail. This implies normal distribution is not a good proposal.

Accuracy and Efficiency of Importance Sampling Weights: To spot problems, plot the logarithm of importance ratios. Estimates will be poor if the largest ratios are too large relative to the average. If the variance of the weights is finite, the effective sample size can be estimated using the following approximation:

$$S_{eff} = \frac{1}{\sum_{s=1}^{S}[\tilde{w}(X^{(s)})]}, \qquad \text{where} \qquad \tilde{w}(X^{(s)}) = \frac{w(X^{(s)})}{\sum_{s=1}^{S}[w(X^{(s)})]}$$

S_{eff} is small, when there are few extremely high weights which unduly influence the distribution.

Example 7: Suppose $X \sim \mathcal{E}(1)$ truncated to [0,1] and we want to evaluate $E[(1+X^2)^{-1}]$. We approximate the integral,

$$\frac{1}{(1-e^{-1})}\int_0^1 \frac{e^{-x}}{1+x^2}\, dx$$

using importance sampling. Let Beta(2,3) be the proposal density. This choice of parameters is a good match for the shape of the distribution over (0,1) interval.

Algorithm 5 (Importance Sampling)

---

1. Generate a sample of G values $X^{(1)}, ..., X^{(G)}$ from Beta(2,3)
2. Calculate $\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{1+(X^{(g)})^2}\right)\left(\frac{e^{-X^{(g)}}}{1-e^{-1}}\right)\left(\frac{B(2,3)}{X^{(g)}(1-X^{(g)})^2}\right)$

Multivariate Simulation

Sample $X \sim N_p(\mu, \Sigma)$. Sample $z \sim N(0, I_p)$ and use the expression, $X = \mu + cz$ where $\Sigma = cc'$ and c is the lower triangular Cholesky matrix.

Chapter 6: Basics of Markov Chains

Finite State Spaces: Consider a stochastic process indexed by t, X_t that takes values in the finite set $S = \{1, 2, \ldots, S\}$. For any pair of integers $i, j \in S$, the transition probabilities p_{ij} are defined as,

$$p_{ij} = P(X_{t+1} = j | X_t = i), \qquad i, j \in S.$$

The assumption that the probability distribution at time $t+1$ depends only on the state of the system at time t is called the Markov property and the resulting stochastic process is called Markov process. Further, not that p_{ij} does not depend on t. This type of stochastic process is called homogenous Markov process.

Properties of Markov process:

- $p_{ij} \geq 0$
- $\sum_{j=1}^{S} p_{ij} = 1$
- We can construct a transition matrix P of dimension $S \times S$

Example of a transition matrix:

$$P = \begin{pmatrix} 0.75 & 0.25 \\ 0.125 & 0.875 \end{pmatrix}$$

Note: The sum across columns equals 1.

Distribution of the state at $t+2$ given that it is in i at t: To go from state i to state j in two steps, the process goes from i at t to any other state k at time $t+1$ and then from k to j at $t+2$. This transition occurs with probability: $p_{ij}^{(2)} = \sum_k p_{ik}\, p_{kj}$.

You may verify that matrix of $p_{ij}^{(2)}$ is given by $PP = P^2$. So by induction the values of $p_{ij}^{(n)}$ are $ij - th$ entries in the matrix P^n where n is any integer. We will be concerned with what happens to $p_{ij}^{(n)}$ as n becomes larger.

If $p_{ij}^{(n)} > 0$ for some $n \geq 1$, then:

- $i \longrightarrow j \implies j$ is accessible from i
- $i \longleftrightarrow j \implies i$ and j communicate i.e. $i \longrightarrow j$ and $j \longrightarrow i$

It can be shown that the communicating relationship between states define an equivalence relationship i.e.,

- Reflexivity: $i \longleftrightarrow i$
- Symmetry: $i \longleftrightarrow j \Longleftrightarrow j \longleftrightarrow i$
- Transitivity: $i \longleftrightarrow j$ and $j \longleftrightarrow k$ implies $i \longleftrightarrow k$

Irreducibility

A Markov process is irreducible if there is just one equivalence class. What this means is that starting from state i, the process can reach any other state with positive probability. The transition matrix P_R below is NOT irreducible:

$$P_R = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix},$$

where both P_1 and P_2 are of dimension $m \times m$. The matrix P_R is not irreducible because if you are in the first 'm' states, you can never leave them.

Periodicity

Consider the transition matrix of the form,

$$P_P = \begin{pmatrix} 0 & P_1 \\ P_2 & 0 \end{pmatrix},$$

Suppose, at $t = 1$ the process is in one of the first m states. Then, at $t = 2$ it must go to one of the second m and must return to the first m states at $t = 3$ and so on. This is described by saying that the period of the chain is 2. If the period is 1 for all states, then the chain is called aperiodic.

More formally, if d_i is the period of i, then $P_{ii}^{(n)} = 0$ whenever n is not a multiple of d_i and d_i is the largest integer with this property. Note that a chain is a aperiodic if $P_{ii}^{(n)} > 0$ for all i and for sufficiently large n.

Invariant Distribution

The probability distribution $\pi = (\pi_1, ..., \pi_n)$ is an invariant distribution for P if,

$$\begin{aligned} \pi' &= \pi' P \qquad\qquad \text{or,} \\ \pi_j &= \sum_i \pi_i p_{ij}, \qquad \text{for } j = 1, \ldots, S, \end{aligned}$$

where π' is a characteristic vector of P with a characteristic root equal to 1.

Lets look at $\pi_j = \sum_i \pi_i p_{ij}$: The left hand side of the equation is the probability that the process is in state j at any t marginalized over the states at $t - 1$, it can be interpreted as the probability of starting the process at state i with probability π_i and then moving to state j with probability p_{ij}.

The fact that the value on LHS is π_j is what makes π an invariant distribution: if the states are chosen according to π, the probability is π_j that the system is in state j at any time.

Example: Consider the transition matrix,

$$P = \begin{pmatrix} 0.750 & 0.250 \\ 0.125 & 0.875 \end{pmatrix},$$

From $\pi' P = \pi'$, we have

$$\begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} 0.750 & 0.250 \\ 0.125 & 0.875 \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix},$$

or $0.750\pi_1 + 0.125\pi_2 = \pi_1$ and because $\pi_2 = 1 - \pi_1$, we have $\pi_1 = 0.750\pi_1 + 0.125(1 - \pi_1)$ which implies $\pi = \big(1/3, 2/3\big)$.

An important topic in Markov chain theory is the existence and uniqueness of invariant distribution. A special case of an irreducible and aperiodic Markov chain is when $p_{ij} > 0$: For these chains you have the following theorem:

Theorem 6.1:

Suppose S is finite and $p_{ij} > 0$ for all i, j. Then there exists a unique probability distribution $\pi_j, j \in S$ such that $\sum_i \pi_i p_{ij} = \pi_j$ for all $j \in S$. Moreover, $|p_{ij}^n - \pi_j| \leq r^n$ where $0 \leq r \leq 1$ for all i, j and $n \geq 1$.

The theorem says that, in a finite state space with all probabilities positive, not only there is a unique invariant distribution, but also that $p_{ij}^{(n)}$ converges at a geometric rate (r^n) to π_j. For large n, the initial state plays no role.

Example: For our transition matrix P, we have

$$P^{10} = \begin{pmatrix} 0.339 & 0.661 \\ 0.330 & 0.670 \end{pmatrix}, \qquad \text{and} \qquad P^{20} = \begin{pmatrix} 0.333 & 0.667 \\ 0.333 & 0.677 \end{pmatrix}.$$

So, P^n has reached its invariant distribution for n = 20.

This theorem, in more general forms, is the basis of MCMC methods. This implies that if a Markov chain satisfies certain conditions, the probability distribution of its n^{th} iterate is, for very large n, very close to its invariant distribution. This has implication for simulation. If we can find a Markov process for which the invariant distribution is the distribution from which we wish to simulate (the target distribution), we can simulate draws from the process to generate values from the target distribution.

The restriction $p_{ij} > 0$ for all i, j is unnecessarily restrictive. Theorem 6.1 can be generalized to the following.

Theorem 6.2

Let P be irreducible and aperiodic over a finite state space. Then there is a unique probability distribution π such that $\sum_i \pi_i p_{ij} = \pi_j$ for all $j \in S$ and $|p_{ij}^n - \pi_j| \leq r^{n/v}$ for all $i, j \in S$ where $0 \leq r \leq 1$ for some positive integer v

Proof: See Bhattacharya and Waymire (1990).

Although Theorem 6.2 states that irreducibility and aperiodicity are sufficient to yield a result that justifies the MCMC method for finite state spaces, we need to consider more general state spaces because most applications involve continuous distributions. Before, we turn to these and the associated complications, lets have a look at Markov chains with a countable number of states.

Countable State Spaces

An example of a countable state space is the random walk. In this process, $S = \{0, \pm 1, \pm 2, ...\}$ is a countable state space and the transformation probabilities are:

$$p_{ij} = \begin{cases} p, & \text{if} \quad j = i + 1 \\ q, & \text{if} \quad j = i \\ r, & \text{if} \quad j = i - 1, \end{cases}$$

where $p + q + r = 1$. The figure in the book depicts that for $p = q = 0.5$, the process drifts with no clear pattern. If all three probabilities are positive, the process is irreducible and aperiodic.

Irreducibility and aperiodicity no longer imply the existence of a unique invariant distribution when S is a countable set but not finite. Consider a case, $p = 0.55, r = 0, q = 0.45$. Then the process drifts to ∞ in the sense $p_{ij}^{(n)} \to 0$ for all i, j. To salvage the counterparts of Theorem 6.1 and Theorem 6.2, we need the concept of "recurrence".

Let $P_j(A)$ denote the probability that event A occurs given that the process starts at j. Then state j is called recurrent if $P_j(X_n = j \quad i.o.) = 1$ where i.o. means infinitely often. In other words, the process returns to state j an infinite number of times with probability 1

If the state is not recurrent, it is transient. In the random walk with $p > q$ none of the states are recurrent. However, all the states are recurrent if $p = q$.

Note: If a process is irreducible, all states are either recurrent or transient.

Recurrence is now not strong enough to imply a unique invariant distributions. We need positive recurrence.

Positive and Null Recurrence:

Let $\tau_j^{(1)}$ be the time it takes for the process to make its first return to state j: $\tau_j^{(1)} = min\{n > 0 : X_n = j\}$. A state j is positive recurrent if $E(z_j^{(i)}) < \infty$. Else, it is null recurrent.

Let $E_j(\cdot)$ be the expected value of a random variable given that the process starts in state j, and $E_\pi(\cdot)$ be the expected value of a random variable with respect to the distribution.

Theorem 6.3

Assume that the process is irreducible. Then,

- If all the states are recurrent, they are either all positive recurrent or null recurrent.
- There exists an invariant distribution if and only if all states are positive recurrent. In that case, the invariant distribution π is unique and is given by $\pi_j = [E_j(\tau_j^{(1)})]^{-1}$

- In case the states are positive recurrent, for any initial distribution, if $E_\pi|f(X_1)| < \infty$, $\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{M} f(X_m) = E_\pi f(X_1)$.

This theorem has many implications for MCMC methods and generalizes readily to continuous state spaces. Under the conditions stated in the theorem, we know that:

- There is a unique invariant distribution
- Averages of functions evaluated at sample values converge to their expected values under the invariant distribution.

Since a possible function is the indicator function $1(X_n = i)$ which has expected value π_i, the invariant distribution can be estimated from sample data.

Convergence of P^n to a matrix whose rows are the invariant distribution requires aperiodicity.

Theorem 6.4

If P is an aperiodic recurrent chain, $\lim_{n\to\infty} P^n$ exists. If P is an aperiodic positive recurrent chain, then $\lim_{n\to\infty} P^n = A$, where A is a matrix whose rows are the invariant distribution.

Markov Chain Theory: Starts with transition probabilities P and determines conditions under which the rows of P^n converge to an invariant distribution.

MCMC Theory: Starts with an invariant distribution and finds a P that converges to it.

Because the existence of an invariant distribution is not in doubt, another theorem can be applied.

Theorem 6.5

Suppose P is a π-irreducible and that π is an invariant distribution for P. Then, P is a positive recurrent and π is the unique invariant distribution of P. If P is also aperiodic, then for π-almost all x, $||P^n(x_j.) - \pi|| \to 0$.

This theorem is also applicable to continuous case. π-irreducible means that for some n, $P^n(x, A) > 0$ for any set A such that $\pi(A) > 0$. Implication: recurrence need not be assumed explicitly if it is known an invariant distribution exists.

Continuous State Spaces

Now suppose that states of a Markov process take values in $\mathbf{R}$. With continuous states, transition probabilities give way to transition kernel or transition density $p(x, y)$.

The Markov property is captured by assuming that the joint density conditional on the initial value $X_0 = x_0$ is given by:

$$f(X_1, \ldots, X_n|X_0 = x_0)(x_1, ..., x_n) = p(x_0, x_1)p(x_1, x_2)\ldots p(x_{n-1}, x_n).$$

Starting at x, the probability that a process moves to a point in $A \subset R$ is given by,

$$P(x, A) = \int_A P(x, y)dy.$$

So, the n step ahead iteration is given by

$$P^{(n)}(x, A) = \int_A P(x, y)\ P^{(n-1)}(y, A)dy.$$

An invariant density $\pi(y)$ for the transition kernel $p(x, y)$ is a density that satisfies:

$$\pi(y) = \int_R \pi(x)p(x, y)dx.$$

Example: An invariant density is the AR(1) process. Consider the autoregressive process of order 1.

$$y_t = \theta y_{t-1} + u_t, |\theta| < 1 \qquad \text{and} \qquad u_t \sim N(0, \sigma^2).$$

For this process, $E(y_t) = 0$ and $V(y_t) = \frac{\sigma^2}{(1-\theta^2)}$. We now verify that the invariant distribution is Gaussian with these parameters i.e. 0 and $\frac{\sigma^2}{(1-\theta^2)}$.

$$\begin{aligned}
\pi(y_t) &= \int \pi(y_{t-1})p(y_{t-1}, y_t)dy_{t-1} \\
&\propto \int \exp\left[-\frac{(1-\theta^2)}{2\sigma^2}\ y_{t-1}^2 \right] \exp\left[-\frac{1}{2\sigma^2}(y_t - \theta y_{t-1})^2 \right] dy_{t-1} \\
&= \int \exp\left[-\frac{1}{2\sigma^2}\Big\{ (1-\theta)^2 y_{t-1}^2 + y_t^2 - 2\theta y_{t-1}y_t + \theta^2 y_{t-1}^2 \Big\} \right] dy_{t-1} \\
&= \int \exp\left[-\frac{1}{2\sigma^2}\Big\{ (y_{t-1} - \theta y_t)^2 + (1-\theta^2) y_t^2 \Big\} \right] dy_{t-1} \\
&\propto \exp\Big[-\frac{1-\theta^2}{2\sigma^2}\ y_t^2 \Big] \sim N\Big(0, \sigma^2/1-\theta^2\Big)
\end{aligned}$$

For continuous state spaces, the definitions of irreducibility and aperiodicity are as previously given, except p_{ij} is replaced with $p(x, y)$.

Recurrence for continuous state spaces: Let $P_x(A)$ denote the probability of event A given that the process starts at x. Then, $\pi-$ irreducible chain with invariant distribution π is recurrent if, for each B with $\pi(B) > 0$, we have

- $P_x(X_n \in B \ \ i.o.) > 0$ for all x
- $P_x(X_n \in B \ \ i.o.) = 1$ for $\pi-$ almost all x

The chain is Harris recurrent if $P_x(X_n \in B \ \ i.o.) = 1$ for all x. Let, total variation norm of a bounded, signed measure λ be denoted by $||\lambda|| = sup_A\lambda(A) - inf_A\lambda(A)$. Also, let total variation distance between λ_1 and λ_2 be denoted by $||\lambda_1 - \lambda_2||$.

Tierney(1994) states the following theorem:

Theorem 6.6

Suppose that P is $\pi-$ irreducible and that π is an invariant distribution for P. Then, P is positive recurrent and π is the unique invariant distribution of P. If P is also aperiodic, then for $\pi-$ almost all x.

$$||P^n(x,.) - \pi|| \to 0,$$

with $||.||$ denoting the total variation distance. If P is Harris recurrent then the convergence occurs for all x.

Theorem 6.7

If $||P^n(x,.) - \pi|| \to 0$ for all x, the chain is π-irreducible, aperiodic, positive recurrent and has invariant distribution π.

These theorems form the basis of MCMC methods. In practice, a researcher attempts to construct an irreducible, aperiodic and positive recurrent transition kernel for which the invariant distribution is the target distribution.

Note that all the preceding results generalize immediately to the case in which the random variables X_n are vectors.