NRES 779: Bayesian Hierarchical Modeling in Natural Resources

Lab 02: Probability

1 Learning Objectives

The purpose of this lab is to exercise what you have learned (or need to learn) about some important topics in Probability, including:

1. Converting DAGs to joint distributions,
2. Converting joint distributions to DAGs,
3. Simplifying,
4. Interpreting and factoring,
5. Probability distributions,
6. Marginal distributions,
7. Moment matching.

2 Motivation

Bayesian analysis is predicated on the idea that we learn about unobserved quantities from quantities we are able to observe. All observed quantities (i.e. parameters, latent states, missing data, and even the data themselves before they are observed) are treated as random variables in the Bayesian approach. All Bayesian analysis extends from the laws of probability, that is, from the "mathematics of random variables."

Random variables are quantities whose value is determined by chance. Statistical distributions represent how "chance" works by specifying the probability that a random variable takes on a value (in the discrete case) or falls within a range of values (in the continuous case). The goal of Bayesian analysis is to discover the characteristics of probability distributions that govern the behavior of random variables of interest, for example, the size of a population, the rate of nitrogen accumulation in a stream, the diversity of plants on a landscape, the change in lifetime income that occurs with changing level of education, the stress levels of juvenile elephants.

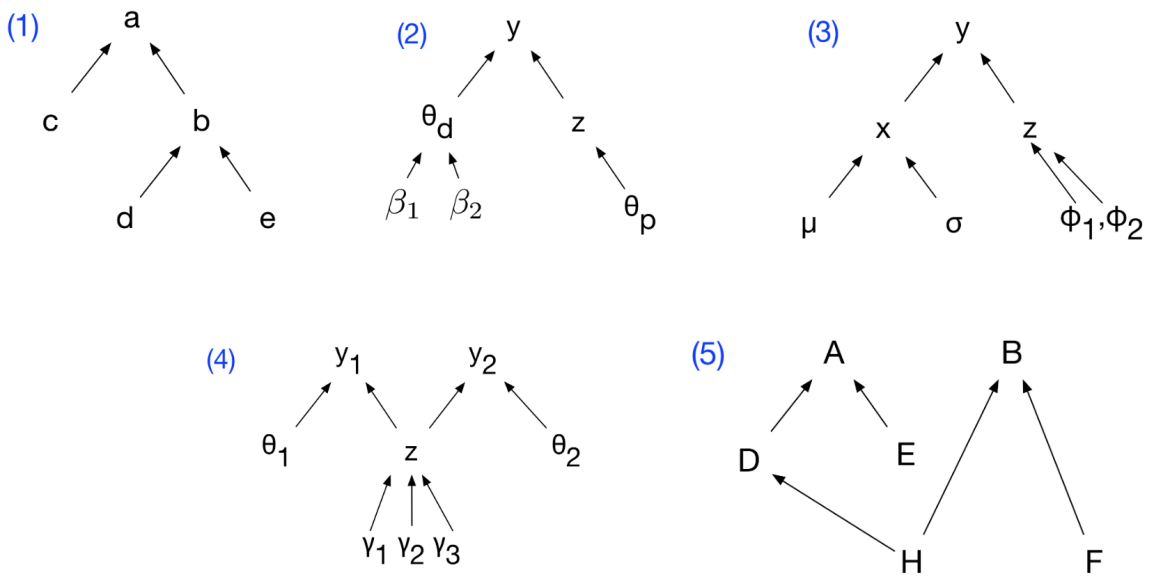
It follows that understanding the laws of probability and statistical distributions provides the foundation for Bayesian analysis. Keep in mind the following learning objectives

- Understand the concepts of conditional and independent random variables.

- Be able to write out joint distributions of random variables given Bayesian networks (directed acyclic graphs).
- Become familiar with frequently used statistical distributions representing discrete and continuous random variables.
- Learn R functions for calculating properties of distributions and for sampling from them.
- Understand discrete and continuous marginal distributions.
- Use moment matching, a procedure key to linking models to data in the Bayesian framework.

3 Converting DAGs to joint distributions

Write out the joint and conditional distributions for the following Bayesian networks. For discrete random variables, $[A]$ is equivalent to $Pr(A)$. For continuous random variables $[A]$ is the probability density of A.



4 Converting joint distributions to DAGs

Draw Bayesian networks (DAGs) for the joint and conditional distributions, below.

1. $Pr(A, B) = Pr(A|B)Pr(B)$
2. $Pr(A, B, C) = Pr(A|B, C)Pr(B|C)Pr(C)$
3. $Pr(A, B, C, D) = Pr(A|C)Pr(B|C)Pr(C|D)Pr(D)$
4. $Pr(A, B, C, D, E) = Pr(A|C)Pr(B|C)Pr(C|D, E)Pr(D)Pr(E)$
5. $Pr(A, B, C, D) = Pr(A|B, C, D)Pr(B|C, D)Pr(C|D)Pr(D)$
6. $Pr(A, B, C, D) = Pr(A|B, C, D)Pr(C|D)Pr(B)Pr(D)$

5 Simplifying

Simplify the expression below, given that z_2 and z_3 are independent random variables.

$$Pr(z_1, z_2, z_3) = Pr(z_1|z_2, z_3)Pr(z_2|z_3)Pr(z_3)$$

6 Interpreting and factoring

The probability of an observation y depends on a true ecological state of interest, z, and the parameters in a data model, $\boldsymbol{\theta}_d$. The probability of the true state z depends on the parameters in an ecological process model, $\boldsymbol{\theta}_p$. We know that $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_p$ are independent. Write out a factored expression for the joint distribution, $Pr(y, z, \boldsymbol{\theta}_d, \boldsymbol{\theta}_p)$. Drawing a Bayesian network will help.

7 Probability distributions

1. We commonly represent the following general framework for linking models to data:

   $$[y_i|f(x_i, \boldsymbol{\theta}), \sigma^2],$$

   which represents the probability of obtaining the observation y_i given that our model predicts the mean of a distribution $f(x_i, \boldsymbol{\theta})$ with variance σ^2. Assume we have count data. What distribution would be a logical choice to model these data? Write out a model for the data.

2. Choose the appropriate distribution for the types of data shown below and justify your decision.

   (a) The mass of carbon in above ground biomass in square meter plot.

   (b) The number of seals on a haul-out beach in the gulf of Alaska.

   (c) Presence or absence of an invasive species in forest patches.

   (d) The probability that a white male will vote republican in a presidential election.

   (e) The number of individuals in four mutually exclusive income categories.

   (f) The number of diseased individuals in a sample of 100.

   (g) The political party affiliation (democrat, republican, independent) of a voter.

3. Find the mean, variance, and 95% quantiles of 100,000 random draws from a Poisson distribution with $\lambda = 33.32$.

4. Simulate one realization of survey data with five categories on a Liker scale, (i.e. strongly disagree to strongly agree). Assume a sample size of 80 respondents and the following probabilities:

   - Strongly disagree = 0.07
   - Disagree = 0.13
   - Neither agree nor disagree = 0.15
   - Agree = 0.23
   - Strongly agree = 0.42

5. The average above ground biomass in a 1 km^2 of sagebrush grassland is 103.4 g/m^2, with a standard deviation of 23.3. You clip a 1 m^2 plot. Write out the model for the data. What is the probability density of an observation of 94 assuming the data are normally distributed? What is the probability that your plot will contain between 90 and 110 gm of biomass? Is there a problem using normal distribution?

   The normal distribution is not an ideal choice because it extends below 0, which is not possible for measurements of above ground biomass.

6. The prevalence of a disease in a population is the proportion of the population that is infected with the disease. Prevalence of chronic wasting disease in male mule deer on winter range near Georgetown, CA is 12 percent. A sample of 24 male deer included 4 infected individuals. Write out a model for the data. What is the probability of obtaining these data conditional on the given prevalence ($p = 0.12$)?

7. Researchers know that the true proportion of related age-sex classifications for elk in the Ruby Mountains are: Adult females ($p = 0.56$), Yearling males ($p = 0.06$), Bulls ($p = 0.16$), and Calves ($p = 0.22$). What is the probability of obtaining the classification data conditional on the known sex-age population proportions given the following counts?

   - Adult females (count=65)
   - Yearling males (count=4)
   - Bulls (Count = 25)
   - Calves (count = 26)

8. Nitrogen fixation by free-living bacteria occurs at a rate of 1.9 g/N/ha/yr with a standard deviation (σ) of 1.4. What is the lowest fixation rate that exceeds 2.5% of the distribution? Use a normal distribution for this problem, but discuss why this might not be a good choice.

   The normal distribution is not an ideal choice because it extends below 0, which is not possible for measurements of nitrogen fixation.

8 Marginal distributions

8.1 Discrete random variables: Diamond's pigeons

The holy grail in Bayesian analysis is to discover the marginal posterior distribution of unobserved quantities (parameters, latent states, missing data, forecasts) from quantities we are able to observe (data). It follows that we must understand what marginal distributions are. The following is an example of a discrete case that also exercises your newly gained familiarity with the laws of probability.

Jared Diamond studied the distribution of fruit pigeons (*Ptilinopus rivoli* and *P. solomonensis* on 32 islands in the Bismark archipelago northeast of New Guinea (Table 1). Define the event $\boldsymbol{R}$ as an island being occupied by *P. rivoli*, and the event $\boldsymbol{S}$ as an island being occupied by *P. solomonensis*. The complementary events are that an island is not occupied by *P. solomonensis* ($\boldsymbol{S}^c$) and not occupied by *P. rivoli* ($\boldsymbol{R}^c$).

Table 1: Data on distribution of species of fruit pigeons on islands

Status	Number of Islands
P. rivoli present, *P. solomonensis* absent	9
P. solomonensis present, *P. rivoli* absent	18
Both present	2
Both absent	3
Total	32

1. Fill in Table 2 to estimate the marginal probabilities of presence and absence of the two species. The cells show the joint probability of the events specified in the row and column. The right column and the bottom row show the marginal probabilities.

Table 2: Estimates of marginal probabilities for island occupancy

Events	S	S^c	Marginal
R	$\Pr\left(S, R\right) =$	$\Pr\left(S^c, R\right) =$	$\Pr\left(R\right) =$
R^c	$\Pr\left(S, R^c\right) =$	$\Pr\left(S^c, R^c\right) =$	$\Pr\left(R^c\right) =$
Marginal	$\Pr\left(S\right) =$	$Pr\left(S^c\right)$	$=$

   (a) What is the sum of the marginal rows?
   (b) What is the sum of the marginal columns?
   (c) Why? Note, when we marginalize over $\boldsymbol{R}$ we are effectively eliminating $\boldsymbol{S}$ and vice versa.

2. Use the data in Table 1 and the probabilities in Table 2 to illustrate the rule for the union of two events, the probability that an island contains either species, $Pr(R \cup S)$.

3. Use the marginal probabilities in Table 2 to calculate the probability that an island contains both species i.e., $Pr(R, S)$, assuming that R and S are independent. Compare the results from those calculations with the probability that both species occur on an island calculated directly from the data in Table 1.

4. Interpret the results ecologically. What is $Pr(R|S)$? What is $Pr(S|R)$.

5. Based on the data in Table 1, the probability that an island is occupied by both species is $232 = 0.062$. Diamond interpreted this difference as evidence of niche separation resulting for interspecific competition, an interpretation that stimulated a decade of debate. What are the conditional probabilities, $Pr(R|S)$ and $Pr(S|R)$?

8.2 Continuous random variables

We now explore marginal distributions for continuous random variables. This requires introducing a new distribution, the multivariate normal:

$$\mathbf{z} \sim \text{multivariate normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

or simply,

$$\mathbf{z} \sim \text{normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

because "multivariate" is implied when we have a vector in the mean location and covariance matrix in the variance location.

In equation 1, $\mathbf{z}$ is a vector of random variables, $\boldsymbol{\mu}$ is a vector of means (which can be the output of a deterministic model), and $\boldsymbol{\Sigma}$ is a covariance matrix. The diagonal of $\boldsymbol{\Sigma}$ contains the variances and the off-diagonal contains the covariance of $\boldsymbol{\Sigma}_{i,j}$. The covariance can be calculated as $\sigma_i \sigma_j \rho$, where σ_i is the standard deviation of the i^{th} random variable, σ_j is the standard deviation of the j^{th} random variable, and ρ is the correlation between the random variable i and j. The covariance matrix is square and symmetric (and also non-negative definite). We will learn more about these matrices later in the course. For now, an example will go a long way toward helping you understand the multivariate normal distribution.

The rate of inflation and the rate of return on investments are know to be positively correlated. Assume that the mean rate of inflation is 0.03 with a standard deviation of 0.015. Assume that the mean rate of return is 0.0531 with a standard deviation of 0.0746. Assume the correlation between inflation and rate of return is 0.5.

You can simulate interest rate and inflation data reflecting their correlation using the following function:

```
> library(MASS)
> DrawRates=function(n,int,int.sd,inf,inf.sd,rho.rates){
+   covar=rho.rates*int.sd*inf.sd
+   Sigma=matrix(c(int.sd^2,
+                  covar,
+                  covar,
+                  inf.sd^2),2,2)
+
+   mu=c(int,inf)
+   x=(mvrnorm(n=n,mu=mu,Sigma))
+
+   #x[x[,2]<0]=0 #do not allow for deflation
+
+   return(x)
+ }
> mu.int=0.0531
> sd.int=0.0746
> mu.inf=0.03
> sd.inf=0.015
> rho=0.5
> n=10000
> x=DrawRates(n,int=mu.int,int.sd=sd.int,inf=mu.inf,inf.sd=sd.inf,rho.rates=rho)
> plot(x[,1],x[,2],pch=19,cex=.05,xlab="Rate of return",ylab="Rate of inflation")
```

What would this cloud look like if the rates were not correlated?

9 Moment Matching

1. You assumed a normal distribution for the earlier above ground biomass problem. Redo the problem with a more appropriate distribution.

When we say *support*, we are referring to the values of a random variable for which probability density or probability exceed 0 and are defined. The support of lognormal distribution is continuous and strictly non-negative, which makes it particularly useful in ecology. Moreover, it is often useful because it is asymmetric, allowing for values that are extreme in the positive direction. Finally, it is useful for representing products of random variables. The central limit theorem would predict that the distribution of sums of random variables will be normal, no matter how each is individually distributed. The products of random variables will be lognormally distributed regardless of their individual distributions.

If a random variable is lognormally distributed then the log of that random variable is normally distributed (conversely, if you exponentiate a normal random variable it generates a lognormal random variable). The first parameter of the lognormal distribution is the mean of the random variable on the log scale (i.e., α on cheat sheet, meanlog in R) and the second parameter is the variance (or sometimes the standard deviation) of the random variable on the log scale (i.e., β on cheatsheet, sdlog in R). We often predict the median of the distribution with our deterministic model, which means that we can use the log of the median as the first parameter because

$$\begin{aligned} z &\sim \text{lognormal}(\alpha, \beta) \\ \text{median}(z) &= e^{\alpha} \\ \log(\text{median})(z) &= \alpha \end{aligned}$$

2. Simulate 10,000 data points from a normal distribution with mean 0 and standard deviation 1 and another 10,000 data points from a log normal distribution with first parameter (the mean of the random variable on the log scale) = 0 and second parameter (the standard deviation of the parameter on the log scale) = 1. Display side-by-side histograms scaled to the density. Find the mean and variance of the lognormal distribution using moment matching. Check your moment-matched values empirically with the simulated data. The moment-matched values and the empirical values are close for the mean, but less so for the variance. Why? What happens when you increase the number or draws? Explore the two distributions by repeating with different means and standard deviations of your choice.

3. We are interested in the proportion (ϕ) of patches within landscapes that are occupied by a rare plant. Existing literature shows that that this proportion has a mean of $\mu = 0.04$ with a standard deviation of $\sigma = .01$. Write out a model for the distribution of ϕ, conditional on μ and σ. The challenge here is to use moment matching for a random variable with support on 0-1.

   (a) Plot the probability distribution of ϕ.

   (b) If you visited 50 patches, what is the probability that 5 would be occupied, conditional on the hypothesis the $\phi = 0.04$?

   (c) Plot the probability of the data for $y = 1, \ldots, 10$ occupied patches out of 50 patches visited conditional on the hypothesis $\phi = 0.04$.

   (d) What is the probability that at least 5 are occupied, conditional on the hypothesis that $\phi = 0.04$?

   (e) Plot the cumulative probability that at least $y = 1, \ldots,$ patches are occupied, conditional on the hypothesis that $\phi = 0.04$?

   (f) Simulate data for 75 patches (empty patch = 1, occupied patch = 0).

4. You are modeling the relationship between plant growth rate and soil water. Represent plant growth (μ_i) as linear function of soil water, $\mu_i = \beta_0 + \beta_1 x_i$. Write out the model for the data. Simulate a data set of 20, strictly non-negative pairs of y and x values. Assume that:

   - Soil water, the x values, varies randomly and uniformly between 0.01 and 0.2
   - $\beta_0 = 0.01$ and $\beta_1 = 0.09$
   - The standard deviation of the model prediction is 0.03. That is, $[y|f(\mu, \sigma = 0.03)]$

Plot the data and overlay the generating model.

5. The Poisson distribution is often used for count data, despite the fact that one must assume that the mean and variance are equal. The negative binomial distribution is a more robust alternative, allowing the variance to differ from the mean. There are two parameterizations for the negative binomial. The first is more frequently used by ecologists:

$$[z|\lambda, r] = \frac{\Gamma(z+r)}{\Gamma(r)z!}\left(\frac{r}{r+\lambda}\right)^{r}\left(\frac{\lambda}{r+\lambda}^{z}\right),$$

where z is a discrete random variable, λ is the mean of the distribution and r is the dispersion parameter. The variance of z equals $\lambda + \frac{\lambda^2}{r}$.

The second parameterization of the negative binomial distribution (and most common) is used to model the number of failures that occur in a sequence of Bernoulli trials, before r successes are obtained *e.g., how many times do you loose at Black Jack, before you win.* This parameterization is more often implemented in coding environments (i.e., JAGS):

$$[z|\phi, r] = \frac{\Gamma(z+r)}{\Gamma(r)z!}\phi^{r}(1-\phi)^{z}$$

where z is the discrete random variable representing the number of failures that occur before r successes are obtained. The parameter ϕ is the probability of success on a given trial. Note that $\phi = \frac{r}{\lambda+r}$.

(a) Simulate 100,000 observations from a negative binomial distribution with mean of 100 and variance of 400 using the **first** parameterization that has a mean and a dispersion parameter (remember to moment match).

(b) Do the same simulation using the **second** parameterization.

(c) Plot side-by-side histograms of the simulated data.