Multivariate Distributions and Gibbs Sampling

Ivan Yi-Fan Chen

2023 Fall

1 Review of Joint, Marginal, and Conditional Distributions

- Start with a simple superficial example: Dining Choice v.s. Gender
  - You ask around the campus of NUK about what students are going to have for lunch later on.
  - Can have rice, noodle or burger.
  - Students are either male or female, physically.
  - According to their answers, you have the following table.

	Male	Female	Dining Choice
Rice	0.25	0.15	0.40
Noodle	0.10	0.35	0.45
Burger	0.10	0.05	0.15
Gender of Students	0.45	0.55	1.00

- Marginal probability:
  - The probability that the student is male / female, i.e., $\Pr(Gender = Male)$ and $\Pr(Gender = Female)$.
  - The probability that the student is going to have rice / noodle / burger for lunch, i.e., $\Pr(DinningChoice = Rice)$, $\Pr(DinningChoice = Noodle)$, and $\Pr(DinningChoice = Burger)$.
- Joint probability:
  - The probability that a specific choice of lunch **and** a specific gender happen at the same time, $\Pr(Gender, DinningChoice)$
  - The probability that a **male** student chooses to have **noodle** is *0.1*, i.e., $\Pr(Gender = Male, DinningChoice = Noodle) = 0.1$
  - The probability that a **female** student chooses to have **rice** is *0.15*, i.e., $\Pr(Gender = Female, DinningChoice = Rice) = 0.15$
- Conditional probability:
  - The probability over the student's dinning choice **given** his/her gender.
  - The probability of the student's gender **given** his/her choice of lunch.
- So what is the probability that a female student is going to have noodle for lunch?
  - This is exactly a conditional probability:

$$\Pr(Choice = Noodle|Gender = Female) = \frac{\Pr(Choice = Noodle, Gender = Female)}{\Pr(Gender = Female)} = \frac{0.35}{0.55} = \frac{7}{11}$$

  - In fact,

$$\Pr(Choice = Rice|Gender = Female) = \frac{\Pr(Choice = Rice, Gender = Female)}{\Pr(Gender = Female)} = \frac{0.15}{0.55} = \frac{3}{11}$$

$$\Pr(Choice = Burger|Gender = Female) = \frac{\Pr(Choice = Burger, Gender = Female)}{\Pr(Gender = Female)} = \frac{0.05}{0.55} = \frac{1}{11}$$

$$\begin{aligned}\Pr(Choice = Noodle|Gender = Female) + \Pr(Choice = Rice|Gender = Female)\\ + \Pr(Choice = Burger|Gender = Female) \quad &= 1\end{aligned}$$

- Similarly, the probability that the student is male if the student have rice is also a conditional probability:

$$\Pr(Gender = Male|DinningChoice = Rice) = \frac{\Pr(Gender = Male, DinningChoice = Rice)}{\Pr(DinningChoice = Rice)} = \frac{0.25}{0.4} = \frac{5}{8}$$

$$\Pr(Gender = Male|DinningChoice = Rice) + \Pr(Gender = Female|DinningChoice = Rice) = 1$$

Let’s restrict ourselves to discrete random variables X and Y.

- **Marginal Probability:** $\Pr(X = x)$ and $\Pr(Y = y)$.
- **Joint Probability:** $\Pr(X = x, Y = y)$, i.e., the events $X = x$ and $Y = y$ happen simultaneously.
    - Marginal probability of an event is obtained by *exhaustively* summing across joint probabilities with respect to other events, i.e.,

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y)$$

$$\Pr(Y = y) = \sum_x \Pr(X = x, Y = y)$$

- **Conditional Probability:** $\Pr(X = x|Y = y)$ and $\Pr(Y = y|X = x)$ which are respectively computed as

$$\Pr(X = x|Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} = \frac{\Pr(X = x, Y = y)}{\sum_x \Pr(X = x, Y = y)}$$

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{\Pr(X = x, Y = y)}{\sum_y \Pr(X = x, Y = y)}$$

Now we consider *continuous* random variables X and Y.

- Denote the *joint* PDF by $f_{X,Y}(x, y)$, and denote the domain for X (Y) by $\mathbf{X}$ ($\mathbf{Y}$).
- **Marginal Probability Density:**

$$f_X(x) = \int_{y \in \mathbf{Y}} f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int_{x \in \mathbf{X}} f_{X,Y}(x, y) dx$$

- **Conditional Density:**

$$f_X(x|Y = y) = \frac{f_{X,Y}(x, Y = y)}{f_Y(Y = y)} = \frac{f_{X,Y}(x, Y = y)}{\int_{x \in \mathbf{X}} f_{X,Y}(x, Y = y) dx}$$

$$f_Y(y|X = x) = \frac{f_{X,Y}(X = x, y)}{f_X(X = x)} = \frac{f_{X,Y}(X = x, y)}{\int_{y \in \mathbf{Y}} f_{X,Y}(X = x, y) dy}$$

- All are quite similar to the discrete case. But to really get the (cumulated) probability we need to perform yet another integration.
    - Example 1: The CDF of the marginal distribution of X is

$$F_X(\overline{x}) = \int_{\underline{x}}^{\overline{x}} \int_{y \in \mathbf{Y}} f_{X,Y}(x, y)\, dy dx$$

  - Example 2: The CDF of the distribution of X *conditional on* $Y = y$ is

$$F_X\left(\overline{x}|Y=y\right) = \int_{\underline{x}}^{\overline{x}} f_X\left(x|Y=y\right)dx$$

  - Example 3: Probability “sum” up to 1

$$1 = \int_{x\in\mathbf{X}}\int_{y\in\mathbf{Y}} f_{X,Y}\,(x,y)dydx$$

2 Bivariate Normal Distribution

- Univariate Normal Distribution: The pdf is given by

$$f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  where mu is the mean and $\sigma > 0$ is the standard deviation.

  - Usually we denote a Normal random variable by $X \sim N(\mu, \sigma^2)$.
  - Linear Combination Property: If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
  - The Linear Combination Property implies that we can represent $X \sim N(\mu, \sigma^2)$ by letting $X = \sigma Z + \mu$ where $Z \sim N(0, 1)$ is Standard Normal.

- Bivariate Normal Distribution: a normal distribution that involves in two potentially correlated random variables X and Y.
- PDF of bivariate normal:

$$f_{X,Y}\left(x,y\right) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}e^{-\frac{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}{2\left(1-\rho^2\right)}}$$

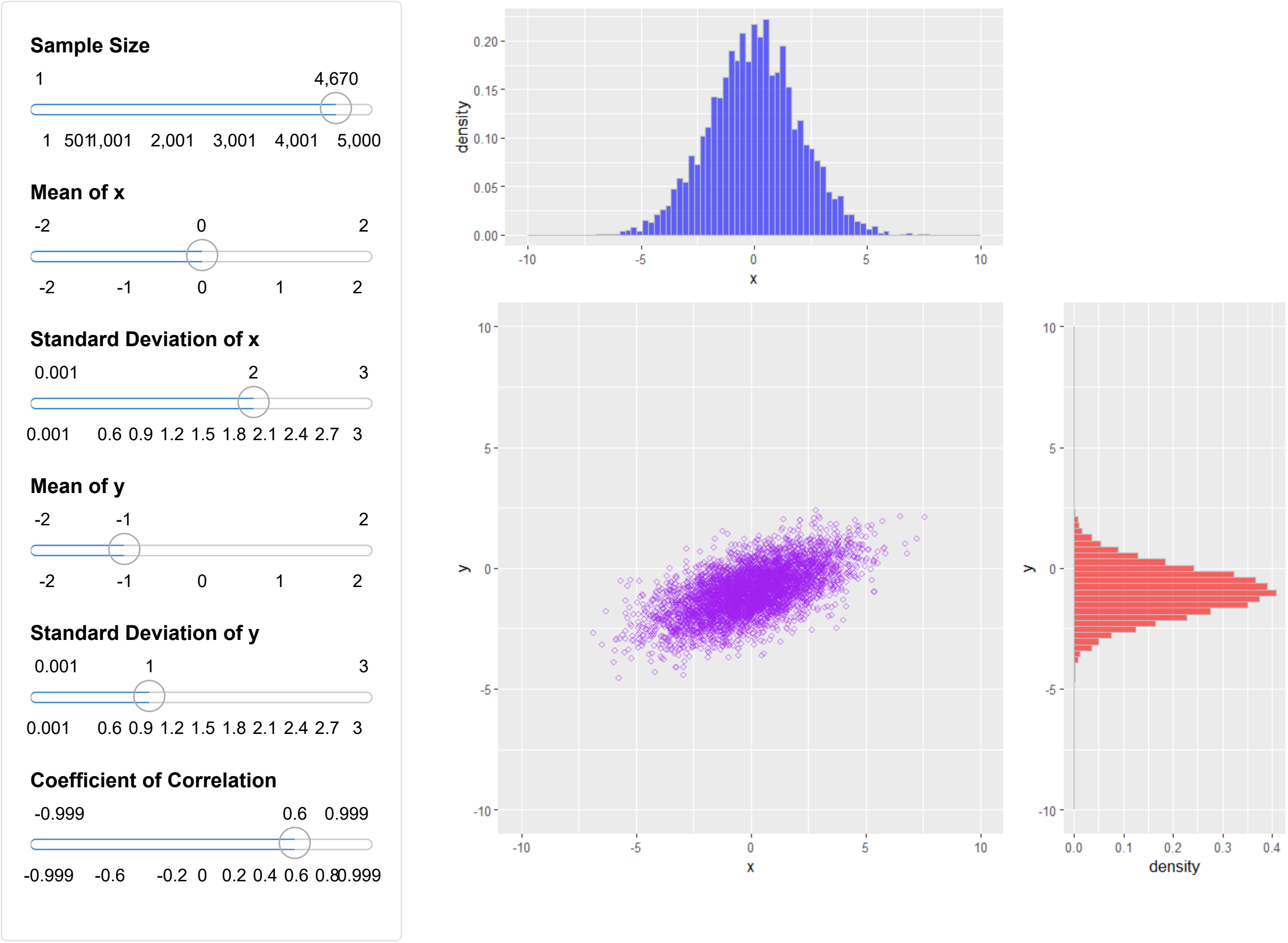
  where μ_X, μ_Y, σ_X and σ_Y are similar to univariate normal, and $\rho \in (-1, 1)$ is the coefficient of correlation between X and Y.

  - In particular, $\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$ where σ_{XY} is the covariance between the random variables.

- The **marginal** distributions of X and Y are respectively normal, i.e., $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$.
- The **conditional** distributions are

$$(X|Y = y) \sim N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_y), \sigma_X^2(1 - \rho^2))$$

$$(Y|X = x) \sim N(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_Y), \sigma_Y^2(1 - \rho^2))$$



Sample Size

1 4,670

Mean of x

-2 0 2

Standard Deviation of x

0.001 2 3

Mean of y

-2 -1 2

Standard Deviation of y

0.001 1 3

Coefficient of Correlation

-0.999 0.6 0.999

How to sample a Bivariate Normal Distribution?

- Recall that in a univariate environment with $Z \sim N(0,1)$, we have $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$. An analogous version of this property also holds in a bivariate environment.
- Consider X and Y following a Bivariate Normal Distribution

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho \\ \sigma_X \sigma_Y \rho & \sigma_Y^2 \end{bmatrix} \right)$$

- Let $\Sigma \equiv \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho \\ \sigma_X \sigma_Y \rho & \sigma_Y^2 \end{bmatrix}$ denote the variance-covariance matrix of this bivariate distribution, it is easily checked that

$$\Sigma = \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix}$$

- Conceptually we can think of Σ as the matrix version of variance, hence the “square-root” of Σ, can be think of as the standard deviation. Namely,

$$\Sigma = \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} CC^T \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix}$$

  where the *lower triangle matrix* C and its transpose matrix C^T are such that

$$CC^T = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

  Then

$$S \equiv \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} C$$

  is the “square-root of Σ” we want.

- By Cholesky Decomposition, we have

$$C = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}$$

- Theorem: Let $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ and are independent. We have

$$\begin{bmatrix} X \\ Y \end{bmatrix} \equiv \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_X\sigma_Y\rho \\ \sigma_X\sigma_Y\rho & \sigma_Y^2 \end{bmatrix} \right)$$

Procedure to draw from Bivariate Normal:

1. Independently draw two vectors of equal length N from standard normal $N(0,1)$. Let $\mathbf{z}_1|_{1\times N}$ and $\mathbf{z}_2|_{1\times N}$ be the vectors we drawn, and bind them into a 2-by-N matrix as

$$\mathbf{z}_{2\times N} \equiv \begin{bmatrix} \mathbf{z}_1|_{1\times N} \\ \mathbf{z}_2|_{1\times N} \end{bmatrix} \equiv \begin{bmatrix} z_{1,1} & z_{1,2} & \dots & z_{1,N} \\ z_{2,1} & z_{2,2} & \dots & z_{2,N} \end{bmatrix}$$

2. Pick μ_X, μ_Y, σ_Y, σ_Y, and ρ to your need. Then do the following matrix algebra in R

$$\begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \mathbf{z}_{2\times N} + \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

   You will get a 2-by-N matrix, with the first row being the sample of X and the second row being the sample of Y. Therefore, each column represents an observation (x, y) and in total the sample size is N.

2.1 In-class Exercise

1. Draw a sample of size 5000 from a bivariate normal distribution. You can choose whatever means, variances and coefficient of correlation.
2. Compute the means, variances, and coefficient of correlation from the sample you just have drawn. Compare with the parameters you provided.

3 Gibbs Sampling

- Sampling from Bivariate, and even Multivariate Normal is simple owing to the Linear Combination Property.
  - For the bivariate case, we have a closed-form solution to the Cholesky Decomposition.
  - For general multivariate cases, we simply construct the correlation matrix with diagonal elements being 1, and the off-diagonal elements being coefficients of correlation for each variable pair. Then we get C by performing a

Cholesky Decomposition with it using `chol()` and then take a transpose using `t()`.

- What if the multivariate distribution is NOT normal? Need other ways around. **Gibbs Sampling** is a possible answer.

- **Requirement:**

  1. Need to know the **conditional densities**. In a bivariate case, $f_X(x|Y=y)$ and $f_Y(y|X=x)$.
  2. The underlying distribution needs to be easy enough to work with.

- **Procedure:**

  Step 1. Pick an arbitrary initial value x_0 and y_0.

  Step 2. Draw y_1 conditional on x_0 from the conditional density $f_Y(y|X=x_0)$.

  Step 3. Draw x_1 conditional on y_1 from the conditional density $f_X(x|Y=y_1)$.

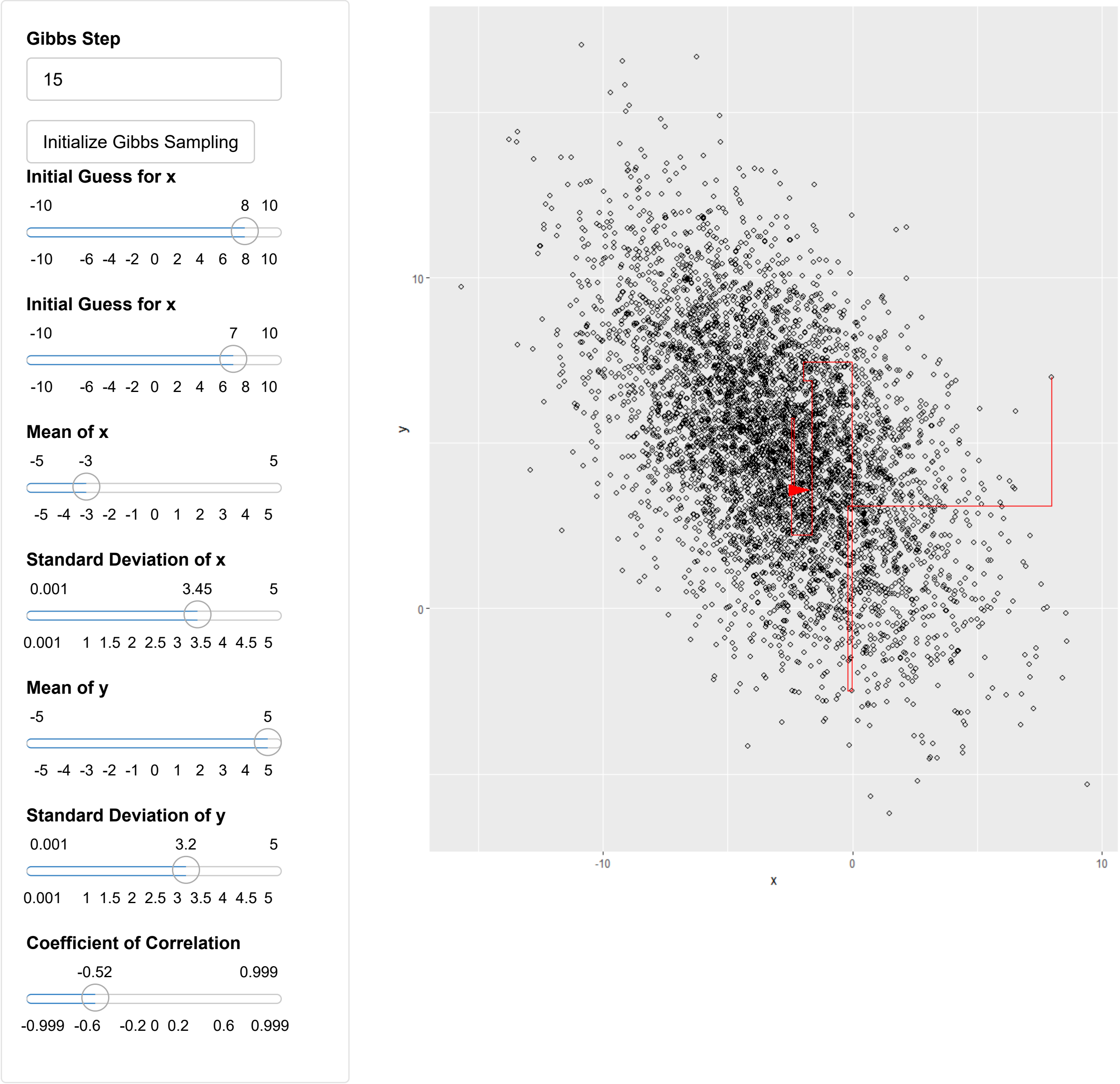
  Step 4. Use x_1 and y_1 to repeat Steps 2 and 3 to obtain x_2 and y_2, so on and so forth until we draw enough of observations (x, y).

- Put it differently, in this procedure we first move along the y-axis, and then along the x-axis, and then repeat. The steps are thus

$$(x_0, y_0) => (x_0, y_1) => (x_1, y_1) => (x_1, y_2) => (x_2, y_2) =>\dots$$

  and eventually we take $\{(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots\}$ as the output.

Let’s see it visually. Each point represents an observation we keep. The intermediate observations are not shown, but can be seen from the path we travel.



Gibbs Step

15

Initialize Gibbs Sampling

Initial Guess for x: 8

Initial Guess for x: 7

Mean of x: -3

Standard Deviation of x: 3.45

Mean of y: 5

Standard Deviation of y: 3.2

Coefficient of Correlation: -0.52

There are several restrictions when using Gibb’s Sampling.

1. We take an initial *guess* which we thought to be the point where most of the density lies. But obviously good guesses don’t happen everyday. The procedure involves in a lot of *traveling*, and you are likely to travel through unimportant regions. This traveling is prolonged if we start with a bad guess. In a more jargon way, this procedure is *auto-correlated* (the current point reached completely depends on the previous starting point as we have seen from the interactive figure). Such an auto-correlation goes away only when we have performed a good amount of drawn.

   The moral is, we should **not** use the observations obtained in the early steps of the procedure. For example, if we draw with Gibbs Sampling for 10^6 observations, we may want to drop the very first 6×10^3 observations. This is so-called **burn-in**. But how much to drop, and how much to draw? It is an art……

2. Gibbs Sampling can be trapped / fail to reach certain regions. Consider the two examples.

Example 1: The joint distribution of (x, y) is defined on the rectangles $[1, 2] \times [1, 2]$ and $[-0.5, 0] \times [-0.5, 0]$. We can think of the density to be falling on two unconnected "islands". If we initiate a guess within, for example, $[1, 2] \times [1, 2]$, we will never be able to get to the other island.

Example 2: The joint distribution of (x, y) is defined on a rectangle $(0, 10) \times (0, 10)$ and at a point $(0, 0)$. Suppose further that the distribution is atomistic such that the point $(0, 0)$ happens by probability 0.9, and the rest of the probability is uniformly distributed on the rectangle. Gibbs Sampling is very likely to get stuck at point $(0, 0)$. This two-dimensional case might not cause real problem as long as we perform a lot of draws. But in a high-dimensional case, say a 100 variable version of this distribution, Gibbs Sampling will certainly take astronomically many draws to really get a meaningful sample!

4 Assignment

1. Perform Gibbs Sampling to draw a sample of size 100000 from a Bivariate Normal Distribution. You are free to set μ_x, μ_y, σ_x, σ_y, and ρ.

   You will need to use `for` loop to make the draw step-by-step. You may want to create a matrix / data.frame with 100000 rows, and replace the corresponding elements with your draw during the loop.

2. Compute ρ with the sample you have drawn and compare with your setting. Specifically, we want to find the *optimal burning-in* such that $diff \equiv |\rho_{computed} - \rho_{setting}| < 10^{-6}$. You will need to setup a `while` loop that iterates up to 100000 times. In the first iteration you compute ρ with the full sample, and in the second iteration you do the same with the first observation dropped, and so on. The iteration stops either when you run out of iteration, or when $diff$ becomes small enough upon n-th iteration. This n is the number you want.