Problem Set 1

MAS 622J/1.126J: Pattern Recognition and Analysis

Due: 5:00 p.m. on September 20

[Note: All instructions to plot data or write a program should be carried out using Matlab. In order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.]

If you collaborated with other members of the class, please write their names at the end of the assignment. Moreover, you will need to write and sign the following statement: "In preparing my solutions, I did not look at any old homeworks, copy anybody's answers or let them copy mine."

Problem 1: Why? [5 points]

Limit your answer to this problem to a page.

a. Describe an application of pattern recognition related to your research. What are the features? What is the decision to be made? Speculate on how one might solve the problem.

b. In the same way, describe an application of pattern recognition you would be interested in pursuing for fun in your life outside of work.

Solution: Refer to examples discussed in lecture.

Problem 2: Probability Warm-Up [20 points]

Let x and y be discrete random variables, and a and b are constant values. Let μ_x denote the expected value of x and σ_x^2 denote the variance of x. Use excruciating detail to answer the following:

a. Show $\mathrm{E}[ax + by] = a\mathrm{E}[x] + b\mathrm{E}[y]$.

b. Show $\sigma_x^2 = \mathrm{E}[x^2] - \mu_x^2$.

c. Show that independent implies uncorrelated.

d. Show that uncorrelated does not imply independent.

e. Let $z = ax + by$. Show that if x and y are uncorrelated, then $\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2$.

f. Let x_i $(i = 1, ..., n)$ be random variables independently drawn from the same probability distribution with mean μ_x and variance σ_x^2. For the sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, show the following: (i) $\mathrm{E}[\overline{x}] = \mu_x$. (ii) $\mathrm{Var}[\overline{x}]$ (variance of the sample mean) $= \sigma_x^2/n$. Note that this is different from the sample variance $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

g. Let x_1 and x_2 be independent and identically distributed (i.i.d) continuous random variables. Can $\Pr[x_1 \leq x_2]$ be calculated? If so, find its value. If not, explain. Hint 1: Remember that for a continuous variable $\Pr[x_1 = k] = 0$, for any value of k. Hint 2: Remember the definition of i.i.d. variables.

h. Let x_1 and x_2 be independent and identically distributed discrete random variables. Can $\Pr[x_1 \leq x_2]$ be calculated? If so, find its value. If not, explain.

Solution:

a. The following is for continuous random variables. A similar argument holds for continuous random variables.

$$\begin{aligned}
\mathrm{E}[ax + by] &= \sum_{x\in X, y\in Y} (ax + by)\, p(x, y) \\
&= a \sum_{x\in X, y\in Y} x\, p(x, y) + b \sum_{x\in X, y\in Y} y\, p(x, y) \\
&= a \sum_{x\in X} x\, p(x) + b \sum_{y\in Y} y\, p(y) \\
&= a\mathrm{E}[x] + b\mathrm{E}[y]
\end{aligned}$$

b. Making use of the definition of variance and the previous part, we have:

$$\begin{aligned}
\sigma_x^2 &= \mathrm{E}[(x - \mu_x)^2] \\
&= \mathrm{E}[x^2 - 2\mu_x x + \mu_x^2] \\
&= \mathrm{E}[x^2] - \mathrm{E}[2\mu_x x] + \mathrm{E}[\mu_x^2] \\
&= \mathrm{E}[x^2] - 2\mu_x \mathrm{E}[x] + \mu_x^2 \\
&= \mathrm{E}[x^2] - 2\mu_x\mu_x + \mu_x^2 \\
&= \mathrm{E}[x^2] - 2\mu_x^2 + \mu_x^2 \\
&= \mathrm{E}[x^2] - \mu_x^2
\end{aligned}$$

c. In order to check if two discrete random variables x and y are uncorrelated, we have to prove $\sigma_{xy} = 0$ (the same holds for continuous random variables.)

From the previous question:

$$\sigma_{x,y}^2 = \mathrm{E}[xy] - \mu_x\mu_y$$

If two variables are independent:

$$\begin{aligned}
\mathrm{E}[xy] &= \sum_{x\in X, y\in Y} xy\, p(x,y) \\
&= \sum_{x\in X, y\in Y} xy\, p(x)\, p(y) \\
&= \sum_{x\in X} x\, p(x) \sum_{y\in Y} y\, p(y) \\
&= \mathrm{E}[x]\, \mathrm{E}[y]
\end{aligned}$$

Finally,

$$\sigma^2_{x,y} = \mathrm{E}[x]\,\mathrm{E}[y] - \mu_x\mu_y = 0$$

d. To prove this, we need to find one case where 1) $p(x,y) \neq p(x)p(y)$ and 2) $\sigma_{xy} = 0$ are satisfied. One possible solution is as follows:

Suppose we have the discrete random variables x and y, and we observed all possibilities:

x	y
1	1
1	1
-1	1
-1	1
0	0
0	0

If we look at the case where $x = 1$ and $y = 1$, 1) is satisfied:

$$\begin{aligned}
p(x=1, y=1) &= \frac{2}{6} = \frac{1}{3} \\
p(x=1)p(y=1) &= \frac{2}{6}\frac{4}{6} = \frac{2}{9}
\end{aligned}$$

Now it is easy to verify that 2) is also satisfied:

$$\sigma^2_{x,y} = \mathrm{E}[xy] - \mu_x\mu_y = 0 - 0\frac{4}{6} = 0$$

e. Given that $z = ax + by$ and that x and y are uncorrelated, we have

$$\begin{aligned}
\sigma^2_z &= \mathrm{E}[(z-\mu_z)^2] \\
&= \mathrm{E}[z^2] - \mu_z^2 \\
&= \mathrm{E}[(ax+by)^2] - (a\mu_x + b\mu_y)^2 \\
&= \mathrm{E}[a^2x^2 + 2abxy + b^2y^2] - (a^2\mu_x^2 + 2ab\mu_x\mu_y + b^2\mu_y^2)
\end{aligned}$$

$$
\begin{aligned}
&= a^2\mathrm{E}[x^2] + 2ab\mathrm{E}[xy] + b^2\mathrm{E}[y^2] - a^2\mu_x^2 - 2ab\mu_x\mu_y - b^2\mu_y^2 \\
&= a^2(\mathrm{E}[x^2] - \mu_x^2) + 2ab(\mathrm{E}[xy] - \mu_x\mu_y) + b^2(\mathrm{E}[y^2] - \mu_y^2) \\
&= a^2\sigma_x^2 + 2ab\sigma_{xy}^2 + b^2\sigma_y^2 \\
&= a^2\sigma_x^2 + b^2\sigma_y^2,
\end{aligned}
$$

where only the last equality depends on x and y being uncorrelated.

f. Using the result of (a) and the fact that $\mathrm{E}[x_i] = \mu_x$,

$$\mathrm{E}[\bar{x}] = \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[x_i] = \frac{1}{n}\; n\; \mu_x = \mu_x$$

Also, using the result of (d) and the fact $\mathrm{Var}[x_i] = \sigma_x^2$

$$\mathrm{Var}[\bar{x}] = \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i] = \frac{1}{n^2}\; n\; \sigma_x^2 = \sigma_x^2/n$$

g. Given that x_1 and x_2 are continuous random variables, we know that $\Pr[x_1 = k] = 0$ and $\Pr[x_2 = k] = 0$ for any value of k. Thus,

$$\Pr[x_1 \leq x_2] = \Pr[x_1 < x_2].$$

Given that x_1 and x_2 are i.i.d., we know that replacing x_1 with x_2 and x_2 with x_1 will have no effect on the world. In particular, we know that

$$\Pr[x_1 < x_2] = \Pr[x_2 < x_1].$$

However, since probabilities or the space of possible values must sum to one, we have

$$\Pr[x_1 < x_2] + \Pr[x_2 < x_1] = 1.$$

Thus,

$$\Pr[x_1 \leq x_2] = \frac{1}{2}.$$

h. For discrete random variables, unlike the continuous case above, we need to know the distributions of x_1 and x_2 in order to find $\Pr[x_1 = k]$ and $\Pr[x_2 = k]$. Thus, the argument we used above fails. In general, it is not possible to find $\Pr[x_1 \leq x_2]$ without knowledge of the distributions of both x_1 and x_2.

Problem 3: Teatime with Gauss and Bayes [20 points]

Let $p(x, y) = \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{(y-\mu)^2}{2\alpha^2} + \frac{(x-y)^2}{2\beta^2}\right)}$

a. Find $p(x)$, $p(y)$, $p(x|y)$, and $p(y|x)$. In addition, give a brief description of each of these distributions.

b. Let $\mu = 0$, $\alpha = 15$, and $\beta = 3$. Plot $p(y)$ and $p(y|x = 9)$ for a reasonable range of y. What is the difference between these two distributions?

Solution:

a. To find $p(y)$, simply factor $p(x, y)$ and then integrate over x:

$$\begin{aligned}
p(y) &= \int_{-\infty}^{\infty} p(x, y)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{(y-\mu)^2}{2\alpha^2} + \frac{(x-y)^2}{2\beta^2}\right)}\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\frac{(y-\mu)^2}{2\alpha^2}}\, e^{-\frac{(x-y)^2}{2\beta^2}}\, dx \\
&= \frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{(y-\mu)^2}{2\alpha^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(x-y)^2}{2\beta^2}}\, dx \\
&= \frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{(y-\mu)^2}{2\alpha^2}} \\
&= \mathcal{N}(\mu, \alpha^2)
\end{aligned}$$

The integral goes to 1 because it is of the form of a probability distribution integrated over the entire domain. To find $p(x|y)$, divide $p(x, y)$ by $p(y)$:

$$\begin{aligned}
p(x|y) &= \frac{p(x, y)}{p(y)} \\
&= \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(x-y)^2}{2\beta^2}} \\
&= \mathcal{N}(y, \beta^2)
\end{aligned}$$

Finding $p(x)$ and $p(y|x)$ follows essentially the same procedure, but the algebra is more involved and requires completing the square in the exponent.

$$\begin{aligned}
p(x) &= \int_{-\infty}^{\infty} p(x, y)\, dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{(y-\mu)^2}{2\alpha^2} + \frac{(x-y)^2}{2\beta^2}\right)}\, dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{\beta^2(y-\mu)^2 + \alpha^2(x-y)^2}{2\alpha^2\beta^2}\right)}\, dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{\beta^2 y^2 - 2\beta^2\mu y + \beta^2\mu^2 + \alpha^2 x^2 - 2\alpha^2 xy + \alpha^2 y^2}{2\alpha^2\beta^2}\right)}\, dy
\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{(\alpha^2+\beta^2)y^2-2(\alpha^2x+\beta^2\mu)y+(\beta^2\mu^2+\alpha^2x^2)}{2\alpha^2\beta^2}\right)} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{y^2-2\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}y+\frac{\beta^2\mu^2+\alpha^2x^2}{\alpha^2+\beta^2}}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{y^2-2\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}y+\left(\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2-\left(\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2+\frac{\beta^2\mu^2+\alpha^2x^2}{\alpha^2+\beta^2}}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{\left(y-\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2-\left(\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2+\frac{\beta^2\mu^2+\alpha^2x^2}{\alpha^2+\beta^2}}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{\left(y-\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} e^{-\left(\frac{\frac{\beta^2\mu^2+\alpha^2x^2}{\alpha^2+\beta^2}-\left(\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} dy \\
&= \frac{1}{2\pi\alpha\beta}\sqrt{2\pi\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}} e^{-\left(\frac{\frac{\beta^2\mu^2+\alpha^2x^2}{\alpha^2+\beta^2}-\left(\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}} e^{-\left(\frac{\left(y-\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} dy \\
(Note &: \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}} e^{-\left(\frac{\left(y-\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} dy = \int_{-\infty}^{\infty} p(y|x)\,dy) \\
&= \frac{1}{\sqrt{2\pi(\alpha^2+\beta^2)}} e^{-\left(\frac{\frac{\beta^2\mu^2+\alpha^2x^2}{\alpha^2+\beta^2}-\left(\frac{\alpha^2x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} \\
&= \frac{1}{\sqrt{2\pi(\alpha^2+\beta^2)}} e^{-\left(\frac{(\alpha^2+\beta^2)(\beta^2\mu^2+\alpha^2x^2)-(\alpha^2x+\beta^2\mu)^2}{2\alpha^2\beta^2(\alpha^2+\beta^2)}\right)} \\
&= \frac{1}{\sqrt{2\pi(\alpha^2+\beta^2)}} e^{-\left(\frac{\alpha^2\beta^2\mu^2+\alpha^4x^2+\beta^4\mu^2+\alpha^2\beta^2x^2-\alpha^4x^2-2\alpha^2\beta^2\mu x-\beta^4\mu^2}{2\alpha^2\beta^2(\alpha^2+\beta^2)}\right)} \\
&= \frac{1}{\sqrt{2\pi(\alpha^2+\beta^2)}} e^{-\left(\frac{\alpha^2\beta^2x^2-2\alpha^2\beta^2\mu x+\alpha^2\beta^2\mu^2}{2\alpha^2\beta^2(\alpha^2+\beta^2)}\right)} \\
&= \frac{1}{\sqrt{2\pi(\alpha^2+\beta^2)}} e^{-\left(\frac{(x-\mu)^2}{2(\alpha^2+\beta^2)}\right)}
\end{aligned}
$$

$$= \mathcal{N}(\mu, \alpha^2 + \beta^2)$$

To find $p(y|x)$ we simply divide $p(x,y)$ by $p(x)$. In finding $p(x)$, we already know the form of $p(y|x)$ (see the longest line in the derivation of $p(x)$ above):

$$\begin{aligned} p(y|x) &= \frac{p(x,y)}{p(x)} \\ &= \frac{1}{\sqrt{2\pi\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}} e^{-\left(\frac{\left(y-\frac{\alpha^2 x+\beta^2\mu}{\alpha^2+\beta^2}\right)^2}{2\frac{\alpha^2\beta^2}{\alpha^2+\beta^2}}\right)} \\ &= \mathcal{N}\left(\frac{\alpha^2 x + \beta^2\mu}{\alpha^2+\beta^2}, \frac{\alpha^2\beta^2}{\alpha^2+\beta^2}\right) \end{aligned}$$

Note that all the above distibutions are Gaussian.

b. The following Matlab code produced Figure 1:

```
close all;
clear all;

%Initialize variables
m = 0.0;
a = 15.0;
b = 3;
x = 9;

%Compute parameters
y = -100:1:100;
mean = ((a^2)*x + (b^2)*m)/(a^2 + b^2);
var = ((a*b)^2)/(a^2 + b^2);
p_y_given_x = (1.0/sqrt(2*pi*var))*exp(-((y-mean).^2)/(2*var));
var2 = a^2;
p_y = (1.0/sqrt(2*pi*var2))*exp(-((y-m).^2)/(2*var2));

%Show information
figure;
plot(y,p_y_given_x,'b');
hold on
plot(y,p_y,'r');
legend('p(y|x)', 'p(y)');
sy = size(y);
axis([y(1),y(sy(2)),0,0.2]);
xlabel('y');
```

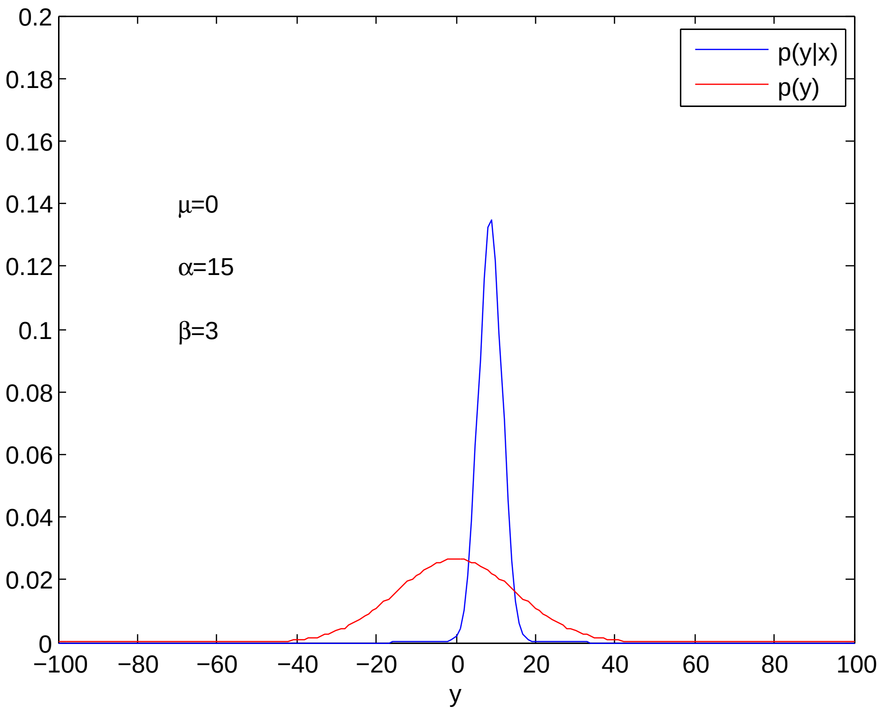

Figure 1: The marginal p.d.f. of y and the p.d.f. of y given x for a specific value of x. Notice how knowing x makes your knowledge of y more certain.

```
text(-70,0.14, '\mu=0');
text(-70,0.12, '\alpha=15');
text(-70,0.1, '\beta=3');
```

Problem 4: Covariance Matrix [15 points]

Let $\mathbf{\Sigma} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$

a. Find the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ by hand (include all calculations.) Verify your computations with MATLAB function *eig*.

b. Verify that $\mathbf{\Sigma}$ is a valid covariance matrix.

c. We provide 500 data points sampled from the distribution $\mathcal{N}([0\,0], \mathbf{\Sigma})$. Download the dataset from the course website and project the data onto the eigenvectors of the covariance matrix. What is the effect of this projection? Include MATLAB code and plots before and after the projection.

Solution:

a. We can find the eigenvectors and eigenvalues of $\mathbf{\Sigma}$ by starting with the definition of an eigenvector. Namely, a vector $\mathbf{e}$ is an eigenvector of $\mathbf{\Sigma}$ if it satisfies

$$\mathbf{\Sigma e} = \lambda \mathbf{e}$$

for some constant scalar λ, which is called the eigenvalue corresponding to $\mathbf{e}$. This can be rewritten as

$$(\mathbf{\Sigma} - \lambda I)\mathbf{e} = 0$$

This is equivalent to

$$\det(\mathbf{\Sigma} - \lambda I) = 0$$

Thus, we require that

$$(5 - \lambda)^2 - 4^2 = 0$$

By inspection, this is true when $\lambda = 9$ and $\lambda = 1$.

To find the eigenvectors, we plug the eigenvalues back into the equation above to get

$$(\mathbf{\Sigma} - 9I)\mathbf{e} = \begin{bmatrix} 5-9 & 4 \\ 4 & 5-9 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -4 & 4 \\ 4 & -4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which gives $a = b$. Normalized, this results in the eigenvector

$$\mathbf{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Similarly, $\lambda = 1$ gives

$$(\mathbf{\Sigma} - 1I)\mathbf{e} = \begin{bmatrix} 5-1 & 4 \\ 4 & 5-1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which gives $a = -b$. Normalized, this results in the eigenvector

$$\mathbf{e}_1 = \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

b. The matrix $\mathbf{\Sigma}$ is a valid covariance matrix if it is symmetric and positive semi-definite. Clearly, it is symmetric, since $\mathbf{\Sigma}^T = \mathbf{\Sigma}$. One way to prove it is positive semi-definite is to show that all its eigenvalues are non-negative. This is indeed the case, as shown in the previous question.

c. Figure 2 shows how projecting the data onto the eigenvectors of the data's covariance matrix reduces the correlation between x and y. The MATLAB code is as follows:

```
clear all;
close all;

%Initialize covariance matrix
S = [5 4; 4 5];

%Load data
load('points');
%Show correlation of original points
corrcoef(X)

%Compute eigenvectors and eigenvalues
[V D] = eig(S);
D
V

%Project data
pX = (V * X')';

%Show correlation of projected points
corrcoef(pX)

%Showing data
figure;
subplot(121)
scatter(X(:,1),X(:,2),'filled');
axis equal
xlabel('x');
ylabel('y');
title('Original Data');

subplot(122);
scatter(pX(:,1),pX(:,2),'filled');
axis equal
xlabel('x');
ylabel('y');
title('Projected Data');
```

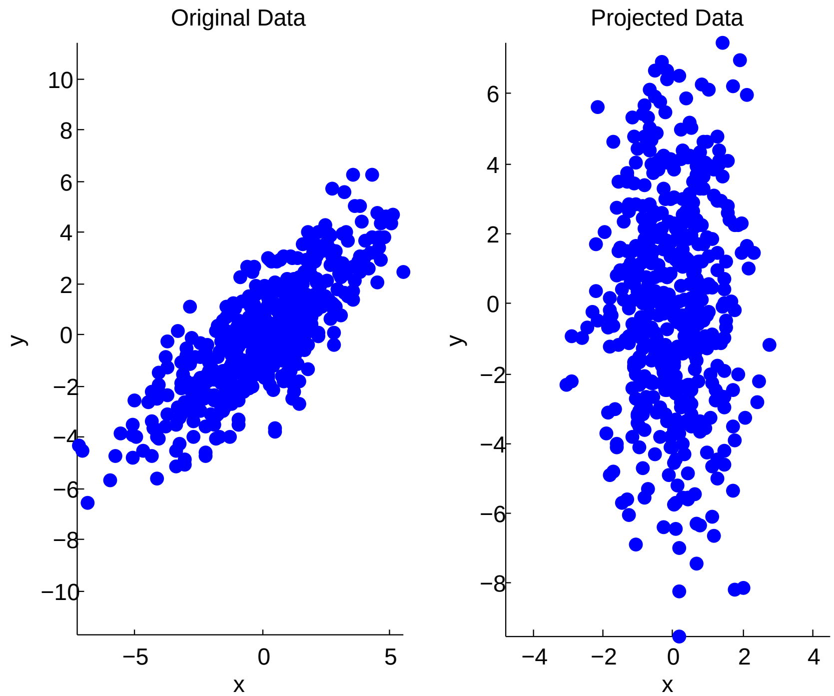

Figure 2: The original data and the data transformed into the coordinate system defined by the eigenvectors of their covariance matrix.

Problem 5: Probabilistic Modeling [20 points]

Let $x \in \{0, 1\}$ denote a person's affective state ($x = 0$ for "positive-feeling state", and $x = 1$ for "negative-feeling state"). The person feels positive with probability θ_1. Suppose that an affect-tagging system (or a robot) recognizes her feeling state and reports the observed state, y, to you. But this system is unreliable and obtains the correct result with probability θ_2.

a. Represent the joint probability distribution $P(x, y|\theta)$ for all x, y (a 2x2 matrix) as a function of the parameters $\theta = (\theta_1, \theta_2)$.

b. The Maximum Likelihood estimation criterion for the parameter θ is defined as:

$$\widehat{\theta}_{ML} = \arg\max_{\theta} L(t_1, ..., t_n; \theta) = \arg\max_{\theta} \prod_{i=1}^{n} p(t_i|\theta)$$

where we have assumed that each data point t_i is drawn independently from the same distribution so that the likelihood of the data is $L(t_1, ..., t_n; \theta) =$

$\prod_{i=1}^{n} p(t_i|\theta)$. Likelihood is viewed as a function of the parameters, which depends on the data. Since the above expression can be technically challenging, we maximize the log-likelihood $\log L(t_1, ..., t_n; \theta)$ instead of likelihood. Note that any monotonically increasing function (i.e., log function) of the likelihood has the same maxima. Thus,

$$\widehat{\theta}_{ML} = \arg\max_{\theta} \log L(t_1, ..., t_n; \theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(t_i|\theta)$$

Suppose we get the following joint observations $t = (x, y)$.

x	y
1	0
1	1
0	0
1	1
0	0
0	1
0	0
1	1

What are the maximum-likelihood (ML) values of θ_1 and θ_2? (*Hint.* Since $P(x, y|\theta) = P(y|x, \theta_2)P(x|\theta_1)$, the estimation of the two parameters can be done separately in the log-likelihood criterion.)

Solution:

a. The probability mass function (pmf) of $x \in \{0, 1\}$ is

$$P(x) = \left\{ \begin{array}{ll} \theta_1, & x = 0 \\ 1 - \theta_1, & x = 1 \end{array} \right\}$$

The conditional pmf of $y \in \{0, 1\}$ given that $x = 0$ is

$$P(y|x = 0) = \left\{ \begin{array}{ll} \theta_2, & y = 0 \\ 1 - \theta_2, & y = 1 \end{array} \right\}$$

The conditional pmf of y given that $x = 1$ is

$$P(y|x = 1) = \left\{ \begin{array}{ll} 1 - \theta_2, & y = 0 \\ \theta_2, & y = 1 \end{array} \right\}$$

Use $P(x, y) = P(y|x)P(x)$ to tabulate the joint pmf of (x, y).

$$P(x, y) = \begin{pmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{pmatrix} = \begin{pmatrix} \theta_2\theta_1 & (1-\theta_2)\theta_1 \\ (1-\theta_2)(1-\theta_1) & \theta_2(1-\theta_1) \end{pmatrix}$$

b. We select (θ_1, θ_2) to maximize the log-likelihood of the samples $\{(x_i, y_i), i = 1, ..., n\}$ which may be expressed as

$$\begin{aligned} J(\theta_1, \theta_2) &= \sum_i \log P(x_i, y_i) \\ &= \sum_i (\log P(y_i|x_i) + \log P(x_i)) \\ &= \left(\sum_i \log P(y_i|x_i)\right) + \left(\sum_i \log P(x_i)\right) \\ &= J_2(\theta_2) + J_1(\theta_1) \end{aligned}$$

Hence, we choose θ_1 to maximize

$$\begin{aligned} J_1(\theta_1) &= \sum_i \log P(x_i) \\ &= N(x=1)\log(1-\theta_1) + (n - N(x=1))\log\theta_1 \end{aligned}$$

where $N(x=1) = \sum_i x_i$. Differentiating w.r.t. θ_1 gives

$$\frac{\partial J_1}{\partial \theta_1} = \frac{-N(x=1)}{1-\theta_1} + \frac{n - N(x=1)}{\theta_1}$$

We set this derivative to zero and solve for θ_1 to obtain

$$\widehat{\theta}_1 = 1 - \frac{N(x=1)}{n}$$

Similarly, we choose θ_2 to maximize

$$\begin{aligned} J_2(\theta_2) &= \sum_i \log P(y_i|x_i) \\ &= N(x=y)\log\theta_2 + (n - N(x=y))\log(1-\theta_2) \end{aligned}$$

where $N(x=y) = \sum_i (x_i y_i + (1-x_i)(1-y_i))$. Differentiating J_2 w.r.t. θ_2, setting to zero and solving for θ_2 gives

$$\widehat{\theta}_2 = \frac{N(x=y)}{n}$$

For the example data, $\widehat{\theta}_1 = \frac{4}{8}$, $\widehat{\theta}_2 = \frac{6}{8}$. Thus,

$$\widehat{P}(x,y) = \begin{pmatrix} \widehat{\theta}_2\widehat{\theta}_1 & (1-\widehat{\theta}_2)\widehat{\theta}_1 \\ (1-\widehat{\theta}_2)(1-\widehat{\theta}_1) & \widehat{\theta}_2(1-\widehat{\theta}_1) \end{pmatrix}$$

The maximum likelihood of the data under this model is

$$\prod_i \widehat{P}(x_i, y_i) = \left(\frac{6}{8}\right)^6 \left(\frac{4}{8}\right)^8 \left(\frac{2}{8}\right)^2 \approx 4.34 \times 10^{-5}$$

Problem 6: Ring Problem [20 points]

To get credit for this problem, you must not only write your own correct solution, but also write a computer simulation of the process of playing this game:

Suppose I hide the ring of power in one of three identical boxes while you weren't looking. The other two boxes remain empty. After hiding he ring of power, I ask you to guess which box it's in. I know which box it's in and, after you've made your guess, I deliberately open the lid of an empty box, which is one of the two boxes you did not choose. Thus, the ring of power is either in the box you chose or the remaining closed box you did not choose. Once you have made your initial choice and I've revealed to you an empty box, I then give you the opportunity to change your mind – you can either stick with your original choice, or choose the unopened box. You get to keep the contents of whichever box you finally decide upon.

- What choice should you make in order to maximize your chances of receiving the ring of power? Justify your answer using Bayes' rule.
- Write a simulation. There are two choices in this game for the contestant in this game: (1) choice of box, (2) choice of whether or not to switch. In your simulation, first let the host choose a random box to place the ring of power. Show a trace of your program's output for a single game play, as well as a cumulative probability of winning for 1000 rounds of the two policies (1) to choose a random box and then switch and (2) to choose a random box and not switch.

Solution:

- Always switch your answer to the box you didn't choose the first time. This reason is as follows. You have a 1/3 chance of initially picking the correct box. That is, there is a 2/3 chance the correct answer is one of the other two boxes. Learning which of the two other boxes is empty does not change these probabilities; your initial choice still has a 1/3 chance of being correct. That is, there is a 2/3 chance the remaining box is the correct answer. Therefore you should change your choice.

  Using Bayes:

  R = Box with the Ring.
  O = Box Opened.
  S = Box Selected.

$$P(R|O,S) = \frac{P(O|R,S)P(R|S)}{P(O|S)} \tag{1}$$

  Where:

$$P(O|S) = \Sigma_{R=1}^{3} P(O,R|S) = \Sigma_{R=1}^{3} P(O|R,S)P(R|S) \tag{2}$$

$$P(R|O,S) = \frac{1 * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0 * \frac{1}{3}} = \frac{2}{3} \quad (3)$$

Another way to understand the problem is to extend it to 100 boxes, only one of which has the ring of power. After you make your initial choice, I then open 98 of the 99 remaining boxes and show you that they are empty. Clearly, with very high probability the ring of power resides in the one remaining box you did not initially choose.

- Here is a sample simulation output for the Ring problem:

```
actual:  1
guess1:  2
reveal:  3
swap  :  0
guess2:  2

actual:  3
guess1:  3
reveal:  1
swap  :  0
guess2:  3

actual:  2
guess1:  3
reveal:  1
swap  :  0
guess2:  3

swap           :  0
win            :  292
lose           :  708
win/(win+lose):  0.292

actual:  3
guess1:  1
reveal:  2
swap  :  1
guess2:  3

actual:  1
guess1:  1
reveal:  2
swap  :  1
guess2:  3
```

```
actual:  3
guess1:  2
reveal:  1
swap  :  1
guess2:  3

swap             : 1
win              : 686
lose             : 314
win/(win+lose):  0.686
```

Here is a Matlab program that simulates the Ring simulation output above:

```
for swap = 0:1
  win  = 0;
  lose = 0;
  for i = 1:1000
    actual = floor(rand()*3)+1;
    guess1 = floor(rand()*3)+1;
    if guess1 == actual
      reveal = floor(rand()*2)+1;
      if reveal == actual
        reveal = reveal + 1;
      end
    else
      if guess1 == 1 && actual == 2
        reveal = 3;
      elseif guess1 == 1 && actual == 3
        reveal = 2;
      elseif guess1 == 2 && actual == 1
        reveal = 3;
      elseif guess1 == 2 && actual == 3
        reveal = 1;
      elseif guess1 == 3 && actual == 1
        reveal = 2;
      elseif guess1 == 3 && actual == 2
        reveal = 1;
      end
    end
    if swap == 1
      if guess1 == 1 && reveal == 2
        guess2 = 3;
      elseif guess1 == 1 && reveal == 3
        guess2 = 2;
      elseif guess1 == 2 && reveal == 1
        guess2 = 3;
```

```
      elseif guess1 == 2 && reveal == 3
        guess2 = 1;
      elseif guess1 == 3 && reveal == 1
        guess2 = 2;
      elseif guess1 == 3 && reveal == 2
        guess2 = 1;
      end
    else
      guess2 = guess1;
    end
    if guess2 == actual
      win = win + 1;
    else
      lose = lose + 1;
    end
    %% only print trace for first 3 games
    if i <= 3
      actual
      guess1
      reveal
      swap
      guess2
    end
  end
  %% print results for each game play policy
  swap
  win / (win + lose)
end
```