Boston University

Department of Electrical and Computer Engineering

EC505 STOCHASTIC PROCESSES

~~Problem Set No. 1 Solutions~~

Fall 2016

Issued: Wednesday, Sept. 7, 2016 **Due:** Monday, Sept. 19, 2016

Problem 1.1

A random experiment consists of tossing a die and observing the number of dots on the top face. Let A_1 be the event that 3 comes up, A_2 the event that an even number comes up, and A_3 the event that an odd number comes up.

(a) Find $P(A_1), P(A_1 \cap A_3)$.

(b) Find $P(A_2 \cup A_3)$, $P(A_2 \cap A_3)$, $P(A_1|A_3)$.

(c) Are A_2 and A_3 disjoint?

(d) Are A_2 and A_3 independent?

Solution:

(a) $P(A_1) = 1/6, P(A_1 \cap A_3) = P(A_1) = 1/6$.

(b) $P(A_2 \cup A_3) = P(\Omega) = 1; P(A_2 \cap A_3) = P(\emptyset) = 0$, $P(A_1|A_3) = \dfrac{P(A_1 \cap A_3)}{P(A_3)} = 1/3$

(c) Yes.

(d) Clearly not. $P(A_2 \cap A_3) = P(\emptyset) = 0 \neq P(A_2)P(A_3)$.

Problem 1.2

A group of students is taking a multiple-choice test. For a particular question on the test, the fraction of students who know the answer is p. The fraction that will have to guess the answer is $(1 - p)$. If a student knows the answer, then he will certainly answer the question correctly. If a student doesn't know the answer and must guess, then the probability of answering the question correctly is $1/n$, where n is the number of choices for the given question.

(a) Compute the probability P_c that a student who answers the question correctly actually knew the answer.

(b) Suppose that the professor believes that $p = .85$, i.e. that 85% of the students actually know the answer. Further, suppose that he wants to design the multiple choice question such that $P_c = .95$, i.e. so that correct answers on the question indicate actual knowledge with 95% probability. How many parts n should the problem have?

Solution:

(a) The events K and C correspond to the events of student knowing the answer and anwering it correctly. Thus:

$$P_c = P(K|C)$$

And by Bayes' Rule:

$$P_c = \frac{P(C|K)P(K)}{P(C)} = \frac{P(C|K)P(K)}{P(C|K)P(K) + P(C|\bar{K})P(\bar{K})}$$

Subsitute $P(C|K) = 1$, $P(C|\bar{K}) = \frac{1}{n}$, $P(K) = p$, and $P(\bar{K}) = 1 - p$

$$P_c = \frac{p}{\frac{1-p}{n} + p}$$

(b) Solve for n in the above equation.

$$n = \frac{P_c(1-p)}{p(1-P_c)}$$

Then subsitute $P_c = .95$ and $p = .85$

$$n = \frac{.95(.15)}{.85(.05)} = 3.353$$

But it is best to have a question with an integer number of choices, so round n up to 4.

Problem 1.3

You are a contestant on a game show. There are three closed doors leading to three rooms. Two of the rooms contain nothing, but the third contains a prize of \$ 100000 which is yours if you pick the right door. You are asked to pick a door by the compere who knows which room contains the car. After you pick a door, the compere opens a door (not the one you picked) to show an empty room. Show that, even without any further knowledge, you will greatly increase your chances of winning the car if you switch your choice from the door you originally picked.

Solution:

I realized that the solution was online and much better than what I can say here. Here it is `http://en.wikipedia.org/wiki/Monty_Hall_problem`

Problem 1.4

A random variable x has probability distribution function

$$P_X(x) = [1 - e^{-2x}]u(x)$$

where $u(\cdot)$ is the unit-step function.

(a) Calculate the following probabilities:

$$P[X \leq 1], \qquad P[X \geq 2], \qquad P[X = 2].$$

(b) Find $p_X(x)$, the probability density function for X.

(c) Let Y be a random variable obtained from X as follows:

$$Y = \begin{cases} 0, & \text{if } X < 2 \\ 1, & \text{if } X \geq 2 \end{cases}$$

Find $p_Y(y)$, the probability density function for Y. Find $E(Y)$.

Solution:

(a) Note that $P(X \leq x) \equiv P_X(x)$. Further, since $P_X(x)$ is continuous $P(X < x) \equiv P_X(x)$ and $P[X = x] = 0$ for any x. Thus:

$$\begin{aligned} P[X \leq 1] &= P_X(1) = 1 - e^{-2} \\ P[X \geq 2] &= 1 - P[X < 2] = 1 - P_X(2) = 1 - (1 - e^{-4}) = e^{-4} \\ P[X = 2] &= 0 \end{aligned}$$

(b)

$$p_x(X) = \frac{dP_X(x)}{dx} = 2e^{-2x}u(x) + [1 - e^{-2x}]\delta(x) = 2e^{-2x}u(x)$$

(c) By construction Y is a discrete random variable that only takes on values 0 and 1. Thus

$$\begin{aligned} p_Y(y) &= [1 - P(Y=1)]\delta(y) + P(Y=1)\delta(y-1) \\ &= [1 - P(X \geq 2)]\delta(y) + P(X \geq 2)\delta(y-1) \\ &= [1 - e^{-4}]\delta(y) + e^{-4}\delta(y-1) \end{aligned}$$

using the results from part (a). To find $E(Y)$ we observe that Y is a discrete random variable and:

$$E(Y) = \sum_k kP(Y=k) = 1(P(Y=1)) + 0(P(Y=0)) = P(Y=1) = e^{-4}$$

Problem 1.5

Let X and Y be statistically independent random variables with probability density functions

$$p_X(x) = \frac{1}{2}\delta(x+1) + \frac{1}{2}\delta(x-1), \qquad p_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)$$

and let $Z = X + Y$, and $W = XY$.

(a) Find the conditional probability density functions $p_{Z|X}(z|x=-1)$ and $p_{Z|X}(z|x=1)$.

(b) Find the probability density function $p_Z(z)$ of Z.

(c) Find the mean values $m_X = E(X)$, $m_Y = E(Y)$, the variances, σ_Y^2, σ_W^2, and the covariance σ_{YW}. Are Y and W uncorrelated random variables? Are Y and W statistically independent random variables?

Solution:

(a)

$$\begin{aligned} p_{Z|X}(z \mid -1) &= p_{Y|X}(z - x \mid x = -1) = p_Y(z+1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z+1)^2}{2\sigma^2}\right] = \mathcal{N}(-1, \sigma^2) \\ p_{Z|X}(z \mid x = +1) &= p_{Y|X}(z - x \mid x = +1) = p_Y(z-1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z-1)^2}{2\sigma^2}\right] = \mathcal{N}(+1, \sigma^2) \end{aligned}$$

where we have used the fact that Y and X are independent so that $p_{Y|X}(y \mid x) = p_Y(y)$.

(b)

$$\begin{aligned} p_Z(z) &= \int_{-\infty}^{\infty} p_{Z,X}(z,x)\,dx \\ &= \int_{-\infty}^{\infty} p_{Z|X}(z|x) p_X(x)\,dx \\ &= p_{Z|X}(z|-1)\,P[X=-1] + p_{Z|X}(z|+1)\,P[X=+1] \\ &= p_{Z|X}(z|-1)\,\frac{1}{2} + p_{Z|X}(z|+1)\,\frac{1}{2} \\ &= \frac{1}{2\sigma\sqrt{2\pi}}\left\{\exp\left[-\frac{(z+1)^2}{2\sigma^2}\right] + \exp\left[-\frac{(z-1)^2}{2\sigma^2}\right]\right\} \end{aligned}$$

Note this is bimodal with two Gaussians centered at -1 and +1. Another (perhaps easier) way to get this result is to use the fact that the pdf of the sum of independent random variables is the convolution of their individual pdfs. Since in this case we get the convolution of two impulses with a Gaussian we again get two Gaussians centered at the impulses at $+1$ and -1.

(c)
- $E(X) = (-1)\frac{1}{2} + (1)\frac{1}{2} = 0$
- $E(Y) = 0$ by inspection, since its Gaussian.
- $\sigma_Y^2 = \sigma^2$ again by inspection since its Gaussian.
- $\sigma_W^2 = E[(W - m_W)^2] = E[W^2] - (E[W])^2$ Now:

$$E[W] = E[XY] = \underbrace{E[X]E[Y]}_{\text{by independence}} = 0$$

Thus

$$\sigma_W^2 = E[W^2] = E[X^2Y^2] = \underbrace{E[X^2]E[Y^2]}_{\text{by independence}}$$

$$\begin{aligned} E[X^2] &= \frac{1}{2}(-1)^2 + \frac{1}{2}(+1)^2 = 1 \\ E[Y^2] &= E[(Y-0)^2] = E[(Y-m_Y)^2] = \sigma_Y^2 = \sigma^2 \end{aligned}$$

so $\boxed{\sigma_W^2 = \sigma^2}$.

- $$\sigma_{YW} = E[(Y - m_Y)(W - m_W)] = E[YW] = E[XY^2] = E[X]E[Y^2] = 0\sigma^2 = 0$$
- From the preceding we see that Y and W are uncorrelated random variables.
- For Y and W to be statistically independent we must have that $p_{W|Y}(w|y) = p_W(w)$, i.e. that knowing Y does not affect W.

$$\begin{aligned} p_{W|Y}(w|y) &= p_{X|Y}\left(\frac{w}{y}\middle| y\right) = p_X\left(\frac{w}{y}\right) = \frac{1}{2}\left[\delta\left(\frac{w}{y}+1\right) + \delta\left(\frac{w}{y}-1\right)\right] \\ p_W(w) &= p_{W|X}(w|-1)P(X=-1) + p_{W|X}(w|+1)P(X=+1) \\ &= \frac{1}{2}\left[p_Y(-w) + p_Y(w)\right] \\ &= \left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{w^2}{2\sigma^2}\right)\right] \end{aligned}$$

Clearly $p_{W|Y}(w|y)$ and $p_W(w)$ are not equal.

Problem 1.6

The outcome of a random experiment is known to have an exponential distribution, but the parameter α of that distribution is not known. We estimate the parameter α from the sample mean:

$$\widehat{\alpha} = \frac{1}{\widehat{m}} \qquad \widehat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_i are independent trials of the experiment.

(a) Note that $\widehat{m}$ is itself a random variable. Find the mean and variance of $\widehat{m}$.

(b) Use the Chebychev bound to estimate the minimum number of experiments N that are required to guarantee that

$$P[|\widehat{m} - m| > .01m] \le .001$$

where $m = 1/\alpha$.

(c) Use the Central Limit Theorem to approximate the distribution of the random variable $\widehat{m} - m$ as a Gaussian random variable. Using the Gaussian approximation, estimate the minimum number of experiments N which would be required to guarantee that

$$P[|\widehat{m} - m| > .01m] \leq .001$$

Solution:

(a) The mean is given by:

$$E[\widehat{m}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{\alpha}$$

The variance is given by:

$$\sigma_{\widehat{m}}^2 = \text{Var}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}\text{Var}\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}[x_i] = \frac{1}{N}\frac{1}{\alpha^2}$$

where we have used the fact that $\text{Var}[cX] = c^2\text{Var}[X]$ and the variance of the sum of iid random variables is the sum of the variances of the random variables.

(b) The Chebychev bound states that for a random variable Z with mean m and variance σ^2:

$$\Pr[|Z - m_Z| \geq a] \leq \sigma^2/a^2$$

Now to apply this statement to $P[|\widehat{m} - m| > .01m]$ we would use $a = .01m$ and $\sigma = \sigma_{\widehat{m}}$ found above. This yields:

$$P[|\widehat{m} - m| > .01m] \leq \frac{\sigma_{\widehat{m}}^2}{.0001m^2} = \frac{\alpha^2}{.0001N\alpha^2} = \frac{1}{.0001N}$$

In order for $1/(.0001N) = .001$, we need $\boxed{N = 10^7}$.

(c) The Central Limit Theorem states that for any iid random variables x_i with mean m and variance σ^2 the quantity:

$$\frac{\sum_{i=1}^{N}(x_i - m)}{\sigma\sqrt{N}}$$

approaches a $N(0,1)$ random variable. Note carefully that this variance is the variance of each sample, *not* the variance of the estimator $\sigma_{\widehat{m}}^2$ which you found for part (a). Now we want to make a statement about the quantity:

$$(\widehat{m} - m) = \frac{1}{N}\sum_{i=1}^{N} x_i - m = \frac{1}{N}\sum_{i=1}^{N}(x_i - m) = \frac{\sigma\sqrt{N}}{N}\frac{\sum_{i=1}^{N}(x_i - m)}{\sigma\sqrt{N}} = \frac{\sigma}{\sqrt{N}}\underbrace{\left[\frac{\sum_{i=1}^{N}(x_i - m)}{\sigma\sqrt{N}}\right]}_{Z}$$

Now from the central limit theorem we know that the term in brackets in the last expression, i.e. Z, approaches a $N(0,1)$ random variable. Now with a bit of algebra we get:

$$\begin{aligned}
\Pr\{|\widehat{m} - m| > .01m\} &= \Pr\left\{\frac{\sigma}{\sqrt{N}}\left|\frac{\sum_{i=1}^{N}(x_i - m)}{\sigma\sqrt{N}}\right| > .01m\right\} = \Pr\left\{\left|\frac{\sum_{i=1}^{N}(x_i - m)}{\sigma\sqrt{N}}\right| > \frac{\sqrt{N}}{\sigma}.01m\right\} \\
&= \Pr\left\{\left|\frac{\sum_{i=1}^{N}(x_i - m)}{\sigma\sqrt{N}}\right| > .01\sqrt{N}\right\} \\
&= \Pr\left\{|Z| > .01\sqrt{N}\right\}
\end{aligned}$$

where $Z \sim N(0,1)$ and we have used the fact that for an exponential random variable $m = 1/\alpha$ and $\sigma = 1/\alpha$. For such a normalized Gaussian random variable Z with zero mean and unit variance:

$$Q(\beta) \equiv \Pr(Z > \beta) = \int_{\beta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{y^2/2}\, dy$$

which is tabulated in Appendix D of Shanmugan & Breipohl or in many other places. Thus in terms of the "Q" function, we want:

$$\Pr\left\{|\widehat{m} - m| > .01m\right\} \quad = \quad \Pr\left\{|Z| > .01\sqrt{N}\right\} = 2Q\left(.01\sqrt{N}\right) \leq .001$$

where because of the absolute value we multiply $Q(\cdot)$ by two. Now going to the tables we find $Q(3.3) = .001/2 = .0005$. Thus

$$\begin{aligned} .01\sqrt{N} &= 3.3 \\ N &= 108900 \end{aligned}$$

Notice that this number is much smaller than that found using the bound. This is becausing we are only using an approximation and not a bound. Often bounds lead to overly conservative answers.

Problem 1.7

In the mid to late 1980's, in response to the growing AIDS crisis and the emergence of new, highly sensitive tests for the virus, there were a number of calls for widespread public screening for the disease. The focus at the time was the sensitivity and specificity (roughly, 1-false positive rate) of the tests at hand. For the tests in question the sensitivity was Pr(Positive Test | Infected) ≈ 1 and the false positive rate was Pr(Positive Test | Uninfected) $\approx .00005$ – an unusually low false positive rate. What was generally neglected in the debate, however, was the low prevelence of the disease in the general population: Pr(Infected) ≈ 0.0001. Since being told you are HIV positive has dramatic ramifications, what clearly matters to you as an individual is the probability that you are uninfected given a positive test result: Pr(Uninfected | Positive test). Calculate this probability. Would you volunteer for such screening?

References

- K. B. Meyer and S. G. Pauker, "Screening for HIV: Can we afford the False Positive Rate," The New England Journal of Medicine, Vol 317, No 4, pg 238–241, 1987.
- R. Weiss and S. O. Thier, "HIV Testing is the Answer – Whats the Question?," The New England Journal of Medicine, Vol 319, No 15, pg 1010–1012, 1988.
- J. B. Jackson et al, "Absence of HIV Infection in Blood Donors with Indeterminate Wester Blot Tests for Antibody to HIV-1," The New England Journal of Medicine, Vol 322, No 4, pg 217–222, 1990.

Solution:

$$\begin{aligned} &\Pr(\text{Uninfected} \mid \text{Positive test}) \\ &\quad = \frac{\Pr(\text{Positive test} \mid \text{Uninfected}) \Pr(\text{Uninfected})}{\Pr(\text{Positive test})} \\ &\quad = \frac{\Pr(\text{Positive test} \mid \text{Uninfected}) \Pr(\text{Uninfected})}{\Pr(\text{Positive test} \mid \text{Uninfected}) \Pr(\text{Uninfected}) + \Pr(\text{Positive test} \mid \text{Infected}) \Pr(\text{Infected})} \\ &\quad = \frac{(\text{False Positive Rate})(1 - \Pr(\text{Infected}))}{(\text{False Positive Rate})(1 - \Pr(\text{Infected})) + (\text{Sensitivity}) \Pr(\text{Infected})} \\ &\quad = \frac{0.00005(1 - .0001)}{0.00005(1 - .0001) + 1(.0001)} = 0.3333 \end{aligned}$$

So there is a 1/3 probability that you are actually healthy if are in a low risk population and your test is positive!

If Pr(Infected) is significantly higher the probability you are actually healthy given a positive test result rapidly decreases to essentially zero. For example if Pr(Infected) = 0.001 (.1% prior probability of infection), Pr(Uninfected | Positive test) = 5% while if Pr(Infected) = 0.01 (1% prior probability of infection), Pr(Uninfected | Positive test) = .5%.

Computer Projects

In the course we will focus on *models* of stochastic phenomona. These models are useful abstractions of reality. In this set of projects we will begin to connect our models to data. The class web site contains two functions, `pdf1d.m`, and `pdf2d.m` which generate estimates of the probability density function (pdf) and probability distribution function (PDF/CDF) of a single and 2 joint random variables, respectively, given vectors of independent observations of these random variables. The estimate is based on a scaled histogram of the data (how is a pdf estimate different from a simple histogram?). You will use these functions in the following problems.

Problem 1.8 Transformations of Random Variables
The purpose of this problem is to investigate transformations of random variables. One very important practical task is the computer generation of random variables with a *given* distribution. In this project we will see how this can be done.

(a) Let X be a uniformly distributed random variable and define the transformed random variable $Y = g(X)$, where the mapping $g(\cdot) : [0,1] \mapsto [-\infty, \infty]$ (i.e. $g(\cdot)$ maps the interval $[0,1]$ to the entire y axis). Suppose that $g(\cdot)$ is also a monotonically increasing function of X, so that there is a unique X corresponding to each Y. What is $P_Y(y)$ in terms of the function $g(\cdot)$? (Note: this solves the analysis or forward problem, in that if a $g(\cdot)$ is specified it explains what the transformed PDF is).

Use this answer to specify a way to generate a random variable Y with a *given* PDF/CDF $P_Y(y)$ starting from a uniformly distributed random variable X. (Note: This solves the design problem, in that the target PDF is given and an appropriate function $g(\cdot)$ to achieve it is sought).

(b) An important random variable is the Cauchy random variable. This random variable (and the Laplace random variable) is often used as a model of impulsive noise in systems. Such distributions are sometimes termed "heavy tailed", referring to the sizeable probability mass in the tails of the distribution (i.e. far from the peak). Write a program `randcau.m` to generate Cauchy random variables. Generate and plot a few thousand Cauchy random variables to see the impulsive behavior. How does the heavy tailed nature of the distribution exhibit itself when trying to generate a pdf estimate using e.g. `pdf1d.m`?

(c) The above method only works when an analytic expression for the inverse of $P_Y(y)$ can be found. One important situation where this is not the case is the Gaussian random variable. Suggest different ways you might generate (approximations to) zero mean, unit variance Gaussian random variables from uniform random variables. Try out your schemes and verify the results using `pdf1d.m`.

Solution:

(a) The first part of this problem asks the question, *given* the function $g(x)$, what is the PDF of the output variable $P_Y(y)$. To find the answer it is helpful to refer to a picture of the situation, such as that in Figure 1. Now we want to find $P_Y(y_0) = \Pr(Y \le y_0)$. From the figure we can see that the equivalent event in terms of X is that $0 \le X \le g^{-1}(y_0)$ so that:

$$P_Y(y_0) = \Pr(Y \le y_0) = \Pr(0 \le X \le g^{-1}(y_0)) = \int_0^{g^{-1}(y_0)} p_X(x)\, dx = g^{-1}(y_0)$$

Thus $P_Y(y) = g^{-1}(y)$ when the input random variable is uniformly distributed on [0, 1] and $g(x)$ has the given properties. This solves the first part.

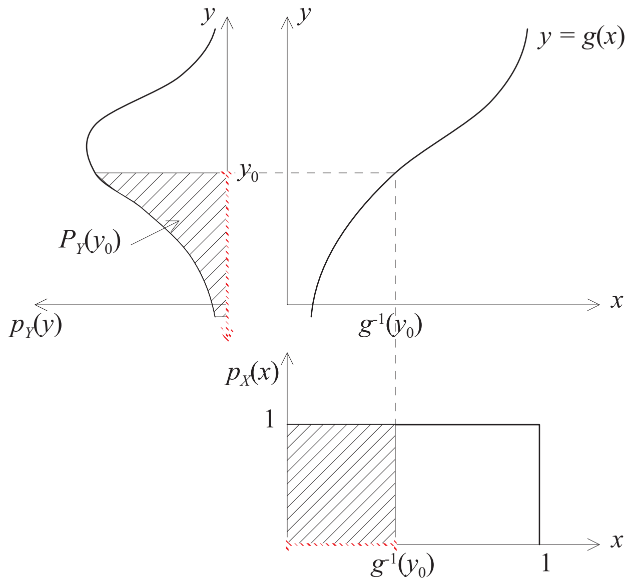


Figure 1: Finding $P_Y(y)$ given $g(x)$

For the second part, note that by inverting both sides of this equation we see that the random variable Y may be obtained from a uniform random variable X by putting the samples x through the transformation $Y = P_Y^{-1}(X)$. Basically, we want to choose $g(X) = P_Y^{-1}(X)$.

(b) This function is given below:

```
function y = randcau(m,n,a)
% y = randcau([m,n],a)
% y = randcau(m,n,a)
%
%       m,n     : Dimensions of matrix generated (mxn).
%       a       : Parameter of Cauchy pdf. Optional. Default: a=1.
%
%       Produces variates from the Cauchy PDF defined by:
%
%                 1       a
%       P_X(x) =  - * ---------
%                 pi  a^2 + x^2
%
%       Here a is the spread of the distribution. Generates an mxn matrix of
%       Cauchy random variates.
%
%       The random numbers are generated using the transform method.

% W. C. Karl 1/6/98

if max(size(m))>1 & nargin == 1
```

```
  a = 1;
  n = m(2);
  m = m(1);
elseif max(size(m))>1 & nargin == 2
  a = n;
  n = m(2);
  m = m(1);
elseif max(size(m))==1 & nargin == 2
  a = 1;
elseif max(size(m))==1 & nargin == 3
else
  error(['Unrecognized input configuration'])
end;

% Generate uniform random variates
x = rand(m,n);

% Perform Transformation
y = a*tan(pi*(x-.5));
```

A plot of Cauchy random variables is shown in Figure 2(a). The experimental pdf is shown in Figure 2(b), where we have only generated the histogram in the range $[-3, 3]$. The difficulty with the heavy tailed distribution when using a uniform cell histogram estimator is that a single large sample value skews the entire histogram! Note that the most of the variates are close to zero, but not infrequently large values also occur.

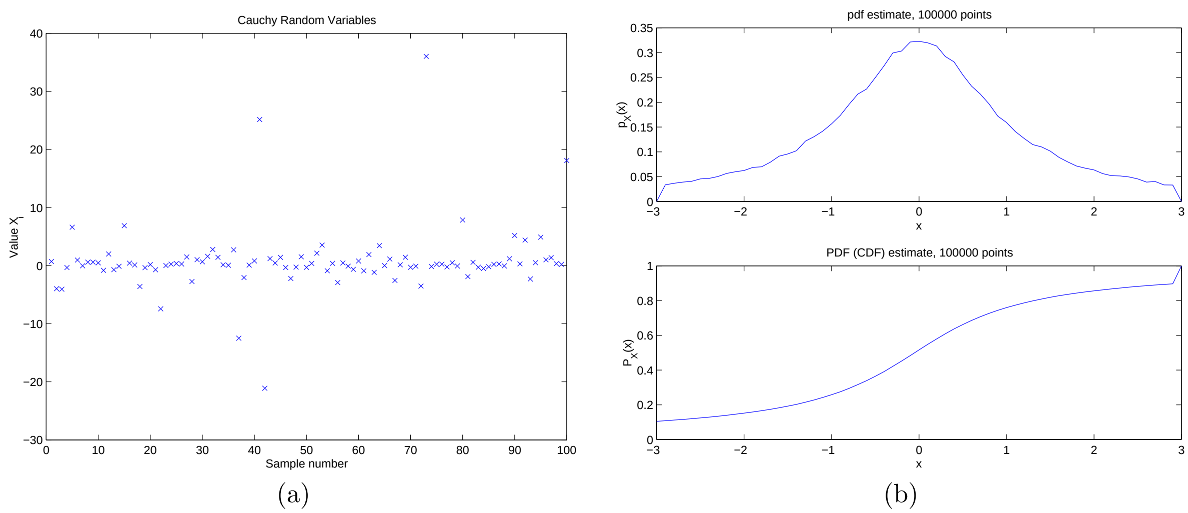
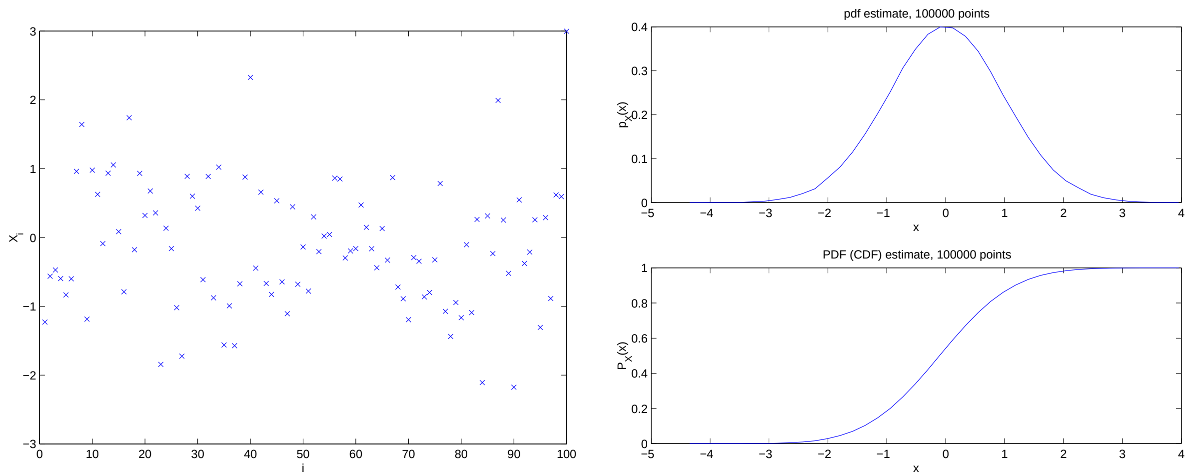


Figure 2: Cauchy random variables

(c) In this problem I just wanted you to think about how to use the tools you know at this point to accomplish something useful. A number of answers are possible, from using a series expansion to the pdf to using the central limit theorem and using a sum of IID random variables. One easily implementable approach is the following. First, use the transform method described above to generate a random variable R that is Rayleigh distributed and a random variable Θ that is uniformly distributed on $[0, 2\pi]$. Then, take these special random "polar" coordinates and generate new random "cartesian" variables as: $X = R\cos(\Theta)$ or $Y = R\sin(\Theta)$ (either will work). One can show that these corresponding X and Y components of the polar points will be independent Gaussian random variables! In summary, points (X, Y) which are uniformly distributed in angle and Rayleigh distributed in radius have independent Gaussian coordinates. Figure 3 shows an empirical pdf/PDF generated from random variables using this method. It works quite well!



(a) (b)

Figure 3: Variates and pdf/PDF of Gaussian random variables

Problem 1.9 Joint Probability Density Functions, Conditionals, and Marginals
In this problem we will investigate joint random variables and their properties.

(a) On the class web site there is a data set `2rvdata.mat` with 4 different sets of joint random variables pairs `(xi,yi)`. Our goal is to understand these random variable pairs better. For each of these sets of joint random variables perform the following calculations:

(i) Generate and plot the empirical joint pdf $p_{X_i,Y_i}(x_i, y_i)$ and PDF $P_{X_i,Y_i}(x_i, y_i)$.

(ii) Generate and plot the empirical marginal distributions $p_{X_i}(x_i)$ and $p_{Y_i}(y_i)$.

(iii) Calculate the empricial covariance matrix and the corresponding correlation coefficient. For your estimate of the covariance matrix approximate the expection operation with the sample mean.

(iv) Estimate and plot the empirical conditional pdf at e.g. $p_{X_i|Y_i}(x_i|y_i = .5)$ and $p_{X_i|Y_i}(x_i|y_i = .75)$.

(b) Based on the information in (a), specify which of the data sets you think came from uncorrelated random variables and which came from independent random variables. For example, you may want to compare $p_{X_i,Y_i}(x_i, y_i)$ to $p_{X_i}(x_i)p_{Y_i}(y_i)$ when deciding independence, etc. Explain your reasoning.

Solution:

(a) The pdf/PDF plots for the different data sets are shown in Figure 4. The marginal pdf plots are shown in Figure 5. The conditional pdfs are plotted in Figure 6. In estimating the conditional pdfs we need to use a tolerance around the given value in finding the points. I used a tolerance of .01 around the values of 0.5 and 0.75. Note that the y values of data set 2 do not (appear to) extend to 0.75 – i.e. there are no values that are this large. Thus since the $y = 0.75$ event never happens, the conditional density really isn't well defined. We simply plot it as zero.

The estimates of the covariance matrices and the correlation coefficients are shown below:

$$\widehat{\Sigma}_{X_1Y_1} = \begin{bmatrix} 0.4309 & -0.0003 \\ -0.0003 & 3.2937 \end{bmatrix}, \quad \widehat{\Sigma}_{X_2Y_2} = \begin{bmatrix} 0.0835 & -0.0003 \\ -0.0003 & 0.0828 \end{bmatrix},$$
$$\widehat{\Sigma}_{X_3Y_3} = \begin{bmatrix} 2.4705 & 1.4910 \\ 1.4910 & 2.4953 \end{bmatrix}, \quad \widehat{\Sigma}_{X_4Y_4} = \begin{bmatrix} 0.0829 & 0.0011 \\ 0.0011 & 1.0009 \end{bmatrix}$$

$$\widehat{\rho}_{X_1Y_1} = -2.3623e-04, \quad \widehat{\rho}_{X_2Y_2} = -0.0041, \quad \widehat{\rho}_{X_3Y_3} = 0.6005, \quad \widehat{\rho}_{X_4Y_4} = 0.0039$$

(b) From the small value of the correlation coefficient it appears that data sets 1, 2, and 4 are uncorrelated. From the joint pdf plots it appears that only data sets 1 and 4 are independent, since knowledge of y does not seem to affect the cross-section of the joint pdf, for these random variables.

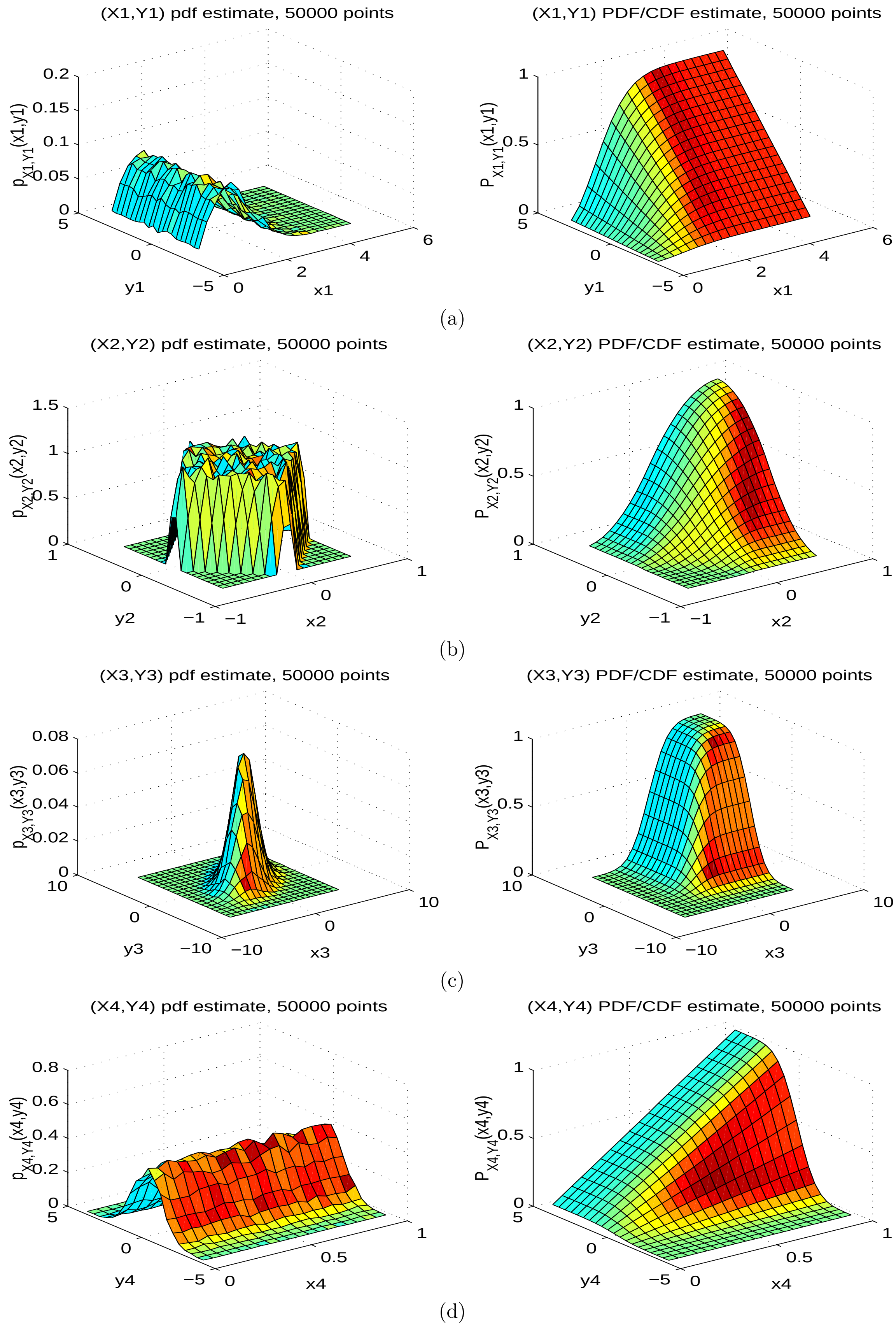
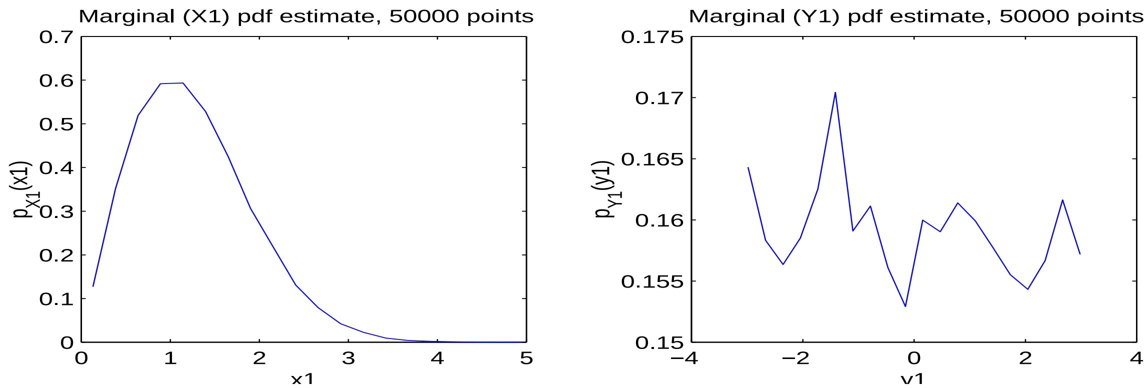
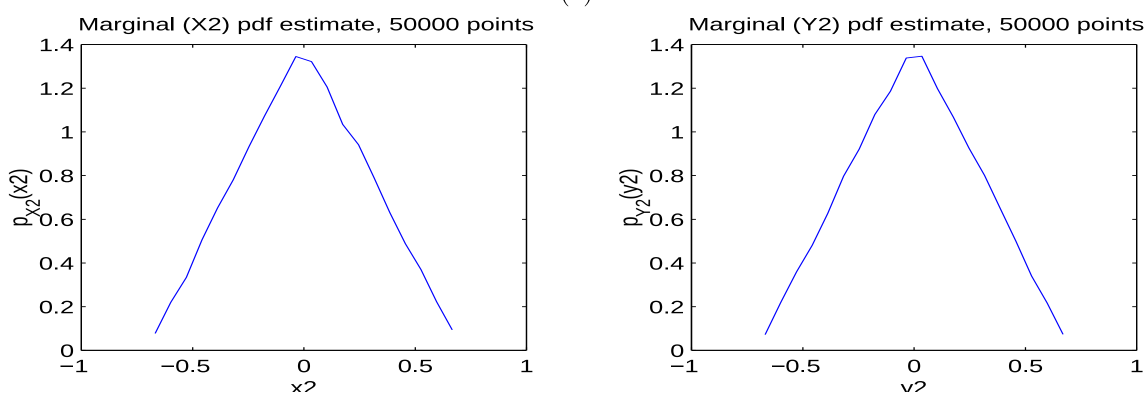


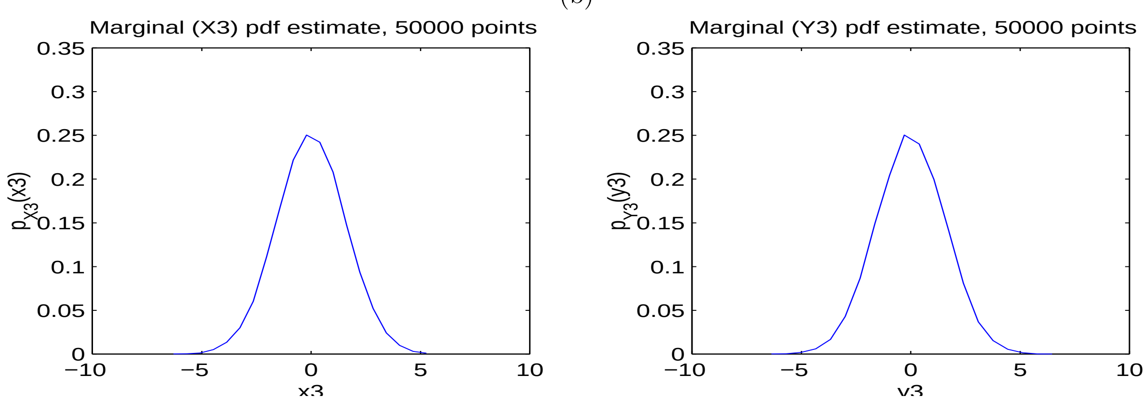
Figure 4: Empirical joint pdf and PDF for data (a) (X_1, Y_1), (b) (X_2, Y_2), (c) (X_3, Y_3), and (d) (X_4, Y_4).



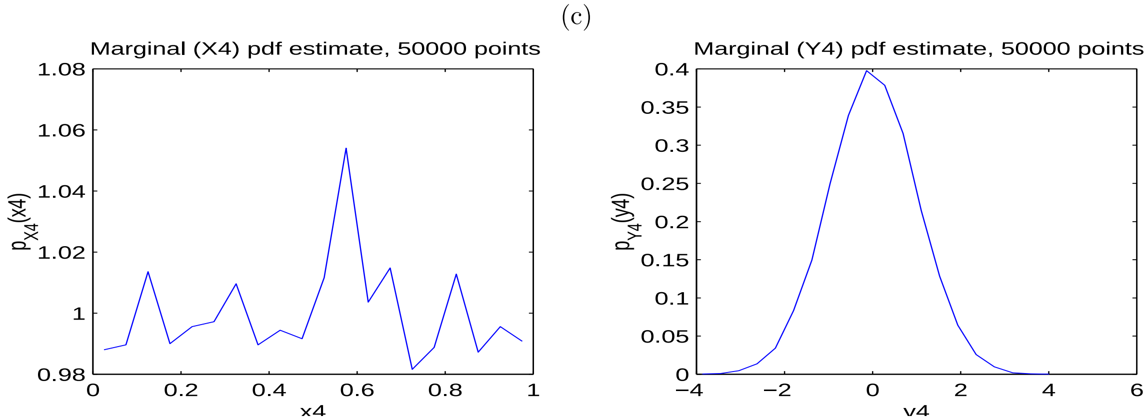
(a)



(b)

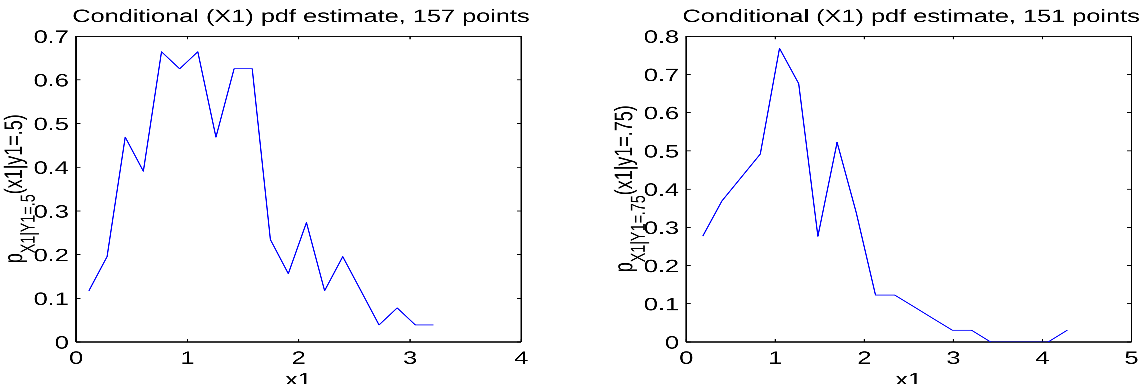


(c)

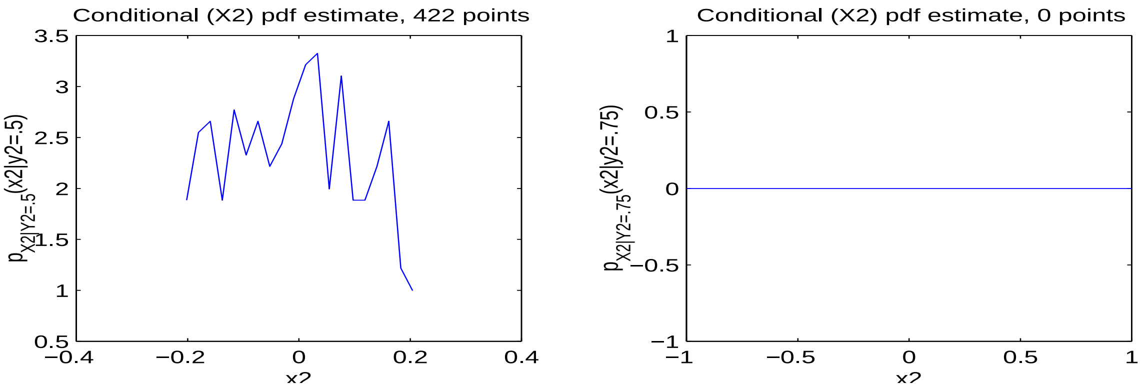


(d)

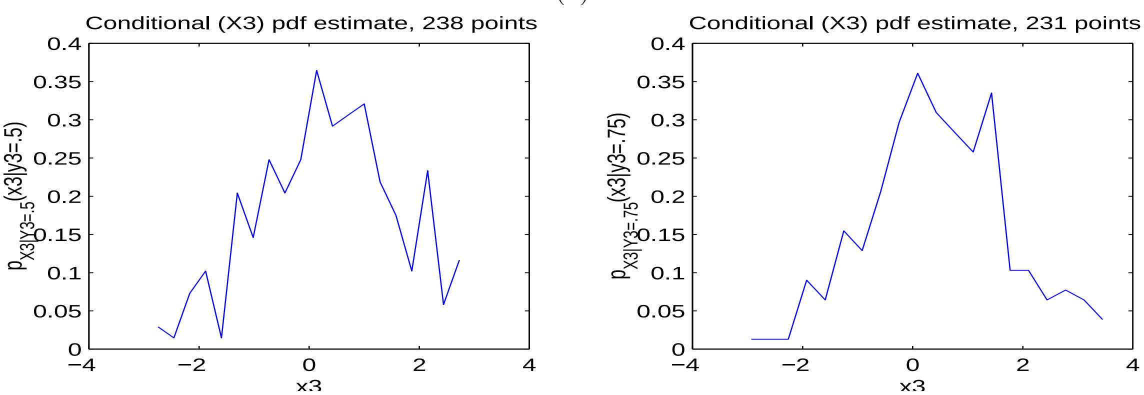
Figure 5: Empirical marginal pdfs for data (a) (X_1, Y_1), (b) (X_2, Y_2), (c) (X_3, Y_3), and (d) (X_4, Y_4).



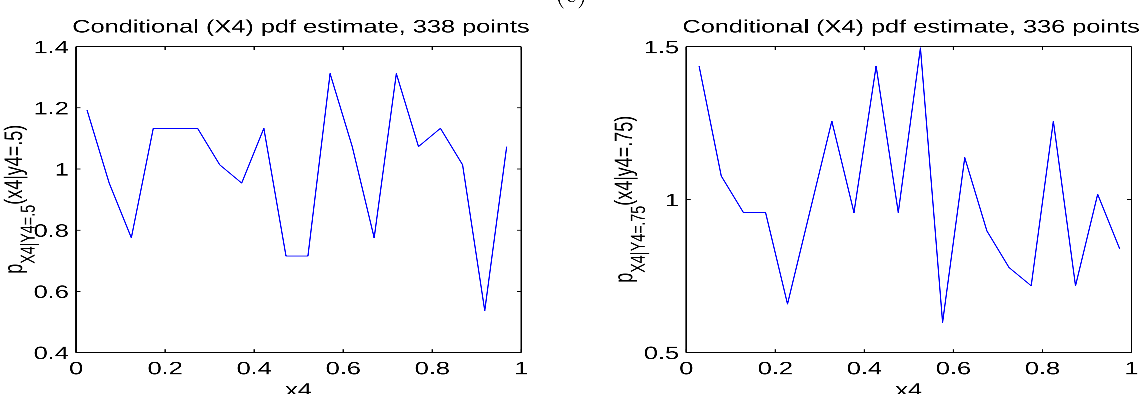
(a)



(b)



(c)



(d)

Figure 6: Empirical conditional pdfs for data (a) (X_1, Y_1), (b) (X_2, Y_2), (c) (X_3, Y_3), and (d) (X_4, Y_4).