Chapter 1

Pairs of Random Variables

1.1 Joint Probability Density Function. Statistical Moments

As with the definition of a pair of events, one may define a pair of random variables. An ouput of a pair of random variables is an array of two numbers. The two random variables that form the pair can be studied separately or together. Individual study of random variables forming the pair may be sufficient to fully characterize the ensemble if and only if the two variables are independent. If the two variables are not independent then the pair is not fully characterized by the description individual statistics.

For a concise presentation, let us consider the pair of random variables (ξ, η). It is described by

- Joint Cumulative Density Function :

$$F_{\xi\eta}(x, y) = P\left((\xi \leq x), (\eta \leq y)\right) \tag{1.1}$$

The two events discussed, $(\xi \leq x)$ and $(eta \leq y)$, occurr simultaneously. The properties of CDF are:

  - Boundary values:

$$F_{\xi\eta}(-\infty, -\infty) = 0 \tag{1.2}$$

$$F_{\xi\eta}(+\infty, +\infty) = 1 \tag{1.3}$$

  - The marginal CDF are :

$$F_{\xi}(x) = P(\xi \leq x) = P\left((\xi \leq x), (\eta \leq +\infty)\right) = F_{\xi\eta}(x, +\infty) \tag{1.4}$$

$$F_{\eta}(y) = P(\eta \leq y) = F_{\xi\eta}(+\infty, y) \tag{1.5}$$

  - The probability for the random variable(RV) to be a rectangular domain $(D_1 = (x_1, x_2] \times (y_1, y_2])$ is:

$$\begin{aligned} &P\left((\xi \in [x_1, x_2]), (\eta \in [y_1, y_2])\right) = \\ =&F_{\xi\eta}(x_2, y_2) + F_{\xi\eta}(x_1, y_1) - F_{\xi\eta}(x_1, y_2) - F_{\xi\eta}(x_2, y_1) \end{aligned} \tag{1.6}$$

- The Joint Probability Density Function is:

$$w_{\xi\eta}(x, y) = \frac{d^2 F_{\xi\eta}(x, y)}{dxdy} \tag{1.7}$$

Main properties:

  - Normalization condition:

$$\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} w_{\xi\eta}(x, y)dxdy = 1 \tag{1.8}$$

  - Marginal PDF:

$$\int\limits_{-\infty}^{\infty} w_{\xi\eta}(x, y)dy = w_{\xi}(x) \tag{1.9}$$

$$\int\limits_{-\infty}^{\infty} w_{\xi\eta}(x, y)dx = w_{\eta}(y) \tag{1.10}$$

  - The probability for the RV to be in a rectangular domain $D_1 = (x_1, x_2] \times (y_1, y_2]$ is:

$$P\left((\xi \in (x_1, x_2]), (\eta \in (y_1, y_2])\right) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} w_{\xi\eta}(x, y)dxdy \tag{1.11}$$

Thi relation may be developed for any type of planar domain D_1:

$$P\left((\xi, \eta) \in D_1)\right)) = \int\int_{D_1} w_{\xi\eta}(x, y)dxdy \tag{1.12}$$

Two random variables are independent if the product of the marginal PDFs equals of the joint PDF:

$$(\xi, \eta) - \text{ independent } \Longleftrightarrow w_{\xi\eta}(x, y)dx = w_{\xi}(x) \cdot w_{\eta}(y), \forall x, y \tag{1.13}$$

This relationship is an extension of the independence of events: 'two events are independent if the probability of happening simultaneously is equal to the product of individual probabilities'. Extending the relationship is immediate if it is envisaged the significance of the RV density. Independence of two random variables is the only case when the deduction 2nd order statistics from statistical measurements of order 1 is doable.

- Second order statistical moments :

  - The raw moment of order p, q is:

$$m_{\xi\eta}^{(p,q)} = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x^p y^q w_{\xi\eta}(x, y)dxdy \tag{1.14}$$

- The moment with respect to the mean of order p, q is :

$$M_{\xi\eta}^{(p,q)} = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(x-\overline{\xi})^p(y-\overline{\eta})^q w_{\xi\eta}(x,y)dxdy \tag{1.15}$$

Widely used are:

- Correlation - the raw moment of order (1,1):

$$R_{\xi\eta} = m_{\xi\eta}^{(1,1)} = \overline{\xi\eta} = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} xyw_{\xi\eta}(x,y)dxdy \tag{1.16}$$

- Covariance - the moment with respect to the mean of order (1,1):

$$K_{\xi\eta} = M_{\xi\eta}^{(1,1)} = \overline{(\xi-\overline{\xi})(\eta-\overline{\eta})} = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(x-\overline{\xi})(y-\overline{\eta})w_{\xi\eta}(x,y)dxdy \tag{1.17}$$

Two random variables are decorrelated if their covariance is zero:

$$K_{\xi\eta} = R_{\xi\eta} - \overline{\xi}\overline{\eta} = \overline{\xi\eta} - \overline{\xi}\overline{\eta} \tag{1.18}$$

If independence is based on a relationship between PDFs, the uncorrelation is given by a similar relationship but based on means, so it is less restrictive. If two random variables are correlated with each other, then there is a linear connection between them. The connection between independence and uncorrelation is:

- Two **independent** random variables are **uncorrelated**
- If two RV are **dependent**, one may conclude nothing about their **correlation**.
- If the two RV are **uncorrelated** one may conclude nothing about their **independence**

These rules are an exception: Gaussian random variables: an uncorrelated pair is also independent.

As the first order pdf was approximated by the histogram, one may construct an estimate of the second order PDF. The process is similar, only that the defined intervals on a linear support for the histogram are replaced by rectangular intervals (defined on a support plan). The result is in the form of a matrix. This matrix is called matrix co-occurrence.

1.2 Regression line

According to dictionary, one of the meanings of 'regression' is 'the act of going back to a previous place or state'. In mathematics a regression is a technique for replacement of a point with a parameterized geometric shape. The cloud of points is hard to be described and thus it is substituted by a form characterized by some parameters of geometrical shapes. This form can be a line, a curve (given by a polynomial or by an arbitrary function), a circle etc. What is the procedure for obtaining popular form?

The procedure to find the parameters is the following: the shape is defined by a set of parameters [1]. We are looking for that set of parameter values that minimize the sum of square distances from each point of the original set to the regressed set of points

The regression line of the results of a pair of random variables allows the study of their correlation. Why? The graphical representation of a cloud of points of a strongly correlated random variables is resemblance a line (i.e. the regression line).

Our problem is to determine the slope (a) and offset (b) of the line that minimizes the distance from it to a set of given points. In this case, the set of points to the particular embodiment consists of a pair of random variables. This problem can be solved by exhaustive search or analytically.

Specifically, the point set is $(x_i, y_i), i = 1, \ldots N$, where x_i are N realizations of the ξ random variable and y_i of the η variable. The analytical solution is found by minimizing the mean square error:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}(y_i - ax_i - b)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i^2 + a^2x_i2^+b^2 - 2ax_iy_i - 2by_i + 2abx_i) \tag{1.19}$$

If we equal the two partial derivatives of the error with zero, we obtain:

$$\frac{\partial\epsilon}{\partial a} = \frac{1}{N}\sum_{i=1}^{N}2ax_i^2 - \frac{1}{N}\sum_{i=1}^{N}2x_iy_i - \frac{1}{N}\sum_{i=1}^{N}2bx_i = 0$$

and respectively:

$$\frac{\partial\epsilon}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}2b - \frac{2}{N}\sum_{i=1}^{N}y_i + \frac{2}{N}\sum_{i=1}^{N}x_i = 0$$

Daca

We denote:

[1] For example, a circle is defined by three parameters: the coordinates center (the abscissa and ordinate) and the radius. A polynomial curve from a function of degree 2 has three parameters: the coefficients of degree 2, 1 and 0

$$
\begin{aligned}
S_X &= \sum_{i=1}^{N} x_i \qquad & S_Y &= \sum_{i=1}^{N} y_i \\
S_{XX} &= \sum_{i=1}^{N} x_i^2 \qquad & S_{YY} &= \sum_{i=1}^{N} y_i^2 \\
& & S_{XY} = \sum_{i=1}^{N} x_i y_i &
\end{aligned}
\tag{1.20}
$$

The system of partial equation is:

$$
\begin{cases}
aS_{XX} + bS_X = S_{XY} \\
aS_X + bN = S_Y
\end{cases}
$$

By solving this system, one fill determine the parameters of the regression line:

$$
\begin{aligned}
a &= \frac{NS_{XY} - S_X S_Y}{NS_{XX} - S_X S_X} \\
b &= \frac{S_Y - aS_X}{N}
\end{aligned}
\tag{1.21}
$$

A consequence of this way of defining the regression line is that it will pass through the center of the point cloud. If one considers two random variables of mean 0, then the point cloud will be centered around the origin and the offset of the regression will also be void. Thus only the slope of the regression is to be found.

A quick analysis of the values calculated in relation ref Eq: RegrParam reveal their statistical significance: S_X and S_Y are proportional to the mean of the two random variables, S_{XX} and S_{YY} with mean of square, and S_{XY} with their correlation. The proportionality constant is N, the number of realizations. In this slope can be rewritten according to statistical measurements :

$$
a = \frac{K_{\xi\eta}}{\sigma_\xi^2}
$$

We recall that the degree of correlation of two random variables is given by the correlation coefficient:

$$
\rho_{\xi\eta} = \frac{K_{\xi\eta}}{\sigma_\xi^2 \sigma_\eta^2}
\tag{1.22}
$$

One may show that the correlation coefficient has the absolute value sub-unitary $|\rho_{\xi\eta}| \leq 1$. A correlation coefficient of 1 means a perfect correlation mode. The degree of correlation is given by the approximation error of a point cloud right [2]. This is done by replacing the relationship 1.19 values calculated for the two parameters:

[2] Typically, the sequence of operations requires first the calculation of regression parameters and then to define the correlation coefficient as a measure of the approximation error

$$
\begin{aligned}
\varepsilon_{min} &= \frac{NS_{YY} - S_Y S_Y}{NS_{XX} - S_X S_X} - \frac{(NS_{XY} - S_X S_Y)^2}{(NS_{XX} - S_X S_X)^2} \\
&\approx \sigma_\eta^2 \left[1 - \left(\frac{K_{\xi\eta}}{\sigma_\xi \sigma_\eta} \right)^2 \right] = \sigma_\eta^2 (1 - \rho_{\xi\eta}^2)
\end{aligned}
$$

Given a set of pairs of points, to find the regression we start by calculating the amounts given by the relation 1.20. Using those, the two parameters that describes the line are found. Last step contains the error of the approximation error of the a cloud with a line.

1.3 Practical work

1. The study of the second order of the probability density function of a pair of Gaussian random variables. This is given by:

$$w_{\xi\eta}(x,y) = \frac{1}{2\pi\sigma_\xi\sigma_\eta\sqrt{1-\rho_{\xi\eta}}}\cdot \exp\left[-\frac{1}{2(1-\rho_{\xi\eta})}\left(\frac{(x-\mu_\xi)^2}{2\sigma_\xi^2} - 2\rho_{\xi\eta}\frac{(x-\mu_\xi)(y-\mu_\eta)}{\sigma_\xi\sigma_\eta} + \frac{(y-\mu_\eta)^2}{2\sigma_\eta^2}\right)\right]$$

The graphical representation of the 2-nd order PDF for various values of the two means, dispersion and correlation coefficient must be done. Then calculate and plot the marginal distributions.

Code.

```
x=-50:50;
y=-50:50;
mx=0;my=0;
sx=5;sy=6;
rxy=0.2;
const1=1/(2*pi*sx*sy*sqrt(1-rxy));
const2=1/(2*(1-rxy));
sx2=2*sx^2 ;
sy2=2*sy^2;
rsxsy=2*rxy/(sx*sy);
for i=1:length(x)
     for j=1:length(y)
          % compute the joint pdf
          w(i,j)=const1*exp(-const2*....
            ((x(i)-mx)^2/sx2-rsxsy*(x(i)-mx)*(y(j)-my)+(y(j)-my)^2/sy2));
     end
end % compute the marginals
wx=sum(w);
wy=sum(w´);
% plot the joint pdf
figure(1);mesh(x,y,w);
% plot the marginals
figure(2);
subplot(1,2,1);plot(x,wx);
subplot(1,2,2);plot(y,wy);
```

One possible result can be seen in Figure 1.1.

Homework: Implement calculation of the joint PDF by vector operations (i.e. remove the *fors*) .

2. Generate two Gaussian random variables with given mean and variances. It is assumed that the two variables are independent. This assumption is met if they are generated separately. Represent graphically co-occurence matrix

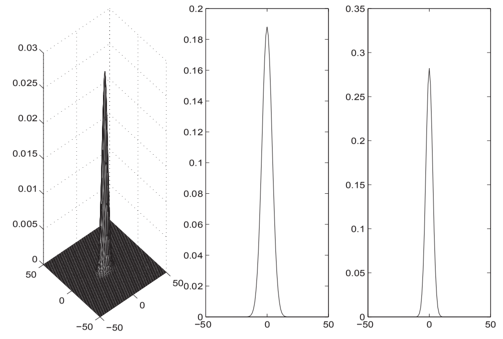

Figure 1.1: Joint pdf for a Gaussian pair of RV.

(probability density approximation order 2). Compare with the results of the previous section.

Solution: We recall that for Gaussian random variables (and only for them) de-correlation implies independence. The code is:

```
 x=randn(1,400);%400 de number from a gaussian with 0 mean and 1 variance.
y=-2+2*randn(1,400);%mean -2 and variance 4
%co-occurrence matrix in 11x11 points
mx=min(x);Mx=max(x);
my=min(x);My=max(y);
absx=mx:(Mx-mx)/10:Mx;
absy=my:(My-my)/10:My;
for i=1:10
    for j=1:10
        M(i,j)=sum((x<absx(i+1)&x>=absx(i))&(y<absy(j+1)&y>=absy(j)));
    end;
end;
figure;surf(absx(1:end-1),absy(1:end-1),M);
```

An example of the co-occurrence matrix may be seen in 1.2.

3. Generate 200 realizations of two random uniform variables x_i, y_i $(i = 1, \ldots 200)$. Plot the 200 points in the plane xOy. Draw the regression line and compute the approximation error. Study two cases: the random variables are independent and correlated (generate the first variable outputs x_i and the achievements of the second will be obtained by a linear transformation.

Code.

```
 N=200;
x=rand(1,N);
```

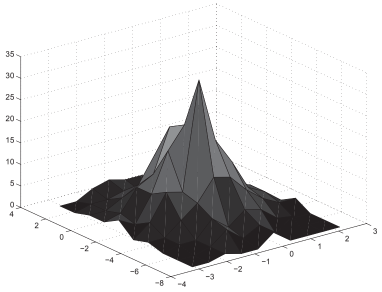

Figure 1.2: Approximation of the joint PDF of a pair of random variables.

```
y=rand(1,N);
plot(x,y,´.´);
Sx=sum(x);
Sy=sum(y);
Sxx=sum(x.*x);
Syy=sum(y.*y);
Sxy=sum(x.*y);
%regression line slope
a=(N*Sxy-Sx*Sy)/(N*Sxx-Sx*Sx);
b=(Sy-a*Sx)/N;
%the line points
xd=sort(x); yd=a*xd+b;
%plot the line
hold on; plot(xd,yd,´r´); %approximation error
corcoef=(N*Sxy-Sx*Sy)/((N*Sxx-Sx*Sx)*(N*Syy-Sy*Sy)));
e=((N*Syy-Sy*Sy)/(N*Sxx-Sx*Sx))-a^2;
fprintf(´correlation coefficient is =%d´,corcoef);
fprintf(´error is= %d´,e);
```

One possible outcome, if the latter is obtained by a linear function applied to the RV x, plus an error: $y = ax + \varepsilon$), where ε is a random variable small uniform dispersion) can be seen in Figure 1.3.

4. Load the matrices $A_1, \ldots, A_6$ stored in the file *Regresie.mat* (see Figure 1.4). Each array has two rows and a number N of columns. A column represents the coordinates of a point in space $(x_i, y_i)^T$. Knowing that the points are from a pair of random variables ξ, η determine, in each case, the degree of correlation between two random variables.

   A Matlab data file is load with the *load* command. *save* command to write on the disk.

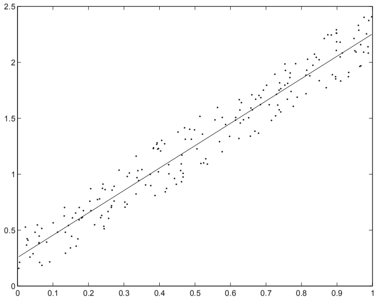

Figure 1.3: The regression line for a pair of random variables strongly correlated.

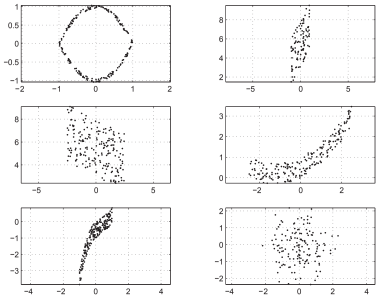

Figure 1.4: Strongly or weaker correlation between data