Multivariate Statistical Modeling (CDT-26)

Preprint · April 2020

DOI: 10.13140/RG.2.2.29085.67045/2

Luciano da F. Costa
University of São Paulo

Multivariate Statistical Modeling (CDT-26)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

4th April 2020

Abstract

Multivariate statistics can be applied to a vast number of scientific and technological problems, including analysis, modeling, hypothesis testin, prediction. In this work, we develop an introduction to some of the main concepts in multivariate statistics, including random vectores, joint probability density functions, joint variation measurements (including covariance and Pearson correlation coefficient), generic multivariate moments, as well as parametric and non-parametric estimation of multivariate probability densities.

'As good luck would have it.'

W. Shakespeare.

1 Introduction

Probability and statistics are fascinating areas with a wide range of theoretical and application potential in virtually every scientific and technological area. It can be used for data normalization and enhancement, characterization, analysis, modeling, simulation, hypothesis testing, prediction... The list is particularly long, because almost every scientific and technological problem can be approached from the probabilistic point of view. The very fact that you are reading these lines provides further indication that probability and statistics are interesting and important.

In a previous work, namely CDT-13 [1], a concise introduction to univariate statistics was presented. By the term *univariate*, it is typically meant random experiments involving a single measurement or random variable. While this type of situations is certainly important, oftentimes we deal with random experiments involving several random variables, which gives rise to the concept of *multivariate statistics* (e.g. [2, 3, 4, 5]), which corresponds to the main subject developed in the present work.

Figure 1 depicts a random experiment that is being characterized in terms of M respective random variables $X_1, X_2, \ldots, X_M$. At each realization of the random experiment, a whole set of instances of these variables can be observed. As a simple example, one can imagine an apple orchard from which fruits are obtained and have properties such as weight, length, sugar index, etc., measured and represented as a respective random variable X_i, $i = 1, 2, \ldots, M$.

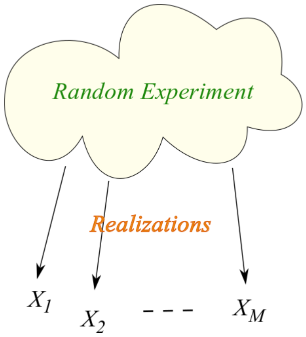


Figure 1: A *multivariate random experiment* being characterized in terms of M respective measurements or random variables $X_1, X_2, \ldots, X_M$, which can be represented as a *random vector* $\vec{X}$. Each time the random experiment is performed, a respective set of these variables is obtained as samples.

We start by presenting the concept of random vectors, and follow by describing joint (multivariate) density functions, multivariate moments, measurements of joint variation of data (correlation, covariance, Pearson correlation coefficient), as well as the interesting problem of parametric and non-parametric multivariate probability density function estimation.

Multivariate statistics involves long integral and differential expressions, reflecting the multiple involved random variables. In order to provide a more progressive approach, and as a didactic resource, in this text several of the related equations are presented first with respect to only two random variables, X_1 and X_2, followed by the respective generic expression considering M random variables.

It is strongly recommended that the previous CDT-13 [1] be read first and/or jointly as a preparation for the current text.

2 Random Vectors

Given a random experiment involving M measurements or random variables X_i, each of its realizations can be expressed in terms of a respective *random vector* $\vec{X}$:

$$\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} \tag{1}$$

For instance, let's consider that we are interested in studying an apple orchard. The random experiment can be defined as collecting a sample of N apples at a given autumn day. At each realization of this experiment, performed on different days, we have a new sample of fruits, which can be characterized in terms of M respective measurements, such as weight, length, sugar degree (e.g. Brix index), etc. Each of these measured properties is understood as a respective random variable X_i, $i = 1, 2, \ldots, M$.

Multivariate statistics is to a great extent concerned about the characterization and prediction of the values of the random variables constituting the respective random vector $\vec{X}$. An interesting related concept is the *ensemble* of a multivariate random experiment, corresponding the complete set of possible respective realizations.

3 Joint Probability Density Functions

One possible means to better understand $\vec{X}$ is to consider the respective *joint probability density function* $p(X_1, X_2, \ldots, X_M)$. Indeed, these densities can provide all statistical information as possible about the respective random vectors $\vec{X}$.

Joint (or multivariate) probability density functions can be known *a priori*, or need to be estimated, such as by using multivariate relative frequency histograms or other methods to be briefly discussed in Section 7.

Any probability density function $p(X_1, X_2)$ that satisfies the following conditions, and every function that satisfy these conditions are a potential candidate for probability density function of the random vector in some random experiment:

$$p(X_1, X_2) \geq 0 \text{ for any } X_1, X_2; \tag{2}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(X_1, X_2) dX_1 dX_2 = 1 \tag{3}$$

These conditions imply that $p(X_1, X_2)$ be non-negative and that the total volume under it be equal to 1.

One of the reason probability density functions are so important is that they allow us to calculate probabilities of observing intervals of values of the random variables and also can be used to define statistical moments (see Section 6). Indeed, it is important to keep in mind that probability density functions are *densities*, implying that the probability of observing any specific random vector is null, i.e. $P(X_1 = a, X_2 = b) = 0$, for $a, bin\Re$.

However, it is possible to calculate probabilities of observing random variable values within given intervals, i.e.:

$$P(a_1 \leq X_1 < b_a, a_2 \leq X_2 < b_2) = \\ = \int_{a_1}^{b_1} \int_{a_2}^{b_2} p(X_1, X_2) dX_1 dX_2 \tag{4}$$

which corresponds to the volume between the probability density surface and the domain $X_1 \times X_2$ within the region defined by the intervals of interest $[a_1 \leq X_1 < a_2]$ and $[b_1 \leq X_2 < b_2]$.

When extended to M random variables, the conditions in Equation 2 become:

$$p(X_1, X_2, \ldots, X_M) \geq 0 \text{ for any } X_1, X_2, \ldots, X_M; \tag{5}$$

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(X_1, X_2, \ldots, X_M) dX_1 dX_2 \ldots dX_M = 1, \tag{6}$$

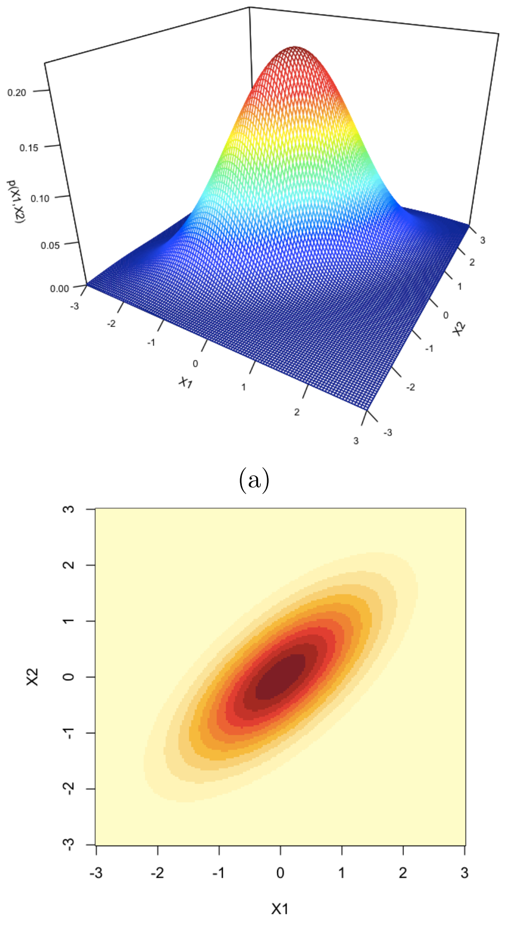
corresponding to the same requisites as for the case $M = 2$ above.

An important joint probability density function is the *multivariate normal density*, whose definition for generic M random variables is:

$$g_{\vec{\mu},K}(\vec{X}) = \\ = \frac{1}{(2\pi)^{M/2}} |K|^{-1/2} exp \left\{ -\frac{1}{2} \left(\vec{X} - \vec{\mu} \right)^T K^{-1} \left(\vec{X} - \vec{\mu} \right) \right\} \tag{7}$$

which has as *parameters* the column vector $\vec{\mu}$ containing the averages of each respective random variables and the data covariance matrix K. In the above equation, $|K|$ is the determinant of K.

Figure 2 illustrates a bivariate normal density ($M = 2$), with $\vec{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $K = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$ as a 3D visualization (a) and as an image with level-sets (b).



(a)

(b)

Figure 2: Visualization of a multivariate normal density for $M = 2$ and specific average column vector $\vec{\mu}$ and data variance matrix K as a 3d surface (a) and as an image with level-sets (b).

Observe that $\vec{\mu}$ defines the position of the peak of the normal density, and K specifies its elongation and inclination. Also, it is interesting to note that each of these two types of visualizations allow complementary information about the probability density function. For instance, while the 3D visualization provides a good overall idea of the surface, including its height, the image allows a good idea of the (X_1, X_2) structure of the density function.

Given a joint probability density function $p(X_1, X_2)$, one is often interested in deriving the respective *marginal univariate density* $p(X_1)$ and $p(X_2)$, which can be obtained by integrating along all other variables other than the one that is being considere, i.e.:

$$
\begin{aligned}
p(X_1) &= \int_{-\infty}^{\infty} p(X_1, X_2) dX_2 \\
p(X_2) &= \int_{-\infty}^{\infty} p(X_1, X_2) dX_1
\end{aligned} \tag{8}
$$

Figure 3 depicts the marginal density $p(X_1)$ obtained from the normal density in Figure 2. In this particular case, the other marginal density $p(X_2)$ can be verified to be identical to $p(X_1)$.

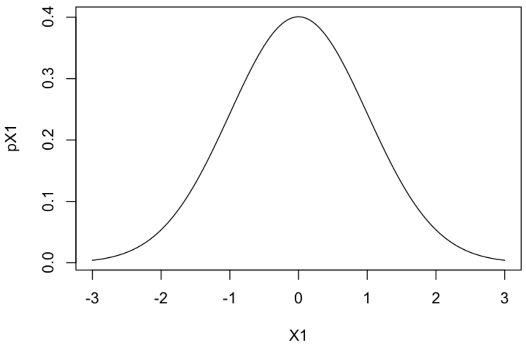


Figure 3: The marginal probability density function $p(X_1)$ obtained from the bivariate normal density in Fig. 2.

In the more generic case involving generic M, we have:

$$
\begin{aligned}
p(X_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} & p(X_1, X_2, \dots, X_M) \\
& dX_1 dX_2 \dots dX_{i-1} dX_{i+1} \dots X_M
\end{aligned} \tag{9}
$$

4 Conditional Densities

Given a probability density function $p(X_1, X_2, \dots, X_M)$, we may be interested in fixing one of the random variables, i.e. $X_i = c)$ for some real constant C, in order to obtain a new density with $M - 1$ random variables. This act of fixing a given variable is related to the concept of *conditional probability.*

In the case of bivariate densities, the above objective can be achieved by applying the following equation while making $X_2 = c$:

$$
p(X_1 | X_2 = c) = \frac{p(X_1, X_2 = c)}{\int_{-\infty}^{\infty} p(X_1, X_2 = c) dX_1} \tag{10}
$$

which ensures proper normalization of the obtained probability density function, in the sense that its total area will be equal to 1.

5 Correlation, Covariance, and Pearson Correlation Coefficient

One of the main senses in which multivariate statistics differs from its univariate counterpart regards the fact that in the former case we have, as we have more than a single random variable, it is necessary to consider possible *interactions* between them, e.g. in the sense of their joint variation.

For instance, in the case of generic fruits, it is reasonable to expect that their weight will tend to increase with their volume, which provides a possible example of *positive* joint variation. On the other hand, it could be expected that the smaller the temperature, the higher the expenses with heating, which characterizes a *negative* joint variation.

Because we are often interested in considering joint variations, it is important to have more objective, mathematical means for quantifying this property. Three concepts are often adopted for this finality: correlation, covariance and Pearson correlation coefficient between two given random variables X_i and X_j. Each of them will be presented and discussed in this section.

The simplest of these, here called *correlation*, is defined from the respective joint probability density as:

$$Corr(X_i, X_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X_i X_j \; p(X_i, X_j) dX_i dX_j \quad (11)$$

When a set of observations of the two random variables $X_{i,p}$ and $X_{j,p}$, $p = 1, 2, \ldots, N$ is available instead of the joint probability $p(X_i, X_j)$, the following equation can be considered for obtaining an *estimation* of the correlation between X_i and X_j:

$$Corr(X_i, X_j) \approx \frac{\sum_{p=1}^{N} (X_{i,p})(X_{j,p})}{N-1} \quad (12)$$

Figure 4 illustrates the interpretation of correlation for 3 possible situations, which include a circular uniform distribution of points centered at the coordinate origin, the previous distribution centered at $(3, 3)$, and the former distribution elongated by a factor of 2 along the X_1 axis.

In Figure 4(a), we have a uniformly distributed set of points with circular borders centered at the origin of the coordinate system. As the points are uniformly distributed around $(0, 0)$, there will be a strong tendency that the products $X_1 X_2$ implied by Equation 12 for the points at the quadrant $(X_i \geq 0, X_j \geq 0)$ will cancel with the products obtained for the quadrant $(X_i \leq 0, X_j \geq 0)$, a similar trend occurring between the quadrants $(X_i \geq 0, X_j \leq 0)$ and $(X_i \leq 0, X_j \leq 0)$. As a consequence, the resulting correlation value could be expected to be very small. This is precisely what happens for the points in Figure 4(a).

However, when the point distribution in Figure 4(a) is shifted to $(3, 3)$, resulting the situation shown in Figure 4(b), the above mentioned products no longer tend to cancel, and a relatively high correlation value is obtained. A similar situation is verified for the point distribution in Figure 4, yielding a slightly larger correlation value as a consequence of the scaling along the X_1 axis.

Another possibility to quantify the joint variation between two random variables X_1 and X_2 is in terms of their *covariance*, defined as:

$$Cov(X_i, X_j) = \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X_i - \mu_{X_i})(X_j - \mu_{X_j}) \; p(X_i, X_j) dX_i dX_j \quad (13)$$

The (unbiased) estimator for the covariance is given as:

$$Cov(X_i, X_j) \approx \frac{\sum_{p=1}^{N} (X_{i,p} - \mu_{X_i})(X_{j,p} - \mu X_j)}{N-1} \quad (14)$$

where the averages μ_{X_i} and μ_{X_j} can be respectively estimated as:

$$\mu_{X_i} \approx \frac{\sum_{p=1}^{N} X_{i,p}}{N}; \qquad \mu_{X_j} \approx \frac{\sum_{p=1}^{N} X_{j,p}}{N} \quad (15)$$

Figure 4 also shows the covariances obtained for each of the three considered cases. We observe that, as a consequence of the subtraction of the mean of each of the variables in Equation 14, it resulted identical for situations (a) and (b), which differ only by their center of mass (given by the averages of X_i and X_j). However, the elongation of the point distribution in (c) still influences the covariance, which resulted different for this case. Informally speaking, we could say that the covariance does not 'sense' the position of the points, being invariant to translation.

The third joint variation statistical measurement considered in this work is the *Pearson correlation coefficient*, given as

$$PCorr(X_i, X_j) = \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{X_i - \mu_{X_1}}{\sigma_{X_i}} \frac{X_j - \mu_{X_j}}{\sigma_{X_j}} \; p(X_i, X_j) dX_i dX_j \quad (16)$$

which can be estimated from samples of X_i and X_j as:

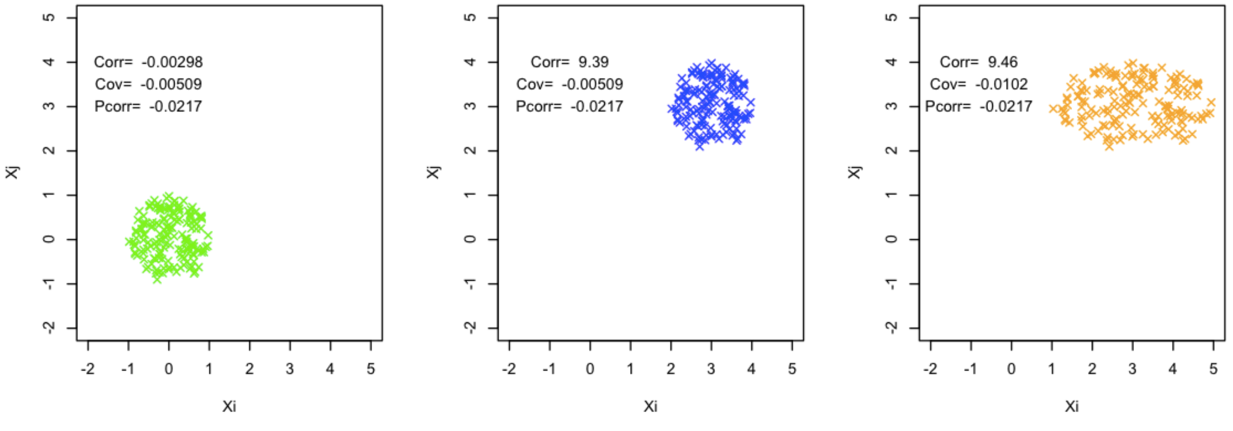
$$Pcorr(X_i, X_j) \approx \frac{1}{N-1} \sum_{p=1}^{N} \frac{X_{i,p} - \mu_{X_i}}{\sigma_{X_i}} \frac{X_{j,p} - \mu X_j}{\sigma_{X_2}} \quad (17)$$

where the two standard deviations can be estimated as $\sigma_{X_i} = +\sqrt{Cov(X_i, X_i)}$ and $\sigma_{X_j} = +\sqrt{Cov(X_j, X_j)}$.

The *standardized* version of a random variable X_i can be obtained as:

$$\tilde{X}_i = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}} \quad (18)$$

Interestingly, the Pearson correlation coefficient between X_i and X_j can be understood as the correlation



(a) (b) (c)

Figure 4: Values of correlation, covariance and Pearson correlation coefficient for three instances of points distribution: circular centered at $(0,0)$, the former centered at $(3,3)$, and then former elongated by a factor of 2 along the x_1 axis. The correlation values vary for each case, the covariance is not influenced by the translation in (b), and the Pearson correlation coefficient is the same for every case.

(i.e. Equation 12) between the respective standardized versions of those two original variables. Also, observe that the Pearson correlation coefficient is bound within the interval $[-1,1]$.

The Pearson correlation coefficient is illustrated for the situations in Figure 4, resulting in identical values in all cases. In particular, the scaling by 2 along the X_i axis in (c) was normalized by the divisions by the standard deviations, being therefore overlooked by the Pearson correlation coefficient.

Figure 5 illustrates four bivariate normal distributions and their respectively estimated Pearson correlation coefficients. It is interesting to observe how the Pearson correlation coefficient values tend to 'saturate', in the sense of varying more slowly, as the points become more and more aligned.

Given a more generic situation involving M random variables, it is interesting to define matrices respectively to each of the above three measurements of joint variation. For instance, in the case of the covariance, we can define the *covariance matrix* of the variables $X1, X_2, \ldots, X_M$ as

$$K_{i,j} = Cov(X_i, X_j) \tag{19}$$

Observe that the three respective matrices, as a consequence of their respective definitions, are necessarily *symmetric.*

6 Multivariate Moments

The correlation and covariance statistical measurements of joint variation discussed in the previous section can be understood as particular cases of the more general concept of statistical moments.

The $M^{[k_1,k_2]}-$moment of the random variables X_1 and X_2 can be defined as:

$$M^{[k_1,k_2]}(X_1,X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X_1^{k_1} X_2^{k_2}\ p(X_1,X_2) dX_1 dX_2 \tag{20}$$

The correlation as seen above can be understood as the $M^{[k_1=1,k_2=1]}-$moment of the variables X_i and X_j.

In the more general case of M random variables, we have:

$$M^{[k_1,k_2,\ldots,k_M]}(X_1,X_2,\ldots,X_M) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} X_1^{k_1} X_2^{k_2}\ldots \ldots X_N^{k_M}\ p(X_1,X_2,\ldots,X_M) dX_1 dX_2 \ldots dX_M \tag{21}$$

The *centered* version of the above moments can be defined as:

$$C^{[k_1,k_2,\ldots,k_M]}(X_1,X_2,\ldots,X_M) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} (X_1 - E[X_1])^{k_1} (X_2 - E[X_1])^{k_2} \ldots \ldots (X_N - E[X_1])^{k_M}\ p(X_1,X_2,\ldots,X_M) dX_1 dX_2 \ldots \ldots dX_M \tag{22}$$

Now, we have that the covariance between two random variables X_i and X_j corresponds to their $C^{[k_1,k_2]}-$moment. The other moments above are also potentially interesting and useful, though tending to have a less direct interpretation.

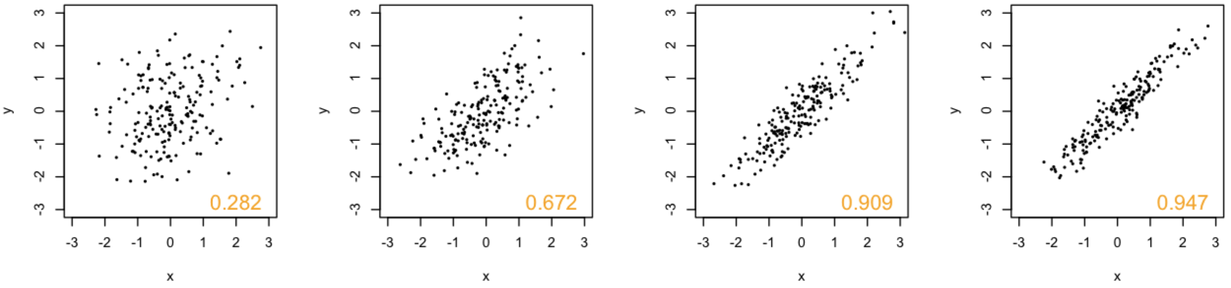


Figure 5: Four bivariate normal distributions with increasing Pearson correlation coefficientes, shown in orange.

7 Parametric and Non-Parametric Estimation

As with univariate probability densities, multivariate can be specified from a formula $p(X_1 X_2, \ldots, x_M)$ or from samples of the involved variables $X_1, X_2, \ldots, X_M$. The former is only possible through theoretical means. In the case we have a set of samples of the random variables, we need to consider a *parametric* or *non-parametric* respective method. In this section, we briefly presents, through illustrations, these two possibilities.

First, consider the table of values sampled for two respective random variables X_1 and X_2.

Table 1: Samples of random variables X_1 and X_2 considered in the estimation examples.

sample	X_1	X_2
1	0.42	0.58
2	-0.31	-0.49
3	0.01	-0.22
4	0.93	0.78
5	0.71	0.59
6	0.88	0.69
7	0.47	0.38
8	0.20	-0.15

We have the points, and we need a respective estimation of the probability density. The parametric way is to assume some type of probability density function, let's say multivariate normal, and to estimate its respective parameters average vector and data covariance matrix. By using Equations 15 and 14, we obtain:

$$\vec{\mu} \approx \begin{bmatrix} 0.413 \\ 0.270 \end{bmatrix} \tag{23}$$

and

$$K \approx \begin{bmatrix} 0.186 & 0.199 \\ 0.199 & 0.234 \end{bmatrix} \tag{24}$$

which completely define a bivariate normal density (Equation 7) that corresponds to the estimation based on the 8 samples.

Figure 6 depicts the so-obtained bivariate probability density function in terms of level-sets. A reasonable adherence can be observed between the original samples (shown in orange) and the obtained density.

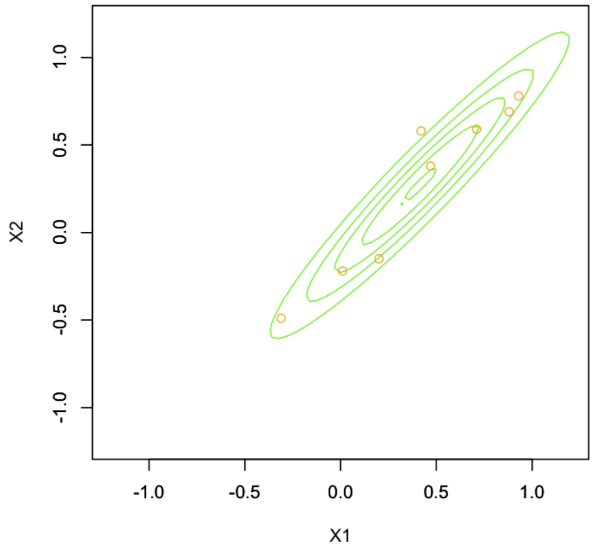


Figure 6: The parametrically estimated bivariate normal probability density function, illustrated in terms of respective level-sets, obtained for the samples in Table 1, also shown in orange.

Now, we briefly address the *non-parametric* estimation of multivariate probability density functions. Unlike *parametric* estimation, here we do not make any assumption on the formula of the likely probability density, but instead rely on the own distribution of original points, which are 'padded' or statistically interpolated by some means.

A possible method for multivariate parametric estimation consists of convolving (e.g. [6]) the original set of points, represented as respective Dirac's deltas, with a suitable *kernel*. In a sense, this method can therefore be understood as if the original samples imprinted themselves, through the adopted kernel, into the respectively estimated density. The following example illustrates the non-parametric estimation of a possible bivariate probability density function for the points in Table 1 considering a circularly symmetric normal kernel.

In absence of additional information about the sought probability density function, a circularly symmetric normal kernel can be considered. Such a kernel is defined by a covariance matrix of the type:

$$K \approx \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \tag{25}$$

being characterized by perfectly circular level-sets. Figure 7 illustrates the result of convolving the circularly symmetric normal kernel ($a = 1$) with Dirac's deltas defined by the position of the samples in Table 1. As expected, the resulting density is more circular than that obtained in Figure 6, as a consequence of the adopted circularly symmetric kernel.

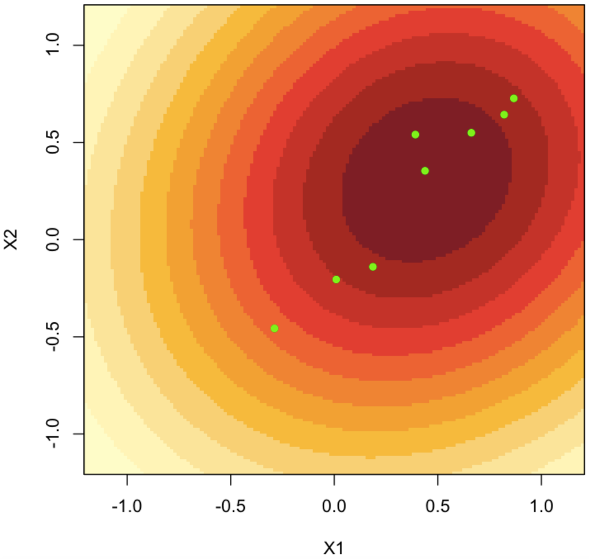


Figure 7: The non-parametrically estimated bivariate normal probability density function, illustrated in terms of respective level-sets, obtained for the samples in Table 1, also shown in green. A circularly symmetric normal function has been adopted in this example.

Other types of kernels can be considered. For instance, Figure 8 illustrates a non-parametric probability density estimation for the same previous samples but using covariance matrix equal to:

$$K \approx \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{26}$$

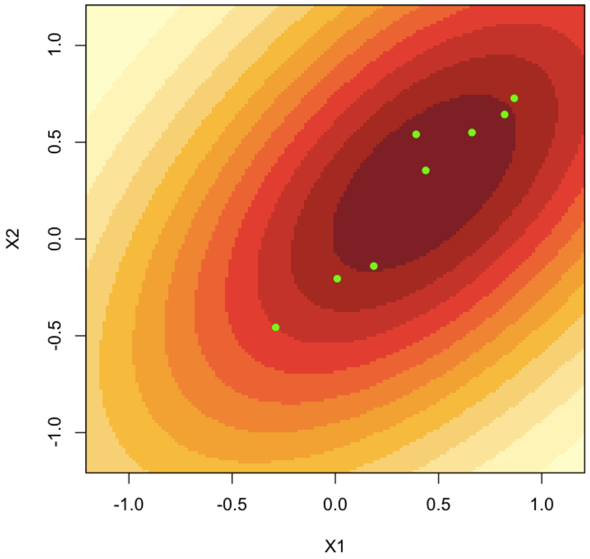


Figure 8: The non-parametrically estimated bivariate normal probability density function, illustrated in terms of respective level-sets, obtained for the samples in Table 1, also shown in green. An elongated normal kernel, aligned to the original points, has been adopted in this example.

Having obtained an estimation of a probability density function for a given set of data, it is important to statistically validate it.

8 Concluding Remarks

Multivariate statistics constitutes a particularly important area in science and technology, mainly as a consequence that almost any problem that is treated deterministically can be extended to a statistical counterpart. It is also essential for dealing with real-world measurements and data.

In the present work, we developed an introductory presentation of some of the main concepts in multivariate statistics, including random vectors, joint probability densities, joint variation measurements, and parametric and non-parametric estimation.

The research area of multivariate statistics is ample and with vast applications and interfaces with several other areas (e.g. signal processing, computer vision, biology, physics, etc.), and it is hoped that the reader will be motivated to probe further into it (e.g. [2, 3, 4, 5, 7]).

Acknowledgments.
Luciano da F. Costa thanks CNPq (grant

no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2.

Costa's Didactic Texts – CDTs

CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also incorporate new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Each CDT focuses on a limited set of interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provided in order to complement the covered subjects.

Observe that CDTs, which come with absolutely no warranty, are non distributable and for non-commercial use only.

The complete set of CDTs can be found at: https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs.

References

[1] L. da F. Costa. Statistical modeling. https://www.researchgate.net/publication/334726352_Statistical_Modeling_CDT-13, 2019. [Online; accessed 4-Apr-2020.].

[2] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson, 2018.

[3] R. M. Warner. *Applied Statistics*. Sage Publication, 2012.

[4] A. V. Rencher and W. F. Christensen. *Methods of Multivariate Analyis*. Wiley, 2012.

[5] S. M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice Hall, 1993.

[6] L. da F. Costa. Convolution! Researchgate, 2019. https://www.researchgate.net/publication/336601899_Convolution_CDT-14. Online; accessed 04-Apr-2020.

[7] L. da F. Costa. Features transformation and normalization: A visual approach. Researchgate, 2020. https://www.researchgate.net/publication/340114268_Features_Transformation_and_Normalization_A_Visual_Approach_CDT-24. Online; accessed 08-Apr-2020.