Forth Seminar on
Information Theory and its Applications
October 14-15, 2020

MULTIVARIATE STATISTICAL CONTROL LIMITS ACCORDING TO DEPENDENCIES

S. A. FALLAH MORTEZANEJAD[1], G. MOHTASHAMI BORZADARAN[2] AND B. SADEGHPOUR GILDEH[3]

[1,2,3]*Department of Statistics, Ferdowsi University of Mashhad, P. O. Box 1159, Mashhad 91775, Iran;*
azadeh.falllah@mail.um.ac.ir; grmohtashami@um.ac.ir; sadeghpour@um.ac.ir

Abstract. The main problem when working with multivariate data sets is how to find the multivariate distribution of data that maintains the principal dependency between the variables. Based on the information we have, the copula function guarantees the dependence on the result distribution function. When there is no other basic information about the statistical structure, this method alone does not meet the needs. For this purpose, we use the concept of maximum entropy to deal with this problem. In this paper, we first consider the distribution of data sets. Data is extracted from a production process that simultaneously controls several different features of a product. Then we obtained an elliptical control range through the maximum entropy applying T^2-Hotelling statistic.

Keywords: Maximum entropy, Copula function, Spearman's rho, Blest measure, Control chart, T^2-Hotelling statistic.

1. Introduction

Shannon entropy has been introduced first by [9] since 1948. Afterward, it is used in many different fields. The maximum entropy principle has been presented by [4] since 1957. Jaynes exerted Lagrange function according to some constraints to find the distribution of maximum entropy. Such paper as [6], [11], and [5] have studied more on the maximum entropy concept. After that, it used wieldy by authors until recent years, and some of them can be mentioned as [1], [3], [12].
The maximum entropy principle is a good manner to find the unknown distribution of a univariate data set, because it does not need any strong presumption about distribution and work well with ill-posed conditions which are not required large sample sizes. Although all benefits of the maximum entropy concept, it can be difficult for some researchers to define some more constraints for multivariate data

set to preserve the original dependency between different variables of multivariate data, and a specialist needs to save it in the result distribution function. Some papers like [2], [14], [8], [10], [13], and [7] have made a link between the maximum entropy principle and copula function. Generally, by the aim of both concepts, we can get a copula density function by a maximum copula entropy just by adding some simple constraints according to intended dependency measures. So, the maximum copula function has the same dependency on the existing data. Finally, the Sklar theorem helps easily to find out the multivariate distribution function whose dependency is the same as the available data.

In this paper, we would like to peruse on manufacturing process data where there are many processes with multivariate data set with unknown distributions which are assumed normal distribution. This assumption is incorrect in general cases. So, technical assistants need to know the distribution. In this regard, the main point is to transfer the basic dependency to the result density function. Thus, we are working on this issue to combine the maximum entropy principle and copula function. As we mentioned before, the maximum entropy principle is applied to find the empirical multivariate distribution and the copula function cares about the dependency. Our predestinate data is bivariate and also dependence, so we estimated its distribution by the maximum entropy principle for some simulated dependency measures which are based on Spearman's rho, Blest measures. In the next step, we apply the T^2-Hotelling statistic which is common to use while dealing with multivariate data set. Afterward, we compute the statistical quality control for these kinds of data. These control limits are reliable because the dependency is paid attention while calculating them.

The proceeds of this paper are: in section 2, first of all, we explain the procedure of finding a bivariate maximum entropy distribution respect to some intended constraints, then we clarify the way of obtaining the maximum copula entropy according to corresponding constraints. In both of them, Shannon entropy is used. In the process of acquiring the maximum copula entropy, we apply some dependence measures to transfer the dependence of an available data set to the final maximum copula function. In the following section, we exert the maximum copula entropy to get the joint density function of the data set using the Sklar theorem, in section 3, we represent the T^2-Hotelling statistic and illustrate how to find the statistical control limits for bivariate data set with its original dependency which save with the maximum copula function in its joint density function, in section 4, we calculate the coefficients of the maximum copula entropy for some instance values of dependence measures whose surface plots are represented in some figures, then we estimate the upper control limit for some different means, in section 5, we make a conclusion and statements of the paper.

2. Joint distribution function via maximum copula entropy method

In this section, we would like to present a feasible method of finding multivariate distribution. For simplicity of calculation and notation, we discuss the bivariate data set. In this manner, we use the maximum entropy principle which has many advantages like being unbias, suitable for small sample size, no need for strong summation, etc. The maximum entropy concept is an applied way to find the unknown distribution of a real data set. It gets a compatible distribution for available information. There is the main question while working with a multivariate data set whose distribution is unknown that how it can be found the distribution with the same original dependency between corresponding variables. Copula function replies to the question as well. Here we suppose to mix these two major concepts to estimate a fitted distribution. In the following, we describe how to find the maximum copula entropy based on Shannon entropy. By the way, we will make a density function based on the maximum copula entropy. [9] introduced Shannon entropy as:

$$\mathcal{H}_S(h) = \int\int_{\mathbb{S}(X,Y)} -\log h(X,Y)dH(X,Y),$$

which is Shannon entropy of random variables X and Y whose density and distribution function are $h(X,Y)$ and $H(X,Y)$ respectively and $\mathbb{S}(X,Y)$ is the joint support set. [4] has introduced the principle of maximum entropy. Then, [6] has extended some constraints on this principle. To find the maximum entropy distribution, some intended constraints are needed as well:

$$\begin{cases} \int\int_{\mathbb{S}(X,Y)} dH(X,Y) = 1, \\ E(g_i(X,Y)) = m_i(x,y), \;\; j = 1,\ldots,k, \end{cases}$$

where $m_i(x,y)$s for $j = 1,\ldots,k$ are some known moments which are calculated based on available data set, $g_i(X,Y)$s for $j = 1,\ldots,k$ are corresponding functions to m_is, k is the number of constraints on moments, and $dH(X,Y)$ is the full differential of $H(X,Y)$. Then the maximum entropy distribution is gotten by applying the lagrange function made of the Shannon entropy and its corresponding constraints as well:

$$\begin{aligned} L(h,\lambda_0,\ldots,\lambda_r) &= -\int_{\mathbb{S}(X,Y)} \log h(X,Y)dH(X,Y) - \lambda_0\{\int_{\mathbb{S}(X,Y)} dH(X,Y) - 1\} \\ &- \Sigma_{i=1}^{k}\lambda_i\{\int_{\mathbb{S}(X,Y)} g_i(X,Y)dH(X,Y) - m_i(x,y)\}. \end{aligned}$$

Then the Lagrange function should be differentiated respect to $h(.)$ and by using the Kuhn-Tucker method the maximum entropy distribution is found out:

$$h(X,Y) = \exp(-\lambda_0 - \Sigma_{i=1}^{k}\lambda_i g_i(X,Y)), \;\; (X,Y) \in \mathbb{S}(X,Y).$$

Now, we are keen on representing how to find the maximum copula entropy via entropy principle as well. First of all, the copula entropy based on the Shannon definition is:

$$\mathcal{H}_S(c) = \int\int_{I^2} -c(u,v)\log c(u,v)dudv,$$

where

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}.$$

The maximum copula entropy has to be found out based on some constraints which ensure the result function to be copula. These essential constraints according to [2] are:

$$\begin{cases} \int\int_{I^2} c(u,v)dudv = 1, \\ \int\int_{I^2} u^r c(u,v)dudv = \frac{1}{r+1}, \\ \int\int_{I^2} v^r c(u,v)dudv = \frac{1}{r+1}, \end{cases}$$

where r is the counter of constraints and the bigger choice of r, the more accurate creature of the result function compared with copula functions. Here, we would like to add some other equations based on some measures of dependence that they should be estimated while dealing with real data set to get a copula function which has the same dependency as the available data set:

$$\begin{cases} \int\int_{I^2} uv\ c(u,v)dudv = \frac{\rho+3}{12}, \\ \int\int_{I^2} u^2v\ c(u,v)dudv = \frac{2\rho-\nu_1+2}{12}, \\ \int\int_{I^2} u^2v^2\ c(u,v)dudv = \frac{\eta+\frac{1}{5}}{6}, \end{cases} \tag{2.1}$$

where ρ, ν_1, and η are Spearman's rho, Blest measure I and III respectively. It is worth to mention that the value of (2.1) is Blest measure II. To find the maximum copula entropy, we have to apply the Lagrange function and Kuhn-Tucker method as well and the result function is:

$$\begin{aligned} c'(u,v) = \exp\big(&-1-\lambda_0-\Sigma_{r=1}^{n}\lambda_r(u^r+v^r) \\ &-\lambda_{n+1}uv-\lambda_{n+2}(u^2v+uv^2)-\lambda_{n+4}u^2v^2\big)\ ,\ \forall u,v\in[0,1]. \end{aligned}$$

The values of λs are gotten by applying $c'(u,v)$ in the intended constraints and a system of equations has to be solved. In practice, measures of dependence exerted in the constraints have to be estimated, because the copula function is required in their computations. In section 4, we are going to find some copula function under some estimated dependence measures. By the way, after finding the copula density functions related to the dependence measures, their joint density functions can be obtained by this formula:

$$f'_{X,Y}(x,y) = c'(u,v)f'_X(x)f'_Y(y), \tag{2.2}$$

where $f'_X(\cdot)$ and $f'_Y(\cdot)$ are marginal functions gotten by the maximum entropy principle based on Shannon definition. So, the joint density function of a dependence data set is gotten via the maximum entropy and copula function. The maximum entropy principle is applied, because it is the best choice when enough information is not available, and it can help us to find a fitted distribution while there is not sufficient information or the sample size is not large enough. The copula function keeps the original dependency between variables of the data set as well. Thus, the result of the joint density function is reasonable for our goal.

3. Elliptical control chart using T^2-Hotelling

In the previous section, we had some estimated dependence measures and based on them, we got the unknown joint density function of a data set. Our goal is to work on data sets obtained from a manufacturing process and we would like to control the process by the time. To control a production process, we need the statistical control limits $[LCL, UCL]$. These limits straightly depend on the joint density function of the process which is usually unknown in practice. The purpose is to use the density function (2.2) to compute the suitable control limits. For this aim, we suppose to apply T^2-Hotelling statistic to deal with multivariate data set, but to find the proper control limits, we need the joint density function represented in (2.2) as well. First of all, we preset the T^2-Hotelling statistic for a random vector $\underset{\sim}{X}$ with mean vector $\underset{\sim}{\mu}$ and the matrix variance-covariance matrix Σ as:

$$T^2_{Hotelling} = (\underset{\sim}{X}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}).$$

In our case of study, we have:

$$\begin{cases} \underset{\sim}{X} = \begin{pmatrix} X \\ Y \end{pmatrix}, \\ \underset{\sim}{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \\ \Sigma^{-1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \end{cases}$$

where $a_{12} = a_{21}$. It is obvious that T^2-Hotelling is a positive statistic and the corresponding LCL is 0. So, we have to solve this equation to get the UCL:

$$P(T^2_{Hotelling} \leqslant UCL) = 1 - \alpha,$$

where α is the level of confidence. Then, we have:

$$\begin{aligned} 1 - \alpha &= P((\underset{\sim}{X} - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}) \leqslant UCL) \\ &= P(a_{11}(X - \mu_X)^2 + a_{22}(Y - \mu_Y)^2 + 2a_{12}(X - \mu_X)(Y - \mu_Y) \leqslant UCL) \\ &= \int\int_{\{(x,y)|a_{11}(X-\mu_X)^2 + a_{22}(Y-\mu_Y)^2 + 2a_{12}(X-\mu_X)(Y-\mu_Y) \leqslant UCL\}} f'_{X,Y}(x,y)dx\ dy. \end{aligned}$$

We need the value of UCL which satisfied the last equation. These control limits are based on the dependency of two variables X and Y whose dependence reflects on $f'_{X,Y}(\cdot,\cdot)$. In the next section, we are going to use this method to find the statistical control limits for a simulation study.

4. Practical example of a manufacturing process

In many studies consisting of numerical data, the main questions are how to find out the distribution of the data set which can be univariate or multivariate, but in almost all researches, the distribution of the existing data set is unknown and should be estimated via a statistical method. The entropy concept is well known and used in many different fields of study such as Mathematics, Physics, Computer Science, Economics, etc. The maximum entropy principle is a statistical method to find the best distribution dealing with inadequate information. Moreover, it acts acceptable with small sample sizes as well. Some intended constraints are required for maximum entropy method which is based on such available information as moments. So, no strong assumptions are needed which is another benefit of this method. While dealing with the univariate data set, it is easy to use the entropy procedure, and we are not worried about the loss of dependency between variables. An important question is how intended constraints have to be defined to keep the original dependency between multivariate data set, or which kinds of constraints guarantee the original dependency in the result distribution function. One way to reply to these kinds of questions is by using the copula function. So we have to find a copula function with the same dependency as the data set has. To do this, we use the maximum entropy to get a copula function named the maximum copula entropy. We define some constraints under some dependency measures of data. Then, the result copula function has the same dependency on the available data set. Just by using the Sklar theorem, the joint distribution of the data set can be obtained with the same dependency.

In this paper, we introduced a feasible way to find out the distribution of an available data set by applying the maximum copula entropy. Afterward, we get statistical control limits by exerting the joint density function, so the control limits are based on the original dependency between variables. Thus, the decision according to the limits is reliable. In the following, we calculate some coefficients of $c'(u,v)$ respect to some value of dependence measures which are represented in Table 1 as well as their corresponding surface plots in Figure 1. We use these copula functions to get the joint density functions of some samples with different means. In Tables 2 and 3, there are three groups with $\mu_X = 2,\ 3,\ 7$ and $\mu_Y = 1,\ 5,\ 6$. These means are applied to find the marginal functions of X and Y. The marginal functions are estimated via a univariate maximum entropy method.

TABLE 1. Coefficients of the maximum entropy

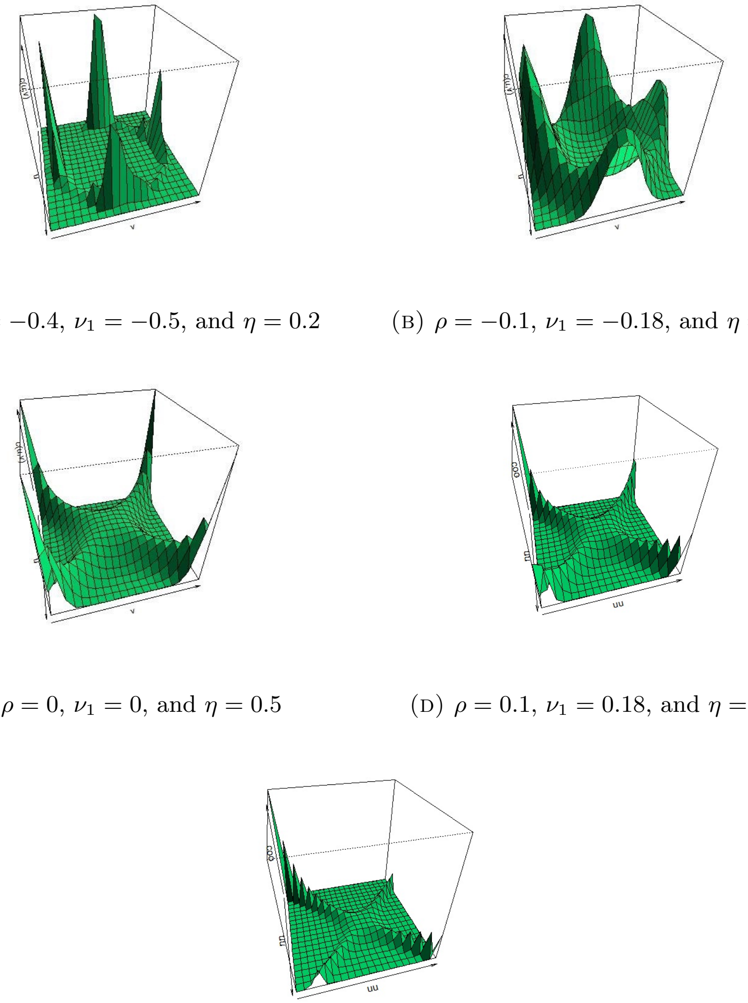
λs	ρ =-0.4 $\nu_1 = -0.5$ η =0.2	$\rho = -0.1$ ν_1 =-0.18 η =0.45	ρ =0 $\nu_1 = 0$ η =0.5	ρ =0.1 $\nu_1 = 0.18$ η =0.55	$\rho = 0.4$ $\nu_1 = 0.5$ $\eta = 0.8$
λ_0	407.1356	30.132266	-3.1513221	-3.846356	-4.821063
λ_1	-2177.948	-55.72411	7.6161238	8.759351	5.748566
λ_2	4386.5280	296.108893	172.0824301	536.398532	2708.47968
λ_3	-3672.824	-293.5333	-358.67218	-1043.037	-4764.8248
λ_4	429.5542	149.434091	178.0929512	479.692338	2069.49012
λ_5	973.9363	-7.548779	0.8739453	19.045948	20.27722
λ_6	6847.6116	479.951059	-397.02907	-1137.668	-5447.5555
λ_7	-6611.443	-441.8196	396.8804704	1104.027338	4800.81162
λ_8	6394.9248	407.919066	-396.68008	-1071.220	-4230.9714

TABLE 2. UCL with confidence level of $1-\alpha$ for some means and different measures of dependence whose copula coefficients are in Table 1. The first type of error is approximating 0.05 respectively to each case.

Groups	$\mu_x = 2,\ \mu_y = 1$		$\mu_x = 3,\ \mu_y = 5$		$\mu_x = 7,\ \mu_y = 6$	
Measure of dependence	$1-\alpha$	UCL	$1-\alpha$	UCL	$1-\alpha$	UCL
ρ =-0.4, ν_1 =-0.5, η =0.2	0.950221	2.57147	0.95025	2.55468	0.950093	2.66865
ρ =-0.1, ν_1 =-0.18, η =0.45	0.950009	6.65817	0.950427	2.55568	0.950271	6.76327
ρ =0, ν_1 =0, η =0.5	0.950084	8.51353	0.950045	8.50172	0.950129	8.49472
ρ =0.1, ν_1 =0.18, η =0.55	0.950169	8.6212	0.950074	8.59518	0.950285	8.60911
ρ =0.4, ν_1 =0.5, η =0.8	0.950216	8.11819	0.950027	8.04674	0.950587	8.17985

5. CONCLUSION

In manufacturing processes, several procedures release a multivariate data set. They reflect the quality of some different product specifications. In the statistical quality control, the main goal is to monitor such data, but their distribution is unknown, so it is difficult to define fitting control limits to the process. In this paper, we find a joint density function and then get suitable control limits. In this regard, we apply T^2-Hotelling statistics which is used while dealing with multivariate data



(A) $\rho = -0.4$, $\nu_1 = -0.5$, and $\eta = 0.2$

(B) $\rho = -0.1$, $\nu_1 = -0.18$, and $\eta = 0.45$

(C) $\rho = 0$, $\nu_1 = 0$, and $\eta = 0.5$

(D) $\rho = 0.1$, $\nu_1 = 0.18$, and $\eta = 0.55$

(E) $\rho = 0.4$, $\nu_1 = 0.5$, and $\eta = 0.8$

FIGURE 1. Surface plot of Copula density function for Table 1

to combine the information. It is common to estimate T^2-Hotelling distribution via Fisher distribution when the data set has a normal distribution, but it is not true in general. By the aim of this paper, we estimate T^2-Hotelling distribution through the maximum copula entropy.

In the end, we add simulation study and find some copula function based on some dependency measures such as Spearman's rho and Blest. The goal is to get the unknown distribution of a manufacturing process data that is multivariate and the variables are dependence. Then, we would like to find statistical quality control limits according to this distribution for all data, because in some quality control

Table 3. UCL with confidence level of $1-\alpha$ for some means and different measures of dependence whose copula coefficients are in Table 1. The first type of error is approximating 0.0027 respectively to each case.

Groups	$\mu_x = 2,\ \mu_y = 1$		$\mu_x = 3,\ \mu_y = 5$		$\mu_x = 7,\ \mu_y = 6$	
Measure of dependence	$1-\alpha$	UCL	$1-\alpha$	UCL	$1-\alpha$	UCL
$\rho = -0.4,\ \nu_1 = -0.5,\ \eta = 0.2$	0.997322	3.75579	0.997381	3.65014	0.99734	3.72115
$\rho = -0.1,\ \nu_1 = -0.18,\ \eta = 0.45$	0.99732	19.3664	0.997448	3.64308	0.997444	30.9215
$\rho = 0,\ \nu_1 = 0,\ \eta = 0.5$	0.997357	34.8288	0.997477	36.5968	0.997323	34.7383
$\rho = 0.1,\ \nu_1 = 0.18,\ \eta = 0.55$	0.997441	35.8291	0.997513	34.6084	0.997447	34.6711
$\rho = 0.4,\ \nu_1 = 0.5,\ \eta = 0.8$	0.997354	33.6949	0.99731	34.2322	0.997385	34.1226

methods, the dependency of variables is not paid attention. The T^2-Hotelling statistic is applied for this aim, and then statistical control limits are obtained exerting the maximum copula entropy as well.

References

1. Cesari, A., Reier, S., Bussi, G. (2018). Using the maximum entropy principle to combine simulations and solution experiments. Computation, 6(1), 15.
2. Chu, B. (2011). Recovering copulas from limited information and an application to asset allocation. Journal of Banking and Finance, 35(7), 1824-1842.
3. Fallah Mortezanejad, S. A., Borzadaran, G. M., Gildeh, B. S. (2019). An entropic structure in capability indices. Communications in Statistics-Theory and Methods, 1-11.
4. Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical review, 106(4), 620.
5. Johnson, R., Shore, J. (1983). Comments on and correction to" Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy"(Jan 80 26-37)[Corresp.]. IEEE transactions on Information Theory, 29(6), 942-943.
6. Kagan, A. M., Linnik, Y. V., Rao, C. R. (1973). Extension of darmois-skitcvic theorem to functions of random variables satisfying an addition theorem. Communications in Statistics-Theory and Methods, 1(5), 471-474.
7. Mortezanejad, S. A. F., Borzadaran, G. M., sadeghpour Gildeh, B. (2019). Joint dependence distribution of data set using optimizing Tsallis copula entropy. Physica A: Statistical Mechanics and its Applications, 121897.
8. J. Piantadosi, P. Howlett, J. Borwein, Copulas with maximum entropy, Optimization Letters 6 (1) (2012) 99–125.
9. Shannon, C. E. (1948). A mathematical theory of communication. Bell system technical journal, 27(3), 379-423.
10. Singh, V. P., Zhang, L. (2018). Copulaentropy theory for multivariate stochastic modeling in water engineering. Geoscience Letters, 5(1), 6.
11. Shore, J., Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on information theory, 26(1), 26-37.
12. Sutter, T., Sutter, D., Esfahani, P. M., Lygeros, J. (2019). Generalized maximum entropy estimation. Journal of Machine Learning Research, 20, 138.
13. Rahmani Shamsi, J., Dolati, A. (2018). Rank based Least-squares Independent Component Analysis. Journal of Statistical Research of Iran JSRI, 14(2), 247-266.
14. N. Zhao, W. T. Lin, A copula entropy approach to correlation measurement at the country level, Applied Mathematics and Computation 218 (2) (2011) 628–642.