entropy

MDPI

Article

Estimating Multivariate Discrete Distributions Using Bernstein Copulas †

Victor Fossaluza *, Luís Gustavo Esteves and Carlos Alberto de Bragança Pereira

Institute of Mathematics and Statistics, Universidade de São Paulo, São Paulo, SP 05508-090, Brazil; lesteves@ime.usp.br (L.G.E.); cadebp@gmail.com (C.A.d.B.P.)

* Correspondence: victor.ime@gmail.com; Tel.: +55-11-3091-6187

† This paper is an extended version of our conference paper presented at the 37th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2017), Jarinu/SP, Brazil, 9–14 July 2017.

Received: 18 December 2017; Accepted: 7 March 2018; Published: 14 March 2018

Abstract: Measuring the dependence between random variables is one of the most fundamental problems in statistics, and therefore, determining the joint distribution of the relevant variables is crucial. Copulas have recently become an important tool for properly inferring the joint distribution of the variables of interest. Although many studies have addressed the case of continuous variables, few studies have focused on treating discrete variables. This paper presents a nonparametric approach to the estimation of joint discrete distributions with bounded support using copulas and Bernstein polynomials. We present an application in real obsessive-compulsive disorder data.

Keywords: Bernstein polynomial; copula; nonparametric inference; Aitchison's distance

1. Introduction

The association between random variables is a subject of interest in many scientific fields. The most complete method of characterizing the association between random variables is to determine the joint distribution of these random variables. Multivariate density functions, for absolutely continuous variables, and multivariate probability mass functions, for discrete variables, have become the focus of researchers interested in evaluating such associations (see, for example, [1–4]).

The motivation for the present paper was a study performed as part of the Obsessive-Compulsive Spectrum Disorder Program of the Institute of Psychiatry, University of São Paulo Medical School. A group of 1001 consecutive adult outpatients diagnosed with primary obsessive-compulsive disorder (OCD) according to the DSM-IV criteria [5] were recruited, and some of these patients were submitted to psychiatric treatment. Their OCD severity was evaluated using the Yale-Brown Scale (YBOCS; [6,7]) at the beginning of the project. At the time when the data records were accessed, only 213 patients participated in the re-evaluation using the same scale. The YBOCS is composed of two sub-scales, obsession (O) and compulsion (C), and each sub-scale assumes values in the set of integers $\{0, 1, \ldots, 20\}$. To measure the OCD severity of the patients, we considered the maximum value between the O and C sub-scale measures, max{O;C}; this method of scoring is known as the M-YBOCS scale (see the discussions in [8,9]).

Figure 1 presents the initial (X) and final (Y) M-YBOCS scores for all 213 patients for whom both initial and final scores were obtained. In this graphic, darker colors and larger dots represent higher cell frequencies. Our first objective is to estimate the marginal distributions of the initial and final scores. For this purpose, all available information should be used: all 1001 patients included in the first evaluation and all 213 remaining patients at the end of the study. If we use only the complete pairs of observation, omitting missing marginal values, we obtain only 213 pairs of measurements to

be used in the estimation of the joint distribution of interest, the support of which possibly contains $441 (= 21^2)$ points, nearly double the sample size. As a consequence, standard methods of estimation (like maximum likelihood) would unavoidably yield estimates equal to zero for most cell probabilities. It is then reasonable to consider the whole dataset (including the available incomplete pairs) in order to improve such estimates.

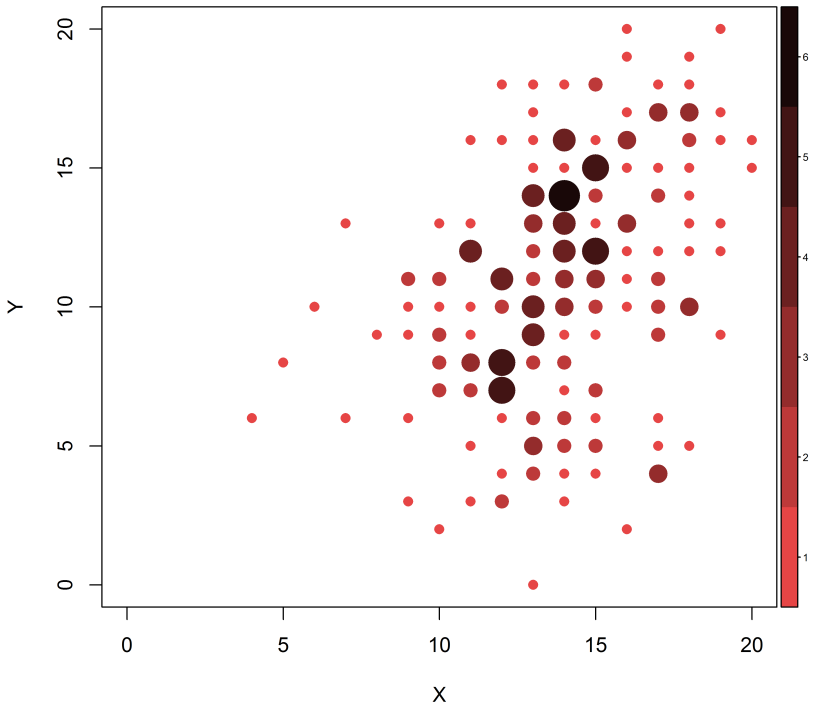

Figure 1. Frequencies of each observed OCD severity before and after treatment.

The objective of the present paper is to introduce a method of estimating multivariate discrete probability mass functions in the presence of (marginal) missing data. For this purpose, we developed an estimation method that uses both empirical distribution functions and Bernstein polynomials. The procedure consists of estimating a smooth joint distribution function, followed by applying a method that transforms this function into a discrete function, i.e., the estimated joint probability mass function. The results of this new method are compared with those of alternative methods, both graphically and by evaluating standard distances.

Section 2 describes the existing methods found in the literature that will be considered for comparison. Section 3 describes our estimator for the joint probability functions. Section 4 presents a discussion of the new method and comparisons of this method to the alternative methods using both simulated samples and the real OCD example. Finally, in Section 5, we present our final comments and considerations for future work.

2. Existing Solutions

First, we introduce the mathematical framework for our problem. Let F be the unknown distribution function of a random vector $\boldsymbol{X}$ that takes values in a subset of $\mathbb{R}^p$. A sample of size n of $\boldsymbol{X}$ is represented by $\boldsymbol{X_1}, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X_i} = (X_{i1}, \ldots, X_{ip})$ and $i = 1, \ldots, n$. In other words, the $\boldsymbol{X_i}$'s are conditionally independent and identically distributed random variables, given any distribution function F. Observations of $\boldsymbol{X_i}$ are denoted by $\boldsymbol{x_i}$.

Assuming that the distribution F is drawn from a known family of distributions, we represent the statistical model by $(\mathcal{X}, \mathcal{F}, \mathcal{P})$, where $\mathcal{X}$ is the sample space, $\mathcal{F}$ is a sigma-algebra of its subsets

and $\mathcal{P} = \{P(\cdot|\theta) : \theta \in \Theta\}$ is a family of distributions indexed by the parameter θ that belongs to the parameter space Θ. The estimation of F is then reduced to that of the parameter θ, and the dependence structure is limited to that supported by the underlying statistical model. For many years, the multivariate normal distribution has been used for most multivariate analyses (see, for example, [10,11]). Recently, for many random phenomena whose distributions are skewed and possess heavier tails than those of the normal distribution, alternative distributions, such as multivariate skew-elliptical distributions, have been adopted [12,13].

In recent approaches, copulas have become a popular tool for modeling multivariate dependence structures and for obtaining new multivariate distributions with given marginals. In short, a copula is a multivariate distribution whose marginals are uniform over the entire range $[0, 1]$. There are many parametric families of copulas, allowing for the modeling of many different dependence structures [1–4]. Let F be a p-dimensional distribution function with the margins $F_1, \ldots, F_p$. Sklar [14] first showed that there exists a p-dimensional copula C such that:

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$$

for all $\boldsymbol{x} = (x_1, \ldots, x_p)$ in the domain of F. If the variables $X_1, \ldots, X_p$ are absolutely continuous, then the copula C is unique; otherwise, C is uniquely determined on $Ran(F_1) \times \ldots \times Ran(F_p)$, where $Ran(F_i)$ is the image of the function F_i, $i = 1, \ldots, p$ [14]. Thus, the copula can be used to separately model the margins and the dependence structure. The non-unique representation of a copula for discrete distributions is a theoretical issue that must be considered in the context of an analytical proof, but this does not limit its empirical applications [15]. However, the above theorem [14] does not tell us how to find the copula C. This problem is widely discussed in the literature, and several solutions to this problem have been proposed (see, for example, [16]). The most widely-used approach is to adjust several families of (parametric) copulas and choose one of them using certain selection criteria or a goodness-of-fit test [17–21].

Nonparametric techniques may also be applied to estimate a multivariate distribution. A popular solution using this approach is the application of the empirical distribution function $F^{(n)} : \mathbb{R}^p \to [0, 1]$, which is defined, for $(t_1, \ldots, t_p) \in \mathbb{R}^p$, as:

$$F^{(n)}(t_1, \ldots, t_p) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{x_{1i} \leq t_1, \ldots, x_{pi} \leq t_p\right\}, \tag{1}$$

where $\mathbb{I}\{A\}$ is the indicator of the set A. This approach is equivalent to using the relative frequencies to estimate the joint probability mass function. The relative frequencies coincide with the maximum-likelihood estimate under the assumption that the data are drawn from a multinomial distribution. One shortcoming of such approaches is that the probability of any non-observed cells will be estimated to be zero.

Another possible approach is to use some function to smooth the empirical distribution. We can consider the Bernstein polynomials [22,23] for this purpose because of their simplicity and good mathematical properties [24,25]. Let $h : [0, 1]^p \to \mathbb{R}$ be a continuous function. The m^{th}-degree (multivariate) Bernstein polynomial for the function h, namely, $B_h^m : [0, 1]^p \to \mathbb{R}$, is defined as:

$$B_h^m(x_1, \ldots, x_p) = \sum_{j_1=0}^{m} \cdots \sum_{j_p=0}^{m} h\left(\frac{j_1}{m}, \ldots, \frac{j_p}{m}\right) \prod_{i=1}^{p} \binom{m}{j_i} x_i^{j_i} (1 - x_i)^{m - j_i}. \tag{2}$$

The multivariate Bernstein polynomials for the function h converge uniformly to the function h as $m \to \infty$ [26,27], and its derivatives are simple to obtain. The function h must be defined in $[0, 1]^p$, and therefore, for practical purposes, data that do not take values in $[0, 1]^p$ must first be transformed [24]. To apply this method to the OCD data, for example, we consider the transformation $Y = X/20$. Moreover, the polynomial degree adopted here is $m = n/log(n)$, as suggested by [24]. Bernstein

polynomials have been used to approximate a copula C by simply replacing the function h with the copula. The resulting Bernstein polynomial, B_C^m, which is also a copula that strongly converges to C, is called a Bernstein copula [28–32]. When the true copula is unknown, the empirical copula can be used instead, and the resulting function is called the empirical Bernstein copula [16,33–37]. The empirical copula is defined as:

$$C_n(u_1, \ldots, u_p) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{F_1(x_{1i}) \leq u_1, \ldots, F_p(x_{pi}) \leq u_p\right\}.$$

Note that even when F_i, $i \in 1, \ldots, n$, is unknown, we can use the empirical marginal distribution $F_i^{(n)}$ as a consistent estimator of F_i, according to the Glivenko–Cantelli theorem (e.g., [38]). Other estimators for marginal distributions could be considered instead, as in the procedure proposed in the next section.

We have so far obtained a continuous function as an estimate while our objective is clearly to estimate a (discrete) probability mass function. Hence, this function must be discretized to obtain an adequate estimate. This can be achieved as follows: suppose, with no loss of generality, that $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a random vector such that all its components X_i, $i = 1, 2, \ldots, p$, assume values in the set $\Omega = \{0, 1, \ldots, k\}$ with probability one. In addition, there always exists a continuous random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ with distribution function F such that $P(0 \leq Z_i \leq k) = 1$, $i = 1, \ldots, p$, and $X_i = \sum_{j=0}^{k} j\, \mathbb{I}\{j - 0.5 < Z_i \leq j + 0.5\}$, i. It follows that:

$$P\left(X_1 = x_1, \ldots, X_p = x_p\right) = P\left(x_1 - 0.5 < Z_1 \leq x_1 + 0.5, \ldots, x_p - 0.5 < Z_p \leq x_p + 0.5\right).$$

Let F (or an estimate $\hat{F}$) be the continuous joint distribution function of the random vector $\boldsymbol{Z}$, and let $B = [\boldsymbol{a}, \boldsymbol{b}] = [a_1, b_1] \times \ldots \times [a_p, b_p]$ be a p-dimensional rectangle with all its vertices in Ω. The F-volume of B [4] is then given by:

$$V_F(B) = \sum_{\boldsymbol{c}} sgn(\boldsymbol{c}) F(\boldsymbol{c}), \tag{3}$$

where the sum is taken over all vertices $\boldsymbol{c} = (c_1, \ldots, c_p)$ of B, and $sgn(\boldsymbol{c})$ is given by:

$$sgn(\boldsymbol{c}) = \begin{cases} 1, & \text{if } c_j = a_j \text{ for an even number of } j\text{'s}, \\ -1, & \text{if } c_j = a_j \text{ for an odd number of } j\text{'s}. \end{cases}$$

In particular, suppose $\boldsymbol{b} = (b_1, \ldots, b_p) \in \{0, 1, \ldots, k\}^p$, and take $B = [\boldsymbol{b} - \frac{1}{2}\,\mathbf{1}, \boldsymbol{b} + \frac{1}{2}\,\mathbf{1}] = [b_1 - 0.5, b_1 + 0.5] \times [b_2 - 0.5, b_2 + 0.5] \times \ldots \times [b_p - 0.5, b_p + 0.5]$, with $b_i \in \Omega$, $\forall i = 1, \ldots, p$, then the probability of the event $\{\boldsymbol{X} = \boldsymbol{b}\} = \left\{X_1 = b_1, \ldots, X_p = b_p\right\}$ can be calculated (estimated) as:

$$P(\boldsymbol{X} = \boldsymbol{b}) = P(b_1 - 0.5 < Z_1 \leq b_1 + 0.5, \ldots, b_p - 0.5 < Z_p \leq b_p + 0.5]) = V_F(B).$$

Because weak convergence occurs at the points of continuity of the limiting distribution function F and because our goal is to estimate a discrete probability mass function, we consider sets of the form $B = [\boldsymbol{b} - \frac{1}{2}\,\mathbf{1}, \boldsymbol{b} + \frac{1}{2}\,\mathbf{1}]$, with $b_i \in \Omega$, so that the vertices of the p-dimensional rectangle B are always points of continuity of the distribution function of the discrete random vector $\boldsymbol{X}$. Thus, such discretization yields satisfactory estimates for the probability mass function of $\boldsymbol{X}$.

3. Proposed Solutions

Our proposed method for estimating the joint distribution of a discrete random vector consists of using Bernstein polynomials to estimate both the marginals and the copula. The advantage of this method is that it allows all observations to be used, even in the case of missing values in some

variable. Furthermore, this method is a nonparametric approach, and there are few restrictions on the dependence structure.

First, for each random variable X_i, we estimate the marginal distributions using the empirical marginal distribution with n_i observations, $F_i^{(n_i)}(x) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{I}(x_{ij} \leq x)$, $i = 1, \ldots, p$; then, the Bernstein polynomial of degree $m_i = n_i/log(n_i)$ is used to smooth this function:

$$B_i^{m_i}(x) = \sum_{j=1}^{m_i} F_i^{(n_i)}\left(\frac{j}{m_i}\right)\binom{m_i}{j}x^j(1-x)^{(m_i-j)}.$$

As this estimator converges to the marginal distribution [24], we estimate the copula using an alternative version of the empirical copula based on the n complete observations and the estimates $B_i^{m_i}$, $i = 1, \ldots, p$,

$$C_n(u_1, \ldots, u_p) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{B_1^{m_1}(x_{1j}) \leq u_1, \ldots, B_p^{m_p}(x_{pj}) \leq u_p\right\},$$

and smooth this function to obtain the corresponding empirical Bernstein copula,

$$B_{C_n}^m(u_1, \ldots, u_p) = \sum_{j_1=0}^{m}\cdots\sum_{j_p=0}^{m} C_n\left(\frac{j_1}{m}, \ldots, \frac{j_p}{m}\right)\prod_{i=1}^{p}\binom{m}{j_i}x_i^{j_i}(1-x_i)^{m-j_i}.$$

Note that the construction of the copula C_n using Bernstein polynomials rather than empirical (marginal) distribution functions yields, at least in the examples to be presented in Section 4, non-zero estimates for non-observed cells. This feature justifies the choice of this alternative version of the empirical copula.

The estimate of the joint distribution function is a discretization (Equation (3)) of the following function:

$$\hat{F}_{m,n}(x_1, \ldots, x_p) = B_{C_n}^m\left(B_1^m(x_1), \ldots, B_p^m(x_p)\right). \tag{4}$$

The algorithm used to obtain the proposed solution is quite simple and is summarized below:

1. for all n_i observations of each variable X_i, estimate the marginal empirical distribution function $F_i^{(n_i)}$;
2. smooth each function $F_i^{(n_i)}$ using a Bernstein polynomial $B_i^{(m_i)}$ of degree m_i;
3. for all complete observations of the random vector $\boldsymbol{X}$, estimate the empirical copula C_n;
4. estimate the Bernstein copula by smoothing the empirical copula C_n using the m^{th}-degree multivariate Bernstein polynomial $B_{C_n}^m$;
5. obtain a continuous estimate of the multivariate distribution function $\hat{F}_{m,n}$ given by Equation (4);
6. discretize $\hat{F}_{m,n}$ using Equation (3) to obtain an estimate of the discrete multivariate probability mass function.

4. Applications

To evaluate the robustness of the method, we simulated datasets from two bivariate discrete distributions generated using copulas (Examples 4.1 and 4.2). For each simulated example, we present the estimated probabilities for three cases:

1. 600 pairs of observations with no censored data;
2. censored data in only one marginal, with 1000 observations in one marginal and 200 in another; and
3. censored data in both variables, with 600 observations for each variable, 300 of which form complete pairs.

After these examples, we present and compare estimates to the observed data from the OCD study (Example 4.3).

For the examples described above, we present the estimates for the probability mass functions considering the proposed method and compare its performance with some existing solutions, similar to those briefly discussed in Section 2 and which are more detailed below:

a. the empirical distribution presented in Equation (1), that is obtained using only the complete pairs and the resulting probability function, coincides with the relative frequencies of the points observed in the sample;
b. the multivariate skew t approximation that is obtained through a discretization of a parametric multivariate continuous distribution, estimated by the maximum likelihood method using only the complete pairs;
c. the discretization of the normal copula with the normal marginal approximation to the distribution function that is obtained by using all observations for marginal distribution estimation and using only the complete pairs for copula estimation. This method is quite similar to that described in (b) considering the normal multivariate distribution rather than the skew t distribution, but here, it is possible to estimate the marginal distributions using all available data, not just the complete pairs;
d. the discretization of the empirical Bernstein polynomial approximation presented in Equation (2), replacing the function h by the empirical distribution $F^{(n)}$ obtained in (a) and using only the complete pairs; and
e. our proposed solution described in the previous section, which is obtained by using the Bernstein polynomial to approximate the margins using all observations and the approximated copula using the complete pairs.

For all examples, we graphically illustrate the estimates of the probability mass distributions and evaluate several distances between the estimated and theoretical distributions. For this purpose, some notation must be introduced. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ be the theoretical probabilities, and let $\hat{\boldsymbol{\theta}} = (\hat{\theta_1}, \ldots, \hat{\theta_k})$ be the estimated probabilities. We consider the following distances for comparison of the estimates:

i. Aitchison's distance:

$$\Delta(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sqrt{\sum_{i=1}^{k} \left[\ln\left(\frac{\hat{\theta}_i}{\theta_i}\right) - \bar{L} \right]^2}, \text{ where } \bar{L} = \frac{1}{k} \sum_{i=1}^{k} \ln\left(\frac{\hat{\theta}_i}{\theta_i}\right)$$

ii. Euclidean distance:

$$\delta(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sqrt{\sum_{i=1}^{k} \left[\hat{\theta}_i - \theta_i\right]^2}$$

iii. Total variation distance:

$$\tau(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{k} \left|\hat{\theta}_i - \theta_i\right|$$

iv. Kullback–Leibler symmetrized divergence:

$$\mathcal{D}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{2} \left[\sum_{i=1}^{k} \theta_i \ln\left(\frac{\theta_i}{\hat{\theta}_i}\right) + \sum_{i=1}^{k} \hat{\theta}_i \ln\left(\frac{\hat{\theta}_i}{\theta_i}\right) \right]$$

Aitchison [39,40] and Pawlowsky [41] have presented many arguments for using Aitchison's distance for compositional vectors, that is when the sum of the vector's components is constant (in our case, the sum of the probabilities is equal to one). Moreover, the orderings implied by these distances agree in most cases.

At the end of this section, we present the estimates for the distribution of the real data described in the Introduction. In this case, we do not know the theoretical distribution; we present only the estimates and the distances calculated from the empirical distribution.

4.1. Simulated Elliptically-Shaped Distribution

In this section, we simulate data from an elliptically-shaped distribution with marginals $X_1 \sim$ beta-binomial $(N_x = 20, \alpha = 5, \beta = 5)$ and $Y_1 \sim$ binomial $(N_y = 20, \pi = 0.5)$ and a normal copula with parameter $\rho = 0.7$.

We can see from Figures 2–4 and from Tables 1–3 that in these examples, the solutions based on elliptical distributions, namely the skew t and normal distributions, yield better estimates. This superior estimation occurs because the theoretical probability mass function is elliptical in shape. However, in practical situations, we have no knowledge of the real shape of the distribution. In such a case, the empirical distribution may be a good basis for evaluating the estimates, despite the existence of many unobserved points that are estimated as zero. When the estimates are compared with the empirical distribution, our proposed solution appears to produce good results, particularly in the presence of censored data.

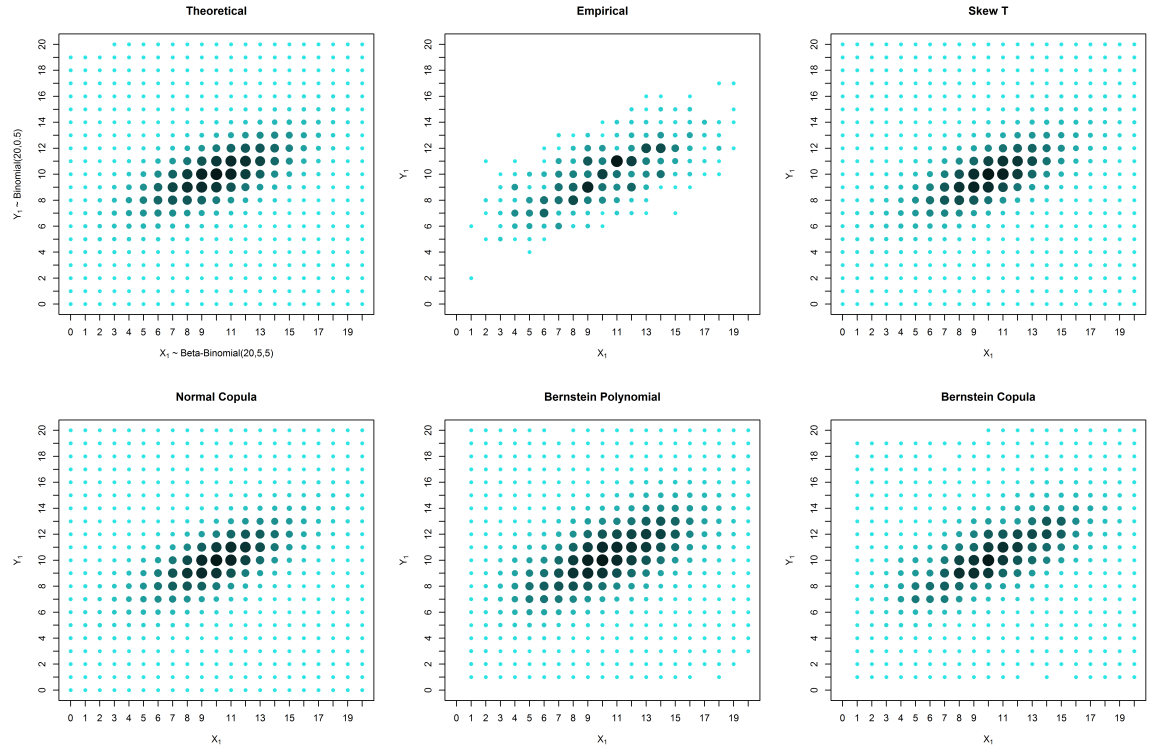

Figure 2. Estimates and theoretical probabilities for 600 complete pairs of observations, simulated from an elliptically-shaped distribution.

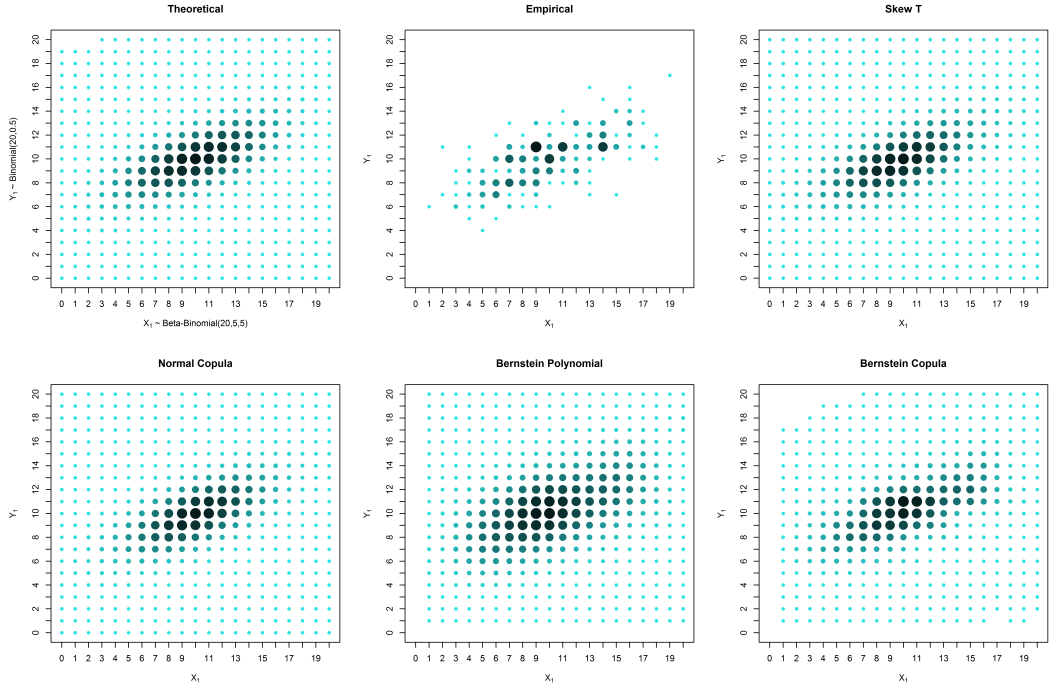

Figure 3. Estimates and theoretical probabilities for the case of censored data in only one marginal, with 1000 observations in one marginal and 200 in the other, simulated from an elliptically-shaped distribution.

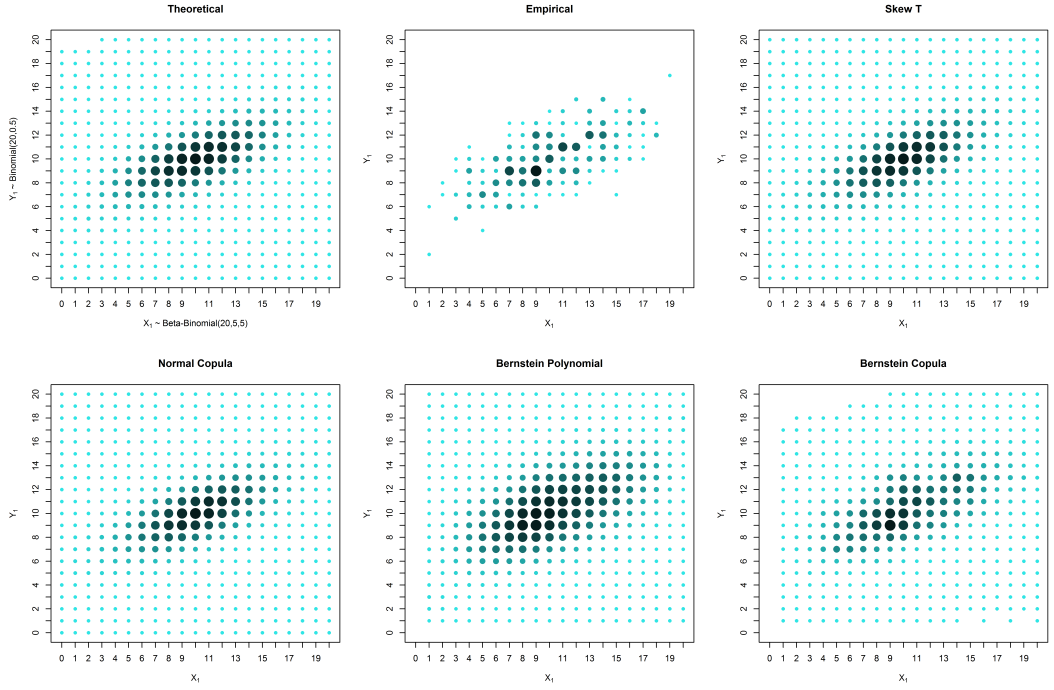

Figure 4. Estimates and theoretical probabilities for the case of censored data in both variables, with 600 observations for each variable, of which 300 form complete pairs, simulated from an elliptically-shaped distribution.

Table 1. Distances between the estimates and theoretical probabilities for 600 complete pairs of observations. The bold values highlight the smaller distances.

Example 4.1.1	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Empirical	4.98521	0.02116	0.09988	0.04154
Skew T	1.44499	0.00629	0.02915	0.00345
Normal Copula	**1.28402**	**0.00476**	**0.02418**	**0.00236**
Bernstein Polynomial	3.45943	0.01388	0.07159	0.01870
Bernstein Copula	3.23712	0.01217	0.06360	0.01578

Table 2. Distances between the estimates and theoretical probabilities for the case of censored data in only one marginal, with 1000 observations in one marginal and 200 in the other. The bold values highlight the smaller distances.

Example 4.1.2	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Empirical	8.97909	0.03454	0.17083	0.12490
Skew T	1.28441	**0.00493**	**0.02291**	**0.00239**
Normal Copula	**1.28040**	0.00554	0.02738	0.00284
Bernstein Polynomial	4.84901	0.02049	0.11110	0.03982
Bernstein Copula	3.28689	0.01171	0.06340	0.01530

Table 3. Distances between the estimates and theoretical probabilities for the case of censored data in both variables, with 600 observations for each variable, of which 300 form complete pairs. The bold values highlight the smaller distances.

Example 4.1.3	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Empirical	7.32955	0.03035	0.13826	0.09162
Skew T	1.12383	0.00419	0.02221	0.00185
Normal Copula	**1.06365**	**0.00375**	**0.01891**	**0.00146**
Bernstein Polynomial	4.54073	0.01934	0.10051	0.03531
Bernstein Copula	3.54526	0.01377	0.06743	0.01963

4.2. Simulated Asymmetrical Distribution

In this section, we present the simulated data for an asymmetrical distribution with margins $X_2 \sim$ beta-binomial $(N_x = 20, \alpha = 0.85, \beta = 1.1)$ and $Y_2 \sim$ binomial $(N_y = 15, \pi = 0.6)$ and a Gumbel copula with the parameter $\theta = 0.7$.

In the case of an asymmetrical distribution, our proposed solution yields a better estimation in all three considered cases: the case with no censored data, the case with missing data in one variable and the case with missing data in both variables. The superior performance of our approach can be observed both graphically (Figures 5–7) and from the calculated distances (Tables 4–6). It is possible to graphically observe that the probabilities of both the smaller and larger values of X_2 are well estimated by our method.

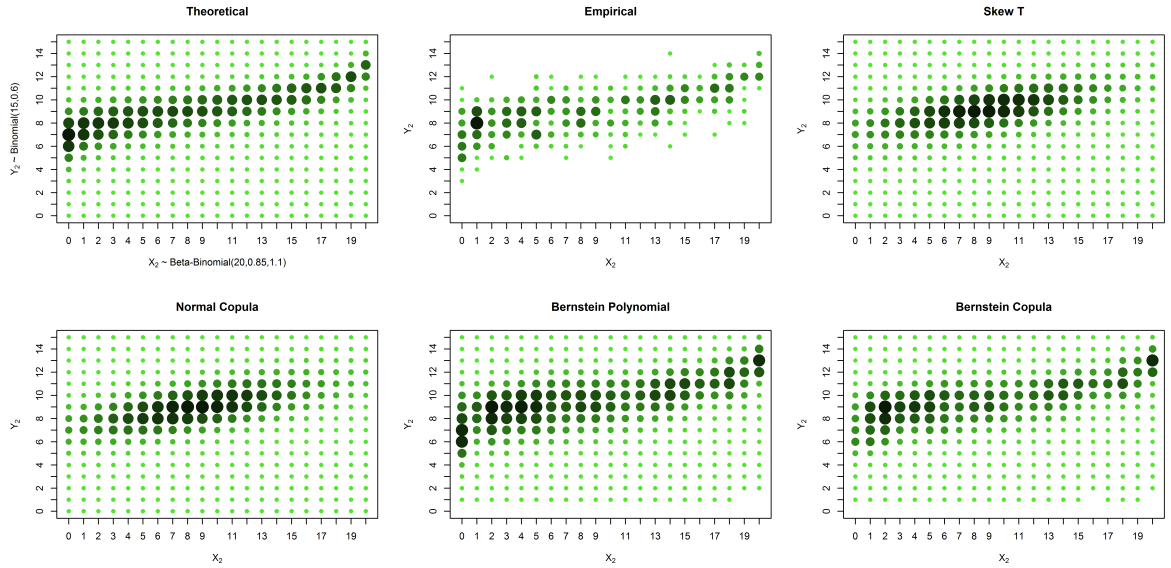

Figure 5. Estimates and theoretical probabilities for 600 complete pairs of observations, simulated from an asymmetrical distribution.

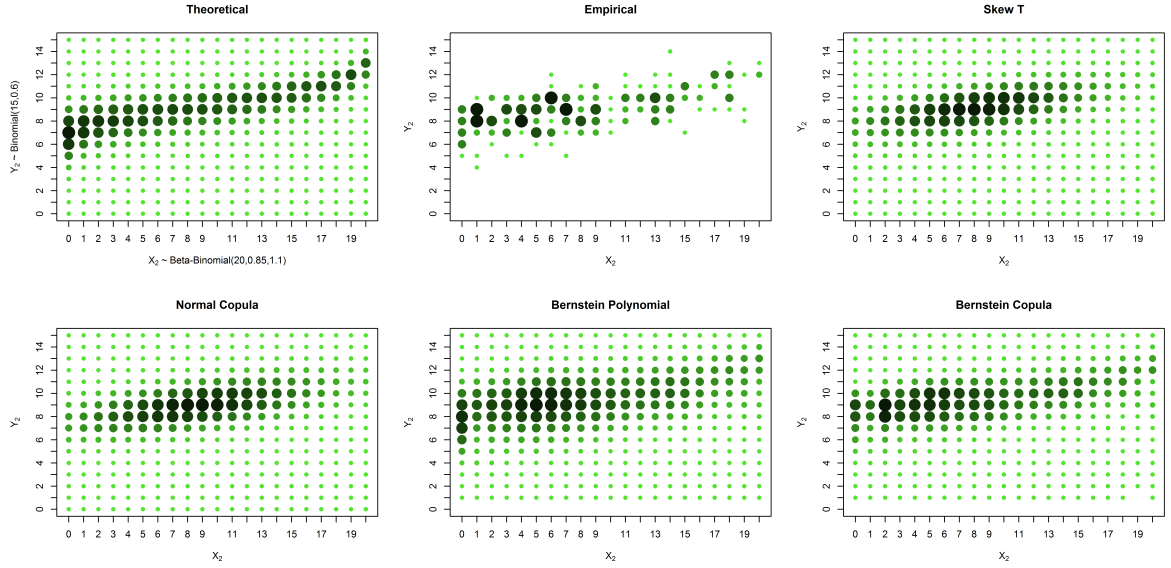

Figure 6. Estimates and theoretical probabilities for the case of censored data in only one marginal, with 1000 observations in one marginal and 200 in the other, simulated from an asymmetrical distribution.

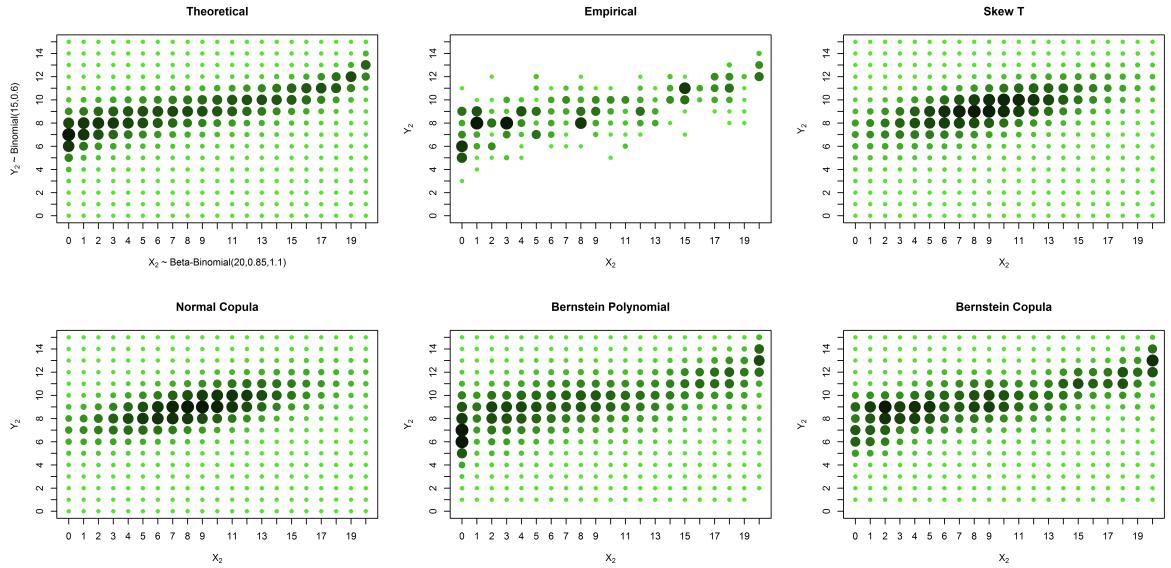

Figure 7. Estimates and theoretical probabilities for the case of censored data in both variables, with 600 observations for each variable, of which 300 form complete pairs, simulated from an asymmetrical distribution.

Table 4. Distances between the estimates and theoretical probabilities for 600 complete pairs of observations. The bold values highlight the smaller distances.

Example 4.2.1	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Empirical	6.17032	0.02767	0.12761	0.06377
Skew T	5.47625	0.03287	0.14429	0.07724
Normal Copula	5.76598	0.03293	0.14785	0.08020
Bernstein Polynomial	5.41325	**0.02436**	0.11969	0.05380
Bernstein Copula	**5.07634**	0.02519	**0.11842**	**0.05068**

Table 5. Distances between the estimates and theoretical probabilities for the case of censored data in only one marginal, with 1000 observations in one marginal and 200 in the other. The bold values highlight the smaller distances.

Example 4.2.2	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Empirical	8.77534	0.03906	0.19104	0.14363
Skew T	5.23773	0.03130	0.13494	0.07356
Normal Copula	5.07437	0.02892	0.12727	0.06549
Bernstein Polynomial	5.65580	0.02626	0.13135	0.06562
Bernstein Copula	**4.86027**	**0.02558**	**0.11172**	**0.05379**

Table 6. Distances between the estimates and theoretical probabilities for the case of censored data in both variables, with 600 observations for each variable, of which 300 form complete pairs. The bold values highlight the smaller distances.

Example 4.2.3	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Empirical	7.32005	0.03284	0.15576	0.09786
Skew T	4.79233	0.02760	0.12321	0.05917
Normal Copula	5.06253	0.02819	0.12855	0.06325
Bernstein Polynomial	5.09522	0.02253	0.11486	0.05028
Bernstein Copula	**4.35547**	**0.01957**	**0.09863**	**0.03595**

The new method performed better than the other presented solutions in the case of asymmetric models. It should be emphasized that the method was developed for cases with a large proportion of censored data or situations in which the number of points to be estimated is larger than the sample size. Figure 8 shows the Aitchison distances between the theoretical distribution and the estimates of the example of Section 4.2 considering censored data in only one marginal. Sample sizes $n \in \{1, \ldots, 2000\}$ with the same proportion of censored data of the example were considered. In this case, the support of the distribution has 336 points. Note that for small samples ($n < 750$), the proposed method presents the smallest distances. Although the method was developed for small samples, it appears that its estimates converge to the theoretical distribution as n increases. However, a detailed investigation on the asymptotic properties of the new method is needed and is the goal of a future work.

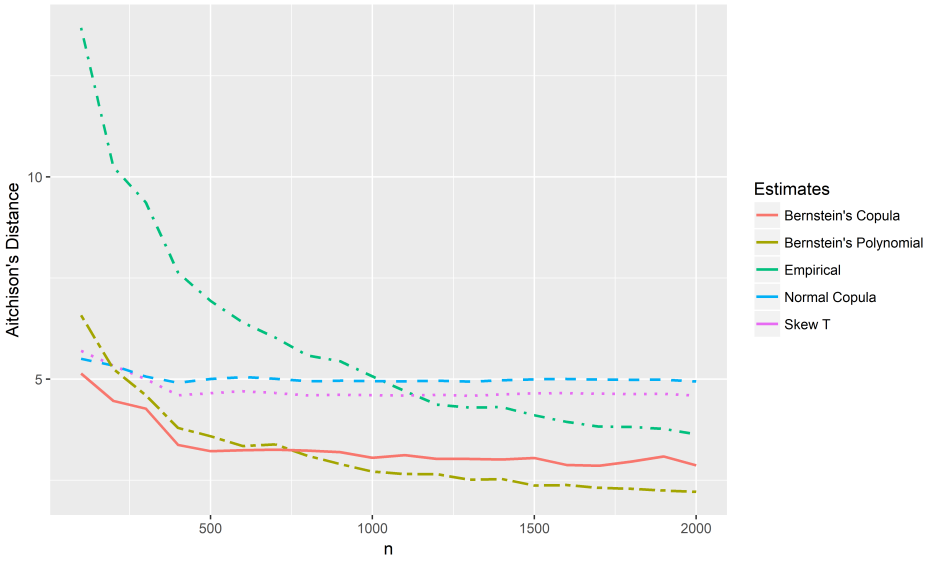

Figure 8. Distances between the estimates and theoretical probabilities for the case of censored data in only one marginal, with n observations in one marginal and 80% of censored data in the other one.

4.3. Real Data

In this section, we present the estimates for the real data described in the Introduction. The YBOCS is one of the most widely-used outcome measures in treatment studies of obsessive compulsive disorder (OCD). The total YBOCS scores comprise an integer number varying from 0–40 and intend to grade the severity of obsessive-compulsive symptoms. The total YBOCS score is the sum of two sub-scales, each ranging from 0–20, one of which measures the severity of compulsion and the other of obsession. The works in [8,9] propose that instead of the sum, it would be better to consider the maximum of these two sub-scales, called M-YBOCS. Thus, psychiatrists have been interested in better understanding the properties of this new scale, such as the probability distribution of M-YBOCS scores before and after patients have received some treatment for OCD.

As already mentioned in the Introduction, the dataset has 1001 observations of the scores at the initial time and only 213 at the final time. This happens because many patients drop out of treatment. The causes of drop out can be extremely different, such as a reduction in symptoms making the patient feel that he/she does not need treatment, or even worsen the symptoms, causing the patient to discredit the treatment. The small number of complete pairs in the database makes it difficult to estimate the joint distribution.

In real problems, there are few cases where the law of probability that generates the data is revealed. In such cases, a fairly common way to assess whether the proposed methods are adequate is to compare estimates with observed data. In predictive models, for example, it is common to verify some distance between predicted and observed values. In this way, we compare the distance between

the estimates and the empirical probability function (which is the relative frequency of each observed point). The proposed solution yields smaller distances than do the existing approaches (Table 7).

Table 7. Distances between the estimates and empirical probabilities for the real data.

Example 4.3	Aitchison	Euclidean	Total Variation	Kullback–Leibler
Skew T	11.56219	0.03941	0.22684	0.19899
Normal Copula	12.07092	0.04143	0.24701	0.21760
Bernstein	11.35361	0.03933	0.22910	0.19125
Bernstein Copula	**10.81475**	**0.03703**	**0.21020**	**0.17184**

The estimation through the empirical distribution presents many zeros due to the small number of observations. The researchers believe that the proportion of unobserved points would decrease if the sample had fewer dropouts. In addition, they believe that common assumptions of normality or even symmetry assumptions make no sense in this case. The proposed method assigns positive probability to non-observed cells and captures the asymmetric nature of the data, which can be observed graphically in Figure 9.

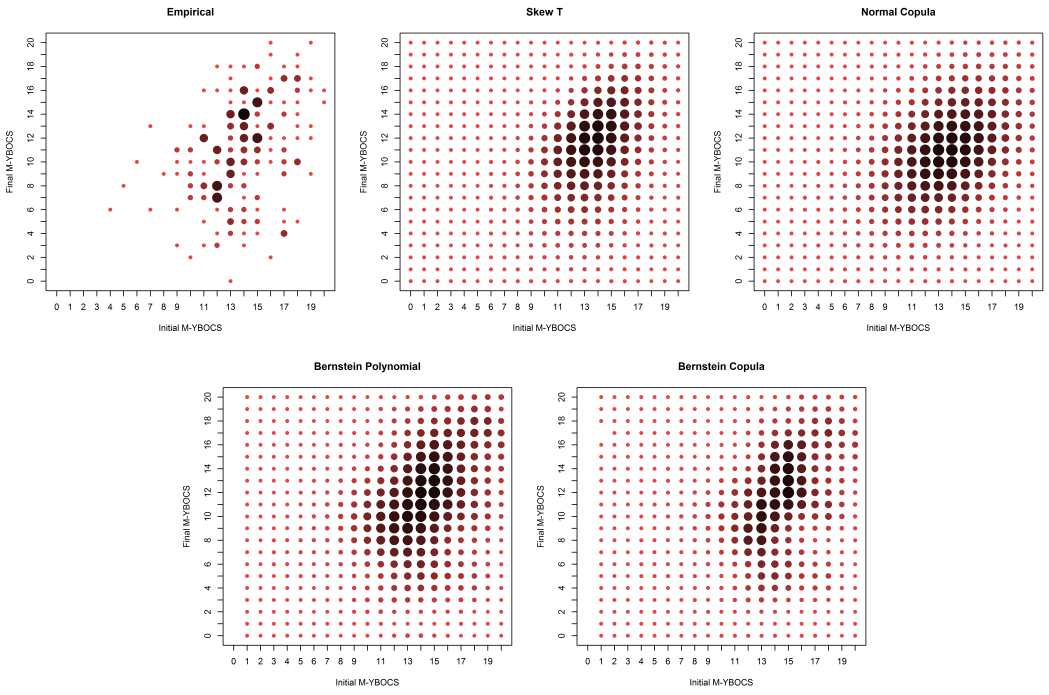

Figure 9. Estimates of probabilities for the real OCD data.

5. Conclusions

In this work, a new approach to the problem of estimating discrete bivariate distributions is presented. The procedure, which essentially consists of estimating both the marginals and the copula using Bernstein polynomials, aims at addressing three important issues: the handling of discrete bivariate data in the presence of marginal missing values (using all available information, including incomplete pairs of observations); the possibility of obtaining positive estimates for non-observed cells, thus yielding "smoother" estimated discrete distributions; and the consideration of a large variety of dependence structures between the relevant random variables. The new approach is suitable for these cases owing to its fairly unrestrictive, nonparametric nature. The use of Bernstein polynomials shows better results of the empirical distribution to estimate the marginal distribution. It is important

to note that the empirical Bernstein copula produces a copula only asymptotically [35], and other methods could be used to estimate Bernstein's copula instead, as the one described in [42]. Anyway, the proposed method showed reasonable estimates for the studied probability mass functions. The new method can be applied also to p-dimensional random variables, $p > 2$: both the mathematical development and the computational implementation are similar to the case $p = 2$.

The new method was applied to several examples of simulated data, and according to a few typical measures of distance (between the estimated and theoretical distributions), it performed better than some of the existing solutions in cases of asymmetrical models, particularly in the presence of censored data and for cases where the number of points to be estimated is larger than the sample size. Although the method was developed for small samples, it appears that the proposed estimates converge to the theoretical distribution, but more detailed studies are still needed.

The new method was also applied to data sampled from adults diagnosed with primary obsessive-compulsive disorder. The estimate obtained by the method was appreciated by researchers in psychiatry.

While the new method has practical advantages over the presented existing alternatives, some aspects were not addressed here, namely: it will yet be necessary to further develop the new procedure in several aspects that were not addressed here: (i) the study of asymptotic properties for large sample size n and/or for higher polynomial degree m; (ii) a formal justification for the new procedure under a decision-theoretical approach; (iii) the development of a more rational approach to the selection of $m, m_1, \ldots, m_p$ (which could depend on n) using the approach suggested in (ii); and (iv) the incorporation of prior knowledge, perhaps as in Petrone [43,44] and Petrone and Wasserman [45], although these authors approached the problem from a univariate Bayesian perspective. These topics will be the focus of future articles.

Acknowledgments: The authors gratefully acknowledge CAPES and CNPq (Grant No. 143240/2009-9) for their financial support and Rafael Izbicki, Adriano Polpo de Campos and Marcio Alves Diniz for their helpful comments.

Author Contributions: All the authors contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dos Anjos, U.U.; Ferreira, F.H.; Kolev, N.V.; Mendes, B.M.V. *Modeling Dependences via Copulas*; 16º SINAPE, Associação Brasileira de Estatística: São Paulo, Brazil, 2004. (In Portuguese)
2. Joe, H. *Multivariate Models and Dependence Concepts*; Chapman & Hall/CRC: London, UK, 1997.
3. Joe, H. *Dependence Modeling with Copulas*; CRC Press: Boca Raton, FL, USA, 2014.
4. Nelsen, R.B. *An Introduction to Copulas*, 2nd ed.; Springer Verlag: New York, NY, USA, 2006.
5. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*; American Psychiatric Association: Washington, DC, USA, 1994.
6. Goodman, W.K.; Price, L.H.; Rasmussen, S.A.; Mazure, C.; Fleischmann, R.L.; Hill, C.L.; Heninger, G.R.; Charney, D.S. The Yale-Brown Obsessive-Compulsive Scale: I. Development, use, and reliability. *Arch. Gen. Psychiatry* **1989**, *46*, 1006.
7. Goodman, W.K.; Price, L.H.; Rasmussen, S.A.; Mazure, C.; Delgado, P.; Heninger, G.R.; Charney, D.S. The Yale-Brown Obsessive-Compulsive Scale: II. Validity. *Arch. Gen. Psychiatry* **1989**, *46*, 1012.
8. Pereira, C.A.B.; Silva, C.B.; Diniz, J.B. Estatística em Psiquiatria. In *Clínica Psiquiátrica*; Miguel, E.C., Gentil, V., Gattaz, W.F., Eds.; Editora Manole: São Paulo, Brazil, 2011; p. 147.
9. Diniz, J.B.; Fossaluza, V.; Belotto-Silva, C.; Shavitt, R.G.; Pereira, C.A.B. The use of Yale-Brown Obsessive-Compulsive Scale: New views of an old measure. *Eur. Neuropsychopharmacol.* **2011**, *21*, S531–S532.
10. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*; Prentice Hall: Upper Saddle River, NJ, USA, 2002; Volume 4.
11. Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*; Academic Press: London, UK, 1980.

12. Branco, M.D.; Dey, D.K. A General Class of Multivariate Skew-Elliptical Distributions. *J. Multivar. Anal.* **2001**, *79*, 99–113.
13. Genton, M.G.; Loperfido, N.M.R. Generalized skew-elliptical distributions and their quadratic forms. *Ann. Inst. Stat. Math.* **2005**, *57*, 389–401.
14. Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statistique Univ. Paris* **1959**, *8*, 229–231.
15. Trivedi, P.K.; Zimmer, D.M. *Copula Modeling: An Introduction for Practitioners*; Foundations and Trends in Econometrics; Now Publishers: Hanover, MA, USA, 2007; Volume 1.
16. Durrleman, V.; Nikeghbali, A.; Roncalli, T. *Which Copula Is the Right One*; Technical Report; Groupe de Recherche Opérationnelle, Crédit Lyonnais: Lyon, France, 2000.
17. Fermanian, J.D. Goodness-of-fit tests for copulas. *J. Multivar. Anal.* **2005**, *95*, 119–152.
18. Genest, C.; Quessy, J.F.; Rémillard, B. Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Stat.* **2006**, *33*, 337–366.
19. Genest, C.; Rémillard, B.; Beaudoin, D. Goodness-of-fit tests for copulas: A review and a power study. *Insur. Math. Econ.* **2009**, *44*, 199–213.
20. Rakonczai, P.; Zempléni, A. Copulas and goodness of fit tests. In *Recent Advances in Stochastic Modeling and Data Analysis*; Skiadas, C.H., Ed.; World Scientific Publishing: Singapore, 2007; pp. 198–206.
21. Berg, D. Copula goodness-of-fit testing: An overview and power comparison. *Eur. J. Finance* **2009**, *15*, 675–701.
22. DeVore, R.A.; Lorentz, G.G. *Constructive Approximation*; Springer: New York, NY, USA, 1993.
23. Lorentz, G.G. *Bernstein Polynomials*, 2nd ed.; Chelsea Pub Co.: New York, NY, USA, 1986.
24. Babu, G.J.; Canty, A.J.; Chaubey, Y.P. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Stat. Plan. Inference* **2002**, *105*, 377–392.
25. Babu, G.J.; Chaubey, Y.P. Smooth estimation of a distribution and density function on a hypercube using Bernstein polynomials for dependent random vectors. *Stat. Probab. Lett.* **2006**, *76*, 959–969.
26. Heitzinger, C.; Hössinger, A.; Selberherr, S. On smoothing three-dimensional Monte Carlo ion implantation simulation results. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2003**, *22*, 879–883.
27. Heitzinger, C.; Hössinger, A.; Selberherr, S. An algorithm for smoothing three-dimensional Monte Carlo ion implantation simulation results. *Math. Comput. Simul.* **2004**, *66*, 219–230.
28. Li, X.; Mikusiński, P.; Sherwood, H.; Taylor, M. *Distributions with Given Marginals and Moment Problems*; Chapter On Approximation of Copulas; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1997; pp. 107–116.
29. Li, X.; Mikusiński, P.; Taylor, M.D. Strong approximations of copulas. *J. Math. Anal. Appl.* **1998**, *255*, 608–623.
30. Kulpa, T. On approximations of copulas. *Int. J. Math. Math. Sci.* **1999**, *22*, 259–269.
31. Sancetta, A.; Satchell, S.E. Bernstein Approximations to the Copula Function and Portfolio Optimization. Cambridge Working Papers in Economics. 2001. Available online: https://ideas.repec.org/p/cam/camdae/0105.html (accessed on 8 March 2018).
32. Taylor, M.D. Bernstein polynomials and n-copulas. *arXiv* **2009**, arxiv:0903.1000.
33. Durrleman, V.; Nikeghbali, A.; Roncalli, T. *Copulas Approximation and New Families*; Technical Report; Groupe de Recherche Opérationnelle, Crédit Lyonnais: Lyon, France, 2000.
34. Sancetta, A. Nonparametric Estimation of Multivariate Distributions With Given Marginals. Cambridge Working Papers in Economics. 2004. Available online: https://www.repository.cam.ac.uk/handle/1810/352 (accessed on 8 March 2018).
35. Sancetta, A.; Satchell, S.E. The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econom. Theory* **2004**, *20*, 535–562.
36. Sancetta, A. Nonparametric estimation of distributions with given marginals via Bernstein–Kantorovich polynomials: L_1 and pointwise convergence theory. *J. Multivar. Anal.* **2007**, *98*, 1376–1390.
37. Bouezmarni, T.; Rombouts, J.V.K.; Taamouti, A. Asymptotic properties of the Bernstein density copula estimator for α-mixing data. *J. Multivar. Anal.* **2010**, *101*, 1–10.
38. Van der Vaart, A.W.; Wellner, J.A. *Weak cOnvergence and Empirical Processes*; Springer Verlag: New York, NY, USA, 1996.
39. Aitchison, J. *The Statistical Analysis of Compositional Data*; Blackburn Press: Caldwell, NJ, USA, 2003.

40. Aitchison, J. *The Single Principle of Compositional Data Analysis, Continuing Fallacies, Confusions and Misunderstandings and Some Suggested Remedies*; Technical Report; Department of Statistics, University of Glasgow: Glasgow, UK, 2008.
41. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. *Lecture Notes on Compositional Data Analysis*; Technical Report; Universitat de Girona: Girona, Spain, 2007.
42. Dou, X.; Kuriki, S.; Lin, G.D.; Richards, D. EM algorithms for estimating the Bernstein copula. *Comput. Stat. Data Anal.* **2016**, *93*, 228–245.
43. Petrone, S. Random Bernstein polynomials. *Scand. J. Stat.* **1999**, *26*, 373–393.
44. Petrone, S. Bayesian density estimation using Bernstein polynomials. *Can. J. Stat.* **1999**, *27*, 105–126.
45. Petrone, S.; Wasserman, L. Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc.* **2002**, *64*, 79–100.