Astron. Astrophys. 188, 258–264 (1987)

ASTRONOMY
AND
ASTROPHYSICS

A three-dimensional extended Kolmogorov-Smirnov test as a useful tool in astronomy

E. Gosset

Institut d'Astrophysique, Université de Liège, 5, avenue de Cointe, B-4200 Cointe-Ougrée, Belgium

Received November 17, 1986; accepted May 11, 1987

Summary. Using Monte Carlo techniques, we derive a three-dimensional version of the well-known Kolmogorov-Smirnov test; the extension is of the type described by Peacock (1983). Such a test is of great practical interest when one wishes to investigate the spatial distribution of a set of data points, particularly for the case of small size samples. A comparison with assumed three-dimensional density laws is made possible for most of current applications. A table of critical values of the new statistic is given for usual significance levels; empirical formulae to simulate the asymptotic behaviour are given as well. The three-dimensional extended two-sided Kolmogorov-Smirnov test is shown to be sufficiently distribution-free such that it can be widely and safely used.

Key words: data analysis – distribution – statistical test

1. Introduction

An everlasting source of enlightenment in astronomy is the study of the distribution of celestial objects. Information can be gained either about the physical properties of the objects themselves or about the structure and the characteristics of the underlying entity that involves them. The study of the distribution of quasars and galaxies provides important clues for our understanding of the physics and the history of the universe. On the opposite, the structure of the clustering of galaxies is an important fact when one wishes to investigate their formation process and consequently to understand their physical properties. The same is true at a smaller scale: some special groups of stars or related objects can indeed be used as tracers of the structure of the Galaxy. The investigated distribution is not necessarily purely spatial but parameter spaces can also be of great interest. Good examples are, for instance, periods of variable stars or shapes of their lightcurves. Using conventional cluster analysis, Mennessier (1985) has tentatively classified Mira variables; the discrimination is based on the distribution of the full amplitudes and asymmetries of the lightcurves relevant to different spectral regions. Many other examples could be cited.

All such problems reduce themselves to a common approach: a statistical study of a distribution of data points in a given space (a point process). Questions of concern are not only the search for the overall characteristics of the distribution from which the data points are drawn but also the intrinsic character of the parent population; is the latter uniformly random, contagious or regular (see the introduction of the paper by Gosset and Louis (1986))? The inverse problem is also considered: given a model for the parent distribution, e.g. *via* a theoretical density law, what is the chance that the observed data points could have arisen from the assumed law? To answer these questions, a variety of tests have been designed but of course most of them are limited to univariate analyses (one-dimensional case). Several new methods to study the distribution of points with respect to deviations from randomness towards clustering or regularity have recently appeared in the astronomical literature. The most interesting ones are: the Statistical Reduction of Population (SRP; see Zięba, 1975), the Multiple Binning Analysis (MBA, classic and/or with randomization tests; Gosset and Louis, 1986), the Nearest Neighbours Analysis (NNA; see Clark and Evans (1954), Thompson (1956), Rose (1977)), the Correlation Function Analysis (CFA; see Fall, 1979, for a review and Sharp, 1979, for some improvements), the Power Spectrum Analysis (PSA; Webster, 1976) and, finally, the Extended Kolmogorov-Smirnov test and the Generalized Power Spectrum Analysis (EKS and GPSA; Peacock, 1983). They all have their own characteristics and a thorough analysis must probably involve the majority of them simultaneously. Initially, the different tests have been designed for two-dimensional applications. The generalization to the third dimension is easy for some of them such as the MBA even with the randomization tests (Gosset and Louis, 1986) or the PSA (Webster, 1982); however, the same is not true for the EKS test. This is disappointing as it is a matter of common knowledge that the one-dimensional Kolmogorov-Smirnov test is very efficient. Its two-dimensional extension seems to have retained most of the attractive features of the original one (Peacock, 1983). This alone would suffice to justify its widespread use. On another side, the two-dimensional Extended Kolmogorov-Smirnov test is very pleasant to utilize as a first step before the application of the GPSA. All that is very encouraging about the necessity and the possibility to conceive the extension to three dimensions.

In this paper, we investigate such an EKS test. In Sect. 2, we recall some features of the one- and two-dimensional tests necessary to the understanding of the remainder of the paper. In Sect. 3, we introduce the three-dimensional version along with an algorithm to be used for the computation of the relevant statistic, whereas in Sect. 4 we deal with its distribution as derived from our simulations. A table of critical values for most usual significance levels and empirical formulae simulating the asymptotic behaviour are also given. In Sect. 5, we investigate the distribution-free character of the statistic.

As a draft of this paper was ready to be submitted, we became aware of a study by Fasano and Franceschini (1986) that

proposes to extend the classical Kolmogorov-Smirnov test in a completely different manner than us: this method is briefly analyzed and commented upon in Sect. 6. Conclusions form the last section.

2. The one- and two-dimensional extended Kolmogorov-Smirnov tests

The one-dimensional Kolmogorov-Smirnov test is the most important general test of goodness-of-fit besides the so-called Pearson χ^2. The latter has the disadvantage of requiring a binning of the data and some information can be lost in such a way. This is due to the discreteness but also to the non-recognition, or at least non-utilization, of the ordering of the sample. Let us define X to be a continuous random variable and the X_j's ($j = 1, n$) the relevant observations of an n points sample. Generality is preserved if we suppose that the X_j's are arranged in an increasing order. The assumed parent distribution corresponding to a null-hypothesis of randomness H_0 can be modelized by a hypothetical probability density function (pdf) $f_X(.)$ and we shall call

$$F_X^C(x) = \int_{-\infty}^{x} f_X(\xi)\, d\xi \tag{1}$$

the cumulative distribution function (cdf). In terms of probability, we obtain

$$F_X^C(x) = P(X \leq x) \tag{2}$$

where the right member denotes the probability of the event written in parentheses. We still have to define the sample distribution function (sdf)

$$\begin{cases} F_n^C(x) = 0 & \text{if } x < X_1\,, \\ F_n^C(x) = j/n & \text{if } X_j \leq x < X_{j+1} \text{ with } j = 1, n-1\,, \\ F_n^C(x) = 1 & \text{if } X_n \leq x\,. \end{cases} \tag{3}$$

As a matter of fact, this function is increased by $1/n$ each time a data point is passed through. It is clear that it can be considered as a tentative approach of the hypothetical cdf if $f_X(.)$ is the true pdf of the variable X. A test can be based on the greatest deviation between the two functions. The statistic

$$D_n^{(1)} = \sup_{\text{all } x} |F_n^C(x) - F_X^C(x)| \tag{4}$$

is the one adopted under the name of two-sided Kolmogorov-Smirnov. The distribution of $D_n^{(1)}$ has the advantage of being perfectly distribution-free, i.e. completely independent of $F_X^C(x)$ when the null-hypothesis holds, whereas the Pearson χ^2 is only asymptotically distribution-free. As $D_n^{(1)}$ is known to be roughly proportional to $1/\sqrt{n}$, it is preferable to use an alternative statistic

$$Z_n^{(1)} = \sqrt{n}\, D_n^{(1)}\,. \tag{5}$$

For more details on the Kolmogorov-Smirnov test, we refer the reader to Kendall and Stuart (1967) for some theory and to Birnbaum (1952) for a tabulation of critical values. We just mention that, for the asymptotic behaviour, we have

$$P(Z_\infty^{(1)} > z) = 2 \sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2i^2 z^2)$$
$$\simeq 2\exp(-2z^2) \quad \text{(for large } z\text{)}\,. \tag{6}$$

Peacock (1983) has shown that, if Z_n' symbolizes the critical value of Z_n for a given significance level, the quantity

$$\delta^{(1)} = 1 - Z_n^{(1)'}/Z_\infty^{(1)'} \tag{7}$$

can be useful in treating a small sample case in the same way as the asymptotic one. For most of the interesting significance levels, i.e. those in the range .20 to .01, we have within a good approximation

$$\delta^{(1)} = 0.20\, n^{-0.6} \tag{8}$$

provided that $n \gtrsim 5$ which is not restrictive at all in astronomy.

In conceiving the two-dimensional extended Kolmogorov-Smirnov test, Peacock (1983) has encountered two difficulties. The first one is the arbitrariness of the chosen direction when cumulating the data. In one dimension, this is not crucial since we have

$$P(X \leq x) + P(X > x) = 1\,, \tag{9}$$

or, similarly

$$F_X^C(x) + F_X^D(x) = 1\,, \tag{10}$$

where C and D denote cumulating in the X increasing and decreasing direction respectively. For the two-dimensional case, we obtain

$$P(X \leq x, Y \leq y) + P(X \leq x, Y > y) + P(X > x, Y \leq y)$$
$$+ P(X > x, Y > y) = 1\,, \tag{11}$$

and there are three independent ways to perform the cumulation. The procedure adopted by Peacock (1983) is to consider each of the four directions in turn (CC, CD, DC, DD) and to adopt the largest of the four differences

$$D_n^{(2)} = \max(D_n^{CC}, D_n^{CD}, D_n^{DC}, D_n^{DD})\,, \tag{12}$$

where, for example,

$$D_n^{CC} = \sup_{\substack{\text{all } x \\ \text{all } y}} |F_n^{CC}(x, y) - F_{X,Y}^{CC}(x, y)|\,, \tag{13}$$

with

$$F_{X,Y}^{CC}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(\xi, \eta)\, d\xi\, d\eta\,, \tag{14}$$

and $F_n^{CC}(x, y)$ is of course a function which is increased by $1/n$ each time a data point appears in the quadrant containing the points (ξ, η) such that $\xi \leq x$ and $\eta \leq y$. The equivalent expressions for the other functions are straightforward. Using Monte Carlo techniques, Peacock (1983) has studied the statistic

$$Z_n^{(2)} = \sqrt{n}\, D_n^{(2)}\,. \tag{15}$$

He has derived its distribution in the framework of a null-hypothesis of uniformity on a square and proposed an analytic expression for the asymptotic distribution:

$$P(Z_\infty^{(2)} > z) \simeq 2\exp(-2(z - 0.5)^2)\,, \tag{16}$$

which holds for significance levels less than 0.20. In order to convert $Z_n^{(2)'}$ to $Z_\infty^{(2)'}$ he has also obtained

$$\delta^{(2)} = 0.53\, n^{-0.9}\,. \tag{17}$$

We would like to point out here that, based on a combination of our own simulations and of the results by Peacock (1983), we

prefer for the asymptotic formula the following expression

$$P(Z_\infty^{(2)} > z) \simeq 2\exp(-2.5(z-0.63)^2). \quad (18)$$

The modification is not essential and would be purposeless if we would limit ourselves to the two-dimensional problem. Its interest will be underlined in Sect. 4.

The second problem encountered by Peacock (1983) concerns the distribution-free character of his test statistic. In the one-dimensional case, this property is a direct consequence of the fact that it is always possible to modify at will the pdf by applying any one to one transformation (preserving the ordering) without affecting the statistic. The same does not hold for the two-dimensional case. Nevertheless, Peacock (1983) has shown that in practice the statistic $Z_n^{(2)}$ turns out to be sufficiently distribution-free for most cases such that it is extremely useful.

A parenthesis ought to be opened here: astronomers usually do not realize that the estimation of a parameter from the sample can induce problems. Generally, the cdf is chosen as belonging to a parametric family of the type $F_X^C(x,\theta)$, where θ is an unknown parameter (not necessarily scalar). Instead of fixing it in advance, one usually prefers to derive from the observed sample a quantity denoted $\hat{\theta}$, the appropriate estimate of θ. By analogy with Eq. 4, one therefore has the statistic

$$\hat{D}_n^{(1)} = \sup_{\text{all } x} |F_n^C(x) - F_X^C(x,\hat{\theta})|, \quad (19)$$

the distribution of which is unknown. However, the tradition is to refer $\hat{D}_n^{(1)}$ simply to tables of $D_n^{(1)}$: such an approach can lead to a strongly conservative test and consequently to a loss of power (Noether, 1967). An example of such an application is available in Sect. 5 of Peacock's (1983) paper when he estimates, by a maximum likelihood method, the parameter α (the centre-to-edge variation in plate sensitivity).

3. The three-dimensional extended Kolmogorov-Smirnov test

3.1. The statistic $Z_n^{(3)}$

Encouraged by Peacock's results, we define the statistic

$$Z_n^{(3)} = \sqrt{n}\, D_n^{(3)}, \quad (20)$$

where

$$D_n^{(3)} = \max(D_n^{\text{CCC}}, D_n^{\text{CCD}}, D_n^{\text{CDC}}, D_n^{\text{CDD}}, D_n^{\text{DCC}}, D_n^{\text{DCD}}, D_n^{\text{DDC}}, D_n^{\text{DDD}}), \quad (21)$$

with, for example,

$$D_n^{\text{CCC}} = \sup_{\substack{\text{all } x \\ \text{all } y \\ \text{all } z}} |F_n^{\text{CCC}}(x,y,z) - F_{X,Y,Z}^{\text{CCC}}(x,y,z)|, \quad (22)$$

and

$$F_{X,Y,Z}^{\text{CCC}}(x,y,z) = \int_{-\infty}^{x}\int_{-\infty}^{y}\int_{-\infty}^{z} f_{X,Y,Z}(\xi,\eta,\mu)\, d\xi\, d\eta\, d\mu \quad (23)$$

and finally $F_n^{\text{CCC}}(x,y,z)$ which is a function increased by $1/n$ each time a data point appears in the three-dimensional octant containing the points (ξ,η,μ) such that $\xi \le x$ and $\eta \le y$ and $\mu \le z$. Let us now see how it is possible to compute the new statistic.

3.2. A comment on the computation of $Z_n^{(3)}$

It is sufficient to limit our investigation to the example of D_n^{CCC}, the transition to $Z_n^{(3)}$ being immediate. The computation of the cdf usually brings no problem; a grid can be computed in advance. The main challenge comes from the sdf. Another difficulty is the necessity to limit strongly the number of loci where the difference sdf-cdf has to be considered without any possibility to miss the true *supremum*. For the one-dimensional case, the cdf being a monotonic function and the sdf a strictly increasing step function, the *supremum* is necessarily located at one of the data points. The investigation is therefore limited to

$$x = X_i \quad \forall i \in [1,n]$$

i.e. performed n times. The two-dimensional case is slightly more complicated because the *supremum* needs not to be located on a data point. However, following similar considerations on the cdf and sdf, we can restrict ourselves to the loci

$$(x,y) = (X_i, Y_j) \quad \forall (i,j) \in [1,n] \times [1,n].$$

Peacock (1983) suggests to perform the search on all the n^2 possibilities which, to extract $Z_n^{(2)}$, will necessitate $4n^2$ computations. The same argument will lead for $Z_n^{(3)}$ to a number of $8n^3$. We believe that some of them are actually needless. We describe in Appendix A an algorithm which restricts the investigation to a smaller number of locations.

4. The distribution of the $Z_n^{(3)}$ statistic

Using Monte Carlo techniques, we have derived the distribution of the $Z_n^{(3)}$ statistic for a grid of values for n. Each $Z_n^{(3)}$ distribution is based on a minimum number of simulations which is never less than – and rarely equal to – 5000. Those results rely of course on constant density laws within a cube, i.e. uniform distributions of the simulated data points. The determined distributions are given in Fig. 1 whereas the critical values for the different n and for some significance levels of interest are included in Table 1.

From a comparison with the work of Peacock (1983), one can see that $Z_n^{(3)}$ is around 1.5 time farther from $Z_n^{(1)}$ than $Z_n^{(2)}$. This statement makes us confident in our results. The offset is practically constant except again that the $P(Z_\infty^{(3)} > z)$ function is slightly steeper. For significance levels of interest (between 0.20 and 0.01), the distribution of $Z_\infty^{(3)}$ (i.e. the asymptotic one) is well represented by

$$P(Z_\infty^{(3)} > z) \simeq 2\exp(-3(z-1.05)^2). \quad (24)$$

This formula is quite convincing when compared with the well-known distribution of $Z_\infty^{(1)}$ (see Eq. 6) and with the one of $Z_\infty^{(2)}$ as proposed above (see Eq. 18). We have illustrated the three curves in Fig. 2.

In the three-dimensional case, it is likewise possible to relate finite sample distributions to the asymptotic behaviour; $Z_n^{(3)}$ can be reduced to $Z_\infty^{(3)}$ by a simple scaling factor. Our simulations show that we have within a good accuracy,

$$\delta^{(3)} = 1 - Z_n^{(3)\prime}/Z_\infty^{(3)\prime} = 0.75\, n^{-0.9} \quad (25)$$

provided that n is greater than 5.

We wish to end by giving an approximate formula relating the size of the sample to the relevant critical value for a significance level of 0.01 and thus propose

$$z(|P(Z_n^{(3)} > z) = 0.01) = 2.38 - 2.1/n. \quad (26)$$

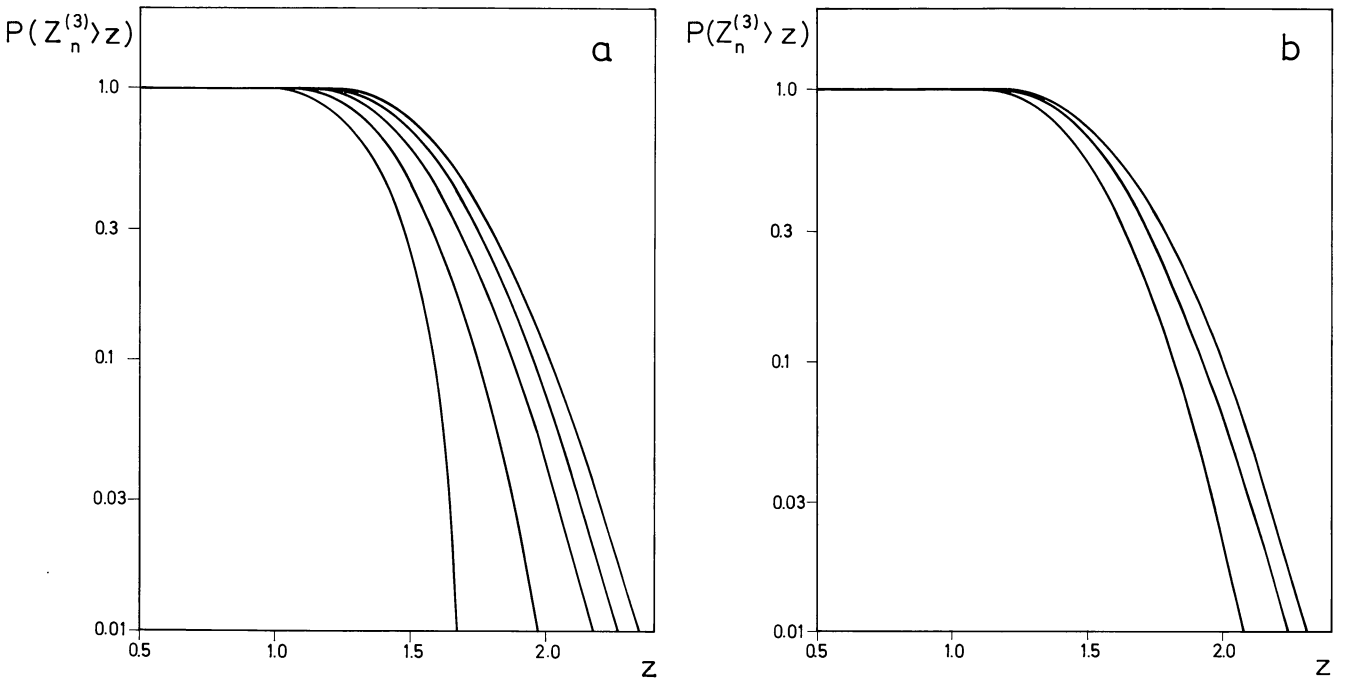

Fig. 1 a and b. Distributions of the $Z_n^{(3)}$ statistic as a function of n: **a** from left to right $n = 3, 5, 10, 20, 50$; **b** from left to right $n = 7, 15, 30$

Table 1. Critical values of $Z_n^{(3)}$ as a function of n and of the significance level.

	Significance levels				
n	0.20	0.15	0.10	0.05	0.01
3	1.53	1.56	1.59	1.63	1.68
5	1.65	1.70	1.74	1.83	1.97
7	1.71	1.76	1.81	1.91	2.08
10	1.76	1.82	1.87	1.98	2.17
15	1.81	1.87	1.92	2.04	2.24
20	1.83	1.90	1.95	2.07	2.27
30	1.86	1.92	1.98	2.10	2.31
50	1.89	1.95	2.01	2.13	2.34
100	1.91	1.97	2.03	2.15	2.36
∞[a]	1.93	1.99	2.05	2.18	2.38

[a] As simulated by our approximate formula

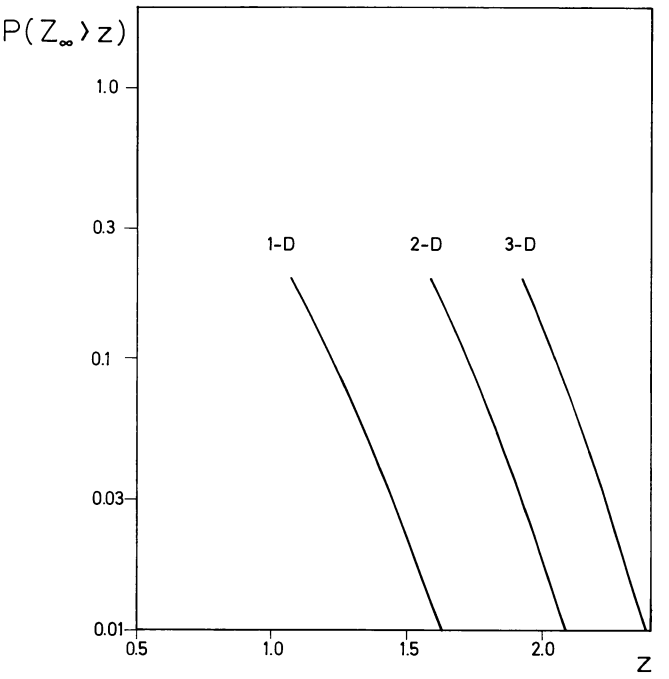

Fig. 2. Diagram of the empirical formulae simulating the asymptotic behaviour; from left to right, the one-, two- and three-dimensional tests

5. The distribution-free character of $Z_n^{(3)}$

As mentioned above, the $Z_n^{(3)}$ statistic is not necessarily distribution-free and we need to investigate this problem in order to demonstrate the generality of the results of Sect. 4. From the work of Peacock (1983) on the two-dimensional test, we know that the statistic $Z_n^{(2)}$ has the desired property as long as the two random variables X and Y are not too strongly correlated.

Again using Monte Carlo techniques, we computed the distribution of a statistic defined in the same way as $Z_n^{(3)}$ but based on different null-hypotheses. Several three-dimensional pdf, exotic or not, have been used (both to compute the cdf and to generate the data); all are freely inspired from the patterns of Fig. 3 of Peacock's (1983) paper. A special effort has been made not to neglect patterns with a marked correlation. We conclude that for $Z_n^{(3)}$, the correlation between the random variables appears as the determining factor too but only extremely correlated variables are problematic and, as will be shown, those cases are not realistic or at least are irrelevant to a three-dimensional test. In fact, only two pdf's give deviations standing out of the statistical fluctuations (Poisson noise of the simulations).

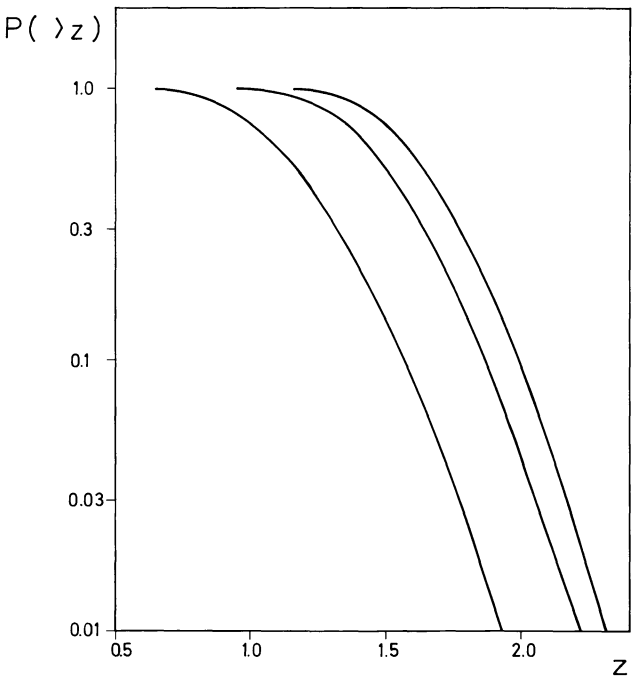

Fig. 3. Investigation of the distribution-free character of the $Z_{30}^{(3)}$ statistic: from left to right, the null-hypothesis is modelized by:
- pdf N°1 (high correlation of three random variables),
- pdf N°2 (high correlation of only two random variables),
- uncorrelated or factorizable pdf (See text in section 5 for more details)

These are

$f_{X,Y,Z}(x, y, z) = \text{const.} \neq 0$ if and only if $x \simeq y \simeq z$

and

$f_{X,Y,Z}(x, y, z) = \text{const.} \neq 0$ if and only if $x \simeq y$, any z

corresponding to zero density everywhere except on a diagonal line of constant density (trial pdf N°1) and on a diagonal plane of constant density (trial pdf N°2) respectively. The two resulting distributions of the statistic along with the assumed one are given in Fig. 3 for the case $n = 30$.

In conclusion, practically all our simulations result in a distribution of the statistic extremely similar to the one derived in Sect. 4. A severe case of deviation, although still realistic, is the one relevant to pdf N°2: it is clear that there is some perturbation of the distribution but from a practical point of view, the resulting uncertainties are negligible (cf. Fig. 3). In fact, the distribution relative to the pdf N°2 can be considered as a border-line case. However, the distribution relevant to the pdf N°1 is more troublesome as the error is no longer negligible. Nevertheless, we think that cases "between" pdf N°1 and pdf N°2 are somewhat unrealistic. At least, they can be considered as irrelevant to a three-dimensional test because the correlation is so strong that one can always reduce the problem to a two-dimensional one by a simple change of variables. The $Z_n^{(3)}$ statistic is therefore sufficiently distribution-free for all cases of practical interest.

6. An alternative way to make the extension

When a first version of this paper was ready to be submitted for publication, we became aware of a paper by Fasano and Franceschini (1986) who introduce an alternative way to conceive the extension of the one-dimensional Kolmogorov-Smirnov test. Their statistic is based on a deviation defined as

$$D_n = \sup_{\substack{(x,y)=(X_i,Y_i) \\ \text{all } i \in [1,n]}} |F_n(x, y) - F_{X,Y}(x, y)| . \quad (27)$$

Clearly, they restrict the search of the *supremum* of the deviation between the sdf and the cdf to loci harbouring a data point. Of course, the true *supremum* will generally be missed but the maximum deviation computed in such a way will have a tendency to vary in the same manner as the true *supremum* does. Therefore, the Fasano and Franceschini's (1986) statistic is probably well-behaved, at least as long as the genuine parent population distribution and the assumed one are not too different. Both Peacock's statistic and the one of Fasano and Franceschini degenerate in the one-dimensional problem to the classical Kolmogorov-Smirnov test. The advantage of the approach by Fasano and Franceschini is the small number of loci (n) where the investigation is to be made. However, their statistic is sensitive to the correlation between the two random variables and therefore requires the publication of three entry tables: the added parameter is the correlation coefficient. The latter can be estimated from the sample but some problems can arise as we state at the end of Sect. 2.

7. Conclusions

We have presented in this paper a three-dimensional version of the Kolmogorov-Smirnov test. An algorithm to systematically explore the space and compute the sdf is presented and an expression for the $Z_n^{(3)}$ statistic is given. Using Monte Carlo techniques, we have derived its distribution; a table of critical values has been given for most usual significance levels as well as empirical formulae to simulate the asymptotic behaviour. The distribution-free character of the $Z_n^{(3)}$ statistic has also been investigated and we conclude that the test is sufficiently distribution-free to be widely and safely used. Only cases of extremely high correlation are not to be treated blindly. The new statistic may be fully used in astronomy as well as in other fields of research; it is an efficient alternative to the Pearson χ^2 test. An example of application has already been evoked in Gosset et al. (1986). Our work is also a first step towards the conception of a more general N-dimensional test.

*Acknowledgements.*The author is greatly indebted to R. Scuflaire, J. Surdej and J.P. Swings for having carefully read and significantly increased the quality of the manuscript. Fruitful comments by an Anonymous Referee have led, we hope, to improvements. A preliminary discussion with J.A. Peacock has been determining as well.

Appendix A

In this appendix, we describe an algorithm to compute the $Z_n^{(3)}$ statistic. The basic philosophy being the same for both the two-dimensional and the three-dimensional cases, we feel preferable for the comprehension to begin with a small example in a plane. We consider the computation of D_n^{CC}. In Fig. A.1, we show a square with five points. Dashed lines are also drawn which delimit the cells where the sdf keeps a constant value: this value of F_n^{CC}

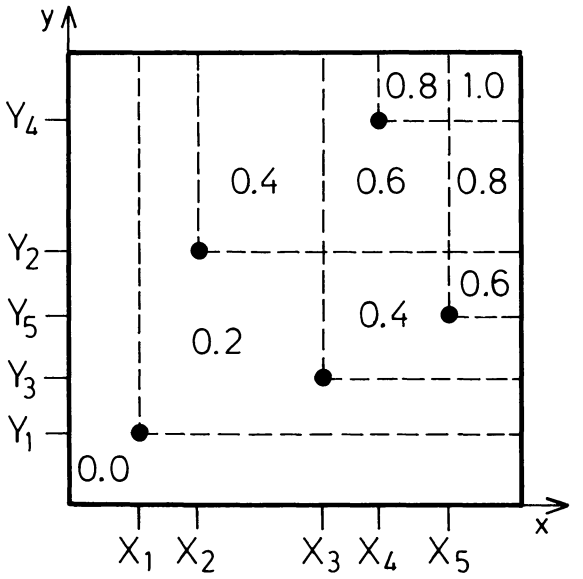

Fig. A1. A square plane with five data points. Dashed lines delimit the cells of constant sdf F_n^{CC} within which the relevant values of the latter are indicated

is indicated at its relevant location. The cdf being monotonic, the search for the *supremum* can be restricted to the data points and to the points at the intersection of two lines. It is preferable to perform a systematic exploration taking the X_j one by one, following an increasing order. As an illustration, let us imagine we arrived at $x = X_3$: we have to position ourselves at

$$(x, y) = (X_3, Y_3)$$

and after that at

$$(x, y) = (X_3, Y_2).$$

At both places, the sdf has two values. The greater one (denoted sdf^+) is a direct function of the order of the relevant Y_i: as we have $Y_2 > Y_3 > Y_1$, the sdf^+ takes respectively the values $2 \times 0.2 = 0.4$ and $3 \times 0.2 = 0.6$ as can be seen in Fig. A.1. The lower values of the sdf (denoted sdf^-) can be computed from the sdf^+ by removing one step if we are on a data point and two steps if we are at the intersection of two lines.

Hereafter, we discuss an algorithm based on the preceding considerations, but in the framework of the three-dimensional extended Kolmogorov-Smirnov test. In what follows, we only consider the computation of D_n^{CCC}; the basic idea is the same *mutatis mutandis* for the seven other intermediate statistics, namely

$$D_n^{CCD}, D_n^{CDC}, D_n^{CDD}, D_n^{DCC}, D_n^{DCD}, D_n^{DDC} \text{ and } D_n^{DDD}.$$

In order to visualize things, let us imagine that the x-axis points towards us, the y-axis to the right and the z-axis to the top. What we hereafter call "space" is some relevant parallelipipedic closed volume such that the pdf can be set to zero outside of it. We then assume that the n data points (n sample)

$$(x, y, z) = (X_i, Y_i, Z_i) \quad \forall i \in [1, n],$$

are dispersed in this space and that the X_i's are ordered in an increasing manner. To draw the sdf (in this case F_n^{CCC}) would of course be impossible as it would require a fourth dimension. Nevertheless, we can imagine the sdf as a partition of the whole space in cells where the sdf takes constant values; in general, those cells have complex shapes with a large number of corners and edges. The inter-cell borders are portions of planes perpendicular to one of the axes; steps of the sdf are of course located there. Since the sdf can be considered as a discontinuous function, we can attribute to it at least two values at a discontinuity point. One will be labelled + and will correspond to a limit approach to the point from the top-right-front corner, the other labelled − being accessed from the bottom-left-rear corner. Both of these values are to be compared with the cdf as computed there. It is interesting to note further that, for the same reason as for the two other tests, the *supremum* is necessarily located on a corner of one of the partitioning cells and nowhere else. A simple proof of this goes as follows. Let us position ourselves at any point in a cell; as the sdf is constant and the cdf monotonic, it is always possible, except if we are in a corner, to find a travel direction such that the cdf will increase or decrease (depending on whether the sdf is respectively inferior or superior to the cdf). This implies an increase of the difference between the two functions. All such types of travel will automatically lead to a corner.

We will now imagine that we are at a corner (x, y, z): the sdf^+ is equal to j/n where j is the number of data points located in the bottom-left-rear three-dimensional octant of (x, y, z). But j is also a serial number of the corner among a collection of points arranged in a very particular way. This statement is worth being explicited. Let us consider the m^{th} data point

$$(x, y, z) = (X_m, Y_m, Z_m), \quad m = 1, n$$

and place ourselves in the plane

$$x = X_m.$$

Taking into account all the points

$$(y, z) = (Y_k, Z_k), \quad \text{with } k = 1, m$$

and ordering them according to increasing y, we obtain

$$(Y_{p(k)}, Z_{p(k)}), \quad \text{with } k \in [1, m]$$

and where p is the relevant new ordered numbering. In fact it would be more rigorous to refer to the permuted Y by using a different symbol such as Y^* ($Y^*_{p(k)} = Y_k$ for all $k \in [1, m]$); this heavy notation has nevertheless been dropped for the remainder of the paper. Points to the left of $Y_{p(m)}$ are of no interest since they correspond to a cell edge, passing through the plane $x = X_m$ perpendicularly to it, and not to a corner. Now, we consider in turn points $Y_{p(i)}$ with $p(i) = p(m)$, m. Let us place ourselves on the line

$$\begin{cases} x = X_m \\ y = Y_{p(i)}, \end{cases}$$

look at the points

$$z = Z_{p(k)} \quad \text{with } p(k) = 1, p(i),$$

and order them according to increasing z. We obtain

$$Z_{q(p(k))} \quad \text{with } p(k) \in [1, p(i)]$$

and where q is the relevant new ordered numbering. Each point below $Z_{q(p(m))}$ corresponds at best to a cell edge, parallel to the x-axis, and not to a corner: they are of no interest. Each point below $Z_{q(p(i))}$ can also be neglected since it is connected to a cell-

edge parallel to the y-axis. Each in turn, we now consider the points

$Z_{q(p(l))}$ with $q(p(l)) = \max(q(p(i)), q(p(m))), p(i)$.

All are corners and the corresponding sdf^+ is simply expressed by

$$(F_n^{CCC})^+ = j/n \quad \text{with } j = q(p(l)) \tag{A.1}$$

i.e. the serial number of the corner as stated above.

The computation of sdf^- is less simple. However, it can be calculated from the sdf^+ by removing some quantity function of the respective arrangement of the data points:

$$(F_n^{CCC})^- = (F_n^{CCC})^+ - \alpha/n \tag{A.2}$$

where α is determined by following a small set of rules. We give those rules below without any demonstration, leaving it to the reader. Using the same notation as above, in the plane $x = X_m$,

- α is 1 if current y is Y_m and current z is Z_m
- α is 2 if current y is Y_m and current $z > Z_m$
- α is 2 if current y is $Y_{p(i)} > Y_m$ and current z is Z_m
- α is 3 if current y is $Y_{p(i)} > Y_m$ and current $z > Z_m$

except that α is 2 if in addition z is $Z_{p(i)}$.

This algorithm permits to combine the cumulation of the sdf with the systematic exploration of the space. Finally, let us point out that it is necessary to take into account loci on the upper borders of the space and that the algorithm is slightly more complicated if we wish to include data points for which at least one of the coordinates is identical.

References

Birnbaum, Z.W.: 1952, *J. Am. Statist. Ass.* **47**, 425
Clark, P.J., Evans, F.C.: 1954, *Ecology* **35**, 445
Fall, S.M.: 1979, *Rev. Modern Physics* **51**, 21
Fasano, G., Franceschini, A.: 1986, *Monthly Notices Roy. Astron. Soc.* (in press)
Gosset, E., Louis, B.: 1986, *Astrophys. Space Sci.* **120**, 263
Gosset, E., Surdej, J., Swings, J.P.: 1986, in *Quasars*, Proceedings IAU Symposium 119, eds: G. Swarup, V.K. Kapahi, pp. 45–46
Kendall, M.G., Stuart, A.: 1967, The Advanced Theory of Statistics, Vol. 2 (C. Griffin and Co., London)
Mennessier, M.O.: 1985, *Astron. Astrophys.* **144**, 463
Noether, G.E.: 1967, Elements of Nonparametric Statistics, John Wiley and Sons, New York
Peacock, J.A.: 1983, *Monthly Notices Roy. Astron. Soc.* **202**, 615
Rose, J.A.: 1977, *Astrophys. J.* **211**, 311
Sharp, N.A.: 1979, *Astron. Astrophys.* **74**, 308
Thompson, H.R.: 1956, *Ecology* **37**, 391
Webster, A.: 1976, *Monthly Notices Roy. Astron. Soc.* **175**, 61
Webster, A.: 1982, *Monthly Notices Roy. Astron. Soc.* **199**, 683
Zięba, A.: 1975, *Acta Cosmologica* **3**, 75