Astron. Astrophys. 289, 640–648 (1994)

ASTRONOMY AND ASTROPHYSICS

Probability density estimation in astronomy

R. Vio[1], G. Fasano[2], M. Lazzarin[3], and O. Lessi[4]

[1] ESA, IUE Observatory, Villafranca del Castillo, Apartado 50727, E-28080 Madrid, Spain
[2] Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122 Padova, Italy
[3] Dipartimento di Astronomia dell'Università di Padova, Vicolo dell'Osservatorio 5, I-35122 Padova, Italy
[4] Dipartimento di Statistica dell'Università di Padova, Via S. Francesco 33, I-35122 Padova, Italy

Received 30 June 1993 / Accepted 30 January 1994

Abstract. In this paper we consider three empirical estimators (Kernel, adaptive Kernel and parametrizing families of Johnson) of the probability density function, which are alternatives to the classical Histogram. By means of numerical simulations the superiority of such estimators is shown in the context of some typical astronomical problems. Given the simplicity of implementation and the efficiency of calculation, the use of these methods is recommended. Indication for the implementation is provided.

Key words: Methods: numerical – methods: statistical

1. Introduction

Histogram is by far the most popular method for the estimation of the Probability Density Function (PDF) underlying a given sample of data. Its use is widespread, in spite of some disappointing drawbacks, first of all the *subjectivity* of the results. It is well known that certain structure of an empirical PDF can be emphasized simply by *playing* with the positions and the sizes of the bins. Certainly this cannot be considered a safe way to proceed. More reliable estimators are needed to provide, at the same time, both "*objective*" results and a good representation of the PDF underlying the data. In this respect, during the last decades, statisticians have elaborated some efficient solutions. These, however, are almost unknown to the astronomical community. On the other hand, the current statistical literature on this field is generally limited to the theoretical aspects of the problem, neglecting the drawbacks connected with the practical applications. This fact limits the applicability of the innovative techniques in the astronomical context. The aim of this paper is to fill this gap by considering three of the most interesting estimators available today. Their performances in experimental situations are explored and compared to each other by means of numerical simulations.

Send offprint requests to: G. Fasano

2. Methods

2.1. Histogram

The Histogram is the most widely used method to estimate an empirical PDF. Its popularity lies principally in its simplicity: given a set of experimental data X_i ($i = 1, 2, \ldots, N$), fixed origin x_0 and bin width h, and defining the bins to be the right open intervals $[x_0+jh,\ x_0+(j+1)h[$ (j = integer), the Histogram $\hat{f}_H(x)$ is defined as:

$$\hat{f}_H(x) = \frac{1}{Nh}(\text{number of } X_i \text{ in the same bin as } x).$$

Implicit in this definition is the choice of both an origin and a bin width. It is well known that $\hat{f}_H(x)$ depends on these two factors. Less clear, however, is that the consequences of an inadequate choice are not only *aesthetic*; in particular $\hat{f}_H$ may become a strongly biased estimator (de Jager et al. 1986).

A possible way to relieve this situation could be the definition of some criterion able to determine the *best* values of h and x_0. At the present, however, a general method has not yet been developed. A solution is available if one fixes the interval, $[a,\ a+L]$, where $\hat{f}_H(x)$ has to be calculated. In this case the *optimal* value of the number of bins (and therefore the *best* value of h) can be found by minimizing the so called Mean Integrated Square Error (MISE). This quantity is defined as:

$$E\{\int_{-\infty}^{\infty}[\hat{f}(x) - f(x)]^2 \mathrm{d}x\},$$

and gives a measure of the *global* distance between the *true* PDF $f(x)$ and its estimate $\hat{f}(x)$. In the case of the Histogram this quantity can be estimated through the formula (Linhart & Zucchini 1986):

$$\mathrm{MISE}(N_\mathrm{B}) = \frac{N_\mathrm{B}}{NL}\left[1 - \frac{N+1}{N-1}\left(\sum_{i=1}^{N_\mathrm{B}}\frac{N_i^2}{N} - 1\right)\right],$$

where N_B is the number of bins and N_i is the number of observations in the i–th bin. Unfortunately numerical simulations

show that, in presence of small samples of data (as often happens), MISE(N_B) is a very noisy function and consequently its utility appears limited.

Another limitation of the Histogram is that $\hat{f}_H(x)$ is a discontinuous function. This fact could be a serious handicap when density estimation constitutes only an intermediate step of some procedure as, for example, a deconvolution.

2.2. Kernel method

In the last decades, statisticians have spent much effort in the study of estimators able to alleviate the drawbacks typical of the Histogram. An interesting solution is represented by the so–called Kernel estimator $\hat{f}_K(x)$, defined as:

$$\hat{f}_K(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right), \quad (1)$$

where h is the *window width* and $K(x)$ is a function, called *kernel*, that satisfies the condition:

$$\int_{-\infty}^{\infty} K(y)\mathrm{d}y = 1. \quad (2)$$

From the definition it appears that $\hat{f}_K(x)$ is constructed through the sum of *bumps*, each corresponding to a single data point. The function $K(x)$ determines the shape of such bumps, whereas the parameter h controls their width and therefore works as a smoothing parameter. The great advantage with respect to the Histogram is that $\hat{f}_K(x)$ does not require the binning of the data. However not all the problems are solved: the question of the *window width* still remains and there is the new problem of the functional form of $K(x)$.

The size of the *window width* is a delicate question, but it can be treated efficiently. In fact, it is possible to obtain an estimate of the optimum value for h, again in the sense of minimizing the MISE, through the minimization of the quantity (Silvermann 1986):

$$P(h) = \int \hat{f}^2(x)\mathrm{d}x - 2N^{-1}\sum_i \hat{f}_{-i}(X_i),$$

where $\hat{f}_{-i}(X_i)$ is the density estimate constructed from all the data but the point X_i. This method is called Least–Squares Cross–Validation. Numerical simulations have shown that, in contrast with the MISE(N_B) defined in the case of the Histogram, the function $P(h)$ generally has little noise and presents a clear convexity even with a limited sample of data. This fact is extremely useful since it allows a reliable automatic, and therefore *objective*, choice of the smoothing parameter.

Concerning the functional form of $K(x)$, it can be shown that the influence of the shape of the kernel on the MISE is small. This is a useful result since it allows the choice of the most convenient functional form. For example, since $\hat{f}_K(x)$ shares the differentiability properties of $K(x)$, it is advantageous to use a kernel having derivatives of all orders.

From the computational point of view, the construction of $\hat{f}_K(x)$ through its definition is unpractical since equation (1) corresponds to a convolution. Actually it is more convenient to pass into the frequency domain by means of the Discrete Fourier Transform (DFT). In such a case it is preferable to use a kernel that makes the DFT calculation computationally efficient. The standard Gaussian density represents a good choice. Moreover (again for computational efficiency), it is useful to discretize the data to a very fine grid with 2^p cells ($p = 7$ or 8 is sufficient) since in this way it is possible to use the Fast Fourier Transform (FFT). A slight modification of this discretization scheme is given in Jones & Lotwick (1984).

Given the present, large circulation of codes for the calculation of the FFT, the implementation of the Kernel algorithm in the frequency domain is particularly simple. In any case a FORTRAN code is provided in Silverman (1982), and a program is also available in the IMSL library. However, the efficient implementation of the Cross–Validation method is not so direct. Even if the FFT approach also permits the solution of this problem, the technical details are too complex and therefore we refer to the work of Silvermann (1986). Here we give a rule–of–thumb which, although general, is adequate for a wide range of densities and is trivially applicable to the data. In effect, assuming

$$h = 0.9\ A\ N^{-1/5}; \quad (3)$$

$$A = \min(\text{st.dev.},\ \text{interquartile range}/1.34),$$

one can obtain a MISE error within 10% for all the t–distributions considered, for the log–normal with skewness up to about 1.8, and for mixtures of normal distributions with separation up to 3 standard deviations (Silvermann 1986).

Although in general the Kernel method works well, its performance can be improved, at least theoretically, in situations such as distributions characterized by very long tails, or in case of mixtures with different spreads. The point is that a unique value of h can over– smooth the parts of the PDF where the data are more concentrated and under– smooth those where the data are sparser. In other words it should be necessary to adapt the window width to the local data density. This target is pursued by the *adaptive* Kernel estimator $\hat{f}_{Ka}(x)$. According to Silverman (1986) its construction proceeds according to the following steps:

1. Determination of a *pilot estimate* $\tilde{f}(x)$. This step is necessary in order to have an idea of the data density. A choice may be $\tilde{f}(x) = \hat{f}_K(x)$, but other solutions are possible.

2. Calculation of the *local bandwidth factors* λ_i:

$$\lambda_i = [\tilde{f}(X_i)/g]^{-\alpha},$$

where g is the geometric mean of the $\tilde{f}(X_i)$:

$$\log_{10}(g) = N^{-1}\sum \log_{10}\tilde{f}(X_i),$$

and α is the *sensitivity parameter*, that is a number satisfying the condition $0 \leq \alpha \leq 1$.

3. Construction of the *adaptive Kernel estimate* $\hat{f}_{Ka}(x)$ through:

$$\hat{f}_{Ka}(x) = N^{-1}\sum_{i=1}^{N} h^{-1}\lambda_i^{-1} K\{h^{-1}\lambda_i^{-1}(x - X_i)\}.$$

At odds with the classical Kernel method, the bandwidth now becomes $h\lambda_i$ and therefore it is no longer constant, depending on the local density of the PDF. Two remarks are necessary:

(a) The parameter α gives the sensitivity of $\hat{f}_{Ka}(x)$ to the details of the *pilot estimate*. The value suggested by Silvermann (1986) is 0.5. However, numerical experiments indicate a dependency of the *optimal* value on the smoothness and characteristics of $\tilde{f}(x)$.

(b) As in the case of $\hat{f}_K(x)$, it is also necessary for $\hat{f}_{Ka}(x)$ to define the value of h. A common solution is to take the same value as used for the *pilot estimate* (this is the solution we use in the following), but it is also possible to extend the cross–validation method to this case, although its implementation is not so computationally efficient.

2.3. Johnson system

The Histogram and the Kernel estimators represent two non–parametric solutions to the PDF estimation problem. A parametric way is, however, viable. The philosophy under this approach is to use the data to estimate the parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ of a set of flexible distribution families $f_i(x; \alpha_1, \alpha_2, \ldots, \alpha_p)$, $(i = 1, 2, \ldots, m)$, which, at varying $\alpha_1, \alpha_2, \ldots, \alpha_p$, are able to provide a wide variety of possible shapes. In this way it is possible to avoid the problems connected with the binning of the data and with the choice of the *optimal* window size, which are the most serious limitations of the methods presented in the previous sections.

Various classes of families have been developed in the past. However, one of the most useful solutions is represented by system of three families $f_{J_i}(x; \eta, \epsilon, \lambda, \gamma)$ $(\eta, \lambda > 0;$ $-\infty < \gamma, \epsilon < \infty)$ introduced by Johnson (1949). These, besides being simple to be used (see below), are able to reproduce all the possible arbitcombinations of the skewness $\pm\sqrt{\beta_1} = E[(x-\mu)^3]/\{E[(x-\mu)^2]\}^{3/2}$ (NB. the square root is given only for convention and the sign is given according to that of the numerator) and kurtosis $\beta_2 = E[(x-\mu)^4]/\{E[(x-\mu)^2]\}^2$ (note that the kurtosis and the skewness of a distribution are not completely free, since they have to satisfy the inequality $\beta_2 > 1 + \beta_1$). As well known, a given distribution is completely characterized by its moments $E[x^k]$, or equivalently by its central moments $E[(x-\mu)^k]$, with $k = 1, 2, \ldots$, and $\mu = E[x]$. The first two moments (e.g. mean and variance) contain information about location and spread, whereas the moments with $k \geq 3$ give information about the shape of the distribution. Unfortunately, unless large samples of data are available, the estimate of these quantities becomes quickly unreliable with increasing k. Therefore, a given distribution can be usually characterized only by the first four moments (e.g. mean, variance, skewness, kurtosis) or, more generally, by four quantities related to them.

The Johnson families are related to the transformations of the normal distribution. If x is the random variable whose distribution is to be represented empirically, the general form of the transformation is given by:

$$z = \gamma + \eta g(x; \epsilon, \lambda)$$

where z is the standard normal variable, g is an arbitrary function and η, $\lambda > \gamma$, ϵ are parameters. Johnson proposed three alternative forms for the function g, according to the fact that the random variable x is unbounded, bounded both above and below and bounded only below, respectively:

$$\begin{aligned} g_1(x) &= \sinh^{-1}\left(\frac{x-\epsilon}{\lambda}\right) \\ g_2(x) &= \ln\left(\frac{x-\epsilon}{\lambda-(x-\epsilon)}\right) \\ g_3(x) &= \ln\left(\frac{x-\epsilon}{\lambda}\right) \end{aligned} \qquad (4)$$

The PDF corresponding to this distributions are:

$$f_{J_1}(x) = \frac{\eta}{\sqrt{2\pi[(x-\epsilon)^2+\lambda^2]}} \times \qquad (-\infty < x < \infty)$$

$$\times \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{(x-\epsilon)}{\lambda} + \sqrt{\left(\frac{x-\epsilon}{\lambda}\right)^2 + 1}\right)\right]^2\right\};$$

$$f_{J_2}(x) = \frac{\eta}{\sqrt{2\pi}} \frac{\lambda}{(x-\epsilon)(\lambda - x + \epsilon)} \times \qquad (\epsilon \leq x \leq \lambda + \epsilon)$$

$$\times \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{x-\epsilon}{\lambda - x + \epsilon}\right)\right]^2\right\}; \qquad (5)$$

$$f_{J_3}(x) = \frac{\eta}{\sqrt{2\pi}(x-\epsilon)} \times \qquad (x \geq \epsilon)$$

$$\times \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{x-\epsilon}{\lambda}\right)\right]^2\right\}.$$

By means of the substitution $\delta = \gamma - \eta \ln \lambda$, the last equation becomes:

$$f_{J_3}(x) = \frac{\eta}{\sqrt{2\pi}(x-\epsilon)} \exp\left\{-\frac{1}{2}[\delta + \eta \ln(x-\epsilon)]^2\right\}. \qquad (5a)$$

Equations (5) define, respectively, the so called S_U, S_B and S_L families. In a plot of the third and fourth standardized moments, $\sqrt{\beta_1}$ and β_2, the S_L family forms a curve which divides the (β_1, β_2) plane in two regions. The S_B distributions lies in one of the regions and the S_U lies in the other.

In using the Johnson system, the first step must be to determine which of the three families should be used. A possible answer to this question could be based on the estimation of β_1 and β_2 and on which of the two regions in the plane (β_1, β_2) the computed point fall into. This method, however, is not advisable given the high variance of the estimates of β_1 and β_2

and their sensitivity to the outlier. Better results are obtainable with percentiles. Slifker & Shapiro (1980) suggest the use of the quantity

$$mn/p^2$$

that can be demonstrated to be < 1, $= 1$ and > 1, for the S_B, S_L and S_U families respectively. Here

$$m = \hat{x}_{\alpha_{3z}} - \hat{x}_{\alpha_z}; \quad n = \hat{x}_{-\alpha_z} - \hat{x}_{-\alpha_{3z}}; \quad p = \hat{x}_{\alpha_z} - \hat{x}_{-\alpha_z}, \quad (6)$$

where x_{α_z} is the sample percentile corresponding to the probability defined by the distribution function at a fixed value of the standard, gaussian variable z. In practice, once the value of z is fixed(and therefore the values of $-3z, -z, 3z$), we obtain, from a table providing the distribution function of the standard normal PDF, the corresponding probabilities $\alpha_{-3z}, \alpha_{-z}, \alpha_z, \alpha_{3z}$. These allow the estimation, from the data, of the corresponding percentiles x_α and then the calculation of the quantity mn/p^2. For example if we choose $z = 0.6$, we have $\alpha_{-3z} \simeq 0.036$, $\alpha_{-z} \simeq 0.274$, $\alpha_z \simeq 0.726$, $\alpha_{3z} \simeq 0.964$, and the quantities in equation (6) are calculated in terms of the empirical percentiles $\hat{x}_{0.036}, \hat{x}_{0.274}, \hat{x}_{0.726}$ and $\hat{x}_{0.964}$. Shapiro & Gross (1981) suggest that, for a small sample of data, z should be 0.524 ($3\alpha = 1.572$) in order to have both a good sampling of the distribution tails and a sufficient statistical accuracy of the estimates. For larger samples this value should be raised. In any case, numerical simulations have shown that, except for pathological choices, the exact value of the parameter is not crucial. Note that, since this quantity is a statistics, the probability that it takes exactly the value 1 is null. Therefore, to have the possibility of selecting the S_L distributions a critical interval of acceptance must be adopted. However, it is our experience that it is better not to use this distributions because, being characterized by only three free parameters, it is less "flexible" than the other two cases.

Once a family has been selected, the next step is to estimate the values of the corresponding characteristic parameters. A possible solution is the matching of the theoretical first four moments of the family with the corresponding sample moments. Again, as previously stressed, this way is not advisable and better results are obtained through the matching of the percentiles. It can be shown that the parameters $\eta, \gamma, \lambda, \epsilon$ of the families (5) can be estimated by means of the formulae:

2.3.1. Johnson S_U family

$$\hat{\eta} = \frac{2z}{\cosh^{-1}\left[\frac{1}{2}\left(\frac{m}{p} + \frac{n}{p}\right)\right]}; \quad (\hat{\eta} > 0)$$

$$\hat{\gamma} = \hat{\eta}\sinh^{-1}\left[\frac{\frac{n}{p} - \frac{m}{p}}{2\left(\frac{m}{p}\frac{n}{p} - 1\right)^{1/2}}\right];$$

$$\hat{\lambda} = \frac{2p\left(\frac{m}{p}\frac{n}{p} - 1\right)^{1/2}}{\left(\frac{m}{p} + \frac{n}{p} - 2\right)\left(\frac{m}{p} + \frac{n}{p} + 2\right)^{1/2}}; \quad (\hat{\lambda} > 0)$$

$$\hat{\epsilon} = \frac{\hat{x}_{\alpha_z} + \hat{x}_{-\alpha_z}}{2} + \frac{p\left(\frac{n}{p} - \frac{m}{p}\right)}{2\left(\frac{m}{p} + \frac{n}{p} - 2\right)}.$$

2.3.2. Johnson S_B family

$$\hat{\eta} = \frac{z}{\cosh^{-1}\left\{\frac{1}{2}\left[\left(1 + \frac{p}{m}\right)\left(1 + \frac{p}{n}\right)\right]^{1/2}\right\}}; \quad (\hat{\eta} > 0)$$

$$\hat{\gamma} = \hat{\eta}\sinh^{-1}\left\{\frac{\left(\frac{p}{n} - \frac{p}{m}\right)\left[\left(1 + \frac{p}{m}\right)\left(1 + \frac{p}{n}\right) - 4\right]^{1/2}}{2\left(\frac{p}{m}\frac{p}{n} - 1\right)}\right\};$$

$$\hat{\lambda} = \frac{p\left\{\left[\left(1 + \frac{p}{m}\right)\left(1 + \frac{p}{n}\right) - 2\right]^2 - 4\right\}^{1/2}}{\frac{p}{m}\frac{p}{n} - 1}; \quad (\hat{\lambda} > 0)$$

$$\hat{\epsilon} = \frac{\hat{x}_{\alpha_z} + \hat{x}_{-\alpha_z}}{2} - \frac{\hat{\lambda}}{2} + \frac{p\left(\frac{p}{n} - \frac{p}{m}\right)}{2\left(\frac{p}{m}\frac{p}{n} - 1\right)}.$$

2.3.3. Johnson S_L family

$$\hat{\eta} = \frac{2z}{\ln\left(\frac{m}{p}\right)}$$

$$\hat{\delta} = \hat{\eta}\ \ln\left[\frac{\frac{m}{p} - 1}{p\left(\frac{m}{p}\right)^{1/2}}\right]$$

$$\hat{\epsilon} = \frac{\hat{x}_{\alpha_z} + \hat{x}_{-\alpha_z}}{2} - \frac{p}{2}\frac{\frac{m}{p} + 1}{\frac{m}{p} - 1}.$$

Even if apparently cumbersome, the implementation of this automatic method is trivial and does not deserve any particular comment. A remark, however, is necessary: in every situation where some analytical curve is fitted to experimental data it is important to check the result. That can be accomplished either by comparing $\hat{f}_J$ with more robust estimators (as, for example, the histogram) or better by transforming the data, via equations (3), to the corresponding normal form and then test for normality.

Besides the Johnson system, in the current literature other parametric families are available. Their utility, however, is more limited due to the difficulty of use. For example, the Pearson's system (Kendall & Stuart 1958), being based on a set of 12 parametric families, presents serious drawback for defining which of them must be used with regard a given sample of data. Another example is represented by the so-called generalized lambda system (Dudewicz et al. 1974) which, for the estimation of the characteristic parameters, requires the solution of a set of non–linear equations.

3. Numerical tests

As shown in the previous section the Kernel and the Johnson estimators appear, at least theoretically, superior to Histogram. In reality their properties have been studied only in the asymptotic case and this gives no certainty about their performances in practical situations. In order to examine this point, we have carried out a set of numerical simulations. In particular we have considered three different PDFs: the standard gaussian, the Chi–Squared with three degrees of freedom and a mixture formed by two gaussians (means 0 and 4 and dispersion 1 and 2 respectively). These distributions correspond to situations with increasing *difficulty*, namely a symmetric and continuous distribution, a strongly asymmetric distribution and an overlap of two distributions characterized by different locations and spreads. For each of the three cases we have considered three different sample sizes (100, 200, 500 points) that are typically found in experimental conditions. For each case and for each sample size we have carried out 1000 numerical simulations. Table 1 gives the result of such simulations trough the parameter,

$$D = \sum_{i=1}^{256} [\hat{f}(x_i) - f(x_i)]^2,$$

whose expected value is proportional to the MISE. In the last equation x_i are the coordinates of a set of points uniformly distributed along the interval in which the distribution $f(x)$ is estimated by $\hat{f}(x)$. Since, as stated before, for the Histogram it is not possible to calculate automatically the *optimal* number of bins, in the simulations this quantity has been kept constant and equal to a value chosen, on the basis of some trials, in such a way as to minimize the quantity D. The parameter h of $\hat{f}_K(x)$ has been automatically determined by means of the cross–validation method. The meaning of Table 1 can be summarized in the following points:

(a) In the case of unimodal distributions, also in the presence of poor samples, the Johnson estimator appears superior to the other ones. This result can be explained by the fact that, avoiding the problems connected to the binning of the data or to the convolution with a Kernel, the Johnson estimator is able to use all the information contained in the data;

(b) In the case of the bimodal distribution the Johnson estimator fails. The reason is that, as seen before, this estimator takes into account only the first four sample moments, whereas

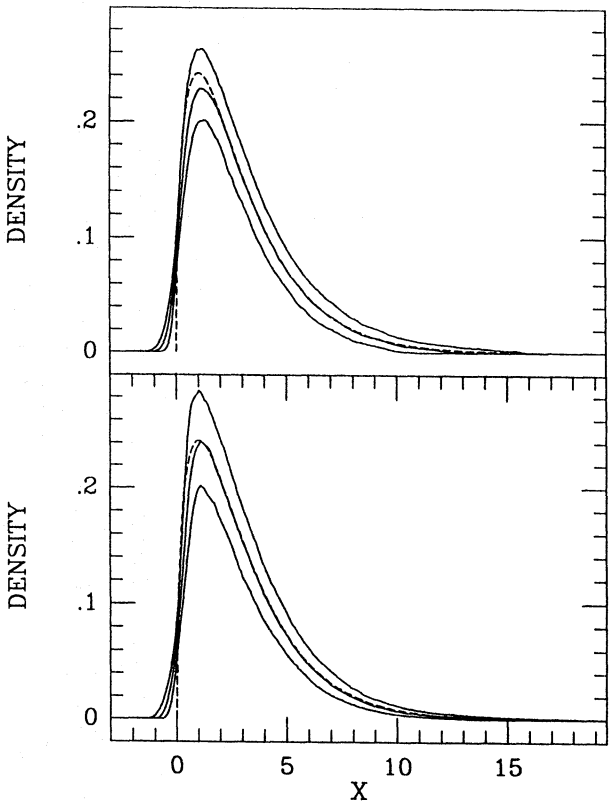

Fig. 1. Top: Mean value and 95% confidence level band (200 points, 1000 simulations) of $\hat{f}_K(x)$ (continuous line) for a χ^2_3 distribution (dotted line). Bottom: The corresponding quantities for $\hat{f}_{K_a}(x)$ ($\alpha = 0.8$)

Table 1. Mean value and dispersion (1000 simulations) of the statistics D for the three distributions discussed in the text. Each pair of columns corresponds to a different size of the data sample used in the numerical experiment (from the left, respectively, 100, 200 and 500)

	Standard Gaussian Distribution					
	$D \times 10^{-1}$	$\sigma_D \times 10^{-1}$	$D \times 10^{-1}$	$\sigma_D \times 10^{-1}$	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$
$\hat{f}_H(x)$	5.123	2.485	3.222	1.323	1.773	5.862
$\hat{f}_K(x)$	2.292	2.030	1.317	1.189	0.636	4.627
$\hat{f}_J(x)$	1.882	1.803	1.005	0.930	0.371	3.326
$\hat{f}_{K_a}(x)$	2.319	1.790	1.357	0.950	0.668	3.993
	χ^2_3 Distribution					
	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$
$\hat{f}_H(x)$	1.544	5.267	8.236	2.931	6.366	1.916
$\hat{f}_K(x)$	1.048	4.558	6.531	2.418	3.704	1.153
$\hat{f}_J(x)$	0.663	6.447	3.439	3.027	1.784	1.197
$\hat{f}_{K_a}(x)$	1.121	4.214	7.985	2.358	5.415	1.254
	Mixture of two Gaussian Distributions					
	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$	$D \times 10^{-1}$	$\sigma_D \times 10^{-2}$
$\hat{f}_H(x)$	1.402	4.255	0.861	2.646	0.050	1.527
$\hat{f}_K(x)$	0.704	4.229	0.413	2.438	0.197	1.045
$\hat{f}_J(x)$	1.361	4.574	1.216	2.759	1.128	1.735
$\hat{f}_{K_a}(x)$	0.513	3.570	0.306	1.930	0.146	1.395

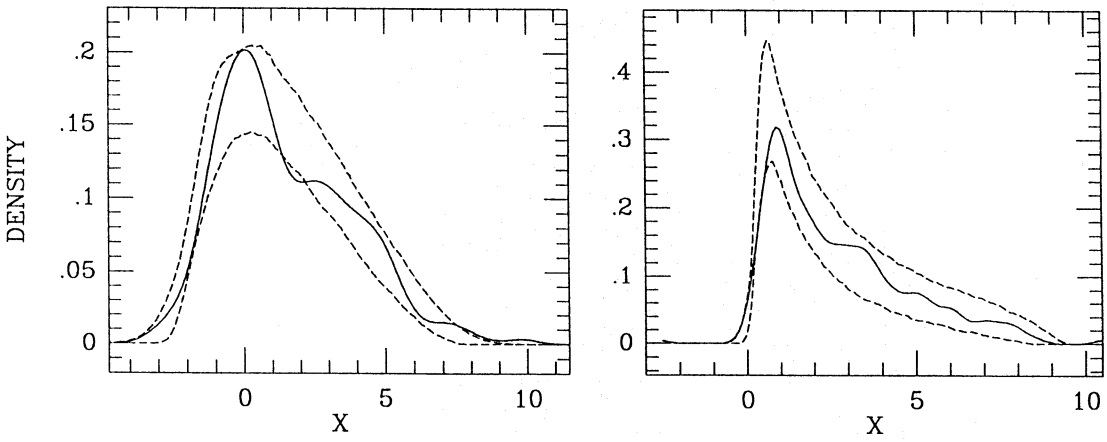

Fig. 2. Left: $\hat{f}_K(x)$ (continuous line) for a sample of 200 data extracted from a mixture of two gaussian distributions having 0 and 3 average values and dispersion 1 and 2 respectively. The dotted lines represent the 90% confidence level calculated as described in the text. Right: The same as before for a χ^2_3 distribution

the information on the bimodality of a distribution is contained also in the higher ones;

(c) Contrary to what was expected on theoretical bases (Silverman 1986), the performance of the adaptive Kernel is better than that of the classical Kernel only in the case of the bimodal distribution. This fact can be explained with the tendency of $f_{Ka}(x)$ to emphasize the structures present in the pilot density. This has been calculated here (as is usual) by means of $\hat{f}_K(x)$, with h provided by the rule–of–thumb (3) (we have also carried out simulations in which h was calculated by means of the cross–validation method but with worse results). In general if the pilot density is either bad or noisy or it already provides a good estimate, the adaptive Kernel makes the final result worse. For example the upper panel of Fig. 1 shows that the Kernel estimator provides an expected estimate of the PDF with a biased mode (a consequence of the convolution of a non–symmetric PDF with a symmetric Kernel), and the bottom panel of Fig. 1 shows that the adaptive Kernel estimator still has this problem. Moreover, even if $E[\hat{f}_{Ka}(x)]$ is better than $E[\hat{f}_K(x)]$, it is characterized by a larger dispersion. The good result in the case of the bimodal distribution is due to the already mentioned difficulties of the classical Kernel estimator in defining a window width suitable for components characterized by different spreads;

(d) In all the cases here considered, the Histogram provides the worst performances.

This last point is partly due to the step behaviour of the Histogram. In order to obtain a continuous distribution it has become a common procedure to interpolate the function $\hat{f}_H(x)$ (by means of polynomials or splines) at coordinates corresponding to the central position of the bins. To test this procedure we have also carried out another set of simulations in which the quantity D has been calculated considering only these points. Given the restricted number of points for calculating D, the number of simulation is 5000 for each case. The results (here not shown) of such simulation are similar to what was found above.

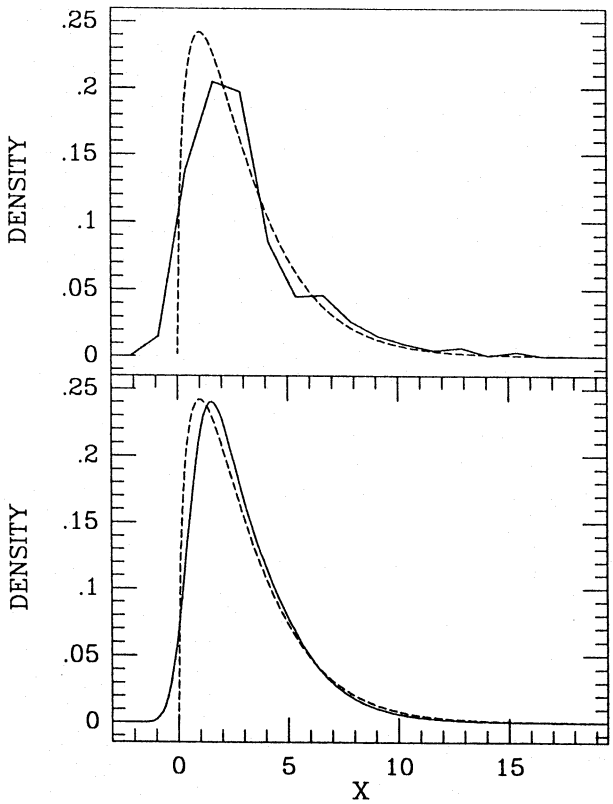

Fig. 3. Top: Deconvolution (continuous line) of $\hat{f}_H(x)$ (20 bins) obtained, on the basis of 200 data points, by means of the Lucy algorithm (10 iterations) in case of a χ^2_3 distribution (dotted line) convolved with a standard gaussian. Bottom: The same as before for $\hat{f}_J(x)$

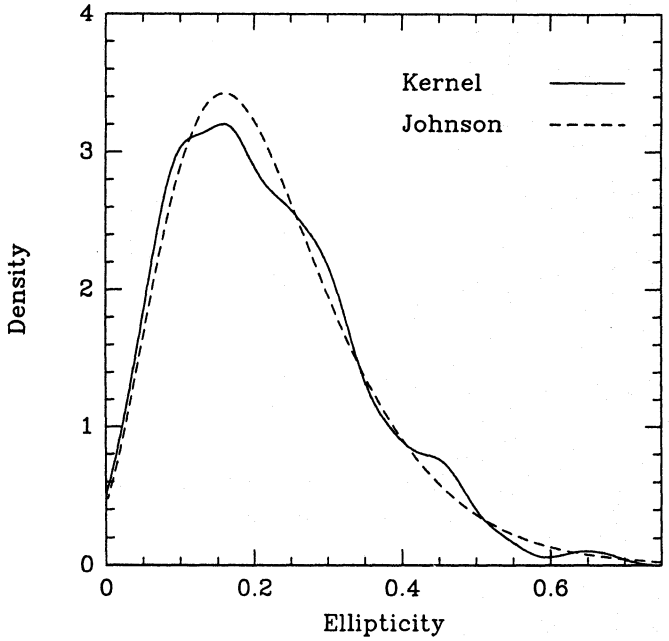

Fig. 4. The Kernel and Johnson's representations of the observed ellipticity distribution for the sample of elliptical galaxies given in Fasano & Vio (1991)

From these facts we can draw the following conclusions:

(a) In case of unimodal distributions, even if the PDF is strongly non–symmetric, the method of Johnson is able to produce very good results;

(b) For more general PDFs it is safer to use the Kernel method;

(c) Due to its strong dependence on the initial pilot estimate and on the free parameter α, the adaptive Kernel method must be used with care, and only in cases when one suspects the inadequacy of the classical Kernel method;

(d) There is no reason to prefer the Histogram since the other methods presented in this paper have nicer properties and are also easy to use and computationally efficient. In particular, since they can work in *automatic* mode, they provide more objective results. Moreover, the Kernel and Johnson methods provide continuous estimates, making it much easier to generate random numbers from an empirical PDF, which is a useful capability in the Monte Carlo studies. This last point is especially true for the Johnson method, for which a fully analytical procedure is available.

We want to stress that this does not mean that the Histogram should no longer be used, but only that it should be used in connection with other methods, given its robustness for checking the results provided by them.

4. Applications

In this section we present two experimental situations which can greatly benefit from the characteristics and flexibility of the previously described methods. A practical astronomical application is also illustrated. Our aim is not to provide an exhaustive treatment of the field, but only to give an idea of the potentialities of the statistical techniques we have considered.

4.1. Bimodality

Frequently, in physical and astronomical research, it is necessary to decide if the empirical PDF of some observed quantity is multimodal or not. In fact the multimodality can imply the contribution of different populations or different physical processes. Lacking of any *a priori* information, the solution of such a problem is complex. In fact, due to the statistical fluctuations, the presence of two or more maxima in a given empirical PDF does not necessarily imply an intrinsic multimodality. On the other hand, in this case, the utility of Histogram is limited because of the *subjectivity* of the results; for example, a bimodality can appear or disappear according to the bin sizes and positions.

By exploiting the above mentioned inadequacy of the Johnson estimator to fit a bimodal PDF, a way to evaluate the possible bimodality of the PSF is represented by the following procedure:

1. Estimation of the PDF through $\hat{f}_K(x)$.

2. Estimation of the PDF through $\hat{f}_J(x)$. Being quite insensitive to the multimodality of the data distribution, this estimator provides a sort of unimodal interpolation of the empirical PDF.

3. Numerical simulation of a large number, say M, of random samples having the same size of the original one and extracted from $\hat{f}_J(x)$.

4. Calculation, for each *synthetic* sample, of $[\hat{f}_K(x)]_i$ ($i = 1, 2, ..., M$).

5. Calculation, at a fixed significance level, of the confidence interval for $\hat{f}_K(x)$, by using for each x the distribution of the values of $[\hat{f}_K(x)]_i$.

The philosophy behind this procedure consists in checking if an unimodal distribution is able to account for the observed data. Figure 2 shows the results obtained with two samples (each of 200 data) which have been extracted respectively from a truly bimodal distribution and from a distribution characterized by a developed tail. The second case reproduces a situation where the statistical fluctuations are important. In both cases the presence of several maxima is evident, but only the intrinsic bimodal distribution shows fluctuations exceeding the 90% confidence level band.

Obviously this procedure can not be considered a true statistical test of multimodality since we are actually testing if the results provided by the two methods are compatible or not. Therefore it must be taken *with a grain of salt.*

4.2. Rectification

Another common problem in astronomy is to estimate the PDF $\psi(\xi)$ when the observations, rather than for the random variable ξ, are available for the random variable x, which is related to ξ via the equation:

$$\phi(x) = \int \psi(\xi) P(x|\xi) \mathrm{d}\xi. \qquad (7)$$

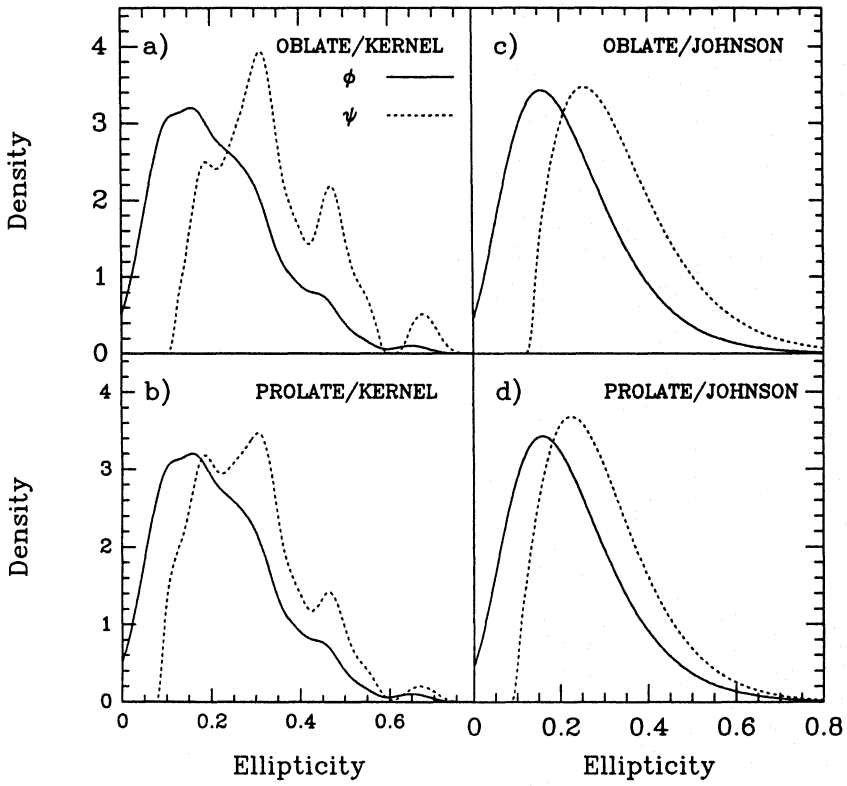

Fig. 5a–d. Deprojection (by means of the Lucy's method, 25 iteration) of the ellipticity distribution shown in Fig. 4 in both oblate and prolate case

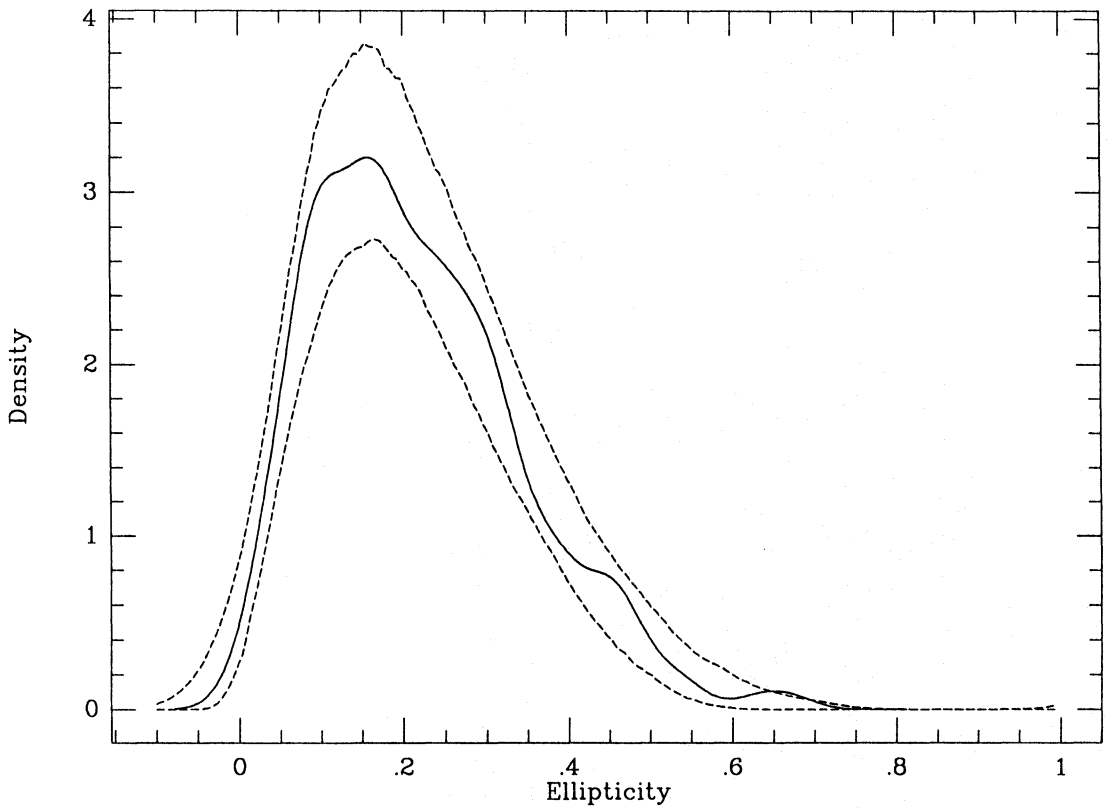

Fig. 6. Test of bimodality, with the method described in the text, for the distribution, obtained with the Kernel method, given in Fig. 4. The dotted lines represent the 90% confidence level interval

In this equation $P(x|\xi)$ is the conditional probability of x *given* ξ and $\phi(x)$ is the PDF of the observable. If $P(x|\xi) = F(x - \xi)$, Eq. (7) becomes a convolution integral. A considerable amount of work (e.g. Lucy 1974) has been devoted to evaluate the best method to solve Eq. (7) with respect to $\psi(\xi)$. However, perhaps it has been not fully understood that, when dealing with data, the most serious limitation in obtaining a reliable distribution $\psi(\xi)$ lies in the estimate of $\phi(x)$, rather than in the rectification algorithm itself. In other words, we are handling a statistical, rather than a mathematical, problem. For example, due to the statistical fluctuations, we could obtain an observed $\hat{\phi}(x)$ which is not compatible with $\psi(\xi)$ and $P(x|\xi)$. Moreover, it is well known that, in the case of the convolution integral, small fluctuations and/or biases of $\hat{\phi}(x)$ will result in much larger fluctuations and/or biases in $\hat{\psi}(\xi)$, with a consequent serious degradation of the results. Therefore, in facing a rectification problem, it is crucial to obtain a good estimate of the empirical distribution. This is clearly shown in Fig. 3, where we used the Lucy algorithm to deconvolve a distribution obtained by convolving a χ^2_3 distribution with a standard Gaussian. The sample size has been assumed to be of 200 data (a small sample for this kind of problem) and the empirical PDF has been estimated by means of the Histogram and Johnson estimators. The superiority of the result obtained with the Johnson method is evident and does not deserve any additional comment.

4.3. An astronomical application

An interesting problem in extragalactic astronomy is the determination of the intrinsic shape of elliptical galaxies. By assuming that elliptical galaxies are oblate or prolate spheroids randomly oriented in the space, the problem consists in evaluating their intrinsic ellipticity distribution $\psi(\varepsilon)$ ($\varepsilon = 1 - q$; $q = intrinsic\ axial\ ratio$) through the knowledge of the corresponding apparent ellipticity distribution $\phi(\varepsilon^\star)$ ($\varepsilon^\star = 1 - q^\star$; $q^\star = apparent\ axial\ ratio$). In formula:

$$\phi(\varepsilon^\star) = \int \psi(\varepsilon) P(\varepsilon^\star|\varepsilon) d\varepsilon \tag{8}$$

(see Noerdlinger 1979 for the analytical forms of $P(\varepsilon^\star|\varepsilon)$.

Figure 4 shows the Kernel and Johnson's estimates of the apparent ellipticity distribution $\phi(\varepsilon^\star)$ obtained with the sample of 204 ellipticals given in Fasano & Vio (1991). The two methods seem provide very similar results. The most significant difference is that the Kernel method provides a more "fluctuating" estimate. Actually, when performing the rectification (Fig. 5), the intrinsic distributions $\psi(\varepsilon)$'s turn out to be quite different. In particular the Kernel method suggests that the intrinsic ellipticity distribution of ellipticals could be multimodal, and this could indicate the existence of two or more different populations of ellipticals (Fasano & Vio 1991). Such point is rather interesting, since it could confirm, from a different point of view, the claim of many authors (see for example Bender et al. 1989) that elliptical galaxies can be roughly divided in two classes, namely the *disky* ellipticals and the *boxy* ones.

However, it is necessary to test if the fluctuations of $\phi(\varepsilon^\star)$, as provided by the Kernel method, are compatible with the statistical fluctuations of this estimator. To do this, we apply the method outlined in previous Sect. 4.1. Figure 6 shows that the multimodality of the rectified distribution could merely be a consequence of statistical fluctuations in the observed distribution. In other word, on the basis of the present small data sample, the intrinsic ellipticity distribution of elliptical galaxies is consistent with an unimodal distribution (see however Fasano & Vio 1991) peaked at about 0.2.

Here we note one important point: in this example the range of the distribution of the physical quantity is known in advance. However, although in principle possible (see Shapiro & Gross 1981), we have not used this information. In effect, from numerical experiments, we have recognized that forcing the solution to be within a fixed interval produces poor results. The explanation for such a (at first sight) paradoxical situation is that fixing the existence range of a PDF means to fix the values of one or two parameters in Eqs. (5) with a consequent loss in the "flexibility" of the curves.

Other applications of the techniques described in this work are in Maceroni et al. (1990) and Pisani (1993).

References

Bender R., Surma P., Döbereiner S., Möllenhoff C., Madejsky R., 1989, A&A 217, 35
Dudewicz E.J., Ramberg J.S., Tadikamalla P.R., 1974, An. Tech. Conf. Trans. Amer. Soc. Qual. Control 28, 407
Fasano G., Vio R., 1991, MNRAS 249, 629
Kendall M.G., Stuart A., 1958, The Advanced Theory of Statistics, Griffin, London
de Jager O.C., Swanepoel J.W.H., Raubenheimer B.C., 1986, A&A 170, 187
Jones M.C., Lotwick H.W., 1984, Applied Statistics 33, 120
Johnson N.L., 1949, Biometrika 36, 149
Linhart H., Zucchini W., 1986, Model Selection, Wiley, New York
Lucy L.B., 1974, AJ 79, 745
Maceroni C., Bianchini A., Rodonò M., Van 't Veer F., Vio R., 1990, A&A 237, 395
Noerdlinger P.D., 1979, ApJ 234, 802
Pisani A., 1993, MNRAS, (in press)
Shapiro S.S., Gross A.J., 1981, Statistical Modelling Techniques, Marcel Dekker
Silverman B.W., 1982, Applied Statistics 31, 93
Silverman B.W., 1986, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, New York
Slifker J.F., Shapiro S.S., 1980, Technometrics 22, 20