Computer Physics Communications 54 (1989) 381–390
North-Holland, Amsterdam

STATISTICAL DECISIONS UNDER NONPARAMETRIC A PRIORY INFORMATION

A.A. CHILINGARIAN

Yerevan Physics Institute, Markarian St., 2, SU-375036, Yerevan 36, Armenia, USSR

Received 3 April 1988; in revised form 12 September 1988

A program is developed for simular and experimental data handling. The main purposes are: the choice of the model most precisely describing the experiment, classification of particles and interaction processes. Procedures used: Bayes error calculation, K nearest neighbour density estimation, "Leave-one-out-at-a-time" test. Used nonparametric methods provide quantitative comparison of multivariate distributions and distribution mixture classification. Applications: high energy physics, cosmic ray physics.

PROGRAM SUMMARY

Title of program: KNN

Catalogue number: ABHS

Program obtainable from: CPC Program Library, Queen's University of Belfast, N. Ireland (see application form in this issue)

Computer: IBM 3090, EC 1045, PDP 11/70, IBM PC-AT

Operating system: VM/CMS, RSX, NORTON

Programming language used: FORTRAN 77

No. of bits in a word: 32, 16

No. of lines in combined program and test deck: 487

Keywords: Monte Carlo statistical inference, nonparametric methods, pattern recognition, multivariate analysis, Bayes risk estimation, probability density estimation, classification

Nature of physical problems
Choice of theoretical model most precisely describing experimental Data, extraction of events of a definite type, classification of particles and interaction processes.

Nature of statistical problems
Quantitative comparison of multivariate distributions, classification of distribution mixture, bump hunting, feature extraction.

Procedures used
Bayes decision making, "leave-one-out" test, K nearest neighbours (KNN) density estimation.

LONG WRITE-UP

0. Introduction

The scientific method is characterized by data classification, the study of their interrelations and relations to past experience, summarized in various theories and hypotheses. Usually, it is impossible either to prove or to refute hypotheses by deductive method. The challenge is to draw sensible conclusions from noisy, discrepant information.

The main aspect of applied statistics is collection and interpretation of data, the interpretative aspect being the one that is now regarded as the essence of the subject [1]. The fundamental idea of statistics is that useful information can be acquired from individual small bits of data. Inductive methods lead to empirical statements, that may be connected with theoretical ones by means of rational inductive conclusion rules [2].

However, it is very important to provide the scientist with an objective criterion by which he can judge the claims of hypotheses (models) under investigation. By model we mean a complete probability statement of what currently is supposed to be known a priori about the mode of generation of data and of uncertainty about the parameters [3].

If this statement consists in the existence of an analytic distribution family (like Poisson or Gaussian), appropriate to the problem at hand, we have a prescribed parametric model. For such parametric models a well known concept of statistical inference consists in obtaining estimates of its parameters and verifying the validity of the chosen family [4].

We shall restrict ourselves to the binary comparison case, that is, comparisons of two from many competing hypotheses at a time. Our example concerns a case where we want to realize the choice of one of two well-defined hypotheses-cosmic ray hadron classification by means of a Transition Radiation Detector (TRD) [5].

The classification problem is traditionally described in terms of null and alternative hypothesis, critical and acceptance regions, and level of significance [6].

The best critical region is constructed by means of the Likelihood Ratio (LR),

$$\mathrm{LR}(x) = \frac{p(x/\theta_{\pi}^{*})}{p(x/\theta_{\mathrm{pr}}^{*})}. \tag{0.1}$$

Here x is a many-dimensional observable, in our case the energy release in TRD layers. $p(x/\theta_{\pi}^{*})$ and $p(x/\theta_{\mathrm{pr}}^{*})$ are conditioned on particle type probability density functions, obtained separately for pions and protons. θ^{*} is a Maximal Likelihood Estimate (MLE)

$$\theta^{*} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{M} \ln\, p(x_i/\theta), \tag{0.2}$$

where set $\{x_i\}$, $i = 1,\ M$, is obtained from TRD callibration or, for superaccelerator energies, by simulation. M is the number of callibration or simulation trials.

Leaving apart the question about effectiveness of MLE for finite samples [7], we want to check, whether the MLE method permits us to do a very powerfull summary of data – we are summarizing a $\{x_i\}$ dataset by a probability density. Maximal Likelihood Summary (MLS) can be used for comparative purposes [8].

But for almost all problems of inference, the crucial question is whether the fitted probability family is in fact consistent with the data. Usually parametric models are chosen for their statistical tractability, rather than for their appropriatness to the real process being studied.

Of course, any statistical inference is conditioned on the model used, and, if the model is oversimplified, so, that essential details are either omitted, or improperly defined, at best only qualitative conclusions may be done. Now, in cosmic ray and accelerator physics very sophysticated models are being used, completely mimicking a stochastic mechanism where by data is generated. An example of such models is the Geant3 system, designed for detector description, simulation and optimization studies [9], and models of cosmic radiation propagation through atmosphere and detectors [10].

Such models are defined on a more fundamental level than parametric models, and provide us with a wide range of outcomes from identical input variable sets- "labled", or "training" samples (TS). These sets of events with known membership representing a general, nonparametric mode of a priori information. For our example, we obtained samples, corresponding to pion and proton traversals through TRD, i.e. ω_π and ω_{pr}.

So, usually, for experimental physics data handling, the likelihood function cannot be written explicitly, and we deal with implicit, nonparametric models, for which no parametric form of underlying distribution is known, or can be assumed.

Although using simulation to analyse data in high energy physics is wide-spread, we are aware of very few systematic investigations of theoretical aspects about how data may be compared with its simulated counterpart [11,12]. What we need is a well defined technique, what one can call Monte-Carlo Inference.

The term "Monte-Carlo Inference" at first appeared in the discussion of the valuable paper by Diggle and Gratton, where analitically intractable model fitting facilities were established [13].

This present paper considers classification and hypothesis testing problems in the framework of Bayesian paradigm [14]. The inference problem in Baysian approach is similarly described in terms of $[X, \Theta, D, P_\theta, p(x/\theta), L(d, \theta)]$, where X is an event (measurement, feature,...) space-collection of possible outcomes of random experiments, $\theta \ni \Theta$ is a parameter or index of various classes, types, hypotheses (further we shall denote these classes by w_i, or explicitely-by w_{pr} and w_π), P_θ is the prior density, $p(x/\theta)$ is the conditional density, D is the decision space, containing possible decisions, and $L(d, \theta)$ is the nonnegative loss function, defined on $D \times \Theta$.

For our example specification of prior density and loss function meets no difficulties: Θ-space includes two values pr and π, with prior probabilities P_{pr} and P_π, $(P_{\mathrm{pr}} + P_\pi = 1)$ which are obtained from a previous most confident experiment. If there is no such experiment, the uniform density can be used: $P_{\mathrm{pr}} = P_\pi = 0.5$.

For classification purposes a simple one to zero loss function is usually used: losses are equal to zero for a correct decision and one for any error.

1. Bayes decision rule, the measures of the closeness of empirical data and model

Bayssian approach provides the general method of incorporation of a priori and experimental information. Bayes theorem,

$$p(\omega_j/x) = \frac{P_j p(x/\omega_j)}{\sum_{i=1}^{L} P_i p(x/\omega_i)}, \tag{1.1}$$

gives us a posterior to the x-density, i.e. the probability of w_j-class(hypothesis) to be truth, if the x-value was observed, and before experiment the P_j prior density was assumed. L is the number of hypothesis under investigation.

The decision rule, that assigns observable x to the class with the highest a posteriori density w^* (Bayes decision rule), takes into account all usefull information and all possible losses due to any wrong decision,

$$w^* = \operatorname*{argmax}_i p(w_i/x), \quad i = 1, L. \tag{1.2}$$

For hadron classification Bayes decision rule takes the form

$$p(w_\pi/x) \lesseqgtr p(w_{pr}/x) \xrightarrow{\text{decision}} w^* \equiv \begin{cases} \mathrm{pr} \\ \pi \end{cases}. \tag{1.3}$$

So, the posterior density is the basis of statistical decisions on particle type and on simular and experimental data closeness. The term closeness refers to the degree of coincidence, similarity, correlation, overlapping or any such variable. Examples of these separability measures are the Kolmogorov variational distance (L1-metric) [15],

$$K_j = \int_X | p(\omega_{\mathrm{exp}}/x) - p(\omega_j/x) | \, \mathrm{d}x, \tag{1.4}$$

and the Bhattacharya distance (Hotteling coefficient),

$$B_j = \int_X \left(p(\omega_{\mathrm{exp}}/x) p(\omega_j/x) \right)^{0.5} \mathrm{d}x. \tag{1.5}$$

Another possibility to compare experimental and different model samples is the loglikelihood function estimation,

$$L_j = \sum_{i=1}^{M} \ln\, p(x_i/w_j), \tag{1.6}$$

where $p(x_i/w_j)$ is the probability density obtained with the w_j training sample, corresponding to the jth class and M is the number of observations.

The most convinient closeness measure, commonly used in pattern recognition problems for feature extraction [16], is Bayes error (Bayes risk for 0–1 loss function). Bayes classifier provides a minimum probability of error among all classifiers for the same feature set. The probability to misclassify the x observable, using Bayes rule, is equal to

$$r(x) = 1 - p(w^*/x), \tag{1.7}$$

or, for our example,

$$r(x) = \min\{\, p(w_{\pi}/x),\ p(w_{\mathrm{pr}}/x)\,\}. \tag{1.8}$$

Finally, Bayes error is determined by the expression

$$R = \int_X r(x)p(x)\,\mathrm{d}x, \tag{1.9}$$

where $p(x)$ is a mixture of distributions, representing the denominator of eq. (1.1).

However, it is impossible to calculate R and other distance measures, as the analytic expression of conditional densities and, hence, the posterior ones, is unknown. Therefore, we are obliged to use their nonparametric estimates. Nonparametric in the sense, that the density function is not a particular member of a previously chosen parametric distribution family, but an estimate based only on sample information and on very mild conditions on the underlying density (usually only continuity).

The nonparametric density estimation will be considered further; now we shall gain some insight into the methods of using the training sample in the procedure of Bayes error calculation.

Three main methods are distinguished [17]: the resubstitution–P method: the classifier is both trained and examined on the same sample i.e., first the conditional densities $p(x/w_i)$ are determined by TS, and then classification is carried out with the same sample; the holdout–H method: the TS is divided into two equal parts, with one half for the training and the other for examination. The P-method decreased Bayes risk; the H, on the contrary, increased it. Besides that the H-method does not use the TS effectively.

The leave-one-out–U method is free of such defects. One element is removed from the sample, the training is performed without it, then this element is classified and replaced in the TS and the procedure is repeated, untill all the TS elements have been classified. The U–method (also refered to as the cross-validation method) has been shown to have a much smaller bias than other methods, and it seems to be insensitive to data departure from normality [18].

The empirical error count was one of the first suggested estimation procedures. Let us introduce a random variable,

$$\epsilon(x) = \begin{cases} 0, & \text{if } x \text{ is classified correctly,} \\ 1, & \text{otherwise.} \end{cases} \tag{1.10}$$

The empirical error R^{e} is determined by

$$R^{\mathrm{e}} = \frac{1}{M}\sum_{i=1}^{M} \epsilon(x_i) = \frac{\#_{\mathrm{err}}}{M}, \tag{1.11}$$

where $\#_{\mathrm{err}}$ is the number of errors, commited during leave-one-out test over TS.

The estimate variance equals

$$\sigma^2_{R^{\mathrm{e}}} = \frac{1}{M}R(1-R). \tag{1.12}$$

Another type of estimate the so called average conditional error R^{p}, is connected with approximation of the expression (1.9),

$$R^{\mathrm{p}} = \frac{1}{M}\sum_{i=1}^{M} \min\{\, p(w_1/x_i),\ p(w_2/x_i)\,\}. \tag{1.13}$$

It is interesting to note, that the variance of this estimate proves to be less, then that of the previous one,

$$\sigma^2_{R^{\mathrm{p}}} = \frac{1}{M}R(1-R) - \frac{R}{2M}. \tag{1.14}$$

This result seems to be paradoxical, as in the second case the information on the x observable true lables is not used. This contradiction is explained by the fact that R^p takes on a continuum of values approaching R, while R^e takes on only discrete values, this quantization causing larger spread around R.

Comparing of experimental and model data can be considered as a two stage process. At first, relevant features must be selected from all available measerments. The feature selection problem can be viewed as an optimization problem, requiring a criterion function and a search procedure. We do not deal with methods of optimal subset selection in this paper; suppose we have an heuristic procedure to generate probable subsets, then we recomend Bayes error as a criterion function. Of course, one can use any other separability measure, but Bayes error is most straightforward and its estimation methods for high dimensionality spaces are well develop. Calculating Bayes error for different subsets of features, we shall select a subset that provides the maximal value of R; this subset will have the maximal differentiation power. With extracted features the tasks of classification, determination of different particle fractions in the mixture distribution, etc., can be carried out.

It must be mentioned, that actual values of both likelihood function and Bayes error have no statistical interpretation; only comparative conclusions may be done. If we want to evaluate statistical significance of our inference, some distribution free test must be used. We can recommend the permutation test [19] and the percentile test, using a bootstrup distribution, providing approximate confidence intervals in small sample nonparametric situations [20].

3. The nonparametric probability density estimation

The nonparametric density estimation methods have received a large development efforts in the last decade [see, e.g. ref. [21], mainly due to their simplicity and absence of excessive requirements on the form of the distribution function. Most popular kernel type estimates were introduced by Rosenblatt [22] and studied by Parzen [23]. In the Parzen procedure every point of TS, w_j, is substituted by a bell-like function, and the density in arbitrary point x of the feature space is obtained as a superposition of numerous "kernels" centered about each TS point,

$$p_h(x) = \frac{1}{M} \sum_{i=1}^{M} (1/h^N) K\{(x - x_i)/h\}, \qquad (2.1)$$

where $K(z)$ is a kernel function, satisfying $\int K(z)\, dz = 1$, kernel size h is a smoothing factor, determining the "spread" of the kernel (for Gaussian kernels $h \equiv$ mean square deviation), N is the dimensionality of the feature space, $\{x_i\} \ni w_j$.

Very close related to the Parzen estimates, KNN estimates were introduced by Fix and Hodges [24] and studied by Loftsgaarden and Quesenberry [25],

$$p_k(x) = \frac{K-1}{MV_k(x)}, \qquad V_k(x) = \frac{2^{N/2} d_k^N(x)}{N\Gamma(N/2)}, \qquad (2.2)$$

where $\Gamma(\cdot)$ is Gamma function, $V_k(x)$ is the volume of the N-dimensional sphere S, containing K elements of TS, nearest, in any convinient metric, to point x, $d_k(x)$ is the distance to the Kth nearest neighbour of point x.

Two metrics are usually used: the Euclidean metric and the Mahalonobis metric [26]. For the latter the distance between observable x and TS element y is equal to

$$D^2_{\text{Mah}} = (x-y)^T \Sigma^{-1} (x-y), \qquad (2.3)$$

where Σ is the covariance matrix, calculated by means of TS, to which y belongs. The use of the Mahalonobis metric allowes one to take into account the correlation information, moreover, the distances, calculated in this metric, are scale invariant, so, no transformation of initial data is necessarily.

The relationship of KNN to the Parzen estimate is brought out if the kernel function is uniform over region S, and zero outside, so only elements of TS within region S, centered at point

x, equally contributed in the density estimate.

Despite underlying density variations in feature space, in every point the density is estimated by K elements, and in low density regions the cell size is much greater than in high density regions. KNN estimates are "fixed event number in the cell" estimates, in contrast with "fixed cell size" histogram estimates.

If the K-value is small relative to TS size M, we have nearly the same probabilities for all TS sample points contributing to the estimate, and averaging over the S region, centered at a x point and containing K neighbours, is valid. But, on the other hand, it is desirable to have K as large as possible to reduce the influence of fluctuations in TS and achieve statistical stability. It is very difficult to choose the optimal K value to cohere these two requirments for different dimensionalities of the observation space, various underlying distributions and TS sizes.

Although many theoretical investigations prove the asymptotically unbiasness and consistence of kernel type estimates [27] and many recommandations have been done on kernel type and width and optimal K value, very little is known under finite sample condition. Recently, K. Fukunaga, the pioneer of KNN -and Parzen method developments for pattern recognition problems, claimed that "unrealiability of the estimators in finite conditions is the major obstacle toward their implementation in practice, and theoretically determined values of h or K gave very discouraging and inconsistent results" [28].

We tried to automate the procedure of optimal K value selection by using such a surprisingly powerfull technique as ordered statistics.

In ref. [29] it was suggested to calculate the KNN density estimates for several, IQ (IQ $< M$, usually IQ $= M/2$), different K values simultaneously. By means of a sequence of estimates, $\{p_k(x)\}$, $k = 1$, IQ, the averaged estimate, KNN3, is constructed,

$$p_{\mathrm{IQ}}(x) = \frac{1}{\mathrm{IQ}} \sum_{k=1}^{\mathrm{IQ}} p_k(x). \tag{2.4}$$

This estimate uses more detailed information on the neighbourhood of point x and is more stable. However, in the sequence of estimates $\{p_k(x)\}$ there may be significant deviations from the true density value, due to large variance of the simple KNN estimate, that distorts the average estimate.

To weight effectively every member of the sequence, we introduce a new KNN estimate in the form of a linear combination of ordered statistics,

$$p_{[\mathrm{IQ}]}(x) = \sum_{k=1}^{\mathrm{IQ}} c_i p_{[k]}(x), \qquad \sum_{k=1}^{\mathrm{IQ}} c_i = 1, \tag{2.5}$$

where square brackets [ ] indicate, that the sequence $\{p_{[k]}(x)\}$ is ordered according to the magnitude of the members.

If we heavily weight the members in the middle of the ordered sequence, we shall obtain a stable, with respect to the fluctuations in TS, estimate. An example of such an estimate is the median estimate, $p^{\mathrm{med}}_{[\mathrm{IQ}]}(x)$, for which

$$c_i = \begin{cases} 1, & \text{if } i = |\mathrm{IQ}/2| + 1, \\ & \mathrm{IQ} \text{ is an odd number}, \\ 0, & \text{otherwise}. \end{cases} \tag{2.6}$$

$|\cdot|$ stands for the whole part of a number. If IQ is an even number, two middle order statistics are weighted with 0.5. So, for every x point, there will be done an unique choice of a K value (or two K values for even IQ). In the middle of an ordered sequence it will appear the most stable member of the sequence. This, self-adjusting character of the median estimate, as we shall see further, leads to estimates better, than one can obtain with any fixed K value.

Our K-dependence investigations [30] show, that for regions with very low density (so called, peripheral regions) simple KNN estimates with $K = 2/7$ (for TS size 50/400), are preferable. It can be explained by the fact, that for points in this region, whenever NN "enters" a high density region, further increase in K will not lead to any significant increase of the KNN volume. Hence, as one can see from (2.2), values of most terms of the ordered sequence will be overestimated, and median estimate will be optimistically biased. Therefore, at the peripheral points, that are chosen according to the relative size of local region it

will be better to calculate the density by a simple KNN rule.

For every point x a local neighbourhood size is defined as,

$$\rho_{\mathrm{LOC}}(x_i) = \frac{1}{\mathrm{MSQ}} \sum_{j=1}^{\mathrm{MSQ}} d_j(x_j), \tag{2.7}$$

where $d_j(x)$ is the distance to the jth nearest neighbour of a point x_i and $\mathrm{MSQ} = \sqrt{M}$. And, if this "local region size" is 3 times greater, than

$$\bar{\rho}_{\mathrm{LOC}} = \frac{1}{M} \sum_{i=1}^{M} \rho_{\mathrm{LOC}}(x_i), \tag{2.8}$$

the density at point x_i is estimated by a simple KNN rule. Median estimates with such peripheral points corrections, we shall call adaptive KNN estimates, $p^{\mathrm{ad}}_{[\mathrm{IQ}]}(x)$.

KNN density estimation enables effective control of the degree of smoothing of empirical distributions. Displaying the same distribution by different modifications of the KNN method, the peaks in the distribution will become more evident.

Recently an evidence of a narrow enhancement in the pp mass distribution from reaction $\gamma d \to pp\pi^-$ was reported [31]. The statistical significance of this enhancement was proved by the KNN density estimation method, showing the observed structure of the mass distribution better, than the Histogram method. So, KNN median and adaptive modifications can also be very usefull for resonance data analysis.

4. Information input and output

The tasks of particle classification and Bayes risk estimation, of obtaining smooth estimates of the probability density, etc. are recomplished with the KNN module. Some simple subroutines are used for nearest neighbour distances calctlation and ordering, local region size determination (DISTNN,DIST,ORDER,RSLOC). Covariance matrix calculation and inverting is done by MISR, CORINV and SMXINV modules.

The exchange of information with the KNN module is realized by means of formal parameters only (no COMMON blocks are used), though it leads to a rather long list of parameters, the undesirable collateral effects of the module on the main program are practically excluded, and high obviousness of operation of the module is provided [32].

All statistical procedures are carried out simultaneously for different modifications of the density estimator. The computation load, however, increased only slightly, since wasting most time ordering NN distances, is done only once. In the presented version of the program the ordering is carried out by simple DO-loops but it can easily be changed by some fast procedure e.g. FLPSOR from the CERN program library, MO1AJF from the NAG library, or by a special routine for finding NN, based on ordered partition of each projection axis [33].

The control parametres and arrays of the module are:

N is the features space dimensionality; NS is the dimensionality of the subset under investigation; M is the size of the TS; L is the number of classes (for simplicity, it is assumed here, that all classes have the same TS size); MP is the control sample size; IQ is the size of ordered densities sequence; METRIC is the code for choosing the metric in which distances to NN are calculated, the value METRIC = 0 corresponds to the Mahalonobis metric; the IMODE parameter chooses the program is operation regime: IMODE = 1, classification and likelihood function calculation, IMODE = 2, Bayes error calculation, IMODE = 3, density estimation.

D(IQ,L,MP) is the basic neighbourhood information, distances from every point of the control sample to each class of the training sample; NL(N) is the code combination, by means of which the feature subset selection from primary observables is realized; code 1 in the ith position signifies the ith feature inclusion in the subset under investigation; AP(L) are the prior probabilities of the models being studied; PST(L) are current values of posterior densities; RM(L) is the mean local region size for TS classes; C(IQ) are the coefficients of the order statistics. During the operation in

"leave-one-out" for Bayes error calculation, the number of TS classes L = 2, and in the subroutine DISTNN the control sample CS(MP) is substituted by the training one.

The number of events, separately of the w_i, $i = 1$, L, classes, classified as representatives of the w_j, $j = 1$, L, classes, by means of the various modifications of the KNN method, are placed in the array CL(L,L,IQ + 2), where

$$\mathrm{CL(L,L,}i\mathrm{)} = \begin{cases} i = 1, 2, \ldots, \mathrm{IQ} \\ \quad \text{error rates for simple KNN,} \\ \quad \text{eq. (2.2)} \\ i = \mathrm{IQ} + 1, \\ \quad \text{rates for median estimate,} \\ \quad \text{eqs. (2.4), (2.5)} \\ i = \mathrm{IQ} + 2, \\ \quad \text{rates for adaptive estimate.} \end{cases} \tag{3.1}$$

This information enables calculation of misclassification rates for every model separately, e.g. $R_{\pi \to \mathrm{pr}}$ and $R_{\mathrm{pr} \to \pi}$, where

$$R = P_\pi R_{\pi \to \mathrm{pr}} + P_{\mathrm{pr}} R_{\mathrm{pr} \to \pi}. \tag{3.2}$$

Using these rates, one can reconstruct the true fraction of iron nuclei in the primary flux [34] or the pion fraction in the hadron flux:

$$P_\pi = \frac{P_\pi^* - R_{\pi \to \mathrm{pr}}}{1 - R_{\pi \to \mathrm{pr}} - R_{\mathrm{pr} \to \pi}}, \tag{3.3}$$

where P_π^* is portion of experimental events classified as pions by Bayes decision rule, $R_{\pi \to \mathrm{pr}}$ and $R_{\mathrm{pr} \to \pi}$ are error rates obtained from the same TS.

Values of the Bayes error, calculated by averaging the posterior error R^{p} are returned in the RP(IQ + 2) array. The R^{e} values, the misclassification rates, are returned in the RE(IQ + 2) array. The order of values in the arrays is analogous to CL.

In the classification regime the class lables, to which the control events are refered, are placed in the array LUM(MP). Likelihood function values are returned in the FLIK(L) array.

In the regime of constructing the density smoothed estimate, density estimates for the L classes of TS, are placed in the DS(MP,IQ + 2) array.

Besides the presented, some operational arrayes are used, PKNN(IQ) and DEN(L,IQ + 2).

5. Program testing

Program testing was performed with the use of samples generated according to the normal (Gaussian) distribution. The choice of this distribution is due to its extensive use as a simple test to compare various density estimation methods, as well as due to the simplicity of calculating the true value of Bayes error, allowing a comparison of the estimates with the calculated values,

$$R = \Phi(-D_{\mathrm{mah}}/2), \tag{4.1}$$

where Φ is the cumulative normal distribution function and D_{mah} is the Mahalonobis distance between mathematical expectations of the two classes.

From the program output one can see results of the KNN operation in different modes: Bayes classification and likelihood estimation of one of the two normal population's, $N(0, 1)$ and $N(1, 1)$, control sample is from the same population as the first class of the training one, the likelihood function for the "true" class (hypothesis) is much greater than for the second class: Bayes error values for different KNN modifications can be compared with the true value, $R = 0.3085$; density KNN estimates are calculated for 27 different estimation modifications.

For investigations of bias and consistency of KNN estimates several random samples of fixed size were generated from the normal distribution; the density was estimated at 51 points, uniform distributed over an interval $(-5, 5)$, then the mean square error (MSE) was calculated,

$$\mathrm{MSE}\{\hat{p}(x)\} = E\{\{p(x) - \hat{p}(x)\}^2\}, \tag{4.2}$$

where the mathematical expectation is taken over all possible samples of fixed size drawn from

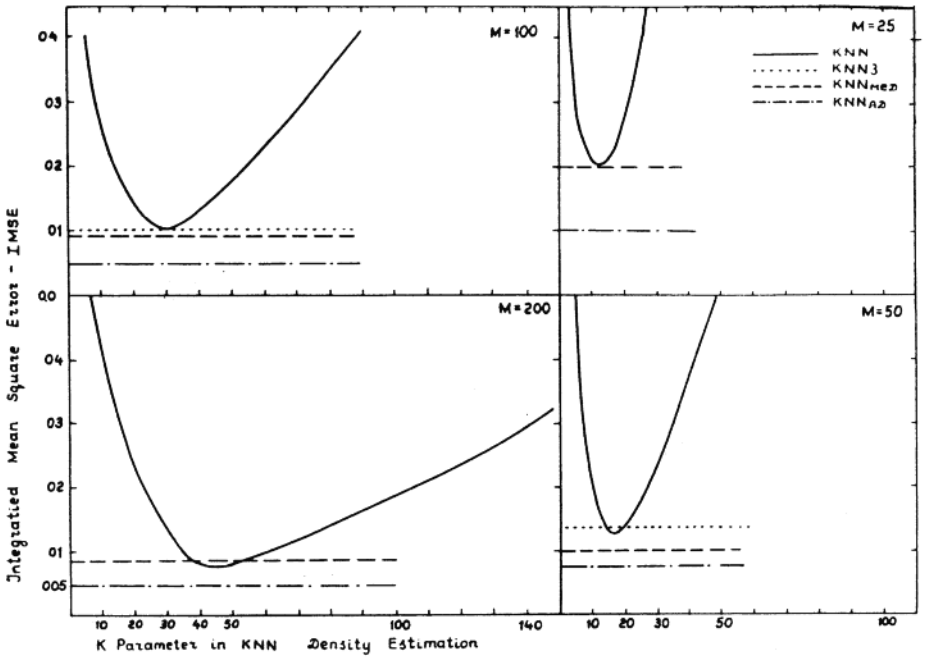

Fig. 1. The test results for the comparison of simple KNN, average, median and adaptive estimates.

general population; $\hat{p}(x)$ is one of considered KNN density estimates.

The integrated mean square error (IMSE) equals

$$\mathrm{IMSE}_{\hat{p}} = \int_X \mathrm{MSE}\{\hat{p}(x)\}\,\mathrm{d}x, \tag{4.3}$$

where $\mathrm{MSE}(x)$ was obtained by averaging density estimates, calculated with 25 independent samples. This procedure was repeated 10 times to evaluate the mean and standard deviation of IMSE.

Figure 1 showes the results of testing for the comparison of simple KNN, average, median and adaptive estimates. Standard normal density $N(0, 1)$ was estimated. Obviously, the adaptive estimates are much more precise than the ones obtained with any fixed parameter K.

Table 1 presents the results of comparing the adaptive KNN method with the Parzen method and MLE with the penalty function (the latter two are taken from ref. [27]. IMSE and its variance (in brackets) are calculated by 25×10 independent samples from $N(0, 1)$ and bimodal distribution $p(x) = 0.5N(-1.5, 1) + 0.5N(1.5, 1)$. The density was estimated in 51 points on the interval $(-5, 5)$.

The adaptive estimates have shown somewhat worse results. However, it should be noted that the parameters of the Parzen and ML methods have been chosen with the aid of information on the true underlying density, while the adaptive KNN

Table 1
IMSE of various nonparametric density estimators

Distribution	TS size M	MLM with penalty	Parzen with Gauss. kern.	Adaptive KNN
$N(0, 1)$	25	0.0100(0.0080)	0.0160(0.0120)	0.0140(0.0100)
$N(0, 1)$	100	0.0037(0.0021)	0.0050(0.0027)	0.0052(0.0020)
$N(0, 1)$	400	0.0015(0.0008)	0.0020(0.0009)	0.0032(0.0012)
Bimodal	25	0.0100(0.0030)	0.0090(0.0700)	0.0012(0.0030)
Bimodal	100	0.0036(0.0007)	0.0036(0.0020)	0.0048(0.0017)

Table 2
Comparison of Bayes error estimation R^e and R^p methods

N	M	R	σ_{R^e}	σ_{R^p}	$\hat{\sigma}_{R^e}$	$\hat{\sigma}_{R^p}$	MSE_{R^e}	MSE_{R^p}
1	100	0.3085	0.046	0.024	0.039	0.026	0.0005	0.0013
1	200	0.3085	0.033	0.017	0.026	0.015	0.0003	0.0008
2	50	0.2340	0.048	0.031	0.044	0.027	0.0030	0.0039
8	50	0.5000	0.070	0.050	0.050	0.012	0.0050	0.0044
10	50	0.5000	0.070	0.047	0.047	0.017	0.0040	0.0046

method uses only the sample information. In processing real data, of course, the analytical density form is unknown, therefore, the slight deterioration of accuracy is compensated with improved procedure stability.

The Bayes risk was calculated for the samples from normal distribution at various feature space dimensionalities and Mahalanobis distances. The calculations were performed with use of 10 independent samples. Table 2 presents the variances, calculated by eqs. (1.12) and (1.14), their estimates and mean square errors. A good agreement of sample and theoretical values of variances is apparent, and although the variances of the R^p values are less than those of R^e, their bias leads to their greater mean square deviation. Therefore, the empirical error count estimator is more preferable, especially for feature spaces of large dimensionality.

References

[1] E. Lederman, Handbook of Applied Mathematics: Statistics (Wiley, New York, 1984).
[2] P. Hajek and T. Havranek, Mechanizing Hypothesis Formation (Springer, Heidelberg, 1979).
[3] G.E.P. Box, Technometrics, 26 (1984) 1.
[4] E.A. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, Statistical Methods in Experimental Physics (North-Holland, Amsterdam, 1971).
[5] A.A. Chilingarian, in: Proc. of the Symp. on High-Energy Particles Transition Radiation, Yerevan, 1984.
[6] S. Zacks, The Theory of Statistical Inference (John Wiley & Sons, New York, 1977).
[7] J. Berkson, Ann. Stat. 8 (1980) 457.
[8] B. Efron, Ann. of Stat., 10 (1982) 340.
[9] R. Brun, F. Bruyant, M. Maire, A.C. McFerson and P. Zanarini, GEANT3, CERN Preprint, DD/EE/84-1 (1986).
[10] A.M. Dynaevsky et al., in: Proc. FIAN, Moscow, 1984.
[11] J.N. Friedman, Data Analysis Techniques for High-Energy Physics, CERN Yellow Report (1974).
[12] A.A. Chilingarian, VANT, Ser. Tecn. Phys. Exp., Kharkov, 1981.
[13] P.J. Diggle and R.J. Gratton, J. Roy. Statist. Soc. B 46 (1984) 193.
[14] D.V. Lindley, Bayesian Statistics, (Soc. for Indust. and Appl. math., Philadelphia, 1978).
[15] M. Yablon and J.T. Chu, IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-2 (1980) 97.
[16] K. Fukunage and R.D. Short, IEEE Trans. on Informormation, IT-26 (1980) 59.
[17] G.T. Toussaint, IEEE Trans. on Information, IT-20 (1974) 472.
[18] S.M. Snappin and J.D. Knoke, Technometrics, 26 (1984) 371.
[19] F. James, Determining the Statistical Significance of Experimental Results, CERN Preprint DD/81-02 (1981).
[20] B. Efron, Canadian J. Statist. 9 (1981) 139.
[21] L. Devroye and L. Gyorfi, Nonparametric Density Estimation. The L1 View, (Wiley, New York, 1985).
[22] M. Rosenblatt, Ann. Math. Stat. 27 (1956) 832.
[23] E. Parzen, Ann. Math. Stat. 33 (1962) 1065.
[24] E. Fix and J.L. Hodges, Project 21-49-004, Report 4, USAF School of Aviation Medicine, Randolf Field, Texas (1951).
[25] D.O. Lofsgaarden and C.D. Quesenberry, Ann. Math. Stat. 36 (1965) 1049.
[26] P.C. Mahalonobis, Proc. of the Nat. Inst. of India 2 (1936) 49.
[27] R.A. Tapia and J.R. Thompson, Nonparametric Probability Density Estimation (The John Hopkins University Press, Baltimore and London, 1978).
[28] K. Fukunaga and D. Himmels, IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI9 (1987) 634.
[29] L.R. Rabiner, E. Levinson, A.E. Rozenberg and J.G. Wilpon, IEEE Trans. on Acoustics, Speech, Signal Processing, ASSP-27 (1974) 336.
[30] A.A. Chilingarian and S.Kh. Galfayan, Stat. Problems of Control, Vilnius, 66 (1984) 66.
[31] B. Bock, W. Ruhw et al, Nucl. Phys. A 459 (1986) 573.
[32] P.E. Gill, W. Murray and S.M. Piches, ACM Trans. of Math. Software, 5 (1979) 266.
[33] B.S. Kim and S.B. Park, On Pattern Analysis and Machine Intelligence, PAMI-8 (1986) 761.
[34] V.G. Denisova, A.M. Dunaevsky, S.A. Slavatinsky et al., in: Proc. of the 20th ICRC, Moscow, 1987, p. 390.