Kernel Density Estimator by Minimizing Bias

kaunas university of technology
ktu 1922

INTRODUCTION

A histogram is one of the simplest and the oldest density estimators. This graphical representation was first introduced by Karl Pearson in 1891. For the approximation of density $f(x)$, the number of observations $X(t)$ falling within the range of Ω is calculated and divided by n and the volume of area Ω. The histogram produced is a step function and the derivative either equals zero or is not defined (when at the cut off point for two bins). This is a big problem if we are trying to maximize a likelihood function that is defined in terms of the densities of the distributions.
It is remarkable that the histogram stood as the only nonparametric density estimator until the 1950's, when substantial and simultaneous progress was made in density estimation and in spectral density estimation. In 1951, in a little-known paper, Fix and Hodges introduced the basic algorithm of nonparametric density estimation; an unpublished technical report was published formally as a review by Silverman and Jones in 1989. They addressed the problem of statistical discrimination when the parametric form of the sampling density was not known. During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt in 1956, Parzen in 1962, and Cencov in 1962. Then followed the second wave of important and primarily theoretical papers by Watson and Leadbetter in 1963, Loftsgaarden and Quesenberry in 1965, Schwartz in 1967, Epanechnikov in 1969, Tarter and Kronmal in 1970 and Kimeldorf and Wahba in 1971. The natural multivariate generalization was introduced by Cacoullos in 1966. Finally, in the 1970's the first papers focusing on the practical application of these methods were published by Scott et al. in 1978 and Silverman in 1978. These and later multivariate applications awaited the computing revolution.

METHODS

The basic kernel estimator $\hat{f}(x)$ with a kernel function *K* and a fixed (global) bandwidth *h* for multivariate data $X \epsilon R^d$ may be written compactly as:

$$f(x) = \frac{1}{nh^d}\sum_{t=1}^{n} K\left(\frac{x - X(t)}{h}\right)$$

The kernel function *K(u)* should satisfy the condition:

$$\int_{-\infty}^{+\infty} K(u)\,du = 1.$$

Usually, but not always, *K(u)* will be a symmetric probability density function $K(u) = K(-u)$ for all values of *u*.
At first, the data is usually prescaled in order to avoid large differences in data spread. A natural approach is first to standardize the data by a linear transformation yielding data with zero mean and unit variance. As a result, first equation is applied to the standardized data. Let *Z* denote the sphered values of random *X*:

$$Z = S^{\frac{-1}{2}} * (X - \bar{X})$$

where $\bar{X}$ is the empirical mean, and $S \epsilon R^{d \times d}$ is the empirical covariance matrix. Applying the kernel density estimator to the standardized data *Z = (Z(1), …, Z(n))* yields the following estimator of density function *f(x)*:

$$f_z(z) = \frac{1}{nh^d}\sum_{t=1}^{n} K\left(\frac{z - Z(t)}{h}\right)$$

$$f(x) = \frac{(\det S)^{\frac{-1}{2}}}{nh^d}\sum_{t=1}^{n} K\left(S^{\frac{-1}{2}}\frac{x - X(t)}{h}\right)$$

The comparative analysis of estimation accuracy was made for four different types of kernels. The first three kernels are classical, whereas the last one is new.
The Gaussian kernel is consistent with the distribution of normal *φ(x)* selection:

$$K_G(x) = \varphi(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} e\left(\frac{-x^T x}{2}\right).$$

The Epanechnikov kernel is the second order polynomial, corrected to satisfy the properties of the density function:

$$K_E(x) = \frac{d+2}{2V_d}(1 - x^T x)\mathbf{1}_{\{|x^T x \le 1|\}}$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the *d*-dimensional unit sphere, and $\Gamma(u) = \int_0^\infty y^{u-i}\, e^{-y} dy$.
The Triweight kernel proposed by Tapia and Thompson in 1978 has better smoothness properties and finite support. It was investigated in detail by Hall in 1985:

$$K_T(x) = \frac{(d+4)(d+6)}{24}\frac{(d+2)}{2V_d}(1 - x^T x)^3 \mathbf{1}_{\{|x^T x \le 1|\}}.$$

The new kernel K_{New} has lighter tails than Gaussian distribution density and was introduced by the authors of this article:

$$K_{New}(x) = \varphi\big(\bar{g}(x)\big)\bar{g}'(x)$$

The main feature of this kernel function is that its form is chosen in such a way as to minimize the bias that occurs in the construction of the criterion using sample values. The construction of the kernel is chosen so that the influence of the used sample point on the constructed estimate is smaller than the environment of that point.
The class of the selected parametric function $\bar{g}$ depends on three parameters:

$$\bar{g} = \left|\prod_{i=1}^{d} x_i\right|^{1/d}\left(c + \left|\left|\prod_{i=1}^{d} x_i\right|^{1/d}\right|^{b}\right)^{a}, \qquad a, b, c > 0,$$

$$\bar{g}' = \left(c + \left|\left|\prod_{i=1}^{d} x_i\right|^{1/d}\right|^{b}\right)^{a} + a\left|\left|\prod_{i=1}^{d} x_i\right|^{1/d}\right|\left(c + \left|\left|\prod_{i=1}^{d} x_i\right|^{1/d}\right|^{b}\right)^{a-1} b\left|\left|\prod_{i=1}^{d} x_i\right|^{1/d}\right|^{b-1},$$

This kernel function depends on the spread *a*, trough *b* and peak shape *c* parameters.

KERNEL FORM

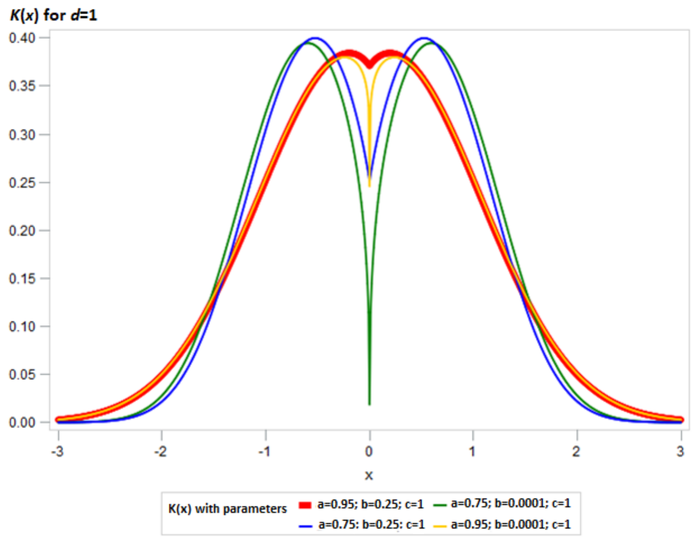

Figure 1 The form of the proposed kernel function with several parameter sets

BANDWIDTH

There are three parameters in kernel density estimator: the sample size *n*, the kernel function *K(·)* and the bandwidth *h*. Quite typically we cannot do anything about the sample size and we have to make the best out of the situation by choosing an appropriate kernel and a suitable bandwidth. It is well known that the bandwidth selection is the most crucial step in order to obtain a good estimate. Unfortunately, bandwidth selection is the most difficult problem in kernel density estimation and a definite and unique solution to this problem does not exist.
It is rather surprising that the most effective bandwidth selection method is a visual assessment by the researcher. The researcher visually compares different density estimates, based upon a variety of bandwidths and then chooses the bandwidth that corresponds to the subjectively optimal estimate. The unfortunate part is that such bandwidths are non-unique; this method will yield different bandwidths when performed by different researchers. This method can also be very time consuming.
The approach based on mathematical analysis is to quantify the discrepancy between the estimate and the target density by evaluated error criterion. The optimal bandwidth will then be the bandwidth value that minimizes the error measured by the error criterion. Such a method is objective and can be time-efficient as computers can solve it numerically.
A global measure of precision is the asymptotic mean integrated squared error (AMISE):

$$AMISE\left(\hat{f}(x)\right) = \frac{K_v^2(K)}{(v!)^2} R(\nabla^v f) h^{2v} + \frac{R(K)^d}{nh^d},$$

where $\nabla^v f(x) = \sum_{k=1}^{d} \frac{\partial^v}{\partial x_k^v} f(x)$ and $R(g) = \int_{-\infty}^{\infty} g(u)^2 du$ is the roughness of a function. The order of a kernel, *v*, is defined as the order of the first non-zero moment $\kappa_j(K) = \int_{-\infty}^{\infty} u^j K(u) du$. For example, if $\kappa_1(K) = 0$ and $\kappa_2(K) > 0$ then *K* is a second-order kernel and *v* = 2. If $\kappa_1(K) = \kappa_2(K) = \kappa_3(K) = 0$ but $\kappa_4(K) > 0$ then *K* is a fourth-order kernel and *v* = 4.The order of a symmetric kernel is always even. Symmetric non-negative kernels are second-order kernels. A kernel is higher-order kernel if *v* > 2. These kernels will have negative parts and are not probability densities.
The optimal bandwidth is:

$$h_0 = \left(\frac{(v!)^2 d R(K)^d}{2vK_v^2(K) R(\nabla^v f)}\right)^{1/(2v+d)} n^{-1/(2v+d)}.$$

The optimal bandwidth depends on the unknown quantity $R(\nabla^{(v)} f)$. For a rule-of-thumb bandwidth, Silverman proposed that it is possible to try the bandwidth computed by replacing *f* in the optimal formula by g_0 where g_0 is a reference density – a plausible candidate for f, and $\hat{\sigma}$ is the sample standard deviation. The standard choice is a multivariate normal density. The idea is that if the true density is normal, then the computed bandwidth will be optimal. If the true density is reasonably close to the normal, then the bandwidth will be close to optimal.
Calculation of that is proceeded according to

$$R(\nabla^v \varphi) = \frac{d}{\pi^{\frac{d}{2}} 2^{d+v}}\left((2v-1)!! + (d-1)\left((v-1)!!\right)^2\right)$$

where the double factorial means (2s + 1)!! = (2s + 1) (2s – 1) ... 5 · 3 · 1. Making this substitution, we obtain

$$h_0 = C_v(K, d) n^{-1/(2v+d)}$$

where $C_v(K,d) = \left(\frac{\frac{d}{\pi^2 2^{d+v-1}(v!)^2 R(K)^d}}{vK_v^2(K)\left((2v-1)!! + (d-1)\left((v-1)!!\right)^2\right)}\right)^{1/(2v+d)}$. and this assumed that variance is equal to 1. Rescaling the bandwidths by the standard deviation of each variable, we obtain the rule-of-thumb bandwidth for the i^{th} variable is

$$h_i = \hat{\sigma}_i\, C_v(K, d) n^{-\frac{1}{2v+d}}.$$

Table 1 provides the normal reference rule-of-thumb constants ($C_v(K,d)$) for the second-order *d*-variate kernel density estimator. First, in the common setting of a second order kernel (*v* = 2) the rule-of-thumb constants are decreasing as d increases. Scott (1992) notes that these reach a minimum when *d* = 11. The *v* = 2 case is the only one he considers. When *v* > 2, it is possible to show that the rule-of-thumb constants are increasing in the dimensionality of the problem. The basic idea behind this is given that higher-order kernels reduce bias, larger bandwidths are needed to minimize AMISE. However, note that the increase is not uniform over *v*.

Table 1 Normal reference rule-of-thumb constants (Cv(K,d)) for the multivariate second-order kernel density estimator

Kernel	d = 1	d = 2	d = 3	d = 4	d = 5	d = 6	d = 7	d = 8	d = 9	d = 10
Gaussian	1.059	1.000	0.969	0.951	0.934	0.933	0.929	0.927	0.925	0.925
Epanechnikov	2.345	2.191	2.120	2.073	2.044	2.025	2.012	2.004	1.998	1.995
Triweight	3.155	2.964	2.861	2.800	2.762	2.738	2.723	2.712	2.706	2.702
New	1.142	1.079	1.045	1.025	1.014	1.007	1.002	1.000	0.998	0.998

FUTURE WORKS

The proposed function of the kernel is still underexplored and requires further in-depth studies. One such study is parameterization to obtain the smallest possible mean integrated squared error when estimating an unknown distribution density.

AUTHORS

T. Ruzgas, K. Pupalaigė

Kaunas University of Technology, Lithuania