Shape-Adaptive Kernel Density Estimation

L.E.N. Baakman

November 28, 2016

1 Introduction

Density estimation tries to find the density $f(\mathbf{x})$ in d-dimensional Euclidean space underlying N points $\mathbf{x}_1 \dots \mathbf{x}_N$, that have been selected independently from $f(\mathbf{x})$.

Kernel density estimation has recently been used to predict dose-volume histograms, these histograms are used to determine radiation doses [5]. Ecologists have explored the habitats of seabirds with density estimation [4]. Density estimation has been described as "a critical first step in in making progress in many areas of astronomy." [3] Astronomers are for example interested in the an estimation of the cosmic density field, which is required for the reconstruction of the large-scale structure of the universe.

One often used method in density estimation is the Parzen approach [6], which gives the following estimate of the density function:

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\sigma^d} K\left(\frac{\mathbf{x} - \mathbf{x}_j}{\sigma}\right). \tag{1}$$

Thus the estimated density is the mean of bumps placed at each observation. The shape of these bumps is determined by the shape of the kernel function $K(\cdot)$, their width is controlled by the bandwidth σ [7]. The Parzen approach requires the kernel to be a probability density function, i.e. $K(\mathbf{x}) \geq 0$ and $\int K(\mathbf{x}) = 1$.

One downside of the Parzen method is that it cannot respond appropriately to variations in the magnitude of the density function, i.e. the peakedness of the kernel is not data-responsive. Consequently in regions of low $f(\mathbf{x})$ that contain only one sample point, $\mathbf{x}$, the estimate will have a peak at $\mathbf{x}$ and be too low in the rest of the region. In areas where the density is high, the sample points are more densely packed together, and the Parzen estimate will tend to spread out the high density region [1]. Breiman, Meisel, and Purcell introduced an variant of the Parzen estimator that addresses this disadvantage by making the sharpness of the kernel responsive to the local data. This variant defines the density estimate as

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} (\alpha d_{j,k})^{-d} K_{\mathscr{G}}\left(\frac{\mathbf{x} - \mathbf{x}_j}{\alpha d_{j,k}}\right), \tag{2}$$

where $K_{\mathscr{G}}(\cdot)$ represents a Gaussian kernel, α is a multiplicative constant and $d_{j,k}$ the distance between $\mathbf{x}_j$ and the k nearest neighbor of $\mathbf{x}_j$. Comparing Equation (1) with (2) we find that the bandwidth σ, has been replaced with $\alpha d_{j,k}$. In low density regions $d_{j,k}$ will be large, and the kernel will be spread out, in high density regions the converse occurs. Breiman, Meisel, and Purcell use a minimization algorithm on a goodness of fit statistic to find suitable values for k and α.

The minimization procedure used by Breiman, Meisel, and Purcell implicitly uses a k-NN pilot estimate. If pilot densities are used explicitly the density estimation process becomes [7]:

(i). Find a pilot estimate $\tilde{f}(\mathbf{x})$ that satisfies $\forall i\, \tilde{f}(\mathbf{x}_i) > 0$.

(ii). Define local bandwidth factors λ_i by

$$\lambda_i = \left(\frac{1}{g}\tilde{f}(\mathbf{x}_i)\right)^{-\beta} \tag{3}$$

where g is the geometric mean of the $\tilde{f}(\mathbf{x}_i)$, and the sensitivity parameter $\beta \in [0, 1]$.

(iii). Compute the adaptive kernel estimate as

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} (\sigma\lambda_i)^{-d} K\left(\frac{\mathbf{x} - \mathbf{x}_j}{\sigma\lambda_i}\right). \tag{4}$$

Often the pilot density is estimated with a fixed kernel method. The approach taken by Breiman, Meisel, and Purcell is computationally expensive, partially due to the use of the Gaussian kernel. The infinite base of this kernel means that an exponential function has to be evaluated for each data point to estimate the density of one data point according to Equation (2). Wilkinson and Meijer [8] propose to reduce this computational complexity in two ways, firstly they replace the infinite base Gaussian kernel with an Epanechnikov kernel, which not only has a finite base, but is also optimal in the sense of the Mean Integrated Square Error (MISE) [2]. Secondly they computed the the pilot densities on a grid including all data points and determined the pilot densities with multi-linear interpolation. Wilkinson and Meijer used Equation (4) with an Epanechnikov kernel, $\lambda_i = 1$ and

$$\sigma = \left(\frac{8\,(d+4)\,(2\sqrt{\pi})^d}{c_d} \right)^{\frac{1}{d+4}} \cdot N^{\frac{-1}{d+4}} \cdot s, \qquad (5)$$

with s the standard deviation of the average of the variances of each of the data points.

Ferdosi et al. [3] were interested in the application of density estimation on datasets that are large, i.e. datasets with more than 50 000 points with dimension ranging from ten to hundreds. Consequently they used the method proposed by Wilkinson and Meijer, but with a more simple estimation of the bandwidth for the pilot estimate kernel, namely

$$\sigma_l = \frac{P_{80}(l) - P_{20}(l)}{\log N}, l = 1, \ldots, d, \qquad (6)$$

where $P_{20}(l)$ and $P_{80}(l)$ are the twentieth and eightieth percentile of the data in dimension l.

Although the widths of the kernels used in both the Breiman estimator and the modified Breiman estimator are sensitive to the data, the shapes of the kernels are dependent of the kernel itself not the data. To further increase the response of the estimator to the data we propose shape-adaptive kernels, kernels of which both the width and the shape are steered by the data.

A disadvantage of these shape-adaptive kernels is that in regions where the density of sample points is low there are not enough data points to compute the shape of the kernel reliably. Consequently we propose to let the amount in which the shape of the kernel is influenced by the local data depend on the local density of the data points.

This paper is organized as follows. Section 2 discusses the proposed shape-adaptive kernels.

Aanvullen

2 Method

References

[1] L. Breiman, W. Meisel, and E. Purcell. “Variable Kernel Estimates of Multivariate Densities”. In: *Technometrics* 19.2 (1977), pp. 135–144.

[2] V.A. Epanechnikov. “Non-Parametric Estimation of a Multivariate Probability Density”. In: *Theory of Probability & Its Applications* 14.1 (1969), pp. 153–158.

[3] B.J. Ferdosi et al. “Comparison of Density Estimation Methods for Astronomical Datasets”. In: *Astronomy & Astrophysics* 531 (2011).

[4] Kirsty J Lees, Andrew J Guerin, and Elizabeth A Masden. “Using kernel density estimation to explore habitat use by seabirds at a marine renewable wave energy test facility”. In: *Marine Policy* 63 (2016), pp. 35–44.

[5] Johanna Skarpman Munter and Jens SjÃűlund. “Dose-volume histogram prediction using density estimation”. In: *Physics in Medicine and Biology* 60.17 (2015), p. 6923. URL: `http://stacks.iop.org/0031-9155/60/i=17/a=6923`.

[6] E. Parzen. “On Estimation of a Probability Density Function and Mode”. In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076.

[7] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probablity. Springer-Science+Business Media, B.V., 1986.

[8] M.H.F. Wilkinson and B.C. Meijer. “DATAPLOT: A Graphical Display Package for Bacterial Morphometry and Fluorimetry Data”. In: *Computer Methods and Programs in Biomedicine* 47.1 (1995), pp. 35–49.