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]. And astronomers have described it as “a critical first step in in making progress in many areas of astronomy." [3] For example they are interested in 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:

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

where d denotes the dimensionality of the data points. 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. It defines the density estimate as

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

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. We shall refer to this estimator as the Breiman Estimator.

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

$$\lambda_i = \left(\frac{\tilde{f}(\mathbf{x}_i)}{\mathrm{GM}\left(\tilde{f}(\mathbf{x}_0), \dots, \tilde{f}(\mathbf{x}_N)\right)}\right)^{-\beta} \qquad (3)$$

where $\mathrm{GM}(\cdot)$ denotes the geometric mean of pilot densities and the sensitivity parameter β must lie in the range $[0, 1]$.

(iii). Compute the adaptive kernel estimate as

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

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]. A disadvantage of this kernel is that it is not continuously differentiable. Secondly they computed the the pilot densities on a grid that covered 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

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

with s the standard deviation of the average of the variances of each of the data series. This estimator will be refered to as the Modifeid Breiman Estimator (MBE).

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 the dimension of the data points ranging from ten to hundreds of elements. They used the MBE with a simpler estimation of the bandwidth for the pilot estimate kernel, namely

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

where $P_{20}\left(l\right)$ and $P_{80}\left(l\right)$ are the twentieth and eightieth percentile of the data in dimension l. The optimal pilot window width, σ, is chosen as the smallest of $\sigma_1, \ldots, \sigma_d$.

Although the widths of the kernels used in estimators proposed by Breiman, Meisel, and Purcell, Wilkinson and Meijer 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 be dependent 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

We propose to compute the pilot density estimate with the method introduced by Wilkinson and Meijer. We use the Epanechnikov kernel for its low computational complexity, as its main disadvantage does not matter in the estimation of pilot densities [7]. The final density is estimated according to:

$$\hat{f}\left(\mathbf{x}\right) = \frac{1}{N} \sum_{i=1}^{N} \left(\sigma \lambda_i\right)^{-d} K_{\mathscr{G}}\left(\mathbf{x}\right), \qquad (7)$$

where $K_{\mathscr{G}}\left(i\right)$ s a Gaussian kernel with mean $\mathbf{x}_j$ and covariance Σ, which is the covariance matrix of the neighborhood of $\mathbf{x}$. The neighbors of $\mathbf{x}$ are determined with the k nearest neighbors algorithm (k-NN) with Euclidean distance. We use this approach rater than a fixed-radius neighborhood to ensure that independent of the sparsity of the data the kernel shape is always based on a reasonable number of data points. Furthermore using k-NN allows us to choose $k > d$, which makes it extremely improbable that the covariance matrix of the neighborhood is singular. We follow Silverman's [7] recommendation of choosing $k = \sqrt{N}$. To ensure that even in high-dimensional data sets $k > d$ we use $k = \max\left(\sqrt{N}, d+1\right)$.

The basic shape of the kernel used for $\mathbf{x}$ is given by the covariance matrix off $C_{\mathbf{x}_i}$, i.e. the union of $\mathbf{x}$ and its neighborhood. To allow the density estimation of each pattern to be influenced by an equal area, before the application of the smoothing factor (λ_i), the basic shapes of the kernels need to be scaled. The eigenvectors and eigenvalues of the covariance matrix define an ellipse, the eigenellipse. We scale the covariance matrix with the factor S, defined as:

$$S = \frac{\sigma^2}{\mathrm{GM}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\right)}, \qquad (8)$$

where λ_j denotes the jth eigenvalue of the jth eigenvector of Σ. The scaling factor S ensures that the shape-adapted covariance matrix has the same scale as the covariance matrix that is implicitly used in the Modified Breiman Estimator with a Gaussian kernel.

Two different values of β are prevalent in the literature. Breiman, Meisel, and Purcell [1] argues in favor of $^1/_d$, whereas Silverman [7] prefers using $\beta = {}^1/_2$, independent of the dimensionality of the data. We have empirically determined that $\beta =?$ is optimal with respect to the ?.

Empirisch vaststellen

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 Sjőlund. “Dose-volume histogram prediction using density estimation”. In: *Physics in Medicine and Biology* 60.17 (2015), p. 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. “DATA-PLOT: A Graphical Display Package for Bacterial Morphometry and Fluorimetry Data”. In: *Computer Methods and Programs in Biomedicine* 47.1 (1995), pp. 35–49.