Shape-Adaptive Kernel Density Estimation

L.E.N. Baakman

February 15, 2017

1 Introduction

The aim of density estimation is 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 $f(\mathbf{x})$.

Estimating densities with kernels has been fairly popular of late. In the medical field this approach has been used to predict dose-volume histograms, which are used to determine radiation doses [5]. Ecologists have explored the habitats of seabirds with density estimation [4]. Ferdosi et al. [3] have described it as "a critical first step in making progress in many areas of astronomy." Within this discipline density estimation is, among other things, used to estimate the density of the cosmic density field, which is required for the reconstruction of the large-scale structure of the universe.

The Parzen approach [6] estimates the density by summing bumps placed at the different observations, formally:

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

where d denotes the dimensionality of the data points. The shape of the placed bumps is determined by the kernel function $K(\cdot)$. Generally these functions are symmetric probability density functions. The Parzen approach requires the kernel to be a probability density function, i.e. $K(\mathbf{x}) \geq 0$ and $\int K(\mathbf{x}) = 1$. The width of the kernels is controlled by the bandwidth σ [7]. Choosing the window with too small, results in a density estimate with spurious fine structures, whereas kernels that are too wide can smooth the density estimate too much. Kernel estimates, such as the Parzen approach, that use kernels of only one width, are called fixed-width estimators.

One downside of fixed-width methods is that they 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 with high density, the sample points are more densely packed together, and the Parzen estimate tends to spread out in the high density region [1]. Adaptive-width methods address this disadvantage of the fixed-width methods. Breiman, Meisel, and Purcell introduced such a method. The sharpness of the kernels used by this method are responsive to the local data, it defines the density estimate as

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

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.

Silverman [7] showed that 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:

(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

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

Where $K(\cdot)$ is symmetric and integrates to unity.

The pilot densities are generally considered to be insensitive to the fine details of the pilot estimate. Therefor a convenient method can be used to estimate the pilot densities. One possible choice for the estimation of the pilot densities would be the Parzen approach. The local bandwidths are depend on the exponent β, if this value is high the λ will be more sensitive to variations in the pilot densities. For $\beta = 0$ we get $\lambda_1 = \ldots = \lambda_d = 1$, consequently Equation (4) reduces to a fixed width kernel density estimation, i.e. the Parzen approach. In the literature two values of β are prevalent. Breiman, Meisel, and Purcell [1] argue that choosing $\beta = {}^1/_d$ will ensure that the number of observations covered by the kernel will be approximately the same in all parts of the data. Whereas Silverman favors $\beta = {}^1/_2$ independent of the dimension of the data, as this value results in a bias that can be shown to be of a smaller order than that of the fixed-width kernel estimate.

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,

$$K_{\mathscr{E}}(\mathbf{x}) = \begin{cases} \frac{d+2}{2c_d}(1-\mathbf{x}\cdot\mathbf{x}) & \text{if } \mathbf{x}\cdot\mathbf{x} < 1 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where c_d denotes the volume of the unit sphere in d dimensions. There are two advantages to using this kernel, firstly it has a finite base and secondly it is optimal in the sense of the Mean Integrated Square Error (MISE) [2]. A disadvantage of this kernel is that it is not continuously differentiable, however as differentiability is not a requirement for the pilot densities, this is not a problem. The second change Wilkinson and Meijer [8] made to the method proposed by Breiman, Meisel, and Purcell [1] is that they computed the densities for points on a grid, that covered the data points, and determined the pilot densities with multi-linear interpolation. Wilkinson and Meijer used Equation (4) with an Epanechnikov kernel and $\lambda_i = 1$ to estimate the pilot densities. They determined the general bandwidth according to

$$\sigma = \left(\frac{8(d+4)\cdot\left(2\sqrt{\pi}\right)^d}{c_d} \right)^{\frac{1}{d+4}} \cdot N^{\left(\frac{-1}{d+4}\right)} \cdot s, \quad (6)$$

where s the square root of the average of the variances of the different dimensions. They estimate the final densities with Equation (4) using the general and local bandwidths estimated in Equations (3) and (6), respectively. The described estimator will be referred to as the Modified Breiman Estimator (MBE).

Ferdosi et al. [3] considered 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

$$\sigma_l = \frac{P_{80}(l) - P_{20}(l)}{\log N}, l = 1, \ldots, d, \quad (7)$$

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

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 on the kernel itself not the data. To further increase the responsiveness of the estimator to the data we propose the use of shape-adaptive kernels in density estimation. Not only the width but also the shape of these kernels would be steered by the data.

A disadvantage of these shape-adaptive kernels could be that in regions where the density of sample points is low, there are insufficient data points to compute the shape of the kernel reliably. Consequently we let the amount of influence exerted by the local data on the shape of the kernel be dependent on the density of the data points in that region.

This paper is organized as follows. Section 2 discusses the proposed shape-adaptive kernels. The datasets used to investigate the performance of these kernels are presented in Section 3. Section 4 presents the results of using both the 'normal' and the 'shape-adaptive' Modified Breiman estimator to estimate the densities of these datasets. The results are discussed in Section 5, this paper is concluded in Section 6.

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, its finite differentiability, does not matter in the estimation of pilot densities [7]. The final density is estimated according to:

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} (\sigma\lambda_i)^{-d} K_{\mathscr{G}}(\mathbf{x}), \tag{8}$$

where $K_{\mathscr{G}}(\cdot)$ is 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 of $C_{\mathbf{x}}$, 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:

$$S = \frac{\sigma^2}{\mathrm{GM}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\right)}, \tag{9}$$

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] argue 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

Discuss performance metric?

3 Datasets

Introduce the section

3.1 Artificial Datasets

Getjat van Ferdosi

Verwijs naar tabel met de definities van de datasets en de plaatjes van de sets

Dataset 1: Verwijs naar plaatje

Dataset 1: Beschrijf dataset

Dataset 1: Waarom is deze interessant

Dataset 1: Wat verwachten we

Dataset 2: Verwijs naar plaatje

Dataset 2: Beschrijf dataset

Dataset 2: Waarom is deze interessant

Dataset 2: Wat verwachten we

Dataset 3: Verwijs naar plaatje

Dataset 3: Beschrijf dataset

Dataset 3: Waarom is deze interessant

Dataset 3: Wat verwachten we

Dataset 4: Verwijs naar plaatje

Dataset 4: Beschrijf dataset

Dataset 4: Waarom is deze interessant

Dataset 4: Wat verwachten we

Dataset 5: Verwijs naar plaatje

Dataset 5: Beschrijf dataset

Dataset 5: Waarom is deze interessant

Dataset 5: Wat verwachten we

3.2 Real World Datasets

4 Results

Introduction into section

4.1 Artificial Datasets

4.2 Real World Datasets

5 Discussion

Introduce section

5.1 Artificial Datasets

5.2 Real World Datasets

5.3 Computational Complexity

Discuss computational complexity of AMBE v.s. SAMBE

5.4 General Discussion

Compare and contrast results of the two datasets

6 Conclusion

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. "DATAPLOT: A Graphical Display Package for Bacterial Morphometry and Fluorimetry Data". In: *Computer Methods and Programs in Biomedicine* 47.1 (1995), pp. 35–49.