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

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

where $K_{\mathcal{G}}(\cdot)$ 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

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

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,

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

where c_d denotes the volume of the unit sphere in d dimensions. Note that the kernel defined in Equation (5) does not have unit variance. this can be corrected by multiplying σ with the square root of the variance of $K_{\mathcal{E}}(\cdot)$.

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 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 use our shape adaptive kernels in combination with the Modified Breiman Estimator introduced by Wilkinson and Meijer [8]. The grid used for the pilot densities is The grid that the pilot densities are computed on We choose to use the method proposed by Ferdosi et al. [3] for computing the general bandwidth because of its lower complexity, compared to the method used by Wilkinson and Meijer [8]. We have empirically determined that using β = works best in our case. The final densities are estimated according to Equation (4) with a reshaped and scaled Epanechnikov kernel. The Epanechnikov kernel reshaped with the matrix Σ is defined as:

$$K_{\mathcal{E}}(\mathbf{x}) = \begin{cases} \frac{d+2}{2c_d}(1-\mathbf{x}\Sigma\mathbf{x}) & \text{if } \mathbf{x}\cdot\mathbf{x} < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

As stated before the matrix Σ is determined based on the neighborhood of the pattern, $\mathbf{x}$, whose density we are estimating.

We determine the neighbors of $\mathbf{x}$ with the k nearest neighbors algorithm (k-NN) with Euclidean distance. This approach is used rather 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

$$k = \max\left(\sqrt{N}, d+1\right).$$

Let $C_{\mathbf{x}}$ denote the union of $\mathbf{x}$ and its neighborhood, the basic shape of the kernel used for $\mathbf{x}$ is then given by $\mathrm{cov}(C_{\mathbf{x}})$.

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. To that end we use the eigenellipse, the ellipse defined by the eigenvectors and eigenvalues of $\mathrm{cov}(C_{\mathbf{x}})$. We scale the covariance matrix with the factor S, defined as:

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

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.

3 Experiment

Describe the experiment: estimate the density of the different datasets with three different methods: MBE, SAMBE, MBE with prewhitened data.

Introduce the subsections

3.1 Datasets

We examine the performance of the estimators on on two groups of datasets: a number of simulated datasets with known desnity fields; a real world dataset with an unkown density field.

Simulated Datasets

The simulated datasets are based on the simulated datasets used by Ferdosi et al. [3]. Figure 1 presents scatter plots of the data sets. The definitions of the these simulated data sets are shown in Table 1.

Dataset 1, shown in Figure 1a, is the most simple set in this group. It is an unimodal Gaussian distribution with random noise added.

The second dataset, depicted in Figure 1b, contains two Gaussian distributions with different covariance matrices and uniform noise. The means and covariance matrices of the Normal distributions are such that they do not overlap.

Dataset 3, represented in Figure 1c, consists of four different normal distributions and uniformly distributed noise. The four Gaussian distributions are placed in such a way that it is unlikely that any overlap occurs.

Figure 1d illustrates dataset 4. This set consists of a horizontal wall-like structure and a vertical filament-like structure.

The fifth dataset, shown in Figure 1e, contains three intersecting walls, each of which has two dimensions that are drawn from a uniform distribution and one that is drawn from a Gaussian distribution.

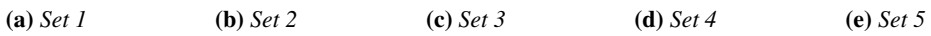
Although dataset one through three differ in complexity, we do not expect any type of estimator to perform better than the other on these datasets, as the spread in all dimensions is approximately equal

Iets over hoe het grid bepaald wordt, of de standaard grootte van het grid.

hoe hebben we dat vastgesteld

Een of andere waarde

The scaling discussed here has to change if the Epanechnikov kernel is used for the final density estimate.



(a) *Set 1* **(b)** *Set 2* **(c)** *Set 3* **(d)** *Set 4* **(e)** *Set 5*

Figure 1: *Scatter plot representation of the simulated datasets defined in Table 1.*

Set	Component	Fraction	Distribution
1	Trivariate Gaussian 1	$^2/_3$	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \mathrm{diag}(30)\big)$
	Uniform random background	$^1/_3$	$(x,y,z) \sim \mathcal{U}\big([0,0,0],[100,100,100]\big)$
2	Trivariate Gaussian 1	$^1/_3$	$(x,y,z) \sim \mathcal{N}\big([25,25,25], \mathrm{diag}(5)\big)$
	Trivariate Gaussian 2	$^1/_3$	$(x,y,z) \sim \mathcal{N}\big([65,65,65], \mathrm{diag}(20)\big)$
	Uniform random background	$^1/_3$	$(x,y,z) \sim \mathcal{U}\big([0,0,0],[100,100,100]\big)$
3	Trivariate Gaussian 1	$^1/_6$	$(x,y,z) \sim \mathcal{N}\big([24,10,10], \mathrm{diag}(2)\big)$
	Trivariate Gaussian 2	$^1/_6$	$(x,y,z) \sim \mathcal{N}\big([33,70,40], \mathrm{diag}(10)\big)$
	Trivariate Gaussian 3	$^1/_6$	$(x,y,z) \sim \mathcal{N}\big([90,20,80], \mathrm{diag}(1)\big)$
	Trivariate Gaussian 4	$^1/_6$	$(x,y,z) \sim \mathcal{N}\big([60,80,23], \mathrm{diag}(5)\big)$
	Uniform random background	$^1/_3$	$(x,y,z) \sim \mathcal{U}\big([0,0,0],[100,100,100]\big)$
4	Wall-like structure 1	$^1/_2$	$(x,z) \sim \mathcal{U}\big([0,0],[100,100]\big)$, $(y) \sim \mathcal{N}\big(5,5\big)$
	Filament-like structure	$^1/_2$	$(x,y) \sim \mathcal{N}\big([50,50], \mathrm{diag}(5)\big)$, $(z) \sim \mathcal{U}\big(0,100\big)$
5	Wall-like structure 1	$^1/_3$	$(x,z) \sim \mathcal{U}\big([0,0],[100,100]\big)$, $(y) \sim \mathcal{N}\big(10,5\big)$
	Wall-like structure 2	$^1/_3$	$(x,y) \sim \mathcal{U}\big([0,0],[100,100]\big)$, $(z) \sim \mathcal{N}\big(50,5\big)$
	Wall-like structure 3	$^1/_3$	$(x,z) \sim \mathcal{U}\big([0,0],[100,100]\big)$, $(y) \sim \mathcal{N}\big(50,5\big)$

Table 1: *The simulated datasets used to test the estimators. The column 'Fraction' indicates for each component of the dataset which fraction of the total number of points of the data set is part of that component.* $\mathcal{N}\big(\mu,\Sigma\big)$ *denotes a Gaussian distribution with mean* μ *and covariance matrix* Σ*. A diagonal matrix with the value x on the diagonal is represented as* $\mathrm{diag}(x)$*.* $\mathcal{U}\big(a,b\big)$ *denotes a uniform distribution with its minimum and maximum set to a and b, respectively.*

for all dimensions. Dataset four and five clearly are spread more in one dimension than in others, thus we expect that the shape adaptive estimator and the estimator on the whitened data will perform better on these sets than the MBE estimator.

The increasing complexity of these dataset allow us to investigate the performance of the classifier on simple situations, one cluster of data with some noise, to complex density fields that should better approximate real world data. The advantage of using simulated data is that the true densities of the data points are known, which allows us to clearly test how well the different methods estimate the densities.

Real World Datasets

3.2 Error Measures

4 Results

5 Discussion

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.