Shape-Adaptive Kernel Density Estimation

L.E.N. Baakman

September 4, 2017

Abstract

Kernel density estimation has gained popularity in the past few years. Generally the methods use symmetric kernels, even though the data of which the density is estimated are not necessarily spread equally in all dimensions. To account for this asymmetric distribution of data we propose the use of shape adaptive kernels: kernels whose shape changes to fit the spread of the data in the local neighborhood of the point whose density is estimated. We compare the performance of the shape adaptive kernels on simulated datasets with known density fields.

Results

Conclusion

1 Introduction

2 Method

3 Experiment

We compare the performance of the shape-adaptive method with that of the Modified Breiman Estimator, on simulated datasets with known density fields. This allows us to test how well the proposed method can recover simple density distributions.

To quantify the performance of the estimators we use the Mean Squared Error (MSE):

$$\mathrm{MSE}\left(\hat{f}(\bullet)\right) = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{f}(\mathbf{x}_j) - f(\mathbf{x}_j)\right)^2.$$

The simulated datasets are a superset of a selection of the sets used by Ferdosi et al. [1]. Figure 1 shows scatter plots of these sets, their definition is given in Table 1.

Dataset one, two, and three, shown in Figures 1a, 1e and 1g, respectively, are taken directly from Ferdosi et al. They consist of a number of spherical Gaussian distributions with random noise added. The means of the Gaussian distribution are chosen in such a way that it is unlikely that the distributions overlap.

Figures 1b, 1f and 1h present dataset four, five, and six. These datasets are created from dataset one, two and, three, respectively. This is done in such a way that the volumes of the eigenspheres of the covariance matrices of the components in the derived datasets have the same volume as the eigenellipsoids of the covariance matrices of the associated components in the original dataset. Furthermore if a is the eigenvalue of the original covariance matrix, the eigenvalues of the covariance matricex of the derived component are a^2, $\sqrt{a}$ and $\sqrt{a}$. Consequently the volumes the eigenspheres of the covariance matrices of the Gaussians in dataset four, five and, six are equal to those of dataset one, two and, three, respectively.

In dataset seven and eight, illustrated in Figures 1c and 1h the semi axes of the ellipsoids all have different lengths. The largest minor axis of the Trivariate Gaussian in dataset seven is a factor two larger than the smallest minor axis in that dataset. Whereas in dataset eight the largest minor axis is exponentially larger than the smallest minor axis.

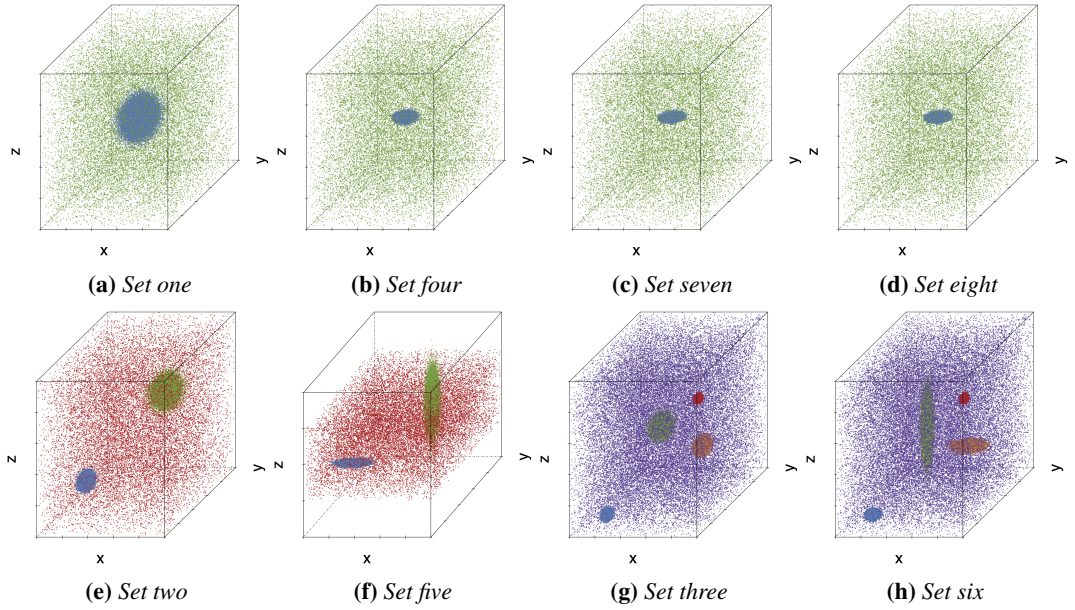
We expect the MBE and shape-adaptive MBE to perform comparable on dataset one through three, as other than the randomly sampled noise these sets only contain data sampled from a Gaussian distribution with a diagonal covariance matrix. Which results in an equal spread of the data in all dimensions for the non-noise data. Given the elongated shape of the non-noise components in dataset four, five, six, seven, and eight we hypothesize that the shape-adaptive estimator outperforms the estimator that is not shape adaptive.

The datasets with a single Gaussian which are increasingly more elongated, i.e. dataset one, four, seven and, eight, allow us to investigate the influence of the ellipitcalness on the performance of the

Set		Component	Number	Distribution
one	●	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \mathrm{diag}(30)\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
two	●	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([25,25,25], \mathrm{diag}(5)\big)$
	●	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([65,65,65], \mathrm{diag}(20)\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
three	●	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([24,10,10], \mathrm{diag}(2)\big)$
	●	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([33,70,40], \mathrm{diag}(10)\big)$
	●	Trivariate Gaussian 3	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([90,20,80], \mathrm{diag}(1)\big)$
	●	Trivariate Gaussian 4	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([60,80,23], \mathrm{diag}(5)\big)$
	●	Uniform random background	4.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
four	●	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \mathrm{diag}([9,\sqrt{3},\sqrt{3}])\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
five	●	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([25,25,25], \mathrm{diag}([25,\sqrt{5},\sqrt{5}])\big)$
	●	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([65,65,65], \mathrm{diag}([\sqrt{20},\sqrt{20},400])\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [150,150,150]\big)$
six	●	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([24,10,10], \mathrm{diag}([4,\sqrt{2},\sqrt{2}])\big)$
	●	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([33,70,40], \mathrm{diag}([\sqrt{10},\sqrt{10},100])\big)$
	●	Trivariate Gaussian 3	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([90,20,80], \mathrm{diag}(1)\big)$
	●	Trivariate Gaussian 4	2.0×10^4	$(x,y,z) \sim \mathcal{N}\big([60,80,23], \mathrm{diag}([25,\sqrt{5},\sqrt{5}])\big)$
	●	Uniform random background	4.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
seven	●	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \mathrm{diag}([9,2*\sqrt{3},{}^1\!/_2\sqrt{3}])\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
eight	●	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \mathrm{diag}([9,3,1])\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$

Table 1: *The datasets used to test the estimators. The column 'Number' indicates for each component of the dataset how many data points are sampled from that component.* $\mathcal{N}(\mu,\Sigma)$ *denotes a Gaussian distribution with mean* μ *and covariance matrix* Σ*. A diagonal matrix with the values* $x_1, \cdots, x_d$ *on the diagonal is represented as* $\mathrm{diag}([x_1, \cdots, x_d])$*, a scalar matrix with* x *on the diagonal is shown as* $\mathrm{diag}(x)$*.* $\mathcal{U}(a,b)$ *denotes a uniform distribution with its minimum and maximum set to* a *and* b*, respectively. The colors shown in the second column correspond with the colors used for these components of the data set throughout the paper.*

Before Final Version: Remove ticks and labels.



(a) *Set one* **(b)** *Set four* **(c)** *Set seven* **(d)** *Set eight*

(e) *Set two* **(f)** *Set five* **(g)** *Set three* **(h)** *Set six*

Figure 1: *Scatter plot representation of the datasets defined in Table 1. The colors of the different components correspond to the colors used in Table 1.*

	Estimator	
	MBE	saMBE
one	4.118×10^{-10}	2.983×10^{-9}
two	5.279×10^{-8}	1.001×10^{-7}
three	4.375×10^{-6}	5.484×10^{-6}
four	4.779×10^{-7}	1.231×10^{-4}
five	5.383×10^{-8}	9.425×10^{-8}
six	4.189×10^{-6}	5.454×10^{-6}
seven	7.323×10^{-7}	4.110×10^{-4}
eight	6.569×10^{-7}	3.306×10^{-4}

Table 2: *The mean squared error of the known densities and the densities estimated by the Modified Breiman Estimator (MBE) and the shape-adaptive MBE (saMBE), respectively, for the datasets in Table 1.*

estimators. Whereas the datasets with multiple ellipsoids, i.e. dataset two, three, five, and six, make it possible to investigate the performance of the classifier on more complex density fields that better approximate real world data.

4 Results

This section presents the results of the experiments described in Section 3. We compare the performance of the two estimators on each dataset with the mean square error, presented in Table 2, and visually with plots. All plots associated with a single dataset have the same domain and range, to allow for easy comparison of the results within a dataset. The horizontal axis is used to represent the known densities, its range is such that each known density can be shown. The estimated densities are shown on the vertical axis, the length of these axes is such that they are long enough to represent every estimated density for that dataset, independent of the used estimator. The black line in each plot illustrates the line all points would lie on if a perfect estimator was used, i.e. the line $x = x$. The colors of the points in these plot correspond to the colors of the elements of the datasets in Table 1 and Figure 1.

Section 4.1 presents the results of the datasets that contain a single Gaussian, in Section 4.2 the results of the datasets that consist of noise and multiple Gaussian distributions are presented.

4.1 Datasets with a Single Gaussian

This section compares the performance of the Modified Breiman Estimator and a shape-adaptive vari-

ant on dataset that contain one Gaussian, i.e. dataset one, four, seven, and eight.

Only on the dataset with a spherical Gaussian is the performance of the two estimators comparable, on all dataset with a single elliptical Gaussian the non-shape adaptive estimator performs significantly better than its shape-adaptive cousin.

Figure 2 presents the results of using the Modified Breiman Estimator and its shape adaptive variant to estimate the densities of the datasets in Table 1 that contain a single Gaussian distribution.

Figure 2a confirms our findings from **??**, namely that the Modified Breiman Estimator gives a good approximation of the densities of dataset one. The densities estimated with the MBE both over, and undershoot the true density. Figure 2a, on the other hand, shows that shape adaptive MBE nearly always overshoots the true density.

Comparing the performance of the two estimators on dataset four with Figures 2b and 2f we find that the Modified Breiman Estimator outperforms the shape-adaptive variant. The second estimator has some extreme outliers, the most extreme of which are 1.072, and -0.5068.

The results of data set seven, shown in Figures 2b and 2f respectively, are comparable to those of four: the original estimator approximates the density pretty well, the shape-adaptive variant has some extreme outliers, the densities estimated by saMBE fall in the range, $[-4.661, 0.9283]$, whereas the true densities all lie within $\left[5.000 \times 10^{-7}, 6.108 \times 10^{-3}\right]$.

Figures 2d and 2h compare the performance of respectively MBE with saMBE on data set eight. We once again observe that the non-shape adaptive estimator approximates the known densities pretty well. Whereas the shape-adaptive estimator returns extreme results with densities that are estimated to be as high as 3.827 and as low as -1.134.

In general we have found that the Modified Breiman estimator works pretty well for data sets that contain a single Gaussian, especially if the Gaussian is spherical. Since the mean square error for dataset eight is lower than the MSE of dataset seven the ellipticalness of the distribution does not seem to influence the performance of this estimator. The shape adaptive MBE results in some extremely high and low estimated densities if used to estimate the densities of non-spherical Gaussian. saMBE overestimated some of the densities of the spherical Gaussian compared to the Modified Breiman Estimator. The range of the values estimated by the shape-adaptive estimator does not seem to be influenced by how electricalness of the Gaussian distribution.

4.2 Datasets with Multiple Gaussians

In this section we present the results of the two estimators on dataset two, five, three, six, i.e. the datasets that contain more than one Gaussian.

The plots in Figure 3 suggest some differences in how the two estimators handle the different components of the dataset, therefore Table 3 presents the mean square error of the different components of the datasets that contain multiple Gaussians.

Table 3 shows some difference in MSE between the different components of dataset two, and the elongated version of this dataset, i.e. dataset five: both estimators perform best on data points that were drawn from the uniform distribution, and worst on Gaussian 1, the Gaussian component with the smallest radius. Furthermore saMBE outperforms MBE on the component named 'Gaussian 2'.

Comparing Figures 3a, 3b, 3e and 3f we find that on both datasets both estimators underestimate the density. Although the ranges of the densities estimated by saMBE are slightly higher than those estimated by the Modified Breiman Estimator the ranges do not differ as extremely as they did when a single non-spherical Gaussian with noise made up the dataset.

Report on Figure 3c and Figure 3d.

General observation of multi sphere datasets.

General observation of the results.

5 Discussion

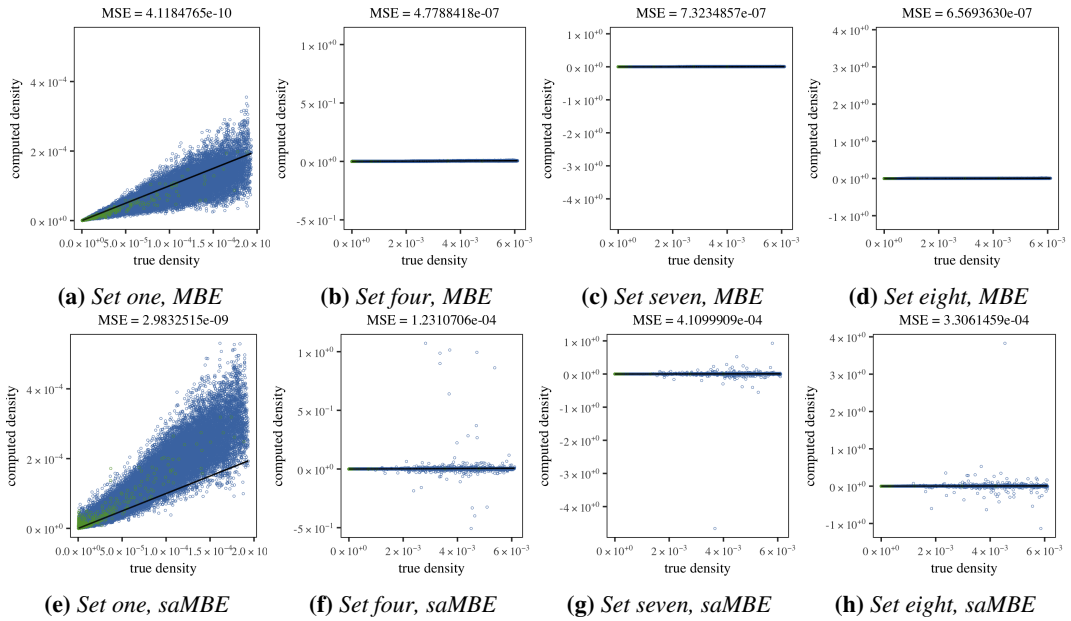
This section discusses the results presented in Section 4 using the structure used in **??**.

5.1 Datasets with a Single Gaussian

The difference in results between the Modified Breiman Estimator and its shape adaptive variant on dataset one, four, seven, eight raise several questions. This section attempts to answer them.

In Figure 2 we observed that saMBE overestimates the densities of dataset one. This could indicate that the kernels are too small, resulting in a too high contribution to the density estimate. Since the Modified Breiman estimator uses the same general and local bandwidth as the shape adaptive version the likely culprit is the shape of the kernel.

Another strange result observed in Figure 2 is that a large number of estimated densities were not in



(a) *Set one, MBE* **(b)** *Set four, MBE* **(c)** *Set seven, MBE* **(d)** *Set eight, MBE*

(e) *Set one, saMBE* **(f)** *Set four, saMBE* **(g)** *Set seven, saMBE* **(h)** *Set eight, saMBE*

Figure 2: *Comparative plots for dataset 1, 4, 7, and 8.*

the expected range $[0,1]$ when saMBE was used. Strangely this effect does not occur when the shape adaptive matrix is not used in a dataset that contains spherical data, i.e. in dataset one. Looking back to **??** we find that since $\forall \mathbf{x}\ K(\mathbf{x}) \in [0,1]$, $\det(\mathbf{H}_i)$ must be smaller than zero to cause a negative density estimate. For the same reason the density estimates that are greater than zero must be the result of $\det(\mathbf{H}_i) < 1$ for some $\mathbf{H}_i$.

The issues above probably explain why the shape adaptive Modified Breiman Estimator does not outperform the non-shape adaptive variant on these data sets.

5.2 Datasets with Multiple Gaussians

What does this section do?

.

Discuss Figure 3a and Figure 3b.

Performs best on Gaussian with largest radius, reason: Most like uniform noise on which both estimators perform best?

Both have difficulty with Gaussian with smallest radius

Why are the densities underestimated?

Discuss Figure 3c and Figure 3d.

Waarom hier weer negatieve resultaten, waarom niet in baakman2?

General discussion of multi sphere datasets.

General Discussion

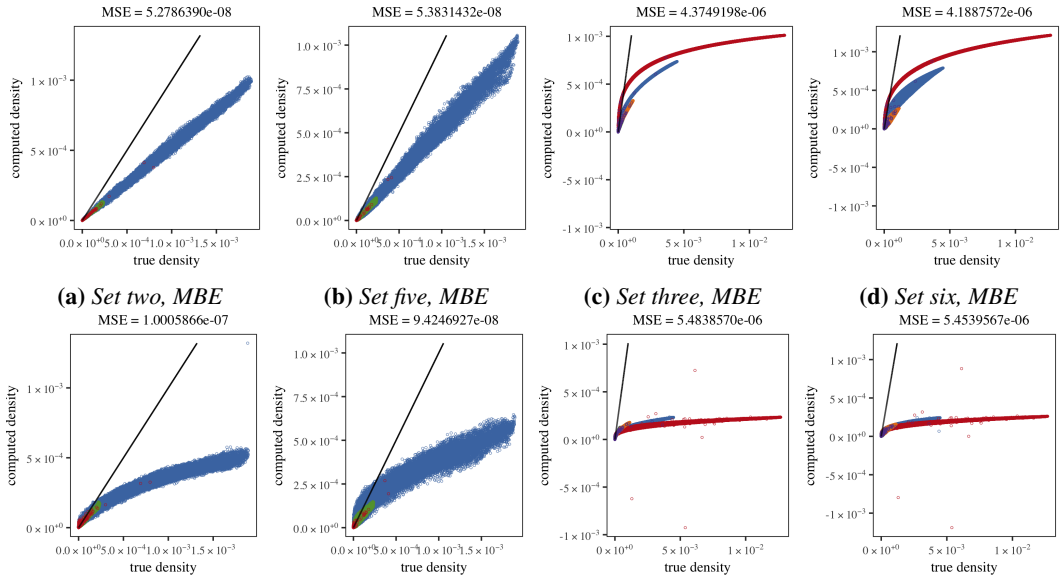
6 Conclusion

References

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

Set		Component	Estimator	
			MBE	saMBE
two	●	Gaussian 1	1.561×10^{-7}	2.993×10^{-7}
	●	Gaussian 2	2.285×10^{-9}	8.345×10^{-10}
	●	Noise	2.021×10^{-11}	2.363×10^{-11}
five	●	Gaussian 1	1.587×10^{-7}	2.810×10^{-7}
	●	Gaussian 2	2.771×10^{-9}	1.769×10^{-9}
	●	Noise	4.829×10^{-12}	4.865×10^{-12}
three	●	Gaussian 1	0.000	0.000
	●	Gaussian 2	0.000	0.000
	●	Gaussian 3	0.000	0.000
	●	Gaussian 4	0.000	0.000
six	●	Gaussian 1	0.000	0.000
	●	Gaussian 2	0.000	0.000
	●	Gaussian 3	0.000	0.000
	●	Gaussian 4	0.000	0.000

Table 3: *The mean squared error of the known densities and the densities estimated by the Modified Breiman Estimator (MBE) and the shape-adaptive MBE (saMBE), respectively, for the different components of the datasets with multiple Gaussians.*



(a) *Set two, MBE* **(b)** *Set five, MBE* **(c)** *Set three, MBE* **(d)** *Set six, MBE*

(e) *Set two, saMBE* **(f)** *Set five, saMBE* **(g)** *Set three, saMBE* **(h)** *Set six, saMBE*

Figure 3: *Comparative plots for dataset 2, 3, 5, and 6.*