Shape-Adaptive Kernel Density Estimation

L.E.N. Baakman

September 6, 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. We distinguish two types of datasets: datasets consisting of a single Gaussian distribution and noise, discussed in Section 3.1 and datasets containing multiple Gaussian distributions next to noise, these are presented in Section 3.2.

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

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

3.1 Datasets with a Single Gaussian

Figure 1 shows a scatter plot representation of the datasets containing a single Gaussian distribution defined in Table 1.

The Gaussian components of these datasets progress from a sphere, i.e. dataset one, to an increasingly more elongated ellipsoid. This makes it possible to investigate the influence of how strongly elongated the distribution is on the density estimate.

Discuss set 1.

Dataset four is created from one by squaring one of the eigenvalues of the covariance matrix, and taking the square root of the other two eigenvalues, without changing the eigenvectors. The resulting covariance matrix defines an eigenellipse with the same volume as the one defined by one.

Discuss set 7.

Discuss set 8.

We expect the Modified Breiman Estimator and its shape-adaptive cousin to perform comparably on dataset one, since due to the symmetric shape of the Gaussian distribution no advantage should be gained by using a shape-adaptive kernel. As the Gaussian distribution is more and more elongated, the advantage of using saMBE should become more pronounced.

3.2 Datasets with Multiple Gaussians

Figure

Table

Dataset 2

Dataset 5

The process of creation is the same as for 4from 1.

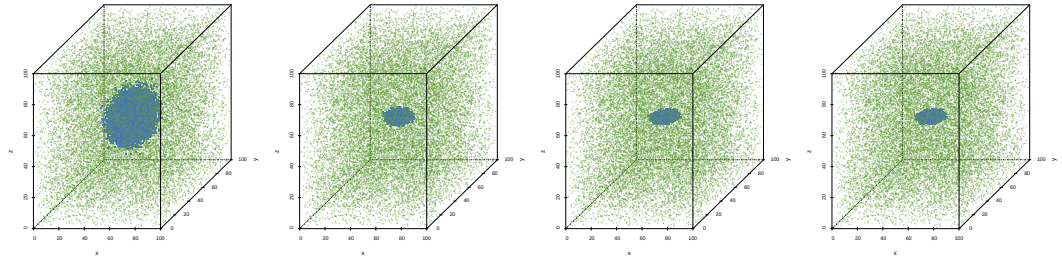
Dataset 3

Dataset 6

The process of creation is the same as for 4from 1.

Before Final Version: Remove ticks and labels.

Before Final Version: Fix the length of the axis of dataset 5.



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

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.*

Set		Component	Number	Distribution
one	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \text{diag}(30)\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
four	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \text{diag}([9,\sqrt{3},\sqrt{3}])\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$
seven	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \text{diag}([9, 2*\sqrt{3}, {}^1\!/_2\sqrt{3}])\big)$
	●	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	$(x,y,z) \sim \mathcal{N}\big([50,50,50], \text{diag}([9,3,1])\big)$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}\big([0,0,0], [100,100,100]\big)$

Table 1: *The containing a single Gaussian distribution next to uniform noise. 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* $\text{diag}([x_1, \cdots, x_d])$*, a scalar matrix with* x *on the diagonal is shown as* $\text{diag}(x)$*.* $\mathcal{U}(a,b)$ *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.*

old

The simulated datasets are a superset of a selection of the sets used by Ferdosi et al. [1]. Figure 2 shows scatter plots of these sets, their definition is given in Table 2.

Dataset one, two, and three, shown in Figures 1a, 2a and 2c, 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, 2b and 2d 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 2d 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 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

5 Discussion

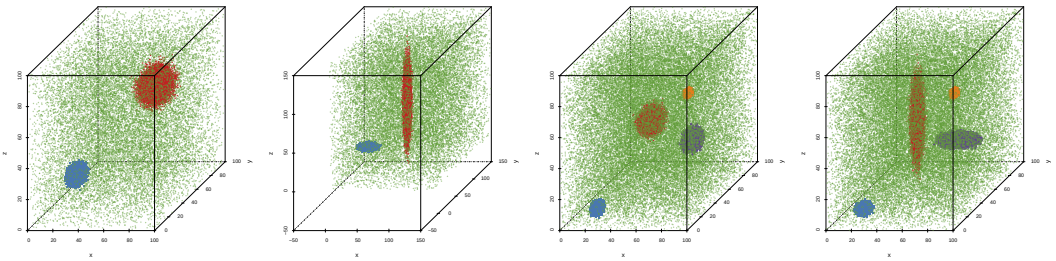
6 Conclusion

References

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

Before Final Version: Remove ticks and labels.

Before Final Version: Fix the length of the axis of dataset 5.



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

Figure 2: *Scatter plot representation of the datasets defined in Table 2. The colors of the different components correspond to the colors used in Table 2.*

Set		Component	Number	Distribution
one	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}([50,50,50], \text{diag}(30))$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$
two	■	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}([25,25,25], \text{diag}(5))$
	▲	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}([65,65,65], \text{diag}(20))$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$
three	■	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}([24,10,10], \text{diag}(2))$
	▲	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}([33,70,40], \text{diag}(10))$
	♦	Trivariate Gaussian 3	2.0×10^4	$(x,y,z) \sim \mathcal{N}([90,20,80], \text{diag}(1))$
	✳	Trivariate Gaussian 4	2.0×10^4	$(x,y,z) \sim \mathcal{N}([60,80,23], \text{diag}(5))$
	●	Uniform random background	4.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$
four	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}([50,50,50], \text{diag}([9,\sqrt{3},\sqrt{3}]))$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$
five	■	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}([25,25,25], \text{diag}([25,\sqrt{5},\sqrt{5}]))$
	▲	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}([65,65,65], \text{diag}([\sqrt{20},\sqrt{20},400]))$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}([-15,-15,-15],[150,150,150])$
six	■	Trivariate Gaussian 1	2.0×10^4	$(x,y,z) \sim \mathcal{N}([24,10,10], \text{diag}([4,\sqrt{2},\sqrt{2}]))$
	▲	Trivariate Gaussian 2	2.0×10^4	$(x,y,z) \sim \mathcal{N}([33,70,40], \text{diag}([\sqrt{10},\sqrt{10},100]))$
	♦	Trivariate Gaussian 3	2.0×10^4	$(x,y,z) \sim \mathcal{N}([90,20,80], \text{diag}(1))$
	✳	Trivariate Gaussian 4	2.0×10^4	$(x,y,z) \sim \mathcal{N}([60,80,23], \text{diag}([25,\sqrt{5},\sqrt{5}]))$
	●	Uniform random background	4.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$
seven	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}([50,50,50], \text{diag}([9,2*\sqrt{3},{}^1\!/_2\sqrt{3}]))$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$
eight	■	Trivariate Gaussian	4.0×10^4	$(x,y,z) \sim \mathcal{N}([50,50,50], \text{diag}([9,3,1]))$
	●	Uniform random background	2.0×10^4	$(x,y,z) \sim \mathcal{U}([0,0,0],[100,100,100])$

Table 2: *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* $\text{diag}([x_1,\cdots,x_d])$*, a scalar matrix with* x *on the diagonal is shown as* $\text{diag}(x)$*.* $\mathcal{U}(a,b)$ *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.*