Comparison of Density Estimation Methods for Astronomical Datasets

B.J. Ferdosi[†], H. Buddelmeijer[‡], A. Helmi[‡], S.C. Trager[‡], E.A. Valentijn[‡], M.H.F. Wilkinson[†], J.M. van der Hulst[‡], J.B.T.M. Roerdink[†]

[†]Institute for Mathematics and Computing Science, University of Groningen

[‡]Kapteyn Astronomical Institute, University of Groningen

{B.J.Ferdosi, M.H.F.Wilkinson, J.B.T.M.Roerdink}@rug.nl,

{H.Buddelmeijer, A.Helmi, S.C.Trager, E.A.Valentijn, J.M.van.der.Hulst}@astro.rug.nl

Abstract

We study the performance of four density estimation techniques. Density estimators are applied to six artificial datasets (ad 1-6) and on two astronomical datasets (mgs 1 and 2) derived from the Millennium galaxy sample (mgs) using a Monte Carlo process. We compared the performance of the methods in two ways: first, by measuring the mean squared error and Kullback–Leibler divergence of each of the methods; second, by the visualization of density fields. The results show that the adaptive kernel based methods perform better than the other methods in terms of calculating the density properly.

1. Introduction

Usage of densities in astronomical data analysis :

- Reconstruction of the field of simulation data [4]
- Analysing structures in phase space [5]
- Finding relations among galaxy color, morphology, environment etc. [1]

2. Density estimation methods

- k-nearest neighbors (kNN)
- adaptive Gaussian kernel density estimation (DEDICA) [6]
- a modified version of the adaptive kernel density estimation of Breiman [2] with Epanechnikov kernel, called the modified Breiman estimator (MBE)
- the Delaunay tessellation field estimator (DTFE) [3]

3. Error measures

- Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^2 \qquad (1)$$

where $\hat{p}_i$ is the density of the ith data point obtained from the density estimator and p_i is the true density of that point.

- Kullback-Leibler divergence (KLD)

For two probability distributions $f(x)$ and $g(x)$ of a random variable X, this is defined as:

$$KLD(f \parallel g) = \int_{-\infty}^{\infty} f(x)\log\left(\frac{f(x)}{g(x)}\right)\,dx \qquad (2)$$

4. Datasets

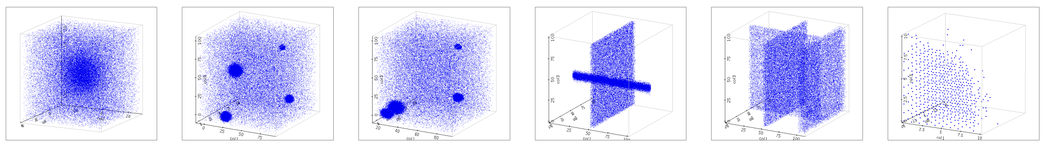


Figure 2: *Scatter plot of artificial datasets. Left to right: ad 1-6.*

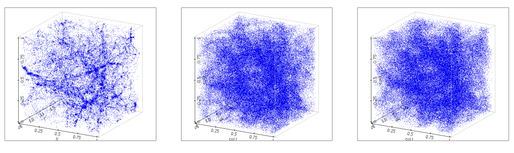
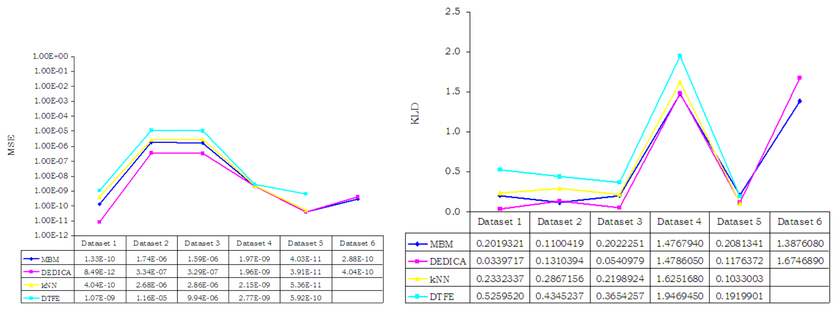


Figure 3: *Scatter plot of galaxy datasets. Left to right: mgs, mgs1 with DTFE generated field, mgs2 with MBE generated field.*

5. Results



	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
MBM	1.33E-10	1.74E-06	1.59E-06	1.97E-09	4.03E-11	2.88E-10
DEDICA	8.49E-12	3.34E-07	3.29E-07	1.96E-09	3.91E-11	4.04E-10
kNN	4.04E-10	2.68E-06	2.86E-06	2.15E-09	5.36E-11	
DTFE	1.07E-09	1.16E-05	9.94E-06	2.77E-09	5.92E-10	

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
MBM	0.2019321	0.1100419	0.2022251	1.4767940	0.2081341	1.3876080
DEDICA	0.0339717	0.1310394	0.0540979	1.4786050	0.1176372	1.6746890
kNN	0.2332337	0.2867156	0.2198924	1.6251680	0.1033003	
DTFE	0.5259520	0.4345237	0.3654257	1.9469450	0.1919901	

Figure 4: *Artificial datasets: MSE and KLD for point densities.*

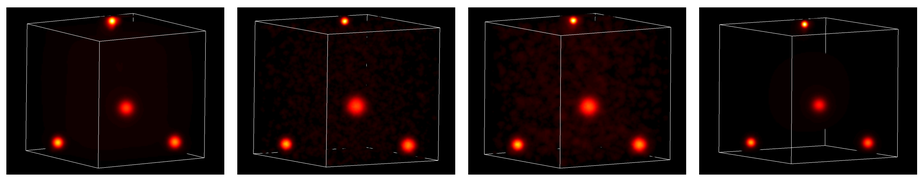


Figure 5: *Volume visualization (ad 4). Left to right: MBE, DTFE, kNN, DEDICA.*

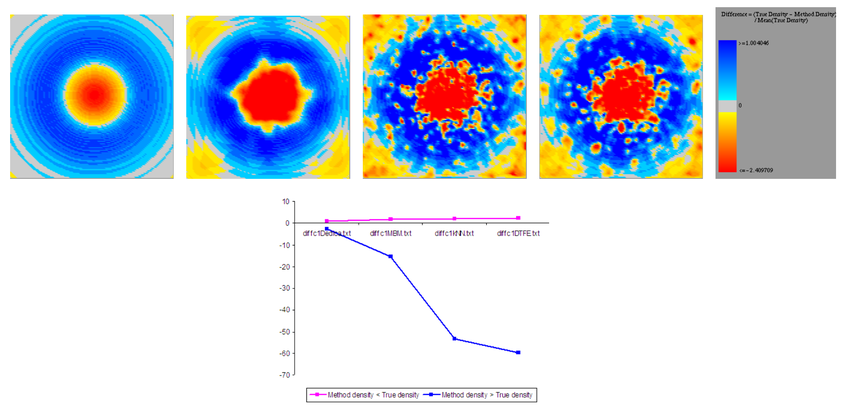


Figure 6: *Top: difference (true density - method density) image visualization of ad 1. From left to right: diffDEDICA, diffMBE, diffDTFE, diffkNN density. Bottom: overestimation vs underestimation.*

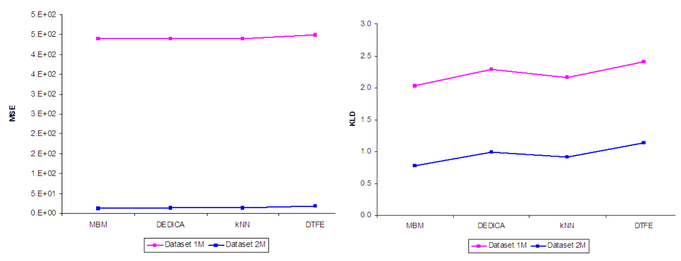


Figure 7: *Derived datasets from the Millennium Simulation, mgs1 and mgs2: MSE and KLD for point densities.*

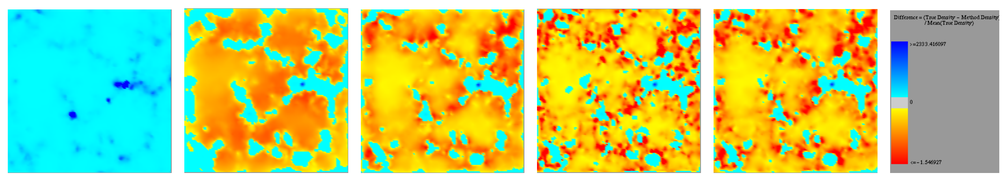


Figure 8: *Difference image visualization for mgs1. Left to right: True density field produced by DTFE, diffDEDICA, diffMBE, diffDTFE, diffkNN.*

6. Conclusion

Choice of methods can depend on the application at hand:

- DEDICA or MBE where proper estimation of densities is required
- DTFE for finding and analyzing structures.

References

[1] I. Baldry, M. L.Balogh, R. Bower, K. Glazebrook, R. C. Nichol, S. P. Bamford, and T. Budavari. *MNRAS*, 373:469–483, 2006.

[2] L. Breiman, W. Meisel, and E. Purcell. *Technometrics*, 19:135–144, 1977.

[3] R. v. d. W. F.I. Pelupessy, W.E. Schaap. *Astronomy and Astrophysics*, 403:389–398, 2003.

[4] R. Gingold and J. Monaghan. *MNRAS*, 181:375, 1977.

[5] M. Maciejewski, S. Colombi, C. Alard, F. Bouchet, and C. Pichon. *MNRAS*, 393:703–722, 2009.

[6] A. Pisani. *MNRAS*, 278:697, 1996.

Netherlands Organisation for Scientific Research

This work was supported by **NWO**, STARE grant, no. 643.200.501.