The Zeldovich Universe:
Genesis and Growth of the Cosmic Web
Proceedings IAU Symposium No. 308, 2014
R. van de Weygaert, S. Shandarin, E. Saar & J. Einasto, eds.

doi:10.1017/S1743921316009947

Adaptive density estimator for galaxy surveys

Enn Saar[1,2]

[1]Tartu Observatory, Tõravere, Tartumaa, Estonia
email: saar@aai.ee

[2]Estonian Academy of Sciences, Kohtu 4, Tallinn, Estonia

Abstract. Galaxy number or luminosity density serves as a basis for many structure classification algorithms. Several methods are used to estimate this density. Among them kernel methods have probably the best statistical properties and allow also to estimate the local sample errors of the estimate. We introduce a kernel density estimator with an adaptive data-driven anisotropic kernel, describe its properties and demonstrate the wealth of additional information it gives us about the local properties of the galaxy distribution.

Keywords. Cosmology: large-scale structure of universe, surveys, galaxies: statistics

1. Introduction

Galaxy position (redshift) surveys give us maps of the universe, formed by the mutual positions of many galaxies. Such maps can be studied as they are, but in many cases the first processed product of the survey would be the continuous (matter or luminosity) density map for the survey volume. Such a map gives us an understanding of the cosmography of the survey, forms a basis for finding and classifying the constituents of the large-scale structure, and is probably useful for many other applications.

Density estimation has become an art in itself in recent years, especially with application of Bayesian methods. One of the most impressive cosmological papers in recent years (Jasche & Wandelt (2012)) demonstrated how one can find the density distribution for a galaxy catalog with photometric redshifts. The errors of photometric redshifts are usually so large that they smear up all the large-scale structure. Jasche & Wandelt showed that using an isotropic covariance matrix as a natural prior one can recover the real structure. Think about it – the only additional requirement is a very natural and simple requirement of local statistical isotropy, and it does practically all the work.

Another long-reaching effort along similar lines is the work by Kitaura *et al.* (2012) who reconstruct the density distribution in the local universe by guessing the initial conditions compatible with the present galaxy distribution, taking them to the present by numerical simulations, and comparing them with real galaxy positions. All this work is one step in a MCMC chain, so the total work is enormous. But the result is certainly the best picture of the local universe for that moment.

Both groups are continuing their work and improving their methods, I recommend to check the literature, searching by the authors. But although this approach is solid and the results are impressive, it is very expensive in terms of computer time. The observational data is represented by numbers of galaxies in spatial cells that cover the survey volume, and every number is an independent variable. The typical number of the cells is about 10^6 to 10^7, and this is the dimension of the space where the MCMC has to work. This demands huge computing power, sophisticated methods, and it is difficult to imagine that it would be possible to sample the posterior distribution uniformly.

As future survey volumes are growing fast, we need to also use fast (simple) methods of density estimation, both as an alternative methods or better inputs for the Bayesian methods.

2. Density estimation

The simplest way to estimate the density is to use histograms, that for a 3-D world translate to disjoint volume elements, usually cubic cells, and to count galaxies in these cells. Although such approach is frequently used, it is not the best way to get the density. The most evident drawback is that the galaxy numbers in adjacent cells may crucially depend on the arbitrary location of cell boundaries – we may as well find a whole galaxy cluster in a cell, as to break it into two halves by a happy boundary. Statisticians have long known that there is a much better way, the kernel density estimation (see, e.g., citeSilverman86).

For a 1-D case, the density $\rho(x)$ for a discrete sample of n points with the positions x_i can be found in any point x as

$$\rho(x) = \frac{1}{nh} \sum_i^n K\left(\frac{x - x_i}{h}\right), \tag{2.1}$$

where x_i are the coordinates of the sample points, summation extends over all these points, $K(x;h)$ is the kernel, and h is the kernel scale. Kernels may be quite arbitrary, there are only four conditions that they must satisfy:

$$K(x) > 0, \tag{2.2}$$

$$\int K(x)dx = 1, \tag{2.3}$$

$$\int xK(x)dx = 0, \tag{2.4}$$

$$\int x^2 K(x)dx < \infty. \tag{2.5}$$

In other words, the kernel must be a symmetric probability distribution of finite variance.

Practice has shown that the exact functional form of a kernel does not matter much, but the scale does – choosing the right scale we can minimize the goodness measure of our density estimate, its MISE (mean integrated squared error):

$$\mathrm{MISE}(h) = E \int \left(\hat{\rho}(x;h) - \rho(x)\right)^2 dx,$$

where $\hat{\rho}(x;h)$ is the estimate of the density found using the scale h, $\rho(x)$ is the true density, and E denotes the expectation value. The MISE is, in fact, the only number that is used to compare how effective the density estimators are.

The formula 2.1 is referred to as a fixed kernel estimate. In practice, the density distributions are frequently non-uniform, and we could get a better estimate by varying the scale h. These estimates are called adaptive estimates, and there are two kinds of them: the sandbox estimate where the kernel size depends on the data points:

$$\rho(x) = \frac{1}{n} \sum_i^n \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right),$$

https://doi.org/10.1017/S1743921316009947 Published online by Cambridge University Press

and the balloon estimate where the kernel size depends on the position where we estimate the density:

$$\rho(x) = \frac{1}{nh(x)} \sum_i^n K\left(\frac{x - x_i}{h(x)}\right)$$

There are several empirical rules and iterative methods to find better h_i or $h(x)$.

In a multidimensional case, the kernel $k(\mathbf{x})$ is also multidimensional:

$$\rho(\mathbf{x}) = \frac{1}{nh} \sum_i^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \tag{2.6}$$

It is usually constructed as a spherical kernel or a direct product of one-dimensional kernels (for Gaussian kernels, this is the same). As for adaptive kernels, the one-dimensional thinking has carried over to the multidimensional case, and for adaptive density estimates, people usually try to construct adaptive kernels as spherical kernels of different scale. This is not good, as it smears up the local galaxy distribution. Another possibility is to use products of one-dimensional kernels of different scale; these scales have to be prescribed, somehow. But there is more freedom in the multidimensional case. For example, a (balloon) density estimate using a Gaussian kernel is

$$\rho(\mathbf{x}) = \frac{1}{(2\pi|\Sigma|)^{D/2}} \sum_i^n \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{x}_i)\right]. \tag{2.7}$$

Here the scaling is fixed by the covariance matrix $\sum(\mathbf{x})$, and this is, in general, not diagonal.

We propose to use a local estimate for this matrix:

$$\mathbf{\Sigma}(\mathbf{x}) = \frac{1}{n} \sum_i^n w(\mathbf{x}, \mathbf{x}_i, R)(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i - \mathbf{x})^T,$$

that is similar to the usual covariance matrix estimate, but is restricted to a region near the point $\mathbf{x}$ by the weight function $w(\mathbf{x}_i, \mathbf{x}_j, R)$. We choose it to be also Gaussian of rms R:

$$w(\mathbf{x}_i, \mathbf{x}_j, R) = \frac{1}{(2\pi)^{D/2} R^D} \exp\left[-\frac{1}{2R^2}(\mathbf{x}_i - \mathbf{x}_j)^2\right].$$

It is easy to see that in the case of a locally constant density, our covariance matrix will be diagonal,

$$\mathbf{\Sigma} = R^2 \mathbf{I},$$

where $\mathbf{I}$ is the unit matrix of dimension D. In a general case, the density distribution is described not only by the scalar $\rho(\mathbf{x})$, but, in addition, by the eigenvalues $S_k, k \in [1, D]$ of the covariance matrix, and by the eigenvectors $\mathbf{v}_k$, the axes of the covariance ellipsoid. This describes the local anisotropy of the galaxy distribution.

Solving for eigensystems is usually a delicate iterative process. But for three-dimensional maps there are direct analytical algoritms giving both the eigenvalues and eigenvectors (see, e.g., Kopp (2008)). The necessary libraries are freely available†. And analytical formulae exist also for two-dimensional matrices. So for the usual 3D and 2D maps the algorithm is fast.

† See `http://www.mpi-hd.mpg.de/personalhomes/globes/3x3/`

https://doi.org/10.1017/S1743921316009947 Published online by Cambridge University Press

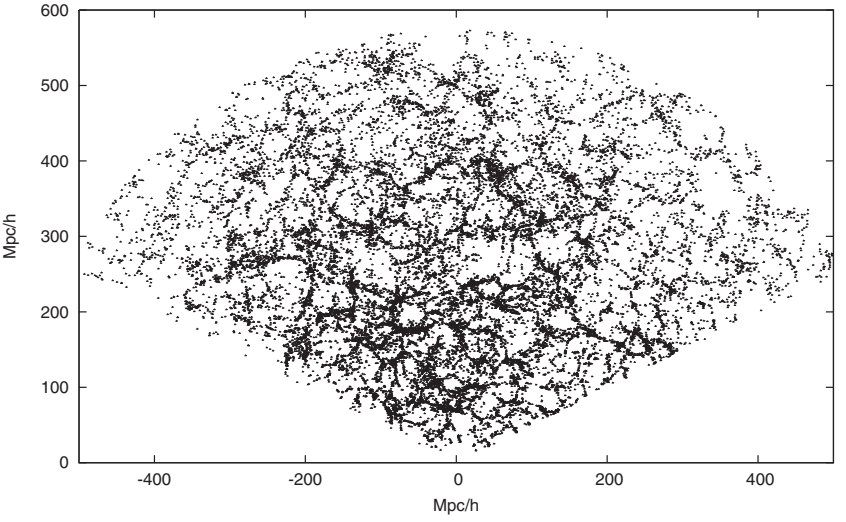

Figure 1. The 2D projection of galaxies for the thin Sloan sample slice used.

3. Examples

For data, we use the SDSS DR8 galaxies, where we have found the groups and spherisized their velocity space fingers (Tempel *et al.* (2012)). To see better how the algorithm works, we show a 2-D example, and select for that galaxies from the equatorial slice, where the SDSS survey coordinate $\eta \in [-2, 2]$ degrees. Fig. 1 shows the projected galaxy distribution in this slice. We compare the Gaussian and adaptive densities for a sensible $R = 10\mathrm{Mpc}h^{-1}$. As expected, adaptive kernels restore the density better than the standard Gaussian one. For the 2D case, we get also the axes ratio that describes the local anisotropy of the galaxy distribution.

4. Summary

The main advantage of the present approach is that it allows us to use the local anisotropy on the galaxy distribution. There are many ways to describe the galaxy maps and to classify the elements of the large-scale structure that rely on the local properties of the smooth density field. It is clear that density maps obtained by counting galaxy numbers in cells or by using isotropic kernels smooth out this information to a very large extent.

The local adaptive estimates of the kernel width are also useful, but not to such extent. We know that galaxy distribution is of a multiscale nature – there are clusters, groups, and filaments of different scale. For one smoothing scale, we get filaments, for another scale, these filaments may form a wall. So there probably is no unique density distribution. These advices us that we should not use our algorithm in an iterative way, although it may be tempting.

The computer code for the algorithm can be found on GitHub†. It is written in C, and includes several tricks to speed up the algorithm – arranging the data for a fast neighbour search, using a compact kernel for the final density estimation, etc.

† `https://github.com/esaar/andens`.

https://doi.org/10.1017/S1743921316009947 Published online by Cambridge University Press

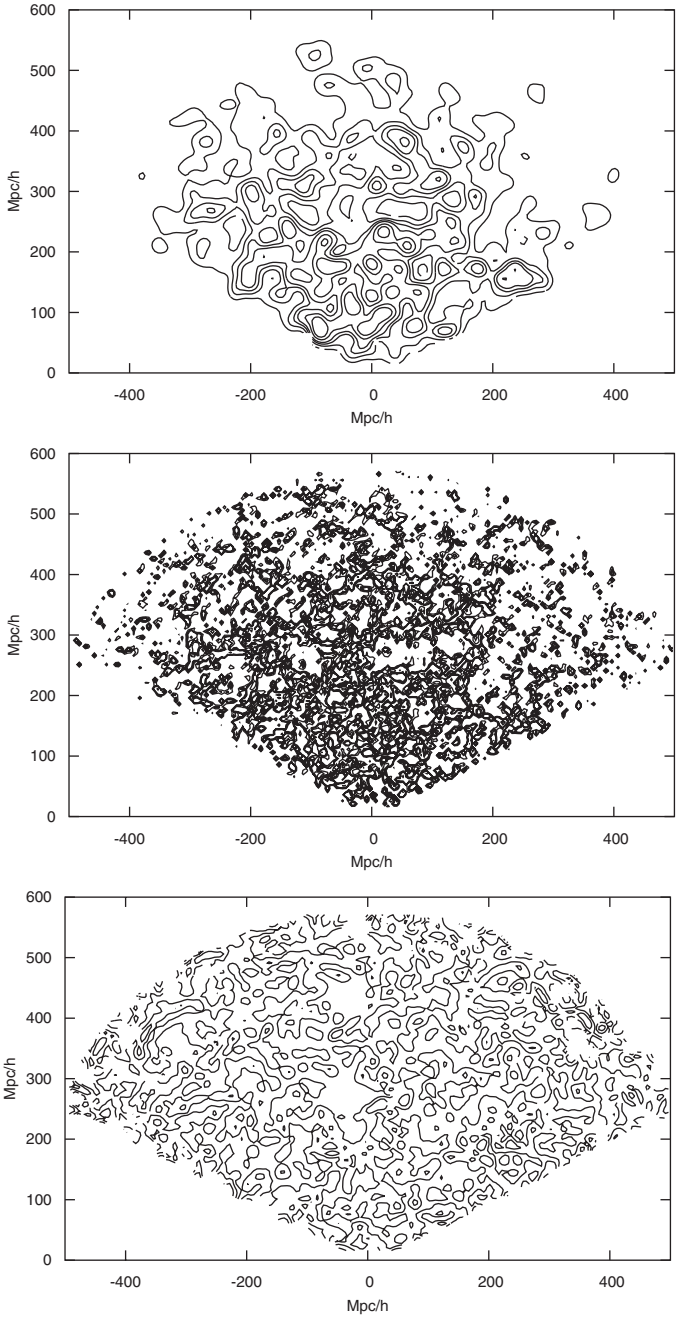

Figure 2. Sloan slice densities for Gaussian smoothing ($R = 10\mathrm{Mpc}h^{-1}$), upper panel, and for adaptive smoothing for the same initial σ, and axes ratios for the adaptive kernels.

References

Jasche, J. & Wandelt, B. D 2012, *Monthly Notices of the Royal Astronomical Society*, 425, 1042

Kitaura, F.-S., Erdoğdu, P., Nuza, S. E., Khalatyan, A., Angulo, R. E, Hoffman, Y., & Gottlöber, S. 2012, *Monthly Notices of the Royal Astronomical Society*, 427, L35

https://doi.org/10.1017/S1743921316009947 Published online by Cambridge University Press

Silverman, B. W. 1986, *Density Estimation for Statistics and Data Analysis. Chapman & Hall, London*

Kopp, J. 2008, *International Journal of Modern Physics C*, 19, 523

Tempel, E., Tago, E.,& Liivamägi, L. J. 2012, *Astronomy and Astrophysics*, 540, A106

https://doi.org/10.1017/S1743921316009947 Published online by Cambridge University Press