Nonparametric Density Estimation and Clustering in Astronomical Sky Surveys

Woncheol Jang[1]

Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708, USA

Abstract

We present a nonparametric method for galaxy clustering in astronomical sky surveys. We show that the cosmological definition of clusters of galaxies is equivalent to density contour clusters (Hartigan , 1975) $S_c = \{f > c\}$ where f is a probability density function. The plug-in estimator $\widehat{S}_c = \{\widehat{f} > c\}$ is used to estimate S_c where $\widehat{f}$ is the multivariate kernel density estimator. To choose the optimal smoothing parameter, we use cross-validation and the plug-in method and show that cross-validation method outperforms the plug-in method in our case. A new cluster catalogue, database of the locations of clusters, based on the plug-in estimator is compared to existing cluster catalogs, the Abell and Edinburgh/Durham Cluster Catalogue I (EDCCI). Our result is more consistent with the EDCCI than with the Abell, which is the most widely used catalogue. We use the smoothed bootstrap to asses the validity of clustering results.

Key words: Galaxy clustering, Density contour cluster, Plug-in estimator, Cross-validation, Smoothed bootstrap.

1 Introduction

It is widely assumed that *the universe is homogeneous and isotropic* which means that it looks the same in all directions and matter is distributed evenly throughout space. This is a key assumption in modern cosmology. There is evidence that supports this assumption on large scales, but on smaller scales,

Email address: `wjang@stat.duke.edu` (Woncheol Jang).

[1] The author would like to thank Larry Wasserman and Bob Nichol for their helpful comments and suggestions

one can find significant deviations from homogeneity and isotropy such as walls, filaments, and clusters of galaxies (Martínez and Saar , 2002).

Knowing how the universe has evolved would lead to the answers for discrepancy of the structure at different scales. According to the classic scenario, the universe has been expanding since the Big Bang. Due to small fluctuations which were present at early epochs, the universe has been clumped and clustered since then. Clusters of galaxies play an important role in finding where the local structure fades away into a homogeneous and isotopic distribution. However, the availability of complete, accurate cluster catalogs for such studies is very limited.

The Abell catalogue has been one of the most widely used catalogs for cosmological research. It was published by George Abell in 1958 and covers the whole northern hemisphere. A catalogue of the southern hemisphere of the sky, was completed by Abell and his colleagues in 1989. It contains 4,073 clusters over the entire sky. Much cosmological research have been done based on the Abell catalogue. However, the Abell catalogue was constructed by a visual scan of photographic material, thus it is subjective and not consistent with other catalogs. Cosmologists have been suspicious that inconsistency of the Abell catalogue would undermine scientific results based on it and they started to build new, objective and accurate large area cluster catalogs which would eventually replace the Abell catalogue. Thus, large area sky surveys using present new technologies, 8-meter optical telescopes, new X-ray and microwave satellites, are currently planned or underway.

The power of modern technology is opening a new era of massive astronomical data that is beyond the capabilities of traditional methods for galaxy clustering. The advent of new massive sky survey brings statisticians to cosmologists and the explosion of the data in cosmology is a blessing and a curse to statisticians. It is a blessing because the amount of data makes nonparametric statistical methods very effective. The same features also limits the use of nonparametric statistical methods without efficient data management. Therefore, efficient automated clustering algorithms will be critical to make the use of nonparametric statistical methods in cosmology.

There has been great progress in the development of new automated clustering algorithms in cosmology. The first generation of automated clustering algorithms are simple variants on the peak-finding algorithm (Lumsden et al , 1992). In recent years, several new, more sophisticated, algorithms have become available including the adaptive matched-filter (AMF) algorithm (Kepner et al , 1999). The idea behind the AMF algorithm is to identify clusters by finding peaks in a kind of likelihood map. To generate the map, a filter, based on a model of the distribution of galaxy, is matched with the data and contour is produced based on the model. To find clusters, one chooses a threshold given by cosmological theory, then all the data below this threshold are removed and what remains are regarded as clusters. The AMF algorithm is very complicated and not rigorously justified. Moreover, the size of the filter (smoothing parameter) is arbitrarily chosen.

The goal of this paper is to produce a catalogue based on nonparametric methods and compare the catalogue with the Abell and EDCCI, the first catalogue based on the AMF algorithm.

This paper is organized as follows. Section 1 is a brief introduction. In Section 2, we explain how to model the spatial locations of galaxy clusters. Then the data, the Edinburgh-Durham Southern Galaxy Catalogue (EDSGC) which are used in our analysis, are described in Section 3. Section 4 outlines nonparametric methods. Section 5 summarizes the results. Finally, we discuss scientific contributions of our results and address possible extensions in Section 6.

2 Stochastic Model of the Galaxy Distribution

Let $Y_1, \ldots, Y_n$ be the positions of the galaxies in a region C. We assume that Y_i is a realization of a Poisson process with the intensity measure $\Lambda_t(C) = \int_C \lambda(y)dy$, the mean number of galaxies inside C at time t. Here $\lambda(y)$ is the intensity function.

Let $\rho_t(y)$ be the mass density function of objects such as galaxies at time t, i.e.,

$$\int\limits_A \rho_t(y)dy \equiv \text{ total mass in a region } A.$$

Cosmologists assume the mean number of galaxies inside a region C is directly proportional to the total mass inside the region. Hence the intensity measure $\Lambda_t(C)$ is

$$\Lambda_t(C) \propto \int\limits_C \rho_t(y)dy.$$

Define the overdensity

$$\delta_t(y) \equiv \frac{\rho_t(y) - \bar{\rho}}{\bar{\rho}},$$

where $\bar{\rho} \equiv \int_U \rho(y)dy / \int_U dy$.

The overdensity is a kind of normalized mass density function and used as a scale free reference threshold. For example, it is believed that galaxies form at overdensity ≈ 1.

Cosmology theory also assumes that the overdensity $\delta_0(x)$ is a realization of a Gaussian process and has evolved, at time t, to $\delta_t = H(\delta_o, t)$. Here, $H(\cdot, \cdot)$ is a complicated nonlinear function.

The early universe was very hot and dense, but began to cool down due to the expansion after the Big Bang. As a result, small fluctuations began to exist due to temperature differences. The fluctuations, large localized overdensities, eventually collapsed to form *virialized* objects or self-gravitating objects such as clusters due to gravitational instability. After overdensites are virialized, the universe is in the form of stability or balance. In other words, the universe is in some form of dynamic equilibrium. To reach dynamic equilibrium, the mass of virialized object must satisfy the following geometric condition.

$$C = \left\{ y \,\middle|\, \rho_t(y) > \delta_c \right\}, \tag{1}$$

where δ_c is the present day $(t = 0)$ overdensity that has collapsed and virialized at time t and can be expressed as a function of unknown cosmological parameters. See Reichart et al (1999) for the details.

Since the mass density function ρ_t is modeled as a random process, the Poisson model results in a double-stochastic process, i.e. the Cox process.

From the condition (1), it is clear that one must estimate the intensity function to understand the spatial locations of clusters of galaxies. In fact, similar problems can be found in spatial statistics literature (Diggle , 1985; Cressie , 1991) where the kernel density estimator was used to estimate the intensity function. Since the mass density function ρ_t can be considered as a probability density function, the cosmological definition of clusters is indeed the same as Hartigan's density contour clusters: the connected components of level sets. In other words, galaxy clustering is equivalent to level set estimation.

A naive estimator for the level set is the plug-in estimator $\widehat{S}_c \equiv \{\widehat{f} > c\}$ where $\widehat{f}$ is a nonparametric density estimator. The consistency of the plug-in estimator with the kernel density estimator was proved by Cuevas and Fraiman (1997) in terms of a set metric such as the symmetric difference d_μ and the Hausdorff metric d_H,

$$d_\mu \equiv \mu(T \Delta S), \quad d_H(T, S) \equiv \inf\{\epsilon > 0 : T \subset S^\epsilon, T^\epsilon \subset S\},$$

where Δ is symmetric difference, μ is Lebesgue measure and S^ϵ is the union of all open balls with a radius ϵ around points of S.

Therefore, we can use the plug-in estimator to estimate a level set given density estimates and use the connected components of level set estimates as clusters of galaxies.

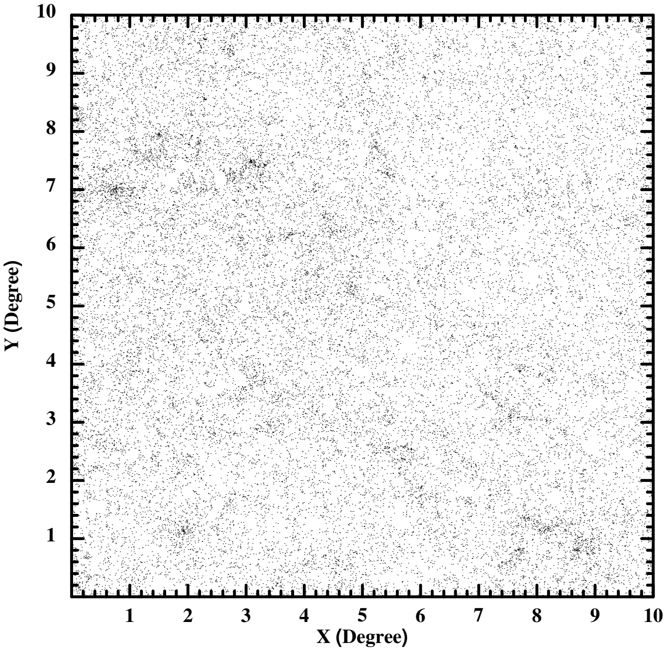

Fig. 1. 10× 10 degree subset of EDSGC

3 Astronomical Sky Survey Data

The source of galaxy data used in our analysis is the EDSGC, which consists of survey plates scanned with the Edinburgh plate measuring machine COSMOS.

The equatorial system is used to list the locations of objects in the catalogue. Each objects list the right ascension (RA) and declination (DEC), the longitude and latitude with respect to the Earth. The right ascension takes values from 0 to 360 degrees of in hours, minutes and seconds (24 hrs = 360 degree) while declines takes values from 0 to 90 degrees for objects in northern hemisphere and -0 to -90 degrees for objects in the southern hemisphere. EDSGC is located from 22 to 3 hrs (through 0 hrs) and -23 to -42 degree. A 10 × 10 degree subset of EDSGC is used in our data analysis and it is located from 0 hrs to 0 hrs 40 mins and -28 to -38 degree.

The catalogue contains 1.5×10^6 galaxies and because of its size, a 10 degree subset of the EDSGC is used in our data analysis. The origin of subset is 0 hr and -28 degree. It contains 41,171 galaxies and each galaxy has 7 attributes.

Figure 1 shows the 10 degree subset of the EDSGC. According to the Abell catalogue, there are 43 clusters in that area while the EDCCI, the first catalogue based the AMF algorithm, shows 42 clusters (Lumsden et al , 1992).

Figure 2 shows the Abell and EDCCI catalogs. The points in Abell and EDCCI catalogs represent the locations of clusters. Both of them find clusters at the

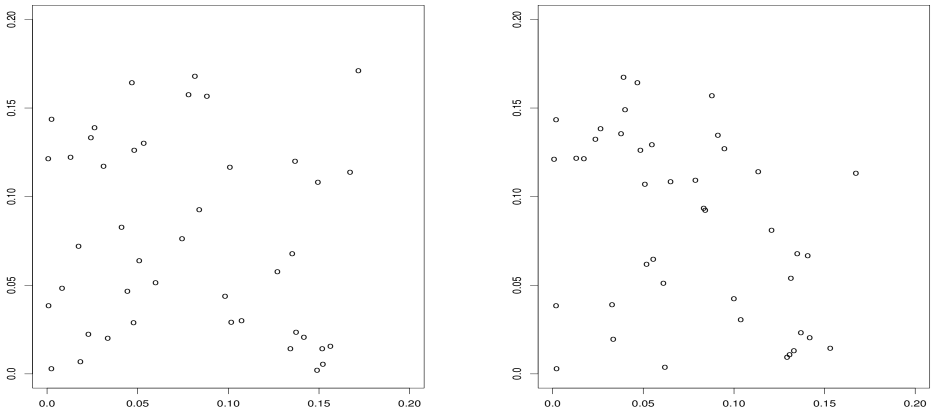

Fig. 2. Abell and EDCCI Catalogue

upper left corner and lower right corner which are obvious from data. However, there is inconsistency between two catalogs. A cluster in the upper right side in the Abell is not found in the EDCCI and the cluster pattern doesn't look similar. By overlapping two catalogs, roughly a half of them are located within a reasonable Euclidean distance according to cosmologists.

4 Methodology

Suppose $Y_1, \ldots, Y_n$ are independent observations from a distribution with density f where $Y_i = (Y_{i1}, \ldots, Y_{id})$, d- dimensional vector. In our case, Y_i is the location of ith galaxy with $d = 2$. We define clusters as connected components of the plug-in estimator. Hence one must estimate the density f first. We use the kernel density estimator for the estimation of f.

4.1 Multivariate kernel density estimation

The general d dimensional kernel estimator is given by

$$\widehat{f}(y) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1/2}(y - Y_i)\right),$$

where $\mathbf{H}$, bandwidth matrix, is a symmetric positive definite $d \times d$ and K is a bounded, compactly supported d-variate kernel satisfying

$$\int K(y)dy = 1, \quad \int yK(y)dy = 0, \quad \text{and} \int yy^t K(y)dy = \mu_2(K)I.$$

Here $\mu_2(K) = \int y_i^2 K(y)$ is independent of i (Wand and Jones , 1995).

We assume that the contour of the kernel is ellipsoidal and elliptical axes of the kernel are aligned with the coordinate axis. In other words, we assume the bandwidth matrix is a diagonal matrix.

$$\mathbf{H} \in \mathcal{H}_2 = \{\mathrm{diag}(h_1^2, \ldots, h_d^2) : h_1, \ldots, h_d > 0\}.$$

Then, the density estimator is given by

$$\widehat{f}(y) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{n} K\left(\frac{y_1 - Y_{i1}}{h_1}, \ldots, \frac{y_d - Y_{id}}{h_d}\right).$$

However, it is a d-dimensional optimization problem and could be very computationally expensive.

We often assume a simpler form of the bandwidth matrix. Suppose that the contour of the kernel is spherically symmetric. Then the bandwidth is a class of

$$\mathcal{H}_1 = \{h^2 I : h > 0\}.$$

Another simple class of the bandwidth matrix is a class of

$$\mathcal{C} = \{h^2 \cdot \mathrm{diag}(\sigma_1^2, \cdots, \sigma_d^2)\},$$

where σ_i^2 is the variance of the ith coordinate variable. As pointed out in Wand and Jones (1993), this scaling approach is not always better than to use $\mathbf{H} \in \mathcal{H}_1$. In our case, the variance of each coordinate are the almost same because the universe is *almost* homogenous. Therefore there was little difference to use either $\mathbf{H} \in \mathcal{C}_1$ or $\mathbf{H} \in \mathcal{H}_1$ in our case.

4.2 Binned Kernel Estimation

To make the nonparametric density estimation effective with huge datasets, one must provide efficient algorithms to calculate the density estimates. The original formula requires $O(n^2)$ evaluation to calculate density estimates at every data point which easily can be a daunting task even with a moderate size of data.

The binned kernel estimation is an appealing way of approximation of kernel estimation with the fast Fourier transform (FFT).

Suppose that K is a symmetric kernel and define bins by equally spaced mesh of points over the support. If the support is infinite, one can replace it with "effective support", whose outside is negligible. For example, we can use $[-4, 4]$ as the effective support of the standard normal distribution.

Let n_j a number of data points in bin B_j which is centered at t_j and $j = 1, \ldots, m$. Then, the binned kernel estimator is given by

$$\widehat{f}(y) = \frac{1}{m} \sum_{j=1}^{m} \frac{n_j}{|\mathbf{H}|^{1/2}} K\left(\mathbf{H}^{-1/2}(y - t_j)\right).$$

The binned kernel estimator only requires $O(m)$ evaluation to calculate a density estimate. Furthermore, with FFT, it only requires $O(m \log m)$ to evaluate density estimates at every grid point.

The rule of thumb for the number of grid points is between 100 and 500 and the approximation is better as m increase (Wand and Jones , 1995).

For higher dimension, one may consider the Weighted Averaging of Rounded Points (WARPing) of Härdle and Scott (1992) to reduce computational complexity. The cost of the WARP and FFT are quite similar for the bivariate case, since both of them are based on equally spaced grid points.

4.3 Bandwidth Selection

Cosmologists have been paid more attention to the shape of the filter (kernel) than the size of the filter (bandwidth). Indeed it is well known in statistics that the choice of bandwidth is crucial in density estimation not the choice of the kernel.

Define the risk function

$$R(f, \widehat{f}) = E[\text{ISE}(\mathbf{H})].$$

Here ISE($\mathbf{H}$) is the integrated square error (ISE) which is given by,

$$\begin{aligned}\text{ISE}(\mathbf{H}) &= \int (\widehat{f}(y; \mathbf{H}) - f(y))^2 dy \\ &= R(\widehat{f}) - 2 \int \widehat{f}(y; \mathbf{H}) f(y) dy + C,\end{aligned}$$

where $C = \int (f(y))^2 dy$ does not depend on $\mathbf{H}$ and $R(\widehat{f}) = \int (\widehat{f}(y; \mathbf{H}))^2 dy$.

In terms of minimizing the risk function, the choice of the optimal bandwidth is far more important than the choice of kernel.

To choose the optimal density estimator (optimal bandwidth) is to find the estimator which minimizes the risk function. Therefore, finding the optimal bandwidth is equivalent to finding the bandwidth matrix which minimizes the first two terms. The idea of cross-validation method is to find bandwidth matrix which minimizes an unbiased estimator of the first two terms using leave-one-out.

Define

$$\mathrm{CV}(\mathbf{H}) = \int \widehat{f}(y;\mathbf{H})^2 dy - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(y_i;\mathbf{H}),$$

where

$$\widehat{f}_{-i}(y;\mathbf{H}) = \frac{1}{(n-1)h_1\cdots h_d}\sum_{j\neq i} K(\mathbf{H}^{-1}(y-Y_j)).$$

Assuming K is the multivariate standard normal density, CV($\mathbf{H}$) is given by

$$CV(h_1,\cdots,h_d) = \frac{1}{(2\sqrt{2\pi})^d n h_1\cdots h_d} + \frac{1}{(2\sqrt{2\pi})^d n^2 h_1\cdots h_d}\times\Delta,$$

where

$$\Delta = \sum_{i\neq j}\left[\exp\left\{-\frac{1}{4}\sum_{k=1}^{d}\left(\frac{y_{ik}-y_{jk}}{h_k}\right)^2\right\} - 2^{(d+2)/2}\times\exp\left\{-\frac{1}{4}\sum_{k=1}^{d}\left(\frac{y_{ik}-y_{jk}}{h_k}\right)^2\right\}\right].$$

See Sain, Baggerly and Scott (1994) for the details.

For $\mathbf{H}\in\mathcal{C}$,

$$\mathrm{CV}(h) = \frac{1}{(2\sqrt{2\pi}h)^d n\sigma_1\ldots\sigma_d} + \frac{1}{(2\sqrt{2\pi}h)^d n^2\sigma_1\cdots\sigma_d}\times\Delta^*,$$

where

$$\Delta^* = \sum_{i\neq j}\left[\exp\left\{-\frac{1}{4h^2}\sum_{k=1}^{d}\left(\frac{y_{ik}-y_{jk}}{\sigma_k}\right)^2\right\} - 2^{(d+2)/2}\times\exp\left\{-\frac{1}{4h^2}\sum_{k=1}^{d}\left(\frac{y_{ik}-y_{jk}}{\sigma_k}\right)^2\right\}\right].$$

Whereas cross-validation method uses a direct approach by finding minimum of ISE, the plug-in method uses asymptotic expansion of MISE, the expectation of ISE.

Assuming $\mathbf{H} \in \mathcal{C}$,

$$\begin{aligned}\text{MISE}(h) &= E\int (\widehat{f}(y;\mathbf{H}) - f(y))^d y \\ &\approx \frac{1}{4}h^4\sigma_1^4\cdots\sigma_d^4 a_0^2 \int \{\nabla f(y)\}^2 dy + \frac{a_1}{nh^d\sigma_1\cdots\sigma_d},\end{aligned}$$

where $\nabla f(y) = \sum_{i=1}^{d}(\partial^2/\partial y_i^2)f(y)$. Here $a_0^2 = \int y^2 K(y)dy$ and $a_1 = \int K^2(y)dy$.

Then, one can find the optimal bandwidth by numerical methods such as the Newton-Raphson method. However the solution still depends on an unknown functional $\nabla f(x)$ which we need to estimate. Wand and Jones (1995) proposed d-stage estimator to address the issue.

4.4 Assessment of Variability

To address the reliability of other catalogs based on our result, we want to assess the variability of our kernel estimates first. The bootstrap is a compelling method to assess variability of estimate in case the distributions of estimators are intractable. The key idea of the bootstrap is to use $\theta(F_n)$ to estimate a functional $\theta(F)$ where F and the cumulative density function and F_n is the empirical cumulative distribution function. Since F_n is discrete, in some situations a smooth estimate of F might be better. Silverman and Young (1987) showed that when the smoothed bootstrap works better.

The usual cases to use the smoothed bootstrap is where the effects of discreetness cause serious problems such as estimating density or sample median. For example, Silverman (1985) used a smoothed bootstrap test for multimodality.

The main idea of the smoothed bootstrap is to resample y^* from the kernel estimate $\widehat{f}$ instead of the raw data. The resampling steps are as follows.

Step 1 Choose integers $I_1, \ldots, I_n$ with equal probability from $1, \cdots, n$.
Step 2 Generate random variable z_i from $K(y_i)$ for $i = 1, \ldots, n$.
Step 3 Let $y_i^* = y_{I_i} + h \cdot z_i$ for $i = 1, \ldots, n$.
Step 4 Repeat step 1-3 N times.
Step 5 Construct bootstrap estimate $\widehat{f_j^*}$ based on each resample $y^{*j} = (y_1^{*j}, \ldots, y_n^{*j})$ for $j = 1, \ldots, N$.
Step 6 Find clusters defined by $\{y : \widehat{f_j^*}(y) > c\}$ in the jth resample for $j = 1, \ldots, N$ where c is a threshold.

If $h = 0$, the smoothed bootstrap estimate is the same as the naive bootstrap estimate.

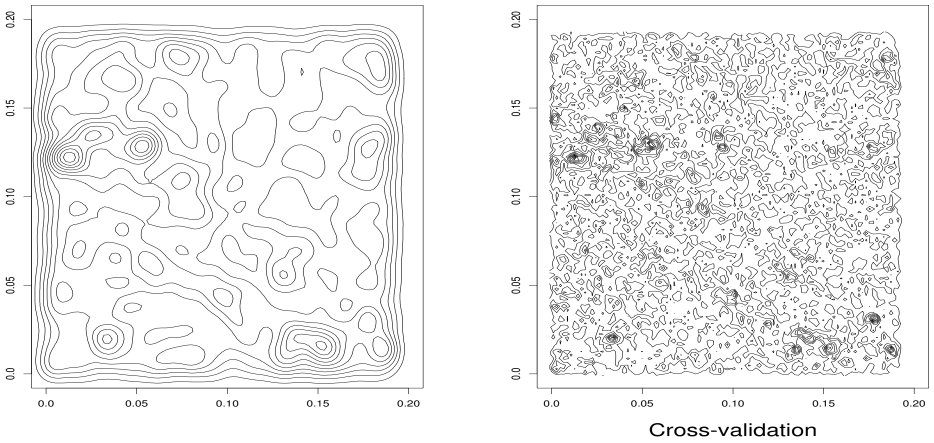

Fig. 3. Contour plot by Plug-in and Cross-validation

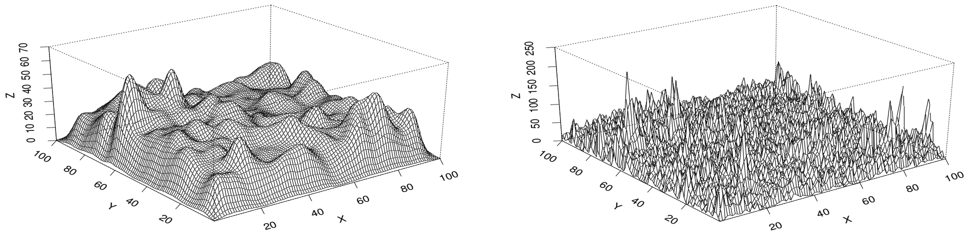

Fig. 4. Density estimates by Plug-in and Cross-validation

Once we have the smoothed bootstrap estimates, we are able to find a cluster map from each estimate. The frequency of appearance of a cluster within a reasonable range from the same position is a measure of consistency of the cluster in that position. In other words, the more we observe a cluster within a range from the same position at each bootstrap estimate, the more confidence we have that the cluster is not due to noise.

5 Results

The optimal bandwidth was selected by cross-validation and the plug-in method. The contour plots of density estimates and 2 dimensional plug-in bandwidth selection were implement by the R library “Kernsmooth” developed by Matt Wand.

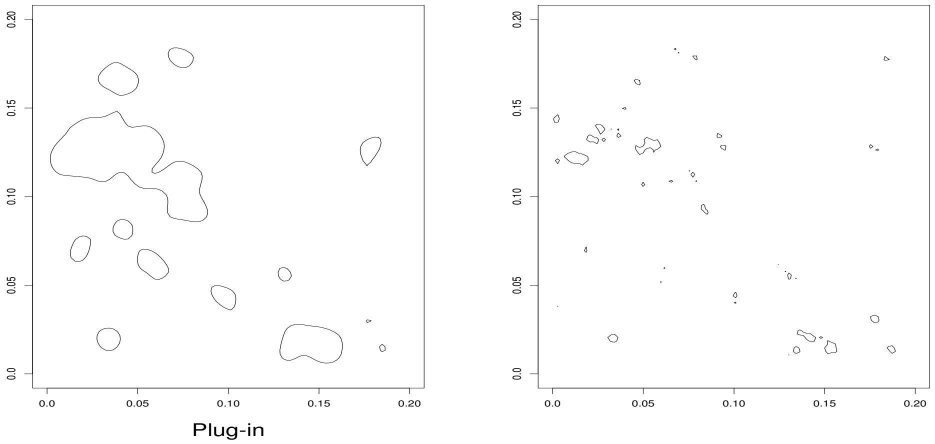

Fig. 5. Density Contours by Plug-in and Cross-validation

To calculate CV(h), R with Fortran subroutine was used and the minimum of CV(h) was found by "nlm" function in R. The program for cross validation method is available upon request.

The optimal bandwidth matrices by the plug-method and cross-validation are diag(0.005046, 0.005065) and diag(0.000734, 0.000701) respectively. Figure 3 and 4 show contour plots and density estimates by the plug-in and cross-validation.

As pointed out by Loader (1999), the "plug-in" method often oversmooths. Finding clusters is equivalent to finding *sharp and high* peaks which requires the smaller bandwidth. To find clusters from the Plug-in estimator $\{\widehat{f} > c\}$, one must determine the threshold c.

While determining the exact threshold from cosmological consideration is an ongoing research problem, contemporary cosmological theory provides a short range of possible threshold values. Within the range of the threshold, we first compare the plug-in method with cross-validation. Figure 4 gives a snapshot of density contours by both methods given a threshold within the range. It is clear that we can find at most a dozens of clusters by the plug-in method while cross-validation captured the feature of the data well within the range.

Neither the Abell nor EDCCI catalogs were built based on the threshold from cosmological theory. Hence, we use an ad-hoc method to choose a threshold to compare our catalogue with them. Specifically, we choose a threshold within the range such that we can find 42-43 clusters, the number of clusters in the Abell and EDCCI by cross-validation method.

To convert density contours into to clusters, we first find a subset of equally spaced grid points belonging to $\{\widehat{f} > c\}$ and find clusters by agglomerating the grid points. In other words, we use $\bigcup_{i=1}^{k_n} B(t_i, r)$ to approximate $\{\widehat{f} > c\}$

Table 1
Number of clusters in 4 quadrants at each catalogue

catalog	Upper Left	Upper Right	Lower Right	Lower Left	Total
Abell	12	5	11	15	43
EDCCI	18	2	12	10	42
Cross-validation	21	3	11	8	43

where $B(t_i, r)$ is a closed ball centered at t_i with radius r. Here t_i's are equally spaced grid points such that $\widehat{f}(t_i) > c$ and r is a half of the grid size. If two balls are adjacent to each other, we consider them to belong to the same cluster. In case a density contour has a region with more than one peak, we consider it as one cluster as long as they are connected components of the density contour. See Jang (2004) for the details.

In practice, it is not easy to check frequencies of appearance of clusters. Furthermore, cosmologists are not interested in the locations of clusters, but the number and size of the clusters, that is the mass distribution of the clusters. Hence we focus on the number of clusters. In figure 5, by the plug-in method, we only find a dozen of clusters, while 43 clusters are found by cross-validation method. The number of common clusters within a reseasonable Euclidean distance between the EDCCI and ours is 27 while the number between the Abell and ours is 20. Also the pattern of clusters in the EDCCI is more similar to ours.

Figure 6 shows 20 smoothed bootstrap estimates. Each bootstrap estimate was constructed based on resample $y^{*j} = (y_1^{*j}, \ldots, y_n^{*j})$ for $j = 1, \ldots, 20$. To resample y^{*j}, we used the following steps.

(1) Choose I_i from $\{1, \ldots, n\}$ with probability $1/n$ for $i = 1, \ldots, n$
(2) z_i is generated from bivariate normal with mean $y_i^T = (y_{i1}, y_{i2})$ and variance matrix $\text{diag}(\sigma_1^2, \sigma_2^2)$. Here σ_j^2 is the variance of $j$the coordinate variable.
(3) Let $y_j^* = y_{I_i} + h \cdot z_i$ where h is the optimal bandwidth by cross-validation.
(4) Find clusters from density contours $\{y : \widehat{f}_j^*(y) > c\}$ where c is the threshold which was used to find clusters in the Abell and EDCCI catalogue and $\widehat{f}_j^*$ is the kernel density estimator based on y_j^*.

It is not easy to match the clusters from smoothed bootstrap estimates to smoothed bootstrap estimates since the number of clusters may not be the same. Instead of comparing the locations of clusters, we count the number clusters at each quadrant and use it as a heuristic measure of consistency.

To compare them more precisely, we divide each catalogue into 4 quadrants. Table 1 reports the number of clusters in each quadrant for the Abell, EDCCI

Table 2
Number of cluster in 4 quadrants at smoothed bootstrap catalogs

	Upper Left	Upper Right	Lower Right	Lower Left	Total
1	16	3	8	7	34
2	17	4	8	12	41
3	20	4	7	11	42
4	19	2	7	11	39
5	17	3	7	9	36
6	19	3	7	10	39
7	18	2	8	11	39
8	15	3	8	8	34
9	22	4	7	13	46
10	12	3	8	8	31
11	17	2	7	11	37
12	18	2	7	12	39
13	19	2	8	8	37
14	10	4	9	11	34
15	13	3	8	7	31
16	19	4	8	11	42
17	18	4	6	13	41
18	17	5	9	7	38
19	22	2	8	7	39
20	19	3	8	10	40
Average	17.35	3.1	7.65	9.85	37.95
Median	18	3	8	10.5	39

and our new catalogue. Since we use a union of balls to approximate clusters, we assigned the clusters to each quadrant depending on the centroid of the union of the ball. While the total number of clusters in each catalogue is almost the same, the number of clusters in each quadrant is quite different. Indeed, almost a half of clusters in the EDCCI and our catalogue are located in upper left quadrant but one can find more clusters in lower left quadrant than any other quadrants in the Abell catalog.

Table 2 shows the smoothed bootstrap results which are consistent with the EDCCI and our catalogue; upper left quadrant is the most crowded area.

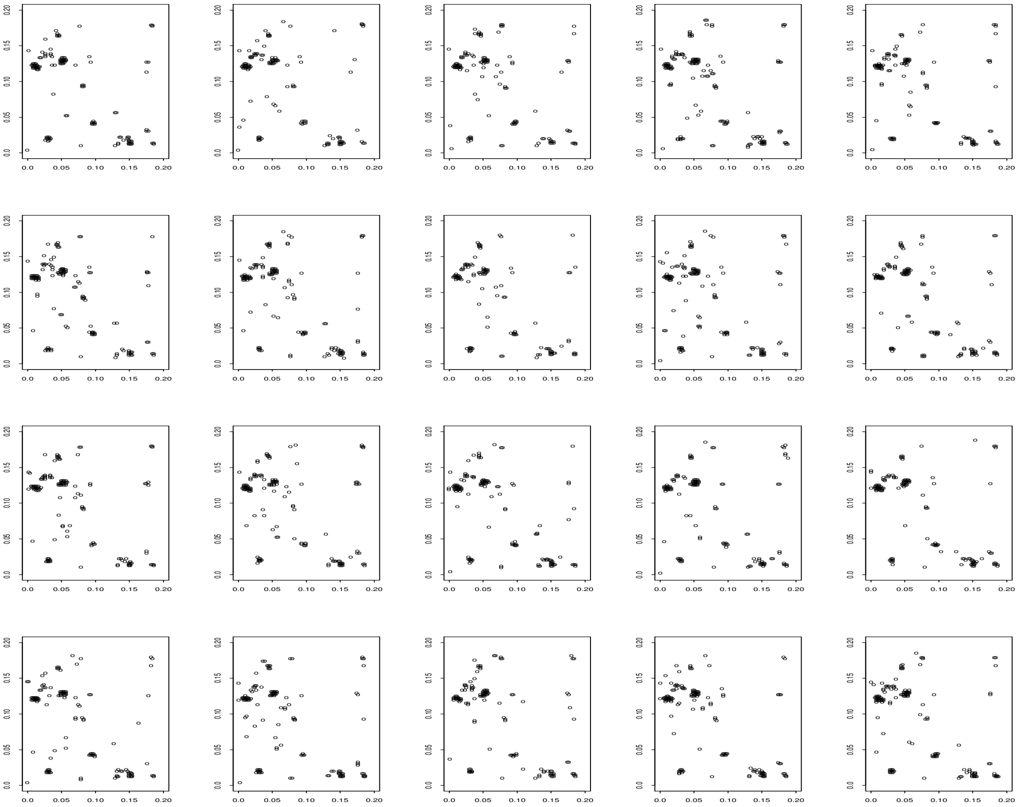

Fig. 6. Smoothed bootstrap estimates

One discrepancy between our result and the smoothed bootstrap is the total number of clusters. The average and median of total number of clusters in the smoothed bootstrap are 37.95 and 39. We found more clusters in the upper left and lower right quadrant, but less in the lower left quadrant in our catalogue. We suspect that some of those tiny clusters in the upper left and lower right quadrant in our catalogue are random noises. For the lower left quadrant, the smoothed bootstrap result shows the bimodality of the number of clusters and certainly our result belongs to the higher mode.

Another interesting result is that there is a cluster near the origin in the Abell and EDCCI, but is not found in our new catalogue. In figure 6, among 20 smoothed bootstrap estimates, only 7 of them have a cluster near origin. Another cluster in the upper right corner which is not found in the EDCCI, is found in every smoothed bootstrap estimates. Based on the smoothed bootstrap we suspect the cluster near origin is a random noise but the cluster on upper right side is real.

6 Discussion

The choice of bandwidth selectors are still widely on debate. For our case, contemporary cosmological theory agrees to the result provided by cross-validation method. While the convergence rate of the cross-validation selector to a true bandwidth is much slower that by plug-in selector's rate (Duong and Hazelton , 2004a), we are more interested in the behavior of plug-in estimator $\{\widehat{f} > c\}$. Also the faster convergence rate of bandwidth selectors does not imply that the plug-in estimator is asymptotically inefficient estimate (Loader , 1999). Furthermore, the difference between two selectors convergence rates for bivariate is small (≈ 2.6) even for $n = 100,000$.

Recently Duong and Hazelton (2004b) suggested to the use of full bandwidth matrix and introduced a new version of a smooth cross-validation method for full bandwidth matrix which might be beneficial for our case.

We also tested the reliability of other catalogs. The Abell catalogue shows inconsistency whereas the EDCCI works reasonably well. We used the smoothed bootstrap to assess the variability of the estimates.

References

Collins, C.A., Heydon-Dumbleton, N.H. and MacGillivray, H.T. 1988. The Edinburgh/Durham Southern Galaxy Catalogue-I. First results on the galaxy angular correlation function. *Mon. Not. R. astr. Soc.* **236**, 7-12.

Cressie, N. 1991. *Statistics for Spatial Data.* Wiley, New York.

Cuevas, A. and Fraiman, R. 1997. A plugin approach to support estimation. *Ann. Statist.* **25**, 2300-2312.

Diggle, P. 1985. A Kernel Method for Smoothing Point Process Data. *Appl. Statist.* **34**, 138-147

Duong, T. and Hazelton, M.L. 2004a. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J. Multivariate Anal.*, To appear.

Duong, T. and Hazelton, M.L. 2004b. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Stat.*, To appear.

Härdle, W.K. and Scott, D.W. 1992. Smoothing by weighted averaging of rounded points. *Computation. Stat.* **7**, 97-128.

Hartigan, J. 1975. *Clustering Algorithm.* Wiley, New York.

Jang, W. 2004. A Fast Clustering Algorithm with Application to Cosmology. Technical Report **803**, Department of Statistics, Carnegie Mellon University.

Kepner, J.,Fan, X., Bahcall, N., Gunn, J. and Lupton, R. 1999. An Automated Cluster Finder: the Adaptive Matched Filer. *Astrophys. J.* **517**, 78-91.

Loader, C. 1999. Bandwidth Selection: Classical or Plug-in? *Ann. Stat.* **27**, 415-438.

Lumsden, S.L., Nichol, R.C., Collins, C.A. and Guzzo, L. 1992. The Edinburgh/Durham Southern Galaxy Catalogue-IV. The Cluster Catalogue. *Mon.Not. R. astr. Soc.* **258**, 1-22.

Martínez, V. and Saar, E. 2002. *Statistics of the Galaxy Distribution.* Chapman and Hall, London.

Reichart, D., Nichol, R., Castander, F., Burker, D., Romer, A.K.,Holden, B., Collins, C., and Ulmer, M. 1999. A Deficit of High-Redshift, High-Luminosity X-Ray Clusters: Evidence for High Value of Ω_m? *Astrophys. J.* **518**, 521-532.

Sain, S.R., Baggerly, K.A., and Scott, D.A. 1994. Cross-validation of multivariate densities. *Journal of the American Statistical Association* **89**, 807-817.

Silverman, B.W. 1985. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Silverman, B.W. and Young 1987. The bootstrap: To smooth or not to smooth? *Biometrika.* **74**, 469-479.

Wand, M.P. and Jones, M.C. 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Stat. Assoc.* **88**, 520-528.

Wand, M.P. and Jones, M.C. 1995. *Kernel Smoothing.* Chapman and Hall, London.