

There are a couple of statistical approaches available to make inferences of level sets  $S_c = \{f > c\}$  - One is excess mass approach (Polonik, 1995), which is pretty theoretical and I am a little bit skeptical about its practical usage.

The other is the plug-in estimator  $\{\hat{f} > c\}$ , simply replacing  $f$  with a nonparametric density estimator such as a kernel density estimator. There is a rich literature on this topic. For the theory such as convergence rates and consistency, see Cuevas and Fraiman (1997).

In practice, it may not be easy to construct the plug-in estimator because of complicated geometrical structure of the estimator. An alternative is use a simplified version of the plug-in estimator. The plug-in estimator can be expressed as a union of some kind of balls. We can approximate the plug-in estimator by a union of "regular" balls - if you use a histogram as  $\hat{f}$  instead of a kernel estimator, you will get this type of the union of balls. Devroye and Wise (1980) provided some details for this type of estimators with applications to statistical quality control.

One more issue here is how to implement "the approximate plug-in estimator" - See Cuevas et al (2000). For high dimensional massive datasets, Wong and Moore (2002); Jang (2006) suggest modified versions of Cuevas et al (2000). I'll send all references to Linsong.

Cheers,  
Woncheol

[1] Cuevas, Febrero and Fraiman (2000) Estimating the number of clusters. Canadian J. of Stat. 28 367-382 [2] Cuevas and Fraiman (1997) A plugin approach to support estimation.

Ann. Stat. 25 2300-2312

[3] Devroye and Wise (1980) Detection of abnormal behavior via nonparametric estimation of the support. SIAM J. Applied Math. 38 480-488 [4] Jang (2006) An efficient clustering algorithm for massive datasets.

preprint.

[5] Polonik (1995) Measuring mass concentration and estimating density contour clusters - an excess mass approach. Ann. Stat. 23 855-881.

[6] Wong and Moore (2002) Efficient algorithms for nonparametric clustering with clutter. In Computing Science and Statistics 34 541-553.