Chapter 27
Level Set Estimation

P. Saavedra-Nieves, W. González-Manteiga, and A. Rodríguez-Casal

Abstract A density level set can be estimated using three different methodologies: Plug-in methods, excess mass methods, and hybrid methods. The three groups of algorithms to estimate level sets are reviewed in this work. In addition, two new hybrid methods are proposed. Finally, all of them are compared through an extensive simulation study and the results obtained are shown.

Keywords Level set estimation • Excess mass • Plug-in • Hybrid • Shape restrictions

27.1 Introduction

Level set estimation theory deals with the problem of reconstructing an unknown set of type $L(\tau) = \{f \geq f_\tau\}$ from a random sample of points $\mathscr{X}_n = \{X_1, \ldots, X_n\}$, where f stands for the density which generates the sample $\mathscr{X}_n$, $\tau \in (0, 1)$ is a probability, fixed by the practitioner, and $f_\tau > 0$ denotes the biggest threshold such that the level set $L(\tau)$ has a probability at least $1 - \tau$ with respect to the distribution induced by f. Figure 27.1 shows the level sets for three different values of the parameter τ. The problem of estimating $L(\tau)$ has been analyzed using three different methodologies in the literature: Plug-in methods, excess mass methods, and hybrid methods. We will present these three groups of automatic methods to reconstruct level sets and we will compare them through a detailed simulation study for dimension 1. We have restricted ourselves to the one-dimensional case because some of these methods have not yet been extended for higher dimension (see [8] or [6] for example). In Sect. 27.2, we will present and compare the plug-in methods. In Sects. 27.3 and 27.4, we will study the behavior of excess mass methods and hybrids methods, respectively. Finally, we will compare the most competitive methods in each group in Sect. 27.5.

P. Saavedra-Nieves (✉) • W. González-Manteiga • A. Rodríguez-Casal
Universidad de Santiago de Compostela, Santiago de Compostela, Spain
e-mail: paula.saavedra@usc.es; wenceslao.gonzalez@usc.es; alberto.rodriguez.casal@usc.es

M.G. Akritas et al. (eds.), *Topics in Nonparametric Statistics*, Springer Proceedings in Mathematics & Statistics 74, DOI 10.1007/978-1-4939-0569-0_27

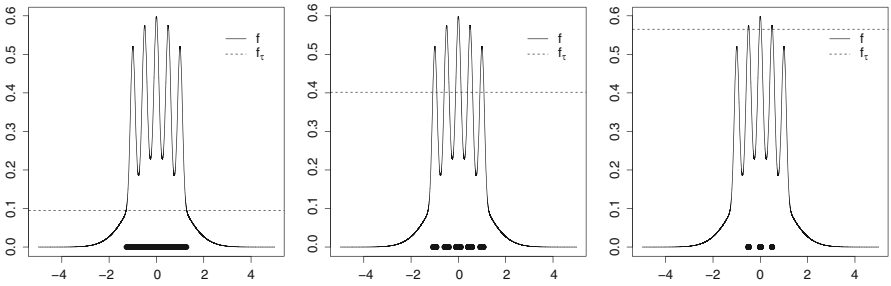

Fig. 27.1 Level sets for a one-dimensional density with $\tau = 0.1$ (*first column*), $\tau = 0.5$ (*second column*) y $\tau = 0.9$ (*third column*)

27.2 Plug-in Methods and Simulations Results

The simplest option to estimate level sets is the so-called plug-in methodology. It is based on replacing the unknown density f by a suitable nonparametric estimator f_n, usually the kernel density estimator. So, this group of methods proposes $\hat{L}(\tau) = \{f_n \geq \hat{f_\tau}\}$ as an estimator, where $\hat{f_\tau}$ denotes an estimator of the threshold. This is the most common approach but its performance is heavily dependent on the choice of the bandwidth parameter for estimating f. Baíllo and Cuevas were interested in choosing the best smoothing parameter to reconstruct a level set in the context of quality control. It was obtained by minimizing a cross-validation estimate of the probability of a false alarm, see [1]. Samworth and Wand proposed an automatic rule to select the smoothing parameter for dimension 1, see [8]. They derived a uniform-in-bandwidth asymptotic approximation of a specific set estimation risk function, $E\{d_{\mu_f}(L(\tau), \hat{L}(\tau))\}$, where $d_{\mu_f}(L(\tau), \hat{L}(\tau)) = \int_{L(\tau)\triangle\hat{L}(\tau)} f(t)dt$ and $\triangle$ denotes the usual difference given by $L(\tau)\triangle\hat{L}(\tau) = (L(\tau) \setminus \hat{L}(\tau)) \cup (\hat{L}(\tau) \setminus L(\tau))$. Of course, it is also possible to consider classical methods such as Seather and Jones or cross validation to select the bandwidth parameter although they are not specific to estimate level sets.

27.2.1 Simulations Results for Plug-in Methods

In this section, we will compare Baíllo and Cuevas' (BC), Samworth and Wand's (SW), Sheather and Jones' (SJ), and cross validation (CV) methods. The first two one are specific bandwidth selectors to estimate level sets. The last two algorithms are general selectors to estimate density functions.

We have generated $1,000$ samples of size $n = 1,600$ for the 15 Marron and Wand's density functions (see [5]) and we have considered three values for the parameter τ: $\tau = 0.2$, $\tau = 0.5$, and $\tau = 0.8$. Although there are several ways to estimate the threshold, we have estimated it by using Hyndman's method,

see [3]. This algorithm estimates the threshold by calculating the τ-quantile of the empirical distribution of $f_n(X_1), \ldots, f_n(X_n)$. We have considered the Sheather and Jones selector to calculate f_n. For each fixed random sample and each method, we have estimated the level set $L(\tau)$ and we computed the error of the estimation by calculating $d_{\mu_f}(L(\tau), \hat{L}(\tau))$. So, for a given model and a value of τ we have calculated $1,000$ errors for each method.

To facilitate the presentation of the results, we use some figures described below. Each figure is divided into rectangles that are painted with different colors according to the method (vertical axis) and the density model (horizontal axis). Colors are assigned as follows: light colors correspond to low errors and vice versa. So, this representation allows to detect the most or less competitive algorithm fixed the value of τ. Given a density, we have ordered the means of the $1,000$ errors calculated by testing if they are equal previously. If we reject the null hypothesis of equality between two means for the same model, then each method will be painted using a different color (darker or lighter according to the mean of the errors is higher or lower). In another case, both algorithms are represented using the same color. We will use this approach in the following sections to compare the methods of the two remaining groups of algorithms.

Figures 27.2 and 27.3 show the plug-in methods comparison for $\tau = 0.5$ and $\tau = 0.8$, respectively. For $\tau = 0.5$, the best results are provided by Sheather and Jones and cross validation selectors. If $\tau = 0.8$, then specific selectors for level sets have better results for the models 1, 2, 3, and 5. All of these densities have an only mode. They are very simple level sets. However, classical selectors are the most competitive for more sophisticated models such as 6, 8, 9, 10, 11, 12, 13, 14, or 15.

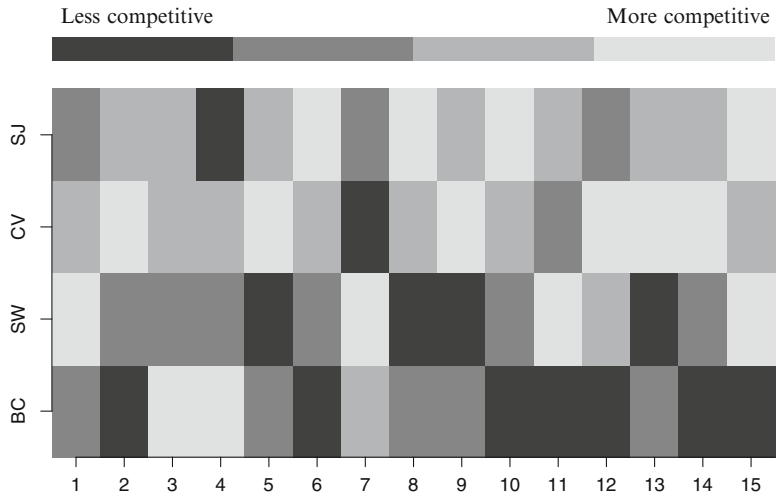

Fig. 27.2 Comparison of plug-in methods (*vertical axis*) with the 15 Marron and Wand's density models (*horizontal axis*), $\tau = 0.5$ and $n = 1,600$. The error criteria is d_{μ_f}

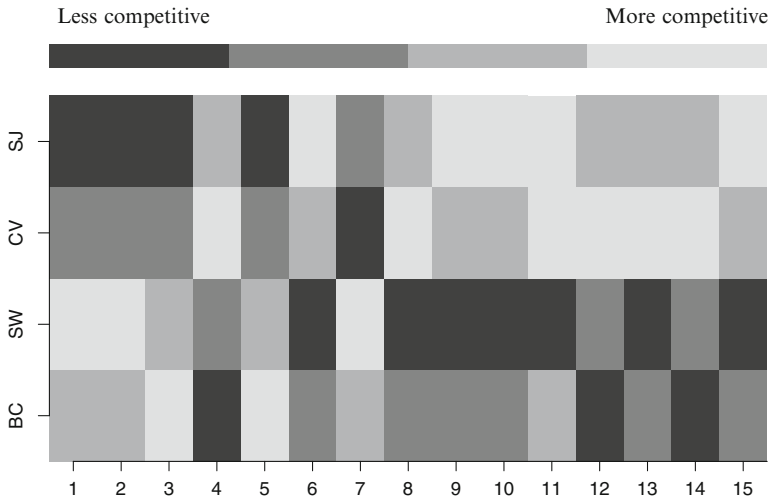

Fig. 27.3 Comparison of plug-in methods (*vertical axis*) with the 15 Marron and Wand's density models (*horizontal axis*), $\tau = 0.8$ and $n = 1,600$. The error criteria is d_{μ_f}.

As a conclusion, specific methods to estimate level sets do not improve the results of the classic bandwidth selection rules. In addition, cross validation and Sheather and Jones methods often provide similar results and they present the best global behavior.

27.3 Excess Mass Methods and Simulations Results

Another possibility consists of assuming that the set of interest satisfies some geometric condition such as convexity. Excess mass approach estimates the level set as the set of greatest mass and minimum volume under the shape restriction considered. For example, Müller and Sawitzki's method for one dimensional level sets assumes that the number of connected components, M, is known, see [6].

27.3.1 Simulations Results for Excess Mass Methods

Müller and Sawitzki's method depends on an unknown parameter M. This is the main disadvantage of this algorithm. We have considered five values for the number of clusters, $M = 1, 2, 3, 4$, and 5. We will denote the Müller and Sawitzki's method with M modes by MS_M.

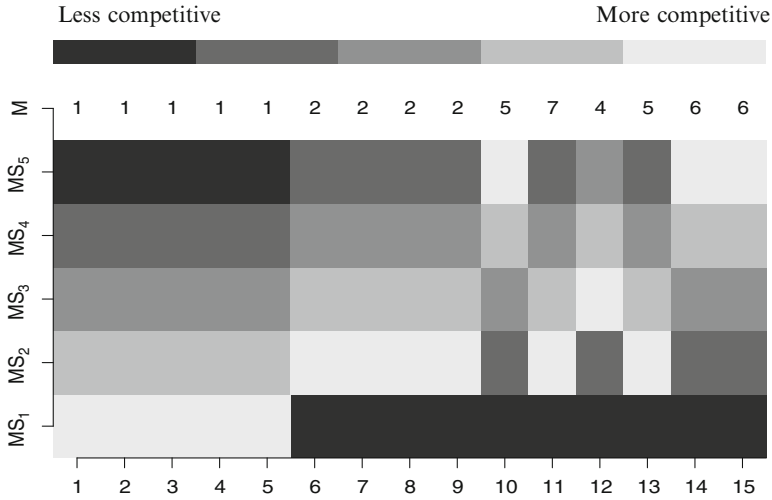

Fig. 27.4 Comparison of Müller and Sawitzki's method for different values of M (*vertical axis*) with the 15 Marron and Wand's density models (*horizontal axis*), $\tau = 0.5$ and $n = 1,600$. The error criteria is d_{μ_f}

To analyze the influence of the parameter M for Müller and Sawitzki's method, we will use Fig. 27.4. In this case, we have written the real number of modes for each density and $\tau = 0.5$ on the vertical axis too.

From Fig. 27.4, it is clear that Müller and Sawitzki's method is very sensitive to the parameter M. For $\tau = 0.5$, densities 1, 2, 3, 4, and 5 are unimodal and $M = 1$ provides the best results. Densities 6, 7, 8, or 9 have two modes and, in this case, the best value of M is $M = 2$. Model 10 has five modes for $\tau = 0.5$ and again $M = 5$ provides the best estimations. However, the best value of M for the Müller and Sawitzki's method is not equal to the real value of M for the models 11, 12, and 13 because some of their modes are not significant. In addition, if misspecification of M occurs, it can be seen that big values of M are better than a small values because the means of errors are lower.

27.4 Hybrid Methods and Simulations Results

As the name suggests, hybrid methods assume geometric restrictions and they use a pilot nonparametric density estimator to decide which sample points can be in the level set, $\mathscr{X}_n^+ = \{f_n \geq \hat{f}_\tau\}$. In this work we proposed two new hybrid methods to estimate convex and r-convex sets with $r > 0$. The last one is a shape condition more general than convexity. In fact, a closed set A is said r-convex with $r > 0$ if $A = C_r(A)$ where $C_r(A) = \bigcap_{\{B_r(x):B_r(x)\cap A=\emptyset\}} (B_r(x))^c$ denotes the r-convex hull of A, $B_r(x)$ denotes the open ball with center in x and radius r and

$(B_r(x))^c$, its complementary. Our two new proposals are based on the convex hull and r-convex hull methods for estimating the support, see [4] and [7], respectively. Under convexity restriction, we suggest estimating the level set as the convex hull of $\mathscr{X}_n^+$ and, under r-convexity, as the r-convex hull of $\mathscr{X}_n^+$. Another classic hybrid method is the so-called the granulometric smoothing method, see [9]. It assumes that the level set $L(\tau)$ and its complementary are r-convex. This method adapts the Devroye–Wise's estimator for the support to the context of level set estimation, see [2]. In this case, the estimator consists of the union of balls around those points in $\mathscr{X}_n^+$ that have a distance of at least r from each point in $\mathscr{X}_n \setminus \mathscr{X}_n^+$.

27.4.1 Simulation Results for Hybrids Methods

Granulometric smoothing method and r-convex hull method depend on an unknown parameter r. This is the main disadvantage of these algorithms. In this work, we have considered five values for the radius of balls, r: $r_1 = 0.01$, $r_2 = 0.05$, $r_3 = 0.1$, $r_4 = 0.2$, and $r_5 = 0.3$. We will denote the methods as follows: Convex hull method by CH, r-convex hull method by CH_r, and granulometric smoothing method with radius r by W_r.

Although these results are not shown here, we have studied the influence of the parameter r for r-convex hull method and granulometric smoothing method. In general, r-convex hull method is less sensitive to the selection to the parameter r. We have compared the three hybrids methods by fixing an intermediate value for r because it is unknown. We have considered $r = r_3$ and use Fig. 27.5 to show the

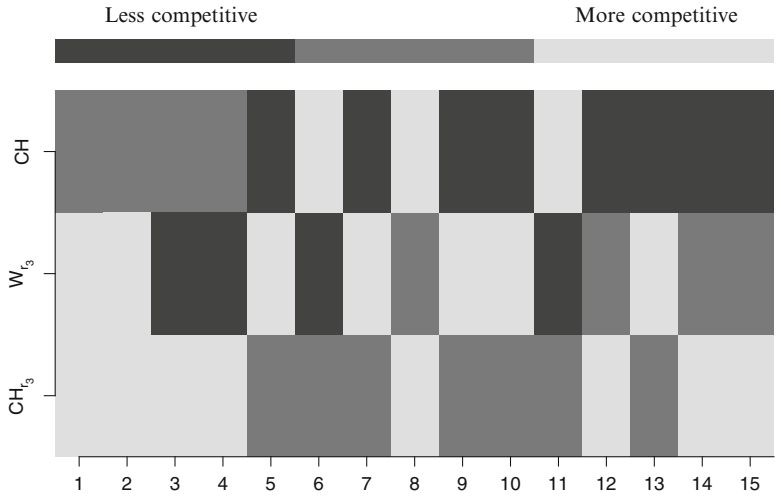

Fig. 27.5 Comparison of hybrid methods (*vertical axis*) with the 15 Marron and Wand's density models (*horizontal axis*), $\tau = 0.2$ and $n = 1,600$. The error criteria is d_{μ_f}

results obtained for $\tau = 0.2$. Each method is represented on the vertical axis and each density model on the horizontal axis.

Some of the density models present convex level sets for $\tau = 0.2$ or $\tau = 0.8$ although they are not unimodal (see, for example, densities 6, 8, or 11 in Fig. 27.5). In this case, when the convexity assumption is true, convex hull method can be very competitive. However, models 1, 2, 3, and 4 have convex level sets for some value of τ and r_3-convex hull method is the most competitive for them. In addition, sometimes convexity hypothesis can be very restrictive (see models 7 or 10, for example) and then, r_3-convex hull or granulometric smoothing methods provide better and similar results although the first one is most competitive for high values of τ.

27.5 Final Conclusions

Finally, we will compare the most competitive methods in each group. So, we will consider cross validation method, Müller and Sawitzki's method, granulometric smoothing method, r-convex hull method, and convex hull method. It is necessary to specify a value for the parameters M and r for Müller and Sawitzki's method and granulometric smoothing method or r-convex hull method. We have fixed $M = 3$ and $r = r_3$ again.

Figures 27.6 and 27.7 show the results for $\tau = 0.2$ and $\tau = 0.5$. Müller and Sawitzki's method with $M = 3$ is not very competitive because most of the models are not trimodal. For low values of τ, cross validation does not present bad results

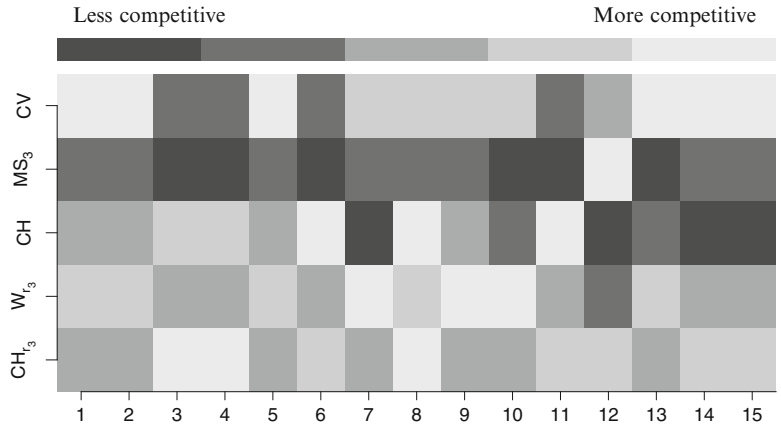

Fig. 27.6 Final comparison of the most competitive methods in each group (*vertical axis*) with the 15 Marron and Wand's density models (*horizontal axis*), $\tau = 0.2$ and $n = 1,600$. The error criteria is d_{μ_f}

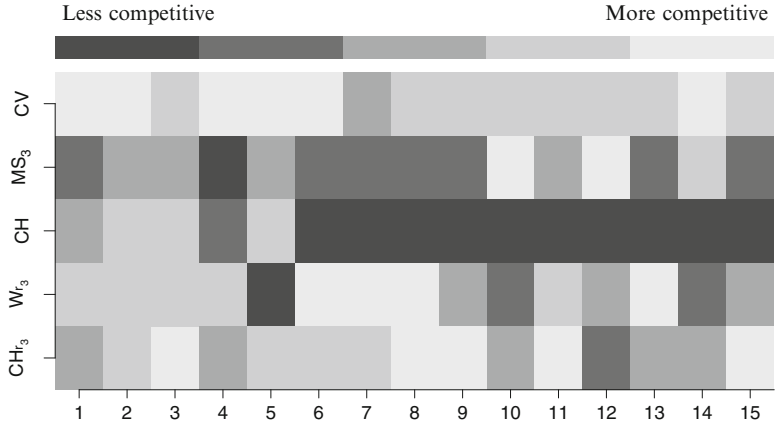

Fig. 27.7 Final comparison of the most competitive methods in each group (*vertical axis*) with the 15 Marron and Wand's density models (*horizontal axis*), $\tau = 0.5$ and $n = 1,600$. The error criteria is d_{μ_f}

but granulometric smoothing or r_3-convex hull methods have a better behavior (see models 3, 4, 6, or 11). But these two methods present a big disadvantage because both depend on an unknown parameter. Convex hull gets worse its results for $\tau = 0.5$ (see models 6, 8, or 11). The rest of the hybrid methods have good results for this value of τ.

In general, if no assumption is made on the shape of the level set, cross validation is a good option. But, if we have some information about the shape of the level set, then hybrid methods can be an alternative. For instance, if τ is small, then convex hull method could be very competitive. Most of these densities have convex level sets for this level. Under more flexible shape restrictions, r-convex hull or granulometric smoothing methods could be used but they depend on an unknown parameter. It would be useful to have a method for selecting it from the sample.

Acknowledgements This work has been supported by Project MTM2008-03010 from the Spanish Ministry of Science and Innovation and the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences) from Belgian Science Policy.

References

1. Baíllo, A., Cuevas, A.: Parametric versus nonparametric tolerance regions in detection problems. Comput. Stat. **21**, 527–536 (2006)
2. Devroye, L., Wise, G.L.: Detection of abnormal behavior via nonparametric estimation of the support. SIAM J. Appl. Math. **38**, 480–488 (1980)
3. Hyndman, R.J.: Computing and graphing highest density regions. Am. Stat. **50**, 120–126 (1996)

4. Korostelëv, A., Tsybakov, A.: Minimax Theory of Image Reconstruction. Lecture Notes in Statistics, vol. 82. Springer, New York (1993)
5. Marron, J., Wand, M. Exact mean integrated squared error. Ann. Stat. **20**, 712–736 (1992)
6. Müller, D.W., Sawitzki, G.: Excess mass estimates and tests of multimodality. J. Am. Stat. Assoc. **86**, 738–746 (1991)
7. Rodríguez-Casal, A.: Set estimation under convexity type assumptions. Annales de l'I.H.P.-Probabilités Statistiques **43**, 763–774 (2007)
8. Samworth, R.J., Wand , M.P.: Asymptotics and optimal bandwidth selection for highest density region estimation. Ann. Stat. **38**, 1767–1792 (2010)
9. Walther, G.: Granulometric smoothing. Ann. Stat. **25**, 2273–2299 (1997)