Statistica Sinica **28** (2018), 1481-1502
doi:https://doi.org/10.5705/ss.202016.0211

BASELINE ZONE ESTIMATION IN TWO DIMENSIONS WITH REPLICATED MEASUREMENTS UNDER A CONVEXITY CONSTRAINT

Atul Mallik[1], Moulinath Banerjee[2] and Michael Woodroofe[2]

[1] *Wells Fargo Securities and* [2] *University of Michigan*

Abstract: We consider the problem of estimating the region on which a non-parametric response function defined on the plane is at its baseline level in a sampling setting where multiple replicates of the response are available at each location. The baseline level typically corresponds to the minimum/maximum of the function and estimating such regions or their complements is pertinent to several problems arising in edge estimation, environmental statistics, fMRI and related fields. We assume the region of interest to be convex and estimate it via fitting a "stump" function to approximate p-values obtained from tests for deviation of the regression function from its baseline level. The shape of the baseline region and the smoothness of the regression function at its boundary play a critical role in determining the rate of convergence of our estimate: for a response function which is "p-regular" at the boundary of the convex baseline region, our estimate converges at a rate $N^{-2/(4p+3)}$, N being the total budget. This is expected to be optimal in light of existing work in related problems. We end with a discussion of various extensions of our approach, as well as connections to existing approaches.

Key words and phrases: Baseline zone estimation, convexity constraint, p-values, replicated measurements.

1. Introduction

Consider a data generating model of the form $Y = \mu(X) + \epsilon$, where μ is a function on $[0,1]^2$ such that

$$\mu(x) = \tau_0 \text{ for } x \in S_0, \text{ and } \mu(x) > \tau_0 \text{ for } x \notin S_0\,, \tag{1.1}$$

τ_0 is unknown, and the covariate X is independent of the error ϵ which has mean zero with finite positive variance σ_0^2. We are interested in estimating the baseline region S_0 beyond which the function deviates from τ_0. Examples of such problems abound. In several *fMRI* studies, one seeks to detect regions of brain activity from cross sectional two-dimensional images. Here, S_0 corresponds to the region of no-activity in the brain with S_0^c being the region of interest. In *LIDAR*

(light detection and ranging) experiments used for measuring concentration of pollutants in the atmosphere, interest often centers on finding high/low pollution zones (see, for example, Wakimoto and McElroy (1986)); in such contexts, S_0 would be the zone of maximal pollution, say in the vicinity of a polluting source. In *dose-response studies*, patients may be put on multiple (interacting) drugs (see, for example, Geppetti and Benemei (2009)), and it is of interest to find the dosage levels (∂S_0) at which the effect of the drugs starts kicking in.

In this paper, we address the problem of estimating $\mathcal{S}_0$ in a *replicated measurement* setting, sometimes called the 'dose–response' setting: n locations are chosen (randomly) from which to sample responses, and at each location m replicates are obtained. Thus, the total number of data-points is $N \equiv m \times n$, and we allow both m and n to grow. Our methodology extends a relatively recent idea from Mallik et al. (2011), developed in a simple 1-dimensional setting, to multiple dimensions. We construct p-value type statistics that detect the deviation of the function μ from its baseline value τ_0 at each covariate level and then fit an appropriate "stump" – a piecewise constant function with two levels – to these p-values. A key motivation for the dose-response setting comes from the minimum effective dose (MED) problems – a one-dimensional version of the problem considered here – where data are available from several patients (multiple replicates) at each dose level (covariate value) and one is interested in finding the lowest dose level where the effect of the concerned drug kicks in. The baseline set is, therefore, an interval $[0, d_0]$ for some unknown $d_0 > 0$. The extension of the dose-response setting to two dimensions as considered in this paper can be viewed as an idealized version of a scenario involving pharmacological studies where subjects are assigned a pair of interacting drugs, and multiple individuals are put at different combinations of drug–levels.

The question of detecting S_0 is also related to the *edge detection* problem which involves recovering the boundary of an image. In edge detection, μ corresponds to the image intensity function with S_0^c being the image and S_0 the background. A number of different algorithms in the computer science literature deal with this problem, though primarily in situations where μ has a jump discontinuity at the boundary of S_0; see Qiu (2007) for a review of edge detection techniques. With the exception of work done by Korostelëv and Tsybakov (1993), Mammen and Tsybakov (1995), and a few others, theoretical properties of such algorithms appear to have been rarely addressed. In fact, the study of theoretical properties of such estimates is typically intractable without some regularity assumption on S_0; for example, Mammen and Tsybakov (1995) discuss

minimax recovery of sets under smoothness assumption on the boundary.

In this paper, we approach the problem from the point of view of a shape-constraint (typically obtained from background knowledge) on the baseline region. We assume that the region S_0 is a closed convex subset of $[0,1]^2$ with a non-empty interior (and therefore, positive Lebesgue measure) and restrict ourselves to the more difficult problem where μ is continuous at the boundary. Convexity is a natural shape restriction to impose, not only because of analytical tractability, but also as convex boundaries arise naturally in several application areas: see, Wang et al. (2007), Ma et al. (2010), Stahl and Wang (2005), and Goldenshluger and Spokoiny (2006) for a few illustrative examples. In the statistics literature, Goldenshluger and Zeevi (2006) provide theoretical analyses of a convex boundary recovery method in a white noise framework.

Our problem also has connections to the level-sets estimation problem since S_0^c is the "level-set" $\{x : \mu(x) > \tau_0\}$ of the function μ. However, because τ_0 is at the *extremity* of the range of μ, the typical level-set estimate $\{x : \hat{\mu}(x) > \tau_0\}$, where $\hat{\mu}$ is an estimate of μ, does not distinguish well between the sets $\{x : \mu(x) \geq \tau_0\}$ and $\{x : \mu(x) > \tau_0\}$ unless μ has a jump at ∂S_0 and, indeed, need not be consistent for S_0^c . Moreover, this plug-in approach does not account for the pre-specified shape of the level-set. We note that the shape-constrained approach to estimate level-sets has also been studied in the literature, e.g., Nolan (1991) studied estimating ellipsoidal level-sets in context of densities, Hartigan (1987) provided an algorithm for estimating convex contours of a density, and Tsybakov (1997) and Cavalier (1997) studied "star-shaped" level-sets of density and regression functions, respectively. These approaches are based on an "excess mass" criterion (or its local version) that yield estimates with optimal convergence rates Tsybakov (1997). It will be seen later that our estimate also recovers the level-set of a transform of μ, but at a level in the *interior* of the range of the transform. More connections in this regard are explored in Section 4.

The smoothness of μ at its boundary plays a critical role in determining the rate of convergence of our estimate: for a regression function which is "p-regular" (formally defined in Section 3) at the boundary of the convex baseline region, our estimate converges at a rate $N^{-2/(4p+3)}$. This coincides with the rate obtained in related level-set estimation problems; see Polonik (1995, Thm. 3.7) and Tsybakov (1997, Thm. 2). It should be pointed out, here, that our convergence rates for estimation of the boundary in the regression context are quite different from the analogous problem of support boundary recovery based on i.i.d. obser-

vations from a multivariate density as studied, for example, in Härdle, Park and Tsybakov (1995), who obtain *faster* convergence rates due to the simpler nature of the problem: namely, there are *no realizations* from outside the support of the density.

In sum, we propose a computationally simple approach to estimate baseline sets in two dimensions in a replicated measurement setting and deduce consistency and rates of convergence of our estimate. Our approach falls at the interface of edge detection and level-set estimation problem as it detects the edge set (S_0^c) through a level-set estimate (see Section 4). While we primarily address the situation where the baseline set is convex, in the presence of efficient algorithms, our approach is extendible beyond convexity (see Section 4).

The rest of the paper is arranged as follows. We formally define our setting, describe the estimation procedure, and list our assumptions in Section 2. We justify consistency and deduce an upper bound on the rate of the convergence of our procedure in Section 3. We end with a discussion on various extensions and connections in Section 4.

2. Estimation Procedure

In this section, we extend a particular variant of the p-value procedure originally developed in a one-dimensional setting in Mallik et al. (2011). Consider a data generating model of the form

$$Y_{ij} = \mu(X_i) + \epsilon_{ij}, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n.$$

Here $m = m_n = m_0 n^\beta$ for some $\beta > 0$, with $N = m \times n$ being the total budget. The covariate X is sampled from a distribution F with Lebesgue density f on $[0,1]^2$ and ϵ is independent of X, has mean 0 and variance σ_0^2.

At each level $X_i = x$, we test the null hypothesis $H_{0,x} : \mu(x) = \tau_0$ against the alternative $H_{1,x} : \mu(x) > \tau_0$ and use the resulting (approximate) p-values to construct an estimate of the set S_0. The (un-normalized[1]) p-values are given by

$$p_{m,n}(x) = 1 - \Phi(\sqrt{m}(\bar{Y}_{i\cdot} - \hat{\tau})),$$

where $\bar{Y}_{i\cdot} = \sum_{j=1}^m Y_{ij}/m$ and $\hat{\tau}$ is some suitable estimate of μ (to be discussed later). These p-values asymptotically have mean $1/2$ for $x \in S_0$ and converge to zero when $x \notin S_0$. This simple observation can be used to construct estimates of S_0. We fit a stump to the observed p–values, with levels $1/2$ and 0 on either side of the boundary of a generic set and prescribe the set corresponding to the

[1]See Remark 1.

best fitting stump (in the sense of least squares) as an estimate of S_0. Formally, we fit a stump of the form $\xi_S(x) = (1/2)1(x \in S)$, minimizing

$$\sum_{i=1}^{n} \{p_{m,n}(X_i) - \xi_S(X_i)\}^2 = \sum_{i:X_i \in S} \left(p_{m,n}(X_i) - \frac{1}{2}\right)^2 + \sum_{i:X_i \in S^c} (p_{m,n}(X_i))^2$$

over appropriate choices of S. The above expression can be simplified and it can be seen that one can alternatively minimize

$$\mathbb{M}_n(S) = \mathbb{P}_n \left\{\Phi\left(\sqrt{m}(\bar{Y} - \hat{\tau})\right) - \gamma\right\} 1_S(X),$$

where $\mathbb{P}_n$ denotes the empirical measure on $\{\bar{Y}_{i\cdot}, X_i\}_{i \le n}$ and $\gamma = 3/4$.

Remark 1. Our procedure uses un-normalized p-values where the test statistic is not normalized by an estimate of the variance. The choice is merely based on notational convenience.The two versions of the procedure, the normalized and the un-normalized one, exhibit similar fundamental features such as the same dichotomous separation over S_0 and S_0^c, and identical rates of convergence.The un-normalized version is notationally more tractable and avoids a few routine justifications required for the normalized version.

The class of sets over which $\mathbb{M}_n$ is minimized should be chosen carefully as very large classes would give uninteresting discrete sets while small classes may not provide a reasonable estimate of S_0. As we assumed S_0 to be convex, we minimize $\mathbb{M}_n$ over $\mathcal{S}$, the class of *closed convex* subsets of $[0,1]^2$. Let $\hat{S}_n = \operatorname{argmin}_{S \in \mathcal{S}} \mathbb{M}_n(S)$. The estimate $\hat{S}_n$ can be computed by an adaptation of a density level-set estimation algorithm from Hartigan (1987) which we state below. If a closed convex set $S^\star$ minimizes $\mathbb{M}_n$, the convex hull of $\{X_i : X_i \in S^\star, 1 \le i \le n\}$ also minimizes $\mathbb{M}_n$. Hence, it suffices to reduce our search to convex polygons whose vertices could only be X_i's. There could be 2^n such polygons, still a computationally expensive collection to search over.

Computing the estimate. We first find the optimal polygon (the convex polygon which minimizes $\mathbb{M}_n$) for each choice of X as its leftmost vertex. We use the following notation. Let this particular X be numbered 1, and let the X_i's not to its left be numbered $2, 3, \ldots, r$. The coordinates of the point i are denoted by z_i, and the line segment $az_i + (1-a)z_j$, $(0 \le a \le 1)$ is written as $[i, j]$. Assume that $1, \ldots, r$ are ordered so that the segments $[1, i]$ move counterclockwise as i increases and so that $i \le j$ if $i \in [1, j]$. Polygons will be built up from triangles for $1 < i < j \le r$; Δ_{ij} is the convex hull of $(1, i, j)$ excluding $[1, i]$. The segment $[1, i]$ is excluded from Δ_{ij} in order to combine triangles without overlap. The quadrilateral with vertices at $1, i, j, k$ for $i < j < k \le r$ is convex if

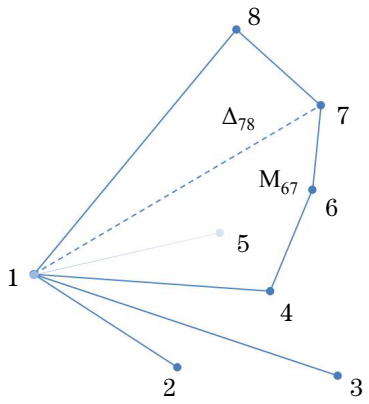

Figure 1. Notation for constructing the convex set estimate. An arbitrary vertex is numbered 1, and those not to its left are numbered $2, 3, \ldots, 8$ in a counterclockwise manner. The triangle Δ_{78} excludes the line segment $[1, 7]$. The optimal polygon (with measure $\mathbb{M}_{67}$) with successive vertices 6, 7 and 1 is depicted as the convex polygon with vertices 1, 4, 6 and 7.

$$D_{ijk} = \begin{vmatrix} z'_i & 1 \\ z'_j & 1 \\ z'_k & 1 \end{vmatrix} \geq 0.$$

Let $\mathbb{M}_{1j}$ be the value of $\mathbb{M}_n$ on the line segment $[1, j]$. Further, for $1 < j < k \leq r$, let $\mathbb{M}_{jk}$ denote the minimum value of $\mathbb{M}_n$ among closed convex polygons with successive counterclockwise vertices j, k and 1. All such convex polygons contain the triangle Δ_{jk} and hence, $\mathbb{M}_n(\Delta_{jk})$, $\mathbb{M}_n$ measure of Δ_{jk}, is a common contributing term to the $\mathbb{M}_n$ measure of all such polygons. This simple fact forms the basis of the algorithm. It can be shown that

$$\mathbb{M}_{jk} = \mathbb{M}_{i^* j} + \mathbb{M}_n(\Delta_{jk}), \tag{2.1}$$

where $i^* = I(k, j)$ is chosen to minimize $\mathbb{M}_{ij}$ over vertices i with $i < j$, $D_{ijk} \geq 0$, i.e,

$$i^* = I(k, j) = \underset{i: i<j, D_{ijk}>0}{\operatorname{argmin}} \mathbb{M}_{ij}. \tag{2.2}$$

Here i^* could possibly be 1, in which case $\mathbb{M}_{jk}$ is simply the $\mathbb{M}_n$ measure of the triangle formed by j, k and 1 (including the contribution of line segment $[1, j]$).

One way to construct an optimal polygon with leftmost vertex 1 is to find the minimum among $\mathbb{M}_{jk}$, $1 \leq j < k$, where $\mathbb{M}_{jk}$'s are computed recursively using (2.1) and (2.2). Hence, one optimal polygon with leftmost vertex 1 has vertices $i_l, i_2, \ldots, i_s = 1$, where either $s = 1$ or $\mathbb{M}_{i_2 i_1} = \min_{1 \leq j < k} \mathbb{M}_{jk}$, $i_3 = I(i_1, i_2)$,

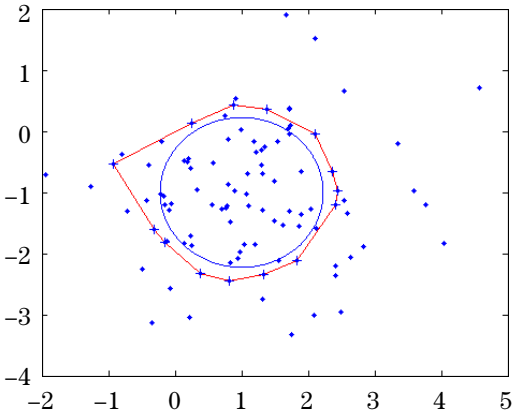

Figure 2. An illustration of the procedure in the dose-response setting with $m = 10$ and $n = 100$. The set S_0 is a circle centered at $(1, -1)$ with radius 1.

$i_4 = I(i_2, i_3)$, ..., $1 = i_s = I(i_{s-2}, i_{s-1})$. Once this is done for each choice of X as the leftmost vertex, the final estimate $\hat{S}_n$ is simply the one with the minimum $\mathbb{M}_n$ value among these n constructed polygons.

Figures 1 and 2 illustrate some aspects of this discussion. There are minor modifications that can be made to the above algorithm so that the over-all implementation involves $O(n^3)$ computations; see Hartigan (1987, Sec. 3) for more details.

2.1. Notations and assumptions

We adhere to the setup above. Let λ denote the Lebesgue measure. The precision of the estimates is measured using the metric

$$d_F(S_1, S_2) = F(S_1 \Delta S_2)\,.$$

For simplicity, we assume τ_0 to be known. It can be shown that our results extend to case where we impute a $\sqrt{mn}$ (dose-response) estimate of τ (more on this in Section 3.1).

Let ρ denote the l_∞ metric on $\mathbb{R}^2$ (which is equivalent to the Euclidean metric but makes some of the subsequent analyses simpler) and for a point $x \in \mathbb{R}^2$ and a set $A \subset \mathbb{R}^2$, let $\rho(x, A) := \inf_{y \in A} \rho(x, y)$.

We make the following assumptions.

1. The function μ is continuous on $[0, 1]^2$.
2. The function μ is p-regular at ∂S_0: for some $\kappa_0, C_0, C_1 > 0$ and for all $x \notin S_0$ such that $\rho(x, S_0) < \kappa_0$,

$$C_0\rho(x,S_0)^p \le \mu(x)-\tau_0 < C_1\rho(x,S_0)^p. \tag{2.3}$$

3. $S_0=\mu^{-1}(\tau_0)$ is convex. For some $\epsilon_0>0$, , $S_0\subset[\epsilon_0,1-\epsilon_0]^2$ and $\lambda(S_0)>0$.
4. The design density f for the dose-response setting is continuous and positive on $[0,1]^2$.

Remark 2. By uniform continuity of μ and compactness of $[0,1]^2$, $\inf\{\mu(x) : \rho(x,S_0)\ge\kappa_0\} > \tau_0$. For a fixed p, τ_0, κ_0, $\delta_0>0$, we **denote** the class of functions μ satisfying Assumptions 1, 2, 3 and

$$\inf\{\mu(x):\rho(x,S_0)\ge\kappa_0\}-\tau_0>\delta_0 \tag{2.4}$$

by $\mathcal{F}_p=\mathcal{F}_p(p,\tau_0,\kappa_0,\delta_0)$.

3. Consistency and Rate of Convergence

As τ_0 is known, we can take $\tau_0=0$ without loss of generality. With $\mathbb{M}_n(S)=\mathbb{P}_n\left\{\Phi\left(\sqrt{m}\bar{Y}\right)-\gamma\right\}1_S(X)$, let P_m denote the measure induced by $(\bar{Y},X)$ and

$$M_m(S)=P_m\left[\left\{\Phi\left(\sqrt{m}\bar{Y}\right)-\gamma\right\}1_S(X)\right].$$

The process M_m acts as a population criterion function and can be simplified as follows. Let

$$Z_{1m}=\frac{1}{\sqrt{m}\sigma_0}\sum_{j=1}^{m}\epsilon_{1j} \tag{3.1}$$

and Z_0 be a standard normal random variable independent of Z_{1m}s. Then

$$\begin{aligned}
E\left\{\Phi\left(\sqrt{m}\bar{Y}_1\right)\middle|\, X_1=x\right\} &= E\left[\Phi\left\{\sqrt{m}\mu(x)+\sigma_0 Z_{1m}\right\}\right]\\
&= E\left(E\left[1\left\{Z_0<\sqrt{m}\mu(x)+\sigma_0 Z_{1m}\right\}\middle|\, Z_{1m}\right]\right)\\
&= P\left\{\frac{Z_0-\sigma_0 Z_{1m}}{\sqrt{1+\sigma_0^2}}<\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right\}=\Phi_m\left\{\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right\},
\end{aligned}$$

where Φ_m denotes the distribution function of $(Z_0-\sigma_0 Z_{1m})/\sqrt{1+\sigma_0^2}$. By Pólya's theorem, Φ_m converges uniformly to Φ as $m\to\infty$. Hence, it can be seen that

$$\lim_{m\to\infty} E\left\{\Phi\left(\sqrt{m}\bar{Y}_1\right)\middle|\, X_1=x\right\}=\frac{1}{2}1_{S_0}(x)+1_{S_0^c}(x).$$

By the Dominated Convergence Theorem, $M_m(S)$ converges to $M(S)$, where

$$\begin{aligned}
M(S)=M_F(S) &= \int_S\left(\frac{1}{2}1_{S_0}(x)+1_{S_0^c}(x)-\gamma\right)F(dx)\\
&= \left(\frac{1}{2}-\gamma\right)F(S_0\cap S)+(1-\gamma)F(S_0^c\cap S).
\end{aligned} \tag{3.2}$$

Note that S_0 minimizes the limiting criterion function $M(S)$. An application of the argmin continuous mapping theorem van der Vaart and Wellner (1996, Theorem 3.2.2) yields a result on the consistency of $\hat{S}_n$.

Theorem 1. *Assume S_0 to be a closed convex set and the unique minimizer of $M(S)$. Then $\sup_{S\in\mathcal{S}} |\mathbb{M}_n(S) - M(S)|$ and $d_F(\hat{S}_n, S_0)$ converge in probability to zero for any $\gamma \in (0.5, 1)$.*

Remark 3. The proof of this theorem appears in Section A.1 of the Appendix, where we actually establish a stronger result: consistency is established in terms of the Hausdorff metric which implies consistency with respect to d_F. Moreover, we do not require m to grow as $m_0 n^\beta$, $\beta > 0$, for consistency. The condition $\min(m, n) \to \infty$ suffices. The result extends to higher dimensions: when μ is a function from $[0,1]^d \mapsto \mathbb{R}$ and $S_0 = \mu^{-1}(0)$ is a closed convex subset of $[0,1]^d$, the corresponding estimate is consistent.

We now proceed to deduce the rate of convergence of $d_F(\hat{S}_n, S_0)$. For this, we study how small the difference $(\mathbb{M}_n - M)$ is and how M behaves in the vicinity of S_0. We split the difference $(\mathbb{M}_n - M)$ into $(\mathbb{M}_n - M_m)$ and $(M_m - M)$ and study them separately. The term $\mathbb{M}_n - M_m$ involves an empirical average of centered random variables, efficient bounds on which are derived using empirical process inequalities. We first establish a bound on the non-random term $(M_m - M)$ in the vicinity of S_0.

Lemma 1. *For any $\delta > 0$, $a_n \downarrow 0$ and $F(S \triangle S_0) < \delta$, we have*

$$|(M_m - M)(S) - (M_m - M)(S_0)| \le \left|\Phi_m(0) - \frac{1}{2}\right| \delta + \min(c_0 a_n, \delta)$$
$$+ \left|\Phi_m\left(\frac{C_0\sqrt{m}a_n^p}{\sqrt{1+\sigma_0^2}}\right) - 1\right| \delta + \left|\Phi_m\left(\frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}}\right) - 1\right| \delta.$$

Here, $c_0 > 0$ is some constant.

Proof. Note that

$$M_m(S) - M_m(S_0) = P_m\left[\left\{\Phi_m\left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right) - \gamma\right\}\{1_S(x) - 1_{S_0}(x)\}\right] \text{ and}$$
$$M(S) - M(S_0) = \int \left\{\left(\frac{1}{2}\right) 1_{S_0}(x) + 1_{S_0^c}(x) - \gamma\right\}\{1_S(x) - 1_{S_0}(x)\} F(dx).$$

Hence, the expression $|(M_m - M)(S) - (M_m - M)(S_0)|$ is bounded by

$$\int_{x\in(S_0\cap S)} \left|\Phi_m(0) - \frac{1}{2}\right| F(dx) + \int_{x\in(S_0^c\cap S)} \left|\Phi_m\left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right) - 1\right| F(dx). \quad (3.3)$$

The first term is bounded by $|\Phi_m(0) - 1/2|\delta$. Let $S_n = \{x : \rho(x, S_0) > a_n\}$. For a set S with a rectifiable boundary, let $P(S)$ denote its perimeter, $P(S) = \lim_{\eta \to 0} \lambda(S^\eta \backslash S)/\eta$. Then $\lambda(S_n^c \backslash S_0) \leq (2P(S_0))a_n$ for large n. Here, $P(S_0)$ is finite Eggleston (1958, pp. 82–89) and is uniformly bounded for $S_0 \in \mathcal{S}$. Using Assumption 4, $F(S_n^c \backslash S_0) \leq c_0 a_n$ for some $c_0 > 0$. Hence, the second sum in (3.3) is bounded by

$$F(S_n^c \backslash S_0) + \int_{x \in (S_n \cap S)} \left| \Phi_m \left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}} \right) - 1 \right| F(dx) \leq \min(c_0 a_n, \delta)$$
$$+ \int_{x \in (S_n \cap S)} \left\{ \left| \Phi_m \left(\frac{C_0 \sqrt{m} a_n^p}{\sqrt{1+\sigma_0^2}} \right) - 1 \right| + \left| \Phi_m \left(\frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}} \right) - 1 \right| \right\} F(dx).$$

As $F(S_n \cap S) < \delta$, we get the result.

To control $\mathbb{M}_n - M_m$, we rely on a version of Theorem 5.11 of van de Geer (2000). The result in its original form is slightly general. In their notation, it involves a bound on a special metric $\rho_K(\cdot)$ (see van de Geer (2000, Eq. (5.23))) which, in light of Lemma 5.8 of van de Geer (2000), can be controlled by bounding the L_2-norm in the case of bounded random variables. This yields the consequence stated below. Here, H_B denotes the entropy with respect to bracketing numbers.

Theorem 2. *Let $\mathcal{G}$ be a class of functions such that $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq 1$. For some universal constant $C > 0$, if C_2, C_3, R and $\underline{N} > 0$ satisfy:*

$$R \geq \sup_{g \in \mathcal{G}} \|g\|_{L_2(P)},$$
$$N \geq C_2 \int_0^R H_B^{1/2}(u, \mathcal{G}, P) du \vee R,$$
$$C_2^2 \geq C^2(C_3 + 1) \text{ and}$$
$$\underline{N} \leq C_3 \sqrt{n} R^2,$$
$$P^* \left\{ \sup_{g \in \mathcal{G}} |\mathbb{G}_n(g)| > \underline{N} \right\} \leq C \exp \left\{ \frac{-\underline{N}^2}{C^2(C_3+1)R^2} \right\}.$$

Here, P^* denotes the outer probability.

Proposition 1. *When $\beta > 0$,*

$$P^* \{ d_F(\hat{S}_n, S_0) > \delta_n \} \to 0$$

for $\delta_n = K_1 \max\{n^{-2/3}, m^{-1/(2p)}\}$, where $K_1 > 0$ is some constant.

Proof. Let k_n be the smallest integer such that $2^{k_n+1}\delta_n \geq 1$. For $0 \leq k \leq k_n$, let $\mathcal{S}_{n,k} = \{S : S \in \mathcal{S},\ 2^k \delta_n \leq d_F(S, S_0) < 2^{k+1}\delta_n\}$. As, $\hat{S}_n$ is the minimizer for

$\mathbb{M}_n$, we have

$$P^*\left\{d_F(\hat{S}_n, S_0) > \delta_n\right\} \leq \sum_{k=0}^{k_n} P^*\left(\inf_{S\in\mathcal{S}_{n,k}} \mathbb{M}_n(S) - \mathbb{M}_n(S_0) \leq 0\right).$$

The sum on the right side can be bounded by:

$$\sum_{k=0}^{k_n} P^*\left\{\sup_{S\in\mathcal{S}_{n,k}} |(\mathbb{M}_n - M)(S) - (\mathbb{M}_n - M)(S_0)| > \inf_{S\in\mathcal{S}_{n,k}} (M(S) - M(S_0))\right\}. \quad (3.4)$$

For $c(\gamma) = \min(\gamma - 1/2, 1 - \gamma)$,

$$M(S) - M(S_0) = \left(\gamma - \frac{1}{2}\right)\{F(S_0) - F(S_0 \cap S)\} + (1-\gamma)F(S_0^c \cap S) \geq c(\gamma)F(S \triangle S_0),$$

and hence, (3.4) is bounded by

$$\sum_{k=0}^{k_n} P^*\left\{\sup_{S\in\mathcal{S}_{n,k}} |(\mathbb{M}_n - M_m)(S) - (\mathbb{M}_n - M_m)(S_0)| > c(\gamma)2^{k-1}\delta_n\right\}$$
$$+ \sum_{k=0}^{k_n} P^*\left\{\sup_{S\in\mathcal{S}_{n,k}} |(M_m - M)(S) - (M_m - M)(S_0)| \geq c(\gamma)2^{k-1}\delta_n\right\}. \quad (3.5)$$

Note that $M_m - M$ is a non-random process and hence, each term in the second sum is either 0 or 1. We now show that the second sum in this display is eventually zero. For this, we apply Lemma 1. We have

$$\begin{aligned}
&\sup_{A\in\mathcal{S}_{n,k}} |(M_m - M)(S) - (M_m - M)(S_0)| \\
&\leq \left|\Phi_m(0) - \frac{1}{2}\right| 2^{k+1}\delta_n + \min(c_0 a_n, 2^{k+1}\delta_n) \\
&\quad + \left|\Phi_m\left(\frac{C_0\sqrt{m}a_n^p}{\sqrt{1+\sigma_0^2}}\right) - 1\right| 2^{k+1}\delta_n + \left|\Phi_m\left(\frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}}\right) - 1\right| 2^{k+1}\delta_n \\
&\leq 4\left\{\left|\Phi_m(0) - \frac{1}{2}\right| + \left|\Phi_m\left(\frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}}\right) - 1\right|\right\} 2^{k-1}\delta_n \\
&\quad + \left\{\frac{2c_0 a_n}{\delta_n} + 4\left|\Phi_m\left(\frac{C_0\sqrt{m}a_n^p}{\sqrt{1+\sigma_0^2}}\right) - 1\right|\right\} 2^{k-1}\delta_n. \quad (3.6)
\end{aligned}$$

Hence, it suffices to show that the coefficient of $2^{k-1}\delta_n$ in (3.6) is smaller than $c(\gamma)$. To this end, fix $0 < \eta < c(\gamma)/8$. For large m,

$$\left|\Phi_m(0) - \frac{1}{2}\right| + \left|\Phi_m\left(\frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}}\right) - 1\right| \leq \eta.$$

Choose c_η such that $a_n = c_\eta m^{-1/(2p)} > \{\Phi_m^{-1}(1-\eta)\sqrt{1+\sigma_0^2}/(C_0\sqrt{m})\}^{1/p}$. For

large n, the coefficient of $2^{k-1}\delta_n$ in (3.6) is bounded by

$$8\eta + \frac{c_0 c_\eta}{K_1} < c(\gamma),$$

for $K_1 > (c_0 c_\eta)/(c(\gamma) - 8\eta)$. Hence, each term in the second sum of (3.5) is zero for a suitably large choice of the constant K_1. The first term can be written as

$$\sum_{k=0}^{k_n} P^* \left\{ \sup_{S \in \mathcal{S}_{n,k}} \left| \mathbb{G}_n g_m(\bar{Y}) 1_{S \triangle S_0}(X) \right| > c(\gamma) 2^{k-1} \delta_n \sqrt{n} \right\}, \tag{3.7}$$

where $g_m(y) = \Phi(\sqrt{m}y) - \gamma$. We are now in a position to apply Theorem 2 to each term of (3.7). In the setup of Theorem 2, $\underline{N} = c(\gamma) 2^{k-1} \delta_n \sqrt{n}$. The concerned class of functions is $\mathcal{G}_{n,k} = \{g_m(\bar{Y})(1_S(X) - 1_{S_0}(X)) : S \in \mathcal{S}_{n,k}\}$. Note that $\|g_m(1_S - 1_{S_0})\|_{L_2(P_m)} \le [E 1_{S \triangle S_0}(X)]^{1/2} \le (2^{k+1}\delta_n)^{1/2}$, so we can pick $R = R_{n,k} = (2^{k+1}\delta_n)^{1/2}$. Let $\mathcal{C}_{n,k} := \{1_S - 1_{S_0} : S \in \mathcal{S}_{n,k}\}$. As $\mathcal{S}_{n,k} \subset \mathcal{S}$, $N_{[\,]}(u, \mathcal{C}_{n,k}, L_2(P)) \le N_{[\,]}(u/2, \mathcal{S}, L_2(P))$ for any $u > 0$, by a simple calculation. Also, starting with a bracket $[f_L, f_U]$ containing a generic function in $\mathcal{C}_{n,k}$ and $\|f_U - f_L\|_{L_2(P)} \le u$, we can obtain a bracket for $g_m(\overline{Y})(1_S(X) - 1_{S_0}(X))$ using the inequality

$$\Phi\left(\sqrt{m}y\right) f_L - \gamma f_U \le g_m(y) 1_B(x) \le \Phi\left(\sqrt{m}y\right) f_U - \gamma f_L.$$

As $\|\Phi(\sqrt{m}y) + \gamma\|_\infty \le 2$,

$$\|(\Phi\left(\sqrt{m}y\right) f_U - \gamma f_L) - (\Phi\left(\sqrt{m}y\right) f_L - \gamma f_U)\|_{L_2(P)} \le 2u.$$

It follows that

$$H_B\{u, \mathcal{G}_{n,k}, L_2(P)\} \le H_B\left\{\frac{u}{2}, \mathcal{C}_{n,k}, L_2(P)\right\} \le H_B\left\{\frac{u}{4}, \mathcal{S}, L_2(P)\right\}.$$

Using the fact that $H_B(u, \mathcal{S}, L_2(P)) = \log(N_{[\,]}(u, \mathcal{S}, L_2(P))) \le A_0 u^{-(d-1)}$ for $d \ge 2$ (see Bronšteĭn (1976)), we get

$$H_B\{u, \mathcal{G}_{n,k}, L_2(P)\} \le 4A_0 u^{-1}$$

for some constant $A_0 > 0$ (depending only on the design distribution). Renaming $4A_0$ as A_0, the conditions of Theorem 2 then translate to

$$\begin{aligned} 2^{k-1} c(\gamma) \delta_n \sqrt{n} &\ge 2C_2 \max(A_0, 1)(2^{k+1}\delta_n)^{1/4}, \\ C_2^2 &\ge C^2(C_3 + 1) \text{ and} \\ c(\gamma) 2^{k-1} \delta_n \sqrt{n} &\le C_3 \sqrt{n} 2^{k+1} \delta_n. \end{aligned}$$

It can be seen that for $K_1 \ge 2^9 (C_2 \max(A_1, 1)/c(\gamma))^{4/3}$, $C_3 = c(\gamma)/4$ and $C_2 = \sqrt{5}C/2$, these conditions are satisfied, and hence, we can bound (3.7) by

$$\sum_{k=0}^{k_n} C \exp\left\{\frac{-2^{k-3}c^2(\gamma)\delta_n n}{C^2(C_3+1)}\right\}.$$

As $\delta_n \gtrsim n^{-2/3}$ (the symbol $\gtrsim$ is used to denote the corresponding $\geq$ inequality holding up to some finite positive constant), the term $\delta_n n$ diverges to ∞ as $n \to \infty$. Hence, the above display converges to zero. This completes the proof.

Remark 4. This result also holds for values of δ_n larger than the one prescribed. Hence, our result delivers consistency though it requires m to grow as $m_0 n^{\beta}$. In terms of the total budget, choosing $\beta = 4p/3$ corresponds to the optimal rate. In this case, δ_n is of the order $n^{-2/3}$ or $N^{-2/(4p+3)}$. This particular rate also appears in related problems involving the estimation of convex level sets for a density function as studied in Polonik (1995) and Tsybakov (1997). Polonik (1995, Thm. 3.7) considers the estimation of density contour clusters using an empirical mass approach under a metric entropy condition on the class of clusters and derives rates of convergence of the estimates in terms of parameters (r, γ), where r describes how quickly the entropy grows in terms of the radius and γ controls the F measure of the set of points at which the density is close to the level of interest. For convex density contour clusters in $\mathbb{R}^2$, the example following Theorem 3.7 shows that for regular situations with $\gamma = 1$, the excess mass estimator is $N^{-2/7}$ consistent for the true level set in d_F distance, which coincides with the rate obtained by our estimator for 1-regular μ functions (i.e $p = 1$). Tsybakov (1997, Thm. 2) demonstrates that his level set estimator of convex level sets exhibits an optimal rate of $N^{-2/(4\alpha+3)}$ for α-regular densities around the level set λ (as defined around (4) of that paper): the α parameter of that paper corresponds to the p parameter in our assumptions. Keep in mind that Polonik (1995) and Tsybakov (1997)] deal with non-replicated settings and use notation n for the total number of observations rather than N.

The bounds deduced for the two sums in (3.5) depend on μ only through p and δ_0, e.g., the exponential bounds from Theorem 2 depend on the class of functions only through their entropy and norm of the envelope which do not change with μ. Hence, we have a result that is similar in flavor to the upper bounds deduced for level-set estimates in Tsybakov (1997).

Corollary 1. *For the choice of δ_n given in Proposition 1, we have*

$$\limsup_{n\to\infty} \sup_{\mu\in\mathcal{F}_p} E_\mu^* \left\{\delta_n^{-1} d(\hat{S}_n, S_0)\right\} < \infty. \tag{3.8}$$

Here, E_μ is the expectation with respect to the model with a particular $\mu \in \mathcal{F}_p$.

The other features of the model such as error distribution and the design distribution do not change.

Proof. Note that

$$E^*_\mu\{\delta_n^{-1} d(\hat{S}_n, S_0)\} \le 1 + \sum_{k\ge 0, 2^k\delta_n \le 1} 2^{k+1} P^*\left\{2^k < \delta_n^{-1} d(\hat{S}_n, S_0) \le s^{k+1}\right\}$$
$$\le 1 + \sum_{k\ge 0, 2^k\delta_n \le 1} 2^k P^*\left\{\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A) - \mathbb{M}_n(S_0) \le 0\right\}.$$

The probabilities $P^*\left(\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A) - \mathbb{M}_n(S_0) \le 0\right)$ can be bounded in an identical manner to that in the proof of the above Proposition and hence, we get

$$\sup_{\mu\in\mathcal{F}_p} E^*_\mu\{\delta_n^{-1} d(\hat{S}_n, S_0)\} \le 1 + \sum_{k=0}^{k_n} C 2^{k+1} \exp\left\{\frac{-2^{k-3} c^2(\gamma)\delta_n n}{C^2(C_3+1)}\right\}.$$

As $\delta_n n \to \infty$, we get the result.

3.1. Extension to the case of an unknown τ_0

While we deduced our results under the assumption of a known τ_0, in applications τ_0 is generally not known. In this situation, quite a few extensions are possible. If S_0 can be safely assumed to contain a positive F-measure set U, then a simple averaging of the $\bar{Y}$ values realized for X's in U would yield a $\sqrt{mn}$-consistent estimator of τ_0. If a proper choice of U is not available, one can obtain an initial estimate of τ_0 in the dose–response setting as

$$\hat{\tau}_{init} = \operatorname*{argmin}_{\tau\in\mathbb{R}} \mathbb{P}_n \left\{\Phi\left(\sqrt{m}(\bar{Y} - \tau)\right) - \frac{1}{2}\right\}^2.$$

This provides a consistent estimate of τ_0 under mild assumptions. A $\sqrt{mn}$-consistent estimate of τ_0 can then be found by using $\hat{\tau}_{init}$ to compute $\hat{S}_n$ and then averaging the $\bar{Y}$ value for the X's realized in ${}_r\hat{S}_n$ for a fixed $r \in (0,1)$. It can be shown that the rate of convergence remains unchanged if one imputes a $\sqrt{mn}$-consistent estimate of τ_0. A sketch of the next result is given in Section A.2.

Proposition 2. *Let $\hat{S}_n$ denote the smallest minimizer of*

$$\mathbb{M}_n(S, \hat{\tau}) = \mathbb{P}_n\left(\left[\Phi\left\{\sqrt{m}(\bar{Y} - \hat{\tau})\right\} - \gamma\right] 1_S(X)\right),$$

where $\sqrt{mn}\,(\hat{\tau} - \tau_0) = O_p(1)$. For $m = m_0 n^\beta$ and δ_n as defined in Proposition 1, we have $P\{d(\hat{S}_n, S_n) > \delta_n\} \to 0$.

4. Discussion

Extensions to non-convex baseline sets. We have addressed the situation

where the baseline set is convex for dimension $d = 2$, but our approach can be extended past convexity and beyond the two-dimensional setting of this paper in the presence of an efficient algorithm and for suitable collections of sets.

For example, let $\tilde{\mathcal{S}}$ be a collection of subsets of $[0,1]^d$ sets and let

$$\tilde{S}_n = \operatorname*{argmin}_{S \in \tilde{\mathcal{S}}} \mathbb{M}_n(S)\,.$$

Here, μ is a real-valued function from $[0,1]^d$ and $S_0 = \mu^{-1}(\tau_0)$ is assumed to belong to the class $\tilde{\mathcal{S}}$. Then, the estimator $\tilde{S}_n$ has the following properties.

Proposition 3. *Assume that S_0 is the unique minimizer (up to F-null sets) of the population criterion function M_F defined in* (3.2)*. Then $d_F(\tilde{S}_n, S_0)$ converges in probability to zero. Moreover, if*

$$H_B(u, \tilde{\mathcal{S}}, P) \lesssim u^{-r} \text{ for some } r < 2\,,$$

$P[d_F(\tilde{S}_n, S_0) > \tilde{\delta}_n]$ converges to zero where $\tilde{\delta}_n = K_1 \max(n^{-2/(2+r)}, m^{-1/(2p)})$ for some $K_1 > 0$.

Remark 5. The proof follows along the lines of Proposition 1. The dependence of the rate on the dimension arises through r which usually grows with d. Some algebra shows that in terms of N, the optimal rate can be written as $N^{-2/(4p+2+r)}$. When $d = 2$, $r = 1$ for the class of convex sets, and the optimal rate is $N^{-2/(4p+3)}$. The assumption on the bracketing entropy in this proposition is essentially a requirement that the class of sets under consideration be a Donsker class. Thus, our rate of convergence result is valid for any collection of such Donsker-type sets in any (fixed) dimension d. From the methodological point of view, the minimizer $\tilde{S}_n$ over $\tilde{\mathcal{S}}$ needs to be computable through an algorithm of reasonable complexity. However, as discussed below, our proposal has close connections to level-set estimation approaches. Since algorithmic approaches to level sets estimation are well-developed in general settings, our p-value based approach is expected to be applicable more broadly. However, level sets estimation techniques typically work better in low dimensions.

Connection with level-set approaches. Minimizing $\mathbb{M}_n(S)$ in the dose-response setting is equivalent to minimizing

$$\begin{aligned}\tilde{\mathbb{M}}_n(S) &= \mathbb{M}_n(S) - \frac{1}{2}\sum_{i=1}^{N}\left\{\frac{1}{4} - p_{m,n}(X_i)\right\} \\ &= \sum_{i=1}^{n} \frac{1/4 - p_{m,n}(X_i)}{2}\left\{1(X_i \in S) - 1(X_i \in S^c)\right\}.\end{aligned}$$

This form is quite similar to an empirical risk criterion function appearing in Willett and Nowak (2007, Eq. (7)) in the context of a level-set estimation procedure. It can be deduced that our baseline detection approach ends up estimating the level set $S_m = \{x : E[p_{m,n}(x)] > 1/4\}$ from i.i.d. data $\{p_{m,n}(X_i), X_i\}_{i=1}^n$ with $0 \le p_{m,n}(X_i) \le 1$. As $m \to \infty$, S_m's decrease to S_0, which is the target set. Hence, in principle, any level-set approach could be applied to the transformed data $\{p_{m,n}(X_i), X_i\}_{i=1}^n$ to yield an estimate for S_m.

In Scott and Davenport (2007), the approach to the level set estimation problem, with the criterion of minimizing the risk criterion in Willett and Nowak (2007), is shown to be equivalent to a *cost-sensitive classification* problem. This problem involves random variables $(X, Y, C) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R}$, where X is a feature, Y a class and C is the cost for misclassifying X when the true label is Y. Cost sensitive classification seeks to minimize the expected cost

$$R(G) = E\{C\ 1(G(X) \neq Y)\}, \tag{4.1}$$

where G, with a little abuse of notation, refers both to a subset of $\mathbb{R}^d$ and $G(x) = 1(x \in G)$. With $C = |\gamma - Y|$ and $\tilde{Y} = 1(Y \ge \gamma)$, the objective of the cost-sensitive classification, based on $(X, \tilde{Y}, C)$, can be shown to be equivalent to minimizing the excess risk criterion in Willett and Nowak (2007). So, approaches like support vector machines (SVM) and k-nearest neighbors (k-NN), which can be tailored to solve the cost-sensitive classification problem (see Scott and Davenport (2007)), are relevant to estimating level sets, and thus provide alternative ways to solve the baseline set detection problem. Since the *loss function* in (4.1) is not smooth, one might prefer to work with its surrogates. Some results in this direction can be found in Scott (2011).

Our developed approach can be also written as a maximization problem: we seek to find the maximizer of:

$$\mathbb{M}_n^{\#}(S) := \sum_{i=1}^{n} \{p_{m,n}(X_i) - \frac{1}{4}\}\, 1(X_i \in S)\,,$$

over all convex S in $[0, 1]^2$ as an estimate of S_m. This is essentially an empirical generalized λ-cluster (with $\lambda = 1/4$) as considered in Polonik and Wang (2005, Definition (2.3)) who seek to estimate level sets in a standard regression function setting (i.e. unreplicated data, or one observation per covariate) via an excess–mass approach. However, apart from the difference of our setting from Polonik and Wang (2005), our paper also differs in that ours derives an explicit convergence rate of the estimate under the restriction of convexity, whereas theirs establishes a weaker result, consistency, but under more general conditions on

the class of sets.

Adaptivity. We have assumed knowledge of the order of the regularity p of μ at ∂S_0, which is required to achieve the optimal rate of convergence, though not for consistency. The knowledge of p dictates the allocation between m and n for attaining the best possible rates. When p is unknown, the adaptive properties of dyadic trees (see Willett and Nowak (2007) and Singh, Scott and Nowak (2009)) could conceivably be utilized to come up with a near-optimal approach. This is a hard open problem and will be a topic of future research.

The baseline zone problem in the conventional regression setting: In the conventional regression setting of the problem (with n covariate–response pairs, a single response per covariate), the use of a p-value based strategy for estimating the baseline requires some sort of spatial smoothing, similar to the one–dimensional case as considered in Mallik et al. (2011). The estimation procedure from Section 2 can be easily adapted from the (m, n) setting to the regression setting. However, it appears difficult to establish the expected optimal rate of convergence $n^{-2/(4p+3)}$ using such a procedure, owing to a bias term arising in the analysis that cannot be adequately controlled. We refer the reader to Mallik, Banerjee and Woodroofe (2013) where this problem is analyzed and a slightly slower rate of $n^{-1/(2p+2)}$ is obtained; see their Remark 5 for more details on the bias issue. We do believe that a more fruitful alternative, one that is likely to produce an estimate with an optimal convergence rate, is to consider the excess–mass–based approach from Polonik and Wang (2005) with appropriate modifications to account for the fact that the level of the function that we deal with in this problem is at the *boundary* of the range of μ. While this promises to be an interesting direction, it is outside the scope of this paper, and we leave it for future research.

Acknowledgment

Supported by NSF Grants DMS-1007751, DMS-1308890 and a Sokol Faculty Award, University of Michigan

Appendix A: Proofs

A.1. Proof of Theorem 1

Here, we establish consistency with respect to the (stronger) Hausdorff metric,

$$d_H(S_1, S_2) = \max\left\{\sup_{x\in S_1} \rho(x, S_2), \sup_{x\in S_2} \rho(x, S_1)\right\}.$$

We only require $\min(m,n) \to \infty$ instead of taking m to be of the form $m_0 n^\beta$, $\beta > 0$.

To exhibit the dependence on m, we denote $\mathbb{M}_n$ by $\mathbb{M}_{m,n}$. Recall that $M_m(S) = E[\mathbb{M}_{m,n}(S)]$ converges to $M(S)$ for each $S \in \mathcal{S}$. Also, $\mathrm{Var}(\mathbb{M}_{m,n}(S)) = (1/n)\mathrm{Var}\left((\Phi(\sqrt{m}\bar{Y}_1) - \gamma)1_S(X)\right) \leq 1/n$. Hence, $\mathbb{M}_{m,n}(S)$ converges in probability to $M(S)$ for any $S \in \mathcal{S}$, as $\min(m,n) \to \infty$.

The space $(\mathcal{S}, d_H)$ is compact (Blaschke Selection Theorem) and M is a continuous function on $\mathcal{S}$. The desired result will be a consequence of argmin continuous mapping theorem van der Vaart and Wellner (1996, Thm. 3.2.2) provided we can justify that $\sup_{S\in\mathcal{S}} |\mathbb{M}_{m,n}(S) - M(S)|$ converges in probability to zero. To this end, let

$$\mathbb{M}^1_{m,n}(S) = \mathbb{M}_{m,n}(S) + \mathbb{P}_n \gamma 1_X(S) = \mathbb{P}_n \Phi\left(\sqrt{m}\bar{Y}\right) 1_S(X)$$

and $\mathbb{M}^1(S) = \mathbb{M}(S) + P\gamma 1_X(S)$. We have

$$\sup_{S\in\mathcal{S}} |\mathbb{M}_{m,n}(S) - M(S)| \leq \gamma \sup_{S\in\mathcal{S}} |(\mathbb{P}_n - P)(S)| + \sup_{S\in\mathcal{S}} |\mathbb{M}^1_{m,n}(S) - M^1(S)|.$$

The first term in this expression converges in probability to zero Ranga Rao (1962). As for the second term, notice that $\mathbb{M}^1_{m,n}(S)$ converges in probability to $M^1(S)$ for each S and $\mathbb{M}^1_{m,n}$ is monotone in S, $\mathbb{M}^1_{m,n}(S_1) \leq \mathbb{M}^1_{m,n}(S_2)$ whenever $S_1 \subset S_2$. As the space $(\mathcal{S}, d_H)$ is compact, there exist $S(1), \ldots, S(l(\delta))$ such that $\sup_{S\in\mathcal{S}} \min_{1\leq l\leq l(\delta)} d_H(S, S(l)) < \delta$, for any $\delta > 0$. Hence,

$$\begin{aligned}
&\sup_{S\in\mathcal{S}} |\mathbb{M}^1_{m,n}(S) - M^1(S)| \\
&= \max_{1\leq l\leq l(\delta)} \sup_{d_H(S,S(l))<\delta} |\mathbb{M}^1_{m,n}(S) - M^1(S)| \\
&\leq 2 \max_{1\leq l\leq l(\delta)} \sup_{d_H(S,S(l))<\delta} |\mathbb{M}^1_{m,n}(S) - M^1(S(l))| \\
&\leq 2 \max_{1\leq l\leq l(\delta)} \max(|\mathbb{M}^1_{m,n}(S_\delta) - M^1(S(l))|, |\mathbb{M}^1_{m,n}(S^\delta) - M^1(S(l))|,
\end{aligned}$$

where $S^\delta = \{x : d(x,S) < \delta\}$ and $S_\delta = \{x; d(x, S^c) > \delta\}$. The right side here converges in probability to $2\max_{1\leq l\leq l(\delta)} \max(|\mathbb{M}^1(S_\delta) - M^1(S_l)|, |\mathbb{M}^1(S^\delta) - M^1(S_l)|$ and can be made arbitrarily small by choosing small δ (as M^1 is continuous). Also, as the map $S \mapsto d_F(S, S_0)$ from $(\mathcal{S}, d_H)$ to $\mathbb{R}$ is continuous, we have consistency in the d_F metric as well. This completes the proof.

A.2. Proof of Proposition 2

Note that $\sqrt{mn}(\hat{\tau} - \tau_0) = O_P(1)$. So, given $\alpha > 0$, there exists $L_\alpha > 0$ such that for $V_{n,\alpha} = (\tau_0 - L_\alpha/\sqrt{mn}, \tau_0 + L_\alpha/\sqrt{mn})$, $P(\hat{\tau} \in V_{n,\alpha}) > 1 - \alpha$. Let $\hat{S}_n(\tau)$

denote the estimate of S_0 based on $\mathbb{M}_n(S,\tau)$. We have

$$P^*\left\{d(\hat{S}_n(\hat{\tau}),S_0)>\delta_n\right\}\le P^*\left\{d(\hat{S}_n(\hat{\tau}),S_0)>\delta_n,\hat{\tau}\in V_{n,\alpha}\right\}+\alpha.$$

Following the arguments for the proof of Proposition 1, the term on the right side can be bounded by

$$\sum_{k\ge 0,2^k\delta_n\le 1}P^*\left\{\inf_{S\in\mathcal{S}_{n,k}}\mathbb{M}_n(S,\hat{\tau})-\mathbb{M}_n(S_0,\hat{\tau})\le 0,\hat{\tau}\in V_{n,\alpha}\right\}.$$

This is further bounded by

$$\sum_{k=0}^{k_n}P^*\Bigg[\sup_{S\in\mathcal{S}_{n,k},\tau\in V_{n,\alpha}}|\{\mathbb{M}_n(S,\tau)-M(S)\}-\{\mathbb{M}_n(S_0,\tau)-M(S_0)\}|$$
$$>\inf_{S\in\mathcal{S}_{n,k}}\{M(S)-M(S_0)\}\Bigg]. \tag{A.1}$$

For $c(\gamma)=\min(\gamma-1/2,1-\gamma)$,

$$M(S)-M(S_0)=\left(\gamma-\frac{1}{2}\right)\{F(S_0)-F(S_0\cap S)\}+(1-\gamma)F(S_0\cap S)\ge c(\gamma)F(S\triangle S_0)$$

and hence (A.1) is bounded by

$$\sum_{k=0}^{k_n}P^*\left\{\sup_{\substack{S\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}}|(\mathbb{M}_n-M_m)(S,\tau)-(\mathbb{M}_n-M_m)(S_0,\tau)|>\frac{c(\gamma)2^k\delta_n}{3}\right\}$$
$$+\sum_{k=0}^{k_n}1\left[\sup_{\substack{S\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}}|\{M_m(S,\tau)-M_m(S,\tau_0)\}-\{M_m(S_0,\tau)-M_m(S_0,\tau_0)\}|\ge\frac{c(\gamma)2^k\delta_n}{3}\right]$$
$$+\sum_{k=0}^{k_n}1\left\{\sup_{S\in\mathcal{S}_{n,k}}|(M_m-M)(S,\tau_0)-(M_m-M)(S_0,\tau_0)|\ge\frac{c(\gamma)2^k\delta_n}{3}\right\}. \tag{A.2}$$

The third term can be shown to be zero for sufficiently large n in the same manner as in the proof of Proposition 1. The first term can be written as

$$\sum_{k=0}^{k_n}P^*\left[\sup_{S\in\mathcal{S}_{n,k},\tau\in V_{n,\alpha}}\left|\mathbb{G}_n g_{m,\tau}(\bar{Y})\{1_S(X)-1_{S_0}(X)\}\right|>\frac{c(\gamma)2^{k-1}\delta_n\sqrt{n}}{3}\right], \tag{A.3}$$

where $g_{m,\tau}(y)=\Phi\left(\sqrt{m}(y-\tau)\right)-\gamma$. We are now in a position to apply Theorem 2 to each term of (A.3). In the setup of Theorem 2, $N=2^{k-1}\delta_n\sqrt{n}$. The concerned class of functions is $\mathcal{G}_{n,k}=\{g_{m,\tau}(\bar{Y})(1_S(X)-1_{S_0}(X)):S\in\mathcal{S}_{n,k},\tau\in V_{n,\alpha}\}$. For any $S\in\mathcal{S}_{n,k}$, $\|g_{m,\tau}(1_S-1_{S_0})\|_{L_2(P_m)}\le[E1_{S\triangle S_0}(X)]^{1/2}\le(2^{k+1}\delta_n)^{1/2}$. So we can

pick $R = R_{n,k} = (2^{k+1}\delta_n)^{1/2}$. By our calculations in the proof of Proposition 1,

$$H_B(u, \{1_S - 1_{S_0} : S \in \mathcal{S}_{n,k}\}, L_2(P)) \leq \tilde{A}_0 u^{-1}$$

for some constant $\tilde{A}_0 > 0$. Also, for the class of functions $\mathcal{T}_n = \{g_{n,\tau}(\cdot) : \tau \in V_{n,\alpha}\}$, it can be shown by a simple partitioning argument that $N_{[]}(u, \mathcal{T}_n, L_2(P)) \leq A_1/u\sqrt{n} \leq A_1/u$, for some constant $A_1 > 0$. As the class $\mathcal{G}_{n,k}$ is formed by multiplying functions from two classes $\mathcal{S}_{n,k}$ and $\mathcal{T}_n$, and the brackets for these two classes can be taken to be bounded in absolute magnitude by 1 and 2, respectively, the bracketing number for $\mathcal{G}_{n,k}$ is bounded above by,

$$H_B\{u, \mathcal{G}_{n,k}, L_2(P)\} \leq 3\,\tilde{A}_0 u^{-1} + \log A_1 + \log\left(\frac{3}{u}\right) \leq A_2 u^{-1}.$$

In light of this, the first term in (A.2) can be shown to go to zero by arguing in the same manner as in the proof of Proposition 1.

For the second term in (A.2), $|\Phi(\sqrt{m}(\bar{Y} - \tau)) - \Phi(\sqrt{m}(\bar{Y} - \tau_0))| \leq \sup_u |\Phi(u + \sqrt{m}(\tau_0 - \tau)) - \Phi(u)|$, which equals $|\Phi(\sqrt{m}(\tau_0 - \tau)/2) - \Phi(-\sqrt{m}(\tau_0 - \tau)/2)|$. As Φ is Lipschitz of order 1, this is further bounded above by $\sqrt{m}|\tau_0 - \tau|/\sigma$. Hence, for sufficiently large n, the supremum appearing in the second line of (A.2) is bounded by

$$2\left(\frac{L_\alpha}{\sqrt{n}}\right) \sup_{S \in \mathcal{S}_{n,k}} P_m|1_S(X) - 1_{S_0}(X)| \leq \frac{L_\alpha 2^{k+2}\delta_n}{\sqrt{n}}.$$

This is eventually smaller that $c(\gamma)2^k\delta_n/3$ and hence, each term in the second sum of (A.2) is eventually zero. This completes the proof.

References

Bronšteĭn, E. M. (1976). ε-entropy of convex sets and functions. *Sibirsk. Mat. Žh.* **17**, 508–514, 715.

Cavalier, L. (1997). Nonparametric estimation of regression level sets. *Statistics* **29**, 131–160.

Eggleston, H. G. (1958). *Convexity. Cambridge Tracts in Mathematics and Mathematical Physics*, No. 47. New York: Cambridge University Press.

Geppetti, P. and Benemei, S. (2009). Pain treatment with opioids: achieving the minimal effective and the minimal interacting dose. *Clin. Drug. Investig.* **29**, 3–16.

Goldenshluger, A. and Spokoiny, V. (2006). Recovering convex edges of an image from noisy tomographic data. *IEEE Trans. Inform. Theory* **52**, 1322–1334.

Goldenshluger, A. and Zeevi, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist.* **34**, 1375–1394.

Härdle, W., Park, B. U. and Tsybakov, A. B. (1995). Estimation of non-sharp support boundaries. *J. Multivariate Anal.* **55**, 205–218.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer.*

Statist. Assoc. **82**, 267–270.

Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction*, vol. 82 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Ma, L., Guo, J., Wang, Y., Tian, Y. and Yang, Y. (2010). Ship detection by salient convex boundaries. In *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 1.

Mallik, A., Banerjee, M. and Woodroofe, M. (2013). Baseline zone estimation in two dimensions. *arXiv preprint arXiv:1312.6414*.

Mallik, A., Sen, B., Banerjee, M. and Michailidis, G. (2011). Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika* **98**, 887–900.

Mammen, E. and Tsybakov, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23**, 502–524.

Nolan, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39**, 348–371.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23**, 855–881.

Polonik, W. and Wang, Z. (2005). Estimation of regression contour clusters–an application of the excess mass approach to regression. *J. Multivariate Anal.* **94**, 227–249.

Qiu, P. (2007). Jump surface estimation, edge detection, and image restoration. *J. Amer. Statist. Assoc.* **102**, 745–756.

Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.* **33**, 659–680.

Scott, C. (2011). Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, L. Getoor and T. Scheffer, eds., ICML '11. New York, NY, USA: ACM.

Scott, C. and Davenport, M. (2007). Regression level set estimation via cost-sensitive classification. *IEEE Trans. Signal Process.* **55**, 2752–2757.

Singh, A., Scott, C. and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.* **37**, 2760–2782.

Stahl, J. S. and Wang, S. (2005). Convex grouping combining boundary and region information. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05. Washington, DC, USA: IEEE Computer Society.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25**, 948–969.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York: Springer-Verlag.

Wakimoto, R. M. and McElroy, J. L. (1986). Lidar observation of elevated pollution layers over los angeles. *J. Climate Appl. Meteor.* **25**, 1583–1599.

Wang, S., Stahl, J., Bailey, A. and Dropps, M. (2007). Global detection of salient convex boundaries. *Int. J. Comput. Vis.* **71**, 337–359.

Willett, R. M. and Nowak, R. D. (2007). Minimax optimal level set estimation. *IEEE Trans. Image Process.* **16**, 2965–2979.

550 S Tryon St, 5th Floor, MAC D1086-051, Charlotte NC 28202, USA.
E-mail: atul.mallik@gmail.com
Department of Statistics and Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
E-mail: moulib@umich.edu
Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
E-mail: michaelw@umich.edu

(Received April 2016; accepted February 2017)