VISUALISING VARIABILITY: CONFIDENCE REGIONS IN LEVEL SET ESTIMATION

Hanna JANKOWSKI[1] and Larissa STANBERRY[2]
[1]York University, Canada [2]Seattle Children's Research Institute, U.S.A.

ABSTRACT: Level sets of a function appear in numerous scientific problems, for example, in the detection of areas of high cancer rates. In practice, the true function is unknown and is therefore estimated. Here, we focus on quantifying the risk of replacing the unknown (true) function with its known estimator in the level set. We describe the variability, or accuracy, of the resulting estimator via the statistical notion of a confidence region, which naturally provides a graphical representation of variation easily visualized by the practitioner.

Keywords: Geometry, graphics, statistical inference, level set

1. INTRODUCTION

A level set of a real-valued function f is a set of points where the values $f(x)$ satisfy some constraint. For example, if $f(x)$ is the intensity of radiation at a point x on a surface, then the level set could be the area of the surface where the radiation intensity exceeds some threshold. Level sets are used in a variety of problems, including "hot spot" detection (e.g. to identify regions with low vegetation growth or areas with high cancer rates) and forecasting of extreme weather events.

In practice, the true function f is often unknown and is estimated from observed data by $\widehat{f}_n$. The level set of f is then estimated by the level set of $\widehat{f}_n$. A key practical question is to characterize and quantify the accuracy/risk of using the estimate $\widehat{f}_n$ in the level set vs. the unknown true function f. In statistics, this type of error is usually quantified using variance or even more appropriately, using standard deviation. However, characterizing, computing, and visualizing the variability of the level set estimate is a particularly difficult statistical problem. This is, in large part, due to the fact that the space of (closed) sets is nonlinear.

Mathematically, we denote the level set as

$$F(c_1,c_2) = \{x \in \mathscr{D} : c_1 \leq f(x) \leq c_2\}, \quad (1.1)$$

where $\mathscr{D} \subset \mathbb{R}^d$, f is a continuous function, $f : \mathscr{D} \mapsto \mathbb{R}$, and $-\infty \leq c_1 \leq c_2 \leq \infty$. The true level set $F(c_1,c_2)$ is then estimated as

$$\widehat{\boldsymbol{F}}_n(c_1,c_2) = \{x \in \mathscr{D} : c_1 \leq \widehat{f}_n(x) \leq c_2\}. \quad (1.2)$$

Sets of this form appear in various statistical problems, such as estimation of contour clusters [18, 19] or the estimation of density support [10]. In Section 5, we show how to apply the proposed method to the estimation of the domain of covariates with specified response level(s). Such a situation arises often in medical studies, such as dose optimization and toxic dose estimation [1, 23, 24], as well as in other fields. Additional examples include mode estimation, estimation of highest density and/or intensity regions [17], and abnormal system behaviour detection. The example given in [4] estimates the spherical density of double stars, and uses level sets to find directions with high densities of these double stars. [7] consider highest density regions in the estimation of the wintering location of the wood thrush songbird. Level sets are also closely related to random closed sets, and this relationship was studied in image inference applications in [8, 9].

Here, we propose to use confidence regions to characterize and visualize the variability of the

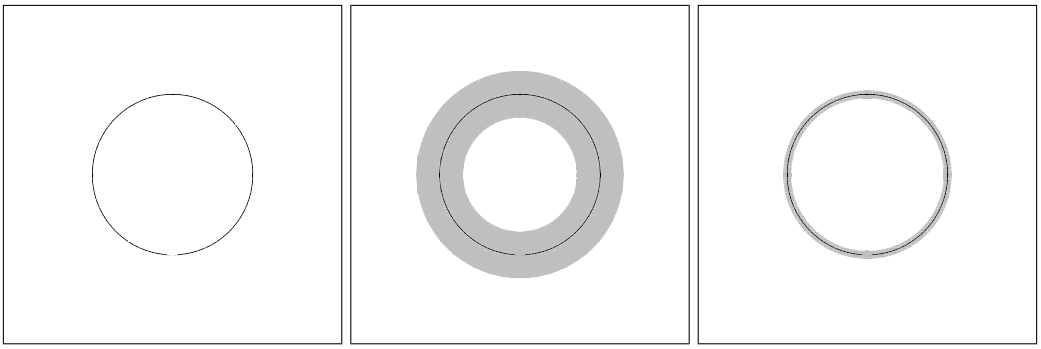

Figure 1: A toy example showing the observed level set (left) and two different confidence regions (centre and right, the confidence region is shown in grey and the estimated level set in black).

level-set estimators. In statistics, a $100(1-\alpha)\%$ confidence region $\boldsymbol{R}$ of the true level set $F(c_1,c_2)$ is a random set that covers $F(c_1,c_2)$ with a probability of at least $1-\alpha$, that is, $\text{pr}(\boldsymbol{R} \supset F(c_1,c_2)) \geq 1-\alpha$, under repeated experimentation. In practice, the common choice is $\alpha = 5\%$ that yields a 95% confidence region. A classical example is the estimation of the population mean μ. Under random samping, the estimate of μ is the sample mean, $\overline{X}_n$. Estimate $\overline{X}_n$ is the point estimate and it gives, statistically, the best guess of μ based on the observed data. The corresponding 95% confidence interval $\boldsymbol{R} = (\overline{X}_n - 1.96\sigma/\sqrt{n}, \overline{X}_n + 1.96\sigma/\sqrt{n})$ reflects the accuracy of the estimator $\overline{X}_n$ (σ is a standard error of the sample), i.e. a narrower confidence interval indicates a higher accuracy. The coverage of $\boldsymbol{R}$ is the probability of $F(c_1,c_2) \subset \boldsymbol{R}$. In the above example, the coverage of $\boldsymbol{R}$ is approxmately 95%.

Figure 1 shows a toy example of a level set $\widehat{\boldsymbol{F}}_n(0,0)$ (left), along with two different 95% confidence regions (middle and right). In practice, the true level set will not be known, and therefore we do not add it to this figure. The two confidence regions greatly differ in size with the larger one (middle) showing the estimator to be less accurate than the smaller one (right). Clearly, it is important for the practitioner to be able to visualize such a difference in the reliability of the estimate.

In set estimation, confidence regions can be used to describe the accuracy of the set estimator as well as its global and local variability. In [8], we proposed a new approach for calculating confidence regions for the mean of a random set. Developed independently, the approach was closely related to that of [3] as both were based on the idea of supremum inversion. This work is motivated by applications to image inference and effective dose estimation [3, 6, 8, 9, 12, 16]. Here, we provide a general framework for calculation of confidence regions for level sets. We focus on the case where the rate of convergence of the estimated function is $\sqrt{n}$, which is most often seen in parametric statistics. For some results in the nonparametric setting, we refer to [13].

The outline of this paper is as follows. In Section 1.1 we outline our notation and key assumptions. In Section 2 we consider consistency of the plug-in estimators and in Section 3 we describe our general approach for confidence region calculation under the assumptions of Section 1.1. Section 4 describes a simulation study, and Section 5 provides some examples. Code for the examples is available at `www.math.yorku.ca/~hkj/`.

1.1 Notation and Assumptions

Unless otherwise stated, we assume that $\mathscr{D}$ is the working domain and write, for example, $F(c_1,c_2) = \{x : c_1 \leq f(x) \leq c_2\}$ without stating

that $x \in \mathscr{D}$ explicitly. We assume that $\mathscr{D} \subset \mathbb{R}^d$, and denote the Euclidean norm of x as $|x|$.

We write $B_r(x_0) = \{x : |x - x_0| \leq r\}$ for the closed ball of radius r centred at x_0. For a set A, we write $A^o, \overline{A}, A^c$ and ∂A to denote its interior, closure, complement and boundary. Unless noted otherwise, set operations are calculated relative to the domain $\mathscr{D}$. That is, $A^c = \mathscr{D} \setminus A$, and so forth. Furthermore, for a set A, we define $A^\delta = \{x : B_\delta(x) \cap A \neq \emptyset\} = \cup_{x \in A} B_\delta(x)$. Deterministic sets are denoted using capital letters $A, B \ldots$, while bold upper-case lettering, $\boldsymbol{A}, \boldsymbol{B}, \ldots$, is used for random sets. We do this to emphasize the difference between the random and observed set. Recall also that the Hausdorff distance between two sets, A and B, is defined as

$$\rho(A,B) \quad = \quad \inf\left\{\delta > 0 : A \subset B^\delta, B \subset A^\delta\right\}.$$

The notation $C(\mathscr{D})$ is used to denote the space of continuous functions $C(\mathscr{D}) = \{f : \mathscr{D} \mapsto \mathbb{R},\ f \text{ continuous}\}$ endowed with the uniform metric. We write $X_n \Rightarrow X$ to say that X_n converges weakly to X. When handling weak convergence of stochastic processes or random fields, we assume that they take values in $C(\mathscr{D})$.

Suppose now that $\widehat{f}_n$ is a random, continuous function such that

$$\text{(A1)} \qquad \sup_{x \in K} |\widehat{f}_n(x) - f(x)| \to 0$$

almost surely (almost everywhere), as $n \to \infty$, for all compact sets $K \subset \mathscr{D}$. To construct confidence regions, the sets (1.1) are restricted to a compact window $\mathscr{W} \subseteq D$, and we require the assumption of weak convergence

$$\text{(A2)} \qquad \sqrt{n}\{\widehat{f}_n(\cdot) - f(\cdot)\} \Rightarrow \mathbb{Z}(\cdot),$$

where $\mathbb{Z}(\cdot)$ is a continuous random field on $\mathscr{W}$. In practice, assumption (A2) can be checked using the techniques described in [2, 25] for $\mathscr{D} \subset \mathbb{R}$ or [11] for $\mathscr{D} \subset \mathbb{R}^d$.

2. CONSISTENCY

An estimator is said to be consistent if it approaches the quantity it is estimating as the sample size increases. Consistency is a core concept in statistics, because if an estimator is biased, this bias becomes negligible for a sufficiently large sample size. Below we provide conditions required for consistency of $\widehat{\boldsymbol{F}}_n(c_1, c_2)$. The proofs appear in the Appendix, and follow from [16] and/or [4].

Let $\mathscr{F}$ be the family of closed sets of $\mathbb{R}^d$ and let $\mathscr{K}$ denote the family of all compact subsets of $\mathbb{R}^d$. For a probability triple $(\Omega, \mathscr{A}, P)$, a random closed set is the mapping $\boldsymbol{A} : \Omega \mapsto \mathscr{F}$ such that for every compact set $K \in \mathscr{K}$

$$\{\omega : \boldsymbol{A}(\omega) \cap K \neq \emptyset\} \in \mathscr{A},$$

(cf. [15]). Note that

$$\begin{aligned} &\{\widehat{\boldsymbol{F}}_n(c_1,c_2) \cap K \neq \emptyset\} \\ &\quad = \quad \left\{\inf_{x \in K} \left|\widehat{f}_n(x) - \frac{c_1 + c_2}{2}\right| \leq \frac{c_2 - c_1}{2}\right\}, \end{aligned}$$

Therefore, since the functions $\widehat{f}_n$ are continuous almost surely, the estimators (1.2) satisfy the measurability requirement and are well-defined.

A random closed set $\boldsymbol{A}_n$ *converges strongly* to a deterministic set A if for any compact set K, $\rho(\boldsymbol{A}_n \cap K, A \cap K) \to 0$ almost surely (almost everywhere) [16]. The key conditions for the consistency of the estimators (1.2) are

$$\{x : f(x) \leq c\} \quad = \quad \overline{\{x : f(x) < c\}} \qquad (2.1)$$

$$\{x : c \leq f(x)\} \quad = \quad \overline{\{x : c < f(x)\}}. \qquad (2.2)$$

Theorem 2.1. *Under assumption (A1), the estimator $\widehat{\boldsymbol{F}}_n(c_1,c_2)$ converges strongly to $F(c_1,c_2)$ if the function f satisfies condition* (2.1) *at $c = c_2$ and condition* (2.2) *at $c = c_1$. Moreover,* (2.1) *and* (2.2) *are necessary in the following sense:*

1. *Suppose that x_0 is a point such that there exists a neighbourhood $B_\delta(x_0)$ and a subsequence n_k such that $\widehat{f}_{n_k}(x) > f(x)$ for all $x \in B_\delta(x_0)$. If $\widehat{\boldsymbol{F}}_n(c_1,c_2)$ is consistent, then* (2.1) *must hold at x_0 for $c = c_2$ in the sense that $x_0 \notin \{x : f(x) \leq c_2\} \setminus \overline{\{x : f(x) < c_2\}}$.*
2. *Suppose that x_0 is a point such that there exists a neighbourhood $B_\delta(x_0)$ and a subsequence n_k such that $\widehat{f}_{n_k}(x) < f(x)$ for all $x \in$*

$B_\delta(x_0)$. *If* $\widehat{\boldsymbol{F}}_n(c_1,c_2)$ *is consistent, then* (2.2) *must hold at* x_0 *for* $c = c_1$ *in the sense that* $x_0 \notin \{x : c_1 \le f(x)\} \setminus \overline{\{x : c_1 < f(x)\}}$.

Example 1: Let $\mathscr{D} = [-2,2]^2 \subset \mathbb{R}^2$ with $f(x) = |x|$ and $F(-\infty,1) = \{x : f(x) \le 1\}$, the disc with radius one centred at the origin. Suppose $U_1,\ldots,U_n$ are independent and identically distributed random variables from the uniform distribution on $[-1,1]^2$, and let $\bar{U}_n$ denote their bivariate sample mean. Then $\widehat{f}_n(x) = |x - \bar{U}_n|$ converges uniformly to $f(x)$ on $\mathscr{D}$, and $f(x)$ satisfies (2.1) and (2.2) at $p = 1$. Therefore, $\widehat{\boldsymbol{F}}_n(-\infty,1)$ and $\widehat{\boldsymbol{F}}_n(1,1)$ are consistent for $F(-\infty,1)$ and $F(1,1) = \{x : |x| = 1\}$. In this case, the Hausdorff distance $\rho(\widehat{\boldsymbol{F}}_n(-\infty,1), F(-\infty,1)) = \rho(\widehat{\boldsymbol{F}}_n(1,1), F(1,1)) = |\bar{U}_n|$ converges to zero almost surely, by the strong law of large numbers.

3. CONFIDENCE REGIONS

Now, assume that the estimating functions $\widehat{f}_n$ satisfy assumption (A2) for some compact window $\mathscr{W} \subset \mathscr{D}$. The confidence regions for the sets (1.1) restricted to $\mathscr{W}$ can be obtained as follows. Let q_1 and q_2 be the quantiles of the process $\sup_{x\in\mathscr{W}} \mathbb{Z}(x)$ such that

$$\begin{aligned} \mathrm{pr}\left(\sup_{x\in\mathscr{W}} \mathbb{Z}(x) \le q_1\right) &= 1-\alpha, \\ \mathrm{pr}\left(\sup_{x\in\mathscr{W}} |\mathbb{Z}(x)| \le q_2\right) &= 1-\alpha. \end{aligned}$$

Then the sets

$$\begin{gathered} \left\{x \in \mathscr{W} : \widehat{f}_n(x) \le c + q_1/\sqrt{n}\right\}, \\ \left\{x \in \mathscr{W} : c_1 - \tfrac{q_2}{\sqrt{n}} \le \widehat{f}_n(x) \le c_2 + \tfrac{q_2}{\sqrt{n}}\right\}, \end{gathered} \tag{3.1}$$

form $100(1-\alpha)\%$ confidence regions for $\mathscr{W} \cap F(-\infty,c)$ and $\mathscr{W} \cap F(c_1,c_2)$, respectively (where $-\infty < c_1 \le c_2 < \infty$ and $c \in \mathbb{R}$). Note that the random variables $\sup_{x\in\mathscr{W}} |\mathbb{Z}(x)|$ and $\sup_{x\in\mathscr{W}} \mathbb{Z}(x)$ are well-defined because $\mathscr{W}$ is compact and $\mathbb{Z}$ has continuous sample paths. This also implies that we may use $\max_{x\in\mathscr{W}} |\mathbb{Z}(x)|$ and $\max_{x\in\mathscr{W}} \mathbb{Z}(x)$ to calculate the quantiles, which is computationally easier. In what follows, we assume that $\mathscr{W} = \mathscr{D}$, where $\mathscr{D}$ is compact, unless otherwise stated.

Recall that the coverage of a confidence region refers to the probability with which it covers the quantity of interest. For a $100(1-\alpha)\%$ confidence region, it is typically considered ideal if the coverage is as close to $100(1-\alpha)\%$ without going over. Such a confidence region is preferred because it is *conservative*, in the sense that it will never under quantify the variability of the estimator. Let us first show that our approach yields such a confidence region.

$$\begin{aligned} &\left\{F(-\infty,c) \subset \{\widehat{f}_n(x) \le c + q_1/\sqrt{n}\}\right\}^c \\ &\quad = \left\{\widehat{f}_n(x) > c + q_1/\sqrt{n} \;\; \exists\, x \in F(-\infty,c)\right\}. \end{aligned}$$

Now, let $\mathbb{Z}_n(x) = \sqrt{n}(\widehat{f}_n(x) - f(x))$. Then

$$\begin{aligned} &\mathrm{pr}\left(\widehat{f}_n(x) > c + q_1/\sqrt{n} \;\; \exists x \in F(-\infty,c)\right) \\ &= \mathrm{pr}\left(\mathbb{Z}_n(x) > \sqrt{n}(c - f(x)) + q_1 \right. \\ &\qquad\qquad \left. \exists x \in F(-\infty,c)\right) \\ &\le \mathrm{pr}\left(\mathbb{Z}_n(x) > q_1 \;\; \exists x \in F(-\infty,c)\right) \\ &\le \mathrm{pr}\left(\mathbb{Z}_n(x) > q_1 \;\; \exists x\right), \end{aligned} \tag{3.2}$$

and taking the limit in n, the latter quantity is less than or equal to α by definition of q_1. It therefore follows that

$$\mathrm{pr}(F(-\infty,c) \subset \{\widehat{f}_n(x) \le c + q_1/\sqrt{n}\}) \ge 1-\alpha,$$

as required, asymptotically. A similar approach works for the case $F(c_1,c_2)$.

The above calculation is illuminating for several reasons:

- Notice that the consistency conditions play no role in the design of the confidence region. Indeed, the confidence region functions as intended even if consistency is violated (see e.g. Example 3).
- The smoothness and variability of the field $\mathbb{Z}$ determines the "size" of the confidence set, which may not be uniform over $\mathscr{W}$. In fact, the larger the window $\mathscr{W}$ is chosen, the wider

the confidence set is. Furthermore, going from the penultimate to the ultimate line in (3.2) we note that the bounding term $\sup \mathbb{Z}(x)$ could be replaced with many other upper bounds, including $\sup_{x \in F(-\infty,c)} \mathbb{Z}(x)$, and asymptotically with $\sup_{x \in F(c,c)} \mathbb{Z}(x)$.

– The calculations in (3.2) give also some ideas on the power of the methodology. In this context, statistical power would indicate an ability to also recognize the regions which should not be in $F(c_1,c_2)$ with high probability. For $x \in F(-\infty,c)^c$, the quantity $\sqrt{n}(c-f(x))$ converges to $-\infty$, and therefore these locations should be easily picked up by the confidence region for large enough sample sizes.

– The confidence regions are conservative, in that the upper bounds in (3.2) mean that over-coverage is possible, even asymptotically. We explore the extent of this via simulations in Section 4.

Finally, we note that our methods are straightforward to implement, and, if exact limiting distributions cannot be found, the confidence sets may be estimated using bootstrap methods, either parametric or nonparametric [5, for additional details on bootstrap methodologies]. In what follows, we used the parametric bootstrap in Example 5 and the nonparametric bootstrap in Example 6.

Example 2: To build 95% confidence regions for Example 1: $F(1,1) = \{x : |x| = 1\}$ with $\mathscr{W} = \mathscr{D} = [-2,2]^2$, we calculate $\mathbb{Z}_n(x) = \sqrt{n}\left(|x-\bar{U}_n| - |x|\right)$, where $\bar{U}_n$ is the average of n independent Uniform$[-1,1]^2$ random variables. Clearly, $\mathbb{Z}_n(x)$ has continuous sample paths. Also, $|\mathbb{Z}_n(x)| \leq \sqrt{n}|\bar{U}_n|$ for all x and, since this is realized at $x = 0$, we obtain that $\sup_{x \in \mathscr{D}} |\mathbb{Z}_n(x)| = \sqrt{n}|\bar{U}_n|$. The limiting distribution is therefore $\sqrt{3^{-1}(Z_1^2 + Z_2^2)}$ where Z_1, Z_2 are independent standard normal variables, and it therefore follows that $q_2 = 1.41$.

4. SIMULATION STUDY

In (3.2), we have shown that the confidence regions (3.1) cover the true level set at least $100(1-\alpha)\%$ of the time, for sufficiently large n. Here, we use Monte Carlo simulations to calculate the empirical coverage probabilities for various examples and sample sizes. The goal is to understand the actual behaviour of the methods, as well as the amount of over-coverage one could expect to see in practice. The cases we consider are (A) Example 2 of Section 3, and (B) Example 3 and (C) Example 4 described below. For each of the three examples, we calculate the $100(1-\alpha)\%$ confidence region with $\mathscr{W} = \mathscr{D}$ and estimate the coverage. We consider the following sets:

		$\mathscr{W} = \mathscr{D}$
(A)	$F(1,1)$	$[-2,2]^2$
(B)	$F(-\infty,1)$	$[-1,2]$
(C)	$F(0,1)$	$[-2,2]^2$

Table 1: Empirical coverage probabilities for 95% confidence region. The standard error due to Monte Carlo sampling is 0.0022.

	$n = 25$	$n = 100$	$n = 1000$
(A)	95.26	94.83	95.12
(B)	94.79	95.34	95.47
(C)	97.81	98.16	98.45

The simulations were done in MATLAB, on a discretized domain $\mathscr{D}$. Because the discretization introduces some error into the calculations [22], we selected a large lattice and calibrated it to give accurate results as follows. Suppose that $\mathscr{D} \subset \mathbb{R}^2$ and $f(x) = |x| - 0.5$ with $\widehat{f}_n(x) = |x| - \bar{U}_n$, where $\bar{U}_n$ is the average of n independent samples from a uniform distribution on $[0,1]$. The confidence regions for $F(0,0)$ are exact because of the special separable form of the function $\widehat{f}_n$ (modulo the sample size approximations). For $n = 1000$ and lattices with $m = 200, 400, 600, 800, 1000$, the empirical coverage probabilities for the 95% confidence region were 95.10, 95.02, 95.30, 95.30, and 94.96s, respectively; the standard error due to Monte Carlo sampling was 0.003. From here,

Table 2: Empirical coverage probabilities for 90% confidence region; the standard error due to Monte Carlo sampling is 0.0030.

	$n=25$	$n=100$	$n=1000$
(A)	89.87	89.70	90.17
(B)	88.49	90.07	90.17
(C)	96.10	96.09	96.71

we selected $m = 600$ for simulations.

Example 3: Let $\mathscr{D} = [-1,2] \subset \mathbb{R}$ and $h(x) = |x-0.5|-0.5$. Let $\widehat{f}_n(x) = \widehat{p}_n|h(x)| + (1-\widehat{p}_n)h(x)$, where $n\widehat{p}_n$ is a binomial random variable with parameters n and $p = 1/2$. Set the true function $f(x) = E[\widehat{f}_n(x)] = \max(h(x),0)$ and $F(-\infty,0) = \{x : f(x) \le 0\} = [0,1]$. Notice that f does not satisfy condition (2.1) at $c = 0$. Indeed, if $\widehat{p}_n > 0.5$ then $\widehat{\boldsymbol{F}}_n(-\infty,0) = \{0,1\}$ and otherwise $\widehat{\boldsymbol{F}}_n(-\infty,0) = [0,1]$. Clearly, consistency does not hold, however, the confidence regions will still behave as expected.

Example 4: Suppose that $\mathscr{D} = [-2,2]^2$ and $f(x) = \beta^T\tilde{x}$, where $\beta^T = (0.5,1,2,-3,1)$ and $\tilde{x} = (1,x_1,x_2,x_1x_2,x_1^2)^T$. This is estimated by the regression function $\widehat{f}_n(x) = \widehat{\beta}_n^T\tilde{x}$, where $\widehat{\beta}_n$ is normally distributed with mean β and variance Σ/n with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.2$ for $i \neq j$. The quantiles of the fluctuation field $\sup_{x\in\mathscr{D}}|\mathbb{Z}(x)|$ were found empirically. The set being estimated is $F(0,1)$.We show examples of the confidence regions for $n = 25, 100$ and 1000 in Figure 2. We refer to Example 5 for a more detailed regression example.

Tables 1 and 2 show the results for $B = 10,000$ Monte Carlo simulations. Although the confidence regions are conservative, (A) and (B) both show almost exact coverage. In (C), the variability of the fluctuation field is reflected in the size of the confidence sets. The effect is compounded because the function f is relatively flat in the neighbourhood of $F(0,1)$.

5. EXAMPLES AND APPLICATIONS

5.1 Covariate Domain Estimation

Least squares regression is a well-known statistical tool ubiquitous across the sciences. In regression analysis, the expected response is modeled as a function of covariates, i.e. we model $E[Y|X = x]$ where Y is the response variable and x is the vector of covariates. Given $f(x) = E[Y|X = x]$, the various sets (1.1) describe the domain of covariates for which the mean response lies within a specified target range. Although we focus on the linear regression model, the method can be easily extended to most generalized linear models. We thus define $f(x) = \beta^T\widetilde{x}$, with $\widetilde{x} = \widetilde{x}(x)$ denoting some continuous function of the covariates. We assume that the covariates are continuous and lie in $\mathscr{D} \subset \mathbb{R}^d$, with $p+1 \ge d$ denoting the number of regression variables. For example, if $f(x) = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$, then $d = 2, p = 3$ and $\widetilde{x}^T = \widetilde{x}^T(x_1,x_2) = (1,x_1,x_2,x_1x_2)$.

Consider the estimating function

$$\widehat{f}_n(x) = \widehat{\beta}_n^T\widetilde{x},$$

where n is the number of observations. Since $\widetilde{x}$ is continuous, the image of $\mathscr{D}$ under $\widetilde{x}$ is compact. It follows that $\widehat{f}_n(x)$ converges uniformly to $f(x)$ as long as $\widehat{\beta}_n$ are consistent estimators, and therefore, $\widehat{f}_n$ satisfies assumption (A1). The conditions (2.1) and (2.2) need to be checked on a case by case basis. For example, for $f(x) = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$, both conditions are satisfied for any value of c, as long as at least one of β_1,β_2,β_3 is non-negative.

Next, let $\mathbb{Z}_n(x) = \sqrt{n}(\widehat{f}_n(x) - f(x))$. If $\widehat{\beta}_n$ is asymptotically normal, we have that $\sqrt{n}(\widehat{\beta}_n - \beta) \Rightarrow Z$, where Z is multivariate normal with mean zero and variance Σ. If unknown, Σ is estimated using standard regression methods. Since $\widetilde{x}$ is continuous, it follows that $\mathbb{Z}_n(x)$ converges weakly in $C(\mathscr{D})$ to a continuous, mean zero Gaussian field, $\mathbb{Z}(x) = Z^T\widetilde{x}$, with covariance structure given by $\mathrm{cov}(\mathbb{Z}(x),\mathbb{Z}(x')) = \widetilde{x}^T\Sigma\widetilde{x}$. Therefore, $\widehat{f}_n$ satisfies assumption (A2), and confidence sets

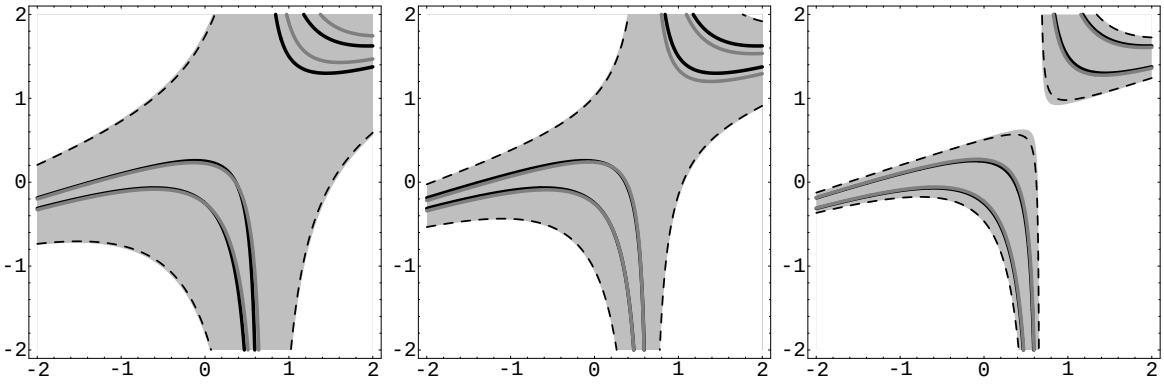

Figure 2: For each of $n = 25, 100, 1000$ (from left to right), the boundary of the true set $F(0,1)$ (black) and the boundary estimate $\widehat{\boldsymbol{F}}_n(0,1)$ (dark grey) are shown along with the 90% confidence region (light grey). The boundary of the modified 90% confidence region is also shown (dashed).

may be formed as described above.

In [3], this problem was considered for a logistic regression in the context of effective dose estimation, and this approach was later studied numerically in [12]. The idea we present here and the approach of [3] are similar in that they both obtain upper bounds on the supremum of the fluctuation process. The upper bounds of [3] are derived via a similar idea to that used for Scheffé's bounds for simultaneous confidence intervals [21]. We refer to [3] and [12] for more details. The general methodology described in Section 3 does not specify the best way in which to obtain these quantiles. The approach of [3] is one which does work for general settings of parametric regression problems, but these bounds are very conservative, and therefore result in much over coverage for the methodology [see 12]. In the example below, we show a more direct approach of estimating the quantiles of $\sup_{x\in\mathscr{D}} \mathbb{Z}(x)$.

Example 5: We illustrate the method on the data set `trees` available with R [20]. Here, girth (in inches), height (in feet) and volume (in cubic feet) of timber were recorded for 31 felled black cherry trees. Set $x_1 =$ girth and $x_2 =$ height. Fitting the model $E[\log Y|x] = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2$, we obtain estimates $\widehat{\beta}_0 = -6.63$ (p-value $= 5.1e-09$), $\widehat{\beta}_1 = 1.98$ (p-value $< 2e-16$) and $\widehat{\beta}_2 = 1.12$ (p-value $= 7.8e-06$). The estimates are not far from the formula volume = height $\times$ girth$^2/4\pi$.

Set $\mathscr{D} = [5,25] \times [50,100]$, and suppose that we are interested in the domain of covariates for which the log-volume is at least $\log 30$ (≈ 3.4), that is,

$$\begin{aligned} & F(-\log 30, \infty) \\ &= \{x : E[\log Y|x] \geq \log 30\} \\ &= \{x : -\beta_0 - \beta_1 \log x_1 - \beta_2 \log x_2 \leq -\log 30\}. \end{aligned}$$

Figure 3 shows the estimator $\widehat{\boldsymbol{F}}_n(-\log 30, \infty) = \{x : \widehat{f}_n(x) \leq -\log 30\}$, where $\widehat{f}_n(x) = \widehat{\beta}^T \widetilde{x}$ and $\widetilde{x}^T = (-1, -\log x_1, -\log x_2)$. Note that $\widetilde{x}$ is continuous on $\mathscr{D}$.

The true function $f(x) = -\beta_0 - \beta_1 \log x_1 - \beta_2 \log x_2$ is strictly decreasing in both x_1 and x_2, and therefore, it satisfies condition (2.1) at $c = -\log 30$, or for any other choice of c. Condition (2.2) is also satisfied, although we do not require it here. It follows that the set $\widehat{\boldsymbol{F}}_n(-\log 30, \infty)$ is consistent for $F(-\log 30, \infty)$.

The 95% confidence set for $F(-\log 30, \infty)$ is $\{x : \widehat{f}_n(x) \leq -\log 30 + q_1/\sqrt{31}\}$, where q_1 is the value such that $\mathrm{pr}(\sup_{x\in\mathscr{D}} \mathbb{Z}(x) \leq q_1) = 0.95$. In this case, the fluctuation process is $\mathbb{Z}(x) = Z^T \widetilde{x}$,

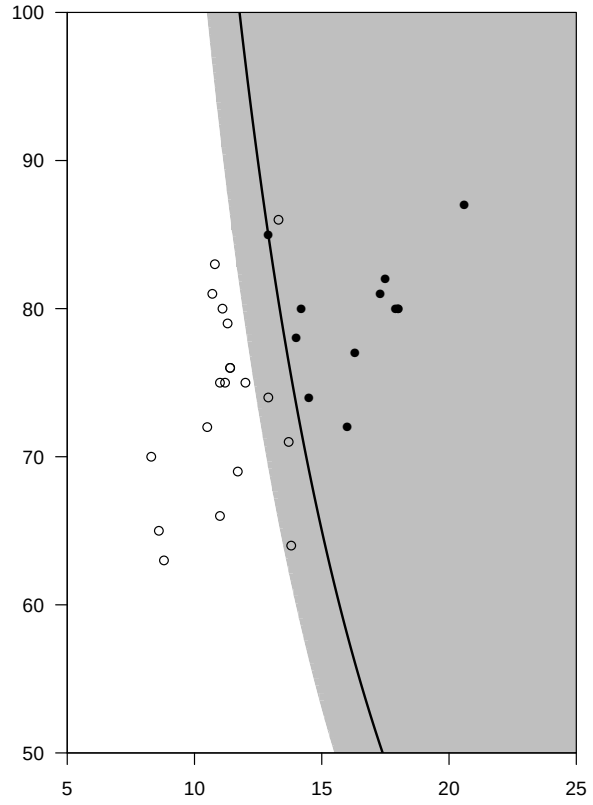

Figure 3: The estimated set $\widehat{F}_n(-\log 30, \infty)$ (the boundary of this set is shown in black) with the 95 % confidence region (light gray). The data points are also shown as circles, with the filled in circles showing those data with volume ≥ 30.

where $Z^T = (Z_0, Z_1, Z_2)$ is a mean-zero multivariate Gaussian. Under the normal linear model, a consistent estimator of the covariance matrix of Z is $\widehat{\Sigma}$ where

$$\begin{aligned}\widehat{\Sigma} &= n\widehat{\sigma}^2(X'X)^{-1} \\ &= 31 \times \begin{bmatrix} 0.6397 & 0.0208 & -0.1601 \\ 0.0208 & 0.0056 & -0.0081 \\ -0.1601 & -0.0081 & 0.0418 \end{bmatrix},\end{aligned}$$

where X is the design matrix of the regression. The supremum of $\mathbb{Z}$ must occur on one of the corners of $\mathscr{D}$,

$$\sup_{z \in \mathscr{D}} \mathbb{Z}(x) = \max_{i=1,2, j=1,2} \left\{ \mathbb{Z}(a_i, b_j) \right\},$$

for $a_1 = \log 5, a_2 = \log 25, b_1 = \log 50$ and $b_2 = \log 100$. We estimate the quantile via Monte Carlo sampling from a multivariate Gaussian with variance $\widehat{\Sigma}$, to obtain

$$\left\{ x : -\beta_0 - \beta_1 \log x_1 - \beta_2 \log x_2 \leq -\log 30 + \tfrac{1.28}{\sqrt{31}} \right\}$$

as the approximate 95% confidence region. The resulting set is shown in Figure 3.

5.2 Estimating the Boundary of a Density Region

Here, our goal is to identify the set $F(c,c) = \{x : f(x) \geq c\}$ for the unknown density f. This problem is closely related to clustering, in the sense that for certain cutoff values c, the set $F(c,c)$ can be used to identify both the number of clusters and their centres.

Example 6: Consider the problem of estimating the level set of density function. Suppose that f is a mixture density given by $f = 0.5g_1 + 0.5g_2$, where g_1 and g_2 are both bivariate Gaussian densities. Figure 4 shows two examples of such density functions with different degrees of separation between the mixture components g_i. Using a random sample of size $n = 1000$, we estimate the level set $F(0.055, 0.055) = \{x : f(x) = 0.055\}$ for both examples. Wider confidence regions in Figure 5 (left) indicate that the lack of separation between the mixture components leads to higher variability (equivalently, less accuracy) of the estimator.

The drastic difference in Figure 5 (top vs. bottom) implies that confidence regions can be effectively used to characterize and visualize the variability and accuracy of level set estimators.

This idea closely relates to estimation of high intensity regions, where one is interested in finding "hot spots", or regions where the estimated intensity crosses some threshold. Important examples include high/low vegetation growth or regions exhibiting high cancer rates [17] or probabilistic forecasting of extreme weather events in meteorology [14].

6. DISCUSSION

We present a method for graphically presenting the risk associated with replacing the unknown

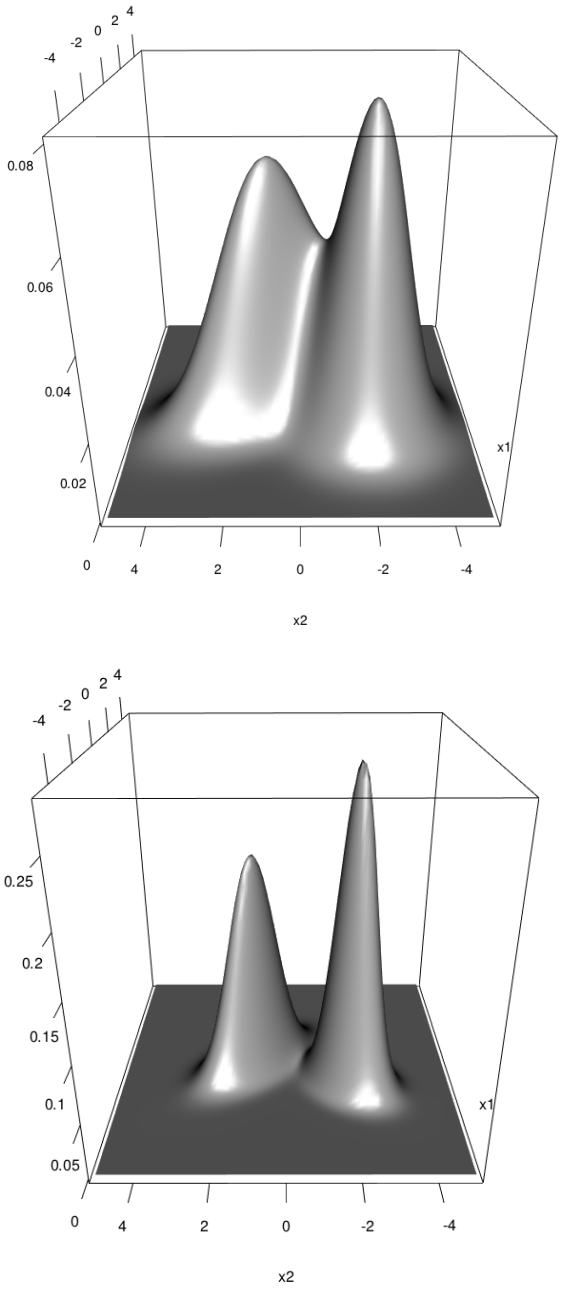

Figure 4: Two Gaussian mixture densities : f (top) and $\widetilde{f}$ (bottom), as defined in the text.

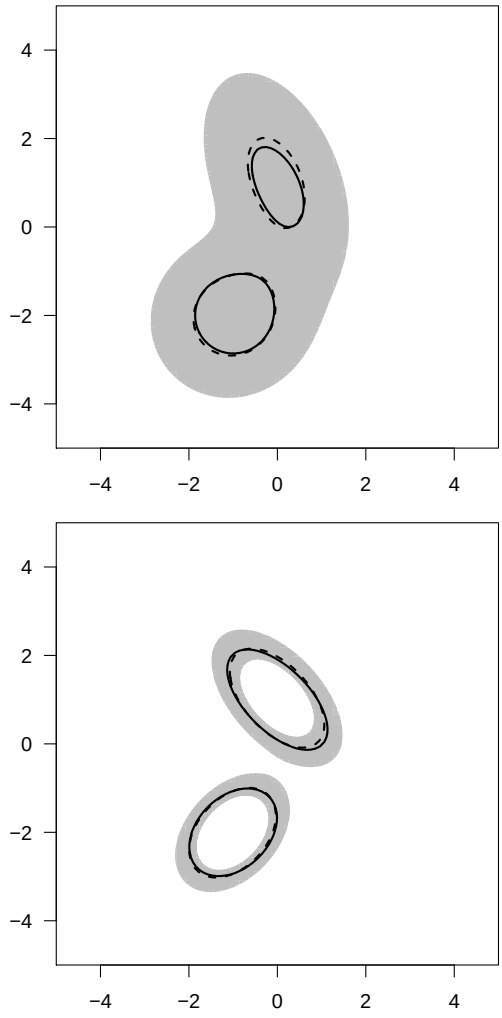

Figure 5: The boundary of true level set $F(0.055, 0.055)$ (solid line) and the boundary of the estimated level set $\widehat{\boldsymbol{F}}_n(0.055, 0.055)$ (dashed line) for densities f (top) and $\widetilde{f}$ (bottom) in Figure 4. The level set estimates are based on a random sample of size $n = 1000$. Visually, both the level sets and their estimates appear quite similar. However, the 95% confidence regions (gray) reveal great variability in the accuracy of the two estimates.

function with its estimator when computing a level set. The method is appropriately conservative, and we have studied the amount of overcoverage in several examples through simulations. We also give some specific examples based on simulated and real data.

APPENDIX

Technical details

Lemma 6.1. *Suppose that f is continuous. Then*

$$\{x : c_1 \le f(x)\}^{\varepsilon} \cap \{x : f(x) \le c_2\}^{\varepsilon} = \{x : c_1 \le f(x) \le c_2\}^{\varepsilon}.$$

Proof. Suppose y is in the set $\{x : c_1 \le f(x)\}^{\varepsilon} \cap \{x : f(x) \le c_2\}^{\varepsilon}$ but not in the set $\{x : c_1 \le f(x) \le c_2\}$. Then one of two possibilities exists: Either $f(y) < c_1$ or $f(y) > c_2$. The argument for both cases is the same, so we present only the first instance.

Assume then that y is such that $f(y) < c_1$. By definition of y, there exists an x_1 such that $c_1 \le f(x_1)$ and $y \in B_{\varepsilon}(x_1)$, or an x_2 such that $f(x_2) \le c_2$ and $y \in B_{\varepsilon}(x_2)$. In the first setting, since f is continuous, there also exists a z such that $c_1 \le f(z) \le c_2$ and $d(y,z) \le \varepsilon$. For example, one such z must fall on the line between y and x_1, which would clearly satisfy $|y-z| \le \varepsilon$. A similar argument shows that if $y \in B_{\varepsilon}(x_2)$, then there exists a z such that $c_1 \le f(z) \le c_2$ and $|y-z| \le \varepsilon$. It follows that $y \in B(z,\varepsilon)$. This proves that

$$\begin{aligned}\{x : c_1 \le f(x)\}^{\varepsilon} \cap \{x : f(x) \le c_2\}^{\varepsilon} \\ \subset \quad \{x : c_1 \le f(x) \le c_2\}^{\varepsilon}.\end{aligned}$$

Containment in the other direction is immediate, completing the proof. □

Proof of Theorem 2.1. The proof here is similar to that of [16]. Without loss of generality, we may assume that $\mathscr{D}$ is compact. If $f : \mathscr{D} \mapsto \mathbb{R}$ is a continuous function satisfying the conditions of the theorem, then

$$\begin{aligned}\widetilde{\varphi}(\pm\varepsilon) &= \rho(\{x : f(x) \le c_2\}, \{x : f(x) \le c_2 \pm \varepsilon\}) \\ \varphi(\pm\varepsilon) &= \rho(\{x : c_1 \le f(x)\}, \{x : c_1 \pm \varepsilon \le f(x)\})\end{aligned}$$

are all continuous for ε near zero, and moreover, they both converge to zero as $\varepsilon \to 0$. Now, by (A1), we know that $\widehat{f}_n$ converges uniformly to f with probability one. Let

$$\eta_n = \sup_{x \in \mathscr{D}} |f(x) - \widehat{f}_n(x)|,$$

and also define

$$\varepsilon_n = \max\{\varphi(\eta_n), \varphi(-\eta_n), \widetilde{\varphi}(\eta_n), \widetilde{\varphi}(-\eta_n)\}$$

which converges to zero as $n \to \infty$ almost surely. We will next show that $\rho(\{x : c_1 \le \widehat{f}_n(x) \le c_2\}, \{x : c_1 \le f(x) \le c_2\}) \le \varepsilon_n$. To this end

$$\begin{aligned}&\{x : c_1 \le f(x) \le c_2\} \\ &\quad \subset \quad \{x : f(x) \le c_2 - \eta_n\}^{\widetilde{\varphi}(-\eta_n)} \\ &\quad \subset \quad \{x : \widehat{f}_n(x) \le c_2\}^{\widetilde{\varphi}(-\eta_n)} \subset \{x : \widehat{f}_n(x) \le c_2\}^{\varepsilon_n}.\end{aligned}$$

Repeating in the other direction, we obtain

$$\begin{aligned}&\{x : c_1 \le f(x) \le c_2\} \\ &\quad \subset \quad \{x : c_1 + \eta_n \le f(x)\}^{\varphi(\eta_n)} \\ &\quad \subset \quad \{x : c_1 \le \widehat{f}_n(x)\}^{\varphi(\eta_n)} \subset \{x : c_1 \le \widehat{f}_n(x)\}^{\varepsilon_n}.\end{aligned}$$

and hence, by Lemma 6.1,

$$\begin{aligned}&\{x : c_1 \le f(x) \le c_2\} \\ &\quad \subset \quad \{x : c_1 \le \widehat{f}_n(x) \le c_2\}^{\varepsilon_n}. \qquad \text{(A-1)}\end{aligned}$$

For the other direction,

$$\begin{aligned}\{x : c_1 \le \widehat{f}_n(x) \le c_2\} \quad &\subset \quad \{x : f(x) \le c_2 + \eta_n\} \\ &\subset \quad \{x : f(x) \le c_2\}^{\varepsilon_n}.\end{aligned}$$

A similar argument shows that $\{x : c_1 \le \widehat{f}_n \le c_2\} \subset \{x : c_1 \le f(x)\}^{\varepsilon_n}$, from which it follows

$$\{x : c_1 \le \widehat{f}_n(x) \le c_2\} \quad \subset \quad \{x : c_1 \le f(x) \le c_2\}^{\varepsilon_n}$$

by Lemma 6.1. Together with (A-1) this proves the result.

To address necessity, suppose that there exists a neighbourhood of x_0, $B_{\delta}(x_0)$, and a subsequence n_k such that $\widehat{f}_{n_k}(x) < f(x)$ for all $x \in B_{\delta}(x_0)$. Assume also that $x_0 \in \{x : c_1 \le f(x)\} \setminus \overline{\{x : c_1 < f(x)\}}$. In particular, this implies that (2.2) is not satisfied, and hence there exists an $\varepsilon > 0$ such that $\rho(x_0, \overline{\{x : c_1 < f(x) \le c_2\}}) > \varepsilon$. It follows that $\rho(\widehat{F}_{n_k}(c_1,c_2), F(c_1,c_2)) > \min(\varepsilon,\delta) > 0$, proving the result. A similar argument proves the other claim. □

ACKNOWLEDGMENTS

HJ would like to thank NSERC for funding support.

REFERENCES

[1] B. N. Bekele and P. F. Thall. Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *J. Amer. Statist. Assoc.*, 99(465): 26–35, 2004.

[2] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons Inc., New York, 1968.

[3] W. Carter, V. Chinchilli, J. Wilson, E. Campbell, F. Kessler, and R. Carchman. An asymptotic confidence region for the ED100p from the logistic response surface for a combination of agents. *The American Statistician*, 40(2): 124–128, 1986.

[4] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Aust. N. Z. J. Stat.*, 48(1): 7–19, 2006.

[5] A. Davison and H. D. *Bootstrap methods and their application*. Cambridge University Press, 1997.

[6] H. Jankowski, X. Ji, and L. Stanberry. A random set approach to confidence regions with applications to the effective dose with combinations of agents. *Statistics in Medicine*, 2014. To appear.

[7] H. Jankowski, V. Sabelnykova, and J. Sheriff. Estimating the wintering location of the wood thrush. Technical report, York University, 2011.

[8] H. Jankowski and L. Stanberry. Confidence regions for means of random sets using oriented distance functions. *Scandinavian Journal of Statistics*, 39(2): 340–357, 2012.

[9] H. K. Jankowski and L. I. Stanberry. Expectations of random sets and their boundaries using oriented distance functions. *Journal of Mathematical Imaging and Vision*, 36(3): 291–303, 2010.

[10] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

[11] H. Kunita. *Stochastic flows and stochastic differential equations*, volume 24 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1990.

[12] J. Li, E. Nordheim, C. Zhang, and C. Lehner. Estimation and confidence regions for multidimensional effective dose. *Biometrical Journal*, 50(1): 110–122, 2008.

[13] E. Mammen and W. Polonik. Confidence sets for level sets. *Journal of Multivariate Analysis*, 122: 202–214, 2013.

[14] C. Mass, S. Joslyn, J. Pyle, P. Tewson, T. Gneiting, A. Raftery, J. Baars, J. M. Sloughter, D. Jones, and C. Fraley. PROBCAST: A web-based portal to mesoscale probabilistic forecasts. *Bulletin of the American Meteorological Society*, 90: 1009–1014, 2009.

[15] G. Matheron. *Random sets and integral geometry*. John Wiley & Sons, New York-London-Sydney, 1975. With a foreword by Geoffrey S. Watson, Wiley Series in Probability and Mathematical Statistics.

[16] I. S. Molchanov. A limit theorem for solutions of inequalities. *Scandinavian Journal of Statistics*, 25: 235–242, 1998.

[17] P. Nguyen, P. Brown, and S. J.E. Mapping cancer risk in Southwestern Ontario with changing census boundaries. *Biometrics*, 68, 2012.

[18] D. Nolan. The excess-mass ellipsoid. *J. Multivariate Anal.*, 39(2): 348–371, 1991.

[19] W. Polonik. Measuring mass concentrations and estimating density contour clusters—an

excess mass approach. *Ann. Statist.*, 23(3): 855–881, 1995.

[20] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`.

[21] H. Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40: 87–104, 1953.

[22] J. Serra. *Image analysis and mathematical morphology*. Academic Press Inc., London, 1984.

[23] P. F. Thall and J. D. Cook. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, 60(3): 684–693, 2004.

[24] P. F. Thall, R. E. Millikan, P. Mueller, and S.-J. Lee. Dose-finding with two agents in Phase I oncology trials. *Biometrics*, 59(3): 487–496, 2003.

[25] S. R. S. Varadhan. *Stochastic processes*, volume 16 of *Courant Lecture Notes in Mathematics*. Courant Institute of Mathematical Sciences, New York, 2007.

ABOUT THE AUTHORS

1. Hanna Jankowski is Associate Professor in the Department of Mathematics and Statistics at York University in Toronto, Canada.

2. Larissa Stanberry is Biostatistician at Predictive Analytics, Seattle Children's Hospital, Seattle, USA.