



## Distribution-Free Prediction Sets

Jing Lei , James Robins & Larry Wasserman

**To cite this article:** Jing Lei , James Robins & Larry Wasserman (2013) Distribution-Free Prediction Sets, Journal of the American Statistical Association, 108:501, 278-287, DOI: [10.1080/01621459.2012.751873](https://doi.org/10.1080/01621459.2012.751873)

**To link to this article:** <http://dx.doi.org/10.1080/01621459.2012.751873>



Accepted author version posted online: 21 Dec 2012.



Submit your article to this journal [↗](#)



Article views: 1161



View related articles [↗](#)

# Distribution-Free Prediction Sets

Jing LEI, James ROBINS, and Larry WASSERMAN

This article introduces a new approach to prediction by bringing together two different nonparametric ideas: distribution-free inference and nonparametric smoothing. Specifically, we consider the problem of constructing nonparametric tolerance/prediction sets. We start from the general conformal prediction approach, and we use a kernel density estimator as a measure of agreement between a sample point and the underlying distribution. The resulting prediction set is shown to be closely related to plug-in density level sets with carefully chosen cutoff values. Under standard smoothness conditions, we get an asymptotic efficiency result that is near optimal for a wide range of function classes. But the coverage is guaranteed whether or not the smoothness conditions hold and regardless of the sample size. The performance of our method is investigated through simulation studies and illustrated in a real data example.

KEY WORDS: Conformal prediction; Consistency; Density level sets; Finite sample; Kernel density.

## 1. INTRODUCTION

### 1.1 Prediction Sets and Density Level Sets

Suppose we observe independent and identically distributed (iid) data  $Y_1, \dots, Y_n \in \mathbb{R}^d$  from a distribution  $P$ . Our goal is to construct a prediction set  $C_n = C_n(Y_1, \dots, Y_n) \subseteq \mathbb{R}^d$  such that

$$\mathbb{P}(Y_{n+1} \in C_n) \geq 1 - \alpha \quad (1)$$

for a fixed  $0 < \alpha < 1$ , where  $\mathbb{P} = P^{n+1}$  is the product probability measure over the  $(n+1)$ -tuple  $(Y_1, \dots, Y_{n+1})$ . In general, we let  $\mathbb{P}$  denote  $P^n$  or  $P^{n+1}$  depending on the context.

The prediction set problem has a natural connection to density level sets and density-based clustering. Given a random sample from a distribution, it is often of interest to ask where most of the probability mass is concentrated. A natural answer to this question is the density level set  $L(t) = \{y \in \mathbb{R}^d : p(y) \geq t\}$ , where  $p$  is the density function of  $P$ . When the distribution  $P$  is multimodal, a suitably chosen  $t$  will give a clustering of the underlying distribution (Hartigan 1975). When  $t$  is given, consistent estimators of  $L(t)$  and rates of convergence have been studied in detail (Polonik 1995; Tsybakov 1997; Baillo, Cuestas-Alberto, and Cuevas 2001; Baillo 2003; Cadre 2006; Willett and Nowak 2007; Rigollet and Vert 2009; Rinaldo and Wasserman 2010). It often makes sense to define  $t$  implicitly using the desired probability coverage  $(1 - \alpha)$ :

$$t(\alpha) = \inf \{t : P(L(t)) \geq 1 - \alpha\}. \quad (2)$$

Let  $\mu(\cdot)$  denote the Lebesgue measure on  $\mathbb{R}^d$ . If the contour  $\{y : p(y) = t(\alpha)\}$  has zero Lebesgue measure, then it is easily shown that

$$C^{(\alpha)} := L(t(\alpha)) = \arg \min_C \mu(C), \quad (3)$$

where the min is over  $\{C : P(C) \geq 1 - \alpha\}$ . Therefore, the density-based clustering problem can sometimes be formulated as estimation of the minimum volume prediction set.

The study of prediction sets has a long history in statistics under various names such as “tolerance regions” and “minimum volume sets;” see, for example, Wilks (1941), Wald (1943), Fraser and Guttman (1956), Guttman (1970), Aichison and Dunsmore (1975), Chatterjee and Patra (1980), Di Bucchianico, Einmahl, and Mushkudiani (2001), Cadre (2006), and Li and Liu (2008). Also related is the notion of quantile contours (Wei 2008). In this article, we study a newer method due to Vovk, Gammerman, and Shafer (2005), which we describe in Section 2.

### 1.2 Main Results

Let  $C_n$  be a prediction set. There are two natural criteria to measure its quality: *validity* and *efficiency*. By validity, we mean that  $C_n$  has the desired coverage for all  $P$  [e.g., in the sense of (1)]. We measure the efficiency of  $C_n$  in terms of its closeness to the optimal (oracle) set  $C^{(\alpha)}$ . Since  $p$  is unknown,  $C^{(\alpha)}$  cannot be used as an estimator but only as a benchmark in evaluating the efficiency. We define the loss function of  $C_n$  by

$$R(C_n) = \mu(C_n \Delta C^{(\alpha)}), \quad (4)$$

where  $\Delta$  denotes the symmetric set difference. We say that  $C_n$  is *efficient at rate  $r_n$*  for a class of distributions  $\mathcal{P}$  if, for every  $P \in \mathcal{P}$ ,  $\mathbb{P}(R(C_n) \geq r_n) \rightarrow 0$  as  $n \rightarrow \infty$ . Such loss functions have been used, for example, by Chatterjee and Patra (1980) and Li and Liu (2008) in nonparametric prediction set estimation and by Tsybakov (1997) and Rigollet and Vert (2009) in density level set estimation.

In this article, we construct  $C_n$  with the following properties.

1. Finite sample validity:  $C_n$  satisfies (1) for all  $P$  and  $n$  under *no* assumption other than iid.
2. Asymptotic efficiency:  $C_n$  is efficient at rate  $(\log n/n)^{c_{p,\alpha}}$  for some constant  $c_{p,\alpha} > 0$  depending only on the smoothness of  $p$ .
3. For any  $y \in \mathbb{R}^d$ , the computational cost of evaluating  $\mathbf{1}(y \in C_n)$  is linear in  $n$ .

Our prediction set is obtained by combining the idea of conformal prediction (Vovk, Gammerman, and Shafer 2005) with

Jing Lei is Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: [jinglei@andrew.cmu.edu](mailto:jinglei@andrew.cmu.edu)). James Robins is Mitchell L. and Robin LaFoley Dong Professor, Department of Epidemiology and Department of Biostatistics, Harvard University, Boston, MA 02115 (E-mail: [robins@hsph.harvard.edu](mailto:robins@hsph.harvard.edu)). Larry Wasserman is Professor, Department of Statistics and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: [larry@stat.cmu.edu](mailto:larry@stat.cmu.edu)). Jing Lei's work is supported by the National Science Foundation grant BCS-0941518 and Larry Wasserman's work by the National Science Foundation grant DMS-0806009 and Air Force grant FA95500910373.

density estimation. We show that such a set, whose analytical form may be intractable, is sandwiched by two kernel density level sets with carefully tuned cutoff values. Therefore, the efficiency of the conformal prediction set can be approximated by those of the two kernel density level sets. As a by-product, we obtain a kernel density level set that always contains the conformal prediction set and satisfies finite sample validity as well as asymptotic efficiency. In the efficiency argument, we refine the rates of convergence for plug-in density level sets at implicitly defined levels first developed in Cadre (2006) and Cadre, Pelletier, and Pudlo (2009), which may be of independent interest. We remark that, while the method gives valid prediction regions in any dimension, the efficiency of the region can be poor in higher dimensions.

### 1.3 Related Work

The *conformal prediction* method (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) is a general approach for constructing distribution-free, sequential prediction sets using exchangeability and is usually applied to sequential classification and regression problems (Vovk, Nouretdinov, and Gammerman 2009). We show that one can adapt the method to the prediction task described in (1). We describe this general method in Section 2 and our adaptation in Section 3.

In multivariate prediction set estimation, common approaches include methods based on statistically equivalent blocks (Tukey 1947; Li and Liu 2008) and plug-in density level sets (Chatterjee and Patra 1980; Hyndman 1996; Cadre 2006). In the former, an ordering function taking values in  $\mathbb{R}^1$  is used to order the data points. Then one-dimensional tolerance interval methods (e.g., Wilks 1941) can be applied. Such methods usually give accurate coverage but efficiency is hard to prove. Li and Liu (2008) proposed an estimator, with a high computational cost, using the multivariate spacing depth as the ordering function. Consistency is only proved when the level sets are convex. On the other hand, the plug-in methods (Chatterjee and Patra 1980) give provable validity and efficiency in an asymptotic sense regardless of the shape of the distribution, with a much easier implementation. As mentioned earlier, our estimator can be approximated by plug-in level sets, which are similar to those introduced in Chatterjee and Patra (1980), Hyndman (1996), Cadre (2006), and Park, Huang, and Ding (2010). However, these methods do not give finite sample validity.

Other important work on estimating tolerance regions and minimum volume prediction sets includes Polonik (1997), Walther (1997), Di Bucchianico, Einmahl, and Mushkudiani (2001), and Scott and Nowak (2006). Scott and Nowak (2006) did have finite sample results but did not have the guarantee given in Equation (1), which is the focus of this article. Bandwidth selection for level sets is discussed in Samworth and Wand (2010). There is also a literature on anomaly detection that amounts to constructing prediction sets. Recent advances in this area include Zhao and Saligrama (2009), Sricharan and Hero (2011), and Steinwart, Hush, and Scovel (2005).

In Section 2, we introduce conformal prediction. In Section 3, we describe a construction of prediction sets by combining conformal prediction with kernel density estimators. The approximation result (sandwich lemma) and asymptotic properties are

also discussed. A method for choosing the bandwidth is given in Section 4. Simulation and a real data example are presented in Section 5. Some technical proofs are given in the Appendix.

## 2. CONFORMAL PREDICTION

Let  $Y_1, \dots, Y_n$  be a random sample from  $P$  and let  $\mathbf{Y} = (Y_1, \dots, Y_n)$ . Fix some  $y \in \mathbb{R}^d$  and let us tentatively set  $Y_{n+1} = y$ . Let  $\sigma_i = \sigma(\{Y_1, \dots, Y_{n+1}\}, Y_i)$  be a “conformity score” that measures how similar  $Y_i$  is to  $\{Y_1, \dots, Y_{n+1}\}$ . We only require that  $\sigma$  be symmetric in the entries of its first argument. We test the hypothesis  $H_0 : Y_{n+1} = y$  by computing the  $p$ -value

$$\pi_n(y) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}[\sigma_j \leq \sigma_{n+1}].$$

By symmetry, under  $H_0$ , the ranks of the  $\sigma_i$  are uniformly distributed among  $\{1/(n+1), 2/(n+1), \dots, 1\}$  and hence for any  $\alpha \in (0, 1)$ , we have  $\mathbb{P}(\pi_n(y) \leq \tilde{\alpha}) \leq \alpha$  where  $\tilde{\alpha} = \lfloor (n+1)\alpha \rfloor / (n+1) \approx \alpha$ . Let

$$\widehat{C}^{(\alpha)}(Y_1, \dots, Y_n) = \{y : \pi_n(y) \geq \tilde{\alpha}\}. \quad (5)$$

It follows that under  $H_0$ , we have  $\mathbb{P}[Y_{n+1} \in \widehat{C}^{(\alpha)}(Y_1, \dots, Y_n)] \geq 1 - \alpha$ . Based on the above discussion, any conformity measure  $\sigma$  can be used to construct prediction sets with finite sample validity, with no assumptions on  $P$ . The only requirement is exchangeability of the data. In this article, we will use  $\sigma_i = \widehat{p}(Y_i)$  where  $\widehat{p}$  is an appropriate density estimator.

## 3. CONFORMAL PREDICTION WITH KERNEL DENSITY

### 3.1 The Method

For a given bandwidth  $h_n$  and kernel function  $K$ , let

$$\widehat{p}_n(u) = \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{u - Y_i}{h_n}\right) \quad (6)$$

be the usual kernel density estimator. For now, we focus on a given bandwidth  $h_n$ . The theoretical and practical aspects of choosing  $h_n$  will be discussed in Sections 3.3 and 4, respectively. For any given  $y \in \mathbb{R}^d$ , let  $Y_{n+1} = y$  and define the augmented density estimator

$$\begin{aligned} \widehat{p}_n^y(u) &= \frac{1}{h_n^d(n+1)} \sum_{i=1}^{n+1} K\left(\frac{u - Y_i}{h_n}\right) = \left(\frac{n}{n+1}\right) \widehat{p}_n(u) \\ &\quad + \frac{1}{h_n^d(n+1)} K\left(\frac{u - y}{h_n}\right). \end{aligned} \quad (7)$$

Now we use the conformity measure  $\sigma_i = \widehat{p}_n^y(Y_i)$  and the  $p$ -value becomes

$$\pi_n(y) := \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}[\widehat{p}_n^y(Y_i) \leq \widehat{p}_n^y(y)].$$

The resulting prediction set is  $\widehat{C}^{(\alpha)} = \{y : \pi_n(y) \geq \tilde{\alpha}\}$ . It follows that  $\mathbb{P}[Y_{n+1} \in \widehat{C}^{(\alpha)}] \geq 1 - \alpha$  for all  $P$  and all  $n$  as required.

Figure 1 shows a one-dimensional example of the procedure. The top left plot shows a histogram of some data of sample size 20 from a two-component Gaussian mixture. The next three

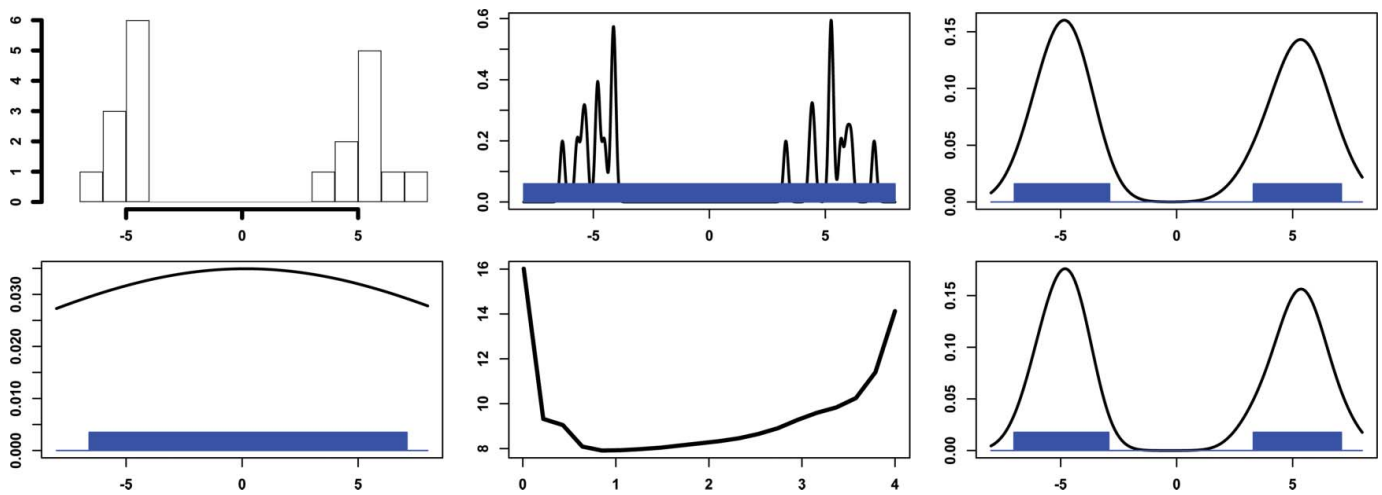


Figure 1. Top left: histogram of some data. Top middle, top right, and bottom left show three kernel density estimators and the corresponding conformal prediction sets with bandwidth 0.1, 1, and 10, respectively. Bottom middle: Lebesgue measure as a function of bandwidth. Bottom right: estimator and prediction set obtained from the bandwidth with smallest prediction set.

plots (top middle, top right, bottom left) show three kernel density estimators with increasing bandwidths as well as the conformal prediction sets derived from these estimators with  $\alpha = 0.05$ . Every bandwidth leads to a valid set, but undersmoothing and oversmoothing lead to larger sets. The bottom middle plot shows the Lebesgue measure of the set as a function of bandwidth. The bottom right plot shows the estimator and prediction set based on the bandwidth whose corresponding conformal prediction set has the minimal Lebesgue measure.

### 3.2 An Approximation

The conformal prediction set is expensive to compute since we have to compute  $\pi_n(y)$  for every  $y \in \mathbb{R}^d$ . Here we derive an approximation to  $\widehat{C}^{(\alpha)}$  that can be computed quickly and maintains finite sample validity. Define the upper and lower level sets of density  $p$  at level  $t$ , respectively:

$$L(t) = \{y : p(y) \geq t\}, \quad \text{and} \quad L^\ell(t) = \{y : p(y) \leq t\}. \quad (8)$$

The corresponding level sets of  $\widehat{p}_n$  are denoted by  $L_n(t)$  and  $L_n^\ell(t)$ , respectively. Let  $Y_{(1)}, \dots, Y_{(n)}$  be the reordered data so that  $\widehat{p}_n(Y_{(1)}) \leq \dots \leq \widehat{p}_n(Y_{(n)})$ . Let  $i_{n,\alpha} = \lfloor (n+1)\alpha \rfloor$ , and define the inner and outer sandwiching sets:

$$L_n^- = L_n(\widehat{p}_n(Y_{(i_{n,\alpha})})), \quad L_n^+ = L_n(\widehat{p}_n(Y_{(i_{n,\alpha})}) - (nh^d)^{-1}\psi_K),$$

where  $\psi_K = \sup_{u,u'} |K(u) - K(u')|$ . Then we have the following “sandwich” lemma, whose proof can be found in Appendix B.

**Lemma 3.1.** (Sandwich Lemma). Let  $\widehat{C}^{(\alpha)}$  be the conformal prediction set based on the kernel density estimator. Assume that  $\sup_u |K(u)| = K(0)$ . Then

$$L_n^- \subseteq \widehat{C}^{(\alpha)} \subseteq L_n^+. \quad (9)$$

According to the sandwich lemma,  $L_n^+$  also guarantees distribution-free finite sample coverage and is easier to analyze. Moreover, it is much faster to compute since it avoids ever having to compute the kernel density estimator based on the augmented data. The inner set,  $L_n^-$ , which is used as an estimate of  $C^{(\alpha)}$  in related work such as in Chatterjee and Patra (1980);

Hyndman (1996); and Cadre, Pelletier, and Pudlo (2009), generally does not have finite sample validity. We confirm this through simulations in Section 5. Next we investigate the efficiency of these prediction sets.

### 3.3 Asymptotic Properties

The inner and outer sandwiching sets  $L_n^-$  and  $L_n^+$  are plug-in estimators of density level sets of the form  $L_n(t_n^{(\alpha)}) = \{y : \widehat{p}_n(y) \geq t_n^{(\alpha)}\}$ , where  $t_n^{(\alpha)} = \widehat{p}_n(Y_{(i_{n,\alpha})})$  for the inner set  $L_n^-$  and  $t_n^{(\alpha)} = \widehat{p}_n(Y_{(i_{n,\alpha})}) - (nh_n^d)^{-1}\psi_K$  for the outer set  $L_n^+$ . Here we can view  $t_n^{(\alpha)}$  as an estimate of  $t^{(\alpha)}$ . In Cadre, Pelletier, and Pudlo (2009), it is shown that, under regularity conditions of the density  $p$ , the plug-in estimators  $t_n^{(\alpha)}$  and  $L_n(t_n^{(\alpha)})$  are consistent with convergence rate  $1/\sqrt{nh_n^d}$  for a range of  $h_n$ . Here we refine the results under more general conditions. We note that similar convergence rates for plug-in density level sets with a fixed and known level are obtained in Rigollet and Vert (2009). The extension to unknown levels is nontrivial and needs slightly stronger regularity conditions.

Intuitively speaking, the plug-in density level set  $L_n(t_n^{(\alpha)})$  is an accurate estimator of  $L(t^{(\alpha)})$  if  $\widehat{p}_n$  and  $t_n^{(\alpha)}$  are accurate estimators of  $p$  and  $t^{(\alpha)}$ , and  $p$  is not too flat at level  $t^{(\alpha)}$ . The following smoothness condition is assumed for  $p$  and  $K$  to ensure accurate density estimation.

- A1.** The density  $p$  is Hölder smooth of order  $\beta$ , with  $\beta > 0$ , and  $K$  is a valid kernel of order  $\beta$ .

Hölder smoothness and valid kernels are standard assumptions for nonparametric density estimation. We give their definitions in Appendix A.

**Remark 3.2.** Assumption A1 can be relaxed in a similar way as in Rigollet and Vert (2009). The idea is that we only need to estimate the density very accurately in a neighborhood of  $\partial C^{(\alpha)}$  (the boundary of the optimal set). Therefore, it would be sufficient to have the strong  $\beta$ -Hölder smoothness condition near  $\partial C^{(\alpha)}$ , together with a weaker  $\beta'$ -Hölder smoothness condition ( $\beta' \leq \beta$ ) everywhere else. For presentation simplicity, we stick with the global smoothness condition in A1.



To control the regularity of  $p$  at level  $t^{(\alpha)}$ , a common assumption is the  $\gamma$ -exponent condition, which was first introduced by Polonik (1995) and has been used by many others (see, e.g., Tsybakov 1997 and Rigollet and Vert 2009). In our argument, such an assumption is also related to estimating  $t^{(\alpha)}$  itself. Specifically, we assume

**A2.** There exist constants  $0 < c_1 \leq c_2$  and  $\epsilon_0 > 0$  such that

$$\begin{aligned} c_1|\epsilon|^\gamma &\leq |P(\{y : p(y) \leq t^{(\alpha)} + \epsilon\}) - \alpha| \\ &\leq c_2|\epsilon|^\gamma, \quad \forall -\epsilon_0 \leq \epsilon \leq \epsilon_0. \end{aligned} \quad (10)$$

The gamma exponent condition requires that the density to be neither flat (for stability of level set) nor steep (for accuracy of  $t_n^{(\alpha)}$ ). As indicated in Audibert and Tsybakov (2007), A1 and A2 cannot hold simultaneously unless  $\gamma(1 \wedge \beta) \leq 1$ . In the common case  $\gamma = 1$ , this always holds.

Assumptions A1 and A2 extend those in Cadre et al. (2009), where  $\beta = \gamma = 1$  is considered. The next theorem states the quality of cutoff values used in the sandwiching sets  $L_n^-$  and  $L_n^+$ .

**Theorem 3.3.** Let  $t_n^{(\alpha)} = \hat{p}_n(Y_{(i_{n,\alpha})})$ , where  $\hat{p}_n$  is the kernel density estimator given by Equation (6), and  $Y_{(i)}$  and  $i_{n,\alpha}$  are defined as in Section 3.2. Assume that A1–A2 hold and choose  $h_n \asymp (\log n/n)^{1/(2\beta+d)}$ . Then for any  $\lambda > 0$ , there exist constants  $A_\lambda, A'_\lambda$  depending only on  $p, K$ , and  $\alpha$ , such that

$$\mathbb{P}\left(|t_n^{(\alpha)} - t^{(\alpha)}| \geq A_\lambda \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + A'_\lambda \left(\frac{\log n}{n}\right)^{\frac{1}{2\gamma}}\right) = O(n^{-\lambda}). \quad (11)$$

We give the proof of Theorem 3.3 in Appendix C. Theorem 3.3 is useful for establishing the convergence of the corresponding level set. Observing that  $(nh_n^d)^{-1} = o((\log n/n)^{\beta/(2\beta+d)})$ , it follows immediately that the cutoff value used in  $L_n^+$  also satisfies (11). The next theorem, proved in Appendix C, gives the rate of convergence for our estimators.

**Theorem 3.4.** Under same conditions as in Theorem 3.3, for any  $\lambda > 0$ , there exist constants  $B_\lambda, B'_\lambda$  depending on  $p, K$ , and  $\alpha$  only, such that, for all  $\hat{C} \in \{\hat{C}^{(\alpha)}, L_n^-, L_n^+\}$ ,

$$\mathbb{P}\left(\mu(\hat{C} \Delta C^{(\alpha)}) \geq B_\lambda \left(\frac{\log n}{n}\right)^{\frac{\beta\gamma}{2\beta+d}} + B'_\lambda \left(\frac{\log n}{n}\right)^{\frac{1}{2}}\right) = O(n^{-\lambda}). \quad (12)$$

**Remark 3.5.** In the most common cases  $\gamma = 1$ , or  $\beta \geq 1/2$ ,  $\gamma\beta \leq 1$ , the term  $(\log n/n)^{\beta\gamma/(2\beta+d)}$  dominates the convergence rate. It matches the minimax risk rate of the plug-in density level set at a known level developed by Rigollet and Vert (2009). As a result, not knowing the cutoff value  $t^{(\alpha)}$  does not change the difficulty of estimation. When  $\beta\gamma/(2\beta+d) > 1/2$ , the rate is dominated by  $(\log n/n)^{1/2}$  and does not agree with the known minimax lower bound and we do not know if the  $\sqrt{\log n/n}$  can be eliminated from the result.

**Remark 3.6.** The theorems above were stated for the optimal choice of bandwidth. The method is still consistent with similar arguments whenever  $nh_n^d/\log n \rightarrow \infty$  and  $h_n \rightarrow 0$ , although the resulting rates will no longer be optimal.

**Remark 3.7.** The same conclusions in Theorems 3.3 and 3.4 hold under a weaker version of Assumption A1. To make this idea more precise, suppose the density function is only  $\beta$ -Hölder smooth in a neighborhood of the level set contour  $\{y : p(y) = t^{(\alpha)}\}$ , but less smooth everywhere else. Then the same proofs of Theorems 3.3 and 3.4 can be used to obtain a slower rate of convergence. After establishing this first consistency result, one can apply the argument again, with the analysis confined in the smooth neighborhood, to obtain the desired rate of convergence. However, in the interest of space and clarity, we will prove our results only under the more restrictive smoothness assumptions that we have stated.

#### 4. CHOOSING THE BANDWIDTH

As illustrated in Figure 1, the efficiency of  $\hat{C}^{(\alpha)}$  depends on the choice of  $h_n$ . The size of estimated prediction sets can be very large if the bandwidth is either too large or too small. Therefore, in practice, it is desirable to choose a good bandwidth in an automatic and data-driven manner. In kernel density estimation, the choice of bandwidth has been one of the most important topics and many approaches have been studied; see Loader (1999), Mammen et al. (2011), Samworth and Wand (2010), and references therein. Here we consider choosing the bandwidth by minimizing the volume of the conformal prediction set.

Let  $\mathcal{H} = \{h_1, \dots, h_m\}$  be a grid of candidate bandwidths. We compute the prediction set for each  $h \in \mathcal{H}$  and choose the one with the smallest volume. To preserve finite sample validity, we use sample splitting as described in Algorithm 1. We state the following result and omit its proof.

**Proposition 4.1.** If  $\hat{C}$  satisfies finite sample validity for all  $h$ , then  $\hat{C}_{\hat{h},2}$ , the output of the sample splitting tuning algorithm, also satisfies finite sample validity.

---

##### Algorithm 1: Tuning With Sample Splitting

---

Input: sample  $\mathbf{Y} = (Y_1, \dots, Y_n)$ , prediction set estimator  $\hat{C}$ , level  $\alpha$ , and candidate set  $\mathcal{H}$

1. Split the sample randomly into two equal-sized subsamples,  $\mathbf{Y}_1$  and  $\mathbf{Y}_2$ .
  2. Construct prediction sets  $\{\hat{C}_{h,1} : h \in \mathcal{H}\}$  each at level  $1 - \alpha$ , using subsample  $\mathbf{Y}_1$ .
  3. Let  $\hat{h} = \arg \min_h \mu(\hat{C}_{h,1})$ .
  4. Return  $\hat{C}_{\hat{h},2}$ , which is constructed using bandwidth  $\hat{h}$  and subsample  $\mathbf{Y}_2$ .
- 

There are two justifications for choosing a bandwidth to make  $\mu(\hat{C}_h)$  small. The first is pragmatic: in making predictions, it seems desirable to have a small prediction set. The second reason is that minimizing  $\mu(C)$  can potentially lead to good risk properties in terms of the loss  $\mu(C \Delta C^{(\alpha)})$  as we now show. Recall that  $R(C) = \mu(C \Delta C^{(\alpha)})$  and define  $\mathcal{E}(C) = \mu(C) - \mu(C^{(\alpha)})$ . To avoid technical complications, we will assume in this section that the sample space is compact and focus on the simple case  $\gamma = 1$  in condition A2.

**Lemma 4.2.** Let  $\hat{C}$  be an estimator of  $C^{(\alpha)}$ . Then  $\mathcal{E}(\hat{C}) \leq R(\hat{C})$ . Furthermore, if  $\hat{C}$  is finite sample valid and A2

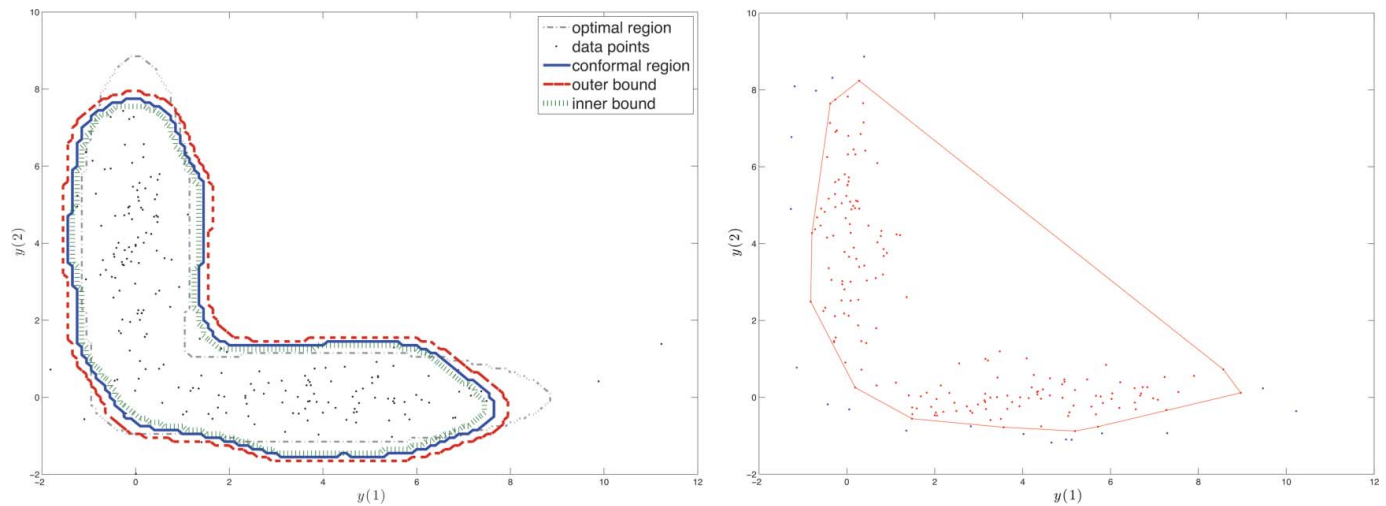


Figure 2. Conformal prediction set (left) and the convex hull of the multivariate spacing depth-based tolerance set (right), with data from a two-component Gaussian mixture. The online version of this figure is in color.

holds with  $\gamma = 1$ , then  $\mathbb{E}(R(\hat{C})) \leq c_1 [\mathbb{E}(\mathcal{E}(\hat{C}))]^{1/2}$  for some constant  $c_1$ .

The bandwidth selection algorithm makes  $\mathcal{E}(\hat{C})$  small. The lemma gives us at least some assurance that making  $\mathcal{E}(\hat{C})$  small will help to make  $R(\hat{C})$  small. The proof of Lemma 4.2 is given in Appendix D. (A similar result can be found in Scott and Nowak (2006).) However, it is an open question whether  $R(\hat{C})$  achieves the minimax rate.

## 5. NUMERICAL EXAMPLES

We first consider simulations on Gaussian mixtures and double-exponential mixtures in two and three dimensions. We apply the bandwidth selector presented in Section 4 to both  $\hat{C}^{(\alpha)}$  and  $L_n^+$ . The bandwidth used for  $L_n^-$  is the same as that for  $L_n^+$ . Therefore, in the results, it is possible to see if  $L_n^-$  is bigger than  $\hat{C}^{(\alpha)}$ , or if  $\hat{C}^{(\alpha)}$  is bigger than  $L_n^+$  because of different bandwidths and data splitting.

### 5.1 2D Gaussian Mixture

We first consider a two-component Gaussian mixture in  $\mathbb{R}^2$ . The first component has mean  $(\sqrt{2\log n} - 2, 0)$  and variance  $\text{diag}(4, 1/4)$ , and the second component has mean  $(0, \sqrt{2\log n} - 2)$  and variance  $\text{diag}(1/4, 4)$  (see Figure 2). This choice of component centers is to make a moderate overlap be-

tween the data clouds from the two components. It makes the prediction set problem more challenging.

Table 1 shows the coverage and Lebesgue measure of the prediction set at level 0.9 ( $\alpha = 0.1$ ) over 100 repetitions. The coverage is excellent and the size of the set is close to optimal. Both the conformal set  $\hat{C}^{(\alpha)}$  and the outer sandwiching set  $L_n^+$  give correct coverage regardless of the sample size. It is worth noting that the inner sandwiching set  $L_n^-$  (corresponding to the method in Hyndman (1996) and Park, Huang, and Ding (2010)) does not give the desired coverage, which suggests that decreasing the cutoff value in  $L_n^+$  is not merely an artifact of proof, but a necessary tuning. The observed excess loss also reflects a rate of convergence that supports our theoretical results on the symmetric difference loss. We compare our method with the approach introduced by Zhao and Saligrama (2009) ( $\hat{C}_{ZS}$ ), where the prediction set is constructed by ranking the distances from each data point to its  $k$ th nearest neighbor. It has been reported that the choice of  $k$  is not crucial and we use  $k = 6$ . (We remark further on the choice of  $k$  at the end of this section.) This method is similar to ours but does not have finite sample validity. We observe that the finite sample coverage of  $\hat{C}_{ZS}$  is less than the nominal level.

Figure 2 shows a typical realization of the estimators. In both panels, the dots are data points when  $n = 200$ . The left panel shows the conformal prediction set with sample splitting (blue solid curve), together with the inner and outer sandwiching sets (red-dashed and green-dotted curves, respectively). Also plotted

Table 1. The simulation results for 2D Gaussian mixture with  $\alpha = 0.1$  over 100 repetitions (mean and one standard deviation). The Lebesgue measure of the ideal set  $\approx 28.02$

	Coverage			Lebesgue measure		
	$n = 100$	$n = 200$	$n = 1000$	$n = 100$	$n = 200$	$n = 1000$
$\hat{C}^{(\alpha)}$	$0.886 \pm 0.005$	$0.897 \pm 0.002$	$0.900 \pm 0.001$	$35.6 \pm 0.7$	$34.3 \pm 0.3$	$31.1 \pm 0.2$
$L_n^-$	$0.861 \pm 0.004$	$0.882 \pm 0.001$	$0.896 \pm 0.001$	$29.8 \pm 0.3$	$34.1 \pm 0.2$	$32.2 \pm 0.1$
$L_n^+$	$0.907 \pm 0.003$	$0.900 \pm 0.001$	$0.907 \pm 0.001$	$36.2 \pm 0.4$	$36.9 \pm 0.2$	$34.1 \pm 0.1$
$\hat{C}_{ZS}$	$0.853 \pm 0.004$	$0.867 \pm 0.002$	$0.881 \pm 0.001$	$28.1 \pm 0.4$	$28.2 \pm 0.2$	$28.0 \pm 0.1$

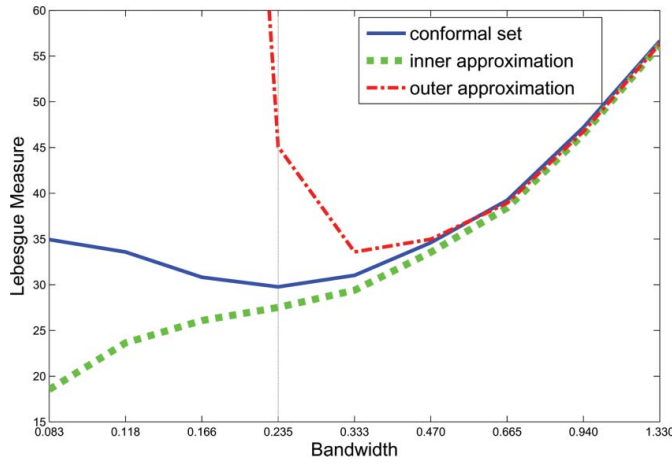


Figure 3. Lebesgue measure of prediction sets versus bandwidth. The online version of this figure is in color.

is the ideal set  $C^{(\alpha)}$  (gray dash-dotted curve). It is clear that all three estimated sets capture the main part of the ideal set, and they are mutually close. On the right panel, we plot a realization of the depth-based approach from Li and Liu (2008). This approach does not require any tuning parameter. However, it takes  $O(n^{d+1})$  time to evaluate  $\mathbf{1}(y \in \hat{C})$  for any single  $y$ . In practice, it is recommended to compute the empirical depth only for all the data points and use the convex hull of all data points with high depth as the estimated prediction set. Such a convex hull construction misses the “L” shape of the ideal set. Moreover, in our implementation, the running time of the kernel density method is much shorter even when  $n = 200$ .

Figure 3 shows the effect of bandwidth on the excess loss  $\mathcal{E}(\hat{C}) = \mu(\hat{C}) - \mu(C^{(\alpha)})$  based on a typical implementation with  $n = 200$ , where the y axis is the Lebesgue measure of the estimated sets. We observe that for the conformal prediction set  $\hat{C}^{(\alpha)}$ , the excess loss is stable for a wide range of bandwidths, especially that moderate undersmoothing does not harm the performance very much. An intuitive explanation is that the data near the contour are dense enough to allow for moderate undersmoothing. Similar phenomena should be expected whenever  $\alpha$  is not too small. Moreover, the selected bandwidth from the outer sandwiching set  $L_n^+$  is close to that obtained from the conformal set. This observation may be of practical interest since it is usually much faster to compute  $L_n^+$ .

*Remark 5.1.* The  $\hat{C}_{ZS}$  method requires a choice of  $k$ . We tried  $k = 2, 3, \dots, 20$ . The coverage increases with  $k$  but does not reach the nominal 0.9 level even when  $k = 20$ . The Lebesgue measure also increases with  $k$  and after  $k = 20$ , it becomes larger than the conformal region.

## 5.2 Further Simulations

We now investigate the performance of our method using distributions with heavier tails and in higher dimensions. These simulations confirm that our method always give finite sample coverage, even when the density estimation is very challenging.

*5.2.1 Double Exponential Distribution.* In this setting, the distribution also has two balanced components. The first component has independent double exponential coordinates:  $Y(1) \sim 2 \text{DoubleExp}(1) + 2.2 \log n$ ,  $Y(2) \sim 0.5 \text{DoubleExp}(1)$ , where  $\text{DoubleExp}(1)$  has density  $\exp(-|y|)/2$ . The second component has the two coordinates switched. The centering at  $2.2 \log n$  is chosen so that there is moderate overlap between data clouds from two components. The results are summarized in Table 2.

*5.2.2 Three-Dimensional Data.* Now we increase the dimension of data. The Gaussian mixture is the same as in the two-dimensional setup, with the third coordinate being an independent Gaussian with mean zero and variance  $1/4$ . The results are summarized in Table 3.

*Remark 5.2.* In the above two simulation settings, the conformal prediction sets are much larger than the ideal (oracle) set unless the sample size is very large ( $n = 1000$ ). This is because of the difficulty of multivariate nonparametric density estimation. In fact, the kernel density estimator may no longer lead to a good conformity score in this case. However, the theory of conformal prediction is still valid as reflected by the coverage. Thus, one may use other conformity scores such as the  $k$ -nearest-neighbor radius, for which a nonconformal version has been reported in Zhao and Saligrama (2009). Other possible choices include Gaussian mixture density estimators and semiparametric models. These extensions will be pursued in future work.

## 5.3 Application to Breast Cancer Data

In this subsection, we apply our method to the Wisconsin Breast Cancer Dataset (available at the University of California, Irvine (UCI) machine learning repository). The dataset contains nine features of 699 patients among which 241 are malignant and 458 are benign. Although this dataset is commonly used to test classification algorithms, it has been used to test prediction region methods in the literature (see, e.g., Park, Huang, and Ding 2010). In this example, we use prediction sets to tell malignant cases from benign ones. Formally, we assume that the benign cases are sampled from a common distribution, and we construct a 95% prediction set corresponding to the high-density region of the underlying distribution. Although the prediction sets are constructed using only the benign cases, the efficiency of the estimated prediction/tolerance set can be measured not only in

Table 2. The simulation results for 2D double exponential mixture with  $\alpha = 0.1$  over 100 repetitions (mean and one standard deviation). The Lebesgue measure of the ideal set  $\approx 55$

	Coverage			Lebesgue measure		
	$n = 100$	$n = 200$	$n = 1000$	$n = 100$	$n = 200$	$n = 1000$
$\hat{C}^{(\alpha)}$	$0.895 \pm 0.005$	$0.916 \pm 0.003$	$0.91 \pm 0.002$	$77.7 \pm 3$	$76.6 \pm 1.6$	$62.3 \pm 0.6$
$L_n^-$	$0.864 \pm 0.006$	$0.897 \pm 0.003$	$0.90 \pm 0.001$	$66.5 \pm 2.3$	$71.7 \pm 1.2$	$58.3 \pm 0.3$
$L_n^+$	$0.893 \pm 0.005$	$0.912 \pm 0.003$	$0.92 \pm 0.001$	$86.1 \pm 7.4$	$78.2 \pm 1.3$	$65.0 \pm 0.4$
$\hat{C}_{ZS}$	$0.871 \pm 0.004$	$0.892 \pm 0.003$	$0.897 \pm 0.001$	$58.2 \pm 1.5$	$60.2 \pm 1.0$	$55.2 \pm 0.4$

Table 3. The simulation results for 3D Gaussian mixture with  $\alpha = 0.1$  over 100 repetitions (mean and one standard deviation). The Lebesgue measure of the ideal set  $\approx 62$

	Coverage			Lebesgue measure		
	$n = 100$	$n = 200$	$n = 1000$	$n = 100$	$n = 200$	$n = 1000$
$\widehat{C}^{(\alpha)}$	$0.917 \pm 0.004$	$0.902 \pm 0.003$	$0.900 \pm 0.002$	$109 \pm 2.4$	$89 \pm 1.5$	$74 \pm 0.7$
$L_n^-$	$0.875 \pm 0.005$	$0.880 \pm 0.003$	$0.889 \pm 0.002$	$109 \pm 2.1$	$98 \pm 1.5$	$81 \pm 0.7$
$L_n^+$	$0.892 \pm 0.004$	$0.898 \pm 0.003$	$0.916 \pm 0.002$	$118 \pm 2.2$	$109 \pm 1.6$	$96 \pm 0.9$
$\widehat{C}_{\text{ZS}}$	$0.869 \pm 0.003$	$0.872 \pm 0.002$	$0.879 \pm 0.001$	$75 \pm 1.3$	$69 \pm 0.8$	$64 \pm 0.4$

terms of its Lebesgue measure, but also in terms of the number of false negatives (i.e., the number of malignant cases covered by the prediction set). Ideally the prediction set shall contain most of benign cases but few malignant cases and hence can be used as a classifier.

In our implementation, the data dimension is reduced to two using standard principal components analysis. Such a dimension reduction simplifies visualization and has also been used in Park, Huang, and Ding (2010). If no dimension reduction is used, the data concentrates near a low dimensional subset of the space, and other conformity scores, such as the  $k$  nearest neighbors radius, can be used instead of kernel density estimation. To test the out-of-sample performance of our method, we randomly choose 100 out of 458 benign cases as testing data. The prediction region is constructed using only the remaining 358 benign cases with coverage level 0.95 and kernel density bandwidth 0.8. We repeat this experiment 100 times. A typical implementation is plotted in Figure 4. In Table 4, we report the mean coverage on the testing data as well as the malignant data. The resulting conformal prediction sets give the desired coverage for the benign cases and low false coverage for the malignant cases. Note that in this case, the inner density level set  $L_n^-$  is equivalent to the method proposed in Park, Huang, and Ding (2010), which in general does not have finite sample validity. In our experiment, the

Table 4. Application to the breast cancer data with  $\alpha = 0.05$  over 100 repetitions. Reported are the mean and one estimated standard deviation of the empirical coverage on the testing benign data and the malignant data

Method	$\widehat{C}^{(\alpha)}$	$L_n^-$	$L_n^+$
Test sample coverage	$0.9514 \pm 0.0012$	$0.9488 \pm 0.0012$	$0.9534 \pm 0.0013$
Malignant data coverage	$0.0141 \pm 0.0002$	$0.0044 \pm 0.0001$	$0.0420 \pm 0.0004$

average out-of-sample coverage is slightly below the nominal level (by about one standard deviation). In this example, we see that the conformal methods ( $\widehat{C}^{(\alpha)}$  and  $L_n^+$ ) give similar empirical performance as the conventional nonconformal method ( $L_n^-$ ), with additional finite sample guarantee.

## APPENDIX A: DEFINITIONS

### A.1 Hölder Smooth Functions

The Hölder class is a popular smoothness condition in nonparametric inferences (Tsybakov 2009, sec. 1.2). Here we use the version given in Rigollet and Vert (2009).

Let  $s = (s_1, \dots, s_d)$  be a  $d$ -tuple of nonnegative integers and  $|s| = s_1 + \dots + s_d$ . For any  $x \in \mathbb{R}^d$ , let  $x^s = x_1^{s_1} \dots x_d^{s_d}$  and  $D^s$  be the differential operator:

$$D^s f = \frac{\partial^{|s|} f}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}(x_1, \dots, x_d).$$

Given  $\beta > 0$ , for any functions  $f$  that are  $\lfloor \beta \rfloor$  times differentiable, denote its Taylor expansion of degree  $\lfloor \beta \rfloor$  at  $x_0$  by

$$f_{x_0}^{(\beta)}(x) = \sum_{|s| \leq \beta} \frac{(x - x_0)^s}{s_1! \dots s_d!} D^s f(x_0).$$

**Definition A.1** (Hölder class). For constants  $\beta > 0$ ,  $L > 0$ , define the Hölder class  $\Sigma(\beta, L)$  to be the set of  $\lfloor \beta \rfloor$ -times differentiable functions on  $\mathbb{R}^d$  such that

$$|f(x) - f_{x_0}^{(\beta)}(x)| \leq L \|x - x_0\|^\beta. \quad (\text{A.1})$$

### A.2 Valid Kernels

A standard condition on the kernel is the notion of  $\beta$ -valid kernels.

**Definition A.2** ( $\beta$ -valid kernel). For any  $\beta > 0$ , function  $K : \mathbb{R}^d \mapsto \mathbb{R}$  is a  $\beta$ -valid kernel if (a)  $K$  is supported on  $[-1, 1]^d$ ; (b)  $\int K = 1$ ; (c)  $\int |K|^r < \infty$ , all  $r \geq 1$ ; and (d)  $\int y^s K(y) dy = 0$  for all  $1 \leq |s| \leq \beta$ .

The last condition is interpreted elementwise. In the literature,  $\beta$ -valid kernels are usually used with Hölder class of functions to derive

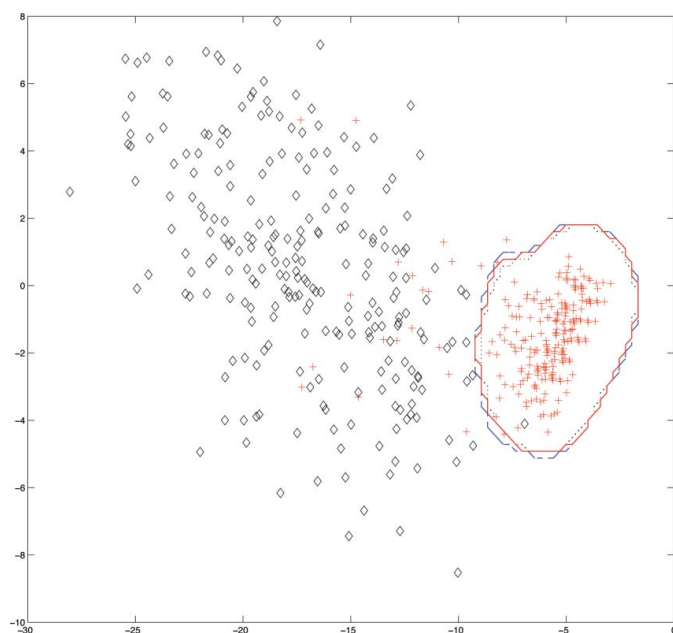


Figure 4. Prediction sets for benign instances. Crosses: benign; diamonds: malignant. Blue dashed curve:  $L_n^+$ ; black dotted curve:  $L_n^-$ ; Red solid curve:  $\widehat{C}^{(\alpha)}$ . The online version of this figure is in color.



fast rates of convergence. The existence of univariate  $\beta$ -valid kernels can be found in section 1.2 of Tsybakov (2009). A multivariate  $\beta$ -valid kernel can be obtained by taking direct product of univariate  $\beta$ -valid kernels.

## APPENDIX B: PROOF OF LEMMA 3.1

*Proof of Lemma 3.1.* Let  $P_n^y = \frac{n}{n+1}P_n + \frac{1}{n+1}\delta_y$ , where  $P_n$  is the empirical distribution defined by the sample  $\mathbf{Y} = (Y_1, \dots, Y_n)$  and  $\delta_y$  is the point mass distribution at  $y$ . Define functions

$$\begin{aligned} G(t) &= P(L^\ell(t)), \\ G_n(t) &= P_n(L_n^\ell(t)) = n^{-1} \sum_{i=1}^n \mathbf{1}(\widehat{p}_n(Y_i) \leq t), \\ G_n^y(t) &= P_n^y(\widehat{p}_n^y(Y) \leq t) \\ &= \frac{1}{n+1} \left( \sum_{i=1}^n \mathbf{1}(\widehat{p}_n^y(Y_i) \leq t) + \mathbf{1}(\widehat{p}_n^y(y) \leq t) \right). \end{aligned}$$

The functions  $G$ ,  $G_n$ , and  $G_n^y$  defined above are the cumulative distribution function (CDF) of  $p(Y)$  and its empirical versions with sample  $\mathbf{Y}$  and  $\text{aug}(\mathbf{Y}, y)$ , respectively, where  $\text{aug}(\mathbf{Y}, y) = (Y_1, \dots, Y_n, y)$ . By (5), the conformal prediction set can be written as

$$\widehat{C}^{(\alpha)} = \{y \in \mathbb{R}^d : G_n^y(\widehat{p}_n^y(y)) \geq \widetilde{\alpha}\}.$$

The proof is based on a direct characterization of  $L_n^-$  and  $L_n^+$ . First, for each  $y \in L_n^-$  and  $i \leq i_{n,\alpha}$ , we have

$$\begin{aligned} \widehat{p}_n^y(y) - \widehat{p}_n^y(Y_{(i)}) &= \frac{n}{n+1} (\widehat{p}_n(y) - \widehat{p}_n(Y_{(i)})) \\ &\quad + \frac{1}{(n+1)h^d} \left( K(0) - K\left(\frac{Y_{(i)} - y}{h}\right) \right) \geq 0. \end{aligned}$$

As a result,  $G_n^y(\widehat{p}_n^y(y)) \geq i_{n,\alpha}/(n+1) = \widetilde{\alpha}$  and hence  $y \in \widehat{C}^{(\alpha)}$ . Similarly, for each  $y \notin L_n^+$  and  $i \geq i_{n,\alpha}$ , we have

$$\begin{aligned} \widehat{p}_n^y(y) - \widehat{p}_n^y(Y_{(i)}) &= \frac{n}{n+1} (\widehat{p}_n(y) - \widehat{p}_n(Y_{(i)})) + \frac{1}{(n+1)h^d} \\ &\quad \times \left( K(0) - K\left(\frac{Y_{(i)} - y}{h}\right) \right) \\ &\leq \frac{n}{n+1} (\widehat{p}_n(y) - \widehat{p}_n(Y_{(i_{n,\alpha})})) + \frac{1}{(n+1)h^d} \psi_K < 0. \end{aligned}$$

Therefore,  $G_n^y(\widehat{p}_n^y(y)) \leq (i_{n,\alpha} - 1)/(n+1) < \widetilde{\alpha}$  and hence  $y \notin \widehat{C}^{(\alpha)}$ .  $\square$

## APPENDIX C: PROOF OF THEOREMS 3.3 AND 3.4

The bias in the estimated cutoff level  $t_n^{(\alpha)}$  can be bounded in terms of two quantities:

$$V_n = \sup_{t>0} |P_n(L^\ell(t)) - P(L^\ell(t))|, \quad R_n = \|\widehat{p}_n - p\|_\infty.$$

Here  $V_n$  can be viewed as the maximum of the empirical process  $P_n - P$  over a nested class of sets and  $R_n$  is the  $L_\infty$  loss of the density estimator. As a result,  $V_n$  can be bounded using the standard empirical process and Vapnik-Chervonenkis (VC) dimension argument and  $R_n$  can be bounded using the smoothness of  $p$  and kernel  $K$  with a suitable choice of bandwidth. Formally, we provide upper bounds for these two quantities through the following lemma.

*Lemma C.1.* Let  $V_n, R_n$  be defined as above, then under Assumptions A1 and A2, for any  $\lambda > 0$ , there exist constants  $A_{1,\lambda}$  and  $A_{2,\lambda}$  depending on  $\lambda$  only, such that

$$\begin{aligned} \mathbb{P}\left(V_n \geq A_{1,\lambda} \sqrt{\frac{\log n}{n}}\right) &= O(n^{-\lambda}), \\ \mathbb{P}\left(R_n \geq A_{2,\lambda} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}\right) &= O(n^{-\lambda}). \end{aligned}$$

*Proof.* First, it is easy to check that the class of sets  $\{L^\ell(t) : t > 0\}$  are nested with VC dimension 2 and hence by classical empirical process theory (see, e.g., van der Vaart and Wellner (1996), sec. 2.14), there exists a constant  $C_0 > 0$  such that for all  $\eta > 0$

$$\mathbb{P}(V_n \geq \eta) \leq C_0 n^2 \exp(-n\eta^2/32). \quad (\text{A.2})$$

Let  $\eta = A\sqrt{\log n/n}$ , we have

$$\mathbb{P}(V_n \geq A\sqrt{\log n/n}) \leq C_0 n^2 \exp(-A^2 \log n/32) = C_0 n^{-(A^2/32-2)}. \quad (\text{A.3})$$

The first result then follows by choosing  $A_{1,\lambda} = \sqrt{32(\lambda+2)}$ . Next we bound  $R_n$ . Let  $\bar{p} = \mathbb{E}[\widehat{p}_n]$  and  $\epsilon_n = (\log n/n)^{\beta/(2\beta+d)}$ . By triangle inequality,  $R_n \leq \|\widehat{p}_n - \bar{p}\|_\infty + \|\bar{p} - p\|_\infty$ . Due to a result of Giné and Guillou (2002) [see also (49) in chapter 3 of Prakasa Rao (1983)], under Assumption A1, there exist constants  $C_1, C_2$ , and  $B_0 > 0$  such that have for all  $B \geq B_0$ ,

$$\mathbb{P}(\|\widehat{p}_n - \bar{p}\|_\infty \geq B\epsilon_n) \leq C_1 \exp(-C_2 B^2 \log(h_n^{-1})) = C_1 h_n^{C_2 B^2}. \quad (\text{A.4})$$

On the other hand, by Assumption A1, for some constant  $C_3$ ,

$$\|\bar{p} - p\|_\infty \leq C_3 h_n^\beta. \quad (\text{A.5})$$

In (A.3), (A.4), and (A.5), the constants  $C_i, i = 0, \dots, 3$ , depend on  $p$  and  $K$  only. Hence,

$$\mathbb{P}(\|\widehat{p}_n - p\|_\infty \geq (C_3 + B)\epsilon_n) \leq C_1 h_n^{C_2 B^2}, \quad (\text{A.6})$$

which concludes the second part by choosing  $A_{2,\lambda} = C_3 + \sqrt{\frac{(2\beta+d)\lambda}{C_2}}$ .  $\square$

*Proof of Theorem 3.3.* Let  $\alpha_n = i_{n,\alpha}/n = \lfloor (n+1)\alpha \rfloor / n$ . We have  $|\alpha_n - \alpha| \leq 1/n$ . Recall that the ideal level  $t^{(\alpha)}$  can be written as  $t^{(\alpha)} = G^{-1}(\alpha)$  where the function  $G$  is the CDF of  $p(Y)$ , as defined in Section 3.2. By the  $\gamma$ -exponent condition, the inverse of  $G$  is well defined in a small neighborhood of  $\alpha$ . When  $n$  is large enough, we can define  $t^{(\alpha_n)}$  as  $t^{(\alpha_n)} = G^{-1}(\alpha_n)$ .

Again, by the  $\gamma$ -exponent condition,  $c_1 |t^{(\alpha_n)} - t^{(\alpha)}|^\gamma \leq |G(t^{(\alpha_n)}) - G(t^{(\alpha)})| = |\alpha_n - \alpha| \leq \frac{1}{n}$ . Therefore, for  $n$  large enough

$$|t^{(\alpha_n)} - t^{(\alpha)}| \leq (c_1 n)^{-1/\gamma}. \quad (\text{A.7})$$

Equation (A.7) allows us to switch to the problem of bounding  $|t_n^{(\alpha)} - t^{(\alpha_n)}|$ . Recall that  $t_n^{(\alpha)} = \widehat{p}_n(Y_{(i_{n,\alpha})})$ . The key of the proof is to observe that  $t_n^{(\alpha)} = G_n^{-1}(\alpha_n) := \inf\{t : G_n(t) \geq \alpha_n\}$ . Then it suffices to show that  $G^{-1}$  and  $G_n^{-1}$  are close at  $\alpha_n$ . In fact, by definition of  $R_n$ , we have for all  $t > 0$ :  $L^\ell(t - R_n) \subseteq L_n^\ell(t) \subseteq L^\ell(t + R_n)$ . As a result, we have

$$P_n(L^\ell(t - R_n)) \leq P_n(L_n^\ell(t)) \leq P_n(L^\ell(t + R_n)).$$

By definition of  $V_n$ ,

$$P(L^\ell(t - R_n)) - V_n \leq P_n(L_n^\ell(t)) \leq P(L^\ell(t + R_n)) + V_n.$$

By definition of  $G$  and  $G_n$ , the above inequality becomes

$$G(t - R_n) - V_n \leq G_n(t) \leq G(t + R_n) + V_n.$$

Let  $W_n = R_n + (2V_n/c_1)^{1/\gamma}$ . Suppose  $n$  is large enough such that

$$\left(\frac{c_1}{n}\right)^{\frac{1}{\gamma}} + \left(\frac{2A_{1,\lambda}}{c_1} \sqrt{\frac{\log n}{n}}\right)^{\frac{1}{\gamma}} < \epsilon_n,$$

then on the event  $V_n \leq A_{1,\lambda} \sqrt{\frac{\log n}{n}}$ ,

$$\begin{aligned} G_n(t^{(\alpha_n)} - W_n) &\leq G(t^{(\alpha_n)} - W_n + R_n) + V_n \\ &= G(t^{(\alpha_n)} - (2V_n/c_1)^{1/\gamma}) - G(t^{(\alpha_n)}) + \alpha_n + V_n \\ &\leq \alpha_n - V_n < \alpha_n. \end{aligned}$$

where the last inequality uses the left side of the  $\gamma$ -exponent condition. Similarly,  $G_n(t^{(\alpha_n)} + W_n) > \alpha_n$ . Hence, for  $n$  large enough, if  $V_n \leq A_{1,\lambda} \sqrt{(\log n)/n}$ , then

$$|t_n^{(\alpha)} - t^{(\alpha_n)}| \leq W_n. \quad (\text{A.8})$$

To conclude the proof, first note that  $(\frac{c_1}{n})^{\frac{1}{\gamma}} = o((\frac{\log n}{n})^{\frac{1}{2\gamma}})$ . Then we can find constant  $A'_\lambda$  such that for all  $n$  large enough,

$$\left(A'_\lambda - \left(\frac{2A_{1,\lambda}}{c_1}\right)^{\frac{1}{\gamma}}\right) \left(\frac{\log n}{n}\right)^{\frac{1}{2\gamma}} \geq \left(\frac{c_1}{n}\right)^{\frac{1}{\gamma}}. \quad (\text{A.9})$$

Let  $A_\lambda = A_{2,\lambda}$ . Combining Equations (A.7) and (A.8), on the event

$$E_{n,\lambda} := \left\{ R_n \leq A_\lambda \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}, \quad V_n \leq A_{1,\lambda} \left(\frac{\log n}{n}\right)^{\frac{1}{2}} \right\}, \quad (\text{A.10})$$

we have, for  $n$  large enough,

$$\begin{aligned} |t_n^{(\alpha)} - t^{(\alpha)}| &\leq |t_n^{(\alpha)} - t^{(\alpha_n)}| + \left(\frac{c_1}{n}\right)^{\frac{1}{\gamma}} \leq W_n + \left(\frac{c_1}{n}\right)^{\frac{1}{\gamma}} \\ &\leq R_n + (2c_1^{-1} V_n)^{1/\gamma} + \left(\frac{c_1}{n}\right)^{\frac{1}{\gamma}} \\ &\leq A_\lambda \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + \left(\frac{2A_{1,\lambda}}{c_1} \sqrt{\frac{\log n}{n}}\right)^{\frac{1}{\gamma}} + \left(\frac{c_1}{n}\right)^{\frac{1}{\gamma}} \\ &\leq A_\lambda \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + A'_\lambda \left(\frac{\log n}{n}\right)^{\frac{1}{2\gamma}}, \end{aligned}$$

where the second last inequality is from the definition of  $E_{n,\lambda}$  and the last inequality is from the choice of  $A'_\lambda$ . The proof is concluded by observing  $\mathbb{P}(E_{n,\lambda}) = O(n^{-\lambda})$ , a consequence of Lemma C.1  $\square$

*Proof of Theorem 3.4.* In the proof, we write  $t_n$  for  $t_n^{(\alpha)}$  as a generic estimate of  $t^{(\alpha)}$  that satisfies (11). Observe that

$$\begin{aligned} \mu(L_n(t_n) \Delta C^{(\alpha)}) &= \mu(\{\widehat{p}_n \geq t_n, p < t^{(\alpha)}\}) \\ &\quad + \mu(\{\widehat{p}_n < t_n, p \geq t^{(\alpha)}\}). \end{aligned} \quad (\text{A.11})$$

Note that

$$\{\widehat{p}_n \geq t_n, p < t^{(\alpha)}\} \subseteq \{t^{(\alpha)} - |t_n - t^{(\alpha)}| - R_n \leq p < t^{(\alpha)}\}, \quad (\text{A.12})$$

and

$$\{\widehat{p}_n < t_n, p \geq t^{(\alpha)}\} \subseteq \{t^{(\alpha)} < p \leq t^{(\alpha)} + |t^{(\alpha)} - t_n| + R_n\}. \quad (\text{A.13})$$

Therefore

$$\begin{aligned} L_n(t_n) \Delta C^{(\alpha)} &\subseteq \{t^{(\alpha)} - |t_n - t^{(\alpha)}| - R_n < p \leq t^{(\alpha)} \\ &\quad + |t^{(\alpha)} - t_n| + R_n\}. \end{aligned} \quad (\text{A.14})$$

Suppose  $n$  is large enough such that

$$2A_{2,\lambda} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + A'_\lambda \left(\frac{\log n}{n}\right)^{\frac{1}{2\gamma}} < \left(\epsilon_0 \wedge \frac{t^{(\alpha)}}{2}\right),$$

where the constant  $A_{2,\lambda}$  is defined as in Lemma C.1 and  $A'_\lambda$  is defined as in Equation (A.9). Then on the event  $E_{n,\lambda}$  as defined in Equation (A.10), applying Theorem 3.3 and condition (10) on the right-hand side of (A.14) yields

$$\begin{aligned} \mu(L_n(t_n) \Delta C^{(\alpha)}) &\leq \frac{P(L_n(t_n) \Delta C^{(\alpha)})}{t^{(\alpha)} - |t_n - t^{(\alpha)}| - R_n} \\ &\leq \frac{2}{t^{(\alpha)}} c_2 \left(2A_{2,\lambda} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + A'_\lambda \left(\frac{\log n}{n}\right)^{\frac{1}{2\gamma}}\right)^\gamma \\ &\leq B_\lambda \left(\frac{\log n}{n}\right)^{\frac{\beta\gamma}{2\beta+d}} + B'_\lambda \left(\frac{\log n}{n}\right)^{\frac{1}{2}}, \end{aligned} \quad (\text{A.15})$$

where  $B_\lambda, B'_\lambda$  are positive constants depending only on  $p, K, \alpha$ , and  $\gamma$ . As a result, both  $L_n^-$  and  $L_n^+$  satisfies the claim of Theorem 3.4. The claim also holds for  $\widehat{C}^{(\alpha)}$  by the sandwich lemma.  $\square$

## APPENDIX D: PROOF OF LEMMA 4.2

*Proof of Lemma 4.2.* The first statement follows since

$$\begin{aligned} \mathcal{E}(C) &= \mu(C) - \mu(C_*) = \mu(C \cap C_*^c) + \mu(C \cap C_*) \\ &\quad - [\mu(C_* \cap C) + \mu(C_* \cap C^c)] \\ &= \mu(C \cap C_*^c) - \mu(C_* \cap C^c) \leq \mu(C \cap C_*^c) + \mu(C_* \cap C^c) \\ &= R(C). \end{aligned}$$

For the second statement, let  $I$  denote the indicator function for  $C$  and let  $I_*$  denote the indicator function for  $C_*$ . Note that, for all  $y$ ,  $(I(y) - I_*(y))(\lambda - p(y)) \geq 0$ . Let  $\lambda = \lambda_\alpha$  and define  $W_\epsilon = \{y : |p(y) - \lambda| > \epsilon\}$ . From Assumption A2 with  $\gamma = 1$ , we have that  $\mu(C \Delta C_*) \leq \mu((C \Delta C_*) \cap W_\epsilon) + c\epsilon$  for some  $c > 0$ . Hence,

$$\begin{aligned} \mu(C \Delta C_*) &\leq \mu((C \Delta C_*) \cap W_\epsilon) + c\epsilon \\ &= \frac{1}{\epsilon} \int_{W_\epsilon} |I(y) - I_*(y)| \epsilon d\mu(y) + c\epsilon \\ &\leq \frac{1}{\epsilon} \int_{W_\epsilon} |I(y) - I_*(y)| |\lambda - p(y)| d\mu(y) + c\epsilon \\ &\leq \frac{1}{\epsilon} \int (I(y) - I_*(y)) (\lambda - p(y)) d\mu(y) + c\epsilon \\ &= \frac{\lambda}{\epsilon} [\mu(C) - \mu(C_*)] - \frac{1}{\epsilon} [P(C) - P(C_*)] + c\epsilon \\ &= \frac{\lambda}{\epsilon} \mathcal{E}(C) - \frac{1}{\epsilon} [P(C) - (1 - \alpha)] + c\epsilon. \end{aligned}$$

Since  $\mathbb{E}(P(C)) \geq 1 - \alpha$ , if we take expected values of both sides, we have that  $\mathbb{E}(R(C)) \leq \frac{\lambda}{\epsilon} \mathbb{E}(\mathcal{E}(C)) + c\epsilon$ . The conclusion follows by setting  $\epsilon = \sqrt{\lambda \mathbb{E}(\mathcal{E}(C)) / c}$ .  $\square$

[Received March 2012. Revised November 2012.]

## REFERENCES

- Aichison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, Cambridge: Cambridge University Press. [278]
- Audibert, J., and Tsybakov, A. (2007), "Fast Learning for Plug-in Classifiers," *The Annals of Statistics*, 35, 608–633. [281]
- Baillo, A. (2003), "Total Error in a Plug-in Estimator of Level Sets," *Statistics & Probability Letters*, 65, 411–417. [278]
- Baillo, A., Cuestas-Alberto, J., and Cuevas, A. (2001), "Convergence Rates in Nonparametric Estimation of Level Sets," *Statistics & Probability Letters*, 53, 27–35. [278]
- Cadre, B. (2006), "Kernel Estimation of Density Level Sets," *Journal of Multivariate Analysis*, 97, 999–1023. [278, 279]
- Cadre, B., Pelletier, B., and Pudlo, P. (2009), "Clustering by Estimation of Density Level Sets at a Fixed Probability," manuscript. Available at <http://hal.archives-ouvertes.fr/docs/00/39/74/37/PDF/tlevel.pdf>. [279, 280, 281]
- Chatterjee, S. K., and Patra, N. K. (1980), "Asymptotically Minimal Multivariate Tolerance Sets," *Calcutta Statistical Association Bulletin*, 29, 73–93. [278, 279, 280]
- Di Buccianico, A., Einmahl, J. H., and Mushkudiani, N. A. (2001), "Smallest Nonparametric Tolerance Regions," *The Annals of Statistics*, 29, 1320–1343. [278, 279]
- Fraser, D. A. S., and Guttman, I. (1956), "Tolerance Regions," *The Annals of Mathematical Statistics*, 27, 162–179. [278]
- Giné, E., and Guillou, A. (2002), "Rates of Strong Uniform Consistency for Multivariate Kernel Density Estimators," *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 38, 907–921. [285]
- Guttman, I. (1970), *Statistical Tolerance Regions: Classical and Bayesian*, London: Griffin. [278]
- Hartigan, J. (1975), *Clustering Algorithms*, New York: Wiley. [278]

- Hyndman, R. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician*, 50, 120–125. [279,280,282]
- Li, J., and Liu, R. (2008), "Multivariate Spacings Based on Data Depth: I. Construction of Nonparametric Multivariate Tolerance Regions," *The Annals of Statistics*, 36, 1299–1323. [278,279,283]
- Loader, C. (1999), "Bandwidth Selection: Classical or Plug-in?," *The Annals of Statistics*, 27, 415–438. [281]
- Mammen, E., Miranda, M. D. M., Nielsen, J. P., and Sperlich, S. (2011), "Do-Validation for Kernel Density Estimation," *Journal of the American Statistical Association*, 106, 651–660. [281]
- Park, C., Huang, J. Z., and Ding, Y. (2010), "A Computable Plug-in Estimator of Minimum Volume Sets for Novelty Detection," *Operations Research*, 58, 1469–1480. [279,282,283,284]
- Polonik, W. (1995), "Measuring Mass Concentrations and Estimating Density Contour Clusters—An Excess Mass Approach," *The Annals of Statistics*, 23, 855–881. [278,281]
- (1997), "Minimum Volume Sets and Generalized Quantile Processes," *Stochastic Processes and Their Applications*, 69, 1–24. [279]
- Prakasa Rao, B. (1983), *Nonparametric Functional Estimation*, New York: Academic Press. [285]
- Rigollet, P., and Vert, R. (2009), "Optimal Rates for Plug-in Estimators of Density Level Sets," *Bernoulli*, 14, 1154–1178. [278,280,281,284]
- Rinaldo, A., and Wasserman, L. (2010), "Generalized Density Clustering," *The Annals of Statistics*, 38, 2678–2722. [278]
- Samworth, R. J., and Wand, M. P. (2010), "Asymptotics and Optimal Bandwidth Selection for Highest Density Region Estimation," *The Annals of Statistics*, 38, 1767–1792. [279,281]
- Scott, C. D., and Nowak, R. D. (2006), "Learning Minimum Volume Sets," *Journal of Machine Learning Research*, 7, 665–704. [279,282]
- Shafer, G., and Vovk, V. (2008), "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research*, 9, 371–421. [279]
- Sricharan, K., and Hero, A. (2011), "Efficient Anomaly Detection Using Bipartite k-NN Graphs," in *Advances in Neural Information Processing Systems 24*, eds. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Red Hook, NY: Curran Associates, Inc. pp. 478–486. [279]
- Steinwart, I., Hush, D., and Scovel, C. (2005), "A Classification Framework for Anomaly Detection," *Journal of Machine Learning Research*, 6, 211–232. [279]
- Tsybakov, A. (1997), "On Nonparametric Estimation of Density Level Sets," *The Annals of Statistics*, 25, 948–969. [278,281]
- (2009), *Introduction to Nonparametric Estimation*, Berlin: Springer. [284]
- Tukey, J. (1947), "Nonparametric Estimation, II. Statistical Equivalent Blocks and Multivariate Tolerance Regions," *The Annals of Mathematical Statistics*, 18, 529–539. [279]
- van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [285]
- Vovk, V., Gammelman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, New York: Springer. [278,279]
- Vovk, V., Nouretdinov, I., and Gammelman, A. (2009), "On-Line Predictive Linear Regression," *The Annals of Statistics*, 37, 1566–1590. [279]
- Wald, A. (1943), "An Extension of Wilks Method for Setting Tolerance Limits," *The Annals of Mathematical Statistics*, 14, 45–55. [278]
- Walther, G. (1997), "Granulometric Smoothing," *The Annals of Statistics*, 25, 2273–2299. [279]
- Wei, Y. (2008), "An Approach to Multivariate Covariate-Dependent Quantile Contours With Application to Bivariate Conditional Growth Charts," *Journal of the American Statistical Association*, 103, 397–409. [278]
- Wilks, S. (1941), "Determination of Sample Sizes for Setting Tolerance Limits," *The Annals of Mathematical Statistics*, 12, 91–96. [278,279]
- Willett, R., and Nowak, R. (2007), "Minimax Optimal Level-Set Estimation," *IEEE Transactions on Image Processing*, 16, 2965–2979. [278]
- Zhao, M., and Saligrama, V. (2009), "Anomaly Detection With Score Functions Based on Nearest Neighbor Graphs," in *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I., Williams, and A. Culotta, Red Hook, NY: Curran Associates, Inc. pp. 2250–2258. [279,282,283]