Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2014)
76, *Part* 1, *pp.* 71–96

Distribution-free prediction bands for non-parametric regression

Jing Lei and Larry Wasserman

Carnegie Mellon University, Pittsburgh, USA

[Received March 2012. Revised December 2012]

Summary. We study distribution-free, non-parametric prediction bands with a focus on their finite sample behaviour. First we investigate and develop different notions of finite sample coverage guarantees. Then we give a new prediction band by combining the idea of 'conformal prediction' with non-parametric conditional density estimation. The proposed estimator, called COPS (conformal optimized prediction set), always has a finite sample guarantee. Under regularity conditions the estimator converges to an oracle band at a minimax optimal rate. A fast approximation algorithm and a data-driven method for selecting the bandwidth are developed. The method is illustrated in simulated and real data examples.

Keywords: Conformal prediction; Finite sample property; Kernel density; Prediction bands

1. Introduction

Given observations $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}^1$ for $i = 1, \ldots, n$, where $\mathcal{X} \subset \mathbb{R}^d$, we want to predict Y_{n+1} given a future predictor X_{n+1}. Unlike typical non-parametric regression methods, our goal is not to produce a point prediction. Instead, we construct a prediction set C_n that contains Y_{n+1} with probability at least $1 - \alpha$. More precisely, assume that $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are independent and identically distributed observations from some distribution P. We construct, from the first n sample points, a set-valued function

$$C_n(x) \equiv C_n(X_1, Y_1, \ldots, X_n, Y_n, x) \subseteq \mathbb{R}^1 \tag{1}$$

such that the next response variable Y_{n+1} falls inside $C_n(X_{n+1})$ with a certain level of confidence. The collection of prediction sets $C_n = \{C_n(x) : x \in \mathbb{R}^d\}$ forms a *prediction band*. The exact form of $\mathcal{X}$ is not critical; for simplicity, we take $\mathcal{X} = [0, 1]^d$.

The problem of prediction sets is well studied in the context of linear regression, where prediction sets are usually constructed under linear and Gaussian assumptions (see DeGroot and Schervish (2012), theorem 11.3.6). The Gaussian assumption can be relaxed by using, for example, quantile regression (Koenker and Hallock, 2001). These linear-model-based methods usually have reasonable finite sample performance. However, the coverage is valid only when the linear (or other parametric) regression model is correctly specified. In contrast, non-parametric methods have the potential to work for any smooth distribution (Ruppert *et al.*, 2003) but only asymptotic results are available and the finite sample behaviour remains unclear.

Recently, Vovk *et al.* (2009) introduced a generic approach, called *conformal prediction*, to construct valid, distribution-free, sequential prediction sets. When adapted to our setting, this

Address for correspondence: Jing Lei, Department of Statistics, 132 Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: jinglei@andrew.cmu.edu

 1369–7412/14/76071

yields prediction bands with a *finite sample coverage guarantee* (or *finite sample validity*) in the sense that

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geqslant 1-\alpha \qquad \text{for all } P, \tag{2}$$

where $\mathbb{P} = P^{n+1}$ is the joint measure of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$. However, the conditional coverage and statistical efficiency of such bands have not been investigated.

In this paper we extend the results in Vovk *et al.* (2009) and study conditional coverage as well as efficiency. We show that, although finite sample coverage defined in expression (2) is a desirable property, this is not enough to guarantee good prediction bands. We argue that the finite sample coverage that is given by expression (2) should be interpreted as *marginal coverage*, which is different from (in fact, weaker than) the *conditional coverage* as usually sought in prediction problems. Requiring only marginal validity may lead to unsatisfactory estimation even in very simple cases. As a result, a good estimator must satisfy something more than marginal coverage. A natural criterion would be conditional coverage. However, we prove that distribution-free conditional coverage is impossible to achieve with a finite sample. As an alternative solution, we develop a new notion, called *local validity*, that interpolates between marginal and conditional validity, and is achievable with a finite sample. This notion leads to our proposed estimator COPS (conformal optimized prediction set). We also show that when the sample size goes to ∞, under regularity conditions, the locally valid prediction band given by COPS can give arbitrarily accurate conditional coverage, leading to an asymptotic conditional coverage guarantee.

We study the efficiency of our estimator by measuring its deviation from an *oracle band*: the band that one should use if the joint distribution P were known. We also give a minimax lower bound on the estimation error so that the efficiency of our method is indeed minimax rate optimal over a certain class of smooth distributions.

To summarize: the method that is given in this paper is the first with both finite sample (marginal and local) coverage, asymptotic conditional coverage and an explicit rate for asymptotic efficiency. The finite sample marginal and local validity is distribution free: no assumptions on P are required. In fact, P does not even need to have a density. Furthermore, all tuning parameters are completely data driven.

The problem of constructing prediction bands resembles that of non-parametric confidence band estimation for the regression function $m(x) = \mathbb{E}(Y|X=x)$. However, these are two different inference problems. First note that non-trivial, distribution-free confidence bands for the regression function $m(x) = \mathbb{E}(Y|X=x)$ do not exist (Low, 1997; Genovese and Wasserman, 2008). However, in this paper we show that consistent prediction bands estimation is possible under mild regularity conditions. Hence there is a distinct difference between confidence bands for the regression function and prediction bands.

1.1. Prior work on non-parametric prediction bands

The usual non-parametric prediction set takes the form

$$\hat{m}(x) \pm z_{\alpha/2}\sqrt{(\hat{\sigma}^2 + s^2)} \tag{3}$$

where $\hat{m}$ is some non-parametric regression estimator, $\hat{\sigma}^2$ is an estimate of the conditional variance of Y given X and s is an estimate of the standard error of $\hat{m}$ and $z_{\alpha/2}$ is either a normal quantile or a quantile determined by bootstrapping. See, for example, section 6.2 of Ruppert *et al.* (2003), section 2.3.3 of Loader (1999) and chapter 5 of Fan and Gijbels (1996). The assumption of constant variance can be relaxed; see, for example, Akritas and Van Keilegom

(2001). Another important class of methods is quantile regression (Koenker and Hallock, 2001) where it is assumed, in a non-parametric form, that the τth quantile of Y given $X = x$ is a smooth function $f_\tau(x)$. If such a quantile regression assumption holds for both $\tau = \alpha/2$ and $\tau = 1 - \alpha/2$, then a prediction band with conditional $1 - \alpha$ coverage is given by $C_n(x) = [\hat{f}_{\alpha/2}(x), \hat{f}_{1-\alpha/2}(x)]$, where $\hat{f}_{\alpha/2}$ and $\hat{f}_{1-\alpha/2}$ are estimated quantile regression functions. Other related work includes Hall and Rieck (2001) on bootstrapping, Davidian and Carroll (1987) on variance estimation and Carroll and Ruppert (1991) on transformation approaches. However, none of these methods yields prediction bands with distribution-free, finite sample validity. Furthermore, these methods always produce a prediction set in the form of an interval which, as we shall see, may not be optimal. In fact, we are not aware of any reference that provides distribution-free finite sample prediction bands with asymptotic optimality properties as we provide in this paper. The only reference that we know of that provides finite sample marginal validity is the work by Vovk *et al.* (2009). However, they focused on linear predictors and did not address efficiency or conditional validity.

1.2. Outline

In Section 2 we introduce various notions of validity and efficiency. In Section 3 we introduce our methods for prediction bands: the COPS estimator. We study the large sample and minimax results of the method in Section 4. We discuss bandwidth selection in Section 5. Section 6 contains several simulation and data examples. Finally, concluding remarks are in Section 7. Some technical details are relegated to Appendix A.

2. Marginal, conditional and local validity

2.1. Marginal validity and prediction sets

Prediction bands are an extension of non-parametric prediction sets (which are also called tolerance regions) which concerns a simple scenario without covariates. Specifically, suppose that we observe n independent and identically distributed copies $Z_1, \ldots, Z_n$ of a random vector $Z \in \mathbb{R}^d$ with distribution P and we want a set $T_n = T_n(Z_1, \ldots, Z_n) \subseteq R^d$ such that $\mathbb{P}(Z_{n+1} \in T_n) \geqslant 1 - \alpha$ for all P. Now we consider prediction with covariates and let $Z_i = (X_i, Y_i)$. Since the probability statement in expression (2) is over the joint distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$, it is equivalent to

$$\mathbb{P}\{(X_{n+1}, Y_{n+1}) \in C_n\} \geqslant 1 - \alpha, \qquad \text{for all } P, \tag{4}$$

i.e. equation (4) is exactly the definition of a prediction set for the joint distribution (X, Y). As a result, any finite sample prediction set for the joint distribution provides a solution to the prediction band problem. In this subsection we pursue this point further. In the following subsections we consider improvements.

We start from the notion of the optimal prediction set, where optimality refers to minimizing the Lebesgue measure while maintaining the probability coverage at the nominal level. The optimal prediction set at level $1 - \alpha$ is an upper level set of the joint density

$$C^{(\alpha)} = \{(x, y) : p(x, y) \geqslant t^{(\alpha)}\}, \tag{5}$$

where $t^{(\alpha)}$ is chosen such that $P(C^{(\alpha)}) = 1 - \alpha$. As illustrated in the following example, an optimal joint prediction can lead to an unsatisfactory prediction band.

Fig. 1 shows the case of a bivariate independent normal distribution. According to equation (5), when X and Y are independent standard normal distributions, the optimal prediction set for

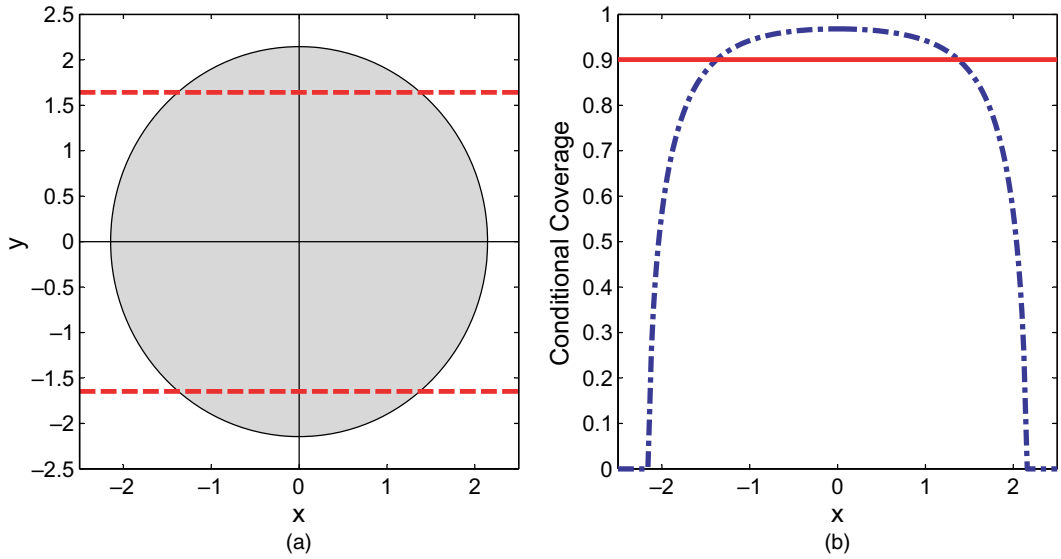

Fig. 1. Joint prediction set and pointwise conditional coverage for bivariate independent Gaussian distributions: (a) optimal (with smallest Lebesgue measure) prediction set with coverage 0.9 (◍) and upper and lower 5% quantiles of the marginal distribution of Y (– – –); (b) $P\{Y \in C(x)|X = x\}$ *versus* x (· – · –) and the desired coverage level (———)

any α is a circle centred at the origin as described by the grey area in Fig. 1(a). But intuitively, since observing X provides no information about Y, the best prediction band at level α should be $C(x) = [-z_{\alpha/2}, z_{\alpha/2}]$, for all x, where z_τ is the τth upper quantile of standard normal. This band is the set between the two broken lines in Fig. 1(a) for $\alpha = 0.1$.

In prediction, another important notion of coverage is the conditional coverage $P\{Y \in C(x)|X = x\}$. The pointwise conditional coverage $P\{Y \in C(x)|X = x\}$ is plotted in Fig. 1(b) for the joint prediction set (the chain curve). We see that the 'optimal' joint prediction set tends to overestimate the set when x is in the high density area and to underestimate for low density x. Let us now consider conditional validity in more detail.

2.2. Conditional validity

Only requiring expression (2) for prediction bands is not enough. We shall refer to expression (2) as *marginal validity* or *joint validity*. This is the type of validity that was used in Shafer and Vovk (2008). As illustrated in the example above, it may be tempting to insist on a more stringent probability guarantee such as the *conditional validity*:

$$\mathbb{P}\{Y_{n+1} \in C_n(x)|X_{n+1} = x\} \geqslant 1 - \alpha \qquad \text{for all } P \text{ and almost all } x. \tag{6}$$

If the joint distribution of (X, Y) is known, we can define an oracle band as the counterpart of expression (2) for conditionally valid bands:

$$C_P(x) = \{y : p(y|x) \geqslant t^{(\alpha)}(x)\} \tag{7}$$

where $t^{(\alpha)}(x) \equiv t_x^{(\alpha)}$ satisfies

$$\int \mathbb{1}\{p(y|x) \geqslant t^{(\alpha)}(x)\}\, p(y|x)\, \mathrm{d}y = 1 - \alpha.$$

We call $C_P = \{C_P(x) : x \in \mathbb{R}^d\}$ the *conditional oracle band*. It is easy to prove that C_P minimizes $\mu\{C(x)\}$ for all x among all bands satisfying $\inf_x P\{Y \in C(x) | X = x\} \geqslant 1 - \alpha$. Note that C_P depends on P but does not depend on the observed data. For an estimator C_n, *asymptotic efficiency* requires that $C_n(x)$ is close to $C_P(x)$ uniformly over all x:

$$\sup_x \mu\{C_n(x) \triangle C_P(x)\} \xrightarrow{\mathrm{P}} 0 \tag{8}$$

where $\triangle$ denotes symmetric set difference. However, we shall show that there do not exist any prediction bands $\hat{C}$ that satisfy both condition (6) and condition (8). In fact, the following claim, which is proved in Appendix A.2, is even stronger.

Let P_X denote the marginal distribution of X under P. A point x is a *non-atom* for P if x is in the support of P_X and if $P_X\{B(x, \delta)\} \to 0$ as $\delta \to 0$, where $B(x, \delta)$ is the Euclidean ball centred at x with radius δ. Let $N(P)$ denote the set of non-atoms. We show that if C_n satisfies condition (6) then the length of $C_n(x)$ is infinite for all $x \in N(P)$.

Lemma 1 (impossibility of finite sample conditional validity). Suppose that an estimator C_n has $1 - \alpha$ conditional validity in the sense of condition (6). For any P and any $x_0 \in N(P)$,

$$\mathbb{P}[\lim_{\delta \to 0} \operatorname{ess\,sup}_{\|x_0 - x\| \leqslant \delta} \mu\{C_n(x)\} = \infty] = 1.$$

Thus, non-trivial finite sample conditional validity is impossible for continuous distributions. We shall instead construct prediction bands with an asymptotic version of condition (6) together with finite sample marginal validity. We say that C_n is *asymptotically conditionally valid* if

$$\sup_x [\mathbb{P}\{Y_{n+1} \notin C_n(x) | X_{n+1} = x\} - \alpha]_+ \xrightarrow{\mathrm{P}} 0 \tag{9}$$

as $n \to \infty$. Here, the supremum is taken over the support of P_X. We note that, if the conditional density $p(y|x)$ is uniformly bounded for all (x, y), then asymptotic conditional validity is a consequence of asymptotic efficiency defined as in condition (8).

In Section 3 we construct a prediction band that satisfies

(a) finite sample marginal validity,
(b) asymptotic conditional validity and
(c) asymptotic efficiency.

Our method is based on the notion of *local validity*, which naturally interpolates between marginal and conditional validity.

Definition 1 (local validity). Let $\mathcal{A} = \{A_j : j \geqslant 1\}$ be a partition of $\operatorname{supp}(P_X)$. A prediction band C_n is locally valid with respect to $\mathcal{A}$ if

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) | X_{n+1} \in A_j\} \geqslant 1 - \alpha, \qquad \text{for all } j \text{ and all } P. \tag{10}$$

Remark 1. The notion of local validity must be considered together with the resolution of partition $\mathcal{A}$. To be specific, let $\delta = \sup_{A \in \mathcal{A}} \operatorname{diam}(A)$. Consider the limiting case of $\delta \to \infty$, which can be thought of as having $\mathcal{A} = \{\operatorname{supp}(P_X)\}$, and local validity becomes marginal validity. In contrast, in the extremal case $\delta \to 0$, A_j shrinks to a single point $x \in \mathbb{R}^d$, and local validity approximates conditional validity. We also note that local validity is stronger than marginal validity but weaker than conditional validity. We state the following proposition whose proof is elementary and omitted.

Proposition 1. If C is conditionally valid, then it is also locally valid for any partition $\mathcal{A}$. If C is locally valid for some partition $\mathcal{A}$, then it is also marginally valid.

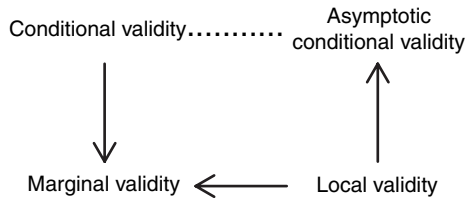

Fig. 2. Relationship between different types of validity

The relationship between local validity and asymptotic conditional validity is more complicated and is one of the technical contributions of this paper. In Section 3 we construct a specific class of locally valid bands. In theorem 1 of Section 4 we show that, under mild regularity conditions, these bands are also asymptotically conditionally valid. To summarize, if C is locally valid then it is also marginally valid. And, under regularity conditions, it can also be asymptotically conditionally valid: Fig. 2.

How can we construct finite sample locally valid prediction bands? A straightforward approach is to apply the method that was developed in Lei *et al.* (2011) to $P_j \equiv \mathcal{L}(X, Y | X \in A_j)$, the joint distribution of (X, Y) conditional on the event $X \in A_j$. Note that we are mostly interested in the case $\max_j \text{diam}(A_j) \to 0$; therefore the marginal density of X within P_j becomes increasingly close to uniform. Therefore, the approach can be simplified to finding $C_{n,j} \in \mathbb{R}^1$, such that $\mathbb{P}(Y \in C_{n,j} | X \in A_j) \geqslant 1 - \alpha$. This approach is detailed in Section 3 and analysed in Section 4.

3. Methodology

3.1. Marginally valid prediction band

We start by recalling the construction of joint prediction sets by using kernel density estimators together with the idea of conformal prediction, as described in Lei *et al.* (2011), using the idea of *conformal prediction* that was developed in Shafer and Vovk (2008) and Vovk *et al.* (2005, 2009). This approach is shown to have finite sample validity as well as asymptotic efficiency under regularity conditions. Suppose that we observe

$$Z_1, \ldots, Z_n \sim P$$

and we want a prediction set for Z_{n+1}. The idea is to test $H_0 : Z_{n+1} = z$ for each z and then to invert the test. Specifically, for any z let $\hat{p}_n^z(\cdot)$ be a density estimator based on the *augmented data* $\text{aug}(\mathbf{Z}; z) = (Z_1, \ldots, Z_n, z)$. Define

$$C_n \equiv C_n(Z_1, \ldots, Z_n) = \{z : \pi_n(z) \geqslant \alpha\} \tag{11}$$

where

$$\pi_n(z) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}\{\sigma_i(z) \leqslant \sigma_{n+1}(z)\}$$

is the p-value for the test, $\sigma_i(z) = \hat{p}_n^z(Z_i)$ for $i = 1, \ldots, n$ and $\sigma_{n+1}(z) = \hat{p}_n^z(z)$. The statistic σ_i is an example of a *conformity measure*. More generally, a conformity measure $\sigma_i(z) = \sigma\{\text{aug}(\mathbf{Z}, z), Z_i\}$ indicates how well a data point Z_i agrees with the augmented data set $\text{aug}(\mathbf{Z}, z)$. In principle $\sigma(\cdot, \cdot)$ can be any function but usually it makes sense to use the fitted residual or likelihood at Z_i with respect to a model estimated from $\text{aug}(\mathbf{Z}, z)$.

The intuition for C_n is as follows. Fix an arbitrary value z. To test $H_0 : Z_{n+1} = z$ we use the heights of the density estimators $\sigma_i(z) = \hat{p}_n^z(Z_i)$ as a test statistic. (Note that $\sigma_1, \ldots, \sigma_{n+1}$ are functions of $\text{aug}(\mathbf{Z}, z)$.) Under H_0, the ranks of the σ_i are uniformly distributed among $\{1, 2, \ldots, n+1\}$, because the joint distribution of $(Z_1, \ldots, Z_n, Z_{n+1})$ does not change under permutations so the vector $(\sigma_1, \ldots, \sigma_{n+1})$ is exchangeable. Therefore, under H_0, $\pi_n(z)$ is uniformly distributed over $\{1/(n+1), 2/(n+1), \ldots, 1\}$ and is a valid p-value for the test in the sense that $\mathbb{P}\{\pi_n(Z_{n+1}) \geqslant \alpha\} \geqslant 1 - \alpha$. The set C_n is obtained by inverting the hypothesis test, i.e. C_n consists of all values z that are not rejected by the test. It then follows that $\mathbb{P}(Z_{n+1} \in C_n) \geqslant 1 - \alpha$ for all P.

In Lei *et al.* (2011), the density $\hat{p}_n^z$ is obtained from kernel density estimators with bandwidth h. It is shown that C_n is also efficient in that it is close to $C^{(\alpha)}$ with high probability where $C^{(\alpha)}$ is the optimal prediction set as defined in expression (5).

Now let $Z = (X, Y)$. The x-slices of a prediction set for Z define a marginally valid band. Specifically, let K_x and K_y be two kernel functions in $\mathbb{R}^d$ and $\mathbb{R}^1$ respectively and consider the kernel density estimator. For any $(u, v) \in \mathbb{R}^d \times \mathbb{R}^1$:

$$\hat{p}_{n;X,Y}(u, v) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n^{d+1}} K_x\left(\frac{u - X_i}{h_n}\right) K_y\left(\frac{v - Y_i}{h_n}\right). \tag{12}$$

For any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^1$, let $(\mathbf{X}, \mathbf{Y}) = (X_1, Y_1, \ldots, X_n, Y_n)$ be the data set and $\text{aug}\{\mathbf{X}, \mathbf{Y}; (x, y)\}$ be the augmented data with $X_{n+1} = x$ and $Y_{n+1} = y$. Define $\hat{p}_{n;X,Y}^{(x,y)}$ as the kernel density estimator from the augmented data:

$$\hat{p}_{n;X,Y}^{(x,y)}(u, v) = \frac{n}{n+1} \hat{p}_{n;X,Y}(u, v) + \frac{1}{(n+1)h_n^{d+1}} K_x\left(\frac{u - x}{h_n}\right) K_y\left(\frac{v - y}{h_n}\right). \tag{13}$$

Define the conformity measure

$$\sigma_i(x, y) := \hat{p}_{n;X,Y}^{(x,y)}(X_i, Y_i) \tag{14}$$

and p-value

$$\pi_i = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbb{1}\{\sigma_j(x, y) \leqslant \sigma_i(x, y)\}, \qquad \text{for } 1 \leqslant i \leqslant n+1. \tag{15}$$

Let $\tilde{\alpha} = \lfloor (n+1)\alpha \rfloor / (n+1)$. Since $(X_i, Y_i)_{i=1}^{n+1}$ are independent and identically distributed by exchangeability, we have, for all i,

$$\mathbb{P}(\pi_i \geqslant \tilde{\alpha}) \geqslant 1 - \alpha. \tag{16}$$

Define

$$\hat{C}^{(\alpha)}(x) = \{y : \pi_{n+1}(x, y) \geqslant \tilde{\alpha}\},$$

where $\pi_{n+1} \equiv \pi_{n+1}[\text{aug}\{\mathbf{X}, \mathbf{Y}; (x, y)\}]$. From inequality (16) we have the following lemma.

Lemma 2. $\hat{C}^{(\alpha)}(x)$ is finite sample marginally valid:

$$\mathbb{P}\{Y_{n+1} \in \hat{C}^{(\alpha)}(X_{n+1})\} \geqslant 1 - \alpha \qquad \text{for all } P.$$

Computing $\hat{C}^{(\alpha)}$ is expensive since we need to find the p-value $\pi_n(z)$ for every z. Lei *et al.* (2011) proposed an accurate approximation C_n^+ to $\hat{C}^{(\alpha)}$—called the sandwich approximation—which avoids the augmentation step altogether but preserves finite sample validity. Let $Z_{(1)}, Z_{(2)}, \ldots,$ denote the data ordered increasingly by $\hat{p}(Z_i)$. Let $j = \lfloor n\alpha \rfloor$ and define

$$C_n^+ = \left\{ z : \hat{p}(z) \geqslant \hat{p}(Z_{(j)}) - \frac{K(0)}{nh^d} \right\}, \tag{17}$$

where $\hat{p}(\cdot)$ is the estimated kernel density with using kernel $K(\cdot)$ and bandwidth h. It can be shown that $\hat{C}^{(\alpha)} \subseteq C_n^+$ and hence C_n^+ also has finite sample validity. Moreover, C_n^+ has the same efficiency properties as C_n if h is chosen appropriately. This result, which is known as the 'sandwich lemma', provides a simple characterization of the conformal prediction set C_n in terms of the plug-in density level set. In this paper, a specific version of the sandwich lemma for the conditional density is stated in lemma 3 in Section 4.2. Thus, using the sandwich approximation we obtain a fast method for constructing a valid band, based on slicing the joint density.

Now we can use the sandwich approximation to the joint conformal region for (X, Y). The resulting band $C_n^+(x)$ is obtained by fixing $X = x$ and taking slices of the joint region and is then a marginally valid band. See algorithm 1.

To summarize: the band given in algorithm 1 is marginally valid. But it is not efficient nor does it satisfy asymptotic conditional validity. This leads to the subject of the next section.

3.1.1. *Algorithm 1: sandwich slicer algorithm*

(a) Let $\hat{p}(x, y)$ be the joint density estimator.
(b) Let $Z_i = (X_i, Y_i)$ and let $Z_{(1)}, Z_{(2)}, \ldots,$ denote the sample ordered increasingly by $\hat{p}(X_i, Y_i)$.
(c) Let $j = \lfloor n\alpha \rfloor$ and define

$$C_n^+(x) = \left\{ y : \hat{p}(x, y) \geqslant \hat{p}(X_{(j)}, Y_{(j)}) - \frac{K_x(0)\, K_y(0)}{nh^{d+1}} \right\}. \tag{18}$$

3.2. *Locally valid bands*

Now we extend the idea of conformal prediction to construct prediction bands with local validity. These bands will also be asymptotically efficient and have asymptotic conditional validity. For simplicity of presentation, we assume that $\text{supp}(P_X) = [0, 1]^d$ where $\text{supp}(P_X)$ denotes the support of P_X and we consider partitions $\mathcal{A} = \{A_k, k \geqslant 1\}$ in the form of equilateral cubes with sides of length w_n, where w_n is a small number to be discussed later. Let $n_k = \Sigma_{i=1}^n \mathbf{1}(X_i \in A_k)$ be the histogram count.

Given a kernel function $K(\cdot) : \mathbb{R}^1 \mapsto \mathbb{R}^1$ and another bandwidth h_n, consider the estimated local marginal density of Y:

$$\hat{p}(v|A_k) = \frac{1}{n_k h_n} \sum_{i=1}^n \mathbb{1}(X_i \in A_k)\, K\left(\frac{Y_i - v}{h_n} \right).$$

The corresponding augmented estimate is, for any $(x, y) \in A_k \times \mathbb{R}^1$,

$$\hat{p}^{(x,y)}(v|A_k) = \frac{n_k}{n_k + 1} \hat{p}(v|A_k) + \frac{1}{(n_k + 1)h_n} K\left(\frac{v - y}{h_n} \right). \tag{19}$$

For any $(x, y) \in A_k \times \mathbb{R}^1$, consider the following *local conformity rank*

$$\pi_{n,k}(x, y) = \frac{1}{n_k + 1} \sum_{i=1}^{n+1} \mathbb{1}(X_i \in A_k)\, \mathbb{1}\{\hat{p}^{(x,y)}(Y_i|A_k) \leqslant \hat{p}^{(x,y)}(Y_{n+1}|A_k)\}, \tag{20}$$

which can be interpreted as the local conditional density rank. It is easy to check that the $\pi_{n,k}(x, y)$ has a subuniform distribution if $(X_{n+1}, Y_{n+1}) = (x, y)$ is another independent sample from P. Therefore, the band

$$\hat{C}_{\text{loc}}^{(\alpha)}(x) = \{y : \pi_{n,k}(x, y) \geqslant \alpha\} \tag{21}$$

for $x \in A_k$ has finite sample local validity.

Proposition 2. For $x \in A_k$, let $\hat{C}_{\text{loc}}^{(\alpha)}(x) = \{y : \pi_{n,k}(x, y) \geqslant \alpha\}$, where $\pi_{n,k}(x, y)$ is defined as in equation (20); then $\hat{C}_{\text{loc}}^{(\alpha)}(x)$ is finite sample locally valid and hence finite sample marginally valid.

Proof. Fix k, let $\{i_1, \ldots, i_{n_k}\} = \{i : 1 \leqslant i \leqslant n, X_i \in A_k\}$. Let $(X_{n+1}, Y_{n+1}) \sim P$ be another independent sample. Define $i_{n_k+1} = n+1$ and $\sigma_{i_l} = \hat{p}^{(x,y)}(Y_{i_l}|A_k)$ for all $1 \leqslant l \leqslant n_k + 1$. Then conditioning on the event $X_{n+1} \in A_k$ and $(i_1, \ldots, i_{n_k})$, the sequence $(\sigma_{i_1}, \ldots, \sigma_{i_{n_k}}, \sigma_{i_{n_k+1}})$ is exchangeable.

We call $\hat{C}_{\text{loc}}^{(\alpha)}$ the conformal optimized prediction set estimator COPS, where the word 'optimized' denotes the effort of minimizing the average set length $\mathbb{E}[\mu\{C_n(X_{n+1})\}]$.

We give a fast approximation algorithm that is analogous to algorithm 1. The resulting approximation also satisfies finite sample local validity as well as asymptotic efficiency as shown in Section 4. See algorithm 2.

Remark 2. In the approach described above, the local conformity measure is $\hat{p}^{(x,y)}(v|A_k)$. In principle one can use any conformity measure that does not need to depend on the partition A_k, as long as the symmetry condition is satisfied. For example, one can use either the estimated joint density $\hat{p}^{(x,y)}(u, v)$ or the estimated conditional density $\hat{p}^{(x,y)}(v|u)$. We note that, when $\text{diam}(A_k)$ is small, these choices of conformity measure are close to each other since $p_X(x)$ and $p(\cdot|x)$ change very little when x varies inside A_k.

Remark 3. Although one can choose any conformity measure, to have local validity the ranking must be based on a local subset of the sample. When A_k is small and the distribution is sufficiently smooth the local sample $(Y_{i_l} : 1 \leqslant l \leqslant n_k)$ approximates independent observations from $p(\cdot|X = x)$ for $x \in A_k$, which can be used to approximate the conditional oracle $C_P(x)$.

3.2.1. Algorithm 2: local sandwich slicer algorithm

(a) Divide $\mathcal{X}$ into bins $A_1, \ldots, A_m$.
(b) Apply algorithm 1 separately on all Y_is within each A_k.
(c) Output $C_n^+(x)$: the resulting set of A_k for all $x \in A_k$.

4. Asymptotic properties

In this section we investigate the asymptotic efficiency of the locally valid prediction band given in equation (21). Again, we focus on cases where $\text{supp}(P_X) = [0, 1]^d$ and $\mathcal{A}$ is a cubic partition with width w_n. The conformity measure is $\hat{p}^{(x,y)}(Y_i|A_k)$ for $x \in A_k$, where $\hat{p}^{(x,y)}(v|A_k)$ is defined as in equation (19) with kernel bandwidth h_n. The argument is similar for other choices of conformity measures that were mentioned in remark 2, such as joint density or conditional density.

4.1. Notation

In the subsequent arguments, $p_X(\cdot)$ denotes the marginal density of X, $p(y|x)$ is the conditional density of Y given $X = x$ and $p(y|A_k)$ is the conditional density of Y given $X \in A_k$. The kernel estimator of $p(y|A_k)$ is denoted by $\hat{p}(\cdot|A_k)$ and $\hat{P}(\cdot|A_k)$ is the empirical distribution of $(Y|X \in A_k)$.

The upper and lower level sets of conditional density $p(y|x)$ are denoted by $L_x(t) \equiv \{y :$

$p(y|x) \geqslant t\}$ and $L_x^l(t) \equiv \{y : p(y|x) \leqslant t\}$ respectively; $\hat{L}_k(t)$, $\hat{L}_k^l(t)$ are the counterparts of $L_x(t)$ and $L_x^l(t)$, defined for $\hat{p}(\cdot|A_k)$. As in the definition of conditional oracle, $t_x^{(\alpha)}$ is the solution to the equation $P_x\{L_x(t)\} = 1 - \alpha$. Its existence and uniqueness are guaranteed if the contour $\{y : p(y|x) = t\}$ has zero measure for all $t > 0$. Finally we let $G_x(t) = P_x\{L_x^l(t)\}$.

4.2. The sandwich lemma

First we show that $\hat{C}_{\text{loc}}^{(\alpha)}(x)$ can be approximated by two plug-in conditional density level sets (lemma 3). For a fixed $A_k \in \mathcal{A}$, conditioning on indices $(i_1, \ldots, i_{n_k})$ of the data in A_k, let $(X_{(k,\alpha)}, Y_{(k,\alpha)})$ be the element of $\{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_{n_k}}, Y_{i_{n_k}})\}$ such that $\hat{p}(Y_{(k,\alpha)}|A_k)$ ranks $\lfloor n_k\alpha \rfloor$ in ascending order among all $\hat{p}(Y_{i_j}|A_k)$, $1 \leqslant j \leqslant n_k$.

Lemma 3 (the sandwich lemma (Lei *et al.*, 2011)). For any fixed $\alpha \in (0, 1)$, if $\hat{C}(x)$ is defined in equation (21) and $\|K\|_\infty = K(0)$, then, for $x \in A_k$, $\hat{C}(x)$ is 'sandwiched' by two plug-in conditional density level sets:

$$\hat{L}_k\{\hat{p}(X_{(k,\alpha)}, Y_{(k,\alpha)}|A_k)\} \subseteq \hat{C}(x) \subseteq \hat{L}_k\{\hat{p}(X_{(k,\alpha)}, Y_{(k,\alpha)}|A_k) - (n_k h_n)^{-1}\psi_K\}, \tag{22}$$

where $\psi_K = \sup_{x,x'} |K(x) - K(x')|$.

The sandwich lemma provides simple and accurate characterization of $\hat{C}_{\text{loc}}^{(\alpha)}(x)$ in terms of plug-in conditional density level sets, which are much easier to estimate. The asymptotic properties of $\hat{C}_{\text{loc}}^{(\alpha)}(x)$ can be obtained by those of the sandwiching level sets.

4.3. Rates of convergence

The following assumption puts boundedness and smoothness conditions on the marginal density p_X, conditional density $p(y|x)$ and its derivatives.

Assumption 1 (regularity of marginal and conditional densities).

(a) The marginal density of X satisfies $0 < b_1 \leqslant p_X(x) \leqslant b_2 < \infty$ for all x in $[0, 1]^d$.
(b) For all x, $p(\cdot|x)$ is in Hölder class $\Sigma(\beta, L)$. Correspondingly, the kernel K is a valid kernel of order β.
(c) For any $0 \leqslant s \leqslant \lfloor\beta\rfloor$, $p^{(s)}(y|x)$ is continuous and uniformly bounded by L for all x and y.
(d) The conditional density is Lipschitz in x: $\|p(\cdot|x) - p(\cdot|x')\|_\infty \leqslant L\|x - x'\|$.

The Hölder class of smooth functions and valid kernels are common concepts in nonparametric density estimation. We give their definitions in Appendix A.1. Assumptions 1, parts (b)–(d), imply that $p(\cdot|A_k)$ is also in a Hölder class and can be estimated well by kernel estimators. Assumption 1, part (d), enables us to approximate $p(\cdot|x)$ by $p(\cdot|A_k)$ for all $x \in A_k$.

The next assumption gives a sufficient regularity condition on the level sets $L_x(t)$.

Assumption 2 (regularity of conditional density level set). There are positive constants ε_0, γ, c_1 and c_2 such that, for all $x \in [0, 1]^d$,

$$c_1\varepsilon^\gamma \leqslant P[\{y : |p(y|x) - t_x^{(\alpha)}| < \varepsilon\}|X = x] \leqslant c_2\varepsilon^\gamma$$

for all $\varepsilon \leqslant \varepsilon_0$. Moreover, $\inf_x t_x^{(\alpha)} \geqslant t_0 > 0$.

Assumption 2 implies that $\mu\{L_x(t_x^{(\alpha)})\} \leqslant \mu\{L_x(t_0)\} \leqslant 1/t_0$. Assumption 2 is related to the notion of the 'γ-exponent' condition that was introduced by Polonik (1995) and widely used in the density level set literature (Tsybakov, 1997; Rigollet and Vert, 2009). It ensures that the conditional density function $p(\cdot|x)$ is neither too flat nor too steep near the contour at

level $t_x^{(\alpha)}$, so the cut-off value $t_x^{(\alpha)}$ and the conditional density level set $C_P(x) = L_x(t_x^{(\alpha)})$ can be approximated from a finite sample. As mentioned in Audibert and Tsybakov (2007), the oracle band $C_P(x)$ is non-empty only if $\gamma(\beta \wedge 1) \leqslant 1$, which holds for the most common case $\gamma = 1$. Assumption 2 also requires that the optimal cut-off values $t_x^{(\alpha)}$ be bounded away from zero.

The following critical rate will be used repeatedly in our analysis:

$$r_n = \left\{ \frac{\log(n)}{n} \right\}^{\beta/\{\beta(d+2)+1\}}. \tag{23}$$

The rate may appear to be non-standard. This is because we are assuming different types of smoothness on y (assumption 1, part (b)) and x (assumption 1, part (d)). This seems to be necessary to achieve both marginal and local validity. More specfically, to achieve marginal and local validity, we use a histogram-like construction over x. However, we used a kernel construction for y given x. Thus, the natural smoothness conditions for theoretical analysis for x are different from those for y given x. This is why the rate is unusual. More traditional rates could be achieved by making assumption 1, part (d), stronger and using a smoother construction. However, we do not know any procedure that uses a smoother construction and still retains finite sample marginal and local validity.

The following theorem gives the convergence rate on the asymptotic efficiency of the locally valid prediction band constructed in Section 3.2.

Theorem 1. Let $\hat{C}_{\mathrm{loc}}^{(\alpha)}$ be the prediction band given by the local conformity procedure as described in equation (21). Choose $w_n \asymp r_n$ and $h_n \asymp r_n^{1/\beta}$. Under assumptions 1 and 2, for any $\lambda > 0$, there is a constant A_λ, such that

$$\mathbb{P}[\sup_{x \in \mathcal{X}} \mu\{\hat{C}_{\mathrm{loc}}^{(\alpha)}(x) \triangle C_P(x)\} \geqslant A_\lambda r_n^{\gamma_1}] = O(n^{-\lambda}),$$

where $\gamma_1 = \min(1, \gamma)$.

Thus, in the common case $\gamma = 1$, the rate is r_n. The following lemma follows easily from the previous result.

Lemma 4. Under assumptions 1 and 2, the local band is asymptotically conditionally valid.

Remark 4. It follows from the proof that the output of algorithm 2 also satisfies the same asymptotic efficiency and conditional validity results.

4.4. Minimax bound

The following theorem says that, in the most common case $\gamma = 1$, the rate given in theorem 1 is indeed minimax rate optimal. Let $\mathcal{P}(\beta, L)$ denote the class of distributions satisfying assumptions 1 and 2 with $\gamma = 1$.

Theorem 2 (lower bound on estimation error). Let $\mathcal{P}_0(\beta, L)$ be the class of distributions on $[0,1]^d \times \mathbb{R}^1$ such that, for each $P \in \mathcal{P}(\beta, L)$, P_X is uniform on $[0,1]^d$ and satisfies assumptions 1 and 2 with $\gamma = 1$. Fix an $\alpha \in (0,1)$; there is a constant $c = c(\alpha, \beta, L, d) > 0$ such that, for all large n,

$$\inf_{\hat{C}_n} \sup_{P \in \mathcal{P}_0(\beta, L)} \mathbb{E}_P[\mu\{\hat{C}_n(x) \triangle C_P(x)\}] \geqslant c r_n,$$

where the infimum is over all estimators $\hat{C}_n$ based on a sample of size n.

Moreover, when $\gamma = 1$, $\mathcal{P}_0(\beta, \Sigma)$ is a subset of $\mathcal{P}(\beta, \Sigma)$ and hence the above result implies a matching minimax rate over $\mathcal{P}(\beta, \Sigma)$ for the case $\gamma = 1$.

5. Tuning parameter selection

In the band given by equation (21), there are two bandwidths to choose: w_n and h_n. Since each bin A_k can use a different h_n to estimate the local marginal density $\hat{p}(\cdot|A_k)$, we can consider $h_{n,k}$, allowing a different kernel bandwidth for each bin.

Since all bandwidths give local validity, we can choose the combination of $(w_n, h_{n,k})$ such that the resulting conformal set has smallest Lesbesgue measure. Such a two-stage procedure of selecting w_n and $h_{n,k}$ from discrete candidate sets $\mathcal{W} = \{w^1, \ldots, w^m\}$ and $\mathcal{H} = \{h^1, \ldots, h^l\}$ is detailed in algorithm 3. To preserve finite sample marginal validity with data-driven bandwidths, we split the sample into two equal-sized subsamples and apply the tuning algorithm on one subsample and use the output bandwidth on the other subsample to obtain the prediction band. Following remark 2, we can use different conformity measures to construct $\hat{C}$. In principle, the above sample splitting procedure works for any conformity measures.

It is straightforward to show that the band $\hat{C}$ constructed as above by using data-driven tuning parameters is locally valid and marginally valid, because the bandwidth (w, h) used is independent of the training data $\mathcal{Z}_2$. From the construction of $\hat{C}$, it will have small excess risk $\mathbb{E}[\mu\{\hat{C}(X)\}] - \mathbb{E}[\mu\{C_P(X)\}]$ if the conformal prediction set is stable under random sampling. Then asymptotic efficiency follows if one can relate the excess risk to the symmetric difference risk. A rigorous argument is beyond the scope of this paper and will be pursued in future work.

5.1. Algorithm 3: bandwidth tuning for COPS

Input data $\mathcal{Z}$, level α and candidate sets $\mathcal{W}$ and $\mathcal{H}$.

(a) Split the data set into two equal-sized subsamples $\mathcal{Z}_1$ and $\mathcal{Z}_2$.
(b) For each $w \in \mathcal{W}$:
  (i) construct partition $\mathcal{A}^w$;
  (ii) for each bin A_k and candidate kernel bandwidth h construct local conformal prediction set $\hat{C}^1_{h,k}$, each at level $1 - \alpha$, using data $\mathcal{Z}_1$;
  (iii) let $h^*_{w,k} = \arg\min_{h \in \mathcal{H}} \mu(\hat{C}^1_{h,k})$, for all k;
  (iv) let $Q(w) = (1/n)\Sigma_k n_k \mu(\hat{C}^1_{h^*_{w,k},k})$.
(c) Choose $\hat{w} = \arg\min Q(w)$; $\hat{h}_{\hat{w},k} = h^*_{\hat{w},k}$.
(d) Construct partition $\mathcal{A}^{\hat{w}}$. For $x \in A_k$, output prediction band $\hat{C}(x) = \hat{C}^2_{\hat{h}_{\hat{w},k},k}$, where $\hat{C}^2_{h,k}$ is the local conformal prediction set estimated from data $\mathcal{Z}_2$ in local set A_k.

6. Data examples

6.1. Synthetic example

The procedure is illustrated by the following example in which $d = 1$, and

$$\begin{aligned} X &\sim \text{Unif}[-1.5, 1.5], \\ (Y|X = x) &\sim 0.5\, N\{f(x) - g(x), \sigma^2(x)\} + 0.5\, N\{f(x) + g(x), \sigma^2(x)\}, \end{aligned} \tag{24}$$

where

$$\begin{aligned} f(x) &= (x-1)^2(x+1), \\ g(x) &= 2\sqrt{(x+0.5)}\, \mathbb{1}(x \geqslant -0.5), \end{aligned}$$

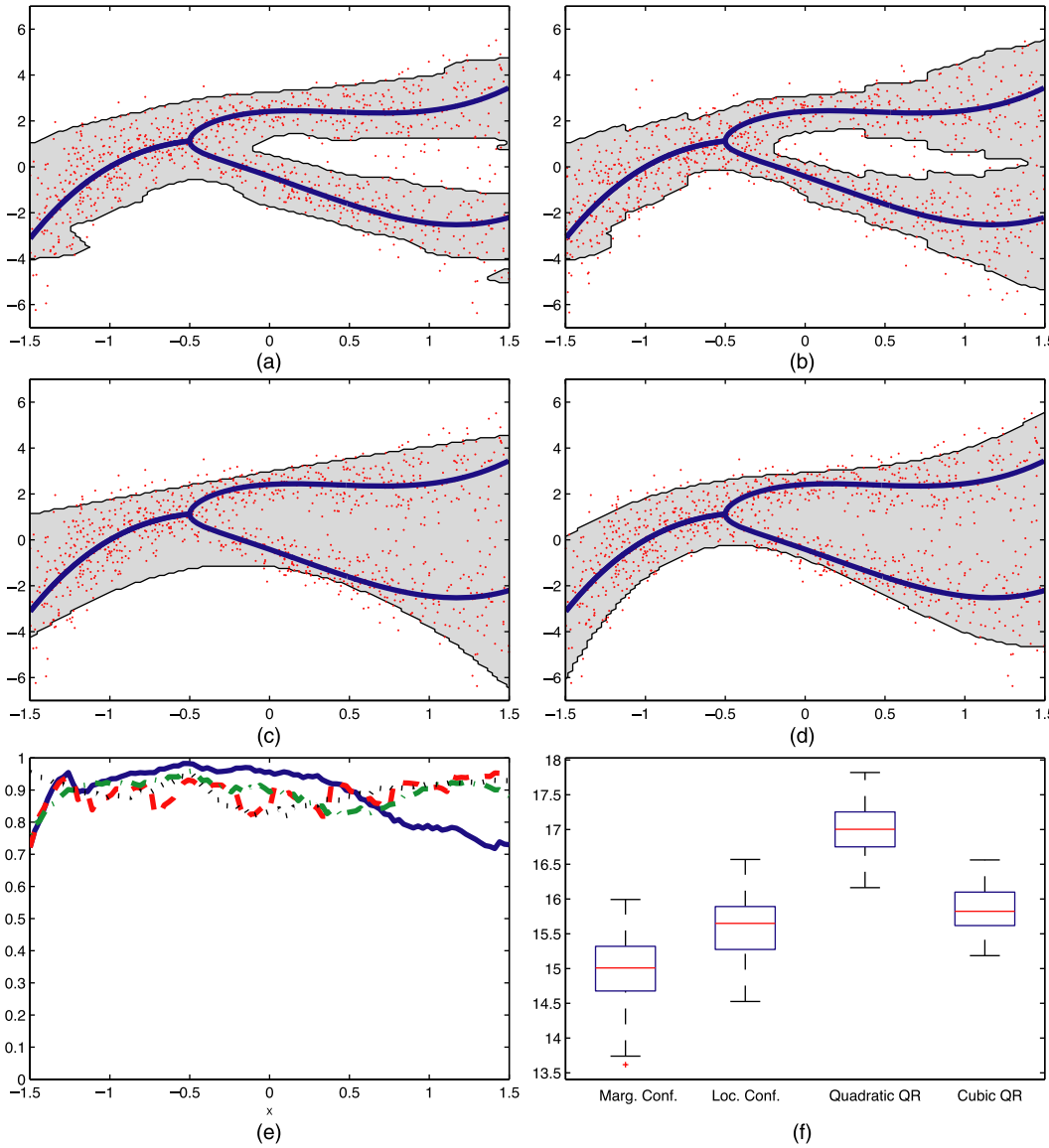

Fig. 3. (a) Marginal conformal prediction bands, (b) local conformal bands, (c) quadratic quantile regression, (d) cubic quantile regression, (e) conditional coverage as a function of x (———, marginal conformal; - - - - - -, local conformal; - · - · - ·, quadratic quantile regression; · · · · · ·, cubic quantile regression) and (f) integrated Lebesgue measure of the prediction regions over 100 repetitions

$$\sigma^2(x) = \tfrac{1}{4} + |x|.$$

For $x \leqslant -0.5$, $(Y|X=x)$ is a Gaussian distribution centred at $f(x)$ with varying variance $\sigma^2(x)$. For $x \geqslant -0.5$, $(Y|X=x)$ is a two-component Gaussian mixture and, for large values of x, the two components have little overlap; Fig. 3.

The performance of prediction bands by using local conformity is plotted and compared with the marginal valid band in Fig. 3, with $n = 1000$ and $\alpha = 0.1$. The conformity measure that is

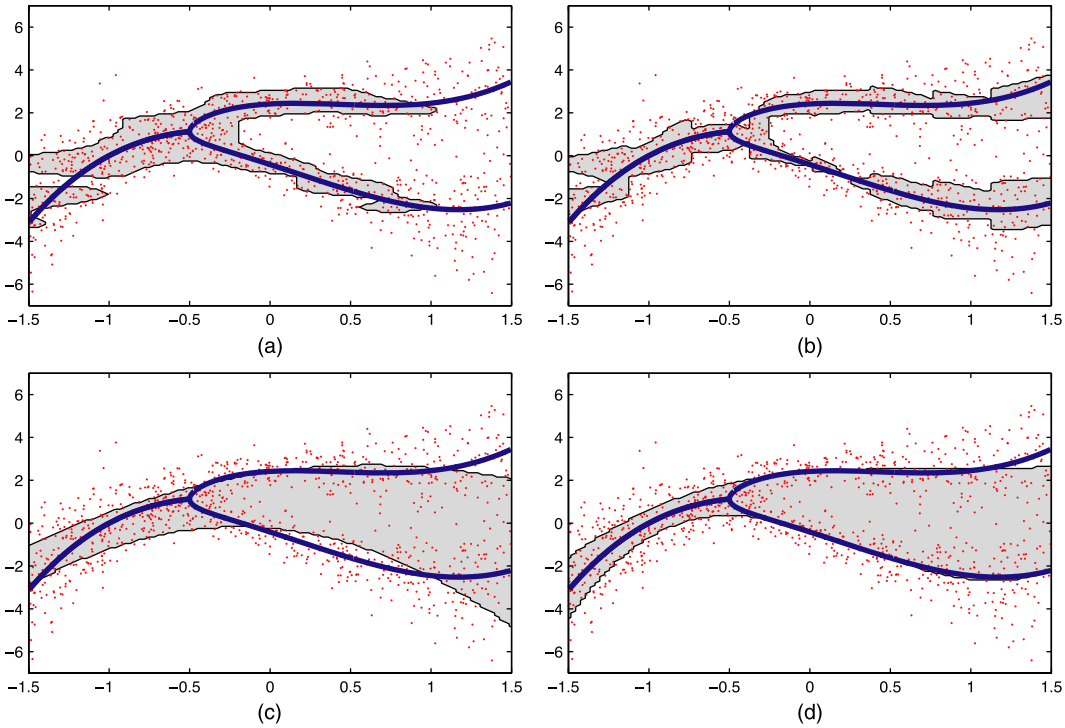

Fig. 4. (a) Marginal conformal prediction bands, (b) local conformal prediction bands, (c) quadratic quantile regression and (d) cubic quantile regression

used here is $\hat{p}^{(x,y)}(Y_i|X_i)$. The locally valid prediction band is constructed by partitioning the support of P_X into 10 equal-sized bins, whereas the marginally valid band is constructed by a global ranking with the same conformity measure. We see that, although the locally valid band has larger Lebesgue measure, it gives the desired coverage for all values x. The marginally valid band overcovers for smaller values of x, and undercovers for larger values of x.

Also shown are quantile regression estimators by assuming quadratic and cubic models. Note that the conformal regions correctly capture the bifurcated structure and have smaller average Lebesgue measure (Fig. 3(f)). And, of course, the conformal method has correct finite sample coverage. Fig. 4 shows results for the case $\alpha = 0.5$. In this case, the difference between the conformal method and quantile regression is more striking.

6.2. *Car data*

Next we consider an example on car mileage. The original data contain features for about 400 cars. For each car, the data consist of miles per gallon, horsepower, engine displacement, size, acceleration, number of cylinders, model year and origin of manufacture. These data have been used in statistics textbooks (e.g. DeGroot and Schervish (2012), chapter 11) to illustrate the art of linear regression analysis. Here we reproduce the linear model that was built in example 11.3.2 of DeGroot and Schervish (2012), where we want to predict the miles per gallon by the horsepower. Clearly, the relationship between miles per gallon and horsepower is far from linear (Fig. 5) so some transformation must be applied before linear model fitting. It makes sense to assume, both from intuition and data plots, that the inverse of miles per gallon, namely gallons per mile, has roughly a linear dependence on the horsepower.

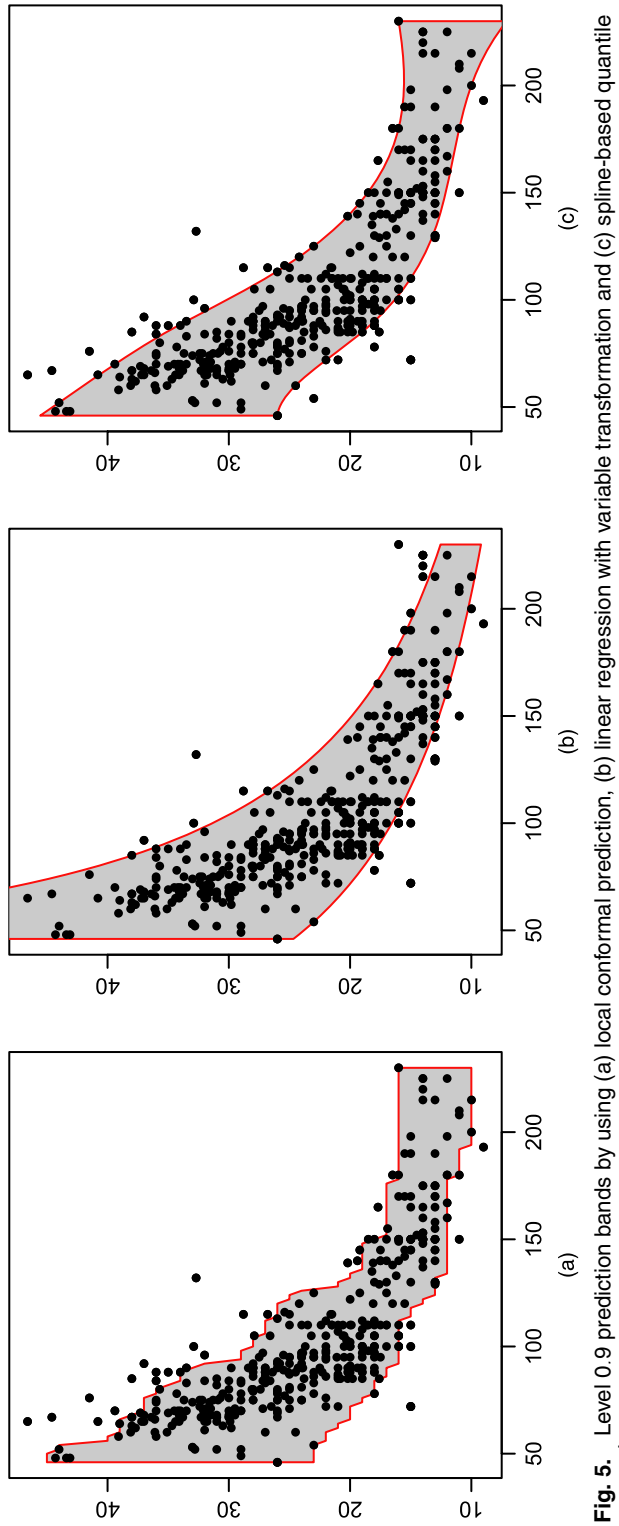

Fig. 5. Level 0.9 prediction bands by using (a) local conformal prediction, (b) linear regression with variable transformation and (c) spline-based quantile regression

In Fig. 5(b) we plot the level 0.9 prediction band that was obtained from the linear regression prediction band. The overall coverage is reasonably close to the nominal level. However, owing to the non-uniform noise level, the band is too wide for small values of horsepower and too narrow for large values. In Fig. 5(a), we plot the non-parametric conformal prediction band by using conformity measure $\hat{p}_h^y(Y_i|X_i)$ to enhance smoothness of the estimated band. Such a

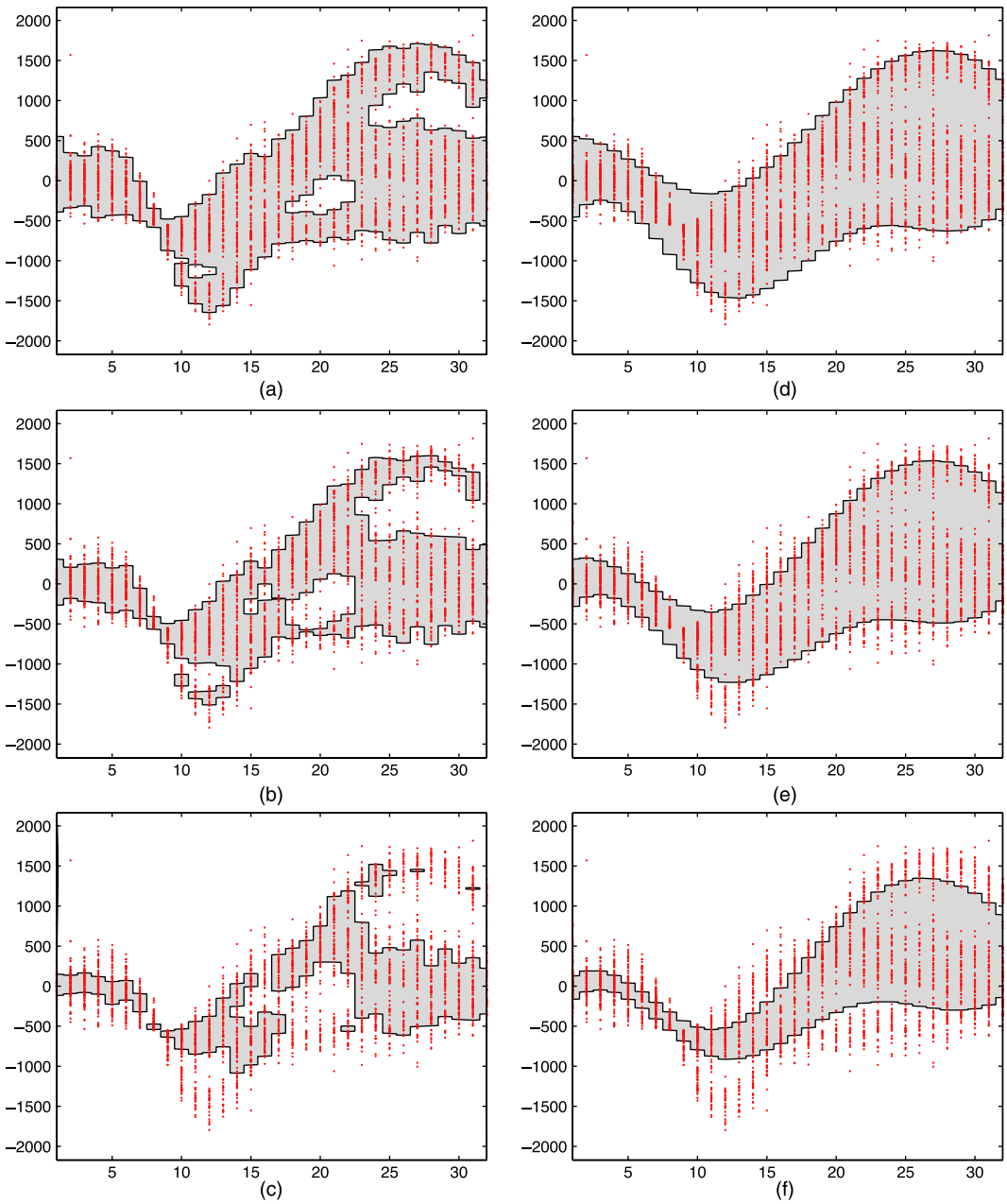

Fig. 6. Conformal prediction bands for (a) $1-\alpha=0.90$, (b) $1-\alpha=0.75$ and (c) $1-\alpha=0.50$ and quantile regressions for (d) $1-\alpha=0.90$, (e) $1-\alpha=0.75$ and (f) $1-\alpha=0.50$

band is asymptotically close to that given in equation (21). The bandwidths are $h_x = 14$ and $h_y = 1.4$. The partition $\mathcal{A}$ is constructed by partitioning the range of horsepower into several intervals to ensure that each set A_k contains roughly the same number of sample points. Here the tuning parameter is the number of partitions and is set to 8. Fig. 5(c) shows spline-based quantile regression. This is similar to the conformal band albeit a little smoother.

The advantage of our method is clear. First, it automatically outputs good prediction bands without involving choosing the variable transformation or specifying a model. The tuning parameters can be chosen in either an automated procedure as described in algorithm 3, or by conventional choices (kernel bandwidth selectors). Second, the conformal prediction band is truly distribution free, with valid coverage for all distributions and all sample sizes.

6.3. Neuron data

Fig. 6 shows an analysis of a macaque monkey's neuron data. (The data were kindly provided by Andy Schwartz, Valerie Ventura and Sonia Todarova.) At each of 32 equally spaced time points, the voltages of 1000 neurons are recorded while a macaque monkey performs a centre-out and out-centre target reaching task with 26 targets in a virtual three-dimensional environment. Technically, these are functional data but for illustration we randomly sampled 10% of the observations so that they can be treated as regression data, with voltage as a response and time as a covariate. The question that we address is how to predict the voltage at a given time.

Figs 6(a)–6(c) are conformal bands for levels $1 - \alpha = 0.90, 0.75, 0.50$ whereas Figs 6(d)–6(f) are quantile regression. We see that the conformal bands are narrower. More importantly, the conformal bands show the structure of the data better. There are clear gaps in the bands, especially for smaller values of α, indicating that the high density regions of the conditional density of Y given X are not connected. The quantile regression approach obscures these features.

7. Final remarks

We have constructed non-parametric prediction bands with finite sample, distribution-free validity. With regularity assumptions, the band is efficient in the sense of achieving the minimax bound. The tuning parameters are completely data driven. We believe that this is the first prediction band with these properties.

An important open question is to establish a rigorous result on the asymptotic efficiency for the data-driven bandwidth. A sketch of such an argument can be given by combining two facts. First, the empirical average excess loss $n^{-1} \Sigma_k n_k \mu(\hat{C}_{h,k})$ is a good approximation to the excess risk $\mathbb{E}[\mu\{\hat{C}(X_{n+1})\}]$ for all w and h. This problem is technically similar to those considered by Rinaldo *et al.* (2010) in the study of stability of plug-in density level sets and prediction sets. Second, one can show that the excess risk provides an upper bound of the symmetric difference risk $\mathbb{E}(\hat{C} \triangle C_P)$, as given in Lei *et al.* (2011) (see also Scott and Nowak (2006)).

The bands are not suitable for high dimensional regression problems. In current work, we are developing methods for constructing prediction bands that exploit sparsity assumptions. These will yield valid prediction and variable selection simultaneously.

Acknowledgements

The authors thank the reviewers for helpful comments and suggestions. The authors also thank Andy Schwartz, Valerie Ventura and Sonia Todarova for providing the neuron data.

Appendix A

A.1. Technical definitions

Now we give formal definitions of some technical terms that are used in the asymptotic analysis, including Hölder class of functions and valid kernel functions of order β. These definitions can be found in standard non-parametric inference textbooks such as Tsybakov (2009), section 1.2.

Definition 2 (Hölder class). Given $L>0$ and $\beta>0$, let $l=\lfloor\beta\rfloor$ be the largest integer strictly less than β. The Hölder class $\Sigma(\beta, L)$ is the family of functions $f:\mathbb{R}^1\mapsto\mathbb{R}^1$ whose derivative $f^{(l)}$ satisfies

$$|f^{(l)}(x)-f^{(l)}(x')|\leqslant L|x-x'|^{\beta-l}, \qquad \forall x, x'.$$

Definition 3 (valid kernels of order β). Let $\beta>0$ and $l=\lfloor\beta\rfloor$. Say that $K:\mathbb{R}^1\mapsto\mathbb{R}^1$ is a valid kernel of order β if the functions $u\mapsto u^j K(u)$, $j=0,\ldots,l$, are integrable and satisfy

$$\int K(u)\,\mathrm{d}u=1,$$

$$\int u^j K(u)\,\mathrm{d}u=0, \qquad j=0,\ldots,l.$$

Remark 5. The relationship between a Hölder class $\Sigma(\beta, L)$ and a valid kernel K of order β is that, for any $p\in\Sigma(\beta, L)$, we have, if $u\mapsto|u|^\beta K(u)$ is integrable, according to proposition 1.2 of Tsybakov (2009), $\|p-p*K_h\|_\infty\leqslant(L/l!)h^\beta\int|u|^\beta|K(u)|\,\mathrm{d}u$, where '$*$' is the convolution operator and $K_h(x)=h^{-1}K(x/h)$.

A.2. Proof of lemma 1

For simplicity we prove the case where $d=1$. For any pair of distributions P and Q let $\mathrm{TV}(P, Q)=\sup_A|P(A)-Q(A)|$ denote the total variation distance between P and Q. Given any $\varepsilon>0$ define $\varepsilon_n=2\{1-(1-\varepsilon^2/8)^{1/n}\}$. From lemma A.1 of Donoho (1988), if $\mathrm{TV}(P, Q)\leqslant\varepsilon_n$ then $\mathrm{TV}(P^n, Q^n)\leqslant\varepsilon$.

Fix $\varepsilon>0$. Let x_0 be a non-atom and choose δ such that $0<P_X\{B(x_0,\delta)\}<\varepsilon_n$ where $\varepsilon_n=2\{1-(1-\varepsilon^2/8)^{1/n}\}$. It follows that $\mathrm{TV}(P^n, Q^n)\leqslant\varepsilon$. Fix $B>0$ and let $B_0=B/\{2(1-\alpha)\}$. Given P, define another distribution Q by $Q(A)=P(A\cap S^{\mathrm{c}})+U(A\cap S)$ where $S=\{(x, y):\ x\in B(x_0,\delta),\ y\in\mathbb{R}\}$ and U has total mass $P(S)$ and is uniform on $\{(x, y):\ x\in B(x_0,\delta),\ |y|<B_0\}$. Note that $P(S)>0$, $Q(S)>0$ and $\mathrm{TV}(P, Q)\leqslant\varepsilon_n$. It follows that $\mathrm{TV}(P^n, Q^n)\leqslant\varepsilon$.

For all $x\in B(x_0,\delta)$, $\int_{C(x)}q(y|x)\,\mathrm{d}y\geqslant 1-\alpha$ implies that $\mu\{C(x)\}\geqslant 2(1-\alpha)B_0=B$. Hence,

$$Q^n[\operatorname*{ess\,sup}_{x\in B(x_0,\delta)}\mu\{C(x)\}\geqslant B]=1.$$

Thus,

$$P^n[\operatorname*{ess\,sup}_{x\in B(x_0,\delta)}\mu\{C(x)\}\geqslant B]\geqslant Q^n[\operatorname*{ess\,sup}_{x\in B(x_0,\delta)}\mu\{C(x)\}\geqslant B]-\varepsilon=1-\varepsilon.$$

Since ε and B are arbitrary, the result follows.

A.3. Proofs of asymptotic efficiency

Lemma 5. Given $\lambda>0$, under condition 1 and $h_n\asymp r_n^{1/\beta}$, there is a numerical constant ξ_λ such that

$$\mathbb{P}\{\sup_k\|\hat{p}(\cdot|A_k)-p(\cdot|A_k)\|_\infty\geqslant\xi_\lambda r_n\}=O(n^{-\lambda}).$$

Proof. For any fixed k, $Y_{i_1},\ldots,Y_{i_{n_k}}$ is a random sample from $P(y|A_k)$ conditioning on n_k. Let $\bar{p}(y|A_k)$ be the convolution density $p(\cdot|A_k)*K_{h_n}(\cdot)$; then, using a result from Giné and Guillou (2002), there are numerical constants C_1, C_2 and ξ_0 such that, for all $\xi\geqslant\xi_0$,

$$\mathbb{P}[\|\hat{p}(\cdot|A_k)-\bar{p}(\cdot|A_k)\|_\infty\geqslant\xi\sqrt{\{\log(n_k)/(n_kh_n)\}}]\leqslant C_1h_n^{C_2\xi^2}. \tag{25}$$

However, by Hölder condition of $p(y|x)$ and hence on $p(\cdot|A_k)$, we have $\|\bar{p}(\cdot|A_k)-p(\cdot|A_k)\|_\infty\leqslant Lh_n^\beta$. Put together with the union bound on all $A_k\in\mathcal{A}_n$

$$\mathbb{P}[\exists k:\|\hat{p}(\cdot|A_k)-p(\cdot|A_k)\|_\infty\geqslant\xi\sqrt{\{\log(n_k)/(n_kh_n)\}}+Lh_n^\beta]\leqslant C_1h_n^{C_2\xi^2}w_n^{-d}.$$

Consider event E_0: $E_0 = \{b_1 n w_n^d/2 \leqslant n_k \leqslant 3b_2 n w_n^d/2, \forall k\}$, where the constants b_1 and b_2 are defined as in assumption 1(a). By lemma 9 we have $\mathbb{P}(E_0^c) \leqslant C_3 w_n^{-d} \exp(-C_4 n w_n^d)$, with constants C_3 and C_4 defined in lemma 9.

On E_0 and for n sufficiently large we have

$$\sqrt{\left\{\frac{\log(n_k)}{n_k h_n}\right\}} \leqslant 2\sqrt{\left[\frac{2\beta+1}{c_1\{\beta(d+2)+1\}}\right]}\sqrt{\left\{\frac{\log(n)}{n w_n^d h_n}\right\}}.$$

When $w_n \asymp r_n$ and $h_n \asymp r_n^{1/\beta}$ as assumed, we have $\sqrt{\{\log(n)/(n w_n^d h_n)\}} = h_n^\beta = r_n$.

Let

$$\xi_\lambda = 2\sqrt{\left[\frac{2\beta+1}{b_1\{\beta(d+2)+1\}}\right]}\left(\sqrt{\left[\frac{\lambda\{\beta(d+2)+1\}+\beta d}{C_2}\right]} \vee \xi_0\right) + L,$$

where the constants b_1 and L are defined in assumption 1, and C_2 is defined in equation (25).

Then we have

$$\begin{aligned}
&\mathbb{P}\{\sup_k \|\hat{p}(\cdot|A_k) - p(\cdot|A_k)\|_\infty \leqslant \xi_\lambda r_n\} \\
&\geqslant \mathbb{P}\left[\sup_k \|\hat{p}(\cdot|A_k) - p(\cdot|A_k)\|_\infty \leqslant (\xi_\lambda - L)\sqrt{\left\{\frac{\log(n)}{n w_n^d h_n}\right\}} + L h_n^\beta, E_0\right] \\
&\geqslant \mathbb{P}\left(\sup_k \|\hat{p}(\cdot|A_k) - p(\cdot|A_k)\|_\infty \leqslant \frac{\xi_\lambda - L}{2\sqrt{((2\beta+1)/[c_1\{\beta(d+2)+1\}])}}\sqrt{\left\{\frac{\log(n_k)}{n_k h_n}\right\}} + L h_n^\beta,\ E_0\right) \\
&\geqslant 1 - \mathbb{P}\left(\exists k : \|\hat{p}(\cdot|A_k) - p(\cdot|A_k)\|_\infty \geqslant \frac{\xi_\lambda - L}{2\sqrt{((2\beta+1)/[c_1\{\beta(d+2)+1\}])}}\sqrt{\left\{\frac{\log(n_k)}{n_k h_n}\right\}} + L h_n^\beta\right) - \mathbb{P}(E_0^c) \\
&= 1 - O(n^{-\lambda}),
\end{aligned}$$

Corollary 1. Let $R_n(x) = \|\hat{p}_n(y|A_k) - p(y|x)\|_\infty$; then, for any $\lambda > 0$, there exists $\xi_{1,\lambda} > 0$ such that

$$\mathbb{P}\{\sup_{x \in B_k} R_n(x) \geqslant \xi_{1,\lambda} r_n\} = O(n^{-\lambda}).$$

Proof. First by Lipschitz condition 1, part (d), on $p(y|x)$,

$$\|p(y|A_k) - p(y|x)\|_\infty \leqslant L w_n \sqrt{d}.$$

Note that $w_n = r_n$; the claim then follows by applying lemma 5 and choosing $\xi_{1,\lambda} = \xi_\lambda + L\sqrt{d}$.

Lemma 6. Let $V_n(x) = \sup_{t \geqslant t_0} |\hat{P}\{L_x^l(t)|A_k\} - P\{L_x^l(t)|x\}|$. Then, for any $\lambda > 0$, there exists $\xi_{2,\lambda}$ such that

$$\mathbb{P}\{\sup_{x \in \mathcal{X}} V_n(x) \geqslant \xi_{2,\lambda} r_n^{\gamma_1}\} = O(n^{-\lambda}),$$

with $\gamma_1 = \min(\gamma, 1)$.

Proof. Consider a fixed A_k and an $x \in A_k$. Note that $\{L_x^l(t) : t \geqslant t_0\}$ is a nested class of sets with Vapnik–Chervonenkis dimension 2. By the Vapnik–Chervonenkis theorem, for all $B > 0$ we have

$$\mathbb{P}\left[\sup_t |\hat{P}\{L_x^l(t)|A_k\} - P\{L_x^l(t)|A_k\}| \geqslant B\sqrt{\left\{\frac{\log(n_k)}{n_k}\right\}}\right] \leqslant C_0 n_k^{-(B^2/32-2)}, \tag{26}$$

for some universal constant C_0. However,

$$\begin{aligned}
|P\{L_x^l(t)|A_k\} - P\{L_x^l(t)|x\}| &= \left|\int_{L_x^l(t)} \{p(y|A_k) - p(y|x)\}\,\mathrm{d}y\right| \\
&\leqslant L w_n \mu\{L_x(t)\}\sqrt{d} \\
&\leqslant L w_n \mu\{L_x(t_0)\}\sqrt{d} \leqslant C L w_n \sqrt{d},
\end{aligned} \tag{27}$$

where $C = t_0^{-1}$ and the last inequality follows from the observation mentioned after assumption 2.

On E_0 we have $\sqrt{\{\log(n_k)/n_k\}} = o(r_n)$ and hence $\sqrt{\{\log(n_k)/n_k\}} \leqslant r_n$ for n sufficiently large. Consider any $x' \in A_k$:

$$\begin{aligned}\left|\hat{P}\{L^l_{x'}(t)|A_k\} - P\{L^l_{x'}(t)|x'\}\right| &\leqslant \left|\hat{P}\{L^l_{x'}(t)|A_k\} - \hat{P}\{L^l_x(t)|A_k\}\right| + \left|\hat{P}\{L^l_x(t)|A_k\} - P\{L^l_x(t)|x\}\right| \\ &\quad + |P\{L^l_x(t)|x\} - P\{L^l_{x'}(t)|x'\}| \\ &\leqslant \|\hat{p}(\cdot|A_k)\|_\infty \mu\{L^l_x(t) \triangle L^l_{x'}(t)\} + V_n(x) + |G_x(t) - G_{x'}(t)| \\ &\leqslant \|\hat{p}(\cdot|A_k)\|_\infty \frac{c_2(2L)^\gamma}{t_0} w_n^\gamma + V_n(x) + CLw_n\sqrt{d} + \frac{2^\gamma c_2 L^{\gamma+1}}{t_0} w_n^\gamma, \end{aligned} \tag{28}$$

where the last step uses lemma 7 below to control $\mu\{L^l_x(t) \triangle L^l_{x'}(t)\}$ and $G_x(t) - G_{x'}(t)$.

Lemma 5 implies that, except for a probability of $O(n^{-\lambda})$, $\sup_k \|\hat{p}(\cdot|A_k)\|_\infty = L + o(1)$ with L defined in assumption 1. Combining inequalities (26), (27) and (28), we have, for some constant $\xi_{2,\lambda}$, $\mathbb{P}\{\sup_x V_n(x) \geqslant \xi_{2,\lambda} r_n^{\gamma_1}\} = O(n^{-\lambda})$, where $\gamma_1 = \min(\gamma, 1)$.

Lemma 7. Under assumptions 1 and 2, $\sup_k \sup_{t \geqslant t_0,\, x,\, x' \in A_k} |G_x(t) - G_{x'}(t)| = O(w_n^{\gamma \wedge 1})$.

Proof.

$$\begin{aligned} L_x(t) \triangle L_{x'}(t) &= \{y : p(y|x) > t, p(y|x') \leqslant t\} \cup \{y : p(y|x) \leqslant t, p(y|x') > t\} \\ &= \{y : t < p(y|x) \leqslant t + Lw_n, p(y|x') \leqslant t\} \cup \{y : t - Lw_n < p(y|x) \leqslant t, p(y|x') > t\} \\ &\subseteq \{y : t - Lw_n < p(y|x) \leqslant t + Lw_n\}, \end{aligned} \tag{29}$$

where the first step uses the fact that $\|p(\cdot|x) - p(\cdot|x')\|_\infty \leqslant L\|x - x'\|$ and the constant L is from assumption 1.

$$\begin{aligned} |G_x(t) - G_{x'}(t)| &\leqslant \left|P\{L^l_x(t)|x\} - P\{L^l_x(t)|x'\}\right| + \left|P\{L^l_x(t)|x'\} - P\{L^l_{x'}(t)|x'\}\right| \\ &= \left|P\{L_x(t)|x\} - P\{L_x(t)|x'\}\right| + \left|P\{L^l_x(t)|x'\} - P\{L^l_{x'}(t)|x'\}\right| \\ &\leqslant \mu\{L_x(t)\}\|p(\cdot|x) - p(\cdot|x')\|_\infty + \|p(\cdot|x')\|_\infty \mu\{L^l_x(t) \triangle L^l_{x'}(t)\} \\ &\leqslant CLw_n\sqrt{d} + L\frac{G_{x'}(t + Lw_n) - G_{x'}(t - Lw_n)}{t_0} \\ &\leqslant CLw_n\sqrt{d} + \frac{2^\gamma c_2 L^{\gamma+1}}{t_0} w_n^\gamma, \end{aligned} \tag{30}$$

where the constant L is from assumption 1, $C = t_0^{-1}$ and (c_2, t_0, γ) are defined in assumption 2.

We complete the argument by using Cadre *et al.* (2009) and Lei *et al.* (2011).

Lemma 8. Fix $\alpha > 0$ and $t_0 > 0$. Suppose that p is a density function satisfying assumption 2. Let $\hat{p}$ be an estimated density such that $\|\hat{p} - p\|_\infty \leqslant \nu_1$, and $\hat{P}$ be a probability measure satisfying $\sup_{t \geqslant t_0} |\hat{P}\{L^l(t)\} - P\{L^l(t)\}| \leqslant \nu_2$. Define $\hat{t}^{(\alpha)} = \inf\{t \geqslant 0 : \hat{P}\{\hat{L}^l(t)\} \geqslant \alpha\}$. If ν_1 and ν_2 are sufficiently small such that $\nu_1 + c_1^{-1/\gamma}\nu_2^{1/\gamma} \leqslant t^{(\alpha)} - t_0$ and $c_1^{-1/\gamma}\nu_2^{1/\gamma} \leqslant \varepsilon_0$ (where c_1 and γ are constants in assumption 2), then

$$|\hat{t}^{(\alpha)} - t^{(\alpha)}| \leqslant \nu_1 + c_1^{-1/\gamma}\nu_2^{1/\gamma}. \tag{31}$$

Moreover, for any $\tilde{t}^{(\alpha)}$ such that $|\tilde{t}^{(\alpha)} - \hat{t}^{(\alpha)}| \leqslant \nu_3$, if $2\nu_1 + c_1^{-1/\gamma}\nu_2^{1/\gamma} + \nu_3 \leqslant \varepsilon_0$, then there are constants ξ_1, ξ_2 and ξ_3 such that $\mu\{\hat{L}(\tilde{t}^{(\alpha)}) \triangle L(t^{(\alpha)})\} \leqslant \xi_1\nu_1^\gamma + \xi_2\nu_2 + \xi_3\nu_3^\gamma$.

Proof. The proof follows essentially from Lei *et al.* (2011), which is a modified version of the argument that was used in Cadre *et al.* (2009).

For $t \geqslant t_0$, let $\hat{L}^l(t) = \{y : \hat{p}(y) \leqslant t\}$. By the assumptions in lemma 8 we have $L^l(t - \nu_1) \subseteq \hat{L}^l(t) \subseteq L^l(t + \nu_1)$ implies that $\hat{P}\{L^l(t - \nu_1)\} \leqslant \hat{P}\{\hat{L}^l(t)\} \leqslant \hat{P}\{L^l(t + \nu_1)\}$ which implies that $P\{L^l(t - \nu_1)\} - \nu_2 \leqslant \hat{P}\{\hat{L}^l(t)\} \leqslant P\{L^l(t + \nu_1)\} + \nu_2$. Hence,

$$\hat{P}\{\hat{L}^l(t^{(\alpha)} - \nu_1 - c_1^{-1/\gamma}\nu_2^{1/\gamma})\} \leqslant P\{L^l(t^{(\alpha)} - c_1^{-1/\gamma}\nu_2^{1/\gamma})\} + \nu_2 \leqslant \alpha,$$

where the last step uses the γ-exponent condition as in assumption 2. Therefore, we must have $\hat{t}^{(\alpha)} \geqslant t^{(\alpha)} - \nu_1 - c_1^{-1/\gamma}\nu_2^{1/\gamma}$. A similar argument gives $\hat{t}^{(\alpha)} \leqslant t^{(\alpha)} + \nu_1 + c_1^{-1/\gamma}\nu_2^{1/\gamma}$. This proves the first part.

For the second part, note that

$$\hat{L}(\tilde{t}^{(\alpha)}) \triangle L(t^{(\alpha)}) = \{y : \hat{p}(y) \geqslant \tilde{t}^{(\alpha)}, p(y) < t^{(\alpha)}\} \cup \{y : \hat{p}(y) < \tilde{t}^{(\alpha)}, p(y) \geqslant t^{(\alpha)}\}.$$

By the assumption on $\tilde{t}^{(\alpha)}$ and the first result,

$$\{\hat{p}(y) \geqslant \tilde{t}^{(\alpha)}\} \subseteq \{p(y) \geqslant t^{(\alpha)} - 2\nu_1 - c_1^{-1/\gamma}\nu_2^{1/\gamma} - \nu_3\},$$
$$\{\hat{p}(y) < \tilde{t}^{(\alpha)}\} \subseteq \{p(y) < t^{(\alpha)} + 2\nu_1 + c_1^{-1/\gamma}\nu_2^{1/\gamma} + \nu_3\}.$$

As a result,

$$\begin{aligned}\mu\{\hat{L}(\tilde{t}^{(\alpha)}) \triangle L(t^{(\alpha)})\} &\leqslant \mu[\{y : |p(y) - t^{(\alpha)}| \leqslant 2\nu_1 + c_1^{-1/\gamma}\nu_2^{1/\gamma} + \nu_3\}] \\ &\leqslant t_0^{-1}c_2(4\nu_1 + 2c_1^{-1/\gamma}\nu_2^{1/\gamma} + 2\nu_3)^{\gamma} \leqslant \xi_1\nu_1^{\gamma} + \xi_2\nu_2 + \xi_3\nu_3^{\gamma},\end{aligned}$$

where (ξ_1, ξ_2, ξ_3) are functions of (t_0, c_1, c_2, γ).

A.3.1. Proof of theorem 1

The proof of theorem 1 is based on a direct application of lemma 8 to the density $p(\cdot|x)$ and the empirical measure $\hat{P}(\cdot|A_k)$ and estimated density function $\hat{p}(\cdot|A_k)$. Here we use $\hat{L}$ for upper level sets of $\hat{p}(\cdot|A_k)$ and omit the dependence on k. Conditioning on $(i_1, \ldots, i_{n_k})$, then one can show that the local conformal prediction set $\hat{C}^{(\alpha)}(x)$ is 'sandwiched' by two estimated level sets:

$$\hat{L}\{\hat{p}(X_{(i_\alpha)}, Y_{(i_\alpha)}|A_k)\} \subseteq \hat{C}^{(\alpha)}(x) \subseteq \hat{L}\{\hat{p}(X_{(i_\alpha)}, Y_{(i_\alpha)}|A_k) - (n_k h_n)^{-1}\psi_K\},$$

where $\psi_K = \sup_{x,x'} |K(x) - K(x')|$. So the asymptotic properties of $\hat{C}^{(\alpha)}(x)$ can be obtained by those of the sandwiching sets.

Recall that $(X_{(i_\alpha)}, Y_{(i_\alpha)})$ is the element of $\{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_{n_k}}, Y_{i_{n_k}})\}$ such that $\hat{p}(Y_{(i_\alpha)}|A_k)$ ranks $\lfloor n_k\alpha \rfloor$ in ascending order among all $\hat{p}(Y_{i_j}|A_k)$, $1 \leqslant j \leqslant n_k$. Let $\hat{t}^{(\alpha)} = \hat{p}(X_{(i_\alpha)}, Y_{(i_\alpha)})$. It is easy to check that

$$\hat{t}^{(\alpha)} = \inf[t \geqslant 0 : \hat{P}\{\hat{L}^{l}(t)|A_k\} \geqslant \alpha].$$

Consider event

$$E_1 = \{\sup_x R_n(x) \leqslant \xi_{1,\lambda} r_n, \sup_x V_n(x) \leqslant \xi_{2,\lambda} r_n^{\gamma_1}\},$$

where ξ_1 and ξ_2 are defined as in the statement of corollary 1 and lemma 6. We have $\mathbb{P}(E_1^{c}) = O(n^{-\lambda})$.

Let $\nu_1 = \xi_{1,\lambda} r_n$ and $\nu_2 = \xi_{2,\lambda} r_n^{\gamma_1}$. Note that $r_n \to 0$ as $n \to \infty$, so for n sufficiently large we have ν_1 and ν_2 satisfying the requirements in lemma 8. Let $\nu_3 = 0$ in this case; then we have, for some constants $\xi'_{1,\lambda}$ and $\xi'_{2,\lambda}$, that

$$\mathbb{P}[\sup_x \mu\{\hat{L}(\hat{t}^{(\alpha)}) \triangle L_x(t^{(\alpha)})\} \geqslant \xi'_{1,\lambda} r_n^{\gamma} + \xi'_{2,\lambda} r_n^{\gamma_1}] = O(n^{-\lambda}),$$

which is equivalent to

$$\mathbb{P}[\sup_x \mu\{\hat{L}(\hat{t}^{(\alpha)}) \triangle L_x(t^{(\alpha)})\} \geqslant \xi'_{\lambda} r_n^{\gamma_1}] = O(n^{-\lambda}),$$

for some constant ξ'_λ independent of n.

Now let $\tilde{t}^{(\alpha)} = \hat{t}^{(\alpha)} - (n_k h_n)^{-1}\psi_K$. Applying lemma 8 with $\nu_3 = \nu_{3,n} = (n_k h_n)^{-1}\psi_K$, we obtain, for some constants $\xi''_{j,\lambda}$, $j = 1, 2, 3$,

$$\mathbb{P}[\mu\{\hat{L}(\hat{t}^{(\alpha)}) \triangle L_x(t^{(\alpha)})\} \geqslant \xi''_{1,\lambda} r_n^{\gamma} + \xi''_{2,\lambda} r_n^{\gamma_1} + \xi''_{3,\lambda} \nu_{3,n}^{\gamma}] = O(n^{-\lambda}).$$

On E_0, $\nu_{3,n} = o(r_n)$, so the above inequality reduces to

$$\mathbb{P}[\mu\{\hat{L}(\hat{t}^{(\alpha)}) \triangle L_x(t^{(\alpha)})\} \geqslant \xi''_{\lambda} r_n^{\gamma_1}] = O(n^{-\lambda}).$$

The conclusion of theorem 1 follows from the sandwiching property:

$$\mu\{\hat{C}^{(\alpha)}(x) \triangle L_x(t_x^{(\alpha)})\} \leqslant \mu\{\hat{L}(\hat{t}^{(\alpha)}) \triangle L_x(t_x^{(\alpha)})\} + \mu\{\hat{L}(\tilde{t}^{(\alpha)}) \triangle L_x(t_x^{(\alpha)})\},$$

where $\hat{t}^{(\alpha)} = \hat{p}(X_{(i_\alpha)}, Y_{(i_\alpha)})$ and $\tilde{t}^{(\alpha)} = \hat{t}^{(\alpha)} - (n_k h_n)^{-1}\psi_K$.

Lemma 9 (lower bound on local sample size). Under assumption 1,

$$\mathbb{P}(\forall k : b_1 n w_n^d/2 \leqslant n_k \leqslant 3b_2 n w_n^d/2) \geqslant 1 - C_3 w_n^{-d}\exp(-C_4 n w_n^d),$$

where $C_3 = 2\,\mathrm{diam}\{\mathrm{supp}(P_X)\}^d$ and $C_4 = b_1^2/(8b_2 + 4b_1/3)$ with b_1 and b_2 defined in assumption 1, part (a).

Proof. Let $p_k = P_X(A_k)$. Use Bernstein's inequality, for each k,

$$\mathbb{P}(|n_k - np_k| \geqslant t) \leqslant \exp\left\{-\frac{t^2/2}{np_k(1-p_k)+t/3}\right\}.$$

The result follows by taking $t = c_1 n w_n^d/2$ and the union bound.

A.4. Proof of theorem 2

We shall use the following version of Fano's method from lemma 3 of Yu (1997). Let $\{P_1, \ldots, P_J\} \subset \mathcal{P}$ be a set of distributions and let $\theta : \mathcal{P} \mapsto \mathcal{D}$. Suppose that $d(\cdot, \cdot) : \mathcal{D}^2 \mapsto \mathbb{R}$ is non-negative and satisfies the triangle inequality. If $d\{\theta(P_j), \theta(P_{j'})\} \geqslant a$ and $\mathrm{KL}(P_i, P_{j'}) \leqslant b$ for all $j \neq j'$, then we have

$$\inf_{\hat{\theta}} \sup_{1 \leqslant j \leqslant J} \mathbb{E}_{P_j}[d\{\hat{\theta}, \theta(P_j)\}] \geqslant \frac{a}{2}\left\{1 - \frac{b + \log(2)}{\log(J)}\right\}, \tag{32}$$

where $\mathrm{KL}(P, P')$ is the Kullback–Leibler distance between P and P' and the infimum is taken over all possible estimators $\hat{\theta}$. In our application, $\theta(P) = \{C_P(x) : x \in [0,1]^d\}$ and

$$d\{\theta(P_1), \theta(P_2)\} = \sup_x \mu\{C_{P_1}(x) \triangle C_{P_2}(x)\}.$$

Now we proceed with the proof of theorem 2. Let p_X be the uniform distribution over $[0,1]^d$ and

$$p_0(y|x) \equiv p(y) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{y^2}{2\sigma^2}\right).$$

For σ sufficiently large, $p_0(y|x) \in \mathcal{P}(\beta, L/2)$. Let $C_0(x) \equiv C_0 = \{y : p(y) > t_0\}$ be such that $\int_{C_0} p(y)\,\mathrm{d}y = 1 - \alpha$. Let y_0 be such that $p(y_0) = t_0$. In particular, $y_0 = \sigma\Phi^{-1}(\alpha/2)$ and $t_0 = p(y_0)$.

Let

$$m_n = \left\lfloor \left\{\frac{n}{\log(n)}\right\}^{1/(d+2+1/\beta)} \right\rfloor,$$

and

$$w_n = \left\{\frac{\log(n)}{n}\right\}^{1/(d+2+1/\beta)}.$$

Then for n sufficiently large there is a constant c_0 such that $w_n \leqslant m_n^{-1} \leqslant c_0 w_n$. We can split $[0,1]^d$ into m_n^d equal-sized cubes of size m_n^{-1} (without much loss of generality, we assume that m_n is an odd integer). The strategy is to construct a collection of 'alternative' conditional distributions $p_1(y|x)$ for $x \in [-w_n/2, w_n/2]^d$. For each $1 \leqslant j \leqslant m_n^d$, the distribution P_j is given by using a uniform marginal of X and using $p_1(y|x - x_j)$ as the conditional density of Y given $X = x$ for x in the jth small cube (where x_j is the centre of the jth cube) and $p(y)$ for all other x.

We first describe the construction of $p_1(y|x)$. Let

$$\kappa(y) = \exp\left(-\frac{1}{1-y^2}\right)\mathbb{1}(|y| \leqslant 1).$$

Then κ is bounded, non-negative and infinitely differentiable and $\kappa \in \Sigma(\beta, \xi)$ for some $\xi > 0$; see Tsybakov (2009), pages 92–93. Hence, $|\kappa^{(l)}(y) - \kappa^{(l)}(y')| \leqslant \xi |y - y'|^{\beta - l}$ for all y and y'.

Define

$$h(x) = w_n \exp\{\kappa(4\|x\|^2/w_n^2)\}$$

Let $c_1 > 0$ be a constant to be chosen later. Consider function

$$p_1(y|x) = p_0(y|x) + c_1 h(x)\,\kappa\left\{\frac{y - y_0}{h^{1/\beta}(x)}\right\} - c_1 h(x)\,\kappa\left\{\frac{y + y_0}{h^{1/\beta}(x)}\right\}. \tag{33}$$

We have the following two observations.

(a) When c_1 is small (and also when n is sufficiently large such that w_n is small compared with $t_0 = p(y_0)$),

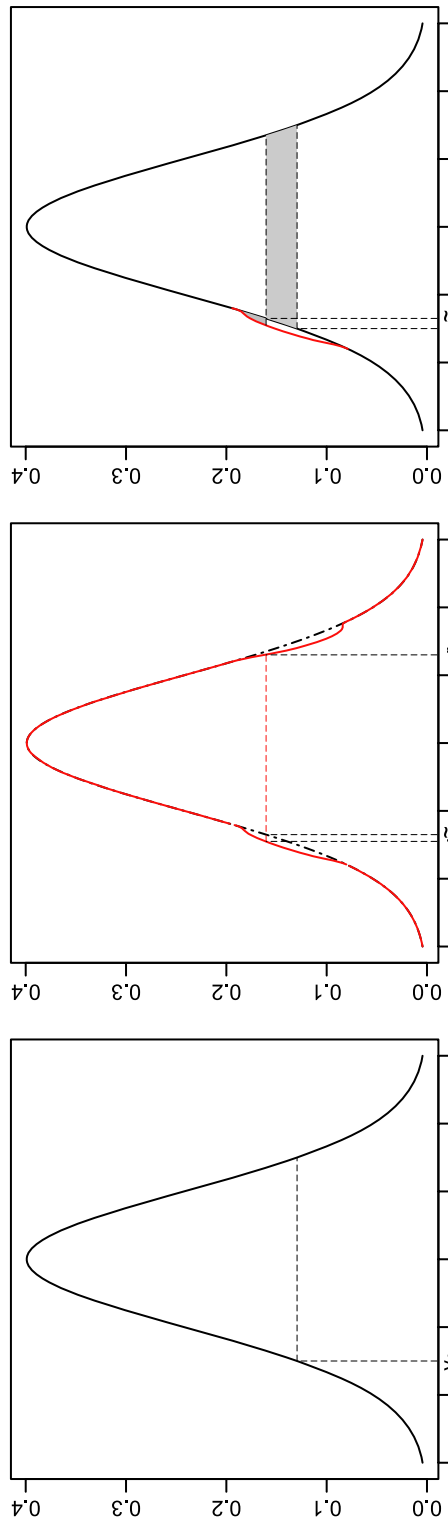

Fig. 7. (a) Density $p_0(\cdot)$ (;, y_0; ------, corresponding $C_0(0)$), (b) $p_1(\cdot|0)$ (;, location of y', $\tilde{y}_0$ and y''; ------, $C_0(1)$) and (c) an explanation of equation (34) (the basic idea is that the upper shaded area should be larger than the lower shaded area (which is not reflected from this picture))

the function $p_1(y|x)$ is non-negative for all (y, x), and hence is a valid density function over y (it is trivial to check that $\int p_1(y|x)\,\mathrm{d}y = 1$).

(b) When $\beta \geqslant 1$, the derivative of $p_1(y|x)$ with respect to y is dominated by $p'(y)$ when the constant c_1 is sufficiently small; therefore $p_1(y|x)$ satisfies assumption 2.

An example of $p(\cdot)$ and $p_1(\cdot|0)$ is given in Figs 7(a) and 7(b).

Next we verify assumption 1. Let $p_0(x, y) = p_0(y|x)\, p_X(x)$ and $p_1(x, y) = p_1(y|x)\, p_X(x)$.

Assumption 3. For $0 < c_1 \leqslant L/(2\xi)$, $p_0(\cdot|x)$ and $p_1(\cdot|x)$ are in $\Sigma(\beta, L)$ for all $x \in [-w_n, w_n]^d$.

Proof. Let $u = (y - y_0)/h_1^{1/\beta}(x)$ and $u' = (y' - y_0)/h_1^{1/\beta}(x)$. When $y_0 \geqslant 2w_n^{1/\beta}$, then, for any y, at most one term in $\kappa\{(y - y_0)/h^{1/\beta}(x)\}$ and $\kappa\{(y + y_0)/h^{1/\beta}(x)\}$ will contribute to the derivative. Taking $\kappa\{(y - y_0)/h^{1/\beta}(x)\}$, we have

$$\begin{aligned}
|p_1^{(l)}(y|x) - p_1^{(l)}(y'|x)| &\leqslant |p_0^{(l)}(y|x) - p_0^{(l)}(y'|x)| + c_1 h^{1-l/\beta}(x)|\kappa^{(l)}(u) - \kappa^{(l)}(u')| \\
&\leqslant \frac{L|y - y'|^{\beta-l}}{2} + h^{1-l/\beta}(x) c_1 \xi |u - u'|^{\beta-l} \\
&\leqslant \frac{L|y - y'|^{\beta-l}}{2} + \frac{L|y - y'|^{\beta-l}}{2} = L|y - y'|^{\beta-l}.
\end{aligned}$$

The same argument also applies to the case of $\kappa\{(y + y_0)/h^{1/\beta}(x)\}$.

Assumption 4. For $j = 0, 1$, $\sup_y |p_1(y|x) - p_1(y|x')| \leqslant L\|x - x'\|$.

Proof. Again, we focus on the part corresponding to the $y - y_0$ part. The same argument also works for the $y + y_0$ part. Note that $|p_0(y|x) - p_0(y|x')| = 0$. Let $\nabla p_1(y|x)$ be the partial derivative of function $p_1(y|x)$ with respect to x. We have, for all large n, that $\|\nabla p_1(y|x)\| \leqslant c_1 c_3$ for some $c_3 > 0$. Choosing c_1 sufficiently small, we have that $\|\nabla p_1(y|x)\| \leqslant L$ for all x and y and all large n. Now

$$p_1(y|x') = p_1(y|x) + (x' - x)^{\mathrm{T}} \nabla p_1(y|\tilde{x})$$

for some $\tilde{x} \in [-1, 1]^d$. Hence, $\sup_y |p_1(y|x) - p_1(y|x')| \leqslant L\|x - x'\|$.

Now we investigate the pairwise separation of between the optimal level sets given by $p_0(y|x)$ and $p_1(y|x)$. Let $C_0(x)$ and $C_1(x)$ be the $1 - \alpha$ optimal prediction sets corresponding to $p_0(y|x)$ and $p_1(y|x)$.

Assumption 5. There is a $c > 0$ such that $\sup_x \mu\{C_0(x) \triangle C_1(x)\} \geqslant c w_n$.

Proof. First, note that $\sup_x \mu\{C_0(x) \triangle C_1(x)\} \geqslant \mu\{C_0(0) \triangle C_1(0)\}$. So fix $x = 0$.

Note that $C_0(0) = (y_0, -y_0)$. By the monotonicity of $p_1(y|0)$ we have, for some $t' > 0$, $C_1(0) = \{y : p_1(y|0) > t'\} = (y', y'')$, where y' and y'' are two numbers that are close to y_0 and $-y_0$ respectively when n is large.

Obviously $t' = p_1(y'|0)$ and we can define $\tilde{y}_0 = g^{-1}(t')$ satisfying $\tilde{y}_0 < 0$. Note that the quantities y', y'' and $\tilde{y}_0$ all depend on n. Then we have $|\tilde{y}_0 - y_0| \leqslant w_n^{1/\beta} = o(1)$. We shall show that $|\tilde{y}_0 - y_0| = O(w_n^{1+1/\beta})$. First consider the case $\tilde{y}_0 \geqslant y_0$. Comparing Figs 7(a) and 7(b), we have the inequality

$$\begin{aligned}
1 - \alpha &= \int_{y'}^{y''} p_1(y|0)\,\mathrm{d}y \\
&\leqslant \int_{y_0}^{-y_0} p(y)\,\mathrm{d}y - 2|y_0|\, p'(y_0)\{1 + o(1)\}(\tilde{y}_0 - y_0) + c_1 w_n^{1+1/\beta} \int \kappa(u)\,\mathrm{d}u,
\end{aligned} \tag{34}$$

where the second term is a lower bound of the loss of coverage comparing the ideal level set of $p_1(y|0)$ with that of $p(y)$ because of the use of a higher cut-off value (roughly corresponding to the lower shaded area in Fig. 7(c)), and the third term is an upper bound of the gain in coverage by adding a small bump of $c_1 w_n \kappa(\cdot/w_n^{1/\beta})$ at y_0 (corresponding to the upper shaded area in Fig. 7(c)).

Inequality (34) can be rewritten as $(\tilde{y}_0 - y_0)_+ = O(w_n^{1+1/\beta})$. Similarly we can show that $(\tilde{y}_0 - y_0)_- = O(w_n^{1+1/\beta})$. Let $\Delta_n = |y' - y_0|$. Combining the above argument with the fact that $p_1(y'|0) = p(\tilde{y}_0)$, we have

$$p(y') + c_1 w_n\, \kappa\left(\frac{\Delta_n}{w_n^{1/\beta}}\right) = p(y_0) + O(w_n^{1+1/\beta}). \tag{35}$$

Now we argue by contradiction. Suppose that there is a subsequence Δ_{k_n} such that $\Delta_{k_n}/w_{k_n} \to 0$. Then $\kappa(\Delta_{k_n}/w_{k_n}^{1/\beta}) \asymp 1$ since $\Delta_{k_n}/w_{k_n}^{1/\beta} = o(1)$. This contradicts equation (35) because $p(y') - p(y_0) = o(w_{k_n})$ along the subsequence k_n. Thus we have proved that there is a constant $c > 0$ such that $\Delta_n \geqslant c w_n$. The result claimed follows by observing that $\mu\{C_0(0) \triangle C_1(0)\} \geqslant \Delta_n$.

Next we bound the Kullback–Leibler deviation between distributions P_j and $P_{j'}$, for $j, j' = 1, \ldots, m_n^d$.

Assumption 6. There is a $C > 0$ such that $\mathrm{KL}(P_j, P_{j'}) \leqslant 2Cc_1^2 w_n^{2+1/\beta+d}$.

Proof. Recall that we assume that n is sufficiently large such that $w_n^{1/\beta} \leqslant |y_0|/2$. Let $c_4 = p(3y_0/2)$; then

$$\begin{aligned}
\int p(y) \log\left\{\frac{p(y)}{p_1(y|x)}\right\} \mathrm{d}y &= -\int_{y_0 - h^{1/\beta}(x)}^{y_0 + h^{1/\beta}(x)} \log\left[1 + \frac{c_1 h(x) \kappa\{(y - y_0)/h^{1/\beta}(x)\}}{p(y)}\right] p(y)\,\mathrm{d}y \\
&\quad - \int_{-y_0 - h^{1/\beta}(x)}^{-y_0 + h^{1/\beta}(x)} \log\left[1 - \frac{c_1 h(x) \kappa\{(y + y_0)/h^{1/\beta}(x)\}}{p(y)}\right] p(y)\,\mathrm{d}y \\
&\leqslant -\int_{y_0 - h^{1/\beta}(x)}^{y_0 + h^{1/\beta}(x)} \left[\frac{c_1 h(x) \kappa\{(y - y_0)/h^{1/\beta}(x)\}}{p(y)} - \frac{c_1^2 h^2(x) \kappa^2\{(y - y_0)/h^{1/\beta}(x)\}}{p^2(y)}\right] p(y)\,\mathrm{d}y \\
&\quad + \int_{-y_0 - h^{1/\beta}(x)}^{-y_0 + h^{1/\beta}(x)} \left[\frac{c_1 h(x) \kappa\{(y + y_0)/h^{1/\beta}(x)\}}{p(y)} + \frac{c_1^2 h^2(x) \kappa^2\{(y + y_0)/h^{1/\beta}(x)\}}{p^2(y)}\right] p(y)\,\mathrm{d}y \\
&\leqslant \frac{2c_1^2}{c_4} h^{2+1/\beta}(x) \int_{-1}^{1} K^2(u)\,\mathrm{d}u \leqslant Cc_1^2 w_n^{2+1/\beta}.
\end{aligned}$$

As a result we have

$$\mathrm{KL}(P_j, P_{j'}) = \int\int p(y) \log\left\{\frac{p(y)}{p_1(y|x)}\right\} \mathrm{d}y\,\mathrm{d}x \leqslant 2Cc_1^2 w_n^{2+1/\beta+d}.$$

Now we apply Fano's lemma. The above argument suggests that we can use $a = cw_n$ as suggested in assumption 5, $b = 2nCc_1^2 w_n^{2+1/\beta+d}$ and $J = m_n^d$ in equation (32). Also observe that $m_n \asymp 1/w_n$. By selecting a sufficiently small c_1 (still independent of n), we can make $1 - \{2Cc_1^2 w_n^{2+1/\beta+d} + \log(2)\}/d \log(m_n)$ bounded away from zero by a constant and hence complete the proof.

References

Akritas, M. G. and Van Keilegom, I. (2001) Non-parametric estimation of the residual distribution. *Scand. J. Statist.*, **28**, 549–567.

Audibert, J. and Tsybakov, A. (2007) Fast learning rate for plug-in classifiers. *Ann. Statist.*, **35**, 608–633.

Cadre, B., Pelletier, B. and Pudlo, P. (2009) Clustering by estimation of density level sets at a fixed probability. *Manuscript*. Institut de Recherche Mathématique de Rennes, Rennes. (Available from `http://hal.archives-ouvertes.fr/hal-00397437/`.)

Carroll, R. J. and Ruppert, D. (1991) Prediction and tolerance intervals with transformation and/or weighting. *Technometrics*, **33**, 197–210.

Davidian, M. and Carroll, R. J. (1987) Variance function estimation. *J. Am. Statist. Ass.*, **82**, 1079–1091.

DeGroot, M. and Schervish, M. (2012) *Probability and Statistics*, 4th edn. Reading: Addison-Wesley.

Donoho, D. L. (1988) One-sided inference about functionals of a density. *Ann. Statist.*, **16**, 1390–1420.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Genovese, C. and Wasserman, L. (2008) Adaptive confidence bands. *Ann. Statist.*, **36**, 875–905.

Giné, E. and Guillou, A. (2002) Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincare* B, **38**, 907–921.

Hall, P. and Rieck, A. (2001) Improving coverage accuracy of nonparametric prediction intervals. *J. R. Statist. Soc.* B, **63**, 717–725.

Koenker, R. and Hallock, K. (2001) Quantile regression. *J. Econ. Perspect.*, **15**, 143–156.

Lei, J., Robins, J. and Wasserman, L. (2011) Distribution free prediction sets. *J. Am. Statist. Ass*, to be published.

Loader, C. (1999) *Local Regression and Likelihood*. New York: Springer.

Low, M. (1997) On nonparametric confidence intervals. *Ann. Statist.*, **25**, 2547–2554.

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

Rigollet, P. and Vert, R. (2009) Optimal rates for plug-in estimators of density level sets. *Bernoulli*, **14**, 1154–1178.

Rinaldo, A., Singh, A., Nugent, R. and Wasserman, L. (2012) Stability of density-based clustering. *J. Mach. Learn. Res.*, **13**, 905–948.
Ruppert, D., Wand, M. and Carroll, R. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Scott, C. D. and Nowak, R. D. (2006) Learning minimum volume sets. *J. Mach. Learn. Res.*, **7**, 665–704.
Shafer, G. and Vovk, V. (2008) A tutorial on conformal prediction. *J. Mach. Learn. Res.*, **9**, 371–421.
Tsybakov, A. (1997) On nonparametric estimation of density level sets. *Ann. Statist.*, **25**, 948–969.
Tsybakov, A. (2009) *Introduction to Nonparametric Estimation*. New York: Springer.
Tukey, J. (1947) Nonparametric estimation: II, Statistical equivalent blocks and multivariate tolerance regions. *Ann. Math. Statist.*, **18**, 529–539.
Vovk, V., Gammerman, A. and Shafer, G. (2005) *Algorithmic Learning in a Random World*. New York: Springer.
Vovk, V., Nouretdinov, I. and Gammerman, A. (2009) On-line predictive linear regression. *Ann. Statist.*, **37**, 1566–1590.
Yu, B. (1997) Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (eds D. Pollard, E. Torgersen and G. L. Yang), pp. 423–435. New York: Springer.