

PRIM ANALYSIS

Wolfgang Polonik

Zailong Wang

Department of Statistics

Mathematical Biosciences Institute

University of California

The Ohio State University

One Shields Ave.

231 West 18th Avenue, #216

Davis, CA 95616-8705

Columbus, OH 43210-1174

July 24, 2007

Abstract

This paper analysis a data mining/bump hunting technique known as PRIM (Fisher and Friedman, 1999). PRIM finds regions in high-dimensional input space with large values of a real output variable. This paper provides the first thorough study of statistical properties of PRIM. Amongst others, we characterize the output regions PRIM produces, and derive some rates of convergence for these regions. Since the dimension of the input variables is allowed to grow with the sample size, the presented results provide some insight about the qualitative behavior of PRIM in very high dimensions. Our investigations also reveal some shortcomings of PRIM, resulting in some proposals for modifications.

The research is support by NSF grant #0406431.

AMS 2000 subject classifications. Primary 62G20, 62G05, 62H12.

Key words and phrases. Asymptotics, bump hunting, data mining, peeling+jittering, VC-classes

1 Introduction

PRIM (Patient Rule Induction Method) is a data mining technique introduced by Friedman and Fisher (1999). Its objective is to find subregions in the input space with relatively high (low) values for the target variable. By construction, PRIM directly targets these regions rather than indirectly through the estimation of a regression function. The method is such that these subregions can be described by simple rules, as the subregions are (unions of) rectangles in the input space.

There are many practical problems where finding such rectangular subregions with relatively high (low) values of the target variable is of considerable interest. Often these are problems where a decision maker wants to choose the values or ranges of the input variables so as to optimize the value of the target variable. Such types of applications can be found in the fields of medical research, financial risk analysis, and social sciences, and PRIM has been applied to these fields.

While PRIM enjoys some popularity, and even several modifications have been proposed (see Becker and Fahrmeier, 2001, Cole, Galic and Zack, 2003, Leblanc et al, 2003, Wu and Chipman, 2003, and Wang et al, 2004), there is according to our knowledge no thorough study of its basic statistical properties. The purpose of this paper is to contribute such a study in order to deepen the understanding of PRIM. Our study also reveals some shortcomings of the algorithm, and proposes some remedies aimed at fixing these shortcomings. Moreover, the methodology developed here should be useful in studying the modifications of PRIM. In particular, we

- provide a rigorous framework for PRIM,
- describe theoretical counterparts of PRIM outcomes,
- derive large sample properties for PRIM outcomes, thereby allowing the dimension of the input space to increase with sample size. These large sample results also provide some information on the choice of one of the tuning parameters involved. Last but not least, we also
- reveal some shortcomings of PRIM and propose remedies.

A formal setup is as follows. Let (X, Y) be a random vector in $d + 1$ dimensional Euclidean space such that $Y \in \mathbb{R}$ is integrable. Suppose that $X \sim F$ with pdf f which is assumed to be continuous throughout the whole paper. Further let m denote the regression function $m(x) := E[Y | X = x]$, $x \in \mathbb{R}^d$. Without loss of generality we assume throughout the paper that $m(x) \geq 0$. Assume that F has support $S \subset \mathbb{R}^d$ also called the input space. Put

$$I(C) := \int_C m(x) dF(x) \quad \text{and} \quad F(C) := \int_C dF(x) \quad C \subset S.$$

The objective of PRIM is to find a subregion $C \subset S$ for which

$$\frac{I(C)}{F(C)} > \lambda$$

where λ is a pre-specified threshold value. This is equivalent to finding sets C with

$$I(C) - \lambda F(C) = \int_C (m(x) - \lambda) dF(x) \geq 0. \quad (1.1)$$

From this point of view an ‘optimal’ outcome (maximizing $I(C) - \lambda F(C)$) is a regression level set

$$C(\lambda) = \{x : m(x) > \lambda\}.$$

Thus it could be said that the conceptual idea behind PRIM is to estimate (or approximate) regression level sets, and this motivation is quite intuitive, as is the algorithm itself.

In order to understand the conceptual idea behind the actual algorithm underlying PRIM, notice that each subset A of $C(\lambda)$ also has the property that $\frac{I(A)}{F(A)} > \lambda$ and each subset A of $S \setminus C(\lambda)$ satisfies $\frac{I(A)}{F(A)} \leq \lambda$. Hence, as an idea for an algorithm to approximate level sets, one might think about iteratively finding ‘small’ (disjoint) subsets B_k satisfying $\frac{I(B_k)}{F(B_k)} > \lambda$, and to use the union of those sets as an approximation of $C(\lambda)$. In fact, this is what the PRIM algorithm is attempting to do. In a greedy fashion the PRIM algorithm iteratively constructs ‘optimal’ axis parallel

rectangles (or boxes) B_1^*, \dots, B_K^* satisfying

$$B_k^* \in \underset{F(B|S \setminus \bigcup_{j=1}^{k-1} B_j^*) = \beta_0}{\operatorname{argmax}} I(B \setminus \bigcup_{j=1}^{k-1} B_j^*), \quad k = 1, \dots, K, \quad (1.2)$$

where β_0 is a (small) tuning parameter to be chosen, and for a set A we denote by $F(\cdot|A)$ the conditional distribution of F given A . Hence, PRIM iteratively construct ‘small’ sets B_k^* , each times removing the outcome B_{k-1}^* of the preceding step and applying the algorithm to what is left over (details are described below). This results in a partition of the input space. The final outcome, R_λ^* , consists of the union of those boxes B_k^* with average $\frac{I(B_k^* \setminus \bigcup_{j=1}^{k-1} B_j^*)}{F(B_k^* \setminus \bigcup_{j=1}^{k-1} B_j^*)}$ exceeding λ .

However, as Figure 1 shows, PRIM in general does *not* approximate level sets, and the individual sets B_j^* are not really ‘small’ in the sense of ‘local’.

PUT FIGURE 1 HERE

(showing a unimodal regression function an some *nested* sets B_k)

The theoretical considerations in the paper mainly concentrate on two substeps of the PRIM algorithm called peeling+jittering (a modification of peeling+pasting to be proposed below), where the empirical version is based on iid observations $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots, n$, from the same populations as (X, Y) . Our main theoretical results provide rates of convergence of outcome regions of empirical PRIM to their theoretical counterparts for a given β_0 . For instance, we will derive conditions under which we have the following type of result. Let \widehat{B}_k , $k = 1, \dots, K$ denote the outcomes of K successive applications of the empirical peeling+jittering,

Suppose that $E|Y|^\gamma < \infty$ for some $\gamma \geq 3$. Let $0 < \beta_0 < 1$ and λ be fixed. Choose the peeling parameter $\alpha = \alpha_n \sim (\frac{d}{n} \log n)^{\frac{1}{3}}$, and suppose that $\frac{d}{n} \log n = o(1)$. Then, under additional assumptions (cf. Theorem 4.3) there exist optimal outcomes B_k^ , $k = 1, \dots, K$ such that*

$$d_F(R_\lambda^*, \widehat{R}_\lambda) = O_P\left(\left(\frac{d^4}{n} \log n\right)^{1/3}\right). \quad (1.3)$$

where $\widehat{R}_\lambda = \bigcup_{\widehat{K}(\lambda)} \widehat{B}_k$ where $\widehat{K}(\lambda) = \{k : \frac{I_n(\widehat{B}_k \setminus \bigcup_{i=1}^{k-1} \widehat{B}_{k-1})}{F_n(\widehat{B}_k \setminus \bigcup_{i=1}^{k-1} \widehat{B}_{k-1})} > \lambda\}$ and $R_\lambda^* = R_\lambda^*(B_1^*, \dots, B_K^*)$ is a theoretical counterpart (cf. Section 4.3).

Here I_n and F_n denote empirical counterparts to I and F , respectively. (See below for precise definitions.) Notice that this result just asserts that there exists an optimal region R_λ^* that is approximated by the peeling+jittering outcome. Except for very special cases (e.g. a unimodal regression function with a uniform F) we cannot hope for a unique optimal outcome R_λ^* , and the above type of result is the best one can hope for. We will, however, present a description of the possible sets B_k^* . It also should be noted that the sets \widehat{B}_k are closely related to so-called minimum volume sets. Rates of convergence of the order $(\frac{\log n}{n})^{1/3}$ have been derived for d -dimensional minimum volume ellipsoids and other minimum volume sets in so-called Vapnik-Cervonenkis (VC) classes with d fixed (see Polonik, 1997, and references therein). This motivates the rates of convergence derived in this paper, since boxes (or rectangles) in \mathbb{R}^d form a VC-class.

Section 3 explores the outcomes of peeling+jittering, thereby also revealing some shortcomings of PRIM. Before we do that, we first describe PRIM in some more detail (Section 2). This is necessary in order to understand the derivations of the large sample results, which are presented in Section 4. These results indicate that tuning of parameters involved in PRIM (see Section 2) should depend on the dimension as well as on moment conditions in terms of the output variable. Proofs of some technical results can be found in Section 5 where also some technical tools are formulated which are needed in the proofs of the main results. Notice again that while the PRIM algorithm is designed to be applicable for both discrete and continuous X -variables, we only study the continuous case.

2 The PRIM algorithm

Peeling. The class of all closed d -dimensional boxes, or axis parallel rectangles

$B \subset [0, 1]^d$ is denoted by \mathcal{B} . Given a subset $S \subseteq [0, 1]^d$ and a value β_0 the goal of peeling is to find

$$B_{\beta_0}^* = \arg \max_{B \subset B} \left\{ \frac{I(B \cap S)}{F(B \cap S)} : F(B|S) = \beta_0 \right\}, \quad (2.1)$$

where $F(\cdot|S)$ denotes the conditional distribution of X given $X \in S$. We always assume that such a set $B_{\beta_0}^*$ exists. Beginning with $B = S$, at each peeling step a small subbox $b \subset B$ is removed. The subbox to be removed is chosen among $2d$ candidate subboxes given by $b_{j1} := \{x \in B : x_j < x_{j(\alpha)}\}$, $b_{j2} := \{x \in B : x_j > x_{j(1-\alpha)}\}$, $j = 1, \dots, d$, where $0 < \alpha < 1$ is another tuning parameter, and $x_{j(\alpha)}$ denotes the α -quantile of $F_j(\cdot|B \cap S)$, the marginal cdf of X_j conditional on $X_j \in B \cap S$. The particular subbox b^* chosen for removal is the one that yields the largest target value (I -measure) among $B \setminus b_{j\pm}$, $j = 1, \dots, d$. The current box is then updated, i.e. B is replaced by $B \setminus b^*$ and the procedure is repeated on this new, smaller box. Notice that the conditional distribution in each current box is used. Hence, in the k th-step the candidate boxes b for removal satisfy $F(b|S) = \alpha(1-\alpha)^{k-1}$. Peeling continues as long as the current box B satisfies $F(B|S) \geq \beta_0$.

The quantity α is usually taken to be quite small so that in each step only a small part of the space in the current box is peeled off (hence the terminology *patient* rule induction). That α cannot be too small is indicated in our theoretical results.

Pasting, has been proposed in order to readjust the outcomes of the peeling strategy. The procedure for pasting is basically the inverse of the peeling procedure. Starting with the peeling solution, the current box is enlarged until the next pasting will cause the average $I(B \cap S)/F(B \cap S)$ in the box to decrease.

Covering. The final output region R^* of the PRIM algorithm is a union of boxes from iterative applications of the peeling procedure. The first box \tilde{B}_1 is constructed using $S = [0, 1]^d$ as described above. The second optimal box \tilde{B}_2 is constructed in the same fashion but using the conditional distribution of F given $[0, 1] \setminus \tilde{B}_1$, and so on, each time removing the optimal outcome of the previous step. The final result

of the PRIM algorithm then is

$$R_\lambda = \bigcup_{\overline{m}_{\tilde{B}_k} \geq \lambda} \tilde{B}_k, \quad (2.2)$$

where $\overline{m}_{\tilde{B}_k} = \frac{I(\tilde{B}_k \setminus \bigcup_{i=1}^{k-1} \tilde{B}_i)}{F(\tilde{B}_k \setminus \bigcup_{i=1}^{k-1} \tilde{B}_i)}$, and λ is pre-specified.

2.1 Jittering

The pasting procedure has the disadvantage that the size (measured by F -measure) of the box resulting from the peeling procedure cannot be controlled, and under certain circumstances this might lead to a relatively large set to be removed in one peeling+pasting step. We propose to replace pasting by what we call *jittering*. Rather than just adding small sets as done in the pasting procedure, we simultaneously add and subtract a box from the $2d$ candidate boxes, as long as we can increase the average of the box. This does not change the F -measure of the box. Of course, the complexity of the algorithm is somewhat increased by doing so.

Jittering is quite important for the below results. It actually enables us to derive a characterization of the boxes resulting from peeling + jittering (cf. Lemma 3.1). This fact makes the use of jittering (rather than pasting) attractive from both a theoretical and a practical perspective. As for the theory, this characterization enables us to derive large sample results for the PRIM outcomes (see below). Another advantage of jittering shows when realizing that peeling might end up in a local minimum. Assuming that this happens, pasting would tend to enlarge the peeling outcome quite significantly. While it might be argued that the covering step following peeling+pasting, or peeling+jittering might eventually remove this set from consideration (since the average of this set might be too low), there is a clear potential that this relatively large set contains interesting parts which in fact carry a high mass concentration. For instance, potential modal regions might be ‘eroded’ from below.

2.2 The empirical version

By definition of $I(C)$ we have

$$I(C) = E\{Y \mathbf{1}\{X \in C\}\}. \quad (2.3)$$

Hence, if $(X_i, Y_i), 1 \leq i \leq n$, is an independent sample with the same distribution as (X, Y) , the empirical analog of I is given by

$$I_n(C) = \frac{1}{n} \sum_{i=1}^n Y_i \mathbf{1}\{X_i \in C\}. \quad (2.4)$$

The empirical analog to F is given by F_n , the empirical distribution of X_1, \dots, X_n . Then the actual PRIM algorithm is performed as described above but with I and F replaced by their empirical versions I_n and F_n , respectively. We replace $\alpha = \alpha_n$ by $\lceil n\alpha_n \rceil/n$, the smallest rational number k/n which is larger than α_n , and we make similar changes to the quantities $\alpha_n(1 - \alpha_n)^j, j = 1, \dots$ and β_0 .

3 PRIM Outcomes

Here we provide a characterization of PRIM outcomes. Some discussions and examples provide some further understanding for what these optimal outcomes are.

Local maximizers.

For a given $O \subseteq [0, 1]^d$ with O open in $S \subseteq [0, 1]^d$ let

$$B_{O, \beta_0}^* = \arg \max \left\{ \frac{I(B \cap S)}{F(B \cap S)}; F(B|S) = \beta_0, B \subseteq O, B \in \mathcal{B} \right\}, \quad (3.1)$$

and define the class of all *local maximizers* of the average $\frac{I(B \cap S)}{F(B \cap S)}$ of size β_0 as

$$\mathcal{M}_{loc}(\beta_0) = \{ B_{O, \beta_0}^*, O \subset S, O \text{ open in } S \}. \quad (3.2)$$

It is important to note here that the sets O are open in S , whereas sets $B \in \mathcal{B}$ by definition are closed. In other words, the sets B_{O, β_0}^* cannot touch boundaries of O , unless this boundary of O also is a (subset of a) boundary of S . Therefore, in

regular situations, for many sets O there will exist no set B_{O,β_0}^* , and there might exist many sets O with the same set B_{O,β_0}^* . As a result the class $\mathcal{M}_{loc}(\beta_0)$ might be small, and often will be finite. Also notice that for the global maximizer $B_{\beta_0}^*$ as defined in (2.1) we have $B_{\beta_0}^* = B_{S,\beta_0}^*$.

We introduce some notation. For a box $B = \bigotimes_{i=1}^d [a_{i1}, a_{i2}]$, $a_{i1} \leq a_{i2}$, $j = 1, \dots, d$, $d \geq 2$, and $k = 1, 2$ let

$$F_j(\cdot, B^{\hat{j}}) = \int_{B^{\hat{j}} \cap S^{\hat{j}}} f(\dots, x_{j-1}, \cdot, x_{j+1}, \dots) d\underline{x}_{\hat{j}} / F(S) \quad (3.3)$$

$$I_j(\cdot, B^{\hat{j}}) = \int_{B^{\hat{j}} \cap S^{\hat{j}}} m(\dots, x_{j-1}, \cdot, x_{j+1}, \dots) f(\dots, x_{j-1}, \cdot, x_{j+1}, \dots) d\underline{x}_{\hat{j}} / F(S) \quad (3.4)$$

where $\underline{x}_{\hat{j}} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)'$, and $B^{\hat{j}} = \bigotimes_{i=1, i \neq j}^d [a_{i1}, a_{i2}] \subset [0, 1]^{d-1}$. Notice the dependence on S which is not reflected in the notation just introduced. Further we denote

$$F_{jk}(\partial B) = F_j(a_{jk}, B^{\hat{j}}) \quad \text{and} \quad I_{jk}(\partial B) = I_j(a_j, B^{\hat{j}}).$$

For $d = 1$ we define $F_{1k}(\partial B) = f(a_{1k}) \mathbf{1}_S(a_{1k})$ and $I_{1k}(\partial B) = m(a_{1k}) f(a_{1k}) \mathbf{1}_S(a_{1k})$.

We also use the notation

$$\partial B_{jk} = \bigotimes_{i=1}^{j-1} [a_{i1}, a_{i2}] \times a_{jk} \times \bigotimes_{i=j+1}^d [a_{i1}, a_{i2}], \quad j = 1, \dots, d, \quad k = 1, 2, \quad (3.5)$$

for the (jk) -th boundary facet of B . We sometimes use the notation ‘ (jk) ’ rather than ∂B_{jk} . ‘Boundary averages’ turn out to be of importance. Therefore we introduce

$$\text{ave}_j(x, B^{\hat{j}}) := \frac{I_j(x, B^{\hat{j}})}{F_j(x, B^{\hat{j}})}, \quad \text{for } F_j(x, B^{\hat{j}}) > 0, \quad (3.6)$$

and use the shorthand notation

$$A_{jk}^{\pm}(\partial B) := \text{ave}_j(a_{jk}^{\pm}, B^{\hat{j}}), \quad (3.7)$$

where $\text{ave}_j(a_{jk}^+, B^{\hat{j}})$ and $\text{ave}_j(a_{jk}^-, B^{\hat{j}})$ denote the limits of $\text{ave}_j(x, B^{\hat{j}})$ as x approaches a_{jk} from the *outside* of the box B and from the *inside* of the box B ,

respectively. Of course, if B is such that its boundary ∂B_{jk} lies at a boundary of S , then $A_{jk}^+(\partial B)$ is not defined. Also, $A_{jk}^\pm(\partial B)$ is only defined for such boxes B with $F_j(a_{jk}, B^{\widehat{j}}) > 0$ for all $j = 1, 2, k = 1, \dots, d$.

Observe that the peeling procedure consists in peeling off that boundary with the smallest average, and hence it has the tendency to keep boundary averages as close as possible. This motivates the following function:

$$\Psi(B) := \max_{(jk)} \left[\max_{(\ell m) \neq (jk)} (A_{\ell m}^+(\partial B)) - A_{jk}^-(\partial B) \right] \quad (3.8)$$

where for any (jk) the max inside $[\dots]$ is taken over all those $(\ell m) \neq (jk)$ for which $A_{\ell m}^+(\partial B)$ exists. In case $A_{\ell m}^+(\partial B)$ does not exist for all (ℓm) we define $\Psi(B) = -\infty$. Lemma 3.1 indicates that boxes B with $\Psi(B) \leq 0$ are potential limits of peeling+jittering. Therefore, for $0 \leq \beta_0 \leq 1$ let

$$\mathcal{N}_{loc}(\beta_0) = \{ B \in \mathcal{B} : F(B|S) = \beta_0, \text{ and } \Psi(B) \leq 0 \}. \quad (3.9)$$

Observe that if $\text{ave}_{jk}(x, B^{\widehat{j}})$ is continuous at all $x = a_{jk}$ (so that $A_{jk}^+(\partial B) = A_{jk}^-(\partial B)$ for all jk), then $\Psi(B) \geq 0$ with $\Psi(B) = 0$ iff all the boundary averages of B are equal. Such boxes are typical candidates for PRIM outcomes. However, $\mathcal{N}_{loc}(\beta_0)$ does not only contain local maximizers, but also minimizers and 'saddlepoints'. A class of typical local minimizers, i.e. boxes minimizing $I(B \cap S)/F(B \cap S)$ is the following. Let

$$m_{loc,\epsilon}(\beta_0) = \{ B^* \in \mathcal{N}_{loc}(\beta_0) \text{ such that (m.i) and (m.ii) hold } \} \quad (3.10)$$

where (m.i) and (m.ii) are as follows. The constant $\epsilon > 0$ such that for each $B^* = \bigotimes_{j=1}^d [a_{j1}^*, a_{j2}^*], \in m_{loc,\epsilon}(\beta_0)$ there exists $\overline{B} = \bigotimes_{j=1}^d [\overline{a}_{j1}, \overline{a}_{j2}], \underline{B} = \bigotimes_{j=1}^d [\underline{a}_{j1}, \underline{a}_{j2}]$, such that $\underline{B} \subseteq B^* \subseteq \overline{B}$ with $\max\{|\overline{a}_{jk} - a_{jk}^*|, |\underline{a}_{jk} - a_{jk}^*|\} \geq \epsilon > 0$ for all $j = 1, \dots, d, k = 1, 2$. Let $\mathcal{U}(\epsilon, B^*) := \{ B \in \mathcal{B}, \underline{B} \subset B \subset \overline{B} \}$, then for all $B \in \mathcal{U}(\epsilon, B^*)$ and all $j \in \{1, \dots, d\}$ we have that

(m.i) $\text{ave}_j(\cdot, B^{\widehat{j}})$ is strictly decreasing in $[\overline{a}_{j1}, \underline{a}_{j1}]$ and strictly increasing in $[\underline{a}_{j2}, \overline{a}_{j2}]$.

(m.ii) For some constant $k_1 > 0$ not depending on B and j ,

$$k_1 |x_1 - x_2| \leq \left| \text{ave}_j(x_1, B^{\hat{j}}) - \text{ave}_j(x_2, B^{\hat{j}}) \right|,$$

for either $x_1, x_2 \in [\bar{a}_{j1}, \underline{a}_{j2}]$ or $x_1, x_2 \in [\underline{a}_{j1}, \bar{a}_{j2}]$.

The crucial assumption is (m.i). It is obvious that (m.ii) is not necessary for a set being a local minimum. Also, as can be seen from the proof of the following result, the ‘global’ constant $\epsilon > 0$ is not really needed here. However, all this is needed for proving our large sample results for PRIM, and it is introduced here for later convenience.

Lemma 3.1 *Suppose that assumptions $(\mathbf{A2})_{\beta_0}$ and $(\mathbf{A3})_{\beta_0}$ hold (cf. Section 4). Then for every $0 < \beta_0 < 1$ and $\epsilon > 0$ we have*

$$\mathcal{M}_{loc}(\beta_0) \subset \mathcal{N}_{loc}(\beta_0) \setminus m_{loc,\epsilon}(\beta_0).$$

In the following we consider some specific examples in order to provide a better feeling for what the sets in $\mathcal{M}_{loc}(\beta_0)$ are.

The one-dimensional case. Although one would likely not use PRIM in the one-dimensional case, a consideration of this simple case provides some insight. Suppose m is a symmetric bimodal regression curve, and let X be uniformly distributed in $[0, 1]$. Figure 2) shows some outcomes of peeling+pasting and peeling+ jittering, respectively for

$$m(x) = \begin{cases} \exp(-30(x - 0.3)^2), & 0 \leq x \leq 0.5; \\ \exp(-30(x - 0.7)^2), & 0.5 < x \leq 1. \end{cases} \quad (3.11)$$

If β_0 is small enough, the solution of (2.1) is one of two intervals with support β_0 each corresponding to each mode. The nature of having two disconnected sets indicates that there are two distinct modes. The population version of the peeling procedure results (when $\alpha \rightarrow 0$) in an interval with support β_0 with one endpoint being a mode (see figure 2, plot 2).

PUT FIGURE 2 HERE

The proposed bottom-up pasting procedure will increase the average value of the box, but it also increases its support (see figure 2, plot 3). If we apply peeling+jittering, then the result approaches the optimal set as $\alpha \rightarrow 0$ (see figure 2, plot 4). An application of the covering strategy (i.e. removal of the just found optimal interval, and a second application of the peeling to the remainder) will result in the analogous interval around the second mode. Thus, the two modes will be resolved.

PUT FIGURE 3 HERE

The multidimensional case. There is a somewhat surprising shortcoming of PRIM in the multidimensional case. In contrast to the one-dimensional case, in two or higher dimensions PRIM might not be able to resolve two distinct modes, even if β_0 is chosen small enough, and also the covering strategy might not help. This is actually what is shown in Figure 3. The long, thin box in plot 2 of Figure 3 is a local maximum, whereas the other two boxes both are global maxima. The covering leads to nested boxes of similar shape, and the two modes are not resolved. A possible remedy is as following.

For a box $B = \bigotimes_{j=1}^d [a_{j1}, a_{j2}]$, let $\ell_i = |a_{j2} - a_{j1}|$, $i = 1, \dots, d$ denote the length of the box in the i -th coordinate and $R_B = \frac{\max_i \ell_i}{\min_i \ell_i}$ the ratio of the box. The idea is to first standardize all the marginals to have same mean and variance. Then, in each peeling step k , to allow only such boxes B_k with $R_{B_k} \leq R_{B_{k-1}}$ or $R_{B_k} \leq r$ as candidates for next peeling. Here r is a pre-chosen parameter to control the final peeled box ratio should avoid boxes which are too thin. A simulation study conducted by the authors indicated that for $d = 2$ a reasonable choice appears to be $1.5 \leq r \leq 2.0$. However, more simulations are necessary to make a more general statement.

4 Convergence results

Next we derive a result about how far the solution of the peeling result differs from its theoretically optimal counterpart. To this end we need some more notation and assumptions.

First we introduce two distance measures between boxes. For two boxes $B = \bigotimes_{j=1}^d [a_{j1}, a_{j2}]$ and $\tilde{B} = \bigotimes_{j=1}^d [\tilde{a}_{j1}, \tilde{a}_{j2}]$ let

$$F(B \Delta \tilde{B}) := F(B \setminus \tilde{B}) + F(\tilde{B} \setminus B),$$

the set-theoretic symmetric difference between B and \tilde{B} , and let

$$\rho_\infty(B, \tilde{B}) := \max_{j=1, \dots, d, k=1, 2} |a_{jk} - \tilde{a}_{jk}|.$$

We also need the following quantities in order to deal with the overlap between two candidate boxes for peeling+jittering. For a box $B = \bigotimes_{i=1}^d [a_{i1}, a_{i2}]$, $0 \leq a_{i1} \leq a_{i2} \leq 1$, $j = 1, \dots, d$, $d \geq 3$, and $k = 1, 2$ let

$$F_{j,\ell}(x_j, x_\ell, B^{\hat{j}, \hat{\ell}}) = \int_{B^{\hat{j}, \hat{\ell}} \cap S^{\hat{j}, \hat{\ell}}} f(\dots, x_j, \dots, x_\ell, \dots) d\underline{x}^{\hat{j}, \hat{\ell}} \quad (4.1)$$

$$I_{j,\ell}(x_j, x_\ell, B^{\hat{j}, \hat{\ell}}) = \int_{B^{\hat{j}, \hat{\ell}} \cap S^{\hat{j}, \hat{\ell}}} m(\dots, x_j, \dots, x_\ell, \dots) f(\dots, x_j, \dots, x_\ell, \dots) d\underline{x}^{\hat{j}, \hat{\ell}} \quad (4.2)$$

where $\underline{x}^{\hat{j}, \hat{\ell}} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_{\ell-1}, x_{\ell+1}, \dots, x_d)'$, and $B^{\hat{j}, \hat{\ell}}$ denotes the $(d-2)$ -dimensional box $\bigotimes_{i=1, i \neq j, \ell}^d [a_{i1}, a_{i2}]$. Further we denote

$$F_{jk,\ell m}(\partial B) = F_{j,\ell}(a_{jk}, a_{\ell m}, B^{\hat{j}, \hat{\ell}}) \quad \text{and} \quad I_{jk,\ell m}(\partial B) = I_{j,\ell}(a_{jk}, a_{\ell m}, B^{\hat{j}, \hat{\ell}}).$$

For $d = 2$ and $j, k, \ell, m = 1, 2$, let $F_{jk,\ell m}(\partial B) = f(a_{jk}, a_{\ell m}) \mathbf{1}_S(a_{jk}, a_{\ell m})$ and $I_{jk,\ell m}(\partial B) = m(a_{jk}, a_{\ell m}) f(a_{jk}, a_{\ell m}) \mathbf{1}_S(a_{jk}, a_{\ell m})$.

Assumptions.

(A1) $_{\beta_0}$ The parameter $0 < \beta_0 < 1$ is such that there exists a $\delta > 0$ and a constant $c_1 > 0$ (only depending on δ) such that

$$\sup_{\beta \in (0,1): |\beta - \beta_0| < \delta} \sup_{B_{\beta_0}^* \in \mathcal{N}_{loc}(\beta_0)} \inf_{B_\beta^* \in \mathcal{N}_{loc}(\beta)} \rho_\infty(B_\beta^*, B_{\beta_0}^*) \leq c_1 \frac{|\beta - \beta_0|}{d}.$$

(A2) $_{\beta_0}$ The parameter $0 < \beta_0 < 1$ is such that

$$0 < \epsilon_0 \leq \inf_{B: F(B|S) \geq \beta_0/2} \|F_j(\cdot, B^{\hat{j}})\|_{\infty} \leq \sup_{B: F(B|S) \geq \beta_0/2} \|F_j(\cdot, B^{\hat{j}})\|_{\infty} < K_0$$

with constants ϵ_0, K_0 neither depending on the index $j \in \{1, \dots, d\}$ nor on d . Here $\|\cdot\|_{\infty}$ denotes sup-norm. Further, there exists an $L \in \mathbb{N}$ such that the following holds. For each j there exists c_{j1}, \dots, c_{jL} such that for all B with $F(B|S) \geq \beta_0/2$ the functions $F_j(\cdot, B^{\hat{j}})$ are uniformly Lipschitz continuous in $(c_{jk}, c_{j,k+1})$, $k = 1, \dots, L-1$, with Lipschitz constant not depending on d .

(A3) $_{\beta_0}$ The parameter $0 < \beta_0 < 1$ is such that

$$\sup_{B: F(B|S) \geq \beta_0/2} \|I_j(\cdot, B^{\hat{j}})\|_{\infty} < K_1$$

for a constant K_1 neither depending on the index j nor on d . Further, for each B with $F(B|S) \geq \beta_0/2$ the functions $F_j(\cdot, B^{\hat{j}})$ are uniformly piecewise Lipschitz continuous (in the sense described in **(A2) $_{\beta_0}$**) with Lipschitz constant not depending on d .

(A4) $_{\beta_0}$ The parameter $0 < \beta_0 < 1$ is such that both

$$\sup_{B: F(B|S) \geq \beta_0/2} \|F_{j,\ell}(\cdot, \cdot, B^{\hat{j}, \hat{\ell}})\|_{\infty} < K_2,$$

$$\sup_{B: F(B|S) \geq \beta_0/2} \|I_{j,\ell}(\cdot, \cdot, B^{\hat{j}, \hat{\ell}})\|_{\infty} < K_2,$$

for a constant K_2 neither depending on the indices j, ℓ nor on d .

(A5) $_{\beta_0}$ The parameter $0 < \beta_0 < 1$ is such that there exists a $\delta > 0$ and a constant $c > 0$ such that for all $B \notin \mathcal{N}_{loc}(F(B|S))$ with $|F(B|S) - \beta_0| < \delta$ we have

$$\Psi(B) \geq c \inf_{B^* \in \mathcal{N}_{loc}(F(B|S))} \rho_{\infty}(B, B^*). \quad (4.3)$$

Discussion of the assumptions. Assumption **(A1) $_{\beta_0}$** says that in a small (enough) neighborhood of an optimal set $B_{\beta_0}^*$ we also find an optimal set B_{β}^* for β close to

β_0 . A scenario where $(\mathbf{A1})_{\beta_0}$ holds is the following. Suppose that F is the uniform distribution and m is rotationally symmetric locally around $B_{\beta_0}^*$ (or possesses other appropriate symmetry properties), and also is monotonically de(in)creasing, again locally around $B_{\beta_0}^*$, when moving outward of an optimal set $B_{\beta_0}^*$ (depending on whether $B_{\beta_0}^*$ is a local minimum of a local maximum). Then there exist optimal sets $\{B_{\beta}^*, |\beta - \beta_0|/d < \epsilon\}$ which form a class of totally ordered sets (with respect to inclusion). It follows that $(\mathbf{A1})_{\beta_0}$ holds.

Assumption $(\mathbf{A5})_{\beta_0}$ is crucial. It implies that when ‘moving away’ from an optimal set B_{β}^* , while keeping the ‘size’ $F(B)$ fixed (equal to β) then the maximal difference of boundary averages increases (recall that for an optimal set B_{β}^* we have $\Psi(B_{\beta}^*) \leq 0$). Condition $(\mathbf{A5})_{\beta_0}$ holds under the scenario given in the discussion of $(\mathbf{A1})_{\beta_0}$ above, provided, for instance, m is differentiable with partial derivatives bounded away from zero.

4.1 Performance of population version of peeling + jittering

The following results presents conditions such that for given $\beta_0 > 0$ the outcome of peeling+jittering is ‘close’ to one of the set in $\mathcal{M}_{loc}(\beta_0) \setminus m_{loc, \frac{\epsilon_1}{d}}(\beta_0)$ (with $S = [0, 1]^d$) as long as $d\alpha$ is small. We will see below that under appropriate assumptions the empirical version of peeling+jittering behaves similarly.

Theorem 4.1 *Let $0 < \beta_0 < 1$, $\epsilon_1 > 0$, and let $\alpha = \alpha_n$ be such that $d\alpha_n = o(1)$ as $n \rightarrow \infty$. Suppose that $(\mathbf{A1})_{\beta_0}$ - $(\mathbf{A5})_{\beta_0}$ hold for $S = [0, 1]^d$. Let \tilde{B} denote the result of the population version of peeling + jittering. Then as $n \rightarrow \infty$ we have*

$$\inf_{B^* \in \mathcal{N}_{loc}(\beta_0) \setminus m_{loc, \frac{\epsilon_1}{d}}(\beta_0)} \rho_{\infty}(B^*, \tilde{B}) = O(\alpha_n), \quad (4.4)$$

and hence

$$\inf_{B^* \in \mathcal{N}_{loc}(\beta_0) \setminus m_{loc, \frac{\epsilon_1}{d}}(\beta_0)} F(B^* \Delta \tilde{B}) = O(d\alpha_n). \quad (4.5)$$

It is straightforward to derive the proof of this theorem from the proof of Theorem 4.2 presented below. Both proofs have the same basic structure, except that the proof

of Theorem 4.1 does not involve stochastic arguments, and this makes it simpler. The proof is hence omitted.

4.2 Empirical performance of peeling + jittering

Let \widehat{B} denote the outcome of empirical peeling+jittering as applied to $S = [0, 1]^d$. The following result shows that \widehat{B} behaves similar to its population version \widetilde{B} .

Theorem 4.2 *Suppose that $E|Y|^\gamma < \infty$ for some $\gamma \geq 2$. Let $\beta_0 > 0$ and $\epsilon_1 > 0$ be fixed, and assume that $(\mathbf{A1})_{\beta_0}$ - $(\mathbf{A5})_{\beta_0}$ hold for $S = [0, 1]^d$. Further suppose that $\alpha_n = o(1)$ satisfies*

$$\max\left(\sqrt{\frac{d}{n} \alpha_n \log \frac{d}{\alpha_n}}, \left(\frac{d}{\alpha_n n} \log \frac{d}{\alpha_n}\right)^{\frac{\gamma-1}{2}}\right) = o(\alpha_n^2). \quad (4.6)$$

Then we have for $\eta_n \geq \alpha_n$ that

$$\inf_{B^* \in \mathcal{N}_{loc}(\beta_0) \setminus m_{loc, \frac{\epsilon_1}{d}}(\beta_0)} \rho_\infty(B^*, \widehat{B}) = O_P(\eta_n), \quad (4.7)$$

and hence

$$\inf_{B^* \in \mathcal{N}_{loc}(\beta_0) \setminus m_{loc, \frac{\epsilon_1}{d}}(\beta_0)} F(B^* \Delta \widehat{B}) = O_P(d\eta_n). \quad (4.8)$$

The above result indicates that one has to balance the choice of α_n with the dimension d and moment conditions on Y in order to obtain good statistical properties. Notice that the extra factor of d in the rates of $F(B^* \Delta \widehat{B})$ reflects the curse of dimensionality. Although all the ‘one-dimensional’ distances might be small, i.e. $\rho_\infty(B^*, \widehat{B})$ is small, the truly d -dimensional distance $F(B^* \Delta \widehat{B})$ might be large due to the fact that d is large.

From Theorem 4.2 we for instance obtain the following two corollaries on consistency and rates of convergence of peeling in d_F -distance.

Corollary 4.1 (Consistency) *Suppose that $E|Y|^\gamma < \infty$ for some $\gamma \geq 2$. Let $0 < \beta_0 < 1$ be fixed and assume that $(\mathbf{A1})_{\beta_0}$ - $(\mathbf{A5})_{\beta_0}$ hold for $S = [0, 1]^d$. If α_n is such*

that

$$\max\left\{\frac{d^3 \log n}{n}, d\left(\frac{\log n}{n}\right)^{\frac{\gamma-1}{\gamma+1}}\right\} = o(\alpha_n) \quad \text{and} \quad d\alpha_n = o(1),$$

then

$$\inf_{B^* \in \mathcal{N}_{\text{loc}}(\beta_0) \setminus m_{\text{loc}, \frac{\epsilon_1}{d}}(\beta_0)} d_F(B^*, \widehat{B}) = o_P(1).$$

For fixed d , for instance, the above result requires α_n to be such that $\alpha_n = o(1)$ with $\alpha_n \left(\frac{n}{\log n}\right)^{\frac{\gamma+1}{\gamma-1}} \rightarrow \infty$. When d is allowed to vary with n , then the strongest condition on d is for $\gamma = 2$ where we need $\frac{d^6}{n} \log n = o(1)$. For $\gamma \geq 3$ we need $\frac{d^4}{n} \log n = o(1)$.

We can go a little further by choosing the ‘optimal’ sequence α_n which balances the two terms in the ‘max’ in (4.6). This leads to the following rates of convergence. We use the notation $a_n \sim b_n$ for sequences of real numbers $\{a_n\}, \{b_n\}$ with $a_n/b_n \rightarrow c > 0$ as $n \rightarrow \infty$.

Corollary 4.2 (Rates of convergence) *Suppose that $E|Y|^\gamma < \infty$ for some $\gamma \geq 2$. Let $0 < \beta_0 < 1$ and $\epsilon_1 > 0$ be fixed, and assume that $(\mathbf{A1})_{\beta_0} - (\mathbf{A5})_{\beta_0}$ hold for $S = [0, 1]^d$, and that $\frac{d}{n} \log n = o(1)$. Then we have the following.*

(i) *If $\gamma \geq 3$ and we choose $\alpha_n \sim \left(\frac{d}{n} \log n\right)^{\frac{1}{3}}$ then*

$$\inf_{B^* \in \mathcal{N}_{\text{loc}}(\beta_0) \setminus m_{\text{loc}, \frac{\epsilon_1}{d}}(\beta_0)} d_F(B^*, \widehat{B}) = O_P\left(\left(\frac{d^4}{n} \log n\right)^{1/3}\right). \quad (4.9)$$

(ii) *If $2 \leq \gamma \leq 3$ and we choose $\alpha_n \sim \left(\frac{d}{n} \log n\right)^{\frac{\gamma-1}{\gamma+3}}$ then*

$$\inf_{B^* \in \mathcal{N}_{\text{loc}}(\beta_0) \setminus m_{\text{loc}, \frac{\epsilon_1}{d}}(\beta_0)} d_F(B^*, \widehat{B}) = O_P\left(\left(\frac{d^{\frac{2(\gamma+1)}{\gamma-1}}}{n} \log n\right)^{\frac{\gamma-1}{\gamma+3}}\right). \quad (4.10)$$

Remarks. (i) The rates in Corollary 4.2 are of the form $O_P(d\alpha_n)$, with α_n as provided in the formulation of the corollary. Hence for those choices of α_n the derived rates are in alignment with the rate of the population version (cf. Theorem 4.1).

(ii) It is known that minimum volume sets in VC-classes converge at rates $n^{-1/3}$ (up to log-terms). This for instance holds for the minimum volume ellipsoid, or the

shorth). Hence, the rate in part (i) does not come as a surprise, since by definition of \widehat{B} it is a kind of minimum volume set in a (sub)class of axis-parallel rectangles which forms a VC-class (e.g. see Polonik 1997)

Proof of Theorem 4.2: The rate for $d_F(B^* \Delta \widehat{B})$ asserted in (4.8) follow from (4.7) because $F(B^* \Delta \widehat{B}) = O(d \rho_\infty(B^*, \widehat{B}))$.

This proof for the rate of $\rho_\infty(B^*, \widehat{B})$ is a ‘stochastic version’ of the proof of Theorem 4.1. For $C_1 > 0$ let

$$\begin{aligned} A_n &:= \left\{ \sup_{B \in \mathcal{B}} |(F_n - F)(B)| \leq C_1 \sqrt{\frac{d}{n}} \right\} \\ &\cup \left\{ \sup_{F_n(B) \leq \alpha_n} |(F_n - F)(B)| \leq C_1 \sqrt{\frac{d}{n} \alpha_n \log \frac{1}{\alpha_n}} \right\} \\ &\cup \left\{ \sup_{F(B) \leq 2\alpha_n} |(I_n - I)(B)| \leq C_1 \max\left(\sqrt{\frac{d}{n} \alpha_n \log \frac{d}{\alpha_n}}, \left(\frac{d}{n \alpha_n} \log \lceil \frac{d}{\alpha_n} \rceil \right)^{\frac{\gamma-1}{2}} \right) \right\}. \end{aligned}$$

It follows from Proposition 5.1 and Proposition 5.2 that for each $\epsilon > 0$ we can choose $C_1 > 1$ such that for n large enough we have $P(A_n) \geq 1 - \epsilon$. We prove the theorem in two steps. First we show that for C large enough we have

$$P\left(\inf_{B^* \in \mathcal{N}_{loc}(\beta_0)} \rho_\infty(B^*, \widehat{B}) > C 2 \eta_n; A_n \right) = 0. \quad (4.11)$$

To complete the proof we then show that

$$P\left(\inf_{B^* \in m_{loc, \epsilon/d}(\beta_0)} \rho_\infty(B^*, \widehat{B}) < c \epsilon_1; A_n \right) = 0, \quad (4.12)$$

for some constant $c > 0$. Let $\widehat{\beta}_n := F(\widehat{B})$. Notice that on A_n we have

$$\begin{aligned} \beta_0 \leq \widehat{\beta}_n &= F_n(\widehat{B}) - (F_n - F)(\widehat{B}) \\ &< (\beta_0 + \alpha_n) + C_1 \sqrt{\frac{d}{n} \log d} \\ &\leq \beta_0 + (C_1 + 1) \delta_n, \end{aligned} \quad (4.13)$$

with $\delta_n := \max(\alpha_n, \sqrt{\frac{d}{n} \log d})$. Using $(\mathbf{A1})_{\beta_0}$ we obtain that for all $B_{\widehat{\beta}_n}^* \in \mathcal{N}_{loc}(\widehat{\beta}_n)$ there exists a $B^* \in \mathcal{N}_{loc}(\beta_0)$ such that on A_n we have

$$\rho_\infty(B_{\widehat{\beta}_n}^*, B^*) \leq \frac{|\widehat{\beta}_n - \beta_0|}{d} \leq \frac{c_1 (C_1 + 1)}{d} \delta_n.$$

Now suppose that for all $B_{\beta_0}^* \in \mathcal{N}_{loc}(\beta_0)$ we have $\rho_\infty(B_{\beta_0}^*, \widehat{B}) > 2C d \eta_n$ for some $C > 0$. Since $\eta_n \geq \delta_n$ we obtain by using triangular inequality that for all $B_{\widehat{\beta}_n}^* \in \mathcal{N}_{loc}(\widehat{\beta}_n)$ we have on A_n for C large enough that

$$\rho_\infty(B_{\widehat{\beta}_n}^*, \widehat{B}) \geq 2C \eta_n - \frac{c_1(C_1 + 1)}{d} \delta_n \geq C \eta_n.$$

In other words, for C large enough

$$\begin{aligned} & \{ \inf_{B^* \in \mathcal{N}_{loc}(\beta_0)} \rho_\infty(B^*, \widehat{B}) \geq 2C \eta_n, A_n \} \\ & \subset \{ \inf_{B_{\widehat{\beta}_n}^* \in \mathcal{N}_{loc}(\widehat{\beta}_n)} \rho_\infty(B_{\widehat{\beta}_n}^*, \widehat{B}) \geq C \eta_n, A_n \} \end{aligned} \quad (4.14)$$

We will show that the probability for the event on the r.h.s. equals zero for both C and n large enough. First notice that because of **(A5)** $_{\beta_0}$ we have $\Psi(\widehat{B}) > \frac{cC}{2} \eta_n$. It follows that there exist two boundary facets of \widehat{B} indexed by, let's say, (jk) and (ℓm) , respectively, with $A_{jk}^+(\partial \widehat{B}) \geq A_{\ell m}^-(\partial \widehat{B}) + \frac{cC}{2} \eta_n$. Let b_{jk} and $b_{\ell m}$ be the two candidate sets for jittering. We have $F_n(b_{jk} \setminus b_{\ell m}) = F_n(b_{\ell m}) = \lceil n \alpha_n (1 - \alpha_n)^{K_n - 1} \rceil / n$, with b_{jk} being outside the box \widehat{B} and $b_{\ell m}$ inside. We will show that for C large enough, $I_n(b_{jk} \setminus b_{\ell m}) > I_n(b_{\ell m})$ which means that adding b_{jk} and removing $b_{\ell m}$ leads to an increase in I_n -measure, while leaving the support of the resulting box constant. In other words, \widehat{B} is not an 'optimal' set, i.e. not a solution to (2.1). A ion.

For a box B and a box b_{jk} adjacent to ∂B_{jk} let

$$r_{jk}^\pm(B) := \frac{I(b_{jk})}{F(b_{jk})} - A_{jk}^\pm(\partial B) \quad (4.15)$$

We have the following result the proof of which can be found in the Appendix.

Lemma 4.1 *Suppose that **(A2)** $_{\beta_0}$ and **(A3)** $_{\beta_0}$ hold and that $F(b_{jk}) > 0$. Then for $\epsilon > 0$ we have uniformly in $j \in \{1, \dots, d\}$, $k \in \{1, 2\}$ we have*

$$\sup_{\{B: F_{jk}(\partial B) \geq \epsilon\}} |r_{jk}^\pm(B)| = O(F(b_{jk})) \quad \text{as } F(b_{jk}) \rightarrow 0. \quad (4.16)$$

It now follows by using (4.16) that

$$\begin{aligned}
I(b_{jk}) &= F(b_{jk}) [A_{jk}^+(\partial\widehat{B}) + r_{jk}^+(\widehat{B})] \\
&= (F(b_{\ell m}) - (F(b_{\ell m}) - F(b_{jk}))) [A_{jk}^+(\partial\widehat{B}) + r_{jk}^+(\widehat{B})] \\
&\geq (F(b_{\ell m}) - (F(b_{\ell m}) - F(b_{jk}))) [A_{\ell m}^-(\partial\widehat{B}) + r_{\ell m}^-(\widehat{B})] \\
&\quad + \frac{cC}{2} \eta_n + r_{jk}^+(\widehat{B}) - r_{\ell m}^-(\widehat{B})
\end{aligned}$$

Hence we obtain

$$I(b_{jk}) - I(b_{\ell m}) \geq (I) - (II),$$

where

$$\begin{aligned}
(I) &:= F(b_{\ell m}) \left[\frac{cC}{2} \eta_n + r_{jk}^+(B) - r_{\ell m}^-(B) \right], \quad \text{and} \\
(II) &:= (F(b_{\ell m}) - F(b_{jk})) \left[A_{\ell m}(\partial\widehat{B}) + r_{\ell m}^-(\widehat{B}) \right. \\
&\quad \left. - \frac{cC}{2} \eta_n + r_{jk}^+(\widehat{B}) - r_{\ell m}^-(\widehat{B}) \right].
\end{aligned}$$

We will now show that for large enough n we have on A_n that

$$I(b_{jk}) - I(b_{\ell m}) \geq (I) - (II) \geq \frac{c\beta_0}{32} C \eta_n \alpha_n \quad (4.17)$$

for C large enough. We prove (4.17) by showing that on A_n we have both

$$(I) \geq \frac{c\beta_0}{16} C \alpha_n \eta_n, \quad \text{and} \quad (4.18)$$

$$(II) \leq \frac{6C_1 K_1}{\epsilon_0} \alpha_n \eta_n, \quad (4.19)$$

with constants $\epsilon_0, K_1 > 0$ from $(\mathbf{A2})_{\beta_0}$ and $(\mathbf{A3})_{\beta_0}$, respectively. Since the constants in (4.19) are fixed these two inequalities imply (4.17) for C large enough.

First we show that on A_n for large enough n we have

$$\beta_0 \alpha_n / 2 < F(b_{\ell m}) < 2 \alpha_n. \quad (4.20)$$

To see the first inequality observe that by construction of the PRIM algorithm we have $F_n(b_{\ell m}) \geq \alpha_n (1 - \alpha_n)^{K_n - 1}$. If at each step of the peeling procedure we were to

cut off *exactly* a fraction of α_n from the current data set then $(1 - \alpha_n)^{K_n - 1} = \beta_0$. However, that due to the discrete nature of the empirical peeling the actual fraction cut off at each peeling step lies in $[\alpha_n, \alpha_n + 1/n)$. Hence, the peeling procedure in general stops after fewer steps (than in the case of cutting off exactly a fraction of α_n), and hence we have $(1 - \alpha_n)^{K_n - 1} \geq \beta_0$. This implies

$$\alpha_n \geq F_n(b_{\ell m}) \geq \alpha_n (1 - \alpha_n)^{K_n - 1} \geq \alpha_n \beta_0.$$

We also have on A_n that $\sup_{(\ell m)} |(F_n - F)(b_{\ell m})| \leq C_1 \sqrt{\frac{d}{n} \alpha_n \log \frac{d}{\alpha_n}} = o(\alpha_n^2) = o(\alpha_n)$. This implies (4.20).

To see (4.18), observe that the upper bound from (4.20) together with Lemma 4.1 imply that $r_{jk}^+(\widehat{B}) - r_{\ell m}^-(\widehat{B}) = O(\alpha_n)$. Adding in the lower bound from (4.20) the assertion now follows for large enough C from the definition of (I) (note that by assumption $\eta_n \geq \alpha_n$).

In order to see (4.19) observe that on A_n

$$\begin{aligned} & |F(b_{jk}) - F(b_{\ell m})| \\ &= |F_n(b_{jk}) - F_n(b_{\ell m})| + |(F_n - F)(b_{jk}) - (F_n - F)(b_{\ell m})| \\ &\leq O\left(\frac{1}{n}\right) + 2C_1 \sqrt{\frac{d}{n} \alpha_n \log \frac{d}{\alpha_n}} \leq 3C_1 \alpha_n \eta_n. \end{aligned} \quad (4.21)$$

Also observe that from $(\mathbf{A2})_{\beta_0}$ and $(\mathbf{A3})_{\beta_0}$ we have $A_{\ell m}(\partial \widehat{B}) \leq K_1/\epsilon_0$. Again using Lemma 4.1 we obtain that the second term in (II) (i.e. the term in $[\dots]$) can be bounded by $2K_1/\epsilon_0 > 0$. Hence, (4.19) follows.

The last step of the proof is to show that the analogue to (4.17) also holds for the difference of the I_n -measures (rather than the I -measures), i.e. we show that on A_n for n large enough

$$I_n(b_{jk}) - I_n(b_{\ell m}) \geq \frac{c\beta_0}{64} C \alpha_n \eta_n. \quad (4.22)$$

Writing

$$I_n(b_{jk}) - I_n(b_{\ell m}) = I(b_{jk}) - I(b_{\ell m}) + ((I_n - I)(b_{jk}) - (I_n - I)(b_{\ell m}))$$

we see that it remains to show that $(I_n - I)(b_{jk}) - (I_n - I)(b_{\ell m}) = O_P(\alpha_n \eta_m)$. We have seen above that on A_n we have $\sup_{jk} F(b_{jk}) \leq 2\alpha_n$, for n large enough. Consequently, on A_n

$$(I_n - I)(b_{jk}) - (I_n - I)(b_{\ell m}) \leq 2C_1 \max \left(\sqrt{\frac{d}{n} \alpha_n \log \lceil \frac{d}{\alpha_n} \rceil}, \left(\frac{d}{n \alpha_n} \log \lceil \frac{d}{\alpha_n} \rceil \right)^{\frac{\gamma-1}{2}} \right).$$

It follows from assumption (4.6) that the r.h.s is $o(\alpha_n^2)$. This completes the proof for the non-overlapping case.

Now consider the case of an overlap of b_{jk} and $b_{\ell m}$. For each (jk) let $b_{jk} = \bigotimes_{i=1}^d [c_{i1}^{jk}, c_{i2}^{jk}]$ and denote the width of b_{jk} by $h_{jk} = c_{j1}^{jk} - c_{j2}^{jk}$. We have uniform upper and lower bounds for these widths, namely,

$$\alpha_n \beta_0 / K_0 \leq h_{jk} \leq \alpha_n / \epsilon_0. \quad (4.23)$$

The first inequality follows from

$$\begin{aligned} \alpha_n \beta_0 \leq \alpha_n (1 - \alpha_n)^{K_n} &= F(b_{jk}) = \int_{c_{j1}^{jk}}^{c_{j1}^{jk} + h_{jk}} F_j(x, \widehat{B}^j) dx \\ &\leq h_{jk} \sup_x F_j(x, \widehat{B}^j) \leq h_{jk} K_0. \end{aligned}$$

The second inequality in (4.23) follows similarly. Next observe that we can write

$$I(b_{jk} \cap b_{\ell m}) = \int_{c_{j1}^{jk}}^{c_{j1}^{jk} + h_{jk}} \int_{c_{\ell 1}^{\ell m}}^{c_{\ell 2}^{\ell m} + h_{\ell m}} I_{j,\ell}(x_j, x_\ell, \widehat{B}^j, \widehat{\ell}) dx_j dx_\ell.$$

This implies with K_2 from **(A4)** $_{\beta_0}$ that

$$I(b_{jk} \cap b_{\ell m}) \leq K_2 h_{jk} h_{\ell m} \leq K_2 \alpha_n^2 / \epsilon_0^2. \quad (4.24)$$

A similar inequality (with different constants) also holds for I replaced by I_n . This follows directly from (4.23) together with the definition A_n and assumption (4.6).

Now observe that we have show in the proof above that if $\rho_\infty(B^*, \widehat{B}) > C 2d \alpha_n$ for some $C > 0$ large enough, then on A_n the difference $I_n(b_{jk}) - I_n(b_{\ell m}) \geq \tilde{C} \alpha_n^2$ where \tilde{C} increases with C . We also have seen that $I_n(b_{jk} \cap b_{\ell m}) \leq C' \alpha_n^2$ for some

constant $C' > 0$ (also on A_n). These two inequalities imply that on A_n the overlap is negligible for C large enough. (Further details are omitted.)

It remains to show (4.12). Write B_{k_n} for the outcome of the peeling procedure (after k_n peels). Suppose there exists a set $B^* \in \mathcal{N}_{loc}(\beta_0) \setminus m_{loc, \epsilon_1/d}(\beta_0)$ with $\rho_\infty(B_{k_n}, B^*) \leq c \alpha_n$ for some $c > 0$. We now show that this implies that on A_n jittering can be applied, leading to a sequence $B_{k_n+1}, B_{k_n+2}, \dots, B_{K_n}$ with $B_{K_n} = \widehat{B}$ being the final outcome of peeling+jittering, satisfying $\rho_\infty(\widehat{B}, B^*) \geq 2C \alpha_n$, with the same C as in (4.11). This essentially completes the proof, because the arguments of the above proof can be repeated to show that this leads to a contradiction.

Without loss of generality assume that $A_{11}^+(\partial B_{k_n})$ exists. We can also assume that $A_{11}^+(\partial B_{k_n}) \geq A_{12}^-(\partial B_{k_n})$ (cf. proof of Lemma 3.1). Further let $b_{11}^{k_n}, b_{12}^{k_n}$ be two corresponding candidate sets for jittering, $b_{11}^{k_n}$ lying outside of B_{k_n} and $b_{12}^{k_n}$ inside, and $F_n(b_{11}^{k_n}) = F_n(b_{12}^{k_n})$. We show that on A_n with n large enough we have

$$I_n(b_{11}^{k_n}) > I_n(b_{12}^{k_n}). \quad (4.25)$$

Consequently, adding $b_{11}^{k_n}$ and subtracting $b_{12}^{k_n}$ leads to B_{k_n+1} with a larger average than B_{k_n} . Repeating this process leads to B_{k_n+2} with $I_n(B_{k_n+2})/F_n(B_{k_n+2}) > I_n(B_{k_n+1})/F_n(B_{k_n+1})$ etc. This process can be repeated (at least) as long as we are still in the neighborhood $\mathcal{U}(\epsilon_1/d, B_{\beta_0}^*)$ which means until for some p we have $(b_{11}^{k_n+p+1} \cup b_{12}^{k_n+p+1}) \setminus (\overline{B} \setminus \underline{B}) \neq \emptyset$. Since $d\alpha_n = o(1)$ and the widths of the candidate sets are of the order $O(\alpha_n)$ it follows that for α_n small enough $d_F(B^*, B_{K_n}) \geq c \epsilon_1/d \geq C \epsilon_1 \alpha_n$.

To complete the proof we now show (4.25). Our assumptions assure that we can assume both $F_1(\cdot, B^{\widehat{1}})$ and $I_1(\cdot, B^{\widehat{1}})$ to be continuous (except in a_{11} and a_{12}).

We have

$$\begin{aligned}
I(b_{11}^{k_n}) &= \int_{a_{11}-h_{11}}^{a_{11}} \frac{I_1(x, B^{\hat{1}})}{F_1(x, B^{\hat{1}})} F_1(x, B^{\hat{1}}) dx \\
&= \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} F(b_{11}^{k_n}) + \int_{a_{11}-h_{11}}^{a_{11}} \left(\frac{I_1(x, B^{\hat{1}})}{F_1(x, B^{\hat{1}})} - \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} \right) F_1(x, B^{\hat{1}}) dx \\
&\geq \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} F(b_{11}^{k_n}) + \int_{a_{11}-h_{11}}^{a_{11}} k_1 |x - a_{11}| F_1(x, B^{\hat{1}}) dx \\
&\geq \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} F(b_{11}^{k_n}) + k_1 \inf_x F_1(x, B^{\hat{1}}) \int_{a_{11}-h_{11}}^{a_{11}} |x - a_{11}| dx \\
&= \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} F(b_{11}^{k_n}) + \frac{k_1 \inf_x F_1(x, B^{\hat{1}})}{2} h_{11}^2. \tag{4.26}
\end{aligned}$$

A similar, but simpler, argument shows that $I(b_{12}^{k_n}) \leq \frac{I_1(a_{12}, B^{\hat{1}})}{F_1(a_{12}^-, B^{\hat{1}})} F(b_{12}^{k_n})$ and hence

$$\begin{aligned}
&I(b_{11}^{k_n}) - I(b_{12}^{k_n}) \\
&\geq \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} F(b_{11}^{k_n}) - \frac{I_1(a_{12}^-, B^{\hat{1}})}{F_1(a_{12}^-, B^{\hat{1}})} F(b_{12}^{k_n}) + \frac{k_1 \inf_x F_1(x, B^{\hat{1}})}{2} h_{11}^2 \\
&\geq \frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} (F(b_{11}^{k_n}) - F(b_{12}^{k_n})) + \frac{k_1 \inf_x F_1(x, B^{\hat{1}})}{2} h_{11}^2 \tag{4.27}
\end{aligned}$$

The first term in the last line could still be negative. However, we can show that on A_n this term actually is of smaller order than the second. This follows from the fact that

$$F(b_{11}^{k_n}) - F(b_{12}^{k_n}) = (F_n - F)(b_{12}^{k_n}) - (F_n - F)(b_{11}^{k_n}),$$

and since $F_n(b_{11}^{k_n}) = F_n(b_{12}^{k_n}) \leq \alpha_n$ we obtain from definition of A_n that

$$|F(b_{11}^{k_n}) - F(b_{12}^{k_n})| \leq C_1 \sqrt{\frac{d}{n} \alpha_n \log \frac{1}{\alpha_n}} = o(\alpha_n^2).$$

The last equality follows from assumption on α_n . On the other hand, we have seen that $h_{11} \geq \text{const.} \alpha_n$ (see (4.23)), and hence on A_n we have (by continuing (4.27)) that (for n large enough)

$$I(b_{11}^{k_n}) - I(b_{12}^{k_n}) \geq -\frac{I_1(a_{11}^+, B^{\hat{1}})}{F_1(a_{11}^+, B^{\hat{1}})} o(\alpha_n^2) + \frac{k_1 \inf_x F_1(x, B^{\hat{1}})}{2} \text{const.} \alpha_n^2 \geq \text{const.}' \alpha_n^2.$$

Hence we have shown (4.25). The fact that on A_n this lower bound translates to a similar bound for the difference of the I_n measures (rather than the I -measures)

follows by using similar arguments used above to show that (4.17) implies (4.22). As has been outlined above, this completes the proof.

□

4.3 Empirical Performance of Covering

The peeling+jittering (or peeling+pasting) procedure is applied iteratively, each time removing the optimal set found by peeling+jittering and applying the procedure to what is left over. In other words, the input space S is different for every iteration step. For this reason we need an additional condition to ensure that it will be possible to compensate to a certain degree the small errors made in each step. We will assume that

(A6)_{β₀} For all $j = 1, \dots, d$ and $k = 1, 2$ there exists a constant $p > 0$ such that we have uniformly in $B_1, B_2 \in \mathcal{B}$ with $F(B_1|S) > \beta_0/2$ and $F(B_2|S) > \beta_0/2$ that

$$\begin{aligned} \sup_x |I_j(x, \widehat{B}_1^j) - I_j(x, \widehat{B}_2^j)| &\leq p \rho_\infty(B_1, B_2), \quad \text{and} \\ \sup_x |F_j(x, \widehat{B}_1^j) - F_j(x, \widehat{B}_2^j)| &\leq p \rho_\infty(B_1, B_2), \end{aligned}$$

This assumption for instance holds if the function $x \rightarrow m(x) \cdot f(x)$, $x \in S = [0, 1]^d$ is Lipschitz continuous.

For a given subset $R \subset [0, 1]$ and $\epsilon > 0$ let $\mathcal{O}_{loc,\epsilon}(\beta_0, R) = \mathcal{N}_{loc}(\beta_0, R) \setminus m_{loc,\epsilon}(\beta_0, R)$ denote a class of ‘optimal’ peeling+jittering outcomes with \mathcal{N}_{loc} and $m_{loc,\epsilon}$ as defined as in (3.9) and (3.10), respectively, but replacing S by R . Recall that here $\epsilon > 0$ indicates some regularity condition on boundary averages required to be satisfied by sets in $m_{loc,\epsilon}$ (see (3.10)). Now set

$$\mathcal{O}_{loc,\epsilon}^{(1)}(\beta_0) = \mathcal{O}_{loc,\epsilon}(\beta_0, [0, 1]^d).$$

Observe that Theorem 4.2 deals with this class. For $k = 2, 3, \dots$ define iteratively

$$\mathcal{O}_{loc,\epsilon}^{(k)}(\beta_0) = \bigcup_{\{B_{(j)}^* \in \mathcal{O}_{loc,\epsilon}^{(j)}(\beta_0), j=1, \dots, k-1\}} \mathcal{O}_{loc,\epsilon}(\beta_0, \overline{S \setminus \bigcup_{j=1}^{k-1} B_{(j)}^*}), \quad (4.28)$$

which are classes of possible outcomes of peeling+jittering after the k -th application of the procedure. Using this notation we define the class of all possible covering outcomes after k iterative peeling+jittering applications as

$$\mathcal{O}_{cover,\epsilon}^{(k)}(\beta_0) = \left\{ \bigcup_{j=1}^k \widehat{B}_{(j)}^*, B_{(j)}^* \in \mathcal{O}_{loc,\epsilon}^{(j)}(\beta_0) \right\}. \quad (4.29)$$

For a given β_0 let $\widehat{B}_{(1)}, \dots, \widehat{B}_{(K)}$ denote outcomes of K consecutive applications of empirical peeling+jittering. With

$$\widehat{K}(\lambda) = \left\{ \widehat{B}_{(k)}^* : \frac{I_n \left(\widehat{B}_{(k)} \setminus \bigcup_{j=1}^{k-1} \widehat{B}_{(j)} \right)}{F_n \left(\widehat{B}_{(k)} \setminus \bigcup_{j=1}^{k-1} \widehat{B}_{(j)} \right)} > \lambda \right\},$$

let

$$\widehat{R}_\lambda = \bigcup_{\widehat{B}_{(k)} \in \widehat{K}(\lambda)} \widehat{B}_{(k)}.$$

denote covering outcomes. Analogously, for $B_{(k)}^* \in \mathcal{O}_{loc,\epsilon}^{(k)}(\beta_0)$, $k = 1, \dots, K$ and a given λ let

$$K(\lambda) = \left\{ B_{(k)}^* : \frac{I \left(B_{(k)}^* \setminus \bigcup_{j=1}^{k-1} B_{(j)}^* \right)}{F \left(B_{(k)}^* \setminus \bigcup_{j=1}^{k-1} B_{(j)}^* \right)} > \lambda \right\},$$

and

$$R_\lambda^* = R_\lambda^*(B_{(1)}^*, \dots, B_{(K)}^*) = \bigcup_{B_{(k)}^* \in K(\lambda)} B_{(k)}^*.$$

Theorem 4.3 *Let $0 < \beta_0 < 1$, $\epsilon_1 > 0$ and λ be fixed. Suppose that for all $B^* \in \mathcal{N}_{loc}^{(k)}(\beta_0)$, $k = 1, \dots, K$ for some $K \geq 1$ the assumptions of Theorem 4.2 and **(A6)** $_{\beta_0}$ hold with $F(\cdot | S)$ replaced by $F(\cdot | B^*)$. Then there exists $R_\lambda^* = R_\lambda^*(B_{(1)}^*, \dots, B_{(K)}^*)$, $B_{(k)}^* \in \mathcal{O}_{loc,\frac{\epsilon_1}{d}}^{(k)}(\beta_0)$ such that the rates of convergence asserted in Theorem 4.2 also hold for $d_F(R_\lambda^*, \widehat{R}_\lambda)$.*

Proof. We present the proof for $K = 2$. Let $\widehat{B}_{(1)}$ and $\widehat{B}_{(2)}$ denote two successive outcomes of empirical peeling+jittering. We have seen in Theorem 4.2 and its proof that there exists a set A_n (with $P(A_n) \rightarrow 1$ as $n \rightarrow \infty$), and a set $B_{(1)}^* \in \mathcal{O}_{loc,\frac{\epsilon_1}{d}}^{(1)}(\beta_0)$

such that on A_n we have $\rho_\infty(\widehat{B}_{(1)}, B_{(1)}^*) \leq 2C\eta_n$ and $d_F(\widehat{B}_{(1)}, B_{(1)}^*) \leq 2Cd\eta_n$. We hence obtain that on A_n

$$\begin{aligned}
& \inf_{B^* \in \mathcal{O}_{\text{cover}, \frac{\epsilon_1}{d}}^{(2)}(\beta_0)} d_F(\widehat{B}_{(1)} \cup \widehat{B}_{(2)}, B^*) \\
& \leq \inf_{B_{(2)}^* \in \mathcal{O}_{\text{loc}, \frac{\epsilon_1}{d}}^{(2)}(\beta_0)} \left[d_F(\widehat{B}_{(1)}, B_{(1)}^*) + d_F(\widehat{B}_{(2)}, B_{(2)}^* \setminus B_{(1)}^*) \right] \\
& \leq 2Cd\eta_n + \inf_{B_{(2)}^* \in \mathcal{O}_{\text{loc}, \frac{\epsilon_1}{d}}^{(2)}(\beta_0)} d_F(\widehat{B}_{(2)}, B_{(2)}^* \setminus B_{(1)}^*)
\end{aligned}$$

Further we have with $S^{(2)} = S \setminus B_{(1)}^*$

$$\begin{aligned}
d_F(\widehat{B}_{(2)}, B_{(2)}^* \setminus B_{(1)}^*) &= F(S^{(2)}) \cdot d_{F(\cdot|S^{(2)})}(\widehat{B}_{(2)}, B_{(2)}^*) \\
&= (1 - \beta_0) \cdot d_{F(\cdot|S^{(2)})}(\widehat{B}_{(2)}, B_{(2)}^*),
\end{aligned}$$

and therefore it suffices to show that for n large enough and some $c^* > 0$ we have

$$P\left(\inf_{B_{(2)}^* \in \mathcal{O}_{\text{loc}, \frac{\epsilon_1}{d}}^{(2)}(\beta_0)} d_{F(\cdot|S^{(2)})}(\widehat{B}_{(2)}, B_{(2)}^*) > c^*Cd\eta_n, A_n\right) = 0.$$

We follow the ideas of the proof of Theorem 4.2 with S replaced by $S^{(2)}$. Let $\widehat{S}^{(2)} = S \setminus \widehat{B}_{(1)}$ and let $\widehat{\beta}_n := F(\widehat{B}_{(2)}|S^{(2)})$. We obtain that on A_n (for n large enough)

$$\begin{aligned}
\beta_0 &\leq \widehat{\beta}_n \leq F_n(\widehat{B}_{(2)}|S^{(2)}) - (F_n - F)(\widehat{B}_{(2)}|S^{(2)}) \\
&\leq F_n(\widehat{B}_{(2)}|\widehat{S}^{(2)}) + F_n(\widehat{B}_{(1)} \Delta B_{(1)}^*) - (F_n - F)(\widehat{B}_{(2)}|S^{(2)}) \\
&= F_n(\widehat{B}_{(2)}|\widehat{S}^{(2)}) + F(\widehat{B}_{(1)} \Delta B_{(1)}^*) + (F_n - F)(\widehat{B}_{(1)} \Delta B_{(1)}^*) - (F_n - F)(\widehat{B}_{(2)}|S^{(2)}) \\
&\leq (\beta_0 + \alpha_n) + 2Cd\eta_n + \frac{2}{1 - \beta_0} C_1 \sqrt{\frac{d}{n} \log d} \\
&\leq \beta_0 + (2Cd + 1 + \frac{1}{1 - \beta_0} C_1) \eta_n \\
&\leq \beta_0 + 3Cd\eta_n
\end{aligned}$$

for C large enough. Similar to (4.14) we now obtain that for C and n large enough we have

$$\begin{aligned} & \{ \inf_{B_{(2)}^* \in \mathcal{O}_{loc, \frac{\epsilon_1}{d}}^{(2)}(\beta_0)} \rho_\infty(B_{(2)}^*, \widehat{B}_{(2)}) \geq c^* C \eta_n, A_n \} \\ & \subset \{ \inf_{B_{\widehat{\beta}_n, (2)}^* \in \mathcal{O}_{loc, \frac{\epsilon_1}{d}}^{(2)}(\widehat{\beta}_n)} \rho_\infty(B_{\widehat{\beta}_n, (2)}^*, \widehat{B}_{(2)}) \geq (c^* - 4) C \eta_n, A_n \}, \end{aligned} \quad (4.30)$$

for any constant $c^* > 0$. Observe further that the function $\Psi(\cdot)$ depends on $F(\cdot|S)$ (through the quantities $A_{jk}^\pm(\partial B)$, and that $F(\cdot|S)$ is now replaced by $F(\cdot|S^{(2)})$). To make this more clear in the notation we denote the corresponding function $\Psi(\cdot)$ by $\Psi(\cdot|S^{(2)})$.

Hence, assuming that

$$\inf_{B_{(2)}^* \in \mathcal{O}_{loc, \frac{\epsilon_1}{d}}^{(2)}(\beta_0)} \rho_\infty(B_{(2)}^*, \widehat{B}_{(2)}) \geq c^* C \eta_n, \quad (4.31)$$

assumption **(A5)** _{β_0} (with $F(\cdot)$ replaced by $F(\cdot|S^{(2)})$) and (4.30) give us that on A_n

$$\Psi(\widehat{B}_{(2)}|S^{(2)}) \geq c(c^* - 4) C \eta_n. \quad (4.32)$$

Without loss of generality assume that $\Psi(\widehat{B}_{(2)}|S^{(2)}) = A_{11}^+(\partial \widehat{B}_{(2)}|S^{(2)}) - A_{22}^-(\partial \widehat{B}_{(2)}|S^{(2)})$, so that

$$A_{11}^+(\partial \widehat{B}_{(2)}|S^{(2)}) - A_{22}^-(\partial \widehat{B}_{(2)}|S^{(2)}) \geq c(c^* - 4) C \eta_n. \quad (4.33)$$

We will show that (4.33) implies the existence of a constant $c' > 0$ with

$$A_{11}^+(\partial \widehat{B}_{(2)}|\widehat{S}^{(2)}) - A_{22}^-(\partial \widehat{B}_{(2)}|\widehat{S}^{(2)}) \geq c'(c^* - 4) C \eta_n. \quad (4.34)$$

We need (4.34) rather than (4.33) (i.e. $\widehat{S}^{(2)}$ rather than $S^{(2)}$), because following the key arguments from the proof of Theorem 4.2 inequality (4.34) implies the existence of two candidate sets b_{jk} and $b_{\ell m}$ for jittering with $I_n(b_{jk}|\widehat{S}^{(2)}) > I_n(b_{\ell m}|\widehat{S}^{(2)})$ which is a contradiction by definition of peeling, and hence (4.31) cannot be true. (The similar inequality with $\widehat{S}^{(2)}$ replaced by $S^{(2)}$ does not lead to a contradiction because empirical peeling+jittering is performed conditional on $\widehat{S}^{(2)}$.)

To see that (4.33) implies (4.34) if $(\mathbf{A6})_{\beta_0}$ holds it is enough to show that there exists a constant $K' > 0$ (not depending on c^*) such that for $(jk) = (11), (22)$ we have

$$|A_{jk}^{\pm}(\partial\widehat{B}_{(2)}|\widehat{S}^{(2)}) - A_{jk}^{\pm}(\partial\widehat{B}_{(2)}|S^{(2)})| \leq K' C \eta_n. \quad (4.35)$$

This is sufficient, because the constant c^* in (4.32) can be chosen large enough. We only present the proof of (4.35) for $(jk) = (11)$. Write

$$B_{(i)}^* = \bigotimes_{j=1}^d [a_{j1}^{(i)}, a_{j2}^{(i)}], \quad \text{and} \quad \widehat{B}_{(i)} = \bigotimes_{j=1}^d [\widehat{a}_{j1}^{(i)}, \widehat{a}_{j2}^{(i)}], \quad i = 1, 2.$$

We have

$$A_{11}^{\pm}(\partial\widehat{B}_{(2)}|S^{(2)}) = \frac{I_1(a_{11}^{\pm}, \widehat{B}_{(2)}^{\widehat{1}} \cap S^{(2), \widehat{1}})}{F_1(a_{11}^{\pm}, \widehat{B}_{(2)}^{\widehat{1}} \cap S^{(2), \widehat{1}})},$$

and a similar equation holds for $S^{(2)}$ replaced by $\widehat{S}^{(2)}$. As for the numerator of the right hand side we then have by using $(\mathbf{A6})_{\beta_0}$

$$\begin{aligned} & |I_1(a_{11}^{\pm}, \widehat{B}_{(2)}^{\widehat{j}} \cap S^{(2), \widehat{1}}) - I_1(a_{11}^{\pm}, \widehat{B}_{(2)}^{\widehat{j}} \cap \widehat{S}^{(2), \widehat{j}})| \\ & \leq |I_1(a_{11}^{\pm}, \widehat{B}_{(1)}^{\widehat{j}}) - I_1(a_{11}^{\pm}, B_{(1)}^{\widehat{j}})| \\ & \leq p \rho_{\infty}(\widehat{B}_{(1)}, B_{(1)}) \leq p 2 C \eta_n \end{aligned}$$

A similar inequality can be shown for $F_1(a_{11}^{\pm}, \widehat{B}_{(2)}^{\widehat{1}} \cap S^{(2), \widehat{1}})$ and this implies that (4.35) holds. In fact, this implication can be seen by using the same arguments as in the proof of Lemma 4.1. Observe here that we have already seen that $\widehat{\beta}_n = F(\widehat{B}_{(2)}|S^{(2)})$ is bounded away from zero, and hence assumption $(\mathbf{A2})_{\beta_0}$ applies which provides us with $F_j(a_{11}^{\pm}, \widehat{B}_{(2)}^{\widehat{j}} \cap S^{(2), \widehat{j}})$ being bounded away from zero uniformly in d and j .

We have shown that under the present assumptions the outcomes of 2 successive applications of peeling+jittering converges to (one of) their theoretical counterparts. Clearly, this also implies that the corresponding I -measures and the corresponding F -measures converge, so that the sets $\widehat{K}(\lambda) \rightarrow K(\lambda)$ in the sense that $\widehat{K}(\lambda)\Delta K(\lambda) \rightarrow 0$ in probability as $n \rightarrow \infty$. This implies the result.

□

5 Appendix

Here we present proofs of some technical results presented above. We also provide some additional technical results which are tools in the proofs of our main results.

Proof of Lemma 3.1: Suppose that B is such that $F(B|S) = \beta_0$ and $\Psi(B) > 0$. Then there exists $\epsilon > 0$ and two boundary facets (jk) and (ℓm) with

$$A_{jk}^+(\partial B) - A_{\ell m}^-(\partial B) > \epsilon. \quad (5.1)$$

Let b_{jk} and $b_{\ell m}$ be two boxes adjacent to ∂B_{jk} and $\partial B_{\ell m}$, respectively, with b_{jk} lying outside the box B and $b_{\ell m}$ inside. Our assumption assure that we can chose those boxes b_{jk} and $b_{\ell m}$ such that $F((b_{jk} \setminus b_{\ell m}) \cap S) = F(b_{\ell m} \cap S) < \frac{\epsilon}{2C} < \frac{\beta_0}{2}$. It follows from Lemma 4.1 that there exists a constant $C > 0$ such that both

$$\begin{aligned} \left| \frac{I((b_{jk} \setminus b_{\ell m}) \cap S)}{F((b_{jk} \setminus b_{\ell m}) \cap S)} - A_{jk}^+(\partial(B \setminus b_{\ell m})) \right| &\leq C F((b_{jk} \setminus b_{\ell m}) \cap S), \quad \text{and} \\ \left| \frac{I(b_{\ell m} \cap S)}{F(b_{\ell m} \cap S)} - A_{\ell m}^-(\partial B) \right| &\leq C F(b_{\ell m} \cap S). \end{aligned} \quad (5.2)$$

This implies $I((b_{jk} \setminus b_{\ell m}) \cap S) > I(b_{\ell m} \cap S)$, and hence removing $b_{\ell m}$ and then adding b_{jk} increases the I -measure but leaves the $F(\cdot|S)$ -measure constant, and hence $B \notin \mathcal{M}_{loc}(\beta_0)$.

The assertion that $\mathcal{M}_{loc}(\beta_0) \cap m_{loc,\epsilon}(\beta_0) = \emptyset$ can be seen as follows. First observe that for $B \in m_{loc,\epsilon}(\beta_0)$ we actually have $\psi(B) = 0$, for if $-\infty < \Psi(B) < 0$ then by definition of $\Psi(B)$ we can further decrease the average via jittering. This follows by a similar argument provided above to show that we can increase the average of a set B if $\Psi(B) > 0$. Similarly we can see that in fact we need $A_{jk}^+ = A_{\ell m}^-$ for all $(\ell m) \neq (jk)$, and thus all $A_{\ell m}^+(\partial B) = A^+$ and $A_{jk}^-(\partial B) = A^-$ for some A^+, A^- , and since $\Psi(B) = 0$ we even need to have $A^+ = A^-$. Hence, if $B \in m_{loc,\epsilon}(\beta_0)$ then, because boundary averages are equal, it is clear that (m.i) implies that adding

a small set to a boundary ($j1$) and subtracting a set of same $F(\cdot|S)$ -measure at boundary ($j2$) leads to an increase in the average and hence $B \notin \mathcal{M}_{loc}(\beta_0)$.

□

Proof of Lemma 4.1 Since the proof for I^+/F^+ is the same as for I^-/F^- we drop the superscripts ‘ \pm ’. Let $B = \bigotimes_{j=1}^d [a_{j1}, a_{j2}]$ and let h_{jk} denote the width of b_{jk} such that we have $F(b_{jk}) = \int_{a_{jk}}^{a_{jk}+h_{jk}} F_j(x, B^{\hat{j}}) dx$. We can assume that there is no discontinuity of F_j and I_j except at a_{jk} (because by assumption we only have finitely many discontinuities and $h_{jk} \rightarrow 0$). By using $(\mathbf{A2})_{\beta_0}$ we thus obtain

$$\begin{aligned} |F(b_{jk}) - h_{jk} F_j(a_{jk}, B^{\hat{j}})| &= \left| \int_{a_{jk}}^{a_{jk}+h_{jk}} (F_j(x, B^{\hat{j}}) - F_j(a_{jk}, B^{\hat{j}})) dx \right| \\ &\leq K_1 \int_{a_{jk}}^{a_{jk}+h_{jk}} |x - a_{jk}| dx \leq K_1 h_{jk}^2, \end{aligned}$$

where K_1 denotes the Lipschitz constant of $F_j(\cdot, B^{\hat{j}})$. By using $(\mathbf{A3})_{\beta_0}$ similar estimates holds for F_j replaced by I_j . We thus obtain

$$\frac{I(b_{jk})}{F(b_{jk})} = \frac{h_{jk} I_j(a_{jk}, B^{\hat{j}}) + O(h_{jk}^2)}{h_{jk} F_j(a_{jk}, B^{\hat{j}}) + O(h_{jk}^2)} = \frac{I_{jk}(\partial B)}{F_{jk}(\partial B)} + O(h_{jk}).$$

The last equality uses the fact that $F_{jk}(\partial B)$ is bounded away from zero. It remains to observe that h_{jk} is of the same order as $F(b_{jk})$ (cf. (4.23)).

□

Next we present two technical results which are important tools in the proofs presented above. The result essentially are Theorem 2.14.1 and Theorem 2.14.2 from van der Vaart and Wellner (1995) which, however, had to be adapted to our situation.

To this end we need to introduce some notation. Let \mathcal{G} be a class of functions with $\|g\|^2 = \text{E}g^2(X, Y) < \infty$, and let $N_B(u, \mathcal{G})$ be defined as the smallest number of L_2 -brackets of size u needed to cover \mathcal{G} . An L_2 -bracket $[g_1, g_2]$ in \mathcal{G} of size u is defined as $[g_*, g^*] = \{g \in \mathcal{G} : g_* \leq g \leq g^*\}$, with $\|g^* - g_*\| \leq u$. The quantity

$N_B(u, \mathcal{G})$ is called the covering number (with bracketing) of \mathcal{G} . The metric entropy with bracketing of \mathcal{G} with respect to the L_2 -norm is defined as

$$H_B(u, \mathcal{G}) := \log N_B(u, \mathcal{G}).$$

Covering numbers for classes of sets are defined analogously by identifying sets with their indicator functions.

Lemma 5.1 *We have*

$$H_B(u, \mathcal{B}) \leq 2d \log \lceil \frac{d}{u^2} \rceil. \quad (5.3)$$

Proof. First partition each of the d coordinate axis into $\lceil \frac{1}{u} \rceil$ intervals each with marginal probability measure $\leq u$. With F_i denoting the i -th marginal distribution function this can be done by using $a_{(i,k)} = F_i^{-1}(ku)$, $k = 1, \dots, \lceil \frac{1}{u} \rceil - 1$ as well as $F_i^{-1}(0)$ and $F_i^{-1}(1)$. Now consider the set of all rectangles determined by picking two of the partitioning points in each coordinate as lower and upper boundary. Notice that there are $\lceil \frac{1}{u} \rceil + 1$ such points in each coordinate. This results in a set \mathcal{B}_u consisting of $(\lceil \frac{1}{u} \rceil + 1)^d$ rectangles. By construction, any rectangle $B \in \mathcal{B}$ has a lower and an upper approximation $B_*, B^* \in \mathcal{B}_u$ with $F(B^* \Delta B_*) \leq du$. Hence we have shown that $H_B(\sqrt{du}, \mathcal{B}) \leq d \log (\lceil \frac{1}{u} \rceil + 1) = d \log \frac{(\lceil \frac{1}{u} \rceil + 1)^{\lceil \frac{1}{u} \rceil}}{2} \leq 2d \log \lceil \frac{1}{u} \rceil$. It follows that $H_B(u, \mathcal{B}) \leq 2d \log \lceil \frac{d}{u^2} \rceil$. \square

Proposition 5.1 *Suppose that $\{(X_i, Y_i), i = 1, \dots, n\}$ are iid and continuous random variables with $E(Y_1^2 | X_1) < M < \infty$ a.s., and $E|Y_1|^\gamma < \infty$ for some $\gamma \geq 2$. Then there exists a universal constant $C_0 > 0$ and a $\delta_0 > 0$ such that for $0 < \delta < \delta_0$ we have*

$$E\left(\sup_{F(B) \leq \delta; B \in \mathcal{B}} |(I_n - I)(B)| \right) \leq C_0 \left(\sqrt{M \frac{d}{n} \delta \log \lceil \frac{d}{\delta} \rceil} + \left(M^2 \frac{d}{n\delta} \log \lceil \frac{d}{\delta} \rceil \right)^{\frac{\gamma-1}{2}} \right).$$

Proof. Write $\sqrt{n}(I_n - I)(B) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_B(X_i, \eta_i) - E g_B(X_i, Y_i)$ with $g_B(x, y) = y \mathbf{1}(x \in B)$. In other words $\sqrt{n}(I_n - I)(B)$ is an empirical process indexed by $\mathcal{G} = \{g_B, B \in \mathcal{B}\}$. Let further

$$g_B^\pm(x, y) = y^\pm \mathbf{1}(x \in B),$$

where $b^+ = b \vee 0$ and $b^- = -(b \wedge 0)$. By definition both g_B^\pm are positive functions.

Since $\text{E}g_B(X_1, Y_1) = \text{E}g_B^+(X_1, Y_1) - \text{E}g_B^-(X_1, Y_1)$ we have

$$\begin{aligned} \sqrt{n}(I_n - I)(B) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g_B^+(X_i, Y_i) - \text{E}g_B^+(X_i, Y_i)] \\ &\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^n [g_B^-(X_i, Y_i) - \text{E}g_B^-(X_i, Y_i)] \\ &=: \nu_n(g_B^+) - \nu_n(g_B^-). \end{aligned}$$

Let $\mathcal{G}^\pm = \{g^\pm, g \in \mathcal{G}\}$. Notice that $F(B) \leq \delta$ implies that $\|g_B^\pm\| \leq \sqrt{M\delta}$. Hence

$$\text{E}\left(\sup_{F(B) \leq \delta} \sqrt{n}|(I_n - I)(B)|\right) \leq \text{E}\left(\sup_{\|g_B^+\| \leq \sqrt{M\delta}} |\nu_n(g_B^+)|\right) + \text{E}\left(\sup_{\|g_B^-\| \leq \sqrt{M\delta}} |\nu_n(g_B^-)|\right).$$

In order to bound the r.h.s. we will apply Theorem 2.14.2 of van der Vaart and Wellner (1995) to both processes $\{\nu_n(g_B^+), g^+ \in \mathcal{G}^+\}$ and $\{\nu_n(g_B^-), g^- \in \mathcal{G}^-\}$. Notice that $|g_B^\pm(x, y)| \leq |y|$. Hence the function $G(y) = y$ is an envelope for both classes \mathcal{G}^\pm and we have $\|G\| \leq \sqrt{M}$.

Also, let

$$a(u) := \frac{\sqrt{M} u}{\sqrt{1 + H_B(\sqrt{M} u, \mathcal{G}^\pm)}}.$$

With this notation Theorem 2.14.2 of van der Vaart and Wellner says that

$$\text{E}\left(\sup_{\|g_B^\pm\| \leq \sqrt{M\delta}} |\nu_n(g_B^\pm)|\right) \leq \int_0^{\sqrt{M\delta}} \sqrt{1 + H_B(u\sqrt{M}, \mathcal{G}^\pm)} du \quad (5.4)$$

$$+ \sqrt{n} \text{E}[|Y_1| \mathbf{1}\{|Y_1| \geq a(\sqrt{M\delta})\sqrt{n}\}]. \quad (5.5)$$

We will now estimate both of the quantities on the r.h.s. This will give the assertion of our theorem. First we estimate the metric entropy of \mathcal{G}^\pm by using (5.3). For $B_* \subset B \subset B^*$ we have

$$g_{B_*}^\pm - \text{E}(g_{B_*}^\pm) \leq g_B^\pm - \text{E}(g_B^\pm) \leq g_{B^*}^\pm - \text{E}(g_{B^*}^\pm),$$

and since $E([g_{B^*}^\pm - E(g_{B^*}^\pm)] - [g_{B_*}^\pm - E(g_{B_*}^\pm)]) \leq M F(B^* \Delta B_*)$ it follows that

$$H_B(u\sqrt{M}, \mathcal{G}^\pm) \leq H_B(u, \mathcal{B}) \leq 2d \log \lceil \frac{d}{u^2} \rceil.$$

We now can estimate the r.h.s. of (5.4). For $0 < \delta < \delta_0 := \frac{1}{\sqrt{e^{1/2}-1}}$ we have $1 < 2d \log \lceil \frac{d}{u^2} \rceil$ and hence we obtain for such δ that

$$\begin{aligned} \int_0^{\sqrt{M\delta}} \sqrt{1 + H_B(u\sqrt{M}, \mathcal{G}^\pm)} du &\leq \int_0^{\sqrt{M\delta}} \sqrt{1 + 2d \log \lceil \frac{d}{u^2} \rceil} du \\ &\leq 2 \sqrt{M\delta d \log \lceil \frac{d}{M\delta} \rceil}. \end{aligned}$$

It remains to estimate (5.5). We have for $\gamma > 1$ and $0 < \delta < \delta_0$ that

$$\begin{aligned} &E[|Y_1| \mathbf{1}\{|Y_1| \geq a(\sqrt{M\delta})\sqrt{n}\}] \\ &\leq \frac{1}{(a(\sqrt{M\delta})\sqrt{n})^{\gamma-1}} E[|Y_1|^\gamma \mathbf{1}\{|Y_1| \geq a(\sqrt{M\delta})\sqrt{n}\}] \\ &\leq (4M^2 \frac{d}{n\delta} \log \lceil \frac{d}{\delta} \rceil)^{\frac{\gamma-1}{2}} E[|Y_1|^\gamma \mathbf{1}\{|Y_1| \geq a(\sqrt{M\delta})\sqrt{n}\}]. \end{aligned}$$

This completes the proof of the proposition, since the last expected value is bounded.

□

Next we present a result for the standard empirical process using random entropy numbers. To this end let \mathcal{G} be a class of functions. Let further $N(u, \mathcal{G}, Q)$ denote the smallest number of $L_2(Q)$ -balls needed to cover \mathcal{G} . Using the fact that \mathcal{B} is a VC-class with VC-index $2d+1$, we have the well-known estimate (for a definition of VC-class and VC-index, as well as the estimate see e.g. van der Vaart and Wellner (1995), section 2.6.1):

$$\sup_Q \log N(u, \mathcal{B}, Q) \leq K d \log \frac{4e}{u}, \quad 0 < u < 1.$$

Here $K > 0$ is a universal constant. W.l.o.g. we assume $K \geq 1$. Using this result the following proposition is a straightforward corollary to Theorem 2.14.1 of van der Vaart and Wellner (1995).

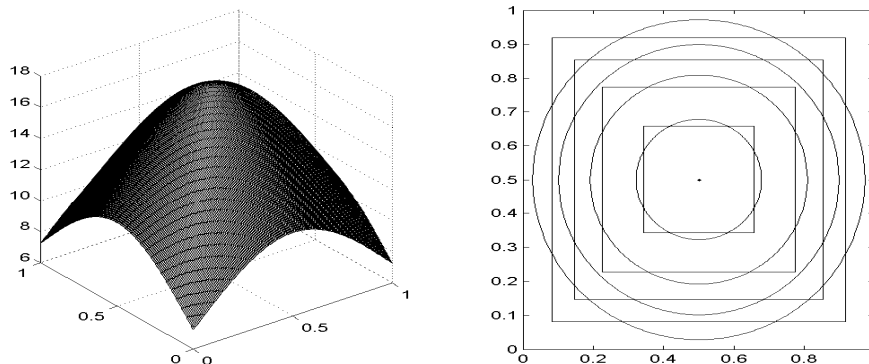
Proposition 5.2 *There exists a universal constant $C_0 > 0$ such that for $0 < \delta < 1$ we have*

$$E\left(\sup_{F_n(B) \leq \delta; B \in \mathcal{B}} |(F_n - F)(B)|\right) \leq C_0 \sqrt{\frac{d}{n} \delta \left(1 + \log \frac{1}{\delta}\right)}.$$

References

- [1] Becker, U. and Fahrmeier, L. (2001) Bump hunting for risk: a new data mining tool and its applications. *Computational Statistics* **16** (3) 373–386.
- [2] Brillinger, R. D. (1994) Examples of Scientific problems and data analyses in demography, neurophysiology, and seismology. *J. Comput. and Graph. Statist.* **3** 1–22.
- [3] Cole, S.W., Galic, Z. and Zack, J.A. (2003) Controlling false negative errors in microarray differential expression analysis: a PRIM approach. *Bioinformatics*, **19** 1808 –1816.
- [4] Friedman, J. H. and Fisher, N. I. (1999) Bump hunting in high-dimensional data. *Statist. & Comp.* **9** 123–143.
- [5] Hastie, T., Tibshirani, R. and Friedman, H.H. (2001) *The elements of statistical learning*. Springer, New York
- [6] Leblanc, M., Jacobson J. and Crowley J., (2002) Partitioning and peeling for constructing prognostic groups. *Statistical Methods for Medical Research*, **11** 247–274.
- [7] Polonik, W. (1997) Minimum volume sets and generalized quantile processes. *Stoch. Proc. and Appl.* **69** 1 - 24.
- [8] van der Vaart, A.W. (1998): *Asymptotic Statistics*. Cambridge Univ. Press.
- [9] van der Vaart, A.W. & Wellner, J. (1996): *Weak convergence and empirical processes*. Springer, New York.

Figure 1: Comparison of theoretical solutions of (2.1) (right, nested squares) and true level sets (right, nested circles) for uni-modal two-dimensional regression curve. Left plot is the regression curve.



- [10] Wang, P., Kim, Y., Pollack, J. and Tibshirani, R. (2004) Boosted PRIM with application to searching for oncogenic pathway of lung cancer. *In: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CBS 2004)*, 604-609.
- [11] Wu, L. and Chipman, H. (2003) Bayesian model-assisted PRIM algorithm. Technical Report

Figure 2: Comparison of optimal solution of (2.1) with peeling/pasting/jittering results for (3.11) by parameters $\alpha=.005$ and $\beta_0=.12$. The optimal intervals are (local) level sets and all the intervals shown in the plots are drawn at the same height corresponding to the level of the level sets. Plot 1: Optimal intervals of (2.1) are $[.24, .36]$ and $[.64, .76]$. Plot 2: Peeling result is $[.2993, .4199]$ with $\beta_0 = .1206$. Plot 3: Peeling + Pasting result is $[.2506, .4199]$ with $\beta_0 = .1693$. Plot 4: Peeling + Jittering result is $[.2396, .3602]$ with $\beta = 0.1206$.

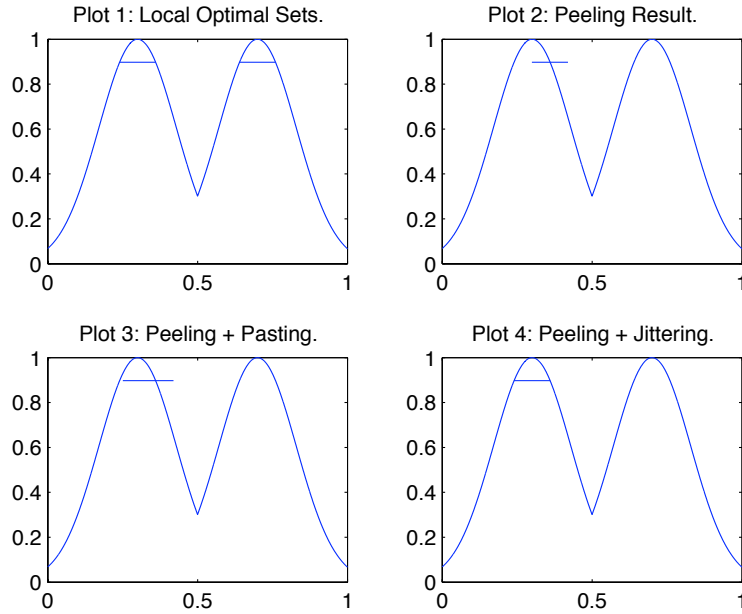
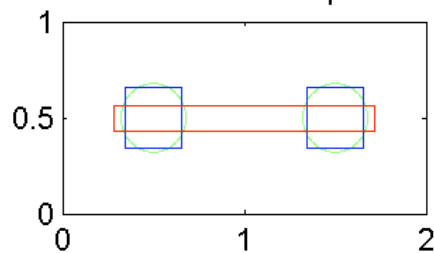
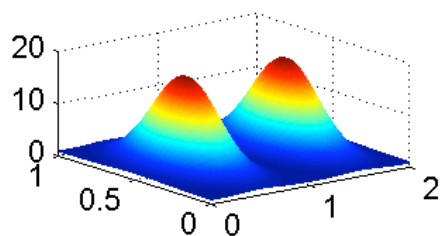
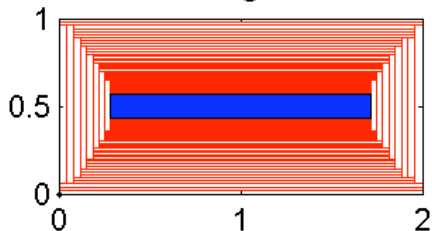


Figure 3: Plot 1 shows the model (symmetric bimodal curve). Plot 2: the level sets (circles) and optimal boxes (squares and rectangles) for $\beta_0 = .1$. Plot 3: indicates the (first) peeling procedure and the outcomes. Plot 4: the covering results for the first 6 peeled boxes. The parameters are $\alpha = .02$ and $\beta_0 = .1$.

Plot 1: Symmetric Bimodal Curve. Plot 2: Level Sets and Optimal Sets.



Plot 3: Peeling Procedure.



Plot 4: Covering Results.

