PRIM ANALYSIS

Wolfgang Polonik
Department of Statistics
University of California
One Shields Ave.
Davis, CA 95616-8705

Zailong Wang
Mathematical Biosciences Institute
The Ohio State University
231 West 18th Avenue, #216
Columbus, OH 43210-1174

July 24, 2007

last modified: July 31, 2009

Abstract

This paper analyzes 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 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, Nannings et al. (2008), 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 remedies aimed at fixing these shortcomings. The methodology developed here should be useful in studying the proposed 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 $m(x) := E\,[\,Y \mid X = x\,]$, $x \in \mathbb{R}^d$. Without loss of generality we assume throughout the paper that $m(x) \geq 0$. Assume that F has support $[0,1]^d \subset \mathbb{R}^d$ also called the input space. Put

$$I(C) := \int_C m(x)\,\mathrm{d}F(x) \quad \text{and} \quad F(C) := \int_C \mathrm{d}F(x), \qquad C \subset [0,1]^d.$$

The objective of PRIM is to find a subregion $C \subset [0,1]^d$ for which

$$ave(C) = \frac{I(C)}{F(C)} > \lambda, \tag{1.1}$$

where λ is a pre-specified threshold value. Property (1.1) is equivalent to

$$I(C) - \lambda\, F(C) = \int_C (m(x) - \lambda)\,\mathrm{d}F(x) \geq 0. \tag{1.2}$$

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 can 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. Nevertheless, as will become clear below, the PRIM algorithm does in general *not* result in an estimate of the level set $C(\lambda)$.

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 $ave(A) > \lambda$ and each subset A of $[0,1]^d \setminus C(\lambda)$ satisfies $ave(A) \leq \lambda$. Hence, as an idea for an algorithm to approximate level sets, one might think about iteratively finding 'small' (disjoint) subsets B_k satisfying $ave(B_k) > \lambda$, 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) $B_1^*, \ldots, B_K^*$, each time removing the outcome B_{k-1}^* of the preceding step(s) and applying the algorithm to the remaining space $S^{(k)} = [0,1]^d \setminus \bigcup_{j=1}^{k-1} B_j^*$, resulting in a partition of $[0,1]^d$. The optimal outcomes satisfy

$$B_k^* \in \underset{F(B \mid S^{(k)}) = \beta_0}{\operatorname{argmax}} \; ave\big(B \cap S^{(k)} \big), \; k = 1, \ldots, K, \tag{1.3}$$

where β_0 is a (small) tuning parameter to be chosen, and $F(\cdot|A)$ denotes the conditional distribution of F given A. The final outcome, R_λ^*, consists of the union of those sets $B_k^* \cap S^{(k)}$ with $ave(B_k^* \cap S^{(k)})$ exceeding λ. (More details on PRIM are given below.)

However, this procedure does *not* lead to approximations of level sets in general. The reason for PRIM not fitting the intuitive an natural conceptual idea laid out above is that the individual sets B_j^*, even though their (conditional) F- measure are all small (equal to β_0), are not really 'small' in the sense of 'local'. This can be seen in Figure 1.

PUT FIGURE 1 HERE

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

One can hope, however, that at least certain features of the level sets are captured by the PRIM outcome. For instance, if the underlying distribution has two modes, then one should hope for PRIM outcomes reflecting the location of the two modes, i.e. for an appropriate threshold λ the outcome should consist of two disjoint sets, each located around one of the two modes. As we will see below, even this not guaranteed. Also characterization of the possible PRIM outcomes is provided in this paper.

Besides providing such more conceptual insight into PRIM (for instance, characterizing the outcomes of the PRIM algorithm), this paper derives theoretical results. These results concern rates of convergence of the outcome regions of empirical PRIM to their theoretical counterparts for a given β_0. For instance, letting $\widehat{R}_\lambda$ denote the empirical counterpart to R^*_λ from above, we will derive conditions under which the following holds:

Suppose that $E|Y|^\gamma < \infty$ for some $\gamma \geq 3$. Let $0 < \beta_0 < 1$ and λ be fixed. Choose the peeling parameter $\alpha = \alpha_n = (\frac{d}{n})^{\frac{1}{3}} \log n$. Then, under additional assumptions (cf. Theorem 5.3) there exists an R^*_λ such that

$$d_F(\widehat{R}_\lambda, R^*_\lambda) = O_P\big(\,\big(\tfrac{d^4}{n}\big)^{1/3}\log n\,\big). \tag{1.4}$$

Here $d_F(A, B)$ denotes the F-measure of the set-theoretic difference of A and B (cf. (5.1)). 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 by their definition the sets $\widehat{B}_k$ are closely related to so-called minimum volume sets. For fixed d, rates of convergence of the order $(\frac{d}{n})^{-1/3}$ times a log-term have been derived for d-dimensional minimum volume ellipsoids and other minimum volume sets in so-called Vapnik-Cervonenkis (VC) classes (see Polonik, 1997, and references therein). Since boxes (or rectangles) in $\mathbb{R}^d$ form a VC-class, the above rates seem plausible.

Section 3 explores the outcomes of peeling+jittering, thereby also discussing some shortcomings of PRIM indicated above. Before that, PRIM is described n some more detail (Section 2). This is necessary to understand the discussions in this paper as well as the derivations of the theoretical results, which are presented and proved in Section 5. 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. Section 4 presents a small simulation study, comparing the original PRIM algorithms with its modifications suggested in this manuscript. Proofs of some miscellaneous technical results related to empirical process theory can be found in Section 6. 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. Given a rectangle B, a peeling step successively peels of small strips along the boundaries of B. The peeling procedure stops if the box becomes too small. More precisely, let the class of all closed d-dimensional boxes, or axis parallel rectangles $B \subset [0,1]^d$ be denoted by $\mathcal{B}$. Given a subset $S \subseteq [0,1]^d$ and a value β_0, the goal of peeling is to find

$$B^*_{\beta_0} = \arg\max_{B \subset \mathcal{B}} \{\, ave(B|S) : \, F(B|S) = \beta_0 \,\}, \tag{2.1}$$

where $F(\cdot|S)$ denotes the conditional distribution of X given $X \in S$, $ave(B|S) = \frac{I(B \cap S)}{F(B \cap S)}$, and $\beta_0 \in [0,1]$ is a tuning parameter to be considered fixed in this paper. We always assume that such a set $B^*_{\beta_0}$ exists. Beginning with $B = S = [0,1]^d$ 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 $b_{j1} := \{x \in B : x_j < x_{j(\alpha)}\}, b_{j2} := \{x \in B : x_j > x_{j(1-\alpha)}\}$, $j = 1, \ldots, d$, where $0 < \alpha < 1$ is a second 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 $X_j \in B \cap S$. By construction, $\alpha = F_j(b_{jk}|B \cap S) = F(b_{jk}|B \cap S)$. The particular subbox b^* chosen for removal is the one that yields the largest target value among $B \setminus b_j$, $j = 1, \ldots, d$, i.e. $b^* = \text{argmin}\{I(b_{jk}|S),\, b_{jk},\, j = 1, \ldots d,\, k = 1, 2\}$. The current box is then updated (shrunk), 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 kth-step the candidate boxes b for removal all 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 chosen too small is quantified 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 outcome the current box is enlarged by pasting along its boundary 'small' strips $b \subset S$. The (at most) $2d$ candidate sets b are boxes alongside the $2d$ boundaries of the current box $B \cap S$ of size $F(b|S) = \alpha \times F(B|S)$. This is done as long as the average increases, i.e. as long as there exists a candidate set b with $ave((B \cup b) \cap S) > ave(B \cap S)$.

Covering. The covering procedure leads to the final output region R^* of the PRIM algorithm as a union of boxes from iterative applications of the peeling+pasting procedure, each time removing the previous outcome, and thus each time changing the input space S for the peeling+pasting procedure. More precisely, the first box B_1^* is constructed via peeling+pasting on $S = [0,1]^d$ as described above. The second optimal box B_2^* is constructed in the same fashion by replacing $S = S^{(1)} = [0,1]^d$ by $S^{(2)} = [0,1]^d \setminus B_1^*$, and so on, each time removing the optimal outcome of the previous step. The hope now is (and as indicated above, in general this is not true) that if the outcome B_k^* of the k-th iterative application of the peeling+pasting procedure is such that its average exceeds a pre-specified λ, then it is a subset of $C(\lambda)$. Thus the final result of the PRIM algorithm is

$$R_\lambda = \bigcup_{ave(B_k^* \cap S^{(k)}) \geqslant \lambda} \big(B_k^* \cap S^{(k)}\big). \tag{2.2}$$

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 after the application of one peeling+pasting procedure. We therefore 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. In fact, since pairs of boxes have to found (and there are of the order d^2 many such pairs, the complexity is increased by a factor of d. (Also the constants in the complexity will increase.)

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 $I(C) = E\{Y\,\mathbf{1}\{X \in C\}\}$. Hence, if $(X_i, Y_i), 1 \leqslant i \leqslant n$, is an independent sample with the same distribution as (X, Y), the empirical analog of I is given by

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

The empirical analog to F is given by F_n, the empirical distribution of $X_1, ..., X_n$, and we denote

$$ave_n(A) = \frac{I_n(A)}{F_n(A)}.$$

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, replacing $\alpha = \alpha_n$ by $\lceil n\alpha_n \rceil / n$, the smallest k/n, $k = 1, 2, \ldots$ which is larger than or equal to α_n.

3 PRIM Outcomes

Here we provide a characterization of PRIM outcomes along with some discussions and examples.

Local maximizers. For a box $B = \bigotimes_{j=1}^{d}[a_{j1}, a_{j2}] \in \mathcal{B}$ consider two bracketing sets $\overline{B} = \bigotimes_{j=1}^{d}[\overline{a}_{j1}, \overline{a}_{j2}] \subset [0,1]^d$, $\underline{B} = \bigotimes_{j=1}^{d}[\underline{a}_{j1}, \underline{a}_{j2}] \subset [0,1]^d$ with $\underline{B} \subseteq B \subseteq \overline{B}$, and assume that

$$|\overline{a}_{jk} - \underline{a}_{jk}| > \epsilon \ \text{ for at least two distinct pairs } (jk),\ 1 \le j \le d,\ k = 1, 2. \tag{3.1}$$

Here we need the 'at least two' (rather than 'at least one') in (3.1) because otherwise we in general would not have other boxes $\widetilde{B}$ of the same size as B in the neighborhood, and (3.3) below would not be useful. Based on such bracketing sets for B, define a neighborhood of B as

$$U(\epsilon, B) := \{\widetilde{B} : \underline{B} \subset \widetilde{B} \subset \overline{B}\}. \tag{3.2}$$

With this type of neighborhood we now define local maximizers $B^*_{\beta_0}$ consisting of sets of size β_0 such that there exists a neighborhood $U(\epsilon, B^*)$ with B^* maximizing the average among all the boxes in this neighborhood:

Definition 3.1 *The class $\mathcal{M}_{\ell oc}(\beta_0)$ consists of all boxes satisfying*

$$\exists\, \epsilon > 0 : \quad B^*_{\beta_0} \in \arg\max\{ave(B \cap S);\ F(B|S) = \beta_0,\ B \in U(\epsilon, B^*).\}. \tag{3.3}$$

For a box $B \subset [0,1]^d$, $d \geq 2$, $1 \leq j \leq d$, and $t \in [0,1]$ let

$$F_j(\,t, B) = \int_{(B \cap S)_t^{\widehat{j}}} f(\ldots, x_{j-1},\, t\,, x_{j+1}, \ldots)\, d\underline{x}_{\widehat{j}}, \tag{3.4}$$

$$I_j(\,t\,, B) = \int_{(B \cap S)_t^{\widehat{j}}} (m \cdot f)(\ldots, x_{j-1}, t, x_{j+1}, \ldots)\, d\underline{x}_{\widehat{j}}, \tag{3.5}$$

where $\underline{x}_{\widehat{j}} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)'$, and for any set $A \subset \mathbb{R}^d$ we let $A_t^{\widehat{j}} = \{(x_1, \ldots, x_{j-1},\, t,\, x_{j+1}, \ldots, x_d)' : (x_1, \ldots, x_d)' \in A\}$ denote the slice through A at the j-th coordinate being equal to t. Further, writing $B = \bigotimes_{i=1}^d [a_{i1},\, a_{i2}]$, let

$$F_{jk}(\partial B) = F_j(a_{jk}, B) \quad \text{and} \quad I_{jk}(\partial B) = I_j(a_{jk}, B), \quad j = 1, \ldots, d,\ k = 1, 2.$$

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

Further denote

$$ave_j(x, B) = \frac{I_j(\,x\,, B)}{F_j(\,x\,, B)}, \qquad \text{for } F_j(x, B) > 0. \tag{3.6}$$

We also use the notation

$$\partial B_{jk} = \bigotimes_{i=1}^{j-1} [a_{i1},\, a_{i2}] \times a_{jk} \times \bigotimes_{i=j+1}^{d} [a_{i1},\, a_{i2}], \qquad j = 1, \ldots, d,\ k = 1, 2, \tag{3.7}$$

for the (jk)-th boundary facet of B. We sometimes use the notation '(jk)' rather than ∂B_{jk}. In case $x = a_{jk}$, the averages in (3.6) become boundary averages, which play a special role here. Therefore we also use the notation:

$$A_{jk}^{\pm}(\partial B) := ave_j(a_{jk}^{\pm}, B), \tag{3.8}$$

where $ave_j(a_{jk}^{+}, B)$ and $ave_j(a_{jk}^{-}, B)$ denote the limits of $ave_j(x, B)$ 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 $j = 1, 2,\ k = 1, \ldots, d$.

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 of the current box during a peeling procedure as close as possible. This motivates the importance of the following function:

$$\Psi(B) := \max_{(jk)} \big[\max_{(\ell m)\neq(jk)} \big(A^{+}_{\ell m}(\partial B) \big) - A^{-}_{jk}(\partial B) \big] \tag{3.9}$$

where for any (jk) the max inside $[\cdots]$ 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 we define for $0 \leq \beta_0 \leq 1$

$$\mathcal{N}_{\ell oc}(\beta_0) = \big\{ B \in \mathcal{B} : F(B|S) = \beta_0 \text{ and } \Psi(B) \leq 0 \big\}. \tag{3.10}$$

Observe that if $ave_j(x, B)$ 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, $\mathcal{N}_{\ell oc}(\beta_0)$ does in general not only contain local maximizers, but also minimizers and 'saddle-points'. In order to define a class of typical local minimizers or saddle points we introduce the properties **(m.i)** and **(m.ii)** below. For this we first need some more notation.

Definition 3.2 *Let B be such that for $\epsilon > \frac{c}{d}$ for some $c > 0$ there exists a neighborhood $U(\epsilon, B)$ as defined in (3.2), and let $(j_1, k_1) \neq (j_2, k_2)$ with $1 \leq j_1, j_2, \leq d$ and $k_1, k_2 \in \{1, 2\}$ denote the two pairs from the definition of $U(\epsilon, B)$. For $(j, k) = (j_1, k_1), (j_2, k_2)$ we define the properties:*

(m.i) *The function $ave_j(\cdot, B)$ is strictly decreasing in $[\overline{a}_{jk}, \underline{a}_{jk}]$ for $k = 1$ (i.e. 'on the left') and increasing for $k = 2$ (i.e.'on the right').*

(m.ii) *For some constant $k_1 > 0$ not depending on B we have*

$$k_1 \, |x_1 - x_2| \leq \big| ave_j(x_1, B) - ave_j(x_2, B) \big| \quad \text{for } x_1, x_2 \in [\overline{a}_{jk}, \underline{a}_{jk}].$$

With these properties let

$$m_{\ell oc}(\beta_0) = \{ B^{*} \in \mathcal{N}_{\ell oc}(\beta_0) \text{ such that (m.i) and (m.ii).} \}. \tag{3.11}$$

The crucial assumption is (m.i). It is obvious that (m.ii) is not necessary for a set being a local minimum or saddle point. It is included here for technical reasons.

The following lemma says that under certain assumptions, PRIM outcomes do not contain local minima or 'saddle points', i.e. sets in $m_{\ell oc}(\beta_0)$.

Lemma 3.1 *Suppose that assumptions* **(A2)** *and* **(A3)** *hold (cf. Section 5). Then for every* $0 < \beta_0 < 1$ *we have*

$$\mathcal{M}_{\ell oc}(\beta_0) \subset \mathcal{N}_{\ell oc}(\beta_0) \setminus m_{\ell oc}(\beta_0).$$

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

$$m(x) = \begin{cases} \exp(-30(x-0.3)^2), & 0 \leqslant x \leqslant 0.5; \\ \exp(-30(x-0.7)^2), & 0.5 < x \leqslant 1. \end{cases} \tag{3.12}$$

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 $\alpha \to 0$) 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 $\alpha \to 0$ (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 separate modes will be recovered.

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.

We want candidate boxes to locate around a single mode. Therefore, one should check whether the conditional distribution of the data falling inside the $2d$ candidate boxes for peeling is unimodal (with decreasing or increasing being special cases). Implementing relevant tests (e.g. Burman and Polonik, 2009) is rather time consuming. A (very) simple shortcut is the following. For a box $B = \bigotimes_{j=1}^{d}[a_{j1}, a_{j2}]$, let $\ell_i = |a_{j2} - a_{j1}|$, $i = 1, \ldots, d$ 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 small simulation study indicated that for $d = 2$ a reasonable choice appears to be $1.5 \leq r \leq 2.0$ as long as one is dealing with mixtures of normals.

4 Simulation study

A small simulation study is presented here for $d = 2$. The two cases considered both are mixtures of two normals with identity covariance matrix and means $(-1, 0)$ and $(0, 1)$ (Model I), and $(-1, -1)$ and $(1, 1)$, respectively (Model II). The tables below show simulated mean and variance of d_{F_n} between the (population) peeling outcomes by using ratio control and no ration control, respectively, and the empirical PRIM with $\beta_0 = 0.05$. We have been using the peeling outcome as comparison because all the four different procedures estimate different targets. The only procedure that estimates the local maximizer with $\beta_0 = 0.05$ is PRIM with jittering and ratio-control in Model I, and PRIM with jittering and either ratio control or no ratio control in Model II.

The simulations shows that in regular cases (mixture of normals) and for $d = 2$, both jittering and pasting seem to behave similar if measured by the variance. As has been discussed above, the target sets are different, however. One can also see no significant differences in terms of variation between ratio control and no ratio control. Notice again, however, that in case 1 the target sets under ratio control and non ration control are different. We have been using the distance to the peeling outcomes (without pasting or jittering). These sets are shown in Figures 4 and 5, respectively.

While this is not surprising that all these procedures behave comparably in the considered regular case, (with different target sets, however) more significant differences might be expected in less regular situations, and such situations are difficult to simulate. An application to real data will certainly lead to differing outcomes, and only a thorough study of the outcomes could reveal whether they are ‘more reasonable’. Such an analysis goes beyond the scope of the manuscript.

Table 1: Mean and variance of d_{F_n} between empirical PRIM outcome and population version of peeling under Model I with $\alpha = 0.01$ based on 1000 simulations.

Sample size	no ratio control				ratio control			
	pasting		jittering		pasting		jittering	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
100	0.0730	0.0325	0.0542	0.0297	0.0915	0.0243	0.0923	0.0238
125	0.0672	0.0315	0.0534	0.0293	0.0908	0.0237	0.0917	0.0225
150	0.0651	0.0312	0.0546	0.0298	0.0904	0.0212	0.0920	0.0200
200	0.0625	0.0320	0.0529	0.0306	0.0881	0.0200	0.0910	0.0195

PUT FIGURE 4 HERE.

Table 2: Mean and variance of d_{F_n} between empirical PRIM outcome and population version of peeling under Model II with $\alpha = 0.01$ based on 1000 simulations

Sample size	no ratio control				ratio control			
	pasting		jittering		pasting		jittering	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
100	0.0841	0.0310	0.0827	0.0305	0.0825	0.0307	0.0852	0.0324
125	0.0806	0.0296	0.0800	0.0297	0.0841	0.0349	0.0820	0.0322
150	0.0773	0.0311	0.0764	0.0303	0.0819	0.0325	0.0794	0.0316
200	0.0743	0.0298	0.0731	0.0286	0.0730	0.0300	0.0770	0.0301

PUT FIGURE 5 HERE.

5 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 $\widetilde{B} = \bigotimes_{j=1}^{d}[\widetilde{a}_{j1},\, \widetilde{a}_{j2}]$ let

$$d_F(B, \widetilde{B}) := F(B \,\Delta\, \widetilde{B}) = F(B \setminus \widetilde{B}) + F(\widetilde{B} \setminus B), \tag{5.1}$$

the F-measure of the set-theoretic symmetric difference between B and $\widetilde{B}$, and let

$$\rho_\infty(B, \widetilde{B}) := \max_{j=1,\ldots,d,\; k=1,2} |\, a_{jk} - \widetilde{a}_{jk}\,|. \tag{5.2}$$

Remark: Notice that boxes are always defined on $[0,1]^d$. However, if we consider $S = [0,1]^d \setminus R$ with R being a union of (non-empty) boxes, then there exist boxes with at least one boundary facet of B lying completely inside R, e.g. let $S = [0,1]^2 \setminus [0,1/2]^2$ and $B = [1/4, 3/4] \times [0, 1/4]$. In such cases it shall be understood that we are working with the smallest of such boxes coinciding with B on S. In the above example this means that we work with $[1/2, 3/4] \times [0, 1/4]$ instead of $B = [1/4, 3/4] \times [0, 1/4]$. Observe that this does not change the conditional quantities $F(\cdot|S)$ and $I(\cdot|S)$. This convention is to be understood in all what follows without further mention, and it is not reflected in the notation (to not further add to the notation). Obviously, this is non-issue for $S = [0,1]^d$.

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

$$F_{j,\ell}(\, s, t\,, B) = \int_{(B \cap S)^{\widehat{j},\widehat{\ell}}_{s,t}} f(\ldots, x_{j-1},\, s,\, x_{j+1}\, \ldots, x_{\ell-1},\, t,\, x_{\ell+1}\, \ldots)\, d\underline{x}^{\widehat{j},\widehat{\ell}} \tag{5.3}$$

$$I_{j,\ell}(\, s, t\,, B) = \int_{(B \cap S)^{\widehat{j},\widehat{\ell}}_{s,t}} (m \cdot f)(\ldots, x_{j-1},\, s,\, x_{j+1}\, \ldots, x_{\ell-1},\, t,\, x_{\ell+1}\, \ldots)\, d\underline{x}^{\widehat{j},\widehat{\ell}} \tag{5.4}$$

where $\underline{x}^{\widehat{j},\widehat{\ell}} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{\ell-1}, x_{\ell+1}, \ldots, x_d)'$, and for a set $A \subset \mathbb{R}^d$, $d \geq 3$, $0 \leq j < k \leq d$, and $0 \leq s, t \leq 1$ we let

$$A_{s,t}^{\widehat{j},\widehat{\ell}} = \{(x_1, \ldots, x_{j-1},\, s,\, x_{j+1}, \ldots, x_{\ell-1},\, t,\, x_{\ell+1}, \ldots, x_d)' :\, (x_1, \ldots, x_d)' \in A\,\}.$$

Further we denote

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

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 $I_{jk,\ell m}(\partial B) = (mf)(a_{jk}, a_{\ell m})\, \mathbf{1}_S(a_{jk}, a_{\ell m})$.

Assumptions. Let $0 < \beta_0 < 1$. For a function $h : [0,1] \to \mathbb{R}$ let $\|h\|_\infty = \sup_{t\in[0,1]} |h(t)|$, and for $\epsilon > 0$ let $U_F(B, \epsilon) = \{B' \subset S : d_F(B, B') < \epsilon\}$.

(A1) There exists a constant $c_1 > 0$ such that for $\delta > 0$ small enough we have

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

(A2) There exists $\epsilon_0, K_0 > 0$ such that for all B with $F(B|S) \geq \beta_0/2$ we have

$$0 < \epsilon_0 \leq F_j(t, B) \leq K_0 \quad \text{for all } t \text{ with } \int_{(B\cap S)_t^{\widehat{j}}} dx^{\widehat{j}} > 0,\ 1 \leq j \leq d,\ d \geq 1.$$

(A3) The function m is bounded (uniformly in d).

(A4) There exists a $K_1 < \infty$ such that for all $1 \leq j, \ell \leq d$ and all d we have

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

(A5) For each $1 \leq j \leq d$ there exists $c_{j1} < \cdots < c_{jL}$, with $L \in \mathbb{N}$ not depending on d, such that the functions $ave_j(\cdot, B)$ are Lipschitz continuous in $[c_{jk}, c_{j,k+1}]$, $k = 1, \ldots, L-1$, uniformly in $B \in U_F(B', \epsilon)$ for some B' with $F(B'|S) \geq \beta_0/2$, and Lipschitz constant not depending on d.

(A6) There exists a $\delta > 0$ and a constant $c > 0$ such that for all $B \notin \mathcal{N}_{\ell oc}(F(B|S))$ with $|\, F(B|S) - \beta_0 \,| < \delta$ we have

$$\Psi(B) \geq c \inf_{B^* \in \mathcal{N}_{\ell oc}(F(B|S))} \rho_\infty(B,\, B^*). \tag{5.5}$$

Discussion of the assumptions. Assumption **(A1)** 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 **(A1)** 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 lcoal 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). If follows that **(A1)** holds. Assumptions **(A2)** - **(A4)** imply in particular that boundary averages for not too small boxes are uniformly bounded, and **(A5)** assumes their Lipschitz continuity.

Assumption **(A6)** 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 $\Psi(B^*_\beta) \leq 0$). It is reminiscent of a margin condition introduced in Polonik (1995), in the context of level set estimation that became popular when used in Tsybakov (2004) in the context of classification. This condition controls the behavior of the target function locally around the level curve, and condition **(A6)** is of a similar flavor. It holds under the scenario given in the discussion of **(A1)** above, provided, for instance, m is differentiable with partial derivatives bounded away from zero.

5.1 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 $\mathcal{N}_{\ell oc}(\beta_0) \setminus m_{\ell oc}(\beta_0)$ (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 5.1 *Let $0 < \beta_0 < 1$, and let $\alpha = \alpha_n$ be such that $d\,\alpha_n = o(1)$ as $n \to \infty$. Suppose that* **(A1)** - **(A6)** *hold for $S = [0,1]^d$. Let $\widetilde{B}$ denote the result of the population version of peeling + jittering. Then as $n \to \infty$ we have*

$$\inf_{B^* \in \mathcal{N}_{\ell oc}(\beta_0) \setminus m_{\ell oc}(\beta_0)} d_F(B^*, \, \widetilde{B}) = O\big(\, d\,\alpha_n \,\big). \tag{5.6}$$

It is straightforward to derive the proof of this theorem from the proof of Theorem 5.2 presented next. Both proofs have the same basic structure. Involving no stochastic elements, the proof of Theorem 5.1 is simpler and hence omitted.

5.2 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 $\widetilde{B}$, and it also shows 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.

Theorem 5.2 *Suppose that $E|Y|^\gamma < \infty$ for some $\gamma \geq 2$. Let $\beta_0 > 0$ be fixed, and assume that* **(A1)** - **(A6)** *hold for $S = [0,1]^d$. Further suppose that*

$$\big(\tfrac{d}{n}\,\log(d\,n)\,\big)^{\min(\frac{1}{3}, \frac{\gamma-1}{\gamma+3})} = o(\alpha_n), \tag{5.7}$$

and that $d\,\alpha_n = o(1)$. Then we have

$$\inf_{B^* \in \mathcal{N}_{\ell oc}(\beta_0) \setminus m_{\ell oc}(\beta_0)} d_F(B^*, \widehat{B}) = O_P(\, d\,\alpha_n). \tag{5.8}$$

Remarks. (i) The rates in Theorem 5.2 are in alignment with the rate of the population version given in Theorem 5.1. (ii) For a motivation of the rates of convergence based on known results for the estimation of minimum volume sets see introduction.

Proof of Theorem 5.2: We prove that

$$\inf_{B^* \in \mathcal{N}_{\ell oc}(\beta_0) \setminus m_{\ell oc}(\beta_0)} \rho_\infty(B^*, \widehat{B}) = O_P(\, \alpha_n). \tag{5.9}$$

The rate for $d_F(B^* \,\Delta\, \widehat{B})$ asserted in (5.8) follow from (5.9) because $d_F(B^*,\, \widehat{B}) = O(\, d\, \rho_\infty(B^*,\, \widehat{B})\,)$. in (5.8) follow from (5.9) 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 5.1. For $C_1 > 0$ let

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

It follows from Proposition 6.1 and Proposition 6.2 that for each $\epsilon > 0$ we can choose $C_1 > 1$ such that for n large enough we have $P(A_n) \ge 1 - \epsilon$. We prove the theorem in two steps. First we show that for C large enough we have

$$P(\inf_{B^* \in \mathcal{N}_{\ell oc}(\beta_0)} \rho_\infty(B^*,\, \widehat{B}) > 2\, C\, \alpha_n;\; A_n\,) = 0. \tag{5.10}$$

To complete the proof we then show that for all $c > 0$ we have for n large enough that

$$P(\inf_{B^* \in m_{\ell oc}(\beta_0)} \rho_\infty(B^*,\, \widehat{B}) < c\, \alpha_n;\; A_n\,) = 0. \tag{5.11}$$

Let $\widehat{\beta}_n := F(\widehat{B})$. By assumption, $\sqrt{\frac{d}{n}} = o(\alpha_n)$, and thus we obtain that on A_n for n large enough

$$\begin{aligned} \beta_0 \le \widehat{\beta}_n = F_n(\widehat{B}) - (F_n - F)(\widehat{B}) &< (\beta_0 + \alpha_n) + C_1 \alpha_n \\ &\le \beta_0 + (C_1 + 1)\, \alpha_n. \end{aligned} \tag{5.12}$$

Using **(A1)** we obtain that for all $B^*_{\widehat{\beta}_n} \in \mathcal{N}_{\ell oc}(\widehat{\beta}_n)$ there exists a $B^* \in \mathcal{N}_{\ell oc}(\beta_0)$ such that on A_n we have

$$\rho_\infty(B^*_{\widehat{\beta}_n},\, B^*) \le\; c_1\, |\beta_n - \beta_0| \le c_1\, (C_1 + 1)\, \alpha_n.$$

Thus, if $\widehat{B} \in \mathcal{N}_{\ell oc}(\widehat{\beta}_n)$ then the assertion follows. We therefore assume from now on that $\widehat{B} \notin \mathcal{N}_{\ell oc}(\widehat{\beta}_n)$. Suppose that for all $B^*_{\beta_0} \in \mathcal{N}_{\ell oc}(\beta_0)$ we have $d_F(B^*_{\beta_0}, \widehat{B}) > 2\,C\,d\,\alpha_n$ for some $C > 0$. We obtain by using triangular inequality that on A_n and for C large enough

$$\rho_\infty(B^*_{\widehat{\beta}_n}, \widehat{B}) \geq 2\,C\,\alpha_n - \ c_1\,(C_1+1)\,\alpha_n\ \geq C\,\alpha_n \quad \text{for all } B^*_{\widehat{\beta}_n} \in \mathcal{N}_{\ell oc}(\widehat{\beta}_n).$$

In other words, for C large enough

$$\{\inf_{B^* \in \mathcal{N}_{\ell oc}(\beta_0)} \rho_\infty(B^*, \widehat{B}\,) \geq 2\,C\,\alpha_n,\ A_n\,\}$$

$$\subset \{\inf_{B^*_{\widehat{\beta}_n} \in \mathcal{N}_{\ell oc}(\widehat{\beta}_n)} \rho_\infty(B^*_{\widehat{\beta}_n}, \widehat{B}\,) \geq C\,\alpha_n,\ A_n\,\} \tag{5.13}$$

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 **(A6)** we have $\Psi(\widehat{B}) > c\,C\,\alpha_n$. It follows that there exist two boundary facets of $\widehat{B}$ indexed by, let's say, (jk) and (ℓm), respectively, with

$$A^+_{jk}(\partial\widehat{B}) \geq A^-_{\ell m}(\partial\widehat{B}) + c\,C\,\alpha_n. \tag{5.14}$$

Let b_{jk} and $b_{\ell m}$ be the two candidate sets for jittering. We almost surely have $F_n(b_{jk} \setminus b_{\ell m}) = F_n(b_{\ell m}) = \alpha_n$, with b_{jk} being outside the box $\widehat{B}$ and $b_{\ell m}$ inside. We will show that for C large enough,

$$I_n(b_{jk} \setminus b_{\ell m}) > I_n(b_{\ell m}) \tag{5.15}$$

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, $\widehat{B}$ cannot be an outcome of the peeling+jittering procedure. A contradiction that verifies (5.10).

In order to show (5.15), we first consider the non-overlapping case and show that $I(b_{jk}) - I(b_{\ell m}) > K\alpha_n^2 > 0$ for some $K > 0$. In a second step we will then show the same with I replaced by I_n (which is (5.15)). Finally we will address the case with an overlap between b_{jk} and $b_{\ell m}$.

For any (jk) let h_{jk} denote the width of b_{jk} in dimension j. Rewrite

$$\begin{aligned} I(b_{jk}) &= \int_{a_{jk}-h_{jk}}^{a_{jk}} ave_j(x, b_{jk}) F_j(x, b_{jk} \cup \widehat{B})\, dx \\ &= F(b_{jk})\, ave_j(a_{jk}^+, \widehat{B}) \;+\; \int_{a_{jk}-h_{jk}}^{a_{jk}} (\, ave_j(x, b_{jk}) - ave_j(a_{jk}^+, \widehat{B})\,)\, F_j(x, b_{jk} \cup \widehat{B})\, dx \\ &=: ave_j(a_{jk}^+, \widehat{B})\, F(b_{jk}) + r_{jk}. \end{aligned} \tag{5.16}$$

Similarly, we can rewrite $I(b_{\ell m})$ as

$$\begin{aligned} &I(b_{\ell m}) \\ &= F(b_{\ell m})\, ave_\ell(a_{\ell m}^-, \widehat{B}) + \int_{a_{\ell m}-h_{\ell m}}^{a_{\ell m}} (\, ave_\ell(x, b_{\ell m}) - ave_\ell(a_{\ell m}^-, \widehat{B})\,)\, F_j(x, b_{jk} \cup \widehat{B})\, dx \\ &=: ave_\ell(a_{\ell m}^-, \widehat{B})\, F(b_{\ell m}) + r_{\ell m}. \end{aligned} \tag{5.17}$$

Using (5.14) we obtain that on A_n we have

$$\begin{aligned} I(b_{jk}) - I(b_{\ell m}) &\geq c\, C\, \alpha_n\, F(b_{jk}) + ave_\ell(a_{\ell m}^-, \widehat{B})\ (F(b_{jk}) - F(b_{\ell m})) + (r_{jk} - r_{\ell m}) \\ &=: (I) + (II) + (III). \end{aligned}$$

We will show that on A_n (and for n large enough) that $(I) \geq \frac{cC\beta_0}{2}\alpha_n^2$, $(II) = o(\alpha_n^2)$ and $|(III)| \leq K\, \alpha_n^2$ for a constant $K > 0$. Since C can be chosen large enough, this then gives

$$I(b_{jk}) - I(b_{\ell m}) \geq \frac{c\, C\beta_0}{4}\alpha_n^2. \tag{5.18}$$

First we show the asserted lower bound for (I). To see this observe that on A_n for large enough n we have almost surely

$$\frac{\beta_0}{2}\ \alpha_n \leq F(b_{\ell m}) \leq 2\ \alpha_n. \tag{5.19}$$

To see this observe that by construction of the PRIM algorithm we have $\alpha_n\, \beta_0 \leq F_n(b_{\ell m}) \leq \alpha_n + \frac{1}{n}$ almost surely (where the $\frac{1}{n}$ comes from the fact that for the empirical PRIM algorithm we peel off a fraction of $\frac{k}{n}$ with k the smallest integer with $\alpha_n \leq \frac{k}{n}$). 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 the asserted inequality for (I).

In order to see that $(II) = o(\alpha_n^2)$ observe that on A_n

$$\begin{aligned} |\,F(b_{jk}) - F(b_{\ell m})\,| &=_{a.s.} |\,(F_n - F)(b_{jk}) - (F_n - F)(b_{\ell m})\,| \\ &\leq 2\,C_1 \sqrt{\tfrac{d}{n}\,\alpha_n\,\log \tfrac{d}{\alpha_n}} \;= o(\alpha_n^2). \end{aligned} \tag{5.20}$$

The last equality holds a.s. because $F_n(b_{jk}) = F_n(b_{\ell m}) = \alpha_n$ a.s.. Further observe that from **(A2)** and **(A3)** we have $A^-_{\ell m}(\partial \widehat{B}) \leq K_1/\epsilon_0$.

In order to finish the proof of (5.18) it remains to show that $|\,(III)\,| = |r_{jk} - r_{\ell m}| \leq K\,\alpha_n^2$ for a fixed constant $K > 0$. To see this observe that by using **(A2)** and **(A5)** that for some $K > 0$

$$\begin{aligned} |r_{jk}| &\leq \int_{a_{jk}-h_{jk}}^{a_{jk}} |\,ave_j(x, b_{jk}) - ave_j(a^+_{jk}, \widehat{B})\,|\,F_j(x, b_{jk} \cup \widehat{B})\,dx \\ &\leq K \int_{a_{jk}-h_{jk}}^{a_{jk}} |x - a_{jk}|\,F_j(x, b_{jk} \cup \widehat{B})\,dx = K\,h_{jk}\,F(b_{jk}) \leq 2\,K\,h_{jk}\,\alpha_n. \end{aligned}$$

It remains to show that $h_{jk} = O(\alpha_n)$. In fact, we have uniform upper and lower bounds for these widths, namely,

$$\frac{\alpha_n \beta_0}{2K_0} \;\leq\; h_{jk} \leq \frac{2\alpha_n}{\epsilon_0}. \tag{5.21}$$

with ϵ_0, K_0 from **(A2)**. To see that let $b_{jk} = \bigotimes_{i=1}^d [c^{jk}_{i1}, c^{jk}_{i2}]$ and denote the width of b_{jk} by $h_{jk} = c^{jk}_{j1} - c^{jk}_{j2}$. By using (5.19), the first inequality follows from

$$\begin{aligned} \alpha_n\,\beta_0/2 \leq F(b_{jk}) &= \int_{c^{jk}_{j1}}^{c^{jk}_{j1}+h_{jk}} F_j(x, b_{jk} \cup \widehat{B})\,dx \\ &\leq h_{jk}\,\sup_x F_j(x, b_{jk} \cup \widehat{B}) \leq h_{jk}\,K_0, \end{aligned}$$

The second inequality in (5.21) follows similarly. This completes the proof of (5.18). The next step of the proof is to show that the analogue to (5.18) 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

$$I_n(b_{jk}) - I_n(b_{\ell m}) \geq \tfrac{c\beta_0}{8}\,C\,\alpha_n^2. \tag{5.22}$$

Writing

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

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

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

It follows from assumption (5.7) that the r.h.s is $o(\alpha_n^2)$. This completes the proof of (5.15), and thus the proof of (5.10), in the non-overlapping case.

Now consider the case of an overlap of b_{jk} and $b_{\ell m}$. We can write

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

This implies with K_1 from **(A4)** that

$$I(b_{jk} \cap b_{\ell m}) \le K_1\,h_{jk}\,h_{\ell m} \le 4\,K_1\,\alpha_n^2/\epsilon_0^2. \tag{5.23}$$

A similar inequality (with different constants) also holds for I replaced by I_n. This follows directly from (5.21) together with the definition A_n and assumption (5.7).

We have shown in the proof above that if $\rho_\infty(B^*,\,\widehat{B}) > C\,2\,\alpha_n$ for some $C > 0$ large enough, then on A_n the difference $I_n(b_{jk}) - I_n(b_{\ell m}) \ge \tilde{C}\ \alpha_n^2$ where $\tilde{C}$ increases with C. We also have seen that $I_n(b_{jk} \cap b_{\ell m}) \le C'\ \alpha_n^2$ 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 . The above arguments can now be repeated mutatis mutandis.

It remains to show (5.11). Write B_{k_n} for the current box at a peeling procedure after k_n peels. Suppose there exists a set $B^* \in m_{\ell oc}(\beta_0)$ with $\rho_\infty(B_{k_n}, B^*) \le c\,\alpha_n \le \frac{c}{d}$ with $c > 0$ from the definition of $m_{\ell oc}(\beta_0)$. We show that on A_n jittering leads to a sequence $B_{k_n+1}, B_{k_n+2}, \ldots, B_{K_n}$ with $\rho_\infty(B_{K_n}, B^*) \ge 2\,C\,\alpha_n$, with the same C as in (5.10). This then completes the proof.

We again first consider the case of non-overlapping case, i.e. $A^{+}_{j1}(\partial B_{k_n}) \geq A^{-}_{j2}(\partial B_{k_n})$ for some j. Without loss of generality assume that $j = 1$. Further let $b^{k_n}_{11}$, $b^{k_n}_{12}$ be two corresponding candidate sets for jittering, $b^{k_n}_{11}$ lying outside of B_{k_n} and $b^{k_n}_{12}$ inside, and $F_n(b^{k_n}_{11}) = F_n(b^{k_n}_{12})$ (a.s.). We show that on A_n with n large enough we have

$$I_n(b^{k_n}_{11}) > I_n(b^{k_n}_{12}). \tag{5.24}$$

For the moment assume (5.24) to hold. Then, adding $b^{k_n}_{11}$ and subtracting $b^{k_n}_{12}$ leads to B_{k_n+1} with a larger average than B_{k_n}. Repeating this process leads to B_{k_n+2} with $ave_n(B_{k_n+2} \cap \widehat{S}^{k_n}) > ave_n(B_{k_n+1} \cap \widehat{S}^{k_n})$ etc. This process can be repeated (at least) as long as we are still in the neighborhood $U(B^*_{\beta_0}, \frac{c}{d})$ (cf. definition of $m_{\ell oc}(\beta_0)$ in (3.11)) which means until for some p we have $(\, b^{k_n+p}_{11} \cup b^{k_n+p}_{12}\,) \setminus (\,\overline{B} \setminus \underline{B}\,) \neq \emptyset$. If this happens, then, by definition of $m_{\ell oc}(\beta_0)$ (see (3.11)) we have $\rho_\infty(B_{k_n+p}, B^*) \geq \frac{c}{d} \geq 2\, C\, \alpha_n$ for any $C > 0$ and n large enough since $d\, \alpha_n \to 0$ by assumption.

To complete the proof it remains to show (5.24). Similarly to (5.22) we show that $I_n(b^{k_n}_{11}) - I_n(b^{k_n}_{12}) \geq K\, \alpha^2_n$. We write $B_{k_n} = \bigotimes_{i=1}^{d} [a_{i1}, a_{i2}]$ so that $A^{\pm}_{1j}(\partial B_{k_n}) = ave_1(a^{\pm}_{1j}, B_{k_n})$. Starting from (5.16) (with b_{jk} replaced by $b^{k_n}_{11}$) we have by using (**m.ii**) from the definition of $m_{\ell oc}(\beta_0)$ (see (3.11)) that

$$\begin{aligned}
I(b^{k_n}_{11}) &\geq ave_1(a^{+}_{11}, B_{k_n})\, F(b^{k_n}_{11}) + \int_{a_{11}-h_{11}}^{a_{11}} k_1\, |\, x - a_{11}\, |\, F_1(x, b_{11} \cup B_{k_n})\, dx \\
&\geq ave_1(a^{+}_{11}, B_{k_n})\, F(b^{k_n}_{11}) + k_1 \inf_x F_1(x, b_{11} \cup B_{k_n}) \int_{a_{11}-h_{11}}^{a_{11}} |\, x - a_{11}\, |\, dx \\
&= ave_1(a^{+}_{11}, B_{k_n}), F(b^{k_n}_{11}) + \tfrac{k_1\, \inf_x F_1(x, b_{11} \cup B_{k_n})}{2}\, h^2_{11}.
\end{aligned} \tag{5.25}$$

A similar, but simpler, argument shows that $I(b^{k_n}_{12}) \leq ave_1(a^{-}_{1,2}, b_{12} \cup B_{k_n})\, F(b^{k_n}_{12})$ and hence

$$\begin{aligned}
&I(b^{k_n}_{11}) - I(b^{k_n}_{12}) \\
&\geq ave_1(a^{+}_{11}, B_{k_n})\, F(b^{k_n}_{11}) - ave_1(a^{-}_{12}, b_{12} \cup B_{k_n})\, F(b^{k_n}_{12}) + \tfrac{k_1\, \inf_x F_1(x, b_{12} \cup B_{k_n})}{2}\, h^2_{11}
\end{aligned}$$

$$\geq ave_1(a_{11}^+, B_{k_n})\left(F(b_{11}^{k_n}) - F(b_{12}^{k_n})\right) + \tfrac{k_1 \inf_x F_1(x,B)}{2}\, h_{11}^2 \tag{5.26}$$

$$= o(\alpha_n^2) + 2K\alpha_n^2 \geq K\,\alpha_n^2. \tag{5.27}$$

The last line follows from (5.21) and (5.20). Hence we have shown (5.24). 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 on A_n, (5.18) implies (5.22). In case we have $A_{jk}^+(\partial B_{k_n}) \geq A_{\ell m}^-(\partial B_{k_n})$ with $j \neq m$ the two candidate boxes overlap. Similar to the above, the overlap is negligible, so that the analog of (5.24) still follows. As has been outlined above, this completes the proof. □

5.3 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 In the following $\mathcal{N}_{\ell oc}^{(k)}(\beta_0)$ and $m_{\ell oc}^{(k)}(\beta_0)$ denote the k-step analogs to $\mathcal{N}_{\ell oc}(\beta_0)$ and $m_{\ell oc}(\beta_0)$ as defined as in (3.10) and (3.11). For a precise define these classes one has to account for the iterative nature of the PRIM algorithm. For a given set $R \subset [0,1]^d$ let $\mathcal{N}_{\ell oc}(\beta_0, R)$ and $m_{\ell oc}(\beta_0, R)$ be defined as in (3.10) and (3.11) with $S = R$. Set $\mathcal{N}_{\ell oc}^{(1)}(\beta_0) = \mathcal{N}_{\ell oc}(\beta_0)$ and $m_{\ell oc}^{(1)}(\beta_0) = m_{\ell oc}(\beta_0)$ and define iteratively

$$\mathcal{N}_{\ell oc}^{(k)}(\beta_0) = \bigcup_{B^* \in \mathcal{N}_{\ell oc}^{(k-1)}(\beta_0)} \mathcal{N}_{\ell oc}(\beta_0, [0,1]^d \setminus B^*), \quad k \geq 2.$$

The classes $m_{\ell oc}^{(k)}(\beta_0)$ are defined analogously. Let

$$\mathcal{O}_{\ell oc}^{(k)}(\beta_0) := \mathcal{N}_{\ell oc}^{(k)}(\beta_0) \setminus m_{\ell oc}^{(k)}(\beta_0), \quad k = 1, \ldots, K.$$

For given β_0, λ and successive peeling+jittering outcomes $B^*_{(k)} \in \mathcal{O}^{(k)}_{\ell oc}(\beta_0)$, $k = 1, \ldots, K$, let $S^{(k)} = [0,1]^d \setminus \bigcup_{j=1}^{k-1} B^*_{(j)}$, and

$$K(\lambda) = \Big\{ 1 \le k \le K : \, ave\big(B^*_{(k)} \cap S^{(k)}\big) \, > \lambda \Big\},$$

and

$$R^*_\lambda = R^*_\lambda(B^*_{(1)}, \ldots, B^*_{(K)}) = \bigcup_{k \in K(\lambda)} \Big(B^*_{(k)} \cap S^{(k)}\Big).$$

Such sets R^*_λ are possible covering outcomes. Analogously, for successive outcomes $\widehat{B}_{(1)}, \ldots, \widehat{B}_{(K)}$ of empirical peeling+jittering let $\widehat{S}^{(k)} = [0,1]^d \setminus \bigcup_{j=1}^{k-1} \widehat{B}_{(j)}$ and

$$\widehat{K}(\lambda) = \Big\{ 1 \le k \le K : \, ave_n\big(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}\big) \, > \lambda \, \Big\},$$

and

$$\widehat{R}_\lambda = \bigcup_{k \in \widehat{K}(\lambda),} \big(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}\big).$$

Sets of the form $\widehat{R}_\lambda$ denote empirical covering outcomes.

We will assume that if, hypothetically, the input space at the $(k+1)$-st peeling step of the population version were somewhat off, while still being close to one of the possible input spaces $[0,1]^d \setminus B^*_{(k)}$ with some $B^*_{(k)} \in \mathcal{O}^{(k)}_{\ell oc, \epsilon}(\beta_0)$, then the corresponding optimal outcomes are also close:

(A7) Let B_j, $j = 1, \ldots, k$, be boxes in $[0,1]^d$, $k \in \mathbb{N}$. Let $R = \bigcup_{j=1}^k B_j$. Suppose that $\inf_{B \in \mathcal{O}^{(k)}_{\ell oc,\epsilon}(\beta_0)} d_F(R, B) \le \epsilon$ for some k and some $\epsilon > 0$. Then, for every $\widetilde{B} \in \mathcal{N}_{\ell oc}(\beta_0, R) \setminus m_{\ell oc}(\beta_0, R)$ there exists $B^* \in \mathcal{O}^{(k)}_{\ell oc}(\beta_0)$, with $d_F(\widetilde{B}, B^*) \le \epsilon$.

Theorem 5.3 *Let $0 < \beta_0 < 1$, $\lambda > 0$, and $K \in \mathbb{N}$ be fixed. Suppose that for any $C > 0$ the assumptions of Theorem 5.2 and* **(A7)** *hold uniformly in $S \in \{[0,1]^d \setminus B,\ B \in U_F\big(B^*, C\, d\, \alpha_n\big)\}$ for $B^* \in \mathcal{O}^{(k)}_{\ell oc}(\beta_0)$ and $k = 1, \ldots, K$. Then there exists $R^*_\lambda = R^*_\lambda(B^*_{(1)}, \ldots, B^*_{(K)})$, $B^*_{(k)} \in \mathcal{O}^{(k)}_{\ell oc}(\beta_0)$, such that the rate of convergence asserted in Theorem 5.2 (for the d_F-pseudo metric) also holds for $d_F(R^*_\lambda, \widehat{R}_\lambda)$.*

Proof. The idea of the proof is as follows. Let $\widehat{B}_{(1)}, \widehat{B}_{(2)}, \ldots, \widehat{B}_{(K)}$ denote the successive outcomes of empirical peeling+jittering. We will show that there exist $B^*_{(k)} \in \mathcal{N}^{(k)}_{\ell oc}(\beta_0) \setminus m^{(k)}_{\ell oc}(\beta_0)$, $k = 1, \ldots, K$ such that

$$d_F(\widehat{B}_{(k)} \cap \widehat{S}^{(k)},\, B^*_{(k)} \cap S^{(k)}) = O_P(d\,\alpha_n), \quad k = 1, \ldots, K, \tag{5.28}$$

where $S^{(1)} = [0,1]^d$ and $S^{(k)} = S^{(k-1)} \setminus B^*_{(k)}$, and similarly $\widehat{S}^{(1)} = [0,1]^d$ and $\widehat{S}^{(k)} = \widehat{S}^{(k-1)} \setminus \widehat{B}_{(k)}$.

Before we prove (5.28), we indicate how the assertion of the theorem follows from (5.28). Observe that (5.28) inplies that $F(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}) - F(B^*_{(k)} \cap S^{(k)}) = o_P(1)$, $k = 1, \ldots, K$, and thus we have on A_n (defined in the proof of Theorem 5.2) that $F_n(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}) - F(B^*_{(k)} \cap S^{(k)}) = o_P(1)$, $k = 1, \ldots, K$. Similarly it follows that $I_n(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}) - I(B^*_{(k)} \cap S^{(k)}) = o_P(1)$, $k = 1, \ldots, K$, and consequently,

$$ave_n\big(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}\big) - ave\big(B^*_{(k)} \cap S^{(k)}\big) = o_P(1), \;\; k = 1, \ldots, K.$$

It follows that $ave_n\big(\widehat{B}_{(k)} \cap \widehat{S}^{(k)}\big) > \lambda$ with probability tending to 1 if and only if for the same k we have $ave\big(B^*_{(k)} \cap S^{(k)}\big) > \lambda$. Therefore, we have $K(\lambda) = \widehat{K}(\lambda)$ with probability tending to one. Thus

$$d_F(\widehat{R}_\lambda,\, R^*_\lambda) \le \sum_{k \in K(\lambda)} d_F(\widehat{B}_{(k)} \cap \widehat{S}^{(k)},\, B^*_{(k)} \cap S^{(k)}).$$

Since $\beta_0 > 0$ was fixed, the number K (and thus the set $K(\lambda)$) is finite, and the assertion of the theorem thus follows from (5.28).

We now indicate how to prove (5.28). The case $K = 1$ is Theorem 5.2. For $K = 2$ observe that from the case $K = 1$ we know that $d_F(B_{(1)}, \widehat{B}_{(1)}) = O_P(d\,\alpha_n)$. In other words, with high probability, $\widehat{B}^{(1)}$ is in a small enough neighborhood around $B^{(1)}$. Therefore, the assumptions of the theorem allows us to use the ideas of the proof of Theorem 5.2 with S replaced by $\widehat{S}^{(2)} = S \setminus \widehat{B}_{(1)}$.

Let $\widetilde{B}_{(2)}$ denote the outcome of the population version of the PRIM algorithm applied with $S = [0,1]^d \setminus \widehat{B}_{(1)}$. By using the same line of proof of Theorem 5.2 we obtain with $\mathcal{O}_{\ell oc}(\beta_0, [0,1]^d \setminus \widehat{B}_{(1)}) = \mathcal{N}_{\ell oc}(\beta_0, [0,1]^d \setminus \widehat{B}_{(1)}) \setminus m_{\ell oc}(\beta_0, [0,1]^d \setminus \widehat{B}_{(1)})$ that

$$\inf_{B_{\widetilde{B}_{(2)}} \in \mathcal{O}_{\ell oc}(\beta_0, [0,1]^d \setminus \widehat{B}_{(1)})} d_F(\widehat{B}_{(2)} \cap \widehat{S}^{(2)},\, \widetilde{B}_{(2)} \cap \widehat{S}^{(2)}) = O_P(d\,\alpha_n). \tag{5.29}$$

Next we utilize assumption $\mathbf{(A7)}_{\beta_0}$. This gives us the existence of a set $B^*_{(2)} \in \mathcal{N}^{(2)}_{\ell oc}(\beta_0) \setminus m^{(2)}_{\ell oc}(\beta_0)$ with $d_F(B^*_{(2)} \cap S^{(2)}, \widetilde{B} \cap \widehat{S}^{(2)}) \leq C\,d\,\alpha_n$. This together with (5.29) gives the assertion of the theorem. The case of an arbitrary K follows analogously. This completes the proof.

□

6 Appendix

Here 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.

Let $\mathcal{G}$ be a class of functions with $\|\, g \,\|^2 = \mathrm{E}g^2(X,Y) < \infty$, 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 6.1 *We have*

$$H_B(u, \mathcal{B}) \leq 2\,d\,\log \left\lceil \frac{d}{u^2} \right\rceil. \tag{6.1}$$

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 $a_{(i,k)} = F_{(i)}^{-1}(k\,u),\ k = 1, \ldots, \lceil \frac{1}{u} \rceil - 1$ as well as $F^{-1}(0)$ and $F^{-1}(1)$. 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 $\binom{\lceil \frac{1}{u} \rceil + 1}{2}^d$ rectangles. By construction, any rectangle $B \in \mathcal{B}$ has a lower and an upper approximation $B_*, B^* \in \mathcal{B}_u$ with $F(B^* \,\Delta\, B_*) \leq d\,u$. Hence we have shown that $H_B(\sqrt{d\,u}\,, \mathcal{B}) \leq d \log \binom{\lceil \frac{1}{u} \rceil + 1}{2} = d \log \frac{\left(\lceil \frac{1}{u} \rceil + 1\right) \lceil \frac{1}{u} \rceil}{2} \leq 2\,d\, \log\lceil \frac{1}{u} \rceil$. It follows that $H_B(u\,, \mathcal{B}) \leq 2\,d\, \log\lceil \frac{d}{u^2} \rceil$. $\Box$

Proposition 6.1 *Suppose that* $\{\, (X_i, Y_i),\ i = 1, \ldots, n \,\}$ *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*

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

Proof. Write $\sqrt{n}\,(I_n - I)(B) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_B(X_i,\, \eta_i) - \mathrm{E} g_B(X_i,\, Y_i)$ 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 $\mathrm{E} g_B(X_1, Y_1) = \mathrm{E} g_B^+(X_1, Y_1) - \mathrm{E} g_B^-(X_1, Y_1)$ we have

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

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

$$\mathrm{E}\big(\sup_{F(B)\le\delta} \sqrt{n}|\,(I_n - I)(B)\,|\big) \le \mathrm{E}\big(\sup_{\|g_B^+\|\le\sqrt{M\,\delta}} |\,\nu_n(g_B^+)\,|\big) + \mathrm{E}\big(\sup_{\|g_B^-\|\le\sqrt{M\,\delta}} |\,\nu_n(g_B^-)\,|\big).$$

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 $|g_B^{\pm}(x,y)| \le |y|$. Hence the function $G(y) = y$ is an envelope for both classes $\mathcal{G}^{\pm}$ and we have $\|G\| \le \sqrt{M}$.

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

$$\mathrm{E}\big(\sup_{\|g_B^{\pm}\|\le\sqrt{M\,\delta}} |\,\nu_n(g_B^{\pm})\,|\big) \ \le\ \int_0^{\sqrt{M\,\delta}} \sqrt{1 + H_B(u\sqrt{M}\,, \mathcal{G}^{\pm})}\, du \tag{6.2}$$

$$+\ \ \sqrt{n}\,\mathrm{E}[\,|Y_1|\,\mathbf{1}\{\,|Y_1| \ge a(\sqrt{M\,\delta}\,)\sqrt{n}\ \}\,]. \tag{6.3}$$

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 (6.1). For $B_* \subset B \subset B^*$ we have

$$g_{B_*}^{\pm} - \mathrm{E}(g_{B^*}^{\pm}) \le g_B^{\pm} - \mathrm{E}(g^{\pm}) \le g_{B^*}^{\pm} - \mathrm{E}(g_{B_*}^{\pm}),$$

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

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

We now can estimate the r.h.s. of (6.2). For $0 < \delta < \delta_0 := \frac{1}{\sqrt{e^{1/2}-1}}$ we have $1 < 2\,d\,\log\lceil\,\frac{d}{u^2}\rceil$ and hence we obtain for such δ that

$$\int_0^{\sqrt{M\,\delta}} \sqrt{1 + H_B(u\sqrt{M}\,, \mathcal{G}^{\pm}\,)}\ du \le \int_0^{\sqrt{M\,\delta}} \sqrt{1 + 2\,d\,\log\lceil\,\tfrac{d}{u^2}\rceil}\ du$$

$$\le 2\sqrt{M\,\delta\,d\,\log\lceil\,\tfrac{d}{M\,\delta}\rceil}.$$

It remains to estimate (6.3). We have for $\gamma > 1$ and $0 < \delta < \delta_0$ that

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

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):

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

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

$$\mathrm{E}\big(\sup_{F_n(B)\leq\delta;B\in\mathcal{B}} |\,(F_n - F)(B)\,|\,\big) \leq C_0\,\sqrt{\tfrac{d}{n}\,\delta\,\big(\,1 + \log\tfrac{1}{\delta}\,\big)}\,.$$

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] Burman, P. and Polonik, W. (2009): Multivariate mode hunting: Data analytic tools with measures of significance. *J. Multivariate Anal.* **100** 1198 – 1218.

[4] 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.

[5] Friedman, J. H. and Fisher, N. I. (1999) Bump hunting in high-dimensional data. *Statist. & Comp.* **9** 123–143.

[6] Hastie, T., Tibshirani, R. and Friedman, H.H. (2001) *The elements of statistical learning.* Springer, New York

[7] Leblanc, M., Jacobson J. and Crowley J., (2002) Partitioning and peeling for constructing prognostic groups. *Statistical Methods for Medical Research,* **11** 247–274.

[8] Nannings, B., Abu-Hanna, A., de Jonge, E. (2008) Applying PRIM (patient rule induction method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. *Int. J. Med. Inf.* **77** 272-279.

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

[10] Polonik, W. (1997) Minimum volume sets and generalized quantile processes. *Stoch. Proc. and Appl.* **69** 1 - 24.

[11] Tsybakov, A.B. (2004): Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135-166.

[12] van der Vaart, A.W. (1998): *Asymptotic Statistics.* Cambridge Univ. Press.

[13] van der Vaart, A.W. & Wellner, J. (1996): *Weak convergence and empirical processes.* Springer, New York.

[14] 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.

[15] Wu, L. and Chipman, H. (2003) Bayesian model-assisted PRIM algorithm. Technical Report

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.

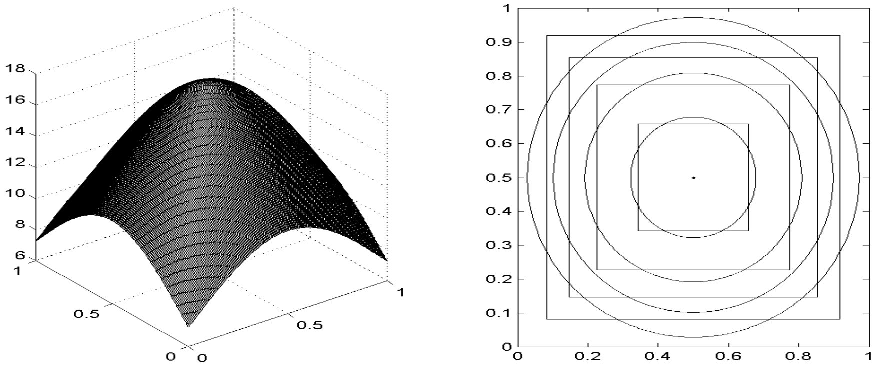


Figure 2: Comparison of optimal solution of (2.1) with peeling/pasting/jittering results for (3.12) by parameters α=.005 and β_0=.12. The optimal intervals are (local) level sets and all the intervals shown in the plots are drawn at the same height corresponding to the level of the level sets. Plot 1: 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 $\beta = 0.1206$.

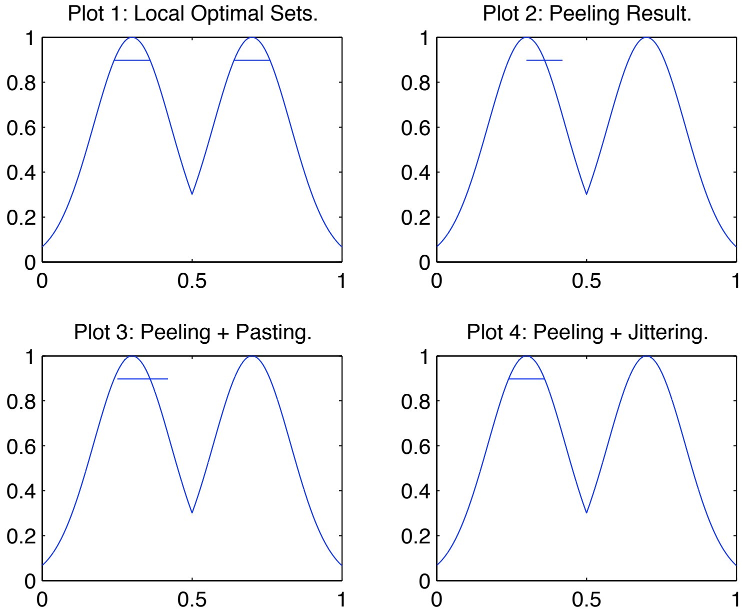
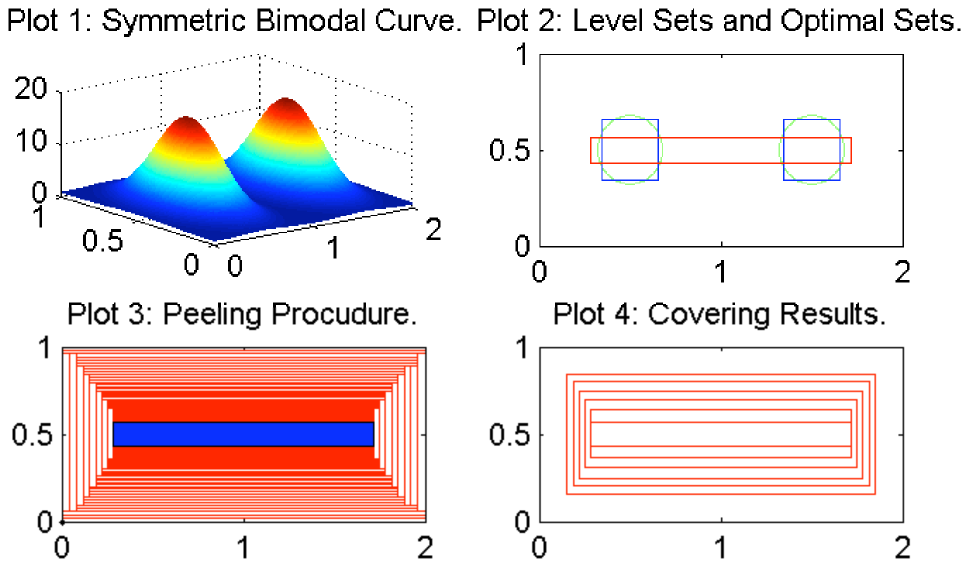


Figure 3: Plot 1 shows the model (symmetric bimodal curve). Plot 2: the level sets (circles) and optimal boxes (squares and rectangles) for $\beta_0 = .1$. Plot 3: indicates the (first) peeling procedure and the outcomes. Plot 4: the covering results for the first 6 peeled boxes. The parameters are $\alpha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



Plot 1: Symmetric Bimodal Curve.

Plot 2: Level Sets and Optimal Sets.

Plot 3: Peeling Procudure.

Plot 4: Covering Results.

Figure 4: Two consecutive peeling outcomes with $\alpha = 0.05$ for Model I (population version); with ratio control (panel 1); without ratio control (panel 2)

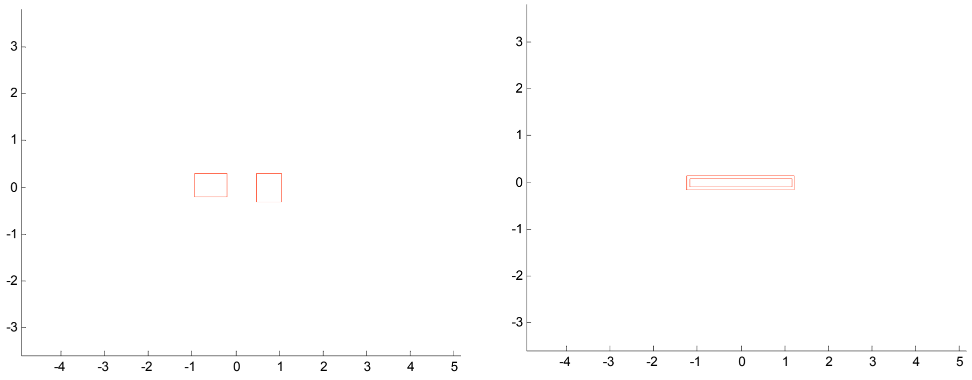
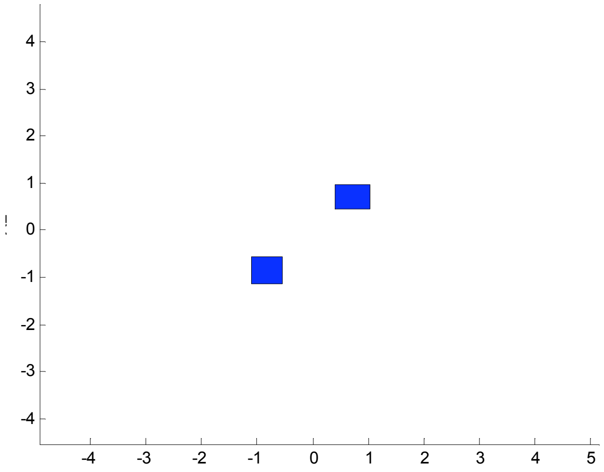


Figure 5: Two consecutive peeling outcomes with $\alpha = 0.05$ for Model II (population version); ratio control and ratio control give the same result


4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5