Parameter Estimation and Inference in a Continuous Piecewise Linear Regression Model

Georg Hahn, Moulinath Banerjee and Bodhisattva Sen

Department of Statistics
Columbia University
New York NY 10027, USA

March 21, 2017

Abstract

The estimation of regression parameters in one dimensional broken stick models is a research area of statistics with an extensive literature. We are interested in extending such models by aiming to recover two or more intersecting (hyper)planes in multiple dimensions. In contrast to approaches aiming to recover a given number of piecewise linear components using either a grid search or local smoothing around the change points, we show how to use Nesterov smoothing to obtain a smooth and everywhere differentiable approximation to a piecewise linear regression model with a uniform error bound. The parameters of the smoothed approximation are then efficiently found by minimizing a least squares objective function using a quasi-Newton algorithm. Our main contribution is threefold: We show that the estimates of the Nesterov smoothed approximation of the broken plane model are also $\sqrt{n}$ consistent and asymptotically normal, where n is the number of data points on the two planes. Moreover, we show that as the degree of smoothing goes to zero, the smoothed estimates converge to the unsmoothed estimates and present an algorithm to perform parameter estimation. We conclude by presenting simulation results on simulated data together with some guidance on suitable parameter choices for practical applications.

Keywords: broken stick model, broken plane model, parameter estimation, smooth approximation, Nesterov smoothing, quasi-Newton algorithm

1 Introduction

We are interested in parameter estimation and inference in a regression model of the type

$$Y = g_\theta(X) + \epsilon, \tag{1}$$

where Y is the response variable, $X \in \mathcal{X} \subset \mathbb{R}^d$ ($d \geq 1$), and $g_\theta : \mathbb{R}^d \to \mathbb{R}$ is *continuous k-piecewise affine* (k-PWA; $k \geq 2$) — there exists a partition of $\mathcal{X}$ into polyhedral sets $\{C_i\}_{i=1}^k$ (i.e., $C_i \subset \mathcal{X}$, $C_i \cap C_j = \emptyset$ for all $i \neq j$) such that

$$g_\theta(x) = a_i^\top x + b_i \qquad \text{if } \; x \in C_i, \tag{2}$$

and that g is continuous on $\mathcal{X}$; see e.g., Scholtes (2012). We assume that $k \in \mathbb{N}$ is given and denote the unknown parameter as $\theta = (a_1, \ldots, a_k, b_1, \ldots, b_k) \in \mathbb{R}^{k(d+1)}$. The unobserved error ϵ is assumed to have zero mean and finite variance.

Given i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ from the above model the goal is to estimate the unknown parameter θ and develop valid inferential procedures for the obtained estimator. A naive approach to solving the above parametric regression problem is to consider the least squares estimator (LSE):

$$\tilde{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^{k(d+1)}} \sum_{i=1}^n (Y_i - g_\theta(X_i))^2, \tag{3}$$

where the minimization is over all continuous PWA g_θ. However, the above estimator is computationally intractable — the optimization problem is non-smooth and non-convex; as noted in Polyak (1987, Chapter 5), virtually no computational guarantees are available for such problems. Such non-smooth functions can only be optimized using gradient-free or subgradient methods which typically attain a square root convergence rate as opposed to the superlinear convergence rate of the quasi-Newton method (under suitable conditions); see e.g., Shor (1985) and Nesterov (2005).

In this paper we resolve this non-smoothness in the estimation of θ by first smoothing g_θ appropriately and then minimizing the least squares criterion with the smoothed approximation of g_θ using a non-linear smooth optimization method such as *BFGS* (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). The novelty of our approach lies in the fact that: (i) we give provable bounds on the (smooth) approximation error of g_θ, and (ii) we theoretically analyze the obtained computationally feasible estimator $\hat{\theta}$ and prove that $\hat{\theta}$ has the same statistical efficiency as $\tilde{\theta}$, the LSE described in (3).

Before we describe our procedure in detail let us look at two motivating real examples where modeling the regression function as in (2) can be useful.

Example 1. *As a first example we consider the PWA in (Hempel et al., 2013, Figure 2)*

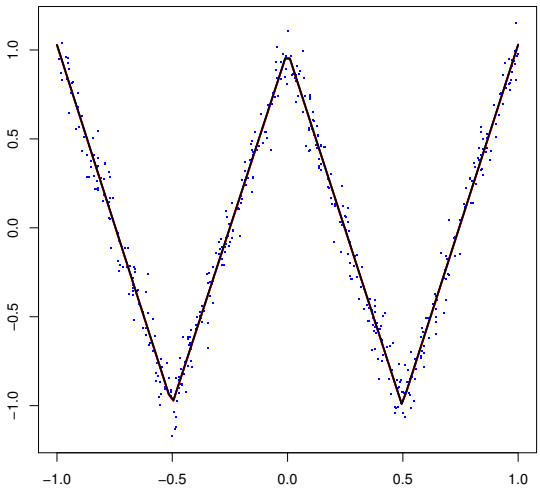

Figure 1: Points generated from the PWA in (Hempel et al., 2013, Figure 2) (blue) and fitted PWA (red) consisting of a difference of two PWAs with three and two lines, respectively.

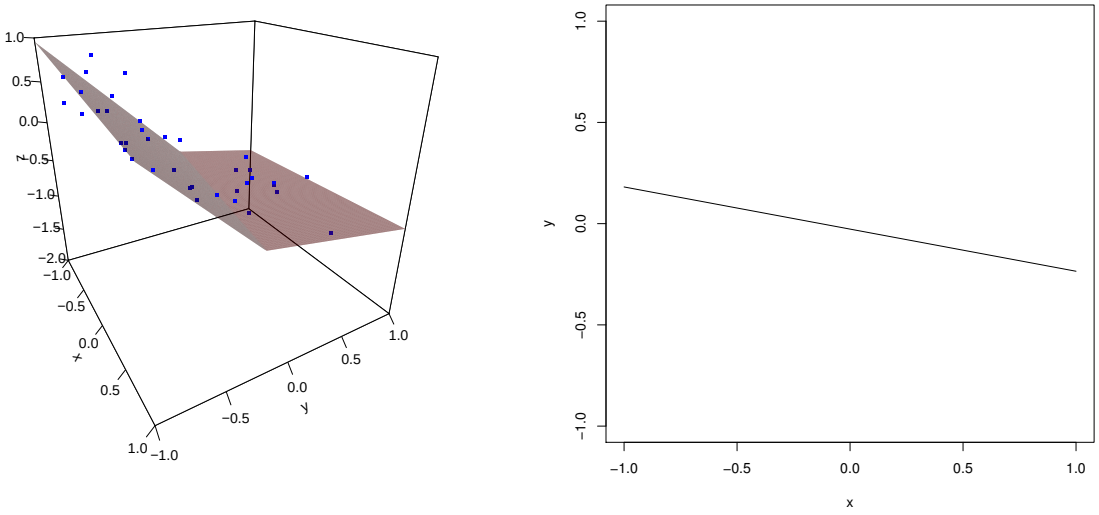

Figure 2: Left: Visual representation of our two scenarios derived from the car dataset in (Henderson and Velleman, 1981, Table 3). Left: Variable *MPG* as a function of the variables describing acceleration *ACCEL* and weight *WT*. Right: Projected intersection line for the best fit displayed on the left.

	$d=1$		$d=2$	
	average of R	time [s]	average of R	time [s]
Nelder and Mead (1965)	0.32	75.5	0.058	1.9
Algorithm 1	0.30	15.9	0.047	0.8

Table 1: Empirical norm and computation time for the two segmented regressions in one (Example 1) and two (Example 2) dimensions. Nelder and Mead (1965) method and Algorithm 1.

displayed in Figure 1 (scaled down to lie in the interval $[-1, 1]$ *for conventional reasons). This PWA is composed of a difference of two PWAs, the first one consisting of three and the second of two lines (see Hempel et al. (2013)). We hence try to fit the PWA* $g_\theta(x) = \max_{i=1,2,3}\{a_i^\top x+b_i\}-\max_{j=1,2}\{c_i^\top x+d_j\}$*, where* $\theta = (a_1, b_1, \ldots, a_3, b_3, c_1, d_1, c_2, d_2)$ *and* $a_i, b_i \in \mathbb{R}$*,* $c_i, d_i \in \mathbb{R}$ *for* $i = 1, 2, 3$*,* $j = 1, 2$*. We first simulated* 500 *points from this PWA by adding Gaussian noise (draws from iid. random variables with zero mean and standard deviation* 0.1*) to uniformly selected function values. We then use our approach (formalized as Algorithm 1 in Section 3) and the gradient-free Nelder and Mead (1965) algorithm to fit* $g_\theta(x)$ *to the data.*

Algorithm 1 has one tuning parameter, the smoothness parameter μ*. In this and the following example we used* $\mu = 0.1$*. For the Nelder and Mead (1965) algorithm we used the implementation in the R function optim. We measure the performance of any algorithm using the empirical norm (the average squared residuals), that is* $R := \frac{1}{n}\sum_{i=1}^n (Y_i - g_{\hat\theta}(X_i))^2$*, where* $(X_i, Y_i)_{i=1}^n$ *are the data,* $g_\theta(x)$ *is as above and* $\hat\theta$ *is the fitted parameter vector.*

Results based on 1000 *repetitions are given in Table 1. As seen from the table, both the Nelder and Mead (1965) and our algorithm perform comparably, yet Algorithm 1 yields results considerably faster even in the one dimensional case.*

Example 2. *As a second example we consider the car dataset in (Henderson and Velleman, 1981, Table 3) which has already been used in various publications to assess the performance of algorithms fitting breakpoint models. This dataset contains consumer report data for* 38 *car models captured in* 11 *variables which include fuel consumption in miles per gallon, number of cylinders or the car weight.*

We attempt to fit the fuel consumption MPG as a function of the acceleration ACCEL and weight WT using a PWA consisting of two planes given by $g_\theta(x) = \max_{i=1,2}\{a_i^\top x+b_i\}$*, where* $x, a_i \in \mathbb{R}^2$*,* $b_i \in \mathbb{R}$ *for* $i = 1, 2$*. The choice of parameters for both algorithms as well as the evaluation using empirical norm and computational time are as in Example 1.*

Table 1 (right) shows results. Our algorithm yields a speed-up of a factor two at a higher accuracy over the method of Nelder and Mead (1965) at comparable accuracy of both methods. A graphical representation of the fitted functional is given in Figure 2 (left).

Importantly, the best fit obtained with our algorithm displayed in Figure 2 (left) finds

two planes with non-trivial intersection line shown in Figure 2 (right) given by $y = -0.20x - 0.06$. *Approaches regressing only one variable at a time would thus not have been able to obtain a fit with comparable precision.*

Many more real-life phenomena are captured by a model of type (1) with continuous piecewise affine g_θ: these include, for instance, the Michigan Bone Health and Metabolism Study (MBHMS) in Das et al. (2015), export prices in six EEC countries in Ginsburgh et al. (1980), leukaemia and lung cancer data or time series of AIDS cases (Stasinopoulos and Rigby, 1992; Rigby and Stasinopoulos, 1992; Molinari et al., 2001; Muggeo, 2003).

We summarize the main contributions of the paper below.

1. Our smoothing approach uses ideas from Nesterov (2005) and yields a smooth approximation g_θ^μ of g_θ that is uniformly close (up to any desired precision depending on the smoothing parameter $\mu > 0$) to g_θ. In fact, we can provide theoretical bounds on the quality of the smooth approximation (i.e., $|g_\theta^\mu - g_\theta|$). Our approach is very different from the usual techniques (e.g., kernel smoothing) employed in non-parametric statistics to obtain smooth approximations of non-smooth functions. To apply our smoothing technique we first express g_θ as the difference of two PWA convex functions which are then smoothed separately while preserving convexity.

2. We give an algorithm based on a quasi-Newton method to estimate θ which is fast and stable in practice. We evaluate our procedure on simulated data and provide empirical guidelines on how to choose the tuning parameter μ in practice. In contrast to existing approaches summarized below, which usually address $d = 1$ or 2, the results in this article are not limited to any dimension. Moreover, existing methods such as the ones of Tishler and Zang (1981) or Muggeo (2003) which are able to handle more than $k = 2$ components are usually based on heuristics and lack a proper theoretical study. To the best of our knowledge, the proposed procedure seems to be the first attempt at a systematic study of the estimation of θ under such generality (i.e., $k \geq 2$ and $d \geq 1$). Can we give any theoretical guarantees on the computation of $\tilde{\theta}$?

3. To describe the theoretical results on the statistical performance of the proposed estimator $\hat{\theta}$ let us consider the case $k = 2$, which has received some attention in the statistical literature — see e.g., Ginsburgh et al. (1980), Smith and Cook (1980), Bacon and Watts (1971), van de Geer (1988). In this situation, g_θ is either convex or concave. Without loss of generality let us assume that g_θ is convex, in which case g_θ can be easily represented as

$$g_\theta(x) = \max\{\alpha^\top x + \gamma, \beta^\top x + \phi\}, \tag{4}$$

where $\theta = (\alpha, \beta, \gamma, \phi) \in \mathbb{R}^{2d+2}$. Even in this simple setup, the computation of $\tilde{\theta}$ (using (3)) is non-trivial: The non-smoothness of g_θ complicates the use of gradient based optimization methods (such as quasi-Newton or Newton-Raphson) to minimize the least squares criterion. In this setting we theoretically analyze the statistical properties of our proposed estimator $\hat{\theta}$. We show that, under proper choices of μ and mild conditions on the distribution of (X, Y), $\hat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal with the *same* limiting distribution as that of the computationally infeasible LSE $\tilde{\theta}$. This immediately yields confidence intervals for θ. Our analysis, in principle, can be extended beyond the setting $k = 2$ where similar results would also hold for the proposed estimator.

Early work on linear regression with change points dates back to Quandt (1958) and Quandt (1960) who consider $d = 1$; also see Blischke (1961), Robison (1964), Hudson (1966). Asymptotic results on the limiting distributions of the LSE in regression models with separate analytical forms in different regions appear in Feder (1975). Siegmund and Zhang (1994) develop conservative confidence regions for the change point of a broken stick model (i.e., $k = 2$) in one dimension.

The idea of smoothing dates back to Tishler and Zang (1981). In order to overcome the lack of smoothness around each change point, they propose to use a smooth quadratic approximation of the model in an interval $(-\beta, \beta)$ around each change point, hence allowing for efficient minimization of the likelihood function using the Newton algorithm. Although both the smooth approximation as well as the dependence on the parameter β are evaluated numerically, no theoretical results or asymptotic distributions are given. The present work is closely related to the work of Das et al. (2015) who consider a broken stick model in two dimensions, possibly with multiple change points. Das et al. (2015) propose a method relying on local smoothing in a neighborhood of each change point and show that the resulting estimate is $\sqrt{n}$ consistent and asymptotically normal.

Nevertheless, all of the aforementioned works suffer from the drawback that only one dimensional problems are considered, and that their results are not extendable to higher dimensions.

The article is structured as follows. Section 2 discusses the smoothing of piecewise affine functions, presents the Nesterov (2005) smoothing technique (Section 2.1) and discusses the choice of the so-called prox function required for smoothing (Section 2.2). We investigate two smoothing techniques based on two different prox functions. An algorithm to compute estimates of both the change point as well as the regression parameters will then be given in Section 3. Section 4 discusses asymptotic results. We show that the least squares estimate of our proposed smoothed problem is also $\sqrt{n}$ consistent, asymptotically normal and moreover converges to the one of the unsmoothed problem as the Nesterov (2005) smoothing parameter goes to zero at an appropriate rate. Section 5 presents selected simulation results demonstrating the accuracy of the estimates obtained with our

approach, verifying convergence rates and highlighting computational issues. The article concludes with a discussion in Section 6. All proofs can be found in the appendix.

2 Smoothing piecewise affine functions

Consider the regression model (1) where g_θ is a continuous piecewise affine (PWA) function. The following result, which states that any continuous PWA function can be expressed as the difference of two convex PWA functions, will be crucial for the rest of the sequel.

Lemma 1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex polyhedral region (i.e., a set formed by the intersection of finitely many hyperplanes). Every continuous PWA function $g : \mathcal{X} \to \mathbb{R}$ defined over a convex polyhedral partition of $\mathcal{X}$ with full dimensional elements C_i, $i = 1, \ldots, k$, can be written as the difference of two convex PWA functions, i.e.,*

$$g(x) = g_1(x) - g_2(x), \qquad \text{for all } x \in \mathcal{X},$$

where $g_i : \mathcal{X} \to \mathbb{R}$, $i = 1, 2$, are convex and PWA functions.

The above result is proved in Kripfganz and Schulze (1987); also see Hempel et al. (2013). Consequently, to find a smooth approximation of a PWA (continuous) g we first consider the case when g is PWA and also *convex*. The advantage of using a convex PWA is that it can be conveniently represented as a maximum of affine functions, i.e.,

$$g(x) = \max_{i=1,\ldots,k} \{a_i^\top x + b_i\} \tag{5}$$

for some $k \geq 1$ where $a_1, \ldots, a_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$. Let A be a $k \times (d+1)$ matrix such that the i-th row of A is $(a_i, b_i) \in \mathbb{R}^{d+1}$. Then g can be succinctly represented as

$$g(x) = \max_{i=1,\ldots,k} (A[x, 1])_i,$$

where for a vector $u \in \mathbb{R}^k$, $(u)_i$ denotes the i-th component of u, and where $[x, 1] \in \mathbb{R}^{d+1}$ is the concatenation of vector x and scalar 1.

In the following subsection we detail an approach to smooth a convex PWA. This procedure will then be used to construct a smooth approximation to any continuous PWA function.

2.1 The smoothing approach

Suppose that $f : \mathbb{R}^q \to \mathbb{R}$ ($q \geq 1$) be a piecewise linear convex function. Then we can express f as

$$f(z) = \max_{i=1,\ldots,p} (Az)_i, \qquad \text{for all } z \in \mathbb{R}^q, \tag{6}$$

for some matrix A of order $p \times q$. In light of the specific representation of interest given in (5), the dimension q will correspond to $q = d + 1$ and likewise for a k-PWA, we will later consider $p = k$. In this subsection we find a smooth convex approximation to any f of the form defined in (6) and study some of its properties.

Let $\|\cdot\|_p$ ($p \geq 1$) be a norm on $\mathbb{R}^p$ and let $\langle \cdot, \cdot \rangle$ denote the usual Euclidean inner product. Let $Q_p \subset \mathbb{R}^p$ be the unit simplex in $\mathbb{R}^p$, i.e.,

$$Q_p := \left\{ w = (w_1, \ldots, w_p) \in \mathbb{R}^p : \sum_{i=1}^{p} w_i = 1, \text{ and } w_i \geq 0, \text{ for all } i = 1, \ldots, p \right\}.$$

Following Nesterov (2005), let ρ be a proximity function (*prox function*) — a nonnegative continuously differentiable *strongly convex* function (with respect to the norm $\|\cdot\|_p$) on Q_p, i.e.,

$$\rho(s) \geq \rho(t) + \langle \nabla \rho(t), t - s \rangle + \frac{1}{2}\|t - s\|_p, \qquad \text{for all } s, t \in Q_p.$$

Consider now the function $f^\mu : \mathbb{R}^q \to \mathbb{R}$ defined as

$$f^\mu(z) := \max_{w \in Q_p} \{\langle Az, w \rangle - \mu \rho(w)\}, \tag{7}$$

where $\mu > 0$ is a tuning parameter. Then f^μ is our *smooth approximation* of f. Observe that when $\mu = 0$, $f^0(z)$ recovers the unperturbed function f (cf. (6)):

$$f(z) = \max_{w \in Q_p} \{\langle Az, w \rangle\} = f^0(z).$$

The following lemma, taken from Nesterov (2005, Theorem 1), shows that f^μ is a smooth convex function.

Lemma 2. *For any $\mu > 0$, the function f^μ defined in (7) is convex and everywhere differentiable in z. The gradient of f^μ is given by $\frac{\partial}{\partial z} f^\mu(z) = A^\top \hat{w}_\mu$, where $\hat{w}_\mu = \arg\max_{w \in Q_p} \{\langle Az, w \rangle - \mu \rho(w)\}$. Moreover, the gradient $z \mapsto \frac{\partial}{\partial z} f^\mu(z)$ is Lipschitz continuous with parameter $\|A\|_{p,q}^2 / \mu$, where*

$$\|A\|_{p,q} := \max_{u,v} \{\langle Au, v \rangle : \|u\|_q = 1, \|v\|_p = 1, u \in \mathbb{R}^q, v \in \mathbb{R}^p\}$$

is a norm on the space of matrices in $\mathbb{R}^{p \times q}$.

The definition of the smoothed estimator f^μ in (7) immediately allows us to obtain

bounds on the approximation error $|f(z) - f^{\mu}(z)|$. Indeed,

$$f^{\mu}(z) \geq \sup_{w \in Q_p} \langle Az, w \rangle - \mu \sup_{w \in Q_p} \rho(w) = f^0(z) - \mu \sup_{w \in Q_p} \rho(w),$$
$$f^{\mu}(z) = \sup_{w \in Q_p} \{\langle Az, w \rangle - \mu\rho(w)\} \leq \sup_{w \in Q_p} \langle Az, w \rangle = f^0(z),$$

using the nonnegativity of the prox function (Nesterov, 2005); also see Mazumder et al. (2015, Section 3.1) for a detailed description of this approach. Summarizing the above considerations we obtain the following two useful results:

1. The function f^{μ} defined in (7) is smooth for any $\mu > 0$, convex and has a Lipschitz continuous gradient which is proportional to μ^{-1}. The original unsmoothed PWA function f is recovered by setting $\mu = 0$.

2. $f^{\mu}(z)$ is a *uniform approximation* to $f^0(z)$ as

$$f^0(z) - \mu \sup_{w \in Q_p} \rho(w) \leq f^{\mu}(z) \leq f^0(z). \tag{8}$$

   Thus the uniform approximation error is upper bounded by

$$\sup_{z \in \mathbb{R}^p} |f(z) - f^{\mu}(z)| \leq \mu \sup_{w \in Q_p} \rho(w) = O(\mu)$$

   which depends only on μ and ρ.

The choice of μ will be important and we give some sufficient conditions on μ in Section 4 for our main theoretical results to hold. We further provide practical guidelines on how to choose μ in Section 5.3. There are various possible choices of the prox function ρ. We will discuss two (standard) choices in detail in the following section which are also used throughout the paper, including in our simulation studies.

We aim to use the approach in (7) to obtain a smooth uniform approximation to any convex PWA f of the form defined in (5).

2.2 Choice of the prox function

We consider two prox functions in this article, one based on the squared error loss and the other based on the entropy loss. As will be shown in Section 4.2 both these prox functions yield $\sqrt{n}$-consistent and asymptotically normal estimators of θ with the *same* limiting distribution as the one of the naive LSE $\tilde{\theta}$.

- **Entropy prox function:** The entropy prox function $\rho : \mathbb{R}^p \to \mathbb{R}$ is given by

$$\rho(w) = \sum_{i=1}^{p} w_i \log(w_i) + \log p,$$

where $w = (w_1, \ldots, w_p)$. Letting $\|\cdot\|_p$ be the ℓ_1-norm in $\mathbb{R}^p$ (i.e., $\|u\|_p = \sum_{i=1}^p |u_i|$) it can be shown that ρ is strongly convex with respect to this norm and satisfies

$$\sup_{w \in Q_p} \rho(w) = \log p;$$

see e.g., Nesterov (2005). More importantly, the entropy prox function allows an analytic expression: for the PWA convex function f given in (5), the prox-smoothed approximation of f is given by

$$\begin{aligned} f^\mu(x) &= \max_{w \in Q_k} \left\{ \sum_{i=1}^k w_i (a_i^\top x + b_i) - \mu \left(\sum_{i=1}^k w_i \log w_i + \log k \right) \right\} \\ &= \mu \log \left(\frac{1}{k} \sum_{i=1}^k e^{\frac{a_i^\top x + b_i}{\mu}} \right). \end{aligned} \tag{9}$$

Further, using (8), we have

$$\sup_{x \in \mathbb{R}^d} |f(x) - f^\mu(x)| \leq \mu \sup_{w \in Q_k} \rho(w) = \mu \log k. \tag{10}$$

- **Squared error prox function:** The squared error prox function $\rho : \mathbb{R}^p \to \mathbb{R}$ is given by

$$\rho(w) = \frac{1}{2} \left\| w - \frac{1}{m} \mathbf{1} \right\|_p^2,$$

where $\mathbf{1} \in \mathbb{R}^d$ is the vector of ones and $\|\cdot\|_p$ denotes the Euclidean norm in $\mathbb{R}^p$. Following Mazumder et al. (2015), the optimization problem in (7) applied to the PWA function in (5) is equivalent to the following convex program:

$$f^\mu(x) = \min_{w \in Q_k} \left\{ \frac{1}{k} \sum_{i=1}^k w_i^2 - \sum_{i=1}^k w_i c_i^{\theta,\mu}(x) \right\}, \tag{11}$$

where $c_i^{\theta,\mu}(x) := (a_i^\top x + b_i)/\mu - 1/k$, $i \in \{1, \ldots, k\}$, and $\theta = (a_1, \ldots, a_k, b_1, \ldots, b_k)$. This problem is identical to the one of finding the Euclidean projection of the vector $(c_i^{\theta,\mu}(x))_{i=1}^k$ onto Q_k, the k-dimensional unit simplex. It can be solved efficiently using the algorithm of Michelot (1986). Denoting the Euclidean projection of the vector $(c_i^{\theta,\mu}(x))_{i=1}^k$ onto Q_k by $\hat{w}^{\theta,\mu}(x) = (\hat{w}_i^{\theta,\mu}(x))_{i=1}^k$, the prox-smoothed approximation for the PWA convex function f given in (5) can be written as

$$f^\mu(x) = \sum_{i=1}^k \hat{w}_i^{\theta,\mu}(x) \cdot (a_i^\top x + b_i) - \mu \rho(\hat{w}^{\theta,\mu}(x)). \tag{12}$$

As $\sup_{w\in Q_k} \frac{1}{k}\left|w-\frac{1}{k}\mathbf{1}\right|^2 = 1-\frac{1}{k}$, we have

$$\sup_{x\in\mathbb{R}^d} |f(x)-f^{\mu}(x)| \le \mu \sup_{w\in Q_k} \rho(w) = \mu\left(1-\frac{1}{k}\right). \tag{13}$$

3 Our Algorithm

In this section we describe our algorithm to estimate the k hyperplanes given in (2). By Lemma 1 we know that the continuous k-PWA function g_θ (in (2)) can be represented as

$$g_\theta(x) = g_{1,\theta_1}(x) - g_{2,\theta_2}(x) \qquad \text{for all } x\in\mathcal{X}, \tag{14}$$

where $\theta=(\theta_1,\theta_2)$ and g_{i,θ_i}, $i=1,2$, is a convex PWA with the representation (see (5))

$$g_{i,\theta_i}(x) = \max_{j=1,\dots,k_i} \{a_{i,j}^\top x + b_{i,j}\} \tag{15}$$

for $\theta_i = (a_{i,1},\dots,a_{i,k_i},b_{i,1},\dots,b_{i,k_i}) \in \mathbb{R}^{k_i(d+1)}$ and nonnegative integers k_i (note that $k_i = \#\theta_i/(d+1)$, where $\#\theta_i$ denotes the length of vector θ_i). We assume here that k_1 and k_2 are specified in advance by the user; see Section 5.6 for a discussion on how to choose these parameters. For identifiability reasons we can take $a_{2,1}=0, b_{2,1}=0$ as formalized in the next lemma.

Lemma 3. *In the model* (14) *with* g_{i,θ_i} *as defined in* (15)*, if* $k_i \ge 1$ *for* $i=1,2$*, we can without loss of generality assume that* $a_{2,1}=0, b_{2,1}=0$.

Proof. We have $g_\theta(x) = \max_{j=1,\dots,k_1}\{a_{1,j}^\top x + b_{1,j}\} - \max_{j=1,\dots,k_2}\{a_{2,j}^\top x + b_{2,j}\}$. We can express g_θ as:

$$g_\theta(x) = \max_{j=1,\dots,k_1} \{a_{1,j}^\top x + b_{1,j}\} - \max_{j=1,\dots,k_2} \{a_{2,j}^\top x + b_{2,j}\} \pm (a_{2,1}^\top x + b_{2,1})$$

and use the fact that $\max\{u,v\} - w = \max\{u-w, v-w\}$ for arbitrary $w,v,w\in\mathbb{R}$. Setting $\bar{a}_{i,j} := a_{i,j} - a_{2,1}$ and $\bar{b}_{i,j} := b_{i,j} - b_{2,1}$ for all $j=1,\dots,k_i$, $i=1,2$, leads to

$$g_\theta(x) = \max_{j=1,\dots,k_1} \{\bar{a}_{1,j}^\top x + \bar{b}_{1,j}\} - \max_{j=1,\dots,k_2} \{\bar{a}_{2,j}^\top x + \bar{b}_{2,j}\}$$

with $\bar{a}_{2,1}=0, \bar{b}_{2,1}=0$. □

With this change in parametrization, the goal is now to estimate the parameter $\theta = (\theta_1,\theta_2) \in \mathbb{R}^{(k_1+k_2)(d+1)}$. Given a set of data points $\{(X_i,Y_i)\}_{i=1}^n$ we use the method of least squares to estimate θ. As g_θ is non-smooth we cannot directly employ a gradient descent algorithm.

As a first step to resolve this difficulty, we compute smoothed convex approximations to g_{1,θ_1} and g_{2,θ_2} using the approach outlined in Sections 2.1 and 2.2. This leads to smoothed functions g^{μ}_{1,θ_1} and g^{μ}_{2,θ_2} and thus to a smoothed approximation of g_θ:

$$g^{\mu}_{\theta} := g^{\mu}_{1,\theta_1} - g^{\mu}_{2,\theta_2}.$$

Let

$$M^{\mu}_{n}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g^{\mu}_{\theta}(X_i)\right)^2$$

be the least squares criterion function we now try to minimize over $\theta \in \mathbb{R}^{(k_1+k_2)(d+1)}$. Thus, the LSE of θ we consider is

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) := \underset{\theta\in\mathbb{R}^{(k_1+k_2)(d+1)}}{\operatorname{argmin}} M^{\mu}_{n}(\theta). \tag{16}$$

We study the computation and the statistical properties of $\hat{\theta}$ is this paper. In this section we focus on the computation of $\hat{\theta}$. To minimize M^{μ}_{n} we first compute its Jacobian matrix:

$$J^{\mu}_{n}(\theta) = 2\sum_{i=1}^{n}\left(Y_i - g^{\mu}_{\theta}(X_i)\right)\cdot\nabla_\theta g^{\mu}_{\theta}(X_i),$$

where the gradient $\nabla_\theta g^{\mu}_{\theta}$ varies depending on the choice of the prox function used for smoothing. As

$$\nabla_\theta g^{\mu}_{\theta}(x) = [\nabla_{\theta_1} g^{\mu}_{1,\theta_1}(x), -\nabla_{\theta_2} g^{\mu}_{2,\theta_2}(x)] \qquad \text{for all } \; x \in \mathcal{X},$$

we have to first find the gradients of the function g^{μ}_{i,θ_i} $(i = 1, 2)$ with respect to θ_i. To this end we state the general form of the gradients for Nesterov (2005) smoothed functions.

Lemma 4. *For $\theta = (a_1, \ldots, a_k, b_1, \ldots, b_k) \in \mathbb{R}^{k(d+1)}$, let*

$$f_\theta(x) := \max_{j=1,\ldots,k}\{a_j^\top x + b_j\} = \max_{j=1,\ldots,k}(A[x,1])_j,$$

where $A \in \mathbb{R}^{k\times(d+1)}$ is a matrix whose j'th row is given by $(a_j, b_j) \in \mathbb{R}^{d+1}$. Then the derivatives (with respect to a_j's and b_j's) of the Nesterov (2005) smooth approximation of $f_\theta(x)$ defined in (7), i.e.

$$f^{\mu}_{\theta}(x) := \max_{w\in Q_p}\{\langle A[x,1], w\rangle - \mu\rho(w)\} = (A[x,1])^\top \hat{w}^{\theta,\mu} - \mu\rho(\hat{w}^{\theta,\mu}) \tag{17}$$

where $\hat{w}^{\theta,\mu} := \arg\max_{w\in Q_p}\{\langle A[x,1], w\rangle - \mu\rho(w)\}$ (as in Lemma 2) are

$$\frac{\partial f^{\mu}_{\theta}}{\partial a_j}(x) = \hat{w}^{\theta,\mu}_j x \qquad \text{and} \qquad \frac{\partial f^{\mu}_{\theta}}{\partial b_j}(x) = \hat{w}^{\theta,\mu}_j \qquad \text{for } j = 1, \ldots, k.$$

Proof. Fix $x \in \mathbb{R}^d$. Indeed, $f_\theta^\mu(x)$ (viewed as a function of θ) is a maximum of functions, which are linear in a_j's and b_j's. Note that $f_\theta^\mu(x)$ admits the representation

$$f_\theta^\mu(x) = \sum_{j=1}^{k} \hat{w}_j^{\theta,\mu}(a_j^\top x) + \sum_{j=1}^{k} \hat{w}_j^{\theta,\mu} b_j - \mu\rho(\hat{w}^{\theta,\mu}).$$

The result now follows because $f_\theta^\mu(x)$ is differentiable in θ since $\hat{w}^{\theta,\mu}$ is unique (note that $\langle A[x,1], w\rangle - \mu\rho(w)$ is a strongly convex function in w). □

We now specialize in the two prox functions used before and give explicit expressions for $\nabla_{\theta_i} g_{i,\theta_i}^\mu(x)$, for $i = 1, 2$.

- **Squared Error prox function:** The approximation g_{i,θ_i}^μ with squared error prox function makes use of the projected vector $\hat{w}^{\theta,\mu}$ maximizing (17) (see the discussion after (11)). It is thus straightforward to apply Lemma 4, leading to

  $$\nabla_{\theta_i} g_{i,\theta_i}^\mu(x) = [x, 1] \otimes \hat{w}^{\theta_i,\mu},$$

  where for two vectors $u = (u_1, \ldots, u_p) \in \mathbb{R}^p$ and $v = (v_1, \ldots, v_q) \in \mathbb{R}^q$ we define $u \otimes v := (u_1 v_1, \ldots, u_p v_1, u_1 v_2, \ldots, u_p v_2, u_1 v_3, \ldots)$ and $\hat{w}^{\theta,\mu}$ is as defined in Lemma 4.

- **Entropy prox function:** Although Lemma 4 gives an explicit derivative of each g_{i,θ_i}^μ for $i = 1, 2$, the vector $\hat{w}^{\theta,\mu} \in Q_p$ maximizing (17) is non-trivial to compute. Instead, the closed form expression

  $$g_{i,\theta_i}^\mu(x) = \mu \log\left(\frac{1}{k_i}\sum_{j=1}^{k_i} e^{\frac{a_{i,j}^\top x + b_{i,j}}{\mu}}\right)$$

  of the smooth approximation of g_{i,θ_i} given in (9) can be differentiated directly. This leads to

  $$\nabla_{\theta_i} g_{i,\theta_i}^\mu(x) = \|r_{i,\theta_i}\|_{k_i}^{-1} \cdot [x \otimes r_{i,\theta_i}, r_{i,\theta_i}],$$

  where $r_{i,\theta_i} = \left(e^{\frac{a_{i,1}^\top x + b_{i,1}}{\mu}}, \ldots, e^{\frac{a_{i,k_i}^\top x + b_{i,k_i}}{\mu}}\right)$ and $\|\cdot\|_{k_i}$ is the l_1-norm in $\mathbb{R}^{k_i}$.

Once the computation of the Jacobian matrix is completed, we use a hill-climbing optimization technique (quasi-Newton method) with random initial starting value to minimize $M_n^\mu(\theta)$. Details are given in Algorithm 1.

The idea behind Algorithm 1 is as follows. Due to the fact that the degree of smoothness of g_θ^μ (and hence of $M_n^\mu(\theta)$) decreases as μ vanishes, minimizing $M_n^\mu(\theta)$ becomes increasingly challenging as $\mu \to 0$. To overcome this problem, we propose to iteratively refine the least squares solution by starting with a large initial value for the smoothness parameter ($\mu_0 > 1$ in Algorithm 1; the value 1 is an arbitrary choice) and by decreasing

Algorithm 1: `Computation of the LSE for the smoothed PWA function`

input: data points $\{(X_i, Y_i)\}_{i=1}^n$, number of planes $k_1, k_2 \in \mathbb{N}$, smoothing parameter $\mu > 0$, tolerance $\tau > 0$ for convergence of quasi-Newton method

output: estimate $\hat{\theta}$

1 Determine $m_0 \in \mathbb{N}$ such that $2^{m_0}\mu > 1$;

2 Sample random initial starting value $\hat{\theta}_0 = (a_{1,1}, \ldots, a_{1,k_1}, b_{1,1}, \ldots, b_{1,k_1}, a_{2,1}, \ldots) \in [-r, r]^{(k_1+k_2)(d+1)}$ for some $r > 0$;

3 **for** $m \leftarrow 0$ **to** m_0 **do**

4 Set $\mu_m := 2^{m_0-m}\mu$;

5 Perform quasi-Newton minimization of $M_n^{\mu_m}(\theta)$ with initial starting value $\hat{\theta}_m$, gradient $J_n^{\mu_m}(\theta)$ and tolerance τ; if the quasi-Newton method fails to converge (within a pre-set number of steps) then restart from line 2;

6 Set minimum found by quasi-Newton step as $\hat{\theta}_{m+1}$;

7 **end**

8 **return** $\hat{\theta} := \hat{\theta}_{m_0+1}$;

the smoothness parameter by a factor of two in every iteration. For this we first determine a $m_0 \in \mathbb{N}$ such that $2^{m_0}\mu > 1$, thus making sure that after m_0 iterations, an estimate of θ for the desired value μ, chosen by the user, is obtained. The optimization itself is carried out using a standard quasi-Newton scheme such as *BFGS* (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). In each iteration m, the estimate $\hat{\theta}_m$ from the last iteration serves as a new initial value for the next call of the minimization (quasi-Newton) method in line 5 with decreased value of μ and a user-specified tolerance τ (typically of the order of 10^{-5}). Alternative criteria for termination can also be employed. If any instance of the quasi-Newton method fails to converge or shows numerical instabilities (for instance due to singularity of the Jacobian matrix), the whole algorithm is restarted using a new random initial value. The estimate $\hat{\theta}$ returned by Algorithm 1 (that is the estimate corresponding to the desired degree of smoothness μ) is the minimum found in last minimization (line 5) for $\mu_{m_0} = \mu$.

4 Statistical properties of the LSE: asymptotic normality and efficiency

4.1 The unsmoothed case

In this section we study the statistical properties of the LSE (obtained from the smoothed PWA function) discussed in Section 3. To keep the presentation simple and the technical arguments less cumbersome, we consider the case when $k_1 = 2$ and $k_2 = 0$, i.e., the

regression function g_θ can be expressed as

$$g_\theta(x) = \max\{\alpha^\top x + \gamma, \beta^\top x + \phi\} \quad \text{for } \theta = (\alpha, \beta, \gamma, \phi) \in \mathbb{R}^{2d+2}. \tag{18}$$

Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. (having joint distribution P on $\mathbb{R}^d \times \mathbb{R}$) from the regression model

$$Y = g_\theta(X) + \epsilon,$$

where Y is the response variable, $X \in \mathcal{X} \subset \mathbb{R}^d$ is the predictor, ϵ is the unobserved error such that $\mathbb{E}(\epsilon|X) = 0$ almost everywhere (a.e.) and has finite variance σ^2, and g_θ has the form given in (18). We assume that the unknown parameter $\theta = (\alpha, \beta, \gamma, \phi) \in \Theta$, where Θ is a compact set in $\mathbb{R}^{2d+2}$.

Let θ_0 be the true value of the parameter θ, which we assume lies in the interior of Θ. Our goal is to estimate θ_0 from the observed data by using the method of least squares:

$$\hat{\theta}_n := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n [Y_i - g_\theta(X_i)]^2. \tag{19}$$

Computing the least squares estimate (19) is challenging in practice. This is due to the lack of smoothness of the functions in the class $\mathcal{G} := \{g_\theta : \theta \in \Theta\}$ under consideration. To remedy this problem, the following subsection investigates the smoothed approximation of the class $\mathcal{G}$ of Nesterov (2005) which allows to compute a gradient for any $g_\theta \in \mathcal{G}$ and hence the use of quasi-Newton methods to compute the least squares estimate (19).

We first aim to establish that the least squares estimator of θ_0 for the unsmoothed class of functions $\mathcal{G}$ is $\sqrt{n}$ consistent and asymptotically normal. This result is summarized in the next lemma and relies on the following assumption.

Assumption 1 (Moment conditions)**.** $\mathbb{E}[\|X\|] < \infty$, $\mathbb{E}[|Y|\|X\|] < \infty$, $\mathbb{E}[Y^2] < \infty$. *We further assume that X does not put all its mass on any hyperplane in $\mathbb{R}^d$.*

Lemma 5. *Under Assumption 1, the following statements hold true.*

1. $\hat{\theta}_n - \theta_0 = O_P(n^{-1/2})$.
2. *As $n \to \infty$, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to a multivariate Normal$(0, V^{-1}WV^{-1})$ distri-*

bution, where

$$W = 4\sigma^2 \int_{g_{\theta_0}(x)=\alpha_0^\top x+\gamma_0} \begin{pmatrix} xx^\top & x & 0 & 0 \\ x^\top & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} dP(x)$$
$$+ 4\sigma^2 \int_{g_{\theta_0}(x)=\beta_0^\top x+\phi_0} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & xx^\top & x \\ 0 & 0 & x^\top & 1 \end{pmatrix} dP(x)$$

and $V = W/(2\sigma^2)$.

Proof. The rates of convergence can be calculated similarly to Proposition 6.4.4 in van de Geer (1988). The limiting distribution follows along the same argument as the one in Example 6.6 of van de Geer (1988). □

4.2 Asymptotic results for the class of smoothed PWAs

We aim to compute the LSE in (19). However, as g_θ is a non-smooth function (of θ) we consider a smooth surrogate of g_θ. For $\mu > 0$ let us define the following smooth approximation of g_θ:

$$g_\theta^\mu(x) = \mu \log\left(e^{\frac{\alpha^\top x+\gamma}{\mu}} + e^{\frac{\beta^\top x+\phi}{\mu}}\right) - \mu \log 2 \quad \text{for } \theta = (\alpha, \beta, \gamma, \phi) \in \Theta \subset \mathbb{R}^{2d+2}.$$

Here are a few important facts about g_θ and g_θ^μ (Nesterov, 2005):

(i) $\sup_{x\in\mathbb{R}^d} |g_\theta(x) - g_\theta^\mu(x)| \leq \mu \log 2$, see (10).

(ii) The function g_θ^μ is continuously differentiable with a gradient that is Lipschitz, given by $\mu^{-1} \max_{\|x\|_{d+1}=1} \langle[\alpha, \gamma], x\rangle^2 + \langle[\beta, \phi], x\rangle^2$ (see Lemma 2).

Suppose that $\{\mu_n\}_{n\geq 1}$ is a sequence of positive numbers such that $\mu_n = o(1)$. Similarly to (19), we estimate θ_0 using the method of least squares:

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g_\theta^{\mu_n}(X_i)]^2. \tag{20}$$

For the asymptotic results presented in the following theorem to hold true, amongst others, it is necessary that the amount of smoothing decreases at a $o\left(n^{-1/2}\right)$ rate, summarized in the next two assumptions:

Overwrite θ here, or indicate dependence on μ? Would overwrite.

Assumption 2 (Parameter space)**.** *Let* $\Theta \subset \mathbb{R}^{2d+2}$ *be a compact set such that* $\theta_0 = (\alpha_0, \beta_0, \gamma_0, \phi_0)$ *belongs to the interior of* Θ*. We assume* $(\alpha_0, \beta_0) \neq (\gamma_0, \phi_0)$*.*

Assumption 3 (Order of μ_n)**.** *Suppose* $\{\mu_n\}_{n\geq 1}$ *is a sequence of constants such that* $\mu_n = o(n^{-1/2})$*.*

The main result is then summarized in the next theorem.

Theorem 6. *Under Assumptions 1–3, the following holds true.*

1. *The least squares estimator is consistent and satisfies* $\hat{\theta}_n - \theta_0 = O_P(n^{-1/2})$*.*
2. *We have*
$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, V^{-1}WV^{-1}),$$
where $V := 2P[\dot{g}_{\theta_0}\dot{g}_{\theta_0}^\top] = \ldots$ (*fill this*) *and* $W = P(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top) = 4P[\epsilon^2\dot{g}_{\theta_0}\dot{g}_{\theta_0}^\top]$*. In particular, when* ϵ *is independent of* X *with variance* $\sigma^2 > 0$*, then* $W = 2\sigma^2 V$*.*

5 Experimental results

This section evaluates the proposed approach (Section 2) from a numerical perspective. We start by exemplarily showing how our proposed approach performs in parameter estimation of a broken stick and a broken plane model (Section 5.1). Further numerical evaluations address the dependence on the smoothing parameter μ (Section 5.3, the ability of Algorithm 1 to find successfully minimize (19), as well as an extension to estimation of PWAs with $k = 3$ components (Section 5.6). In the entire section we used the implementation of the Newton method provided by the `optim` function in R, obtained by setting its parameter *method* to "BFGS". Initial values for BFGS (the parameter r in Algorithm 1 were always generated in the interval $[-1, 1]$.

5.1 Examples of parameter estimation in a broken stick and broken plane model

We exemplarily show how Algorithm 1 can be used in practice to estimate regression parameters. We generated two parameter vectors, one for a broken stick and one for a broken plane model with two components, respectively. We then simulated $n = 200$ points from each model by uniformly selecting points on the lines (planes) and by adding iid. normal noise with mean 0 and standard deviation 0.1.

We used Algorithm 1 as given in Section 3 in connection with squared error loss smoothing and a smoothing parameter $\mu = 0.1$ to compute fitted parameters.

Figure 3 shows the true (generated) model (darkred), its smooth Nesterov (2005) approximation (red) as well as the randomly generated points (blue). The best fit obtained

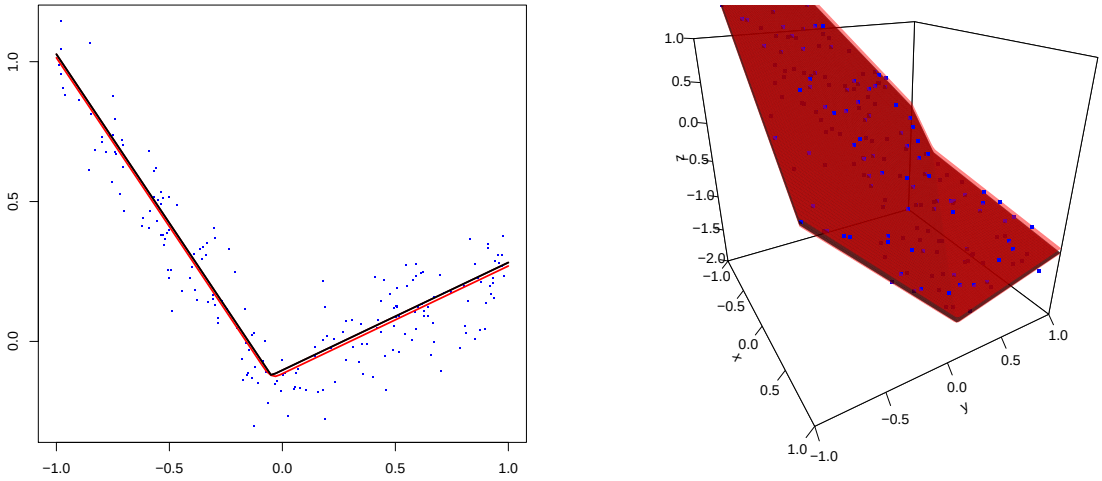

Figure 3: Random points (blue) generated from a broken line (left) and broken plane (right) model with two components. True model (darkred), Nesterov (2005) smooth approximation (red) with $\mu = 0.1$ and estimated model (black) from the data.

with Algorithm 1 (black) shows that the original line (plane) can be reasonably well recovered in both cases. We will elaborate on the quality and computability of such approximations in the following sections.

5.2 Comparison of Algorithm 1 to a gradient free method

We compare the performance of Algorithm 1 to the popular gradient free method of Nelder and Mead (1965). For this we generate two planes in dimensions $d \in \{2, \ldots, 4\}$ and make sure that these planes intersect in the $[-1, 1]^d$ cube with at least a 90 degree angle in order to avoid singular cases. We then generate $n = 10^d$ points from each model in dimension d and apply the Nelder and Mead (1965) method as well as Algorithm 1 with squared error prox function and fixed smoothing parameter $\mu = 0.1$.

We evaluate the quality of the fit using the empirical norm criterion as in Example 1.

Figure 4 shows results. We see that in comparison to a gradient free method, using Nesterov (2005) smoothing to obtain gradients in Algorithm 1 yields a fraction of the computational runtime as the dimension increases. The empirical norm increases for both methods as the dimension increases, although the one of Algorithm 1 seems to consistently stay around one order of magnitude below the one of the Nelder and Mead (1965) method (note that the time axis has a log scale).

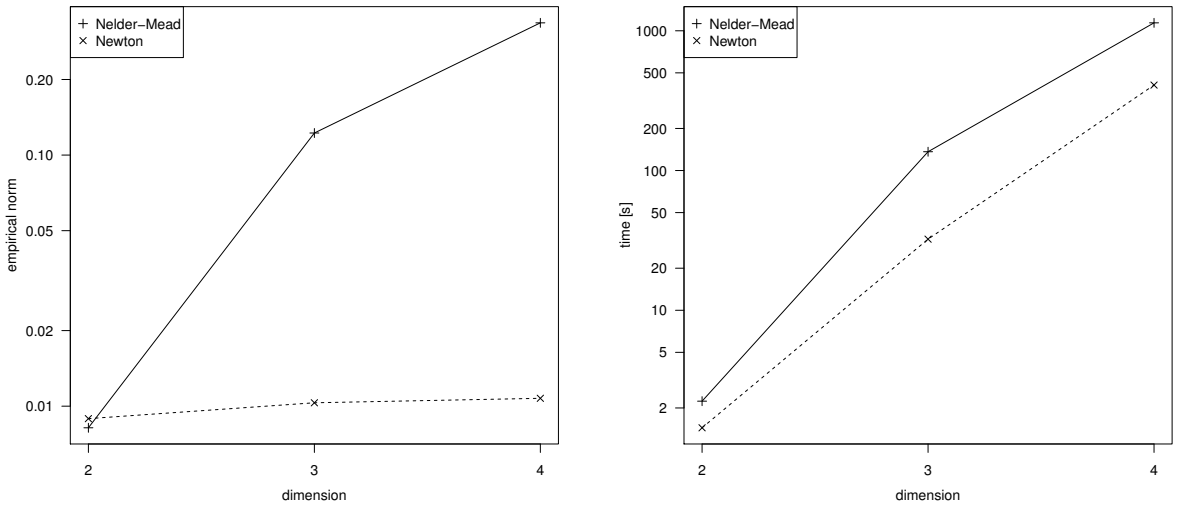

Figure 4: Comparison of Algorithm 1 with the Nelder and Mead (1965) method in terms of average empirical norm (left) and computational time (right) for the best fit. Log scale on the time axis.

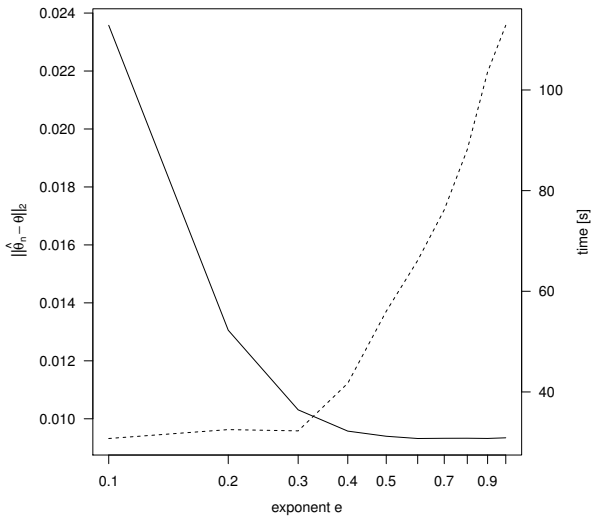

Figure 5: Progression of $\left|\hat{\theta}_n - \theta_0\right|$ (solid line) as a function of $\mu = n^{-e}$ and corresponding time (dashed line) required for its computation. The number of points $n = 1000$ was fixed and the exponent e was varied.

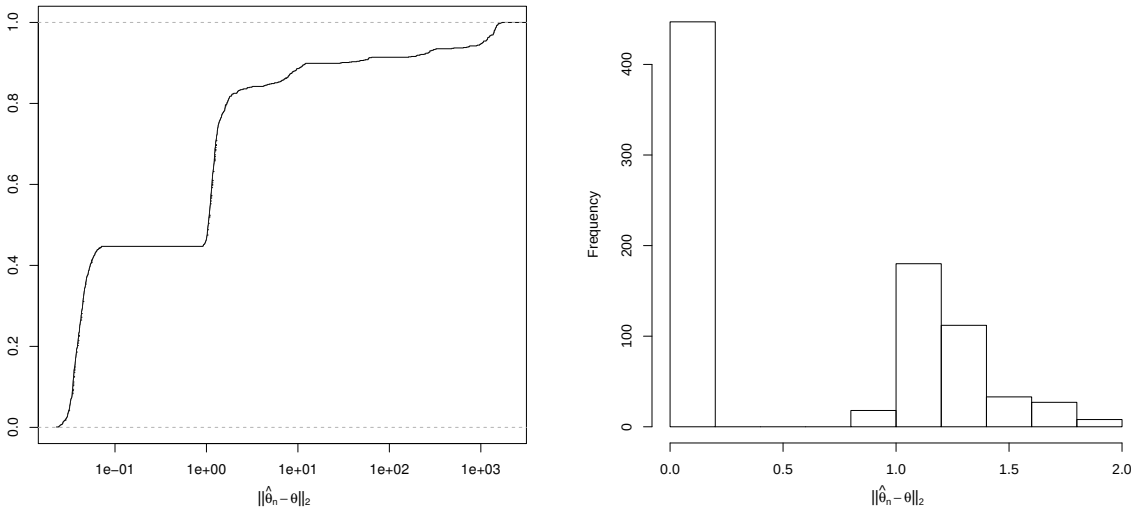

Figure 6: Broken line regression with Algorithm 1 with squared error prox function and $\mu = 0.1$. Results based on 1000 repetitions. Left: Ecdf of $\left|\hat{\theta}_n - \theta_0\right|$, the deviation of the estimated from the generated parameters in the l_2 norm. Right: Histogram of $\left|\hat{\theta}_n - \theta_0\right|$ after truncation at 0.1.

5.3 Dependence on the smoothing parameter

We further evaluate the dependence of the smoothing parameter on the performance of Algorithm 1. For this we fix a parameter vector θ of a plane in two dimensions, generate $n = 1000$ points on it and repeatedly fit a two plane segmented model to the data points using Algorithm 1 with squared error prox loss function. The smoothing parameter was chosen as $\mu = n^{-e}$, where the exponent $e \in \{0.1, 0.2, \ldots, 1\}$ ranged from 0.1 to 1 in steps of 0.1.

For each μ, we repeatedly fit the plane fixed at the start of the experiment a total number of $p = 100$ times. From this pool we select the best estimate $\hat{\theta}_n$ measured in terms of the empirical norm and record its deviation $\left|\hat{\theta}_n - \theta\right|$ from θ as well as the computational time needed to compute $\hat{\theta}_n$. This procedure is repeated a total number of 100 times in order to average both the deviation as well as the computational time.

Figure 5 shows averages of both $\left|\hat{\theta}_n - \theta\right|$ (solid line) as well as of the computational time (dashed line). As expected, the quality of the recovered plane increases (ie. the deviation of the estimate $\hat{\theta}_n$ from the true parameters θ goes to zero) as the parameter μ goes to zero. Likewise, the computational runtime to obtain estimates of this quality increases.

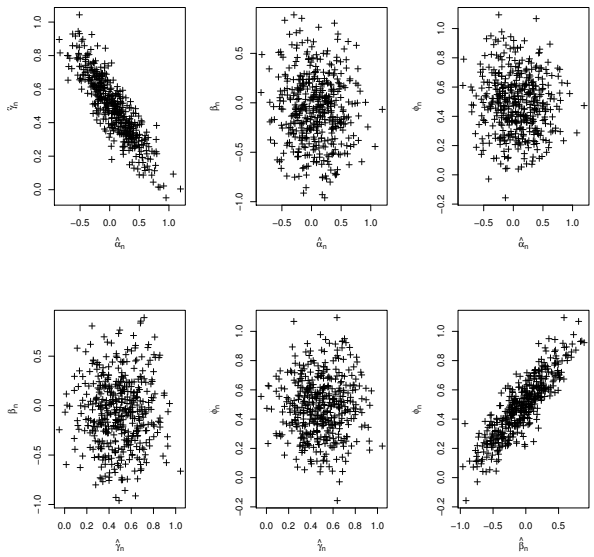

Figure 7: Projections of all six pairs of parameter estimates of $\theta = (\alpha, \gamma, \beta, \phi) \in \mathbb{R}^4$. Plot only shows estimates lying in the truncated range $[-1, 1]$.

5.4 Assessment of Newton's method and local minimization

Figure 6 investigates how likely the Newton method in Algorithm 1 is able to find the global minimum in the case of a broken line regression (dimension one) with two components.

To this end we fix a broken plane model in one dimension and repeatedly apply Algorithm 1 using Nesterov (2005) smoothing in connection with the squared error loss prox function and smoothing parameter $\mu = 0.1$. All results are based on 1000 repetitions.

Figure 6 (left) shows the ecdf of $\left|\hat{\theta}_n - \theta_0\right|$ based on all 1000 runs, that is the deviation of the estimated from the generated parameters in the l_2 norm. As can be seen from the figure, the Newton method converges to an estimated set of parameters very close to the true parameters in l_2 norm roughly in 90% of all cases. A separate histogram for all estimates $\hat{\theta}_n$ satisfying $\left|\hat{\theta}_n - \theta_0\right| < 0.1$ is displayed in Figure 6 (right).

Even though the Newton method converged to a very good approximation in most of the cases, the remaining estimates deviate from the (true) underlying parameters θ in l_2 norm by more than 10^3. An investigation of these cases showed that due to the particular implementation of the Newton method in the R function `optim`, failure of converge results in the Newton run being aborted by some numerical criterion leading to value at the boundary of the parameter domain (of absolute magnitude up 10^3 to 10^4).

Figure 7 shows all six projections of two parameter estimates of $\theta = (\alpha, \gamma, \beta, \phi) \in \mathbb{R}^4$ in the xy-plane for the 1000 runs already analyzed in Figure 6. To capture the majority of estimates actually taking values of around zero, Figure 7 only shows estimates lying in the truncated range of $[-1, 1]$. It can be seen from the plot that a well defined optimum

Parameter	Siegmund and Zhang (1994)		Algorithm 1	
	coverage prob.	length	coverage prob.	length
θ_0	0.879	0.169	0.913	0.184
a_1	0.954	0.208	0.963	0.225
b_1	0.959	0.119	0.972	0.128
a_2	0.957	0.219	0.976	0.242
b_2	0.951	0.128	0.969	0.141

Table 2: Coverage probabilites and lengths of the confidence intervals for $\theta_0 = (a_1, b_1, a_2, b_2)$ and its four components. Model consisting of two lines in one dimension (one parametrized in (a_1, b_1) and the other in (a_2, b_2)). Exact method of Siegmund and Zhang (1994) and Algorithm 1.

exists.

5.5 Assessment of coverage probabilites

We assess the accuracy of confidence intervals computed for all parameters, measured in terms of their coverage probabilites and lengths. To this end, we compare confidence intervals computed with Algorithm 1 to the ones computed with the exact method of Siegmund and Zhang (1994).

Our setup is as follows: We fix a PWA consisting of two lines in one dimension, characterized through a θ_0. We then repeat the following procedure $R = 1000$ times:

1. Using Algorithm 1 (with $\mu = 0.01$) and the exact method by Siegmund and Zhang (1994), we compute the best estimate of $\hat{\theta}_n$, measured with respect to the empirical norm, from a pool of 10 repetitions. For each of the 10 repetitions, we generate 200 points on the PWA. For the method of Siegmund and Zhang (1994), we determine the change point in $[-1, 1]$ of the two lines via a grid search with 1000 equidistant points in $[-1, 1]$.

2. For the best estimate determined in the previous step, we store $\hat{\theta}_n$, the set P of 200 points generated on the PWA as well as the empirical covariance matrix C computed with $\hat{\theta}_n$ and P. The matrix C is computed using plug-in estimates for the integrals in matrix W of Lemma 5, where the points in P are divided up into two subsets corresponding to the two sides of the estimated change-point specified in $\hat{\theta}_n$.

3. Finally, a normal confidence interval is computed for the i'th component of θ_0 as $\hat{\theta}_n^{(i)} \pm 1.96\sqrt{C_{i,i}/N_i}$ for all $i \in \{1, \ldots, 4\}$, where $\hat{\theta}_n^{(i)}$ is the i'th entry of the estimate $\hat{\theta}_n$ and N_i is the number of points in P which fall into the line segment defined through $\hat{\theta}_n^{(i)}$.

In each repetition of the above procedure we record (1) the number of times each of the four confidence intervals contains the truth in θ_0 as well as the number of times all

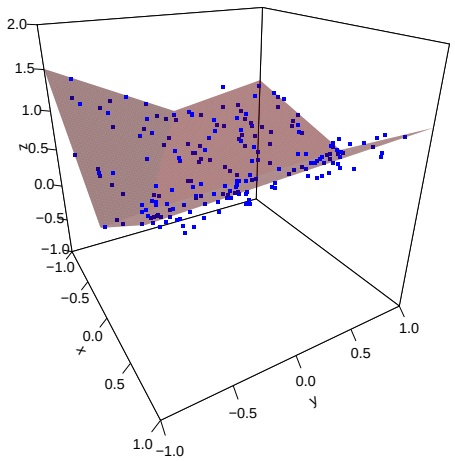

Figure 8: Fit of parameters for a segmented plane consisting of three components. Algorithm 1 with squared error prox smoothed function. Blue points are generated points from the true model (red), fitted segmented planes in black.

confidence intervals simultaneously contains all entries of θ_0 (this allows us to compute coverage probabilites) and (2) the length of all confidence intervals.

Results are shown in Table 5.5. The table demonstrates that for the individual components of θ_0, the exact method of Siegmund and Zhang (1994) yields confidence intervals which keep the 95% coverage while being only slightly conservative. Algorithm 1 yields individual confidence intervals which are slightly too wide, hence resulting in higher lengths than the exact ones and coverage probabilites which exceed 95%. With respect to the simultaneous coverage for the entire vector θ_0, the method of Siegmund and Zhang (1994) yields a confidence interval with a lower coverage probability than the one of Algorithm 1.

5.6 Extension to three and more planes

6 Discussion

7 Acknowledgements

The authors would like to thank the editor, associate editor and the two referees for their helpful comments.

A Proofs of Section 4.2

This section proves that the least squares estimate of the entropy smoothed functions g_θ^μ are $\sqrt{n}$ consistent and asymptotically normal. This is the result of Theorem 6. It will be proven by means of the following lemmas and theorems.

Lemma 7. *Under Assumptions 1-3 we have* $\hat{\theta}_n - \theta_0 = o_p(1)$.

Proof. For $\theta \in \Theta$, let us define

$$M(\theta) := P[(Y - g_\theta(X))^2] \qquad \text{and} \qquad M_n(\theta) = P[(Y - g_\theta^{\mu_n}(X))^2].$$

Also, let

$$\mathbb{M}_n(\theta) := \mathbb{P}_n[(Y - g_\theta^{\mu_n}(X))^2] = \frac{1}{n}\sum_{i=1}^{n}[Y_i - g_\theta^{\mu_n}(X_i)]^2.$$

We will apply Theorem 3.2.3 of VdV&W to prove this result. In particular, we will show that $\sup_{\theta\in\Theta} |\mathbb{M}_n(\theta) - M(\theta)| \to 0$ in probability. Observe that

$$\mathbb{M}_n(\theta) - M(\theta) = [\mathbb{M}_n(\theta) - M_n(\theta)] + [M_n(\theta) - M(\theta)].$$

Now,

$$\begin{aligned}
|(M_n - M)(\theta)| &= \left|P[(Y - g_\theta^{\mu_n}(X))^2] - P[(Y - g_\theta(X))^2]\right| \\
&= |P\left[\{g_\theta^{\mu_n}(X) - g_\theta(X)\} \cdot \{-2Y + g_\theta^{\mu_n}(X) + g_\theta(X)\}\right]| \\
&= |P\left[\{g_\theta^{\mu_n}(X) - g_\theta(X)\} \cdot \{-2g_{\theta_0}(X) + g_\theta^{\mu_n}(X) + g_\theta(X)\}\right]| \\
&\leq P\left[|g_\theta^{\mu_n}(X) - g_\theta(X)| \cdot | - 2g_{\theta_0}(X) + g_\theta^{\mu_n}(X) + g_\theta(X)|\right] \\
&\leq (\mu_n \log 2)\ P[|g_\theta^{\mu_n}(X) - g_\theta(X)| + 2|g_\theta(X) - g_{\theta_0}(X)|] \\
&\leq (\mu_n \log 2)\ P[(\mu_n \log 2) + 2G(X)\|\theta - \theta_0\|] \qquad (21)
\end{aligned}$$

where we have used the facts:

$$\sup_{x\in\mathbb{R}^d} |g_\theta(x) - g_\theta^{\mu_n}(x)| \leq \mu_n \log 2, \qquad \text{for all } \theta \in \Theta,$$

and

$$|g_\theta(x) - g_{\theta_0}(x)| \leq G(x)\|\theta - \theta_0\|, \qquad (22)$$

for $G(x) = \|(1, x)\|$ (see Lemma 10). Thus, by Assumptions 1 and 3,

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{a.s.} 0 \qquad \text{as } \ n \to \infty.$$

Now, for any $\eta > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|\mathbb{M}_n(\theta) - M_n(\theta)| > \eta\right) \leq \frac{1}{\eta}\mathbb{E}\left[\sup_{\theta\in\Theta}|\mathbb{M}_n(\theta) - M_n(\theta)|\right] \leq J_{[\,]}(1)\|F\|_{P,2},$$

where $J_{[\,]}(1) :=$

□

The following result gives the rate of convergence of the estimator $\hat{\theta}_n$.

Theorem 8. *Under Assumptions 1-3 we have* $\hat{\theta}_n - \theta_0 = O_p(n^{-1/2})$.

Proof. We will apply Theorem 3.2.5 of VdV and W to prove the result.

Note that θ_0 minimizes $M(\theta)$, for $\theta \in \Theta$ (show this). Hence, as $M(\theta)$ is strictly convex around θ_0 (show this), we have

$$M(\theta) - M(\theta_0) \geq c\|\theta - \theta_0\|^2, \qquad \text{for } \theta \text{ in a neighborhood of } \theta_0,$$

for some $c > 0$.

We now have to bound

$$\mathbb{E}\left[\sup_{\|\theta-\theta_0\|<\delta}\sqrt{n}|(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta_0)|\right]. \tag{23}$$

Observe that

$$\sqrt{n}|(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta_0)| \leq S_n(\theta) + T_n(\theta)$$

where

$$\begin{aligned} S_n(\theta) &:= \sqrt{n}|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_0)| \\ T_n(\theta) &:= \sqrt{n}|(M_n - M)(\theta) - (M_n - M)(\theta_0)|. \end{aligned}$$

The stochastic process $S_n(\theta)$ can be bounded using metric entropy calculations (do this). Next we have to control the deterministic process $T_n(\theta)$. Similarly, we can show that

$$\begin{aligned} |(M_n - M)(\theta_0)| &= \left|P[(Y - g_{\theta_0}^{\mu_n}(X))^2] - P[(Y - g_{\theta_0}(X))^2]\right| \\ &= P\left[|g_{\theta_0}^{\mu_n}(X) - g_{\theta_0}(X)|^2\right] \qquad (24) \\ &\leq \mu_n^2(\log 2)^2. \qquad (25) \end{aligned}$$

Therefore, the expected supremum in (23) can be upper bounded by

$$\begin{aligned} & \mathbb{E}\left[\sup_{\|\theta-\theta_0\|<\delta} S_n(\theta)\right] + \mathbb{E}\left[\sup_{\|\theta-\theta_0\|<\delta} T_n(\theta)\right] \\ \leq \; & c_0\delta + \sqrt{n}[c_1\mu_n^2 + c_2\mu_n\delta] =: \phi_n(\delta), \end{aligned}$$

where c_0 is ... (find this using the metric entropy result), $c_1 = 2(\log 2)^2$ and $c_2 = 2(\log 2)P[G(X)]$. Note that $\phi_n(\delta)/\delta^\alpha$ is a decreasing function for any $1 < \alpha < 2$. Then with $r_n := n^{1/2}$, we have

$$r_n^2\phi_n(r_n^{-1}) = c_0 r_n + c_1\sqrt{n}r_n^2\mu_n^2 + c_2 r_n\mu_n \lesssim \sqrt{n}, \qquad \text{for every } n,$$

as $r_n\mu_n = O(1)$.

As $\hat{\theta}_n \to_p \theta_0$ (by Lemma 7) then by Theorem 3.2.5 of VdV&W we have

$$n^{1/2}(\hat{\theta}_n - \theta_0) = O_p(1).$$

□

Let us define

$$m_\theta(x, y) := (y - g_\theta(x))^2.$$

Observe that $(x, y) \mapsto m_\theta(x, y)$ is a measurable function for each $\theta \in \Theta$ and that $\theta \mapsto m_\theta(x, y)$ is differentiable at θ_0 for P-a.e. x with derivative

$$\dot{m}_{\theta_0}(x, y) = -2(y - g_{\theta_0}(x))\dot{g}_{\theta_0}(x).$$

Theorem 9. *Under assumptions 1-3 we have*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, V^{-1}WV^{-1}),$$

where

$$V := 2P[\dot{g}_{\theta_0}\dot{g}_{\theta_0}^\top] = \ldots \qquad (\textit{fill this}) \tag{26}$$

and

$$W = P(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top) = 4P[\epsilon^2\dot{g}_{\theta_0}\dot{g}_{\theta_0}^\top].$$

In particular, when ϵ is independent of X with variance $\sigma^2 > 0$, then $W = 2\sigma^2 V$.

Proof. For the convenience of notation, let us write

$$m_\theta^\mu(x, y) := (y - g_\theta^\mu(x))^2, \quad (\text{for } \mu > 0).$$

We will study the stochastic process

$$\tilde{\mathbb{M}}_n(h) := n\mathbb{P}_n(m^{\mu_n}_{\theta_0+hn^{-1/2}} - m^{\mu_n}_{\theta_0}), \qquad \text{for } \; h \in \mathbb{R}^{2d+2}.$$

Observe that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \operatorname*{argmin}_h \tilde{\mathbb{M}}_n(h).$$

We will show that

$$\tilde{\mathbb{M}}_n(h) \xrightarrow{d} h^\top \Delta + \frac{1}{2} h^\top V h =: \tilde{\mathbb{M}}(h) \;\text{ in } \ell^\infty(\{h : \|h\| \le K\})$$

for every $K > 0$, where $\Delta \sim N_d(0, W)$. Then the conclusion of the theorem follows from the argmax (argmin) continuous theorem (see e.g., Theorem 3.2.2 in VdV&W) upon noticing that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \operatorname*{argmin}_h \tilde{\mathbb{M}}_n(h) \xrightarrow{d} \operatorname*{argmin}_h \tilde{\mathbb{M}}(h) = -V^{-1}\Delta \sim N_d(0, V^{-1}WV^{-1}).$$

First observe that

$$\begin{aligned} n\mathbb{P}_n(m^{\mu_n}_{\theta_0+hn^{-1/2}} - m^{\mu_n}_{\theta_0}) \;&=\; n\mathbb{P}_n(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) \\ &\quad + n\mathbb{P}_n[(m^{\mu_n}_{\theta_0+hn^{-1/2}} - m^{\mu_n}_{\theta_0}) - (m_{\theta_0+hn^{-1/2}} - m_{\theta_0})]. \end{aligned} \tag{27}$$

Study of the second term on the right-hand side of (27)**:** The second term on the right-hand side of the above display can also be expressed as

$$n(\mathbb{M}_n - M)(\theta_0 + hn^{-1/2}) - n(\mathbb{M}_n - M)(\theta_0).$$

Using a similar expansion as in (21) with $\theta = \theta_0 + hn^{-1/2}$ (and $\mathbb{P}_n$ instead of P) we have

$$|n(\mathbb{M}_n - M)(\theta_0 + hn^{-1/2})| \le n\mu_n^2(\log 2)^2 + 2n\mu_n(\log 2)\mathbb{P}_n[G(X)]\|n^{-1/2}h\|$$

which converges uniformly to 0 a.s. in $\ell^\infty(\{h : \|h\| \le K\})$ as $n^{-1/2}\mu_n \to 0$ and $\mathbb{P}_n[G(X)] \le P[G(X)] + 1 < \infty$ for all large n a.s. (by Assumption 1). Now using a similar calculation as in (25), we have

$$|n(\mathbb{M}_n - M)(\theta_0)| \le n\mu_n^2(\log 2)^2 \to 0, \qquad \text{as } \; n \to \infty.$$

Thus, we have shown that the second term in (27) goes to zero (a.s.) uniformly in $T_K := \{h \in \mathbb{R}^{2d+2} : \|h\| \le K\}$.

Study of the first term on the right-hand side of (27)**:** Observe that

$$\begin{aligned} n\mathbb{P}_n(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) &= \sqrt{n}(\mathbb{P}_n - P)[\sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0})] \\ &\quad + nP(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}). \end{aligned} \tag{28}$$

By a second order Taylor expansion of $M(\theta) := P[m_\theta]$ about θ_0, we have

$$P[m_\theta] - P[m_{\theta_0}] = P[(g_\theta - g_{\theta_0})^2] = \frac{1}{2}(\theta - \theta_0)^\top V(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

where V is defined in (26). Thus, the second term of the right side of the (28) converges to $(1/2)h^\top Vh$ uniformly on T_K.

To handle the first term in (28) we use the Donsker's theorem with classes of functions changing with n; see e.g., Theorem 2.11.23 of VdV&W. For $K > 0$ fixed, we consider the following class of functions

$$\mathcal{F}_n := \{\sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) : \|h\| \leq K\}.$$

Notice that for all θ_1, θ_2 in a neighborhood $\mathcal{N}$ of θ_0,

$$|m_{\theta_1}(x,y) - m_{\theta_2}(x,y)| = |g_{\theta_1}(x) - g_{\theta_2}(x)| \cdot |-2y + g_{\theta_1}(x) + g_{\theta_2}(x)| \leq F(x,y)\|\theta_1 - \theta_2\|$$

where

$$F(x,y) := G(x)\left[2|y| + \sup_{\theta_1,\theta_2 \in \mathcal{N}} |g_{\theta_1}(x) + g_{\theta_2}(x)|\right]$$

and G is defined after (22). Further note that $F \in L_2(P)$ (by Assumption 1).

Thus, the function class $\mathcal{F}_n$ has envelope $F_n(x,y) \equiv KF(x,y)$ for all n, and since $F(x,y) \in L_2(P)$ the Lindeberg conditions (see Equation 2.11.21 of VdV&W) are satisfied easily. Furthermore, for $s, t \in \mathbb{R}^{2d+2}$ such that $\|s\|, \|t\| < K$, define

$$f_{n,s} := \sqrt{n}(m_{\theta_0+sn^{-1/2}} - m_{\theta_0}), \quad \text{and} \quad f_{n,t} := \sqrt{n}(m_{\theta_0+tn^{-1/2}} - m_{\theta_0}).$$

By the dominated convergence theorem the covariance functions satisfy

$$P(f_{n,g}f_{n,t}) - P(f_{n,s})P(f_{n,t}) \to P(s^\top \dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top t) = s^\top Wt.$$

Lastly, to apply Theorem 2.11.23 of VdV&W we need to that the bracketing entropy condition holds. To this end, observe that

$$N_{[\,]}(2\eta\|F_n\|_{P,2}, \mathcal{F}_n, L_2(P)) \leq N(\eta, T_K, \|\cdot\|) \leq \left(\frac{CK}{\eta}\right)^d,$$

where the first inequality follows from Theorem 2.7.11 of VdV&W (classes that are Lip-

schitz in a parameter) and the second inequality follows from an upper bound on the η-covering number of an Euclidean ball in $\mathbb{R}^{2d+2}$. Thus,

$$\int_0^\delta \sqrt{N_{[\,]}(2\eta\|F_n\|_{P,2}, \mathcal{F}_n, L_2(P))}d\eta \lesssim \int_0^\delta \sqrt{d\log\left(\frac{CK}{\eta}\right)}d\eta \to 0 \qquad \text{as } \delta \to 0,$$

and hence the bracketing entropy condition holds. We conclude that $\tilde{\mathbb{M}}_n(h)$ converges weakly to $h^\top\Delta$ in $\ell^\infty(\{h : \|h\| \leq K\})$, and the desired result holds. □

The proof of Theorem 6 is merely a summary of the aforementioned results:

Proof of Theorem 6. The first statement follows from Lemma 7 and Theorem 8. The second statement follows from Theorem 9. □

A.1 Additional lemmas

Lemma 10. *It holds true that* $|g_\theta(x) - g_{\theta_0}(x)| \leq G(x)\|\theta - \theta_0\|$.

Proof. Consider $|g_\theta(x) - g_{\theta_0}(x)|$ for a fixed x and four cases depending on whether $g_\theta(x) = \max\{\alpha^\top x + \gamma, \beta^\top x + \phi\}$ takes its maximum in the first or second argument.

1. If $g_\theta(x) = \alpha^\top x + \gamma$, $g_{\theta_0}(x) = \alpha_0^\top x + \gamma_0$, then immediately $|g_\theta(x) - g_{\theta_0}(x)| \leq (\|x\| + 1)\|\theta - \theta_0\|$ by triangle and Cauchy-Schwarz inequalities. The norm is the Euclidean norm.

2. The case $g_\theta(x) = \alpha^\top x + \gamma$, $g_{\theta_0}(x) = \beta_0^\top x + \phi_0$. If $\alpha^\top x + \gamma \leq \beta_0^\top x + \phi_0$ then $|g_\theta(x) - g_{\theta_0}(x)| \leq |(\beta^\top x + \phi) - (\beta_0^\top x + \phi_0)| \leq (\|x\| + 1)\|\theta - \theta_0\|$ due to the fact that $\beta^\top x + \phi \leq \alpha^\top x + \gamma$. Likewise, if $\alpha^\top x + \gamma > \beta_0^\top x + \phi_0$ then using $\beta_0^\top x + \phi_0 \geq \alpha_0^\top x + \gamma_0$ one obtains $|g_\theta(x) - g_{\theta_0}(x)| \leq |(\alpha^\top x + \gamma) - (\alpha_0^\top x + \gamma_0)| \leq (\|x\| + 1)\|\theta - \theta_0\|$.

3. The case $g_\theta(x) = \beta^\top x + \phi$, $g_{\theta_0}(x) = \alpha_0^\top x + \gamma_0$ follows analogously to the previous one.

4. The case $g_\theta(x) = \beta^\top x + \phi$, $g_{\theta_0}(x) = \beta_0^\top x + \phi_0$ follows analogously to the first one.

Combined this shows that $|g_\theta(x) - g_{\theta_0}(x)| \leq G(x)\|\theta - \theta_0\|$ with $G(x) = \|x\| + 1$. □

According to van der Vaart and Wellner (2000), the quantity

$$S_n(\theta) := \sqrt{n}|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_0)|$$

is bounded by

$$E^* \int_0^{\theta_n(\delta)} \sqrt{\log N(\epsilon, M_{n,\delta}, L_2(\mathbb{P}_n)}d\epsilon.$$

Lemma 11. *Assuming* $\mathbb{E}(X^2) < \infty$, $H(\delta, \mathbb{G}_n(R), Q_n) \sim \log\left(\frac{C+\delta}{\delta}\right)$ *for a constant* C *independent of* μ.

Proof. Define

$$\mathbb{G}' = \left\{ g_\theta(x) = \frac{\alpha^T x + \gamma}{\mu} : \mathbb{R}^d \to \mathbb{R} \mid (\alpha, \beta, \gamma, \phi) \in \Theta \right\}.$$

Let $\{c_j\}_{j=1}^N$ be a covering of Θ with balls $B_j := B(c_j, \epsilon)$ centered at c_j with radius ϵ, and for each $c_j = (\alpha_j, \beta_j, \gamma_j, \phi_j)$, define the corresponding function in $\mathbb{G}'$ as $C_j(x) := g_{c_j}(x)$.

Let $\delta > 0$ be given. Then according to (van de Geer, 2009, Lemma 2.5), $\Theta \subseteq \mathbb{R}^{2d+2}$ (since it is bounded by a ball of radius $\tilde{R}$) can be covered by

$$N \leq \left(\frac{4\tilde{R} + \epsilon}{\epsilon} \right)^{2d+2}$$

balls of radius ϵ.

For $g_0 \in \mathbb{G}'$, let $g \in \mathbb{G}'_n(R) = \{g \in \mathbb{G}' : \|g - g_0\|_{Q_n} \leq R\}$, where $\|g_0\|_{Q_n} \leq \mu^{-1}\tilde{R}(M+1)$ using the bounds on $\alpha, \beta, \gamma, \phi$ and M. It follows that $\|g\|_{Q_n} \leq \|g - g_0\|_{Q_n} + \|g_0\|_{Q_n} \leq R + \mu^{-1}\tilde{R}(M+1) =: R_2$. As $g \in \mathbb{G}'$, $g(x) = \mu^{-1}(\alpha^T x + \gamma)$ for some $(\alpha, \beta, \gamma, \phi) \in \Theta$.

For the $(\alpha, \beta, \gamma, \phi) \in \Theta$ defining g, there exist $(\alpha_j, \beta_j, \gamma_j, \phi_j)$ such that $\|(\alpha, \beta, \gamma, \phi) - (\alpha_j, \beta_j, \gamma_j, \phi_j)\| \leq \epsilon$, which in turn implies that $\|\alpha - \alpha_j\| \leq \epsilon$, $\|\beta - \beta_j\| \leq \epsilon$, $|\gamma - \gamma_j| \leq \epsilon$, $|\phi - \phi_j| \leq \epsilon$.

It follows immediately that $|(g - C_j)(x)| \leq \frac{1}{\mu}\epsilon(M+1)$, implying

$$\|g - C_j\|_{Q_n}^2 = \frac{1}{n}\sum_i |(g - C_j)(x_i)|^2 \leq \frac{1}{\mu^2}\epsilon^2(M+1)^2 = \delta^2$$

for the choice $\epsilon = \frac{\mu\delta}{M+1}$. Hence, $\{C_j\}_{j=1}^N$ form a δ-cover of $\mathbb{G}'_n(R_2)$ and

$$H(\delta, \mathbb{G}'_n(R), Q_n) \leq (2d+2)\log\left(\frac{4R_2(\mu)(M+1) + \mu\delta}{\mu\delta} \right),$$

which is of the form $\log\left(\frac{C+\delta}{\delta}\right)$ for a constant C.

Equally, the class

$$\mathbb{G}'' = \left\{ g(x) = \frac{\beta^T x + \phi}{\mu} : \mathbb{R}^d \to \mathbb{R} \mid (\alpha, \beta, \gamma, \phi) \in \Theta \right\}$$

has the same entropy.

Let $F_1 = G'$, $F_2 = G''$ and $\phi(x, y) = f(x, y) = \mu \log\left[\frac{1}{2}(e^x + e^y)\right]$. Next, compute the entropy for the class $\phi(F_1, F_2)$ using (Kosorok, 2008, Lemma 9.13).

Both classes F_1 and F_2 are BUEI, and $\phi : \mathbb{R}^2 \to \mathbb{R}$ is locally Lipschitz on Θ, mean-

ing that it satisfies the requirement $|\phi \circ f(x) - \phi \circ g(x)|^2 \leq c^2 \sum_{j=1}^{k} (f_j(x) - g_j(x))^2 = c^2 |f(x) - g(x)|^2$ of the lemma.

The proof of (Kosorok, 2008, Lemma 9.13) then shows that the covering number of $\phi(F_1, F_2)$ can be bounded by the product of the individual covering numbers, and hence that the entropy of $\phi(F_1, F_2)$ satisfies

$$\begin{aligned}&\log N(\delta H, \phi(F_1, F_2), Q_n)\\ &\leq \log N(\delta \|F_1\|_{Q,2}, F_1, Q_n) + \log N(\delta \|F_2\|_{Q,2}, F_2, Q_n),\end{aligned}$$

for the combined envelope $H = |\phi(f_0)| + c(|f_{01}| + F_1 + |f_{02}| + F_2)$.

Hence $H(\delta, \mathbb{G}_n(R), Q_n) \sim \log\left(\frac{C+\delta}{\delta}\right) + \log\left(\frac{C+\delta}{\delta}\right) \sim \log\left(\frac{C+\delta}{\delta}\right)$. □

A.2 Computation of the covariance matrices for squared error prox

Note that $\theta = (\alpha, \gamma, \beta, \phi)$ (in this order), $m_\theta(x) = (y - g_\theta^\mu(x))^2$ and

$$\frac{\partial}{\partial\theta} m_\theta(x) = (-2)(y - g_\theta^\mu(x)) \left(e^{\frac{\alpha^T x+\gamma}{\mu}} + e^{\frac{\beta^T x+\phi}{\mu}}\right)^{-1} \begin{pmatrix} xe^{\frac{\alpha^T x+\gamma}{\mu}} \\ e^{\frac{\alpha^T x+\gamma}{\mu}} \\ xe^{\frac{\beta^T x+\phi}{\mu}} \\ e^{\frac{\beta^T x+\phi}{\mu}} \end{pmatrix}.$$

Let $A := \left(e^{\frac{\alpha^T x+\gamma}{\mu}} + e^{\frac{\beta^T x+\phi}{\mu}}\right)$. Hence, using $\mathbb{E}(y - g_\theta^\mu(x)) = \epsilon$,

$$\begin{aligned}W = P[\dot{m}_\theta \dot{m}_\theta^\top] &= \int_{-1}^{1} 4(y - g_\theta^\mu(x))^2 A^{-2} \begin{pmatrix} xe^{\frac{\alpha^T x+\gamma}{\mu}} \\ e^{\frac{\alpha^T x+\gamma}{\mu}} \\ xe^{\frac{\beta^T x+\phi}{\mu}} \\ e^{\frac{\beta^T x+\phi}{\mu}} \end{pmatrix} \begin{pmatrix} xe^{\frac{\alpha^T x+\gamma}{\mu}} \\ e^{\frac{\alpha^T x+\gamma}{\mu}} \\ xe^{\frac{\beta^T x+\phi}{\mu}} \\ e^{\frac{\beta^T x+\phi}{\mu}} \end{pmatrix}^\top dx \\ &= 4\epsilon^2 \int_{-1}^{1} A^{-2} \begin{pmatrix} xx^\top e^{2\frac{\alpha^T x+\gamma}{\mu}} & xe^{2\frac{\alpha^T x+\gamma}{\mu}} & xx^\top e^{\frac{\alpha^T x+\gamma}{\mu}} e^{\frac{\beta^T x+\phi}{\mu}} & xe^{\frac{\alpha^T x+\gamma}{\mu}} e^{\frac{\beta^T x+\phi}{\mu}} \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \end{pmatrix} dx\end{aligned}$$

assuming independence of ϵ from X.

Similarly, V is the Jacobi matrix of $\mathbb{E}(y - g_\theta^\mu(x))^2$. Here,

$$\nabla P m_\theta = \mathbb{E}\left((-2)(y - g_\theta^\mu(x)) A^{-1} \left[xe^{\frac{\alpha^T x+\gamma}{\mu}}, e^{\frac{\alpha^T x+\gamma}{\mu}}, xe^{\frac{\beta^T x+\phi}{\mu}}, e^{\frac{\beta^T x+\phi}{\mu}}\right]\right).$$

By the product rule, the entry $(1,1)$ in $V = JPm_\theta$ (Jacobi matrix of Pm_θ) is

$$\int_{-1}^{1}(+2)\left(A^{-1}xe^{\frac{\alpha^T x+\gamma}{\mu}}\right)^2+(-2)\epsilon\left[\frac{x^2}{m}A^{-1}e^{\frac{\alpha^T x+\gamma}{\mu}}-\frac{x^2}{m}A^{-2}e^{2\frac{\alpha^T x+\gamma}{\mu}}\right]dx,$$

the $(1,2)$ entry is

$$\int_{-1}^{1}(+2)x\left(A^{-1}e^{\frac{\alpha^T x+\gamma}{\mu}}\right)^2+(-2)\epsilon\left[\frac{x}{m}A^{-1}e^{\frac{\alpha^T x+\gamma}{\mu}}-\frac{x}{m}A^{-2}e^{2\frac{\alpha^T x+\gamma}{\mu}}\right]dx,$$

the $(1,3)$ entry is

$$\int_{-1}^{1}(+2)xA^{-1}e^{\frac{\alpha^T x+\gamma}{\mu}}xA^{-1}e^{\frac{\beta^T x+\phi}{\mu}}+(-2)\epsilon\frac{-x^2}{m}A^{-2}e^{\frac{\alpha^T x+\gamma}{\mu}}e^{\frac{\beta^T x+\phi}{\mu}}dx,$$

and the $(1,4)$ entry is

$$\int_{-1}^{1}(+2)xA^{-1}e^{\frac{\alpha^T x+\gamma}{\mu}}A^{-1}e^{\frac{\beta^T x+\phi}{\mu}}+(-2)\epsilon\frac{-x}{m}A^{-2}e^{\frac{\alpha^T x+\gamma}{\mu}}e^{\frac{\beta^T x+\phi}{\mu}}dx.$$

Similarly for the other entries.

As $\mu \to 0$, in the region where $\max g_\theta(x) = \alpha^T x+\gamma$, that is where $\alpha^T x+\gamma >> \beta^T x+\phi$ (corresponding to the first two rows of the matrices W and V), A grows asymptotically like $e^{\frac{\alpha^T x+\gamma}{\mu}}$, hence $A^{-1}e^{\frac{\alpha^T x+\gamma}{\mu}}$ goes to one and $A^{-2}e^{\frac{\alpha^T x+\gamma}{\mu}}e^{\frac{\beta^T x+\phi}{\mu}}$ goes to zero.

Comparing with V and W in Lemma 5 shows that the matrices for the unsmoothed case are recovered as $\mu \to 0$.

A.3 Computation of the covariance matrices for squared error prox

Note that $\theta = (\alpha, \gamma, \beta, \phi)$ (in this order), $m_\theta(x) = (y - g_\theta^\mu(x))^2$ and

$$\frac{\partial}{\partial\theta}m_\theta(x) = (-2)(y - g_\theta^\mu(x))\begin{pmatrix} w_1 x \\ w_1 \\ w_2 x \\ w_2 \end{pmatrix}.$$

Hence, using $\mathbb{E}(y - g_\theta^\mu(x)) = \epsilon$,

$$W = P[\dot{m}_\theta \dot{m}_\theta^\top] = \int_{-1}^{1} 4(y - g_\theta^\mu(x))^2 \begin{pmatrix} w_1 x \\ w_1 \\ w_2 x \\ w_2 \end{pmatrix} \begin{pmatrix} w_1 x \\ w_1 \\ w_2 x \\ w_2 \end{pmatrix}^\top dx$$

$$= 4\epsilon^2 \int_{-1}^{1} \begin{pmatrix} w_1^2 x x^\top & w_1^2 x & w_1 w_2 x x^\top & w_1 w_2 x \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \end{pmatrix} dx$$

assuming independence of ϵ from X.

Similarly, V is the Jacobi matrix of $\mathbb{E}(y - g_\theta^\mu(x))^2$. Here,

$$\nabla P m_\theta = \mathbb{E}\left((-2)(y - g_\theta^\mu(x))\left[w_1 x, w_1, w_2 x, w_2\right]\right),$$

and the Jacobi matrix is

$$V = JPm_\theta = \int_{-1}^{1} (+2) \begin{pmatrix} w_1^2 x x^\top & w_1^2 x & w_1 w_2 x x^\top & w_1 w_2 x \\ w_1^2 x & w_1^2 & w_1 w_2 x & w_1 w_2 \\ w_1 w_2 x x^\top & w_1 w_2 x & w_2^2 x x^\top & w_2^2 x \\ w_1 w_2 x & w_1 w_2 & w_2^2 x & w_2^2 \end{pmatrix} dx$$

$$= 2 \begin{pmatrix} w_1^2 \int_{-1}^{1} x x^\top dx & w_1^2 \int_{-1}^{1} x dx & w_1 w_2 \int_{-1}^{1} x x^\top dx & w_1 w_2 \int_{-1}^{1} x dx \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \end{pmatrix}$$

As $\mu \to 0$, in the region where $\max g_\theta(x) = \alpha^T x + \gamma$ (corresponding to the first two rows of the matrices W and V) $w_1 \to 1$ and $w_2 \to 0$, and conversely for the case $\max g_\theta(x) = \beta^T x + \phi$. Comparing with V and W in Lemma 5 shows that the matrices for the unsmoothed case are recovered as $\mu \to 0$.

B A generalization of the model in Siegmund and Zhang (1994) for the intersection of two planes

B.1 Two lines

The first part derives the model of Siegmund and Zhang (1994) from two line equations. A priori, consider the model of two lines and a change point θ,

$$y = (\alpha_1 x + \alpha_2)\mathbb{I}(x < \theta) + (\beta_1 x + \beta_2)\mathbb{I}(x \geq \theta).$$

Leaving out either indicator is valid since then, the remaining coefficients will be fit to compensate for the first (fixed) line equation, ie.

$$y = (\alpha_1 x + \alpha_2) + (\beta_1 x + \beta_2)\mathbb{I}(x \geq \theta).$$

Since θ is the change point, both lines must intersect at θ, hence

$$\alpha_1\theta + \alpha_2 = (\alpha_1 + \beta_1)\theta + (\alpha_2 + \beta_2),$$

since the for $x \geq \theta$, the second indicator is one and the new terms are added to the first line. Simplifying the above yields $\beta_2 = -\beta_1\theta$, and plugging this into the line equation yields

$$\begin{aligned} y &= (\alpha_1 x + \alpha_2) + (\beta_1 x - \beta_1\theta)\mathbb{I}(x \geq \theta) \\ &= (\alpha_1 x + \alpha_2) + \beta_1(x - \theta)\mathbb{I}(x \geq \theta) \\ &= \alpha_1 x + \alpha_2 + \beta_1(x - \theta)^+, \end{aligned}$$

where it was used that Siegmund and Zhang (1994) define $(x-\theta)^+ = \max(x-\theta, 0)$, which is $(x - \theta)$ if $x \geq \theta$ and 0 otherwise. Thus it is the same as $(x - \theta)\mathbb{I}(x \geq \theta)$.

B.2 Two planes

For 3D, a similar expression for two intersecting planes can be obtained in a similar fashion. First, a priori the model of two intersecting planes is given by

$$z = (\alpha_1 x + \alpha_2 y + \alpha_3)\mathbb{I}(P1) + (\beta_1 x + \beta_2 y + \beta_3)\mathbb{I}(P2),$$

where the indicators symbolically encode whether (x, y) lie in plane $P1$ or $P2$.

Note first that similar to the 2D case, any two intersecting planes in 3D can be separated by a vertical plane that is inserted at the right spot in-between them. It is not necessary to tilt such a separator plane, similarly to the fact that a one-dimensional pa-

rameter θ was sufficient to separate two lines. Precisely, the separator plane has to be inserted between the two intersecting planes vertically such that it stands on the projection line of the intersection line of the two planes (projected down to the lower square side of the cube in which the two planes are defined).

Let C be the lower square side of the cube in which the two planes are defined. Let P be the rim of this square. By considering any two points $p, q \in \partial P$, all separator planes can be parameterized, and any point (x, y) can be characterized to lie on either the "left" or "right" side of the separator by considering the cross product.

Define for any point (x, y) the side indicator

$$(x, y)^s = \text{sign}\left(\begin{vmatrix} q_x - p_x & x - p_x \\ q_y - p_y & y - p_y \end{vmatrix}\right).$$

This quantity is 1 if (x, y) is one the "left" side of the line through p and q, it is -1 if (x, y) is on the "right" side and it is 0 if (x, y) lies on the line. Using the side indicator,

$$z = (\alpha_1 x + \alpha_2 y + \alpha_3)\mathbb{I}((x, y)^s < 0) + (\beta_1 x + \beta_2 y + \beta_3)\mathbb{I}((x, y)^s \geq 0).$$

Define analogously to Siegmund and Zhang (1994),

$$\phi(x, y)^+ = \phi(x, y)\mathbb{I}((x, y)^s \geq 0) = \begin{cases} \phi(x, y) & (x, y)^s \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

i.e. the result of $\phi(x, y)^+$ is precisely $\phi(x, y)$ if its argument (x, y) is on the + (left) side.

Using the fact that the parameters of the second plane can compensate for the first ones, the first indicator can again be omitted, leaving

$$z = (\alpha_1 x + \alpha_2 y + \alpha_3) + (\beta_1 x + \beta_2 y + \beta_3)\mathbb{I}((x, y)^s \geq 0).$$

Now, suppose both planes intersect above the projected line through p, q, that is all points in the intersection of the planes additionally satisfy the line equation

$$(x - p_x)(q_y - p_y) = (y - p_y)(q_x - p_x) \Rightarrow y = \frac{(x - p_x)(q_y - p_y)}{q_x - p_x} + p_y =: f(x).$$

This assumes that $p_x \neq q_x$. Otherwise, the line through p, q can be parameterized as

$$(x - p_x)(q_y - p_y) = (y - p_y)(q_x - p_x) \Rightarrow x = \frac{(y - p_y)(q_x - p_x)}{q_y - p_y} + p_x = p_x =: f(y).$$

As in the 2D case, use this to save one parameter of the plane equation: On the line $(x, y = f(x))$, the two planes given above by $z = ...$ with one omitted indicator coincide

and hence

$$\alpha_1 x + \alpha_2 f(x) + \alpha_3 = (\alpha_1 + \beta_1)x + (\alpha_2 + \beta_2)f(x) + (\alpha_3 + \beta_3),$$

leading to $\beta_3 = -\beta_1 x - \beta_2 f(x)$. Simplifying the plane equation gives

$$\begin{aligned} z &= (\alpha_1 x + \alpha_2 y + \alpha_3) + (\beta_1 x + \beta_2 y - \beta_1 x - \beta_2 f(x))\mathbb{I}((x,y)^s \geq 0) \\ &= \alpha_1 x + \alpha_2 y + \alpha_3 + \beta_2(y - f(x))\mathbb{I}((x,y)^s \geq 0) \\ &= \alpha_1 x + \alpha_2 y + \alpha_3 + \beta_2(y - f(x))^+, \end{aligned}$$

for $p_x \neq q_x$, and using the definition of $f(x,y)^+$.

This is the analogue of the Siegmund and Zhang (1994) model for intersecting planes. It makes sense to have 4 parameters since for a given projected intersection line, one needs three parameters to define the first plane, the precise non-projected intersection follows from a vertical cut through the first plane at the given projected line, and only one parameter is then needed to define the angle with which the second plane arises from the non-projected intersection line.

Do the two planes in the model coincide at $(x, f(x))$? Clearly, the first plane above the given projected intersection line is given by $\alpha_1 x + \alpha_2 f(x) + \alpha_3$, whereas the other is given by $\alpha_1 x + \alpha_2 f(x) + \alpha_3 + \beta_2(f(x) - f(x))^+ = \alpha_1 x + \alpha_2 f(x) + \alpha_3$. They hence intersect at $(x, f(x))$.

Reversely, if the two planes in the above model intersect then

$$\alpha_1 x + \alpha_2 y + \alpha_3 = \alpha_1 x + \alpha_2 y + \alpha_3 + \beta_2(y - f(x))$$

and hence at the intersection $y = f(x)$, provided $\beta_2 \neq 0$, hence the two planes intersect precisely at the pre-specified projected intersection line $(x, f(x))$).

References

Bacon, D. and Watts, D. (1971). Estimating the transition between two intersecting straight lines. *Biometrika*, 53(3):525–534.

Blischke, W. (1961). Least squares estimators of two intersecting lines. *Technical Report No. 7, Department of Navy, Office of Naval Research, Contract No. Nonr-401(39), Project No. NR 042-212*, BU-135-M(1):1–10.

Broyden, C. (1970). The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA J Appl Math*, 6(1):76–90.

Das, R., Banerjee, M., Nan, B., and Zheng, H. (2015). Fast estimation of regression parameters in a broken-stick model for longitudinal data. *J Am Stat Assoc. To appear.*

Feder, P. (1975). The log likelihood ratio in segmented regression. *Ann Statist*, 3(1):84–97.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322.

Ginsburgh, V., Tishler, A., and Zang, I. (1980). Alternative estimation methods for two-regime models: A mathematical programming approach. *Eur Econ Rev*, 13(2):207–228.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Math. Comp.*, 24:23–26.

Hempel, A. B., Goulart, P. J., and Lygeros, J. (2013). Every continuous piecewise affine function can be obtained by solving a parametric linear program. In *Control Conference (ECC), 2013 European*, pages 2657–2662. IEEE.

Henderson, H. and Velleman, P. (1981). Building multiple regression models interactively. *Biometrics*, 37(2):391–411.

Hudson, D. (1966). Fitting segmented curves whose join points have to be estimated. *J Am Stat Assoc*, 61(316):1097–1129.

Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics.

Kripfganz, A. and Schulze, R. (1987). Piecewise affine functions as a difference of two convex functions. *optimization*, 18(1):23–29.

Mazumder, R., Choudhury, A., Iyengar, G., and Sen, B. (2015). A computational framework for multivariate convex regression and its variants. *arXiv:1509.08165*, pages 1–36.

Michelot, C. (1986). A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50(1):195–200.

Molinari, N., Daurès, J., and Durand, J. (2001). Regression splines for threshold selection in survival data analysis. *Stat Med*, 20(5):237–247.

Muggeo, V. (2003). Estimating regression models with unknown break-points. *Stat Med*, 22(1):3055–3071.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math Program*, 103(1):127–152.

Polyak, B. (1987). *Introduction to Optimization*. Optimization Software, Inc., Publications Division, New York.

Quandt, R. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J Am Stat Assoc*, 53(284):873–880.

Quandt, R. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *J Am Stat Assoc*, 55(290):324–330.

Rigby, R. and Stasinopoulos, D. (1992). Detecting break points in the hazard function in survival analysis. *Statistical Modelling*, pages 303–311.

Robison, D. (1964). Estimates for the points of intersection of two polynomial regressions. *J Am Stat Assoc*, 59(305):214–224.

Scholtes, S. (2012). *Introduction to Piecewise Differentiable Equations*. Springer Briefs in Optimization.

Shanno, D. (1970). Conditioning of quasi-newton methods for function minimization. *Math. Comp.*, 24:647–656.

Shor, N. (1985). *Minimization Methods for Non-Differentiable Functions*. Springer-Verlag Berlin.

Siegmund, D. and Zhang, H. (1994). Confidence regions in broken line regression. *IMS Lecture Notes (Monograph Series)*, 23(1):292–316.

Smith, A. and Cook, D. (1980). Straight lines with a change-point: A bayesian analysis of some renal transplant data. *J.R.S.S. Series C (Applied Statistics)*, 29(2):180–189.

Stasinopoulos, D. and Rigby, R. (1992). Detecting break points in generalised linear models. *Comput Stat Data An*, 13(4):461–471.

Tishler, A. and Zang, I. (1981). A new maximum likelihood algorithm for piecewise regression. *Ann Statist*, 76(376):980–987.

van de Geer, S. (1988). *Regression analysis and empirical processes*. Amsterdam, The Netherlands: Centrum voor Wiskunde en Informatica. CWI Tract 45.

van de Geer, S. (2009). *Empirical Processes in M-Estimation*. Cambridge University Press.

van der Vaart, A. and Wellner, J. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics.