Biometrika (2020), **00**, 0, *pp.* 1–8
Printed in Great Britain
doi: 10.1093/biomet/asaa049

Jump or kink: on super-efficiency in segmented linear regression breakpoint estimation

BY YINING CHEN

Department of Statistics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, U.K.
y.chen101@lse.ac.uk

SUMMARY

We consider the problem of segmented linear regression with a single breakpoint, with the focus on estimating the location of the breakpoint. If n is the sample size, we show that the global minimax convergence rate for this problem in terms of the mean absolute error is $O(n^{-1/3})$. On the other hand, we demonstrate the construction of a super-efficient estimator that achieves the pointwise convergence rate of either $O(n^{-1})$ or $O(n^{-1/2})$ for every fixed parameter value, depending on whether the structural change is a jump or a kink. The implications of this example and a potential remedy are discussed.

Some key words: Changepoint; Minimax rate; Pointwise rate; Structural break.

1. INTRODUCTION

Asymptotic analysis is commonly used to facilitate comparison between different statistical estimators from a frequentist's perspective. Once the consistency of an estimator is established, the focus then naturally moves onto its rate of convergence. In general, statements concern the following two types of rates: the pointwise rate where the limit is taken when the unknown parameter is fixed, and the uniform rate where the limit is taken as the supremum over some or all of the parameter space. In addition, the convergence rate of the estimator that achieves the fastest uniform rate among all the estimators is known as the minimax rate. Often, the global minimax rate is used to characterize the hardness of the problem.

In many settings, the pointwise rate, the uniform rate and the minimax rate are the same, in which case the corresponding estimator is usually regarded as rate optimal. However, there are exceptions where caution must be exercised. A notable example arises from the phenomenon of super-efficiency, first documented by Joseph L. Hodges, Jr. in 1951. This topic was later treated comprehensively by Le Cam (1953) and Hájek (1972), among many others, in the settings of regular parametric models. See Stigler (2007) and Vovk (2009) for excellent reviews of the turbulent history of the early studies. More recently, super-efficiency has been investigated in more complicated settings, for instance in nonparametric function estimation (Brown et al., 1997), mixture models (Heinrich & Kahn, 2018) and isotonic regression (Banerjee et al., 2019).

Let us denote the parameter space of interest by Θ, any estimator of $\theta \in \Theta$ by $\hat{\theta}$, the loss function by $L(\theta, \hat{\theta})$ and the corresponding risk function by $R(\theta, \hat{\theta}) = E_\theta L(\theta, \hat{\theta})$. For every $\theta \in \Theta$, suppose that there exists some $\gamma_\theta > 0$ such that

$$0 < \liminf_{\epsilon \to 0+} \liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\|\theta' - \theta\| \leqslant \epsilon} n^{\gamma_\theta} R(\theta', \hat{\theta}) \leqslant \limsup_{\epsilon \to 0+} \limsup_{n \to \infty} \inf_{\hat{\theta}} \sup_{\|\theta' - \theta\| \leqslant \epsilon} n^{\gamma_\theta} R(\theta', \hat{\theta}) < \infty;$$

then $n^{-\gamma_\theta}$ is known as the local minimax rate at θ. In this context, an estimator $\hat{\theta}$ is super-efficient in its convergence rate if

$$\limsup_{n\to\infty} n^{\gamma_\theta} R(\theta,\hat{\theta}) < \infty \text{ for every } \theta \in \Theta \quad \text{and} \quad \limsup_{n\to\infty} n^{\gamma_\theta} R(\theta,\hat{\theta}) = 0 \quad \text{for some } \theta \in \Theta.$$

The purpose of this short note is to demonstrate that super-efficiency can occur in the setting of segmented linear regression, even with only a single breakpoint. In spite of the popularity of segmented regression in the statistics and econometrics literature, to our knowledge this phenomenon has not been widely understood in these contexts. In particular, since the class of segmented linear regression models is not regular, e.g., not differentiable in quadratic mean, existing results regarding super-efficiency in regular parametric models cannot be immediately applied. By focusing on estimating the location of the single breakpoint and taking the loss function to be the Euclidean distance between the true location and estimated location of the breakpoint, we show that the global minimax convergence rate of the risk is at least $O(n^{-1/3})$, i.e.,

$$\liminf_{n\to\infty} \inf_{\hat{\theta}} \sup_{\theta\in\Theta} n^{1/3} R(\theta,\hat{\theta}) > 0.$$

We then illustrate super-efficiency by constructing an estimator $\hat{\theta}^{\mathrm{S}}$ that, for every fixed $\theta \in \Theta$, depending on whether the breakpoint is a jump or a kink, satisfies either

$$n\, R(\theta,\hat{\theta}^{\mathrm{S}}) < \infty \quad \text{or} \quad n^{1/2}\, R(\theta,\hat{\theta}^{\mathrm{S}}) < \infty.$$

These findings point to an interesting scenario in which the breakpoint can be estimated at the rate of $O_p(n^{-1/3})$ in the minimax sense. However, as long as we are willing to assume that the truth does exist, i.e., the location, as a proportion of the data sequence, and the size of change does not vary with n, the breakpoint can then be estimated at a much faster rate of either $O_p(n^{-1})$ or $O_p(n^{-1/2})$. Consequently, for this particular breakpoint estimation problem, caution must be taken when one uses the global minimax rate to characterize its difficulty, and when one compares and interprets the convergence rates of different estimators.

Here, our focus is on the segmented linear regression with a single breakpoint, with the aim of illustrating super-efficiency. These results also hold in the setting of multiple breakpoints under suitable conditions. There have been a number of recent works on estimating the number and locations of unknown breakpoints in the settings of both continuous and discontinuous piecewise linear mean signals. See Bai & Perron (1998), Muggeo (2003), Das et al. (2016), Maidstone et al. (2019) and Baranowski et al. (2019), to name but a few. In particular, with less stringent spacing conditions between consecutive breakpoints in the setting of a continuous piecewise linear mean signal, Maidstone et al. (2019) and Baranowski et al. (2019) proposed estimators that could achieve within a logarithmic factor of $O_p(n^{-1/3})$ estimation of the locations of all unknown breakpoints. The estimator's convergence rate was further improved to within a logarithmic factor of $O_p(n^{-1/2})$ in Baranowski et al. (2019) under more restrictive assumptions. See also Hansen (2017) for inference in the presence of a kink, Hidalgo et al. (2019) for a test of continuity at the breakpoint, and an as yet unpublished 2017 paper by Y. Dong from the University of California Irvine for a related problem on treatment effect evaluation. There has also been work on kink location estimation in various univariate nonparametric regression settings. See Raimondo (1998), Goldenshluger et al. (2006), Cheng & Raimondo (2008), Wishart & Kulik (2010) and Wishart (2011), and references therein. Finally, we mention the work of Korostelev & Lepski (2008), who investigated a version of the jump location estimation problem with a growing dimension.

2. MODEL SET-UP: SEGMENTED LINEAR REGRESSION WITH A SINGLE BREAKPOINT

Suppose that we observe (X_{ni}, Y_{ni}) for $i = 1, \ldots, n$. Consider the fixed design setting where

$$X_{ni} = i/(n+1),$$
$$Y_{ni} = f_\theta(X_{ni}) + \sigma\varepsilon_{ni}$$

for some $\sigma > 0$ and some function $f_\theta : [0,1] \to \mathbb{R}$. Here, $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ are independent and identically distributed $N(0,1)$ random variables. Furthermore, f_θ is a piecewise linear function indexed by $\theta = (\tau_\theta, \alpha_\theta^-, \alpha_\theta^+, \beta_\theta^-, \beta_\theta^+) \in \Theta \subset [0,1] \times \mathbb{R}^4$ of the form

$$f_\theta(x) = \begin{cases} \alpha_\theta^- + \beta_\theta^-(x - \tau_\theta) & \text{if } x \in [0, \tau_\theta], \\ \alpha_\theta^+ + \beta_\theta^+(x - \tau_\theta) & \text{if } x \in (\tau_\theta, 1]. \end{cases}$$

In other words, f_θ has a single breakpoint at τ_θ, with its linear part over $[0, \tau_\theta)$ determined by the slope β_θ^- and the intercept α_θ^- at $(\tau_\theta)-$, and its linear part over $(\tau_\theta, 1]$ determined by the slope β_θ^+ and the intercept α_θ^+ at $(\tau_\theta)+$. For simplicity, we have assumed that f_θ is left-continuous, so $f_\theta(\tau_\theta) = \alpha_\theta^-$. If $|\alpha_\theta^+ - \alpha_\theta^-| \neq 0$, then we refer to τ_θ as a jump. Otherwise, if $|\alpha_\theta^+ - \alpha_\theta^-| = 0$ but $|\beta_\theta^+ - \beta_\theta^-| \neq 0$, then we call τ_θ a kink.

To asymptotically analyse the breakpoint estimator based on $(X_{n1}, Y_{n1}), \ldots (X_{nn}, Y_{nn})$, it is common to assume that the actual breakpoint does not occur too close to the boundary at $x = 0$ or $x = 1$, and the structural change is noticeable, so at least one of the two quantities $|\alpha_\theta^+ - \alpha_\theta^-|$ and $|\beta_\theta^+ - \beta_\theta^-|$ is reasonably large. As such, it is natural to restrict ourselves to the parameter space of

$$\Theta = \left\{\theta \in [0,1] \times \mathbb{R}^4 \mid \tau_\theta \in [\delta, 1-\delta],\ \max\left(|\alpha_\theta^+ - \alpha_\theta^-|, |\beta_\theta^+ - \beta_\theta^-|\right) \geqslant \delta\right\}$$

for some fixed but perhaps unknown small $\delta > 0$. Here, the dependence of Θ on δ is suppressed.

As mentioned previously, our main focus is on estimating the location of the breakpoint. In a sense, we treat α_θ^-, α_θ^+, β_θ^- and β_θ^+ as nuisance parameters. For any estimator $\hat{\theta} = (\tau_{\hat{\theta}}, \alpha_{\hat{\theta}}^-, \alpha_{\hat{\theta}}^+, \beta_{\hat{\theta}}^-, \beta_{\hat{\theta}}^+)$, we evaluate its performance based on the estimated breakpoint's absolute loss, namely, with the loss function $L(\theta, \hat{\theta}) = |\tau_{\hat{\theta}} - \tau_\theta|$ and the risk function $R(\theta, \hat{\theta}) = E_\theta L(\theta, \hat{\theta})$. Analogous conclusions could also be made under other losses, such as $L(\theta, \hat{\theta}) = |\tau_{\hat{\theta}} - \tau_\theta|^q$ for some $q > 1$.

Finally, we remark that this particular fix design is selected with the aim of better connecting with the existing changepoint detection literature.

3. Minimax rate of convergence

First, we investigate the local minimax rate of convergence. We separate the parameter space into two disjoint sets Θ^{K} and $\Theta \backslash \Theta^{\mathrm{K}}$, where Θ^{K} is the parameter space representing functions with a kink, i.e.,

$$\Theta^{\mathrm{K}} = \left\{\theta \in \Theta \mid \alpha_\theta^- = \alpha_\theta^+\right\}.$$

Theorem 1. *Under the set-up in* § 2*,*

$$\liminf_{\epsilon \to 0+} \liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leqslant \epsilon} n R(\theta', \hat{\theta}) > 0 \qquad \textit{for every } \theta \in \Theta \backslash \Theta^{\mathrm{K}}$$

and

$$\liminf_{\epsilon \to 0+} \liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leqslant \epsilon} n^{1/3} R(\theta', \hat{\theta}) > 0 \qquad \textit{for every } \theta \in \Theta^{\mathrm{K}}. \tag{1}$$

Theorem 1 implies that when τ_θ is a jump, the local minimax rate for estimating the location of the breakpoint in terms of the magnitude of $\tau_{\hat{\theta}} - \tau_\theta$ is at least of $O_p(n^{-1})$. However, this rate slows down considerably to $O_p(n^{-1/3})$ when τ_θ is a kink.

The next corollary concerns the global minimax rate, which immediately follows from Theorem 1.

Corollary 1. *Under the set-up in* § 2*,* $\liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} n^{1/3} R(\theta, \hat{\theta}) > 0$.

Although it is known that when the exact type of breakpoint is given a priori, a jump could be estimated at $O_p(n^{-1})$ and a kink could be estimated at $O_p(n^{-1/2})$, we emphasize that these facts alone are far from

implying the minimax rate of $O_p(n^{-1/3})$ for $\tau_{\hat{\theta}} - \tau_\theta$ as shown above. The $O_p(n^{-1/3})$ rate also appears in Raimondo (1998), which considers a related problem in the nonparametric setting where there is a jump in the first derivative of a continuous mean. However, the class of functions he considered in deriving this rate is different from ours.

On the other hand, if we further constrain the parameter space from Θ to Θ^{K}, then the following minimax result for kink location estimation in the setting of continuous segmented linear regression holds.

THEOREM 2. *Under the set-up in § 2,*

$$\liminf_{n\to\infty} \inf_{\hat{\theta}} \sup_{\theta\in\Theta^{\mathrm{K}}} n^{1/2} R(\theta, \hat{\theta}) > 0. \tag{2}$$

Note that $O_p(n^{-1/2})$ implied by (2) is faster than $O_p(n^{-1/3})$ implied by (1). This seemingly counter-intuitive difference in the rates is due to the different choices of the parameter spaces in their derivations, and can be explained by examining the proofs of Theorems 1 and 2 in the Supplementary Material. Our proofs follow from Le Cam's two-point method. See Le Cam (1986) or Yu (1997).

To give some intuition, we first confine ourselves to $\theta_1 = (1/2, 0, 0, -1, 1)$ and $\theta_2 = (1/2 + \Delta, -\Delta, \Delta, -1, 1)$ for some small $\Delta > 0$, with $\theta_1, \theta_2 \in \Theta$. Denote the distribution of $(Y_{n1}, \ldots, Y_{nn})$ using the data-generating process described in § 2 with θ_1 as $\mathcal{P}^n_{\theta_1}$, and that with θ_2 as $\mathcal{P}^n_{\theta_2}$. Then, breakpoint estimation could be viewed as the problem of differentiating between $\mathcal{P}^n_{\theta_1}$ and $\mathcal{P}^n_{\theta_2}$ based on the observations, whose hardness is dictated by the squared total variation distance between them. In the meantime, this squared total variation distance, denoted by $\|\mathcal{P}^n_{\theta_1} - \mathcal{P}^n_{\theta_2}\|^2_{\mathrm{TV}}$, can be bounded under suitable conditions as follows, with $\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x$ playing a crucial role in characterizing the hardness of the original problem:

$$\begin{aligned}
\|\mathcal{P}^n_{\theta_1} - \mathcal{P}^n_{\theta_2}\|^2_{\mathrm{TV}} &\leqslant 2 - 2\prod_{i=1}^{n} \left\{1 - d^2_{\mathrm{hel}}\big(N\big[f_{\theta_1}\{i/(n+1)\}, \sigma^2\big], \big[N\big(f_{\theta_2}\{i/(n+1)\}, \sigma^2\big]\big)\right\} \\
&= 2 - 2\exp\left(-\frac{1}{8\sigma^2}\sum_{i=1}^{n}\big[f_{\theta_1}\{i/(n+1)\} - f_{\theta_2}\{i/(n+1)\}\big]^2\right) \\
&\to 2 - 2\exp\left[-\frac{n}{8\sigma^2}\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x\right].
\end{aligned}$$

Here, d_{hel} is the Hellinger distance and $d^2_{\mathrm{hel}}\{N(\mu_1, \sigma), N(\mu_2, \sigma)\} = 1 - \exp\{-(\mu_1 - \mu_2)^2/(8\sigma^2)\}$. For this particular pair of θ_1 and θ_2, $\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x = O(\Delta^3)$ for $\Delta \to 0$; see Fig. 1(a). Here, $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ only differ over $x \in (1/2, 1/2 + \Delta]$, meaning that the problem can be viewed as a local one as most pairs of the observations, namely (X_{ni}, Y_{ni}) with $X_{ni} \notin (1/2, 1/2 + \Delta]$, are irrelevant. In contrast, with the same value of θ_1, if we only consider parameters in Θ^{K} we are then unable to find a $\theta_2 \in \Theta^{\mathrm{K}}$ with the corresponding breakpoint at $1/2 + \Delta$ such that $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ only differ over a small neighbourhood. In fact, $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ will have to differ over a substantial interval that does not shrink as $\Delta \to 0$, so the problem of distinguishing between f_{θ_1} and f_{θ_2} appears more global than before. By taking, for example, $\theta_2 = (1/2 + \Delta, -\Delta, -\Delta, -1, 1)$, while keeping the same θ_1, we obtain $\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x = O(\Delta^2)$, as demonstrated in Fig. 1(b). In fact, we can further show that $\inf_{\theta_2\in\Theta^{\mathrm{K}}} \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x = O(\Delta^2)$. This order difference of $\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x$ in Δ implies that the cases of $\theta \in \Theta$ and $\theta \in \Theta^{\mathrm{K}}$ are fundamentally different! To give more details, in Le Cam's method, the minimax rate can be derived by picking Δ such that $\|\mathcal{P}^n_{\theta_1} - \mathcal{P}^n_{\theta_2}\|_{\mathrm{TV}} \leqslant C$ for some constant $C < 1$. In our settings, as derived as above, this roughly amounts to requiring $\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x = O(n^{-1})$. As such, Δ would be taken as $O(n^{-1/3})$ in Fig. 1(a), and $O(n^{-1/2})$ in Fig. 1(b), which are also the minimax convergence rates for breakpoint estimation under $\theta \in \Theta$ and $\theta \in \Theta^{\mathrm{K}}$, respectively, in terms of the expected absolute loss. Finally, for completeness, we also illustrate the case of a noticeable jump in Fig. 1(c), where $\theta_1 = (1/2, -1/2, 1/2, -1, 1)$, i.e., with jump size of 1, and $\theta_2 = (1/2 + \Delta, -1/2 - \Delta, 1/2 + \Delta, -1, 1)$. Since $\int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 \mathrm{d}x = O(\Delta)$ here, the minimax rate for estimating a breakpoint of this type is $O(n^{-1})$.

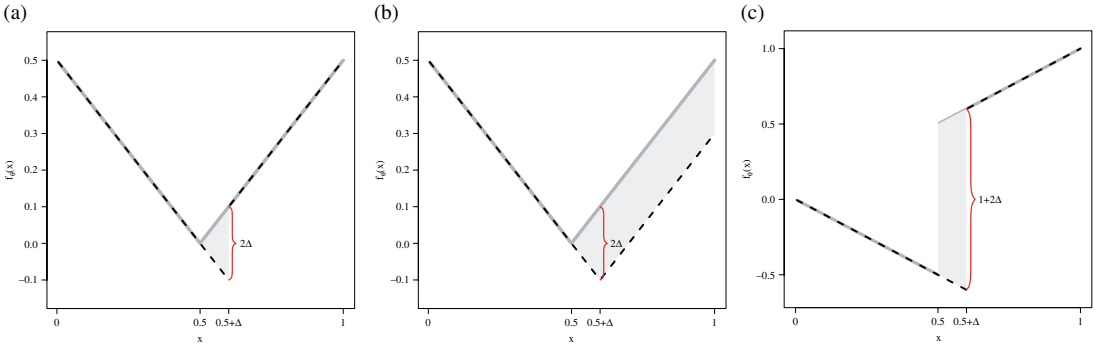

Fig. 1. Plots of f_{θ_1} and f_{θ_2} with their difference shaded in light grey. In (a), $\theta_1 = (1/2, 0, 0, -1, 1)$ and $\theta_2 = (1/2 + \Delta, -\Delta, \Delta, -1, 1)$. In (b), the continuity constraint is enforced with the same θ_1, but $\theta_2 = (1/2 + \Delta, -\Delta, -\Delta, -1, 1)$. Finally, (c) demonstrates the case of a nonvanishing jump with $\theta_1 = (1/2, -1/2, 1/2, -1, 1)$ and $\theta_2 = (1/2 + \Delta, -1/2 - \Delta, 1/2 + \Delta, -1, 1)$. In all the plots, the difference between f_{θ_1} and f_{θ_2} at $x = 1/2 + \Delta$ is highlighted using a curly bracket. Solid lines represent $f_{\theta_1}(x)$ and dashed lines represent $f_{\theta_2}(x)$.

4. A SUPER-EFFICIENT ESTIMATOR

Write $\bar{\Theta} = [0, 1] \times \mathbb{R}^4$ and $\bar{\Theta}^{\mathrm{K}} = \{\theta \in \bar{\Theta} \mid \alpha_\theta^- = \alpha_\theta^+\}$. For notational convenience, henceforth $\hat{\theta}$ is denoted as the least squares estimator satisfying

$$\hat{\theta} := \hat{\theta}_{\mathrm{LS}} \in \operatorname{argmin}_{\theta \in \bar{\Theta}} \sum_{i=1}^{n} \{Y_{ni} - f_\theta(X_{ni})\}^2. \tag{3}$$

Here, we minimize over $\bar{\Theta}$ instead of Θ, because δ of Θ is not always known a priori. Equivalently, we can write

$$\tau_{\hat{\theta}} \in \operatorname{argmin}_{\tau \in [0,1]} \left\{ \min_{\alpha,\beta} \sum_{i=1}^{n} (Y_{ni} - \beta X_{ni} - \alpha)^2 \mathbb{1}_{\{X_{ni} \in [0,\tau]\}} + \min_{\alpha,\beta} \sum_{i=1}^{n} (Y_{ni} - \beta X_{ni} - \alpha)^2 \mathbb{1}_{\{X_{ni} \in (\tau,1]\}} \right\}.$$

Similarly, for kink estimation where we further restrict ourselves to Θ^{K}, we denote the corresponding least squares estimator by $\hat{\theta}^{\mathrm{K}} \in \operatorname{argmin}_{\theta \in \bar{\Theta}^{\mathrm{K}}} \sum_{i=1}^{n} \{Y_{ni} - f_\theta(X_{ni})\}^2$, with

$$\tau_{\hat{\theta}^K} \in \operatorname{argmin}_{\tau \in [0,1]} \left[\min_{\alpha,\beta^-,\beta^+} \sum_{i=1}^{n} \left\{ Y_{ni} - \alpha - \beta^-(X_{ni} - \tau) \mathbb{1}_{\{X_{ni} \in [0,\tau]\}} - \beta^+(X_{ni} - \tau) \mathbb{1}_{\{X_{ni} \in (\tau,1]\}} \right\}^2 \right].$$

For uniqueness, we take $\tau_{\hat{\theta}}$ and $\tau_{\hat{\theta}^K}$ to be the smallest element in the respective sets of minima.

As we shall see, $\tau_{\hat{\theta}}$ achieves $O_p(n^{-1})$ for fixed $\theta \in \Theta \backslash \Theta^{\mathrm{K}}$, but slows down to $O_p(n^{-1/3})$ when $\theta \in \Theta^{\mathrm{K}}$. Meanwhile, $\tau_{\hat{\theta}^{\mathrm{K}}}$ achieves $O_p(n^{-1/2})$ for $\theta \in \Theta^{\mathrm{K}}$. Therefore, making the correct extra assumption of continuity at the breakpoint and using the corresponding estimator could improve the convergence rate from $O_p(n^{-1/3})$ to $O_p(n^{-1/2})$ for $\theta \in \Theta^{\mathrm{K}}$. This motivates us to shrink $\hat{\theta}$ towards Θ^{K} in certain cases to improve the pointwise rate. In particular, we could estimate the breakpoint by $\tau_{\hat{\theta}^{\mathrm{S}}}$, where

$$\hat{\theta}^{\mathrm{S}} = \begin{cases} \hat{\theta}^{\mathrm{K}} & \text{if } |\alpha_{\hat{\theta}}^+ - \alpha_{\hat{\theta}}^-| \leqslant n^{-1/6}, \\ \hat{\theta} & \text{if } |\alpha_{\hat{\theta}}^+ - \alpha_{\hat{\theta}}^-| > n^{-1/6}. \end{cases} \tag{4}$$

We are now in the position to discuss the pointwise and local uniform convergence rates of $\tau_{\hat{\theta}^{\mathrm{S}}}$.

THEOREM 3. *Under the set-up in § 2, $\tau_{\hat{\theta}^{\mathrm{S}}}$ is a super-efficient estimator for τ_θ. In particular, we have*

$$\limsup_{n \to \infty} nR(\theta, \hat{\theta}^{\mathrm{S}}) < \infty \qquad \textit{for every } \theta \in \Theta \backslash \Theta^{\mathrm{K}}$$

and

$$\limsup_{n\to\infty} n^{1/2} R(\theta, \hat{\theta}^{\mathrm{S}}) < \infty \qquad \textit{for every } \theta \in \Theta^{\mathrm{K}}.$$

It follows from Theorem 3 that $\sup_{\theta\in\Theta} \limsup_{n\to\infty} n^{1/2} R(\theta, \hat{\theta}^{\mathrm{S}}) < \infty$, i.e., the pointwise rate of breakpoint estimation via $\hat{\theta}^{\mathrm{S}}$ for every $\theta \in \Theta$ is faster than the global minimax rate of $O(n^{-1/3})$.

THEOREM 4. *Under the set-up in § 2, we have that*

$$\limsup_{\epsilon\to 0+} \limsup_{n\to\infty} \sup_{\theta'\in\Theta:\|\theta'-\theta\|\leqslant\epsilon} nR(\theta', \hat{\theta}^{\mathrm{S}}) < \infty \qquad \textit{for every } \theta \in \Theta\backslash\Theta^{\mathrm{K}}$$

and

$$\liminf_{\epsilon\to 0+} \liminf_{n\to\infty} \sup_{\theta'\in\Theta:\|\theta'-\theta\|\leqslant\epsilon} n^{1/3} R(\theta', \hat{\theta}^{\mathrm{S}}) = \infty \qquad \textit{for every } \theta \in \Theta^{\mathrm{K}}.$$

Theorem 4 implies that the global uniform rate of $\tau_{\hat{\theta}^{\mathrm{S}}}$ in terms of the absolute loss is worse than the global minimax rate of $O(n^{-1/3})$, which could actually be achieved by $\tau_{\hat{\theta}}$. This type of behaviour is typical for super-efficient estimators that tend to achieve better pointwise convergence rates at the cost of worse uniform convergence rates.

In addition, the construction of the super-efficient estimator is by no means unique. In fact, here the threshold $n^{-1/6}$ in (4) can be replaced by $cn^{-\gamma}$ for any fixed $c > 0$ and $\gamma \in (0, 1/3)$. Alternatively, one could replace $|\alpha^+_{\hat{\theta}} - \alpha^-_{\hat{\theta}}|$ in (4) by the difference between the residual sum of squares from fitting the model over either $\bar{\Theta}^{\mathrm{K}}$ or $\bar{\Theta}$, and then choose the cut-off decision boundary accordingly.

5. NUMERICAL EXPERIMENT

We run a small simulation study to compare the behaviour of $\tau_{\hat{\theta}}$ and $\tau_{\hat{\theta}^{\mathrm{S}}}$. Two different scenarios are considered under the settings of § 2: (a) $\theta_1 = (0.5, 0, 0, -1, 1) \in \Theta^{\mathrm{K}}$, i.e., $f_{\theta_1}(x) = |x - 0.5|$; (b) $\theta_2 = (0.5, 0, 0.5, -1, 1) \in \Theta\backslash\Theta^{\mathrm{K}}$, i.e., $f_{\theta_2}(x) = |x - 0.5| + \mathbb{1}_{\{x>0.5\}}$. Here, we take $\sigma = 0.5$ and $n = 100, 200, 500, 1000, 2000$. All experiments are repeated 1000 times. The estimated values of $R(\theta, \hat{\theta})$ and $R(\theta, \hat{\theta}^{\mathrm{S}})$, also known as the mean absolute errors of $\tau_{\hat{\theta}}$ and $\tau_{\hat{\theta}^{\mathrm{S}}}$, are reported in Fig. 2 on a log-log scale.

The super-efficiency phenomenon is visible in Fig. 2(a), where the super-efficient estimator $\tau_{\hat{\theta}^{\mathrm{S}}}$ performs better than the least squares estimator $\tau_{\hat{\theta}}$ in the presence of a kink, especially for large n. It is also evident from the plot that $\tau_{\hat{\theta}^{\mathrm{S}}}$ and $\tau_{\hat{\theta}}$ have different pointwise convergence rates there, as indicated in § 4. Meanwhile, Fig. 2(b) demonstrates that in the presence of a jump, $\tau_{\hat{\theta}^{\mathrm{S}}}$ and $\tau_{\hat{\theta}}$ perform similarly. In particular, for large $n = 2000$, they are exactly the same in all 1000 runs. Finally, Fig. 2 confirms that in terms of pointwise rates, estimating the location of a jump is easier than estimating that of a kink.

6. DISCUSSION

Although the super-efficient estimator $\tau_{\hat{\theta}^{\mathrm{S}}}$ achieves a pointwise rate faster than the global minimax rate for every $\theta \in \Theta$, it is clear that in our example super-efficiency only occurs over Θ^{K}, which is a Lebesgue null set in comparison to Θ. However, if we were to focus solely on τ_θ by treating α^-_θ, α^+_θ, β^-_θ and β^+_θ as nuisance parameters, then super-efficiency could occur at every $\tau_\theta \in [\delta, 1 - \delta]$. Here, $[\delta, 1 - \delta]$ is no longer a Lebesgue null set in comparison to $[0, 1]$.

On the other hand, while $\tau_{\hat{\theta}^{\mathrm{S}}}$ achieves a better pointwise convergence rate for locating the breakpoint, like the Hodges estimator, it is penalized at local neighbouring points. In particular, $\limsup_{n\to\infty} \sup_{\theta\in\Theta} n^{1/3} R(\theta, \hat{\theta}^{\mathrm{S}}) = \infty$, but $\limsup_{n\to\infty} \sup_{\theta\in\Theta} n^{1/3} R(\theta, \hat{\theta}) < \infty$, where $\hat{\theta}$ is taken as the

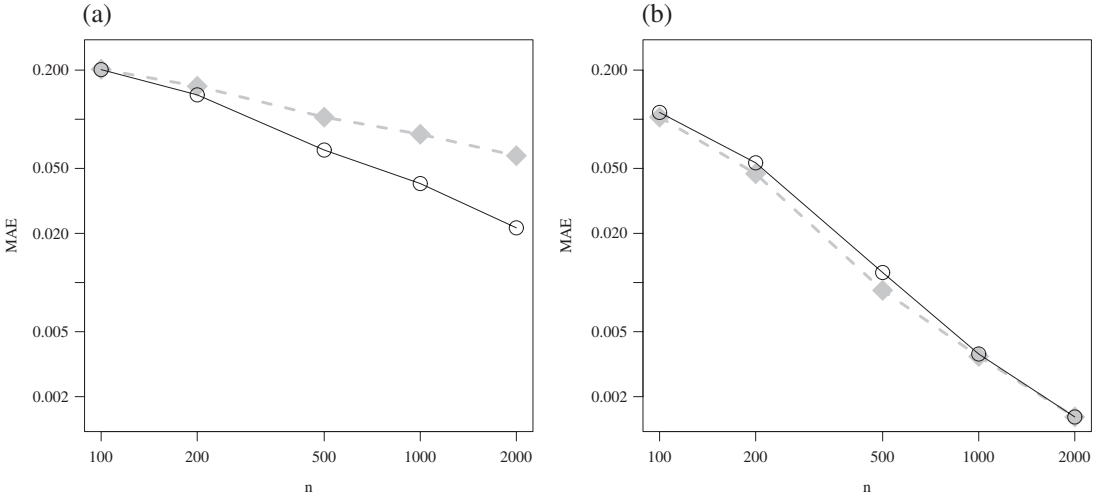

Fig. 2. Estimated mean absolute errors of $\tau_{\hat{\theta}}$ (dashed) and $\tau_{\hat{\theta}^S}$ (solid) for $n = 100, 200, 500, 1000, 2000$ on a log–log scale under different scenarios: (a) θ_1 with a kink; (b) θ_2 with a jump.

least squares estimator defined in (3). Thus, in terms of the uniform convergence rate, $\tau_{\hat{\theta}^S}$ actually performs worse than $\tau_{\hat{\theta}}$.

However, it is useful to think about whether this perspective of uniformity is what we are really interested in. In this breakpoint estimation problem, the global minimax convergence rate is derived by considering alternatives with the jump size $\to 0$ as $n \to \infty$. It is entirely possible that these alternatives might violate the modeller's real brief in practice, whose intention dictates that whenever there is a jump or a change in slope, or both, the corresponding change size has to be significant. Mathematically, this plausibly more appropriate parameter space would be a restricted version of Θ, given by

$$\Theta^* = \left\{\theta \in \Theta \mid \min\left(\mathbb{1}_{\{\alpha_\theta^- = \alpha_\theta^+\}} + \left|\alpha_\theta^+ - \alpha_\theta^-\right|\mathbb{1}_{\{\alpha_\theta^- \neq \alpha_\theta^+\}},\ \mathbb{1}_{\{\beta_\theta^- = \beta_\theta^+\}} + \left|\beta_\theta^+ - \beta_\theta^-\right|\mathbb{1}_{\{\beta_\theta^- \neq \beta_\theta^+\}}\right) \geqslant \delta\right\}.$$

It can then be shown that for every $\theta \in \Theta^{\mathrm{K}}$,

$$\liminf_{\epsilon \to 0+} \liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta' \in \Theta^*: \|\theta' - \theta\| \leqslant \epsilon} n^{1/2} R(\theta', \hat{\theta}) > 0,$$

implying that the local minimax convergence rate for estimating the kink over Θ^* is $O_p(n^{-1/2})$, and thus $\hat{\theta}^{\mathrm{S}}$ is no longer super-efficient over Θ^*.

Besides, one interesting feature of this breakpoint estimation problem is that the local minimax convergence rates are different across the parameter space. In this circumstance, it might not be entirely adequate to summarize the hardness of the problem by a single global minimax rate. See also Donoho et al. (1995).

Our example can be generalized in more complex settings, such as segmented polynomial regression with higher-order polynomials, as well as those with heterogeneous sub-Gaussian errors and multiple breakpoints. It shows the importance of choosing suitable parameter spaces for the calculation of uniform or minimax convergence rates. We hope that it also demonstrates the need to interpret different types of convergence rates and the practical meaning of rate optimality with care and caution.

Acknowledgement

The author thanks Moulinath Banerjee, Haeran Cho and Piotr Fryzlewicz for their helpful comments on an earlier draft, as well as the editors and three reviewers for their valuable suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains proofs of all the theoretical results.

REFERENCES

BAI, J. & PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.

BANERJEE, M., DUROT, C. & SEN, B. (2019). Divide and conquer in non-standard problems and the super-efficiency phenomenon. *Ann. Statist.* **47**, 720–57.

BARANOWSKI, R., CHEN, Y. & FRYZLEWICZ, P. (2019). Narrowest-over-threshold detection of multiple change-points and change-point-like features. *J. R. Statist. Soc.* B **81**, 649–72.

BROWN, L. D., LOW, M. G. & ZHAO, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25**, 2607–25.

CHENG, M-Y. & RAIMONDO, M. (2008). Kernel methods for optimal change-points estimation in derivatives. *J. Comp. Graph. Statist.* **17**, 56–75.

DAS, R., BANERJEE, M., NAN, B. & ZHENG, H. (2016). Fast estimation of regression parameters in a broken-stick model for longitudinal data. *J. Am. Statist. Soc.* **111**, 1132–43.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. & PICHARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. R. Statist. Soc.* B **57**, 301–69.

GOLDENSHLUGER, A., TSYBAKOV, A. & ZEEVI, A. (2006). Optimal change-point estimation from indirect observations. *Ann. Statist.* **34**, 350–72.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, vol. 1, pp. 175–94.

HANSEN, B. E. (2017). Regression kink with an unknown threshold. *J. Bus. Econ. Statist.* **35**, 228–40.

HEINRICH, P. & KAHN, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46**, 2844–70.

HIDALGO, J., LEE, J. & SEO, M. H. (2019). Robust inference for threshold regression models. *J. Economet.* **210**, 291–309.

KOROSTELEV, A. & LEPSKI, O. (2008). On a multi-channel change-point. *Math. Methods Statist.* **17**, 187–97.

LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. Calif. Pub. Statist.* **1**, 277–330.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer.

MAIDSTONE, R., FEARNHEAD, P. & LETCHFORD, A. (2019). Detecting changes in slope with an L_0 penalty. *J. Comp. Graph. Statist.* **28**, 265–75.

MUGGEO, V. M. R. (2003). Estimating regression models with unknown break-points. *Statist. Med.* **22**, 3055–71.

RAIMONDO, M. (1998). Minimax estimation of sharp change points. *Ann. Statist.* **26**, 1379–97.

STIGLER, S. M. (2007). The epic story of maximum likelihood. *Statist. Sci.* **22**, 598–620.

VOVK, V. (2009). Superefficiency from the vantage point of computability. *Statist. Sci.* **24**, 73–86.

WISHART, J. R. (2011). Minimax lower bound for kink location estimators in a nonparametric regression model with long-range dependence. *Statist. Prob. Lett.* **81**, 1871–5.

WISHART, J. R. & KULIK, R. (2010). Kink estimation in stochastic regression with dependent errors and predictors. *Electron. J. Statist* **4**, 875–913.

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen & G. L. Yang, eds. pp. 423–35. New York: Springer.

[*Received on* 25 *April* 2019. *Editorial decision on* 3 *February* 2020]