

ASYMPTOTICS FOR CHANGE-POINT MODELS UNDER VARYING DEGREES OF MIS-SPECIFICATION

BY RUI SONG¹, MOULINATH BANERJEE² AND MICHAEL R. KOSOROK³

*North Carolina State University, University of Michigan
and University of North Carolina at Chapel Hill*

Change-point models are widely used by statisticians to model drastic changes in the pattern of observed data. Least squares/maximum likelihood based estimation of change-points leads to curious asymptotic phenomena. When the change-point model is correctly specified, such estimates generally converge at a fast rate (n) and are asymptotically described by minimizers of a jump process. Under complete mis-specification by a smooth curve, that is, when a change-point model is fitted to data described by a smooth curve, the rate of convergence slows down to $n^{1/3}$ and the limit distribution changes to that of the minimizer of a continuous Gaussian process. In this paper, we provide a bridge between these two extreme scenarios by studying the limit behavior of change-point estimates under varying degrees of model mis-specification by smooth curves, which can be viewed as local alternatives. We find that the limiting regime depends on how quickly the alternatives approach a change-point model. We unravel a family of “intermediate” limits that can transition, at least qualitatively, to the limits in the two extreme scenarios. The theoretical results are illustrated via a set of carefully designed simulations. We also demonstrate how inference for the change-point parameter can be performed in absence of knowledge of the underlying scenario by resorting to sub-sampling techniques that involve estimation of the convergence rate.

1. Introduction. The study of change-point models has a long and rich history in the statistics and econometrics literature. Change-point models, where a signal function shows an abrupt transition at one or more points in its domain can be used to study phenomena that are subject to sudden shock effects, or which show natural phase-transitions at different stages of evolution. Applications are many and varied and arise in the analysis of climate data [Lund and Reeves (2002)], estimation of mixed layer depth from oceanic profile data [Thomson and Fine (2003)], structural breaks in economics [Bai and Perron (1998, 2003)], quality control and dynamical systems in an engineering context [Lai (1995)] and genetics [Shen and Zhang (2012)], to name a few. Sequential methods for change-point

Received May 2014; revised July 2015.

¹Supported in part by NSF Grant DMS-10-07698 and 1309465.

²Supported in part by NSF Grant DMS-10-07751 and 1308890.

³Supported in part by NCI Grant P01 CA142538.

MSC2010 subject classifications. Primary 62G20, 62G05; secondary 62E20.

Key words and phrases. Change-point, model mis-specification.

detection have been around for a very long time; the literature here is truly huge, with a comprehensive treatment in the book by [Basseville and Nikiforov \(1993\)](#) and the excellent review paper by [Lai \(2001\)](#), but see also [Csörgő and Horváth \(1997\)](#) which has an in-depth study of limit theorems in change-point analysis. On the other hand, inference on jump-discontinuities (change-points) in an otherwise smooth curve based on observed or designed data has also received attention in the nonparametric as well as the survival analysis literature: see, for example, [Gijbels, Hall and Kneip \(1999\)](#), [Hall and Molchanov \(2003\)](#), [Kosorok and Song \(2007\)](#), [Lan, Banerjee and Michailidis \(2009\)](#), [Loader \(1996\)](#), [Müller \(1992\)](#), [Müller and Song \(1997\)](#), [Pons \(2003\)](#), [Ritov \(1990\)](#) and references therein. A canonical change-point model which illustrates many important features of this genre of problems is given by

$$Y = \beta_l 1(X \leq \theta) + \beta_u 1(X > \theta) + \varepsilon,$$

where the predictor $X \in [0, 1]$ is assumed to be a continuous random variable, $\beta_l \neq \beta_u$ are fixed constants, ε is a continuous random variable, independent of X with zero expectation and finite variance. The parameters of interest are the change-point parameter θ and the regression parameter $(\beta_l, \beta_u)'$. For this model, the least squares estimator of the change-point parameter converges to the truth at rate n , with the limit distribution being described by the minimizer of a two-sided, compound Poisson process. The asymptotic distribution of the least squares estimates of $(\beta_l, \beta_u)'$ are normal, and are unaltered by estimation of the change-point, that is, they have the same distribution as the least squares estimates that would have been obtained if θ were known. The detailed analysis can be found in [Kosorok \(2008\)](#). A closely related model allows the parameters β_l and β_u to approach each other with increasing sample size n (as opposed to staying fixed in the above display). As long as $\beta_u - \beta_l$ approaches 0 at a rate slower than $n^{-1/2}$, the change-point can be estimated. However, due to the loss of signal in this model, the rate of convergence of the LSE of θ slows to $n^{1-2\xi}$, where $\beta_u - \beta_l = O(n^{-\xi})$; furthermore, the limit distribution is now starkly different and described by the minimizer of Brownian motion plus triangular drift. See, for example, [Bhattacharya and Brockwell \(1976\)](#) for an early treatment of this problem, and [Müller and Song \(1997\)](#) for a nonparametric incarnation. [Hušková \(1999\)](#) considered estimators in location models with various gradual changes and showed that the limit behavior of least squares-type estimators of the change point in these models depends on the type of gradual change.

A natural question, as in most statistical problems, is the effect of misspecification on the change-point estimator. Suppose first that the true model is of the form $Y = f(X) + \varepsilon$ where f is actually smooth but that the model $\beta_l 1(X \leq \theta) + \beta_u 1(X > \theta)$ is fitted instead. This is what happens, for example, in CART where the change-point analysis represents the best approximation of a binary decision tree (piecewise constant function with a single jump, also called a

stump) to f . Bühlmann and Yu (2002) and Banerjee and McKeague (2007) studied the asymptotics of the estimates of the change-point and the regression coefficient in this problem and showed that in this setting cube-root asymptotics with Chernoff limit distributions obtain. As shown by Banerjee and McKeague (2007), all three least squares estimates converge at the slower $n^{1/3}$ rate because the change-point estimation depends on local features of the smooth regression curve, which are more complex in comparison to when the true regression function is a stump model. Therefore, change-point estimation and inference are highly unstable under model mis-specification by a smooth curve due to this sharp fall in the estimator's rate of convergence: from a rate as fast as n under the true change-point model, to only a cube-root rate under a smooth curve.

While this is an interesting finding, this formulation does not quite capture the more subtle issue of how the degree of mis-specification affects the estimates of the stump parameters: to elaborate, consider functions f_1 and f_2 that are both smooth, but suppose that one is linear and the other a sigmoidal function with a sharp ascent. Clearly, f_2 is much closer to a stump-model than f_1 , so fitting the mis-specified change-point model should be less consequential in the case of f_2 than f_1 . But the fixed-model approach described in the above paragraph does not satisfactorily capture this issue. This motivates us, in this work, to consider models where the degree of mis-specification is allowed to change-diminish, in fact—as $n \rightarrow \infty$ and to explore the consequences of this diminishing mis-specification on the behavior of the stump estimates. In particular, how does the rate of mis-specification bear upon this behavior in terms of rates of convergence and limit distributions?

Our strategy considers a sequence of models $Y = f_n(X) + \varepsilon$, where f_n converges to a stump function at a rate controlled by a parameter $\alpha_n \rightarrow \infty$. We find that if the f_n 's converge to a stump slowly enough [$\alpha_n = o(n)$], the limit distribution of the change-point estimator stays identical to the case $f_n \equiv f$, the fixed function setting of Banerjee and McKeague (2007), though the rate of convergence can be accelerated to (almost) n ; if f_n 's approach the stump rapidly [$n = o(\alpha_n)$], the rate and limit distribution are identical to those that obtain when the true function is a stump, whereas, at the boundary $n \sim \alpha_n$, the limit distribution is different from either of the previous scenarios and belongs to a family of distributions that can transition, in a manner to be made precise in Section 4, to the Chernoff distribution [the limit with $\alpha_n = o(n)$] on one end and the minimizer of a compound Poisson process [the limit distribution with $n = o(\alpha_n)$] on the other. The joint limit behavior of the estimates of the levels of the stump and the jump-point, however, show an abrupt change as one changes from $\alpha_n = o(n)$ to $n = O(\alpha_n)$: in the former case, the normalized estimates are asymptotically correlated with correlation 1 (i.e., linear functions of one another), which is also what happens in the “fixed f ” scenario, while for the latter the estimates of the levels are asymptotically independent of that of the change-point. Viewing these f_n 's as a sequence of local

alternatives to the limiting null model, a stump, the above phenomena are qualitatively identical to what transpires with the MLE in regular parametric models under a sequence of local alternatives, depending on how quickly the alternatives approach the null, an analogue we discuss more fully in the final section.

The problem addressed in this paper should be contrasted with the “local alternative”—type models considered in [Bhattacharya and Brockwell \(1976\)](#), [Müller and Song \(1997\)](#) and [Hušková \(1999\)](#). In all these papers, the limit of the sequence of change-point models considered—the so-called “null” model—is a smooth function *without* a change-point, whereas we have the *reverse* scenario: our sequence of models are smooth functions that, in the limit, produce a discontinuous change-point model. Our work is also quite different from inference in settings where the change-point is not a discontinuity but represents a point of smooth and/or gradual change; see, for example, [Vogt and Dette \(2015\)](#). To the best of our knowledge, our work is the first attempt at providing a comprehensive as well as systematic understanding of the behavior of change-point models under local smooth alternatives. We hope that it will stimulate more investigation into this relatively uncharted territory.

An interesting aspect of our approach is the fact that our paradigm approximates fixed function settings as the sample size varies. Consider again, the function f_2 introduced in the previous page, with its sharp ascent. As we show in our simulations (Section 6), with a fixed steep function (not changing with n), each of our three asymptotic regimes takes turns being the best approximation to the sample distribution of the estimated change-point. Specifically, for small sample sizes, the steep function is indistinguishable from the stump model and the compound Poisson process is the best approximation; for moderate sample sizes, the intermediate regime is the best approximation; while for large sample sizes, the fact that the function is not a stump is detectable by the estimator and the Chernoff limit becomes the best approximation. In other words, the proposed contiguous model sequence realistically approximates the range of distributional behavior which can be found in practical data settings.

From the perspective of statistical inference, the key contribution of our work is the formulation of a concrete theoretical framework for adaptive estimation of the change-point parameter under possible mis-specification through a subsampling procedure, elaborated in Section 5. Consider the following basic inference problem: *Given data $\{Y_i, X_i\}$ from a regression model to which a stump model has been fit, how would a practitioner go about constructing a confidence interval for the change-point parameter? The true underlying (unknown) model: $Y = v_0(X) + \varepsilon$ may not even be a change-point model but potentially in the proximity of one.* The framework of our paper then allows a way of making inference on the change-point parameter by (i) *couching* the underlying model in a sequence of models f_n determined by an unknown α_n , (ii) providing a meaningful interpretation to the population change-point parameter and (iii) last but not least, building confidence intervals for the parameter by adaptively estimating the correct regime for

the given dataset through a subsampling procedure which teases out the correct degree of calibration (the convergence rate) from the data itself.

Before we proceed further, a word of clarification regarding the use of the term “change-point” is in order. It is true that the underlying regression functions in the framework of this paper are smooth and therefore do not possess a change-point in the conventional sense of the term. The change-point model is used as a working model to fit the data; in that sense, the change-point is really a split-point in the spirit of Banerjee and McKeague (2007). However, if one takes the point of view that any change-point model fitted by a statistician is really an approximation to some underlying continuous model, with better fits corresponding to continuous models that are close to a model with a discontinuity—so the notion of a change-point is really a convenient idealization—then the term “change-point” can be used without further scope of confusion. Indeed, the title of the paper emphasizes this view in highlighting the mis-specification angle upfront.

The rest of the paper is organized as follows. In Section 2, we formulate the regression problem and list our assumptions with interpretations. In Section 3, we systematically establish the asymptotic results including consistency, convergence rates and weak convergence of the change-point estimator. The connections among the different limiting distributions obtained for different choices of α_n are established in Section 4. Section 5 describes the adaptive inference strategy. Section 6 presents empirical evidence from a simulation study where a stump model is fit to data arising from a smooth regression curve and illustrates when the different asymptotic regimes comes into play: this is seen to depend on the nature of the underlying smooth function as well as the sample size, and the boundary case is seen to provide more robust approximations than the others. Section 7 discusses connections of our results to a number of other problems and scope for furthering this direction of research.

2. Change-point models under model mis-specifications.

2.1. *The model set-up.* At stage n , the observed data (Y_i, X_i) , $i = 1, \dots, n$ are n i.i.d. copies of (Y, X) , where $Y = f_n(X) + \varepsilon$: here $E(\varepsilon) = 0$, ε is independent of X with bounded second moment and X follows some distribution F_X on $[0, 1]$. Thus, we have a sequence of models (changing with n). The functions f_n will be constructed to be smooth but will converge to a stump function as described later in this section.

At each stage n , our working model will be a stump of the form $g(x; \theta, \beta_l, \beta_u) = \beta_l 1(x \leq \theta) + \beta_u 1(x > \theta)$ and the best working model will be determined from the sample via least squares. Denote a generic $(\theta, \beta_l, \beta_u)$ by ψ , and let

$$\begin{aligned} \hat{\psi}^n &\equiv (\hat{\theta}^n, \hat{\beta}_l^n, \hat{\beta}_u^n)^T = \arg \min_{\psi} \sum_{i=1}^n (Y_i - g(X_i, \psi))^2 \\ &= \arg \min_{\psi} \{\mathbb{P}_n [y - g(x; \psi)]^2\}, \end{aligned}$$

with \mathbb{P}_n denoting the empirical measure of the data $\{Y_i, X_i\}_{i=1}^n$. Letting P_n denote the true distribution at stage n , the corresponding population parameter $\psi^n = (\theta^n, \beta_l^n, \beta_u^n)^T$ is defined through the least squares estimation problem:

$$\begin{aligned} \psi^n &= \arg \min_{\psi} M_n(\psi) \\ &= \arg \min_{\psi} P_n m_{\psi}(X, Y) = \arg \min_{\psi} P_n \{Y - \beta_l 1(X \leq \theta) - \beta_u 1(X > \theta)\}^2. \end{aligned}$$

We assume that there is a unique (population) minimizer $\psi^n \equiv (\theta^n, \beta_l^n, \beta_u^n)$, with $\beta_l^n \neq \beta_u^n$ at stage n .

To focus on the main ideas, we consider a sequence f_n of the type

$$f_n(x) = f(\alpha_n(x - \theta^0)),$$

where f is a smooth bounded monotone (increasing) function defined on \mathbb{R} and α_n a sequence going to ∞ . Denote $f(-\infty)$ by β_l^0 and $f(\infty)$ by β_u^0 . As n goes to infinity, $f_n(x)$ then converges to the stump

$$f_0(x) \equiv \beta_l^0 1(x \leq \theta^0) + \beta_u^0 1(x > \theta^0)$$

at all points except θ_0 . We let $\psi^0 = (\theta^0, \beta_l^0, \beta_u^0)^T$ denote this limiting population parameter. Note that the speed of convergence of f_n to f_0 is regulated by the parameter α_n . Define $\xi^n = \alpha_n(\theta^n - \theta^0)$, which can be viewed as a rescaled “bias” term due to model mis-specification.

Let $\xi^n = \alpha_n(\theta^n - \theta^0)$ and $\xi^0 = f^{-1}((\beta_l^0 + \beta_u^0)/2)$. From Theorem 2.2 below, $\lim_{n \rightarrow \infty} \xi^n = \xi^0$.

From the normal equations that characterize ψ^n , we have

$$\begin{aligned} \xi^n &= f^{-1}((\beta_l^n + \beta_u^n)/2), \\ \beta_l^n &= \frac{\int_0^{\theta^n} f(\alpha_n(x - \theta^0)) p_X(x) dx}{P(X \leq \theta^n)}, \\ \beta_u^n &= \frac{\int_{\theta^n}^1 f(\alpha_n(x - \theta^0)) p_X(x) dx}{P(X > \theta^n)}. \end{aligned}$$

Although the working model is an oversimplification of the true model at each fixed n , as n gets larger, the approximation to the true model is better. It will be shown later that ψ^n converges to its limit ψ^0 . The statistic $\hat{\psi}_n$ defined earlier estimates ψ^n and, therefore, indirectly, ψ_0 . We note here that the minimizer of $\mathbb{P}_n[y - g_n(x; \psi)]^2$ is not necessarily unique, so in the case of multiple minimizers, we take $\hat{\psi}_n$ to be the minimizer with the smallest value of the first co-ordinate (if two minimizers have identical first co-ordinates their last two co-ordinates must also coincide). For simplicity of reference, we call this the *smallest* argmin. We will study the asymptotic behavior of $\hat{\psi}^n$ as α_n converges to infinity at different rates.

2.2. *Assumptions.* We now describe our assumptions on the model formulated above.

A. $f(x)$ is continuously differentiable in an open neighborhood N of ξ^0 with $f'(\xi_0) > 0$.

B. The density $p_X(x)$ does not vanish and is continuously differentiable in a neighborhood of θ_0 .

C1.

$$\inf_n \liminf_{|h_1| \rightarrow \infty} \frac{1}{|h_1|} \int_0^{h_1} [f(t + \xi^n) - f(\xi^n)] dt > 0.$$

C2. For any positive constant K , $\inf_n \inf_{u \in [-K + \xi^n, K + \xi^n]} f'(u) > 0$.

C3. The integrals $\int_{-\infty}^{\xi^0} (f(t) - \beta_l^0) dt$ and $\int_{\xi^0}^{\infty} (f(t) - \beta_u^0) dt$ exist and are denoted as ξ^l and ξ^u , respectively.

Assumptions A and B are adapted from the conditions in Banerjee and McKeague (2007). Assumption C1 says that the average increase of f over all sufficiently large finite intervals with ξ^n as an end-point should be bounded away from 0. Assumption C2 is, essentially, a positivity condition on the derivative of f in every compact neighborhood of ξ_0 . Assumption C3 figures in calculating the asymptotic bias of β_l^n and β_u^n for β_l^0 and β_u^0 , respectively. Note that this assumption implies that $\int_{-\infty}^0 \{f(t) - \beta_l^0\}^2 dt$ and $\int_0^{\infty} \{f(t) - \beta_u^0\}^2 dt$ are both $O(1)$.

In the sequel, it should be understood that the proof of any lemma, proposition or theorem that does not appear in the main text has been relegated to the supplementary material [Song, Banerjee and Kosorok (2015)].

2.3. *Limiting behavior of ψ^n .* We establish the asymptotic behavior of ψ^n , the stage n population parameter in two steps. First, we show the consistency of ψ^n for ψ^0 and next, we establish the convergence rates and calculate the limiting (normalized) bias of ψ^n for ψ^0 . Note that the convergence results of ψ^n to ψ^0 in all three steps are deterministic. The following theorem establishes the consistency of ψ^n .

THEOREM 2.1. *Under assumptions A, B and C, $\lim_{n \rightarrow \infty} (|\theta^n - \theta^0| + |\beta_l^n - \beta_l^0| + |\beta_u^n - \beta_u^0|) = 0$.*

The next theorem deals with convergence rates and asymptotic bias.

THEOREM 2.2. *Under assumptions A–C, $\lim_n \alpha_n(\theta^n - \theta^0) = \xi^0$, $\lim_n \alpha_n(\beta_l^n - \beta_l^0) = p_X(\theta^0) F_X(\theta^0)^{-1} \xi^l$, and $\lim_n \alpha_n(\beta_u^n - \beta_u^0) = p_X(\theta^0) \times F_X(\theta^0)^{-1} \xi^u$.*

REMARK. Since $\alpha_n(\theta^n - \theta_0) = f^{-1}((\beta_l^n + \beta_u^n)/2)$, and by Theorem 2.1, β_l^n and β_u^n converge to β_l^0 and β_u^0 , respectively, it is immediate that $\alpha_n(\theta^n - \theta_0)$ converges to ξ_0 as defined earlier in this section. The proofs of the remaining two convergences are given in the Appendix [Song, Banerjee and Kosorok (2015)].

3. Asymptotic results for $\hat{\psi}_n$. We will present the asymptotic results for $\hat{\psi}_n$ in three subsections in the order of consistency, convergence rates and weak convergence.

3.1. *Consistency.* We first establish Euclidean consistency for $\hat{\psi}^n$, where the results are summarized in Theorem 3.1.

THEOREM 3.1. *Under assumptions A–C, $|\hat{\theta}^n - \theta^n| + |\hat{\beta}_l^n - \beta_l^n| + |\hat{\beta}_u^n - \beta_u^n| = o_P(1)$.*

3.2. *Rate of convergence.* In this section, we establish the convergence rates for change-point estimators under different degrees of model mis-specification. As an important first step, we introduce a dichotomous distance that describes the variation of the population criterion function about its minimizer.

$$\begin{aligned}\rho_1(\psi, \psi^n) &= \max\{\alpha_n^{1/2}|\theta - \theta^n|, |\beta_l - \beta_l^n|, |\beta_u - \beta_u^n|\}, \\ \rho_2(\psi, \psi^n) &= \max\{|\theta - \theta^n|^{1/2}, |\beta_l - \beta_l^n|, |\beta_u - \beta_u^n|\}.\end{aligned}$$

The following lemma is about a unified distance which enables a certain expansion of the objective function.

LEMMA 3.1. *Under assumptions A–C, it follows that for ψ in a neighborhood of ψ^n defined as: $\{\psi : d_n(\psi, \psi^n) < \delta_0\}$ for some small $\delta_0 > 0$, there exists a positive constant E_0 such that*

$$(3.1) \quad M_n(\psi) - M_n(\psi^n) \geq E_0 d_n^2(\psi, \psi^n),$$

where $d_n^2(\psi, \psi^n) = \rho_1^2(\psi, \psi^n)1(|\theta - \theta^n| \leq K\alpha_n^{-1}) + \rho_2^2(\psi, \psi^n)1(|\theta - \theta^n| > K\alpha_n^{-1})$.

This dichotomous nature of the distance d_n is really what drives the convergence rate of $\hat{\psi}_n$. It reflects the fact that the magnitude of the fluctuation of M_n around ψ^n is governed by where θ falls with respect to a (shrinking) α_n^{-1} order neighborhood of θ^n . If θ falls in the shrinking neighborhood, the growth of M_n around ψ^n in the first co-ordinate is at least of order $\alpha_n(\theta - \theta^n)^2$; if not, the growth is at least of order $|\theta - \theta^n|$, which appears in the classic correctly specified change-point problem considered in Kosorok (2008). Note that the order of $\alpha_n(\theta - \theta^n)^2$ is dominated by that of $|\theta - \theta^n|$, precisely when $|\theta - \theta^n|$ is $O(1/\alpha_n)$: this slower growth of M_n in its first co-ordinate in a shrinking Euclidean neighborhood is what makes

the convergence rate fall below n for a wide range of α_n . For slow-growing α_n , which can be considered as α_n essentially behaving like a constant, we converge toward the ρ_1 setting and the distance function of Banerjee and McKeague (2007) and approach the $n^{1/3}$ convergence rate for $\hat{\theta}^n$ obtained in their work; for rapidly growing α_n , we move toward the ρ_2 setting and the distance function in Kosorok (2008), and, approach the n -rate of convergence instead. The precise statements of the convergence rates appear in Theorem 3.2 below.

We next calculate bounds on the modulus of continuity of the empirical process with respect to this distance: this is one of the key ingredients that dictates the convergence rate. The dichotomous nature of the distance requires exercising some care via calculating

$$P_n^* \left[\sup_{d_n(\psi, \psi^n) < \delta} |\mathbb{G}_n \{m_\psi(X, Y) - m_{\psi^n}(X, Y)\}| \right],$$

where $\mathbb{G}_n m(\cdot) = (\mathbb{P}_n - P_n)m(\cdot)$ for function $m(\cdot)$. By definition of the distance $d_n(\psi, \psi^n)$, for some $\delta > 0$ we have

$$\begin{aligned} \{d_n(\psi, \psi^n) < \delta\} &= \{\rho_1(\psi, \psi^n) < \delta, |\theta - \theta^n| \leq 1/\alpha_n\} \\ &\cup \{\rho_2(\psi, \psi^n) < \delta, |\theta - \theta^n|^{1/2} > 1/\sqrt{\alpha_n}\}. \end{aligned}$$

For $\delta \leq 1/\sqrt{\alpha_n}$, the second term on the right-hand side is the null set and since for this range $\delta/\sqrt{\alpha_n} \leq 1/\alpha_n$, we have $\{d_n(\psi, \psi^n) < \delta\} = \{\rho_1(\psi, \psi^n) < \delta\}$.

On the other hand, for $\delta > 1/\sqrt{\alpha_n}$, $\delta/\sqrt{\alpha_n} > 1/\alpha_n$ and the set $\{d_n(\psi, \psi^n) < \delta\} = \{\rho_2(\psi, \psi^n) < \delta\}$. In the next lemma, we establish the order of modulus of two function classes which will be used for the convergence rates, as stated in Theorem 3.2.

LEMMA 3.2. *Under assumptions A–C, we have that for $0 < \delta \leq 1/\sqrt{\alpha_n}$,*

$$(3.2) \quad E_n^* \left[\sup_{d_n(\psi, \psi^n) < \delta} |\mathbb{G}_n(m_\psi(X, Y) - m_{\psi^n}(X, Y))| \right] \lesssim \frac{\delta^{1/2}}{\alpha_n^{1/4}},$$

where E_n^* denote the outer expectation at stage n . On the other hand, for $\delta > 1/\sqrt{\alpha_n}$,

$$(3.3) \quad E_n^* \left[\sup_{d_n(\psi, \psi^n) < \delta} |\mathbb{G}_n(m_\psi(X, Y) - m_{\psi^n}(X, Y))| \right] \lesssim \delta.$$

REMARK. The proof of Lemma 3.2 involves reasonably standard arguments that use maximal inequalities to control the expected modulus of continuity of an empirical process via the magnitude of an envelope function and an entropy integral. The proof of Lemma 3.1 needs more careful handling; in particular, it requires analyzing the fluctuation of M_n about ψ^n in terms of two components: the

fluctuation about the first co-ordinate keeping the others fixed plus the fluctuation about the second and third co-ordinates keeping the first fixed. This is formalized in Lemma 0.1 in the Appendix [Song, Banerjee and Kosorok (2015)], the key preparatory result for the proof of Lemma 3.1.

THEOREM 3.2. *Under assumptions A–C, we have:*

(i) *When $\alpha_n = o(n)$, $n^{1/3}\alpha_n^{2/3}|\hat{\theta}^n - \theta^n| + n^{1/3}\alpha_n^{1/6}|\hat{\beta}_l^n - \beta_l^n| + n^{1/3}\alpha_n^{1/6}|\hat{\beta}_u^n - \beta_u^n| = O_P(1)$.*

(ii) *When $\alpha_n = n$, $n|\hat{\theta}^n - \theta^n| + \sqrt{n}|\hat{\beta}_l^n - \beta_l^n| + \sqrt{n}|\hat{\beta}_u^n - \beta_u^n| = O_P(1)$.*

(iii) *When $n = o(\alpha_n)$, $n|\hat{\theta}^n - \theta^n| + \sqrt{n}|\hat{\beta}_l^n - \beta_l^n| + \sqrt{n}|\hat{\beta}_u^n - \beta_u^n| = O_P(1)$.*

PROOF. From Lemma 3.2, we have for $\delta \leq 1/\sqrt{\alpha_n}$,

$$E_n^* \left[\sup_{d_n(\psi, \psi^n) < \delta} |\mathbb{G}_n(m(x, y, \psi) - m(x, y, \psi^n))| \right] \lesssim \frac{\delta^{1/2}}{\alpha_n^{1/4}}.$$

On the other hand, for $\delta > 1/\sqrt{\alpha_n}$,

$$E_n^* \left[\sup_{d_n(\psi, \psi^n) < \delta} |\mathbb{G}_n(m(x, y, \psi) - m(x, y, \psi^n))| \right] \lesssim \delta.$$

To apply Theorem 0.2 in the Appendix [Song, Banerjee and Kosorok (2015)], we are then led to a bounding function $\phi_n(\delta)$ for the modulus of continuity which is given by

$$\phi_n(\delta) = \frac{\delta^{1/2}}{\alpha_n^{1/4}} 1\left(\delta \leq \frac{1}{\sqrt{\alpha_n}}\right) + \delta 1\left(\delta > \frac{1}{\sqrt{\alpha_n}}\right).$$

It is easily seen that $\phi_n(\delta)/\delta^\alpha$ is a decreasing function for $\alpha = 1$. Solving $r_n^2 \phi_n(1/r_n) \leq \sqrt{n}$ yields

$$(3.4) \quad \frac{r_n^{3/2}}{\alpha_n^{1/4}} 1(r_n \geq \sqrt{\alpha_n}) + r_n 1(r_n < \sqrt{\alpha_n}) \leq \sqrt{n}.$$

Next we analyze the rate from (3.4) via isolating three cases for different choices of α_n one by one.

For the first case, considering α_n going to ∞ but no faster than n , that is, $\alpha_n = o(n)$, we seek a solution with $r_n \geq \alpha_n^{1/2}$. To see this, suppose $r_n < \sqrt{\alpha_n}$. Then the solution is $r_n = \sqrt{n}$. Therefore, $\sqrt{\alpha_n} > \sqrt{n}$. This is a contradiction, however, since by our condition, \sqrt{n} is eventually larger than $\sqrt{\alpha_n}$. This leads to: $r_n = (n^{1/2}\alpha_n^{1/4})^{2/3} = n^{1/3}\alpha_n^{1/6}$. We hence conclude that

$$n^{1/3}\alpha_n^{1/6}\{\rho_1(\hat{\psi}_n, \psi_n)1(\alpha_n|\theta^n - \theta^n| \leq 1) + \rho_2(\hat{\psi}_n, \psi_n)1(\alpha_n|\theta^n - \theta^n| > 1)\}$$

is $O_P(1)$. This implies that

$$n^{1/3}\alpha_n^{1/6}|\hat{\beta}_l^n - \beta_l^n| = O_P(1), \quad n^{1/3}\alpha_n^{1/6}|\hat{\beta}_u^n - \beta_u^n| = O_P(1),$$

and that

$$n^{1/3}\alpha_n^{2/3}|\hat{\theta}^n - \theta^n|1(\alpha_n|\theta^n - \theta^n| \leq 1) + n^{2/3}\alpha_n^{1/3}|\hat{\theta}^n - \theta^n|1(\alpha_n|\theta^n - \theta^n| > 1)$$

is $O_P(1)$. Since $\alpha_n = o(n)$ it is strictly slower than both $n^{1/3}\alpha_n^{2/3}$ and $n^{2/3}\alpha_n^{1/3}$, showing that $\alpha_n|\hat{\theta}^n - \theta^n|$ is $o_P(1)$. This then forces $1(\alpha_n|\hat{\theta}^n - \theta^n| > 1)$ to go to 0 in probability. Since this is a zero-one valued random variable, it is easily argued that the second term in the above display must converge to 0 in probability. Given any subsequence, we can find a further subsequence along which the indicator converges almost surely to 0, and is therefore identically 0 in the long run, whence the second term also has to be identically 0. We thus conclude that $n^{1/3}\alpha_n^{2/3}|\hat{\theta}^n - \theta^n|$ is $O_P(1)$.

For the second case, we consider $\alpha_n = cn$, for some positive constant c . We note that $\alpha_n = n$ is equivalent to $\alpha_n = cn$ for any $c \in (0, \infty)$ since the c can be absorbed into the function f without loss of generality. From now on, we will use $\alpha_n = n$ everywhere else. Both rates $n^{1/3}\alpha_n^{2/3}$ and $n^{2/3}\alpha_n^{1/3}$ are equal to n and we conclude that $n|\hat{\theta}^n - \theta^n|$ is $O_P(1)$, $\sqrt{n}|\hat{\beta}_l^n - \beta_l^n| = O_P(1)$ and $\sqrt{n}|\hat{\beta}_u^n - \beta_u^n| = O_P(1)$.

For the third case, we consider $n = o(\alpha_n)$. In this case, the second part in (3.4) becomes relevant, that is, we seek a solution with $r_n < \sqrt{\alpha_n}$. The r_n from the first part— $n^{1/3}\alpha_n^{1/6}$ —is inconsistent with the condition that $r_n \geq \sqrt{\alpha_n}$ and we are led to the solution $r_n = \sqrt{n}$ which is indeed consistent with the condition $r_n < \sqrt{\alpha_n}$. Conclude that

$$\begin{aligned} & (n\alpha_n)^{1/2}\rho_1^2(\hat{\psi}^n, \psi^n)1(\alpha_n|\hat{\theta}^n - \theta^n| \leq 1) + n\rho_2^2(\hat{\psi}^n, \psi^n)1(\alpha_n|\hat{\theta}^n - \theta^n| > 1) \\ & = O_P(1). \end{aligned}$$

Since $n\alpha_n$ is faster than n^2 , it follows that $n|\hat{\theta}^n - \theta^n|$ is $O_P(1)$, $\sqrt{n}|\hat{\beta}_l^n - \beta_l^n| = O_P(1)$ and $\sqrt{n}|\hat{\beta}_u^n - \beta_u^n| = O_P(1)$. On the other hand, by the observation that the least squares estimate $\hat{\theta}^n$ is at least as far as θ^n from the X_i closest to the latter and the fact that this X_i converges to θ^n at rate n (in fact, $n|X_i - \theta^n|$ converges to an exponential distribution), it follows that n must be the nontrivial rate of convergence. \square

3.3. Asymptotic distributions. Having established the rate of convergence, we now determine the asymptotic distribution. In the following, we discuss three different cases. The first result is the asymptotic distribution for $\alpha_n = o(n)$, which follows a rescaled Chernoff distribution. Recall that Chernoff's distribution is the unique minimizer of $W(t) + t^2$ over all real t , where $W(t)$ is two-sided Brownian motion starting from 0.

THEOREM 3.3. *Let $q_n = n^{1/3}\alpha_n^{1/6}(\alpha_n^{1/2}, 1, 1)^T$ and $F_X(\cdot)$ be the cumulative distribution function of X . Denote the pointwise product on Euclidean space as*

“o.” Under assumptions A–C, when $\alpha_n = o(n)$,

$$q_n \circ (\hat{\theta}^n - \theta^n, \hat{\beta}_l^n - \beta_l^n, \hat{\beta}_u^n - \beta_u^n) \rightarrow_d (1, c_1, c_2) \arg \max_h Q(h),$$

where $Q(h)$ has a rescaled Chernoff distribution: $Q(h) = aW(h) - bh^2$, $W(\cdot)$ is a standard two-sided Brownian motion process on the real line, $a^2 = \sigma^2 p_X(\theta^0)$,

$$b = \frac{1}{2} f'(\xi^0) p_X(\theta^0) - \frac{1}{8} (\beta_u^0 - \beta_l^0) p_X(\theta^0)^2 \left(\frac{1}{F_X(\theta^0)} + \frac{1}{1 - F_X(\theta^0)} \right),$$

$$c_1 = \frac{p_X(\theta^0) (\beta_u^0 - \beta_l^0)}{2F_X(\theta^0)} \quad \text{and} \quad c_2 = \frac{p_X(\theta^0) (\beta_u^0 - \beta_l^0)}{2(1 - F_X(\theta^0))}.$$

REMARK. Note the similarity of the above results to that in Theorem 2.1 of Banerjee and McKeague (2007). The regime $\alpha_n = o(n)$ can be interpreted as the *slow regime* which yields asymptotic behavior similar to the situation in that paper where the smooth function $f_n \equiv f$ and does not change with n . The form of the limits is similar to those obtained in Theorem 2.1 but note the difference in convergence rates. While in Banerjee and McKeague (2007) the rate of convergence of all three parameters is $n^{1/3}$, in our current situation we do get an acceleration above this rate: for the change-point parameter, the accelerated rate can (almost) go up to n and for the level parameters it can (almost) go up to \sqrt{n} as α_n gets close to order n , these limiting rates being the rates of convergence for a correctly specified change-point model. Also note that the asymptotic correlation between the least squares estimate of the stump levels and that of the change-point is 1, whereas in the cases to follow, these will be seen to be asymptotically independent.

The next result is the asymptotic distribution for $\alpha_n = n$. This is the most interesting scenario and yields a new limit distribution. To deduce the limit distribution of $\hat{h}_n = (\hat{h}_{1n}, \hat{h}_{2n}, \hat{h}_{3n})^T$, where $\hat{h}_{1n} = n(\hat{\theta}^n - \theta^n)$, $\hat{h}_{2n} = \sqrt{n}(\hat{\beta}_l^n - \beta_l^n)$ and $\hat{h}_{3n} = \sqrt{n}(\hat{\beta}_u^n - \beta_u^n)$, we consider the limit of the process $h \mapsto Q_n(h) = n\mathbb{P}_n(m_{\psi_{n,h}} - m_{\psi^n})$, where

$$\psi_{n,h} = \psi^n + (h_1/n, h_2/\sqrt{n}, h_3/\sqrt{n}) \quad \text{and} \quad h = (h_1, h_2, h_3)^T.$$

The general scheme of argument runs as follows: We first derive a tractable approximation of Q_n , denoted \tilde{Q}_n , that is, uniformly close to Q_n in a sense to be made precise later. The advantage of \tilde{Q}_n is its decomposability into three parts where each represents the contribution of a parameter. Next, the tightness of \tilde{Q}_n is established, which coupled with finite-dimensional convergence furnishes the weak limit of \tilde{Q}_n . This, by the uniform closeness, alluded to above is also the weak limit of Q_n . The final step involves deriving the weak convergence of the normalized estimators by the application of an appropriate continuous mapping theorem for the argmax/argmin functional.

We start with the first step. From the results on convergence rates, we know that $\hat{h}_n = (\hat{h}_{1n}, \hat{h}_{2n}, \hat{h}_{3n})^T$ is uniformly tight and is the smallest argmin of $h \mapsto Q_n(h) = n\mathbb{P}_n(m_{\psi_{n,h}} - m_{\psi^n})$. Observe that $m_{\psi_{n,h}}(X, Y) - m_{\psi^n}(X, Y)$

$$\begin{aligned} &= 2(Y - f_n(\theta^n))(\beta_u^n - \beta_l^n)\{1(X \leq \theta^n + h_1/n) - 1(X \leq \theta^n)\} \\ &\quad + (2Y - 2\beta_l^n - h_2/\sqrt{n})1(X \leq \theta^n + h_1/n)h_2/\sqrt{n} \\ &\quad + (2Y - 2\beta_u^n - h_3/\sqrt{n})1(X > \theta^n + h_1/n)h_3/\sqrt{n}. \end{aligned}$$

Consequently,

$$\begin{aligned} Q_n(h) &= 2(\beta_u^n - \beta_l^n)n\mathbb{P}_n(Y - f_n(\theta^n))\{1(X \leq \theta^n + h_1/n) - 1(X \leq \theta^n)\} \\ &\quad + 2\sqrt{n}[\mathbb{P}_n(Y - \beta_l^n)1\{X \leq \theta^n + h_1/n\}]h_2 - \mathbb{P}_n1\{X \leq \theta^n + h_1/n\}h_2^2 \\ &\quad + 2\sqrt{n}[\mathbb{P}_n(Y - \beta_u^n)1\{X > \theta^n + h_1/n\}]h_3 - \mathbb{P}_n1\{X > \theta^n + h_1/n\}h_3^2 \\ &= T_{1n}(h_1) + \hat{T}_{2n}(h_1, h_2) + \hat{T}_{3n}(h_1, h_3), \end{aligned}$$

where

$$\begin{aligned} T_{1n}(h_1) &= 2(\beta_u^n - \beta_l^n)n\mathbb{P}_n(Y - f_n(\theta^n))\{1(X \leq \theta^n + h_1/n) - 1(X \leq \theta^n)\}, \\ \hat{T}_{2n}(h_1, h_2) &= 2\sqrt{n}[\mathbb{P}_n(Y - \beta_l^n)1(X \leq \theta^n + h_1/n)]h_2 - \mathbb{P}_n1\{X \leq \theta^n + h_1/n\}h_2^2 \end{aligned}$$

and

$$\hat{T}_{3n}(h_1, h_3) = 2\sqrt{n}[\mathbb{P}_n(Y - \beta_u^n)1(X > \theta^n + h_1/n)]h_3 - \mathbb{P}_n1\{X > \theta^n + h_1/n\}h_3^2.$$

We now define $\tilde{Q}_n(h)$ as follows:

$$\begin{aligned} \tilde{Q}_n(h) &= T_{1n}(h_1) + 2\sqrt{n}\mathbb{P}_n[\varepsilon 1\{X \leq \theta^n\}]h_2 - \mathbb{P}_n\{X \leq \theta^n\}h_2^2 \\ &\quad + 2\sqrt{n}\mathbb{P}_n[\varepsilon 1\{X > \theta^n\}]h_3 - \mathbb{P}_n\{X > \theta^n\}h_3^2 \\ &= T_{1n}(h_1) + T_{2n}(h_2) + T_{3n}(h_3) = \mathbb{P}_n T_1(h_1) + \mathbb{P}_n T_2(h_2) + \mathbb{P}_n T_3(h_3), \end{aligned}$$

where

$$\begin{aligned} T_1(h_1) &= 2n(\beta_u^n - \beta_l^n)(Y - f_n(\theta^n))\{1(X \leq \theta^n + h_1/n) - 1(X \leq \theta^n)\}, \\ T_2(h_2) &= 2\sqrt{n}[\varepsilon 1\{X \leq \theta^n\}]h_2 - 1\{X \leq \theta^n\}h_2^2 \quad \text{and} \\ T_3(h_3) &= 2\sqrt{n}[\varepsilon 1\{X > \theta^n\}]h_3 - 1\{X > \theta^n\}h_3^2. \end{aligned}$$

In Lemma 3.3 below, we show that $Q_n(h)$ and $\tilde{Q}_n(h)$ are uniformly close, as random elements in the space \mathcal{D}_K , where \mathcal{D}_K , $K \subset \mathbb{R}^3$ is the space of functions $q : K \mapsto \mathbb{R}$, K being a compact rectangle in \mathbb{R}^3 . Such functions $w(h_1, h_2, h_3)$ are piece-wise constant, hence, cadlag in the first argument, h_1 , and are continuous in the last two arguments (h_2, h_3) . For each compact interval C in \mathbb{R} , define Λ_C

to be the collection of continuous, strictly increasing maps $\lambda : C \mapsto C$ such that $\lambda(C) = C$. Similar to Seijo and Sen (2011), define a norm on Λ_C as follows:

$$(3.5) \quad \lambda \mapsto \|\lambda\| \equiv \sup_{s \neq t, s, t \in C} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right|.$$

Note that $K = I \times A$, necessarily, for a two-dimensional compact rectangle A and a compact interval I . For $w_1, w_2 \in \mathcal{D}_K$, we define the Skorohod topology as the one induced by the metric

$$d_K(w_1, w_2) \equiv \inf_{\lambda \in \Lambda_I} \left\{ \sup_{u \in K} |w_1(u_1, u_2, u_3) - w_2(\lambda(u_1), u_2, u_3)| + \|\lambda\| \right\}.$$

Endowed with this metric, \mathcal{D}_K is a complete separable metric space.

LEMMA 3.3. *Under conditions B–C2, $Q_n - \tilde{Q}_n = o_P^K(1)$ in (\mathcal{D}_K, d_K) for each K above. The superscript K in $o_P^K(1)$ indicates that the norm of the error is in terms of d_K .*

To obtain the limit distribution of $Q_n(h)$, we next establish the uniform tightness of $\{\tilde{Q}_n\}_{n=1}^\infty$.

LEMMA 3.4. *The process $\{\tilde{Q}_n\}_{n=1}^\infty$ is uniformly tight.*

We now define the limit process. Let $\{v^+(h) : h \geq 0\}$ be a homogeneous Poisson process on $[0, \infty)$ with right continuous and left limit (in short RCLL) sample paths and rate parameter $p_X(\theta^0)$. Let $\{\varepsilon_i\}_{i=1}^\infty$ be i.i.d. versions of ε and distributed independently of $v^+(h)$. Let S_i denote the time to the i th arrival for the Poisson process v^+ , that is, $S_i = R_1 + R_2 + \dots + R_i$, where $\{R_j\}_{j=1}^\infty$ are the i.i.d. exponential inter-arrival times corresponding to $v^+(h)$. For $h \geq 0$, define

$$\Lambda_1(h) = \sum_{j=0}^{v^+(h)} (\varepsilon_j + f(S_j + \xi^0) - f(\xi^0)).$$

To define the process for $h \leq 0$, generate $v^-(h)$, an LCRR (left continuous with right limit) homogeneous Poisson process on $[0, \infty)$ with parameter $p_X(\theta^0)$ and $\{\tilde{\varepsilon}_i\}_{i=1}^\infty$ i.i.d. ε again, and independent of $v^-(h)$. Also, v^- and the $\tilde{\varepsilon}_i$'s are generated independently of v^+ and ε_i 's. Let \tilde{S}_i denote the time to the i th arrival for the process v^- . For $h \leq 0$, define

$$\Lambda_1(h) = \sum_{j=0}^{v^-(-h)} (-\tilde{\varepsilon}_j + f(\xi^0) - f(-\tilde{S}_j + \xi^0)).$$

It can be easily seen that the process $\Lambda_1(h)$, thus defined, has independent increments.

We now show that on every compact rectangle K , $\tilde{Q}_n(u)$ converges to the tight process $Q(u)$ in the d_K metric, where

$$\begin{aligned} Q(u) &= 2(\beta_u^0 - \beta_l^0)\Lambda_1(u_1) + 2Z_1u_2 + u_2^2P\{X \leq \theta^0\} + 2Z_2u_3 + u_3^2P\{X > \theta^0\} \\ &\equiv 2(\beta_u^0 - \beta_l^0)\Lambda_1(u_1) + \Lambda_2(u_2) + \Lambda_3(u_3), \end{aligned}$$

where Z_1 and Z_2 are mean zero independent Gaussians with respective variances $\sigma^2P\{X \leq \theta^0\}$ and $\sigma^2P\{X > \theta^0\}$ and Z_1 , Z_2 and Λ_1 are all independent. The result is summarized in Theorem 3.4.

THEOREM 3.4. *Under assumptions A–C, when $\alpha_n = n$, the process \tilde{Q}_n converges weakly to Q in \mathcal{D}_K for every compact rectangle K in \mathbb{R}^3 . Furthermore, via a continuous mapping argument, $\hat{h}_n \rightarrow_d h^*$, where*

$$h^* = (\hat{h}_1^*, \hat{h}_2^*, \hat{h}_3^*)^T = \arg \min_{h \in \mathbb{R}^3} Q(h).$$

Also, $n(\hat{\theta}^n - \theta^n) = \arg \min_{h_1} T_{1n}(h_1) + o_P(1)$ and converges weakly to \hat{v}_{Λ_1} , where $\hat{v}_{\Lambda_1} = \inf\{v : \Lambda_1(v) = \min_v \Lambda_1\}$, while $\sqrt{n}(\hat{\beta}_l^n - \beta_l^n)$ and $\sqrt{n}(\hat{\beta}_u^n - \beta_u^n)$ converge weakly to mean zero Gaussian variables with variances $\sigma^2/P(X \leq \theta^0)$ and $\sigma^2/P(X > \theta^0)$, respectively. Finally, $n(\hat{\theta}^n - \theta^n)$, $\sqrt{n}(\hat{\beta}_l^n - \beta_l^n)$ and $\sqrt{n}(\hat{\beta}_u^n - \beta_u^n)$ are asymptotically independent.

REMARK. Note that, by the argmin of Q , we mean the *smallest* argmin as with M_n in Section 2, since there may be multiple minimizers with differing values of the first co-ordinate.

The next result is the asymptotic distribution for $n = o(\alpha_n)$, when the rate of the rescaling parameter α_n going to infinity, that is, the speed that the working model approaches the true model, is even faster than n . In this scenario, the obtained limiting distribution is identical with that obtained under correct specification, that is, when the true model is $f_0(x; \psi) = \beta_l^0 1(x \leq \theta^0) + \beta_u^0 1(x > \theta^0)$, the limit of the regression functions considered in this paper. The arguments for this case follow exactly the same pattern as the case $n = \alpha_n$, so we omit the details and only describe the limit process and the asymptotic convergence results. Note that the rate of convergence in the two cases: $\alpha_n = n$ and $n = o(\alpha_n)$ are identical, and \hat{h}_n and Q_n are therefore defined in the exact same way as for the case $\alpha_n = n$.

Recall that $\{v^+(h) : h \geq 0\}$ is a homogeneous Poisson process on $[0, \infty)$ with right continuous and left limit (RCLL) sample paths and rate parameter $p_X(\theta^0)$ and $\{\varepsilon_i\}_{i=1}^\infty$ are i.i.d. versions of ε and distributed independently of $v^+(h)$. For $h \geq 0$, define

$$\Lambda(h) = \sum_{j=0}^{v^+(h)} (\varepsilon_j + \beta_u^0 - f(\xi^0)).$$

To define the process for $h \leq 0$, again consider $\nu^-(h)$ and $\{\tilde{\varepsilon}_i\}_{i=1}^\infty$, exactly as defined before and independent of $\nu^+(h)$ and $\{\varepsilon_i\}$. For $h \leq 0$, define

$$\Lambda(h) = \sum_{j=0}^{\nu^-(h)} (-\tilde{\varepsilon}_j + f(\xi^0) - \beta_l^0).$$

It is easily seen that the process $\Lambda(h)$ has independent increments. Also, note that the process only depends on f through its limits at $-\infty$ and ∞ : this follows by recalling that $f(\xi_0) = (\beta_l^0 + \beta_u^0)/2$. The proof of the below theorem is skipped owing to its similarities to the proof of Theorem 3.4.

THEOREM 3.5. *Under assumptions A–C, when $n = o(\alpha_n)$, $n(\hat{\theta}_n - \theta^n)$, $\sqrt{n}(\hat{\beta}_l^n - \beta_l^n)$ and $\sqrt{n}(\hat{\beta}_u^n - \beta_u^n)$ are asymptotically independent. Furthermore, $n(\hat{\theta}_n - \theta^n) = \arg \min_h T_{1n}(h_1) + o_P(1)$ and converges weakly to $2(\beta_u^0 - \beta_l^0)\hat{\nu}_\Lambda$, where $\hat{\nu}_\Lambda = \inf\{\nu : \Lambda(\nu) = \arg \min \Lambda\}$, while $\sqrt{n}(\hat{\beta}_l^n - \beta_l^n)$ and $\sqrt{n}(\hat{\beta}_u^n - \beta_u^n)$ converge weakly to mean zero Gaussian variables with variances $\sigma^2/F_X(\theta^0)$ and $\sigma^2/(1 - F_X(\theta^0))$, respectively.*

4. Connections among the different limit distributions. The goal in this section is to explore the connections between the three limiting regimes that arise when considering the behavior of $\hat{\theta}_n - \theta^n$ (appropriately normalized) for different values of α_n . For $\alpha_n = o(n)$, we get Chernoff’s distribution, up to a constant, whereas minimizers of two-sided compound Poisson processes appear in the other two cases. For $\alpha_n = n$, the limit distribution depends on the entire function f , whereas for $n = o(\alpha_n)$, the distribution depends only on the limiting change-point model f_0 . We show below that the distribution in the intermediate case, $\alpha_n = n$, belongs to a family of “boundary distributions” that can transition, at least qualitatively, to each of the other two limits. For easy exposition, we first restrict attention to the following *one-parameter* version of our problem. The case where β_l and β_u are unknown will be discussed later.

At stage n , consider the model $Y = f(\alpha_n(X - \theta_0)) + \varepsilon$ with the levels β_l^0 and β_u^0 assumed *known*. We estimate θ_0 by

$$\bar{\theta}_n := \arg \min_{\theta} \mathbb{P}_n[(Y - \beta_l^0)^2 1(X \leq \theta) + (Y - \beta_u^0)^2 1(X > \theta)] \equiv \arg \min \mathbb{M}_n(\theta),$$

where $\mathbb{M}_n(\theta) = \mathbb{P}_n[(Y - 1/2)1(X \leq \theta)]$, the equivalence of the two criterion functions being a consequence of some simple algebra. As before, the smallest argmin is used.

The population version of $\mathbb{M}_n(\theta)$ is given by: $M_n(\theta) = P_n[(Y - 1/2)1(X \leq \theta)]$ and $\tilde{\theta}^n = \arg \min_{\theta} M_n(\theta)$. As in the 3-parameter problem, let $\xi_0 = f^{-1}((\beta_l^0 + \beta_u^0)/2)$, let $a_0 = \sqrt{\sigma^2 p_X(\theta_0)}$ and $b_0 = p_X(\theta_0) f'(\xi_0)/2$. It is not difficult to check that $\tilde{\theta}^n = \theta_0 + (1/\alpha_n)\xi_0$. The following theorem gives the distribution of $\bar{\theta}_n$ under the different regimes.

THEOREM 4.1. *In the above one-parameter model:*

(a) *when* $\alpha_n = o(n)$,

$$n^{1/3} \alpha_n^{2/3} (\bar{\theta}_n - \tilde{\theta}^n) \rightarrow_d L \equiv \arg \min_h (a_0 W(h) + b_0 h^2);$$

(b) *when* $\alpha_n = n$,

$$n(\bar{\theta}_n - \tilde{\theta}^n) \rightarrow_d \arg \min_h \Lambda(h),$$

where

$$\begin{aligned} \Lambda(h) = & \left\{ \sum_{j=0}^{v^+(h)} (\varepsilon_j + f(\xi_0 + S_j) - f(\xi_0)) \right\} 1(h \geq 0) \\ & + \left\{ \sum_{j=0}^{v^-(-h)} (-\tilde{\varepsilon}_j + f(\xi_0) - f(\xi_0 - \tilde{S}_j)) \right\} 1(h < 0), \end{aligned}$$

where S_j 's and \tilde{S}_j 's are as defined previously;

(c) *when* $n = o(\alpha_n)$,

$$n(\bar{\theta}_n - \tilde{\theta}_n) \rightarrow_d \arg \min_h \tilde{\Lambda}(h),$$

where

$$\begin{aligned} \tilde{\Lambda}(h) = & \left\{ \sum_{j=0}^{v^+(h)} (\varepsilon_j + \beta_u^0 - f(\xi_0)) \right\} 1(h \geq 0) \\ & + \left\{ \sum_{j=0}^{v^-(-h)} (-\tilde{\varepsilon}_j + f(\xi_0) - \beta_l^0) \right\} 1(h < 0). \end{aligned}$$

REMARK. Note that the limit distributions in (b) and (c) are identical to those obtained for $n(\hat{\theta}_n - \theta^n)$ in the 3-parameter problem, while the limit distribution in case (a) is *different*: the constant b_0 in the drift term is *larger* than b that shows up in the three-parameter problem; see Theorem 3.3. The smaller b leads to a larger variance in the 3-parameter problem, the price of having to estimate the levels β_l^0 and β_u^0 . In the settings (b) and (c), the estimation of the levels has no effect on the distribution of the change-point since the level estimates are asymptotically independent of the change-point estimate and, therefore, the distributions in the 1-parameter and 3-parameter problems coincide. The proof of the above theorem is skipped as it involves easier versions of the arguments required to prove the distributional results in the 3-parameter problem.

We now introduce a family of processes $\{\Lambda_c\}_{c>0}$ that generalizes the process Λ appearing in the central case, (b). For $c > 0$, define

$$\Lambda_c(h) = \left\{ \sum_{j=0}^{v^+(h)} (\varepsilon_j + f(\xi_0 + cS_j) - f(\xi_0)) \right\} 1(h \geq 0) + \left\{ \sum_{j=0}^{v^-(-h)} (-\tilde{\varepsilon}_j + f(\xi_0) - f(\xi_0 - c\tilde{S}_j)) \right\} 1(h < 0).$$

The parameter c that dictates the above family is a scale parameter that regulates the shift of the increments of the generalized compound Poisson process Λ_c . An instructive (statistical) way of thinking about Λ_c is to consider the model: $Y = f_c(n(X - \theta_0)) + \varepsilon$, with $f_c(t) \equiv f(ct)$. By calculations similar to those needed to prove Theorem 4.1, we can show that

$$(4.1) \quad n(\bar{\theta}_{n,c} - \tilde{\theta}^{n,c}) \rightarrow_d \arg \min_h \Lambda_c(h),$$

where $\bar{\theta}_{n,c}$ and $\tilde{\theta}^{n,c}$ are the analogues of $\bar{\theta}_n$ and $\tilde{\theta}^n$ in the one-parameter model above, which corresponds to $c = 1$.

The following results show that the distribution of the minimizer of Λ_c approximates the limit distributions in the cases (a) and (c), as c approaches 0 and ∞ , respectively, for the one-parameter problem.

THEOREM 4.2. *Under assumptions A–C, as $c \rightarrow 0$,*

$$c^{2/3} \arg \min_h \Lambda_c(h) \rightarrow_d L \equiv \arg \min_h [a_0 W(h) + b_0 h^2].$$

THEOREM 4.3. *Under assumptions A–C, as $c \rightarrow \infty$,*

$$\arg \min_h \Lambda_c(h) \rightarrow_d \tilde{L} \equiv \arg \min_h \tilde{\Lambda}(h).$$

Heuristically, Theorem 4.3 is somewhat easier to visualize. As $c \rightarrow \infty$, for every j , $f(\xi_0 - c\tilde{S}_j)$ goes to 0 almost surely and $f(\xi_0 + cS_j)$ to 1 almost surely, and by putting in these limiting values in the expression for Λ_c we recover the process $\tilde{\Lambda}$. This is not a rigorous verification, as we need to show that the convergence of the processes happens in a strong enough topology for distributional convergence of the argmin functional. This is accomplished in the proof of Theorem 4.3. As far as Theorem 4.2 is concerned, the crux of the argument lies in showing that an appropriately scaled version of Λ_c (where scaling appears in the magnitude of the process as well as its argument) converges to a Brownian motion plus a quadratic drift; see Theorem 0.4 in Song, Banerjee and Kosorok (2015).

Define the sequence $c_n := \alpha_n/n$. Consider first, case (c): $n = o(\alpha_n)$, where the statistical model can be written as $Y = f_{c_n}(n(X - \theta_0)) + \varepsilon$ with $c_n \rightarrow \infty$.

By (4.1), conclude that the distribution of $n(\bar{\theta}_n - \tilde{\theta}^n)$ can be approximated by that of $\arg \min \Lambda_{c_n}(h)$. This, of course, is consistent with what we learn in Theorems 4.3 and 4.1: as c_n grows large in this case, by Theorem 4.3, $\arg \min \Lambda_{c_n}(h)$ and $\arg \min_h \tilde{\Lambda}(h)$ are close in a distributional sense, and the latter is indeed the limit of $n(\bar{\theta}_n - \tilde{\theta}^n)$ in case (c) of Theorem 4.1.

Next, consider case (a): $\alpha_n = o(n)$. As above, using (4.1), conclude that the distribution of $n(\bar{\theta}_n - \tilde{\theta}^n)$ can be approximated by that of $\arg \min \Lambda_{c_n}(h)$, as well. Since c_n becomes small in this case, by Theorem 4.2, this can be approximated by $c_n^{-2/3}L$, which is essentially what part (a) of Theorem 4.1 tells us. Thus, the family $\{\Lambda_c\}$ provides a uniform approximation to the limit distributions across the three different situations.

In the 3-parameter problem, when $\alpha_n = o(n)$, we know from Theorem 3.3 that $n^{1/3}\alpha_n^{2/3}(\hat{\theta}_n - \theta^n) \rightarrow_d \arg \min_h (a_0W(h) + bh^2) := L'$, and L' and L have different distributions. The distribution of $n(\hat{\theta}_n - \theta^n)$ can then be approximated by that of $c_n^{-2/3}L'$. Noting that $L \equiv_d (a_0/b_0)^{2/3}\mathbb{C}$ and $L' \equiv_d (a_0/b)^{2/3}\mathbb{C}$, where $\mathbb{C} = \arg \min_h (W(h) + h^2)$ is the Chernoff random variable, the distribution of $n(\hat{\theta}_n - \theta^n)$ can be approximated by that of $c_n^{-2/3}(b_0/b)^{2/3}L$ and, therefore, by $(b_0/b)^{2/3} \arg \min \Lambda_{c_n}(h)$. With $n = o(\alpha_n)$, it is not difficult to see that the distribution of $n(\hat{\theta}_n - \theta^n)$ in the 3-parameter case can still be approximated by $\arg \min_h \Lambda_{c_n}(h)$, as in the 1-parameter case.

5. Adaptive inference for the change-point parameter. Inference on θ^n when α_n is known can be achieved through subsampling or the “ m out of n ” bootstrap. To perform adaptive inference when α_n is *unknown*, which is the case in practice, it is important to estimate it reliably. To this end, we resort to the results in Bertail, Politis and Romano (1999) who proposed a subsampling procedure when the convergence rate is unknown: the key idea is to use the data to first construct an estimate of the rate of convergence and then use this *estimated rate* to produce a confidence interval for the parameter of interest. Following their idea, we describe an adaptive inference procedure for θ^n when α_n is *unknown*.

Consider $\alpha_n = n^\gamma$ where $0 \leq \gamma < \infty$. (We restrict ourselves to this polynomial class as this covers essentially all interesting regimes and is tractable to deal with using the suggested method.) By the asymptotic results of the previous section, we know that $n^{1/3}n^{2\zeta/3}(\hat{\theta}_n - \theta^n)$ converges to a tight random variable, say L , where $\zeta = \gamma \wedge 1$, the minimum of γ and 1. To construct a level $1 - \alpha$ CI, we proceed thus:

(1) Pick subsample sizes $n_1 < n_2 < n$ where $n_j = n^{\beta_j}$, with $1 > \beta_2 > \beta_1 > 0$. For each $j = 1, 2$, collect a subsample of size n_j without replacement r times and run the change-point estimation procedure these r subsamples to obtain change-point estimates $\{\hat{\theta}_{jk}^*\}_{k=1}^r$.

(2) Next, note that for each j , the empirical distribution of the $\{n_j^{(1/3+2\zeta/3)} \times (\hat{\theta}_{jk}^* - \hat{\theta}_n)\}_{k=1}^r$ (conditional on the given data) approximates the distribution of L . Using a moment approximation, we can then write

$$\frac{1}{r} \sum_{k=1}^r n_j^{(1/3+2\zeta/3)} |\hat{\theta}_{jk}^* - \hat{\theta}_n| \approx E(|L|), \quad j = 1, 2,$$

and, therefore,

$$\log E(|L|) \approx \log \left[\frac{1}{r} \sum_{k=1}^r |\hat{\theta}_{jk}^* - \hat{\theta}_n| \right] + \frac{1+2\zeta}{3} \log n_j, \quad j = 1, 2.$$

(3) Equating the right-hand side of the above display for $j = 1$ to that for $j = 2$, a natural estimate of ζ is found by solving the equation

$$\frac{1+2\hat{\zeta}}{3} = \left\{ \log \left(\frac{n_2}{n_1} \right) \right\}^{-1} \left[\log \left\{ \frac{1}{r} \sum_k |\hat{\theta}_{1k}^* - \hat{\theta}_n| \right\} - \log \left\{ \frac{1}{r} \sum_k |\hat{\theta}_{2k}^* - \hat{\theta}_n| \right\} \right].$$

This formula is essentially the same as that in Bertail, Politis and Romano (1999) immediately preceding Theorem 1 (of that paper), with the only difference being that here we use a moment functional instead of a quantile functional.⁴

(4) Estimate the $\alpha/2$ th and $(1 - \alpha/2)$ th quantiles of L , say $q_{\alpha/2}^*$ and $q_{(1-\alpha)/2}^*$ from the empirical distribution of $n_j^{(1+2\hat{\zeta})/3} (\hat{\theta}_j^* - \hat{\theta}_n)$ conditional on the data (either for $j = 1$ or 2). This can be done by drawing a new set of subsamples of size n_j from the original data.

(5) An approximate level $1 - \alpha$ CI for θ^n is $[\hat{\theta}_n - q_{(1-\alpha)/2}^* n^{-(1+2\hat{\zeta})/3}, \hat{\theta}_n - q_{\alpha/2}^* n^{-(1+2\hat{\zeta})/3}]$.

As this is a simple adaptation of an established procedure, we have not presented extensive simulation studies in the paper. However, we present results and figures from limited simulation studies to provide a feel for the procedure. Data are generated from the model $Y = f_n(X) + \varepsilon$, where

$$f_n(x) = \frac{\exp(n^\gamma(x - 0.5))}{1 + \exp(n^\gamma(x - 0.5))},$$

the random noise ε follows a normal distribution with mean zero and standard deviation 1.8 and the covariate X follows a uniform distribution on $[0, 1]$. The sample size n is taken to be 2000. Three values of γ : 0.8, 1 and 1.2 are considered to account for each of the three regimes. For demonstration purposes, we fit a one-parameter change-point model using $\beta_l^0 = 0$ and $\beta_u^0 = 1$ (as at the beginning

⁴For our problem, we found the moment functional to produce somewhat stabler estimates of ζ as compared to the quantiles.

of Section 4) and only estimate the change-point parameter. Given a dataset of size 2000 from the above model with parameter γ , to estimate $\hat{\zeta}$, we consider subsample sizes $n_1 = 100$ and $n_2 = 200$, draw $r = 1000$ subsamples for each subsample size and then apply the formula in step (3) above. The process is repeated for 200 datasets, resulting in 200 estimates $\{\hat{\zeta}_i\}_{i=1}^{200}$ and their median $\hat{\zeta}_m$ is chosen as the final estimate of ζ . For the three settings, the estimated values are 0.76, 1 and 1, respectively, the corresponding true ζ 's being 0.87, 1 and 1.

Figure 1 presents QQ-plots of the empirical distribution of $n^\zeta(\hat{\theta}_n - \theta^n)^5$ (based on 1000 independent datasets) versus the empirical subsampling distribution $\tilde{n}^{\hat{\zeta}_m}(\hat{\theta}^* - \hat{\theta}_n)$ based on 1000 subsamples of size $\tilde{n} = 100$. For each of the three regimes, the plots show an approximate alignment with the $y = x$ line as would be expected. The empirical coverage probabilities for θ^n using subsampling based 95% CIs and the $\hat{\zeta}_m$ values from the previous paragraph [see the formula for the CI in step (5) above] on 200 new datasets (with $n = 2000$) are found to be 96.5%, 93.5% and 93%, respectively, for the three regimes.

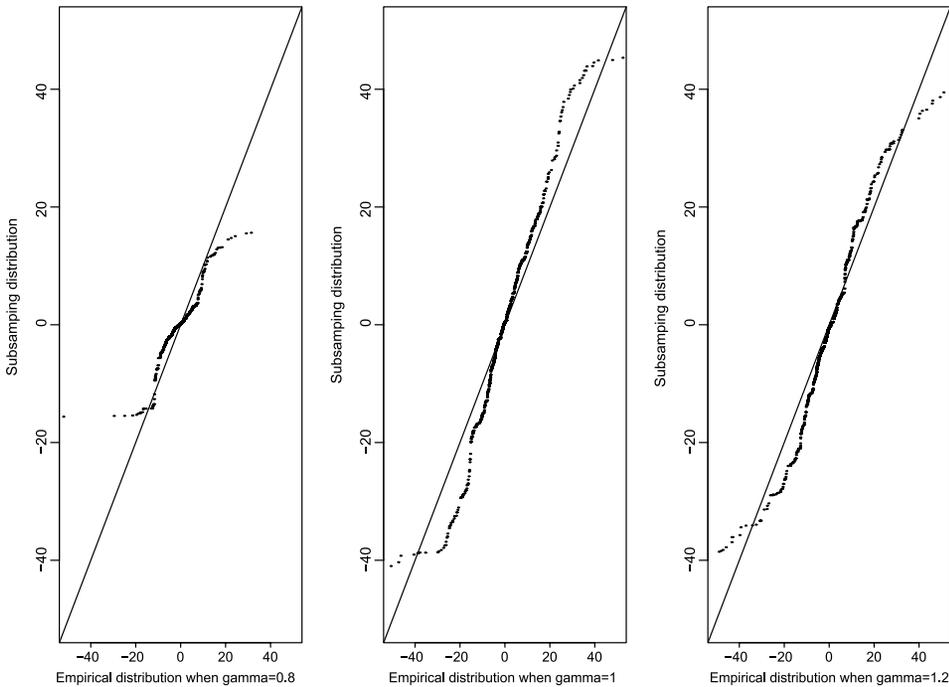


FIG. 1. The QQ-plots of subsampling empirical distributions versus the three empirical distributions for sample size $n = 2000$ and subsample size 100. The true convergence rates are $n^{0.867}$, n and n , respectively. The straight line is a 45 degree line through the origin.

⁵For the 1-parameter model θ^n is referred to as $\tilde{\theta}_n$ in Section 4.

To demonstrate the performance of the *fully adaptive* 5-step procedure described earlier in the section, we discuss results from a second simulation experiment from the model above with $\gamma = 0.5, 0.75, 1$ and 2 and sample size $n = 1500$. Given a dataset of size 1500 from the regression model with parameter γ , in the first step, we resample the data 5000 times with subsample sizes $n_1 = 100$ and $n_2 = 200$, respectively. For each subsample size, we evenly split these 5000 subsamples into 10 groups. We then compute 10 estimates of ζ via the formula in the step 3 of the above 5-step procedure, each estimate using 500 (this is the r from the general description of the procedure) subsamples of size 100 and 500 of size 200 and prescribe the median of these estimates as the value of $\hat{\zeta}$ to be used for the construction of the confidence interval for θ^n . (Using the median provides additional stability to the estimation of ζ .) The confidence interval construction (steps 4 and 5) uses 500 additional subsamples from the same dataset with subsample size 100 . The empirical confidence intervals for the four scenarios based on 200 independent datasets (average lengths of the 200 CIs in brackets) are $94\%(0.129)$, $96.5\%(0.086)$, $97\%(0.086)$ and $95\%(0.081)$, respectively, providing numerical evidence of the proposed adaptive inference procedure. Note that the coverages reported in this paragraph are more realistic than the ones in the previous paragraph, since the CI for each dataset is based on an estimate of ζ *computed from the same dataset* as is always the case in a real application.

The adaptive procedure is computationally fairly intensive owing to the estimation of the convergence rate for each sample. Also note that the subsampling procedure, by the very nature of it, involves tuning parameters (the subsample sizes) and this typically plays an important role in the reliability of the results [see the discussion toward the end of Section 4 in Bertail, Politis and Romano (1999)]. Further investigations to fine tune and objectify the selection of subsample size in the context of subsampling with unknown convergence rates in general problems, and more specifically, in the change point problem we study in this paper would be very interesting but fall outside the scope of the current paper.

6. Simulation studies. In this section, we provide detailed empirical evidence of the theoretical results. Our framework stipulates a sequence of models changing with n and converging to a limiting stump model with α_n regulating how fast the regression functions converge to a stump. We view a changing sequence of models with a given α_n as an asymptotic framework within which to couch a *given fixed regression problem*: $Y = v_0(X) + \varepsilon$ with n data points available, and ask the question: which asymptotic regime: “slow” (i.e., $\alpha_n < n$), “intermediate” ($\alpha_n = n$) or “fast” ($\alpha_n > n$) provides the best description of the behavior of $\hat{\theta}_n$, the least squares estimate of the population parameter $\theta_{0,v}$. Note that the population parameters $(\beta_{l,v}, \beta_{u,v}, \theta_{0,v})$ are given by the minimizer of $E(Y - \beta_l 1(X \leq \theta) - \beta_u 1(X > \theta))^2$ over all $(\beta_l, \beta_u, \theta)$, the expectation being taken with respect to the joint distribution of (Y, X) in the above regression model.

In the interests of a clean exposition, we restrict ourselves to three specific regimes: the one corresponding to $\alpha_n = c_0$ for a constant c_0 , that is, the fixed function set-up of Banerjee and McKeague (2007) which yields a Chernoff limit, the one with $\alpha_n = n$ that gives the intermediate distribution (Theorem 3.4) and the last with $\alpha_n > n$, which produces the compound Poisson process limit which also arises when the true regression model is a fixed change-point model (Theorem 3.5). The case $\alpha_n = c_0$ should be viewed as a representative of the slow regimes corresponding to $\alpha_n < n$: recall that all slow regimes lead to a multiple of Chernoff's distribution, albeit with different convergence rates.

We generate data from the model $Y = v_0(X) + \varepsilon$, where

$$v_0(x) = \frac{\exp(M(x - 0.5))}{1 + \exp(M(x - 0.5))},$$

the random noise ε follows a normal distribution with mean zero and standard deviation 0.6, and M is a constant that we vary for different simulation settings as explained below. The covariate X follows a piecewise uniform distribution on $[0, 1]$ and is symmetric about 0.5. To demonstrate the effect of the density p_X on the limiting distribution, we consider two scenarios for generating the covariate X . In scenario 1, the density is 8 on $[0.45, 0.55]$, while in scenario 2, the density is 4 on $[0.4, 0.6]$. For scenario 1, we consider four different values of M : 35, 60, 100, 1000 and for scenario 2, M assumes values 20, 60, 100, 1000. Larger values of M produce steeper sigmoidal curves which are closer to a change point model than smaller values. For each combination of p_X and M (leading to 8 settings), we generate data for sample sizes $n = 50, 100, 1000, 4000$; for each sample, we generate 500 datasets (replicates) to get the empirical distribution of $\hat{\theta}^n$. For the limiting distribution based on the fixed function setting, the normalized least squares estimate $n^{1/3}(\hat{\theta}^n - \theta^n)$ is calibrated against the appropriate Chernoff limit; see Theorem 2.1 of Banerjee and McKeague (2007). For the limiting distribution based on $\alpha_n = n$, we calibrate $n(\hat{\theta}^n - \theta^n)$ against the quantiles of the argmin of Λ in Theorem 3.4; while, for the third case, we calibrate $n(\hat{\theta}^n - \theta^n)$ against the quantiles of the argmin of Λ_1 in Theorem 3.5. Note that in our asymptotic framework, θ^n is the population minimizer of the change-point parameter at stage n , and since we have a fixed regression model v_0 in our set-up, this is identically equal to the parameter $\theta_{0,v}$.

Simulating from the limit distribution in Theorem 2.1 of Banerjee and McKeague (2007)—the slow regime—requires the $(\beta_l^0, \beta_u^0, d^0)$ appearing in that result. For us, these are simply the population parameters $(\beta_{l,v}, \beta_{u,v}, \theta_{0,v})$ (which depend on M and p_X). To simulate the theoretical limiting distribution based on Theorem 3.4 (intermediate regime), we write our fixed function v_0 as $v_0(x) = f(n(x - 0.5))$ [so as to obtain the representation $Y = f(n(X - 0.5)) + \varepsilon$ based on which the limit is derived]; here, f , of course, becomes dependent on n : $f_n(t) = \frac{\exp(Mt/n)}{1 + \exp(Mt/n)}$. The quantity $\xi_0 = f_n^{-1}((\beta_l^0 + \beta_u^0)/2)$, needed to generate Λ

in Theorem 3.4 is 0.5, since β_u^0 and β_l^0 , the limits of f_n as t goes to ∞ and $-\infty$, respectively, are 1 and 0. To simulate the limit distribution based on Theorem 3.5, we require the value $(\beta_u^0 - \beta_l^0)/2$ in that theorem, and here $\beta_u^0 = 1$ and $\beta_l^0 = 0$ are the levels of the limiting change-point model.

The QQ-plots of the empirical distribution of the normalized least squares estimate based on 500 replicates against that of a sample drawn from the limiting distribution for each regime (the size of the sample from the limiting approximation is 2000 in every simulation setting) are presented in a series of figures: two of these corresponding to scenario 1, $M = 1000$ and scenario 2, $M = 1000$ are presented in Figures 2 and 3 respectively in the main paper and the rest are included in Song, Banerjee and Kosorok (2015). The general pattern is fairly clear: the fast regime provides better approximations at smaller sample sizes than at larger ones, the slow regime improves at higher sample sizes, and the intermediate regime is much more robust to the sample size, though at high sample sizes ($n = 1000, 4000$) the approximation provided by it starts breaking down (see, e.g., the behavior of the intermediate regime for smaller values of M). The fast regime generally completely breaks down at high sample sizes and for smaller values of M (20, 35, 60), which correspond to curves that are farther from a change-point model, tends to behave poorly even for small samples. While the slow regime improves for larger samples, it sometimes provides a decent approximation at smaller samples as well [again, see some of the plots in Song, Banerjee and Kosorok (2015)].

The general pattern can be explained by noting that at small sample sizes the data is typically not adequate to discover the features of the underlying sigmoidal curve; especially for a steep curve (e.g., $M = 1000$ as presented in the paper), at small n , the data only “sees” the change-point type feature and, therefore, an approximation using the fast regime (also the regime for a true change-point model) performs better. For large n , the data is able to “pick out” the overall pattern of the continuous curve quite well and consequently, the setting of Banerjee and McKeague (2007) which deals with fitting a change-point working model to a smooth fixed regression function is apt. The intermediate setting or the boundary case strikes a balance between these two approximations as it uses some features of the underlying regression curve but on the other hand not as local features as the ones used by the asymptotics in Banerjee and McKeague (2007). Hence, it provides an approximation that adapts much better to changes in sample size. This is consistent with the fact that the family of boundary distributions can transition to either of the two extreme limits, as shown in Section 4.

7. Discussion. In this paper, we have studied the asymptotic behavior of change-point models under a wide range of model mis-specification. We end with a discussion of some important aspects of our work and some related problems.

Analogy to classical parametric models: Viewing the f_n 's as a sequence of local alternatives to the limiting null model: the stump function $f_0(x; \psi_0) = \beta_l^0 1(x \leq \theta^0) + \beta_u^0 1(x > \theta^0)$, the phenomena studied in this paper are qualitatively identical

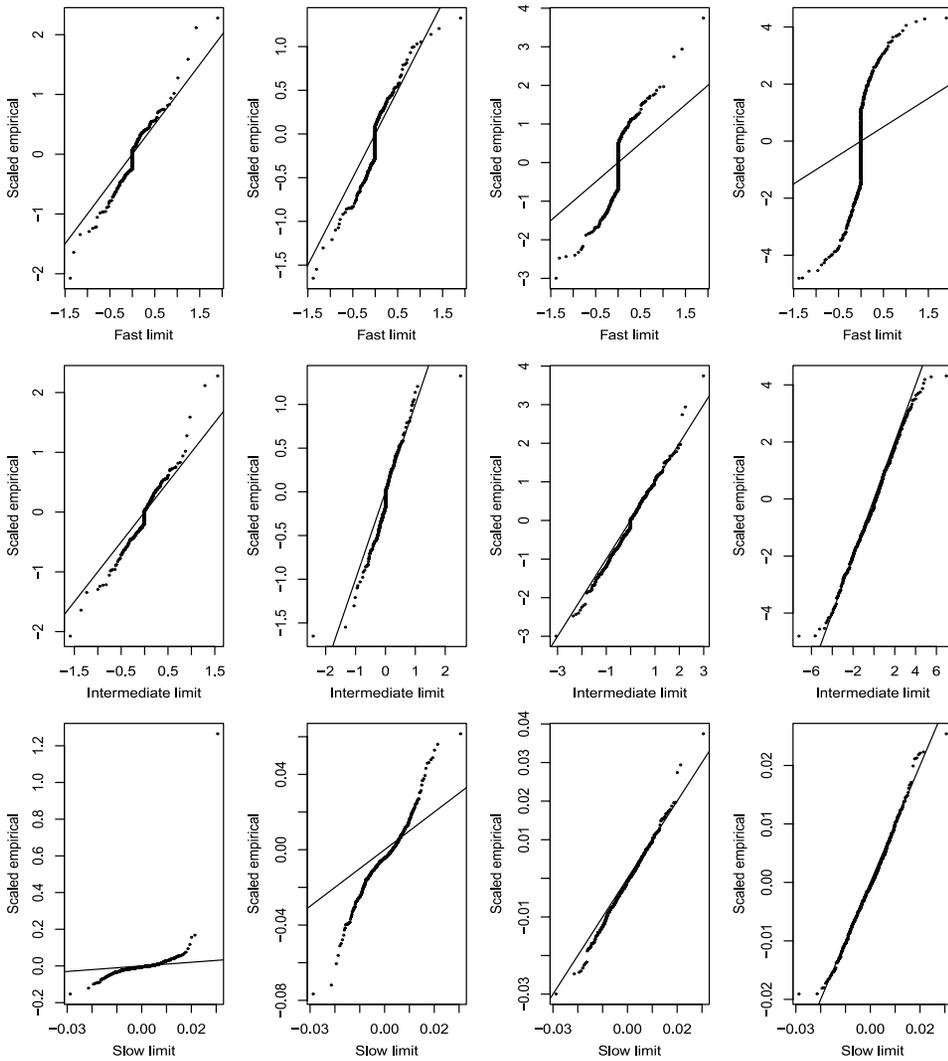


FIG. 2. The QQ -plots of empirical distributions versus the three theoretical limits for $p_X(\theta_0) = 8$ and $M = 1000$. First row and second row: empirical distribution of $n(\hat{\theta}^n - \theta^n)$ vs the fast limit and the intermediate limit, respectively. Third row: empirical distribution of $n^{1/3}(\hat{\theta}^n - \theta^n)$ vs the slow limit. The four columns represent $n = 50, 100, 1000$ and 4000 , respectively. The straight line is a 45 degree line through the origin.

to what transpires with the MLE in regular parametric models under a sequence of local alternatives.

So, consider such a model $\{p(x; \eta)\}$ with the p -dimensional parameter η and let X_1, X_2, \dots, X_n denote i.i.d. observations. Let $\hat{\eta}$ denote the MLE for η . We aim to test the null hypothesis $\eta = \eta_0$. It is well known that under the null, $\sqrt{n}(\hat{\eta} - \eta_0)$ follows an asymptotic normal distribution $N(0, I(\eta_0)^{-1})$, where $I(\eta_0)$

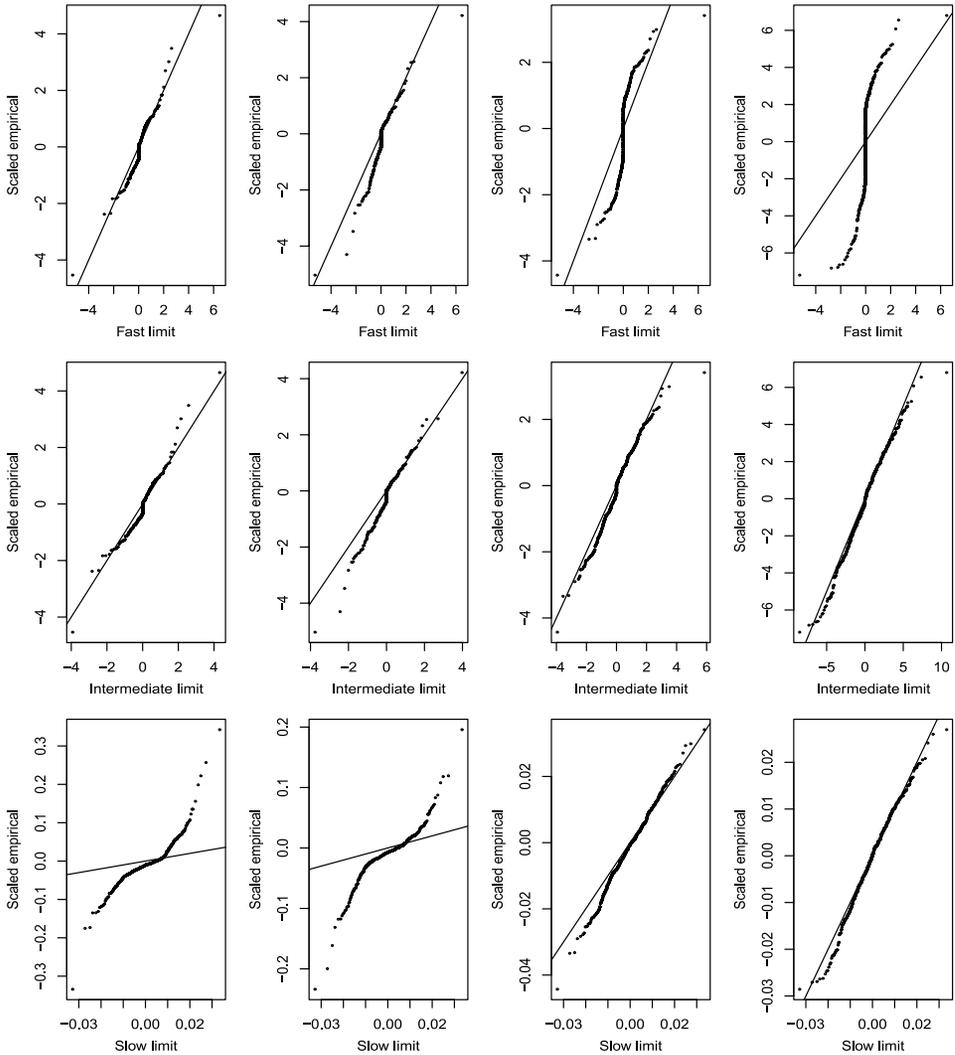


FIG. 3. The QQ-plots of empirical distributions versus the three theoretical limits for $p_X(\theta_0) = 4$ and $M = 1000$. The rest of the captions and notations are the same as those in Figure 2.

is the information matrix for η . With alternatives converging faster than \sqrt{n} , say $\eta_n = \eta_0 + hn^{-\gamma}$ for $\gamma > 1/2$, the limit of $\sqrt{n}(\hat{\eta}_n - \eta_0)$ continues to be identical to that under the null. With alternatives converging at a slower than the “regular” \sqrt{n} rate, that is, when $\gamma < 1/2$, the limit distribution of $\sqrt{n}(\hat{\eta}_n - \eta_0)$ is no longer tight, since the bias term $\sqrt{n}(\eta_n - \eta_0)$ drifts to ∞ . In the change-point problem, $f_0(x, \psi_0)$, of course, plays the role of η_0 , the convergence rate n , which is the natural convergence rate of the least-squares estimate of θ^0 under the null model, plays the role of \sqrt{n} , the f_n 's take on the role of η_n , while α_n becomes the analogue of

$n^{-\gamma}$. As noted in the discussion before the statement of Theorem 3.5, for α_n going to ∞ faster than n (corresponding in the classical case to $\gamma > 1/2$), the asymptotic distribution of $\hat{\theta}^n$ in our problem is identical to that under the null model f_0 . When $\alpha_n = o(n)$ (corresponding in the classical case to $\gamma < 1/2$), Theorem 3.3 in conjunction with Theorem 2.2 tells us that $n^{1/3}\alpha_n^{2/3}(\hat{\theta}^n - \theta_0)$ does not have a tight limit, since the bias term $n^{1/3}\alpha_n^{2/3}(\theta^n - \theta^0)$ goes to ∞ .

It remains to compare the cases where the alternative approaches the null at the natural convergence rate. In the classical scenario, this corresponds to $\gamma = 1/2$ and produces a tight distribution in the limit, namely, $N(h, I(\eta_0)^{-1})$ for $\sqrt{n}(\hat{\eta}_n - \eta_0)$; thus, the direction of approach of the local alternatives figures in the limit. In the change-point scenario, the analogous situation is $\alpha_n = n$, and as Theorem 3.4 shows, now the distribution of $n(\hat{\theta}^n - \theta^0)$ converges to a tight limit which depends upon f , which can be interpreted as the “direction” in which the smooth f_n 's approach the stump f_0 . One important difference between the classical and the change-point scenario is, of course, the differing convergence rates: the α_n parameter influences the rate at which $\hat{\theta}^n$ approaches θ^n in the change-point model, but the γ parameter in the classical scenario does not influence the convergence rate: in fact, $\sqrt{n}(\hat{\eta}_n - \eta_n)$ is $O_p(1)$ in all situations.

An alternative approach for inference: An alternative approach for inference in this problem, kindly brought to our attention by a referee, relies on smoothed least squares estimation along the lines pursued in the papers Seo and Linton (2007) and Seo (2012). Seo and Linton (2007) studies a linear regression model with regime switching, where the form of the linear regression depends on whether a particular subset of covariates lies above or below a hyperplane (whose parameters are also unknown). This can be thought of as a “change-plane” problem. To avoid the nonstandard distributions that would come into play under a regular least squares approach, the authors replace the indicator function appearing in the least squares criterion by a smooth integrated kernel function (analogous to a distribution function) in the spirit of the smoothed maximum score estimator of Horowitz (1992). The corresponding smoothed least squares estimators—even those of the hyperplane parameters—are seen to be asymptotically normal under appropriate conditions on the model and the bandwidth used for the integrated kernel function. Asymptotic normality makes inference more tractable though the rate of convergence is somewhat compromised and can be at most $n^{3/4}$ (up to a logarithmic factor), slower than the n -rate of convergence attained by the regular least squares estimators. While the set-up of Seo and Linton (2007) works under the assumption that the threshold model defined by the hyperplane is true, Seo (2012) (Section 4) explores the behavior of the smoothed least squares estimate under mis-specification in the spirit of our paper and establishes asymptotic normality (Theorem 4), with the rate again depending upon the bandwidth used. These investigations suggest the possibility that using a smoothed least squares approach in

our diminishing mis-specification problem could also lead to asymptotic normality, avoiding the nonstandard distributions that now come into play, at the expense of somewhat reduced convergence rates.

Other potential extensions and connections: A natural question is the extension of this approach to multiple change-points, that is, a situation where the limit of the converging (smooth) models is a piecewise constant function with multiple jumps. It is clear that the properties of the underlying f (i.e., assumptions A through C) which were used to manufacture the converging models would now need to change. Recall that in this paper, the regression function at stage n , $f_n(x) = f(\alpha_n(x - \theta_0))$ and as α_n goes to ∞ , f_n must necessarily converge to a piecewise constant function with a single jump. For example, to take into account the situation where the limiting function is of the form $\alpha_0 1(x < \theta_{10}) + \beta_0 1(\theta_{10} \leq x < \theta_{20}) + \gamma_0 1(\theta_{20} \leq x)$, one possibility for a converging smooth function \tilde{f}_n could be

$$f_n(x) = f_1(\alpha_n(x - \theta_{10})) + f_2(\alpha_n(x - \theta_{20})),$$

for monotone functions f_1, f_2 and appropriate conditions on their limit values at $-\infty$ and ∞ . Note, moreover, that the α_n sitting within the f_2 could be replaced by a different rate parameter ($\beta_n \neq \alpha_n$) going to ∞ . Thus, the multiple change-point problem throws up a number of different challenges which are outside the scope of this paper.

In conclusion, we would like to note an interesting connection of our results to Fryzlewicz (2007), also pointed out by a referee. Section 3.2 of Fryzlewicz (2007) considers approximating functions in different smoothness classes using the Unbalanced Haar transform as basis vectors for the approximating class. These basis functions are piecewise constant by construction and are therefore expected to provide more precise approximations to underlying functions that are structurally similar. Indeed, the result of Theorem 3.1 in that paper shows that when f is in the class of functions of bounded variation the expected IMSE of the Haar transform based estimate attains a rate of $n^{-2/3}$ up to a logarithmic factor that involves the sample size as well as certain features of the approximation basis. On the other hand, when f is in $S[0, 1]$, the space of piecewise constant functions with finitely many jumps, the rate improves to n^{-1} , again up to logarithmic terms and the number of jumps of the function. In our work, the approximating function is a piecewise constant function with a single jump (a stump) and the underlying function a smooth function that can be considered close to a (limiting) stump with a jump at θ_0 . The degree of closeness is measured by the parameter α_n . Our results show that for larger values of α_n , which correspond to the underlying function behaving more like a stump, the rate of convergence of $\hat{\theta}_n$ is faster: for α_n at least as large as n , $|\hat{\theta}_n - \theta_0| = O_p(1/n)$ whereas for $\alpha_n = o(n)$, $|\hat{\theta}_n - \theta_0| = O_p(\alpha_n^{-1})$ and therefore slower than the other case. We note, of course, that in contrast to Fryzlewicz (2007) where a global measure of error is considered, our results are formulated in terms of the convergence of the estimated jump parameter alone.

Acknowledgments. We thank the Editor, the Associate Editor and three referees for providing input and suggestions that vastly improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplemental materials for “Asymptotics for change-point models under varying degrees of mis-specification” (DOI: [10.1214/15-AOS1362SUPP](https://doi.org/10.1214/15-AOS1362SUPP); .pdf). We provide the proofs for all the theoretical results and some figures from the simulation results.

REFERENCES

- BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. [MR1616121](#)
- BAI, J. and PERRON, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18** 1–22.
- BANERJEE, M. and MCKEAGUE, I. W. (2007). Confidence sets for split points in decision trees. *Ann. Statist.* **35** 543–574. [MR2336859](#)
- BASSEVILLE, M. and NIKIFOROV, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ. [MR1210954](#)
- BERTAIL, P., POLITIS, D. N. and ROMANO, J. P. (1999). On subsampling estimators with unknown rate of convergence. *J. Amer. Statist. Assoc.* **94** 569–579. [MR1702326](#)
- BHATTACHARYA, P. K. and BROCKWELL, P. J. (1976). The minimum of an additive process with applications to signal estimation and storage theory. *Z. Wahrsch. Verw. Gebiete* **37** 51–75. [MR0426166](#)
- BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961. [MR1926165](#)
- CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, Chichester. [MR2743035](#)
- FRYZLEWICZ, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Statist. Assoc.* **102** 1318–1327. [MR2412552](#)
- GIJBELS, I., HALL, P. and KNEIP, A. (1999). On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.* **51** 231–251. [MR1707773](#)
- HALL, P. and MOLCHANOV, I. (2003). Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *Ann. Statist.* **31** 921–941. [MR1994735](#)
- HOROWITZ, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60** 505–531. [MR1162997](#)
- HUŠKOVÁ, M. (1999). Gradual changes versus abrupt changes. *J. Statist. Plann. Inference* **76** 109–125. [MR1673343](#)
- KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York. [MR2724368](#)
- KOSOROK, M. R. and SONG, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Ann. Statist.* **35** 957–989. [MR2341694](#)
- LAI, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *J. Roy. Stat. Soc. Ser. B. Stat. Methodol.* **57** 613–658. [MR1354072](#)
- LAI, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statist. Sinica* **11** 303–408. [MR1844531](#)
- LAN, Y., BANERJEE, M. and MICHAELIDIS, G. (2009). Change-point estimation under adaptive sampling. *Ann. Statist.* **37** 1752–1791. [MR2533471](#)

- LOADER, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.* **24** 1667–1678. [MR1416655](#)
- LUND, R. and REEVES, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate* **15** 2547–2554.
- MÜLLER, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761. [MR1165590](#)
- MÜLLER, H.-G. and SONG, K.-S. (1997). Two-stage change-point estimators in smooth regression models. *Statist. Probab. Lett.* **34** 323–335. [MR1467437](#)
- PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.* **31** 442–463. [MR1983537](#)
- RITOV, Y. (1990). Asymptotic efficient estimation of the change point with unknown distributions. *Ann. Statist.* **18** 1829–1839. [MR1074438](#)
- SEIJO, E. and SEN, B. (2011). A continuous mapping theorem for the smallest argmax functional. *Electron. J. Stat.* **5** 421–439. [MR2802050](#)
- SEO, M. H. (2012). Forecasting with a regime-switching model. Unpublished manuscript.
- SEO, M. H. and LINTON, O. (2007). A smoothed least squares estimator for threshold regression models. *J. Econometrics* **141** 704–735. [MR2413485](#)
- SHEN, J. J. and ZHANG, N. R. (2012). Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.* **6** 476–496. [MR2976479](#)
- SONG, R., BANERJEE, M. and KOSOROK, M. (2015). Supplement to “Asymptotics for change-point models under varying degrees of mis-specification.” DOI:[10.1214/15-AOS1362SUPP](#).
- THOMSON, R. E. and FINE, I. V. (2003). Estimating mixed layer depth from oceanic profile data. *Journal of Atmospheric and Oceanic Technology* **20** 319–329.
- VOGT, M. and DETTE, H. (2015). Detecting gradual changes in locally stationary processes. *Ann. Statist.* **43** 713–740. [MR3319141](#)

R. SONG
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
BOX 8203
2311 STINSON DRIVE, 5120 SAS HALL
RALEIGH, NORTH CAROLINA 27695-8203
USA
E-MAIL: rsong@ncsu.edu

M. BANERJEE
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
451, WEST HALL
1085 SOUTH UNIVERSITY
ANN ARBOR, MICHIGAN 48109-1107
USA
E-MAIL: moulib@umich.edu

M. R. KOSOROK
DEPARTMENTS OF BIostatISTICS
AND STATISTICS AND OPERATION RESEARCH
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
3101 MCGAVRAN-GREENBERG HALL
CB 7420
CHAPEL HILL, NORTH CAROLINA 27599-7420
USA
E-MAIL: kosorok@unc.edu