Journal of Econometrics 167 (2012) 274–294

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Likelihood estimation and inference in threshold regression

Ping Yu

Department of Economics, 12 Grafton Road, University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 18 January 2009
Received in revised form
10 April 2011
Accepted 14 December 2011
Available online 21 December 2011

JEL classification:
C11
C13
C21

Keywords:
Threshold regression
Structural change
Nonregular models
Boundary
Efficiency bounds
Bayes
Middle-point MLE
Compound Poisson process
Wiener–Hopf equation
Local asymptotic minimax
Credible set

ABSTRACT

This paper studies likelihood-based estimation and inference in parametric discontinuous threshold regression models with i.i.d. data. The setup allows heteroskedasticity and threshold effects in both mean and variance. By interpreting the threshold point as a "middle" boundary of the threshold variable, we find that the Bayes estimator is asymptotically efficient among all estimators in the locally asymptotically minimax sense. In particular, the Bayes estimator of the threshold point is asymptotically strictly more efficient than the left-endpoint maximum likelihood estimator and the newly proposed middle-point maximum likelihood estimator. Algorithms are developed to calculate asymptotic distributions and risk for the estimators of the threshold point. The posterior interval is proved to be an asymptotically valid confidence interval and is attractive in both length and coverage in finite samples.

1. Introduction

Since its invention by Tong Howell in the 1970s, the threshold regression model is popular in both statistics and econometrics.[1] Particularly, it has many applications in economics, e.g., Potter (1995), Durlauf and Johnson (1995), Savvides and Stengos (2000), Huang and Yang (2006) and Boetel et al. (2007) among others; see also Lee and Seo (2008) for other examples. The typical setup of threshold regression models is

$$y = \begin{cases} x'\beta_1 + \sigma_1 e, & q \le \gamma; \\ x'\beta_2 + \sigma_2 e, & q > \gamma; \end{cases} \tag{1}$$
$$E[e|x,q] = 0,$$

where q is the threshold variable used to split the sample, γ is the threshold point, $x \in \mathbb{R}^k$, $\beta \equiv (\beta_1', \beta_2')' \in \mathbb{R}^{2k}$ and $\sigma \equiv (\sigma_1, \sigma_2)'$ are threshold parameters on the mean and variance in the two regimes. We set $E[e^2] = 1$ as a normalization of the error variance and allow for conditional heteroskedasticity. All the other variables have the same definitions as in the linear regression framework.

There are two asymptotic frameworks for statistical inference on γ. The first is introduced by Chan (1993) in a nonlinear time series context, where $(\beta_1', \sigma_1)' - (\beta_2', \sigma_2)'$ is a fixed constant. The second is introduced by Hansen (2000), where no threshold effect on variance exists and the threshold effect in mean diminishes asymptotically. This paper uses the discontinuous framework of Chan (1993) with i.i.d. data. The results developed in this paper can serve as a benchmark for more complicated data generating processes in time series and panel data.

Both Chan (1993) and Hansen (2000) use least squares criteria to estimate γ, and derive the asymptotic distributions of the corresponding least squares estimators (LSEs), but the efficiency theory has never been studied. As Andrews (1993) concludes in the related structural change context, "no optimality properties are known for the ML estimator of π", where π is the structural change point and plays a similar role as γ in threshold regression. This paper intends to fill this gap in a parametric setting. In this environment, the density of e conditional on (x, q) is assumed to be $f_{e|x,q}(e|x, q; \alpha)$, where $\alpha \in \mathbb{R}^{d_\alpha}$ is some nuisance parameter affecting the shape of the error distribution. The joint distribution

E-mail addresses: p.yu@auckland.ac.nz, whistle.yu@gmail.com.

[1] See Howell (2007) for the birth of the threshold time series model.

0304-4076/$ – see front matter
doi:10.1016/j.jeconom.2011.12.002

of (x, q) is $f_{x,q}(x, q)$, the marginal distribution of q is $f_q(q)$, and the unknown parameter is $\theta = \left(\beta_1', \beta_2', \sigma_1, \sigma_2, \alpha', \gamma\right)' \equiv \left(\underline{\theta}', \gamma\right)'$.

In regular models, it is well known that the Bayes estimator (BE) and the maximum likelihood estimator (MLE) are asymptotically equivalent; see, e.g., Theorem 10.8 of Van der Vaart (1998). In nonregular models, however, Hirano and Porter (2003) and Chernozhukov and Hong (2004) show that the BE can be more efficient than the MLE in boundary estimation. By interpreting γ as a "middle" boundary of q, this paper finds a similar result about the estimation of γ. It is worth pointing out that the threshold regression model is more general than the models in the above-mentioned boundary literature. As illustrated in the following Section 2, the conventional boundary problems are special cases of (1) in an extremely simplified setup. Like the usual boundary literature, the results of this paper are developed using the framework in the seminal book by Ibragimov and Has'minskii (1981). Their Chapter 5 also discusses the statistical inference when densities have jumps, but their arguments seem more relevant to Hirano and Porter (2003) and Chernozhukov and Hong (2004).

The models considered in this paper are very general. For example, we allow heteroskedasticity and threshold effects in both mean and variance, error distributions with general parametric forms, and general loss functions. Independent work by Chan and Kutoyants (2010) considers a problem similar to this paper in threshold autoregressive models under very restrictive specifications; e.g., the error term is i.i.d. normal, the slope parameters are known, and only the mean square error loss is considered. They use a simulation method to find critical values for the confidence intervals (CIs) of the threshold point, which is, as argued in Section 3.4, not practical in reality. Instead, we suggest to use the posterior interval as the CI for the threshold point. Most importantly, they give little details on the efficiency problem.

This paper is organized as follows. Section 2 illustrates the main idea of this paper using a simple threshold regression model. Section 3 presents the main result of this paper, in which the asymptotic distributions of the MLE and BE are derived, and the BE is proved to be most efficient among all estimators. Also, the posterior interval is proven to be an asymptotically valid confidence interval. Section 4 shows some simulation results, and Section 5 concludes. All assumptions, proofs, lemmas and algorithms are given in Appendices A–D, respectively.

Before closing this introduction, it should be pointed out that the framework of this paper is essentially frequentist in the sense that while Bayes procedures are used, the randomness is confined to the data and does not include parameters. Correspondingly, we do not intend to propose a new Bayesian simulation method; such methods can be found in Geweke and Terui (1993). A word on notation: the letter C is used as a generic positive constant, which need not be the same in each occurrence. ℓ is always used for indicating the two regimes in (1), so is not written out explicitly as "$\ell = 1, 2$" throughout the paper. The code for simulations is available at http://homes.eco.auckland.ac.nz/pyu013/research.html.

2. No error term: an illustration

In this section, a simple threshold regression model is used to illustrate the main result of this paper: the threshold point is essentially a "middle" boundary. In the following discussion, $q_{(m)}$ denotes the mth order statistic of a sequence of random variables $\{q_i\}_{i=1}^n$.

Suppose the population model is

$$y = \mathbf{1}(q \le \gamma), \quad q \sim U[0, 1], \tag{2}$$

where $U[0, 1]$ is the uniform distribution on $[0, 1]$, $\mathbf{1}(\cdot)$ is the indicator function, γ is the parameter of interest, and $\gamma_0 = 1/2$.[2] This is equivalent to $x = 1$, $\beta_{10} = 1$, $\beta_{20} = 0$, and $\sigma_{10} = \sigma_{20} = 0$ in the general setup (1). There is no error term e in (2), so the observed y value can only be 0 or 1. Such a simple model can be viewed as a treatment rule in social program evaluation. If q is interpreted as the percentiles of income, then people below the median income are enrolled in the program with y taking value 1. Otherwise, people are not enrolled with y taking value 0. Such a treatment rule is too simple in reality, as the propensity score is a step function dropping from 1 (q below γ_0) to 0 (q above γ_0).[3] Here, the task is to find the step treatment rule, given the income of people and whether they are enrolled in the program.

For this simple model, the likelihood function is

$$p\left(W_n|\gamma\right) = \prod_{i=1}^{n} \left(\mathbf{1}\left(q_i \le \gamma\right)^{\mathbf{1}(y_i=1)} \cdot \mathbf{1}\left(q_i > \gamma\right)^{\mathbf{1}(y_i=0)}\right), \tag{3}$$

where $W_n = (w_1, \ldots, w_n)$ with $w_i \equiv (y_i, q_i)$ is the dataset, and 0^0 is defined to be 1. A simple calculation shows that the MLE is $[q_{(m)}, q_{(m+1)})$, where m is the number of y_i's with value 1. When there is an interval maximizing this likelihood function, following the literature, the left endpoint (i.e., $q_{(m)}$) is taken as the estimator.[4] Such an estimator is called the left-endpoint MLE (LMLE), and is denoted as $\widehat{\gamma}_{\text{LMLE}}$.

First, $\widehat{\gamma}_{\text{LMLE}}$ is n consistent. Notice that $\widehat{\gamma}_{\text{LMLE}}$ is the q_i that is closest to $1/2$ from the left. Since $\{q_i\}_{i=1}^n$ are sampled from $U[0, 1]$, n data points of q are randomly put into an interval with length 1, and thus the average distance between contiguous q_i's is around $1/n$. $1/2$ is in the interval $[q_{(m)}, q_{(m+1)})$, so $n(\widehat{\gamma}_{\text{LMLE}} - 1/2)$ is expected to be $O_p(1)$. Second,

$$n(\widehat{\gamma}_{\text{LMLE}} - 1/2) \xrightarrow{d} -\text{Exp}(1), \tag{4}$$

where Exp(1) is a standard exponential distribution. Since $\widehat{\gamma}_{\text{LMLE}}$ is smaller than $1/2$, for any $t \le 0$,

$$P\left(n(\widehat{\gamma}_{\text{LMLE}} - 1/2) \le t\right) = P\left(q_i \notin \left(1/2 + \frac{t}{n}, 1/2\right] \text{ for all } i\right) = \left(1 + \frac{t}{n}\right)^n \to e^t.$$

To further appreciate (4), suppose we want to estimate γ in the distribution of yq. yq picks out the q's such that $y = 1$. Its distribution is a point mass at 0 plus a uniform distribution on $(0, \gamma]$. Since γ is the right endpoint of this distribution, it is well known that the MLE is the maximum of the data and follows the exponential distribution asymptotically. Similarly, γ can be treated as the left boundary of $(1 - y)q$ and estimated by the minimum of the nonzero $(1 - y_i)q_i$. In short, γ can be viewed as a boundary (of both yq and $(1 - y)q$) although it is in the middle of q's support.

[2] I would like to thank Jack Porter for providing this example. I also want to thank an associate editor and a referee for improving its exposition.

[3] Such a treatment rule is called the sharp design, as opposed to the fuzzy design where the treatment is not deterministic in the two regimes, by Trochim (1984) in the regression discontinuity design (RDD) literature; see Bajari et al. (2010) for a similar analysis as below when using RDD to study contracting in health care. Usually, γ_0 is set by the policy-maker, and is publicly known.

[4] Most of the literature uses the left endpoint instead of the middle point. A possible reason is that these two estimators are thought to bear similar properties. For example, the sample splitting based on either point is the same; the maximizing interval shrinks at rate $1/n$ as shown in the following paragraph, so both methods generate almost the same point estimate in practice. The only exception to use the middle point, to my knowledge, is Gijbels et al. (1999) in the nonparametric environment, but they do not provide any theoretical justification.

For comparison, the asymptotic distribution of the BE is developed as follows. Suppose the prior $\pi(\gamma) = \mathbf{1}(0 \le \gamma \le 1)$, then the posterior density is

$$p(\gamma|W_n) = \frac{1}{q_{(m+1)} - q_{(m)}} \mathbf{1}\left(q_{(m)} \le \gamma < q_{(m+1)}\right).$$

Under the square error loss $l(v - \gamma) = (v - \gamma)^2$ or the absolute deviation error loss $|v - \gamma|$, the BE $\widehat{\gamma}_{BE}$ is uniquely defined and equal to $\frac{q_{(m)}+q_{(m+1)}}{2}$. It is still in the interval $[q_{(m)}, q_{(m+1)})$, but equals the middle-point MLE (MMLE) $\widehat{\gamma}_{MMLE}$. Note that $n(\widehat{\gamma}_{BE} - 1/2) = \frac{1}{2}\left[n\left(q_{(m)} - 1/2\right) + n\left(q_{(m+1)} - 1/2\right)\right]$. Since $n\left(q_{(m)} - 1/2\right)$ and $n\left(q_{(m+1)} - 1/2\right)$ are asymptotically independent with asymptotic distributions $-$Exp(1) and Exp(1), respectively, $n(\widehat{\gamma}_{BE} - 1/2)$ is asymptotically distributed as the difference between two independent exponential distributions, which is double exponential:

$$n(\widehat{\gamma}_{BE} - 1/2) \xrightarrow{d} \text{DExp}(0,1/2), \tag{5}$$

where DExp(0, 1/2) is a double exponential distribution with location 0 and scale 1/2. As a result,

$$\lim_{n\to\infty} E[(n(\widehat{\gamma}_{BE} - 1/2))^2] = 1/2 < 2 = \lim_{n\to\infty} E[(n(\widehat{\gamma}_{LMLE} - 1/2))^2],$$

and

$$\lim_{n\to\infty} E[|n(\widehat{\gamma}_{BE} - 1/2)|] = 1/2 < 1 = \lim_{n\to\infty} E[|n(\widehat{\gamma}_{LMLE} - 1/2)|],$$

meaning that the BE is more efficient than the LMLE.

To provide more insight about the general model which will be discussed in the next section, it is useful to derive the above asymptotic results through the conventional procedure: derive the limit likelihood ratio process first, and then apply the argmax continuous mapping theorem (see, e.g., Theorem 3.2.2 of Van der Vaart and Wellner (1996)) to find the asymptotic distribution. The localized likelihood ratio process is defined as

$$\begin{aligned}\frac{dP_v^n}{dP_0^n} &= \frac{\prod_{i=1}^n \left(\mathbf{1}\left(q_i \le \gamma_0 + \frac{v}{n}\right)^{\mathbf{1}(y_i=1)} \cdot \mathbf{1}\left(q_i > \gamma_0 + \frac{v}{n}\right)^{\mathbf{1}(y_i=0)}\right)}{\prod_{i=1}^n \left(\mathbf{1}(q_i \le \gamma_0)^{\mathbf{1}(y_i=1)} \cdot \mathbf{1}(q_i > \gamma_0)^{\mathbf{1}(y_i=0)}\right)} \\ &= \mathbf{1}\left(q_i \notin \left(\gamma_0, \gamma_0 + \frac{v}{n}\right] \text{ for all } i\right) \\ &\quad \times \mathbf{1}\left(q_i \notin \left(\gamma_0 + \frac{v}{n}, \gamma_0\right] \text{ for all } i\right),\end{aligned}$$

which converges weakly to

$$\exp\{D(v)\} \equiv \begin{cases} 1, & \text{if } -T_1 \le v < T_2; \\ 0, & \text{otherwise;} \end{cases}$$

where T_ℓ are independent standard exponential variables. The process $D(v)$ takes only two values: 0 and $-\infty$. By the argmax theorem, $n(\widehat{\gamma}_{LMLE} - 1/2)$ converges to the left endpoint of $\arg\max_{v\in\mathbb{R}} \exp\{D(v)\}$ which follows $-$Exp(1), and $n(\widehat{\gamma}_{BE} - 1/2)$ converges to the middle point of $\arg\max_{v\in\mathbb{R}} \exp\{D(v)\}$ which follows DExp(0, 1/2). The left two panels of Fig. 1 display a typical sample path of $\exp\{D(v)\}$ and the asymptotic distributions of $\widehat{\gamma}_{LMLE}$ and $\widehat{\gamma}_{BE}$.

To compare γ_0 with the conventional boundary, the MLE and BE of γ_0 in the distributions of yq and $(1-y)q$ are examined again. From Example 2 on p. 272 of Ibragimov and Has'minskii (1981) or Section 2.1 of Hirano and Porter (2003), the limit likelihood ratio process is $\exp\{D^-(v)\} \equiv \exp\{-v\}\mathbf{1}(v \ge -T_1)$ for yq and is $\exp\{D^+(v)\} \equiv \exp\{v\}\mathbf{1}(v \le T_2)$ for $(1-y)q$. Accordingly, the asymptotic distribution of the MLE is $-$Exp(1) for yq and Exp(1) for $(1-y)q$. Note that the MLE of γ_0 for both yq and $(1-y)q$ is unique, which contrasts threshold regression where the MLE is always an interval. Also, the posterior mean and posterior median are different. From Theorem 3.3 of Chernozhukov and Hong (2004), the mean of the asymptotic distribution of the posterior mean is zero, and the median of the asymptotic distribution of the posterior median is zero. From their Remark 3.7, both are location shifts of the MLE, so it is not surprising that the BE is more efficient than the MLE whose mean (median) is ∓ 1 ($\mp \ln 2$) for yq and $(1-y)q$, respectively. A typical sample path of $\exp\{D^-(v)\}$ ($\exp\{D^+(v)\}$) and the asymptotic densities of the MLE and BE of γ_0 are shown in the middle (right) two panels of Fig. 1. Since T_1 is sufficient in $\exp\{D^-(v)\}$, T_2 is sufficient in $\exp\{D^+(v)\}$, and $\{T_1, T_2\}$ are sufficient in $\exp\{D(v)\}$, the threshold point in this simple example can be treated as two boundaries which coincide.

Although this example is simple, it reveals the essence of the threshold point: it is a boundary. In boundary estimation, the BE is more efficient than the MLE. Also, only the local data around the boundary are informative. For example, only $q_{(m)}$ and $q_{(m+1)}$ are used in the estimation of γ in this simple example. This observation is generally true as shown in the next section.

3. Threshold regression: the general model

This section presents the general results for the parametric threshold regression model. It begins with the asymptotic properties of the likelihood ratio process, followed by the analysis of the MLEs and the BE, and concludes with the construction of confidence intervals.

3.1. Limit likelihood ratio process

As mentioned in Section 2, a common first step in deriving the asymptotic distribution of likelihood-based estimators is to find the finite-dimensional marginal limit of the localized likelihood ratio process. Such an initial step is also called the convergence of experiments in the literature such as Van der Vaart (1998).

The calculation in Section 2 shows that the normalization sequence for γ in the localized likelihood ratio process is $1/n$. It is also well known that the normalization sequence for regular parameters is $1/\sqrt{n}$. Define φ_n as the normalization matrix, which is a diagonal matrix with $1/\sqrt{n}$ in the first $2k + 2 + d_\alpha$ diagonal entries and $1/n$ in the remaining one diagonal entry.

Suppose the dataset $W_n = (w_1, \ldots, w_n)$ with $w_i \equiv (y_i, x_i', q_i)'$ is observed, then the localized likelihood ratio function with the true local parameter sequence $\theta_n \equiv \theta + \varphi_n h_0$ is

$$\begin{aligned}Z_{n,\theta_n}(h) &= \frac{L_n(\theta_n + \varphi_n h)}{L_n(\theta_n)} = \exp\left(\sum_{i=1}^n \ln \frac{f(w_i|\theta_n + \varphi_n h)}{f(w_i|\theta_n)}\right) \\ &= \exp\left(\sum_{i=1}^n \ln \frac{f_{y|x,q}(w_i|\theta_n + \varphi_n h)}{f_{y|x,q}(w_i|\theta_n)}\right),\end{aligned}$$

and $Z_n(h) \equiv Z_{n,\theta_0}(h)$. Here, $h = (u', v)$ with $u = \left(u_{\beta_1}', u_{\beta_2}', u_{\sigma_1}, u_{\sigma_2}, u_\alpha'\right)'$ being the local parameter for $\underline{\theta} = \left(\beta_1', \beta_2', \sigma_1, \sigma_2, \alpha'\right)'$, and v being the local parameter for γ, and

$$\begin{aligned}f_{y|x,q}(w|\theta) &= \frac{1}{\sigma_1} f_{e|x,q}\left(\frac{y - x'\beta_1}{\sigma_1}\middle| x, q; \alpha\right) \mathbf{1}(q \le \gamma) \\ &\quad + \frac{1}{\sigma_2} f_{e|x,q}\left(\frac{y - x'\beta_2}{\sigma_2}\middle| x, q; \alpha\right) \mathbf{1}(q > \gamma),\end{aligned}$$

$$f(w|\theta) = f_{y|x,q}(w|\theta) \cdot f_{x,q}(x, q),$$

$$L_n(\theta) = \prod_{i=1}^n f(w_i|\theta) \text{ is the likelihood function.}$$

The following assumption is imposed on the range of data and parameters. All other regularity conditions are collected in Appendix A.

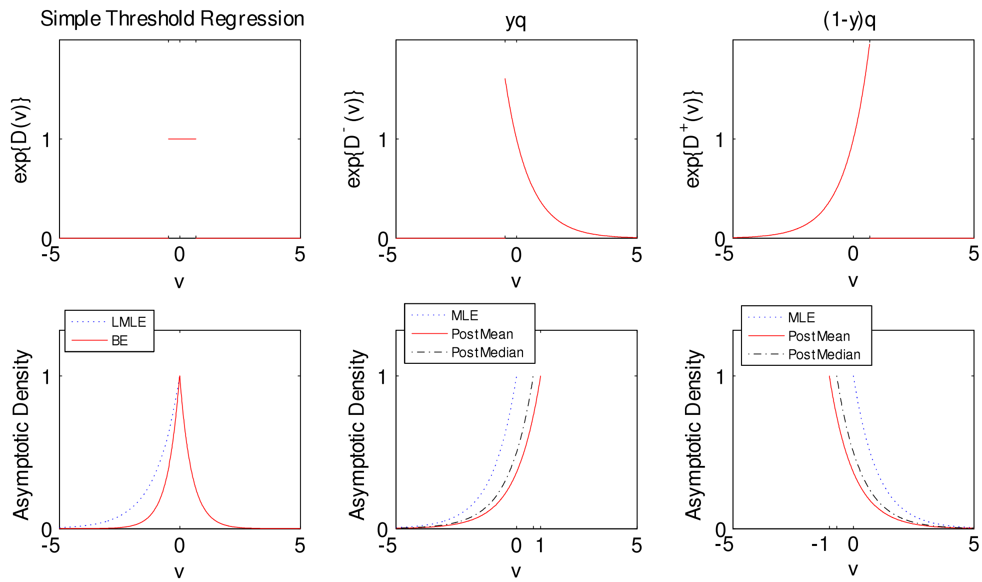

Fig. 1. $\exp\{D(v)\}$, $\exp\{D^{\mp}(v)\}$, and asymptotic densities of the MLE and BE of γ_0.

Assumption D0. $w_i \in \mathbb{W} \equiv \mathbb{R} \times \mathbb{X} \times \mathbb{Q} \subset \mathbb{R}^{k+2}$, $\beta_1 \in B_1 \subset \mathbb{R}^k$, $\beta_2 \in B_2 \subset \mathbb{R}^k$, $0 < \sigma_1 \in \Omega_1 \subset \mathbb{R}$, $0 < \sigma_2 \in \Omega_2 \subset \mathbb{R}$, $\alpha \in \Lambda \subset \mathbb{R}^{d_\alpha}$, $\gamma \in \Gamma = [\underline{\gamma}, \overline{\gamma}]$, $\Theta = B_1 \times B_2 \times \Omega_1 \times \Omega_2 \times \Lambda \times \Gamma \equiv \underline{\Theta} \times \Gamma$ is compact, $\theta_0 \in \Theta_0$, where Θ_0 is the interior of Θ, $\left(\beta_{10}', \sigma_{10}\right)' \neq \left(\beta_{20}', \sigma_{20}\right)'$, where $\neq$ means all corresponding coordinates of two vectors are different.

Define the local parameter spaces: $U_n = \sqrt{n}\left(\underline{\Theta} - \underline{\theta}_0\right)$, $V_n = n\left(\Gamma - \gamma_0\right)$, and $H_n = U_n \times V_n$. Under Assumption D0, H_n converges to $H = \mathbb{R}^{2k+3+d_\alpha}$. Further define

$$\bar{z}_1\left(\mathrm{w}|\widetilde{\theta}_2, \theta_1\right) = \ln \frac{\frac{\sigma_1}{\widetilde{\sigma}_2} f_{e|x,q}\left(\left.\frac{\sigma_1 e + x'(\beta_1 - \widetilde{\beta}_2)}{\widetilde{\sigma}_2}\right| x, q; \widetilde{\alpha}\right)}{f_{e|x,q}\left(e|x, q; \alpha\right)} = \ln \frac{f_{e|x,q}\left(e|x, q; \widetilde{\theta}_2, \underline{\theta}_1\right)}{f_{e|x,q}\left(e|x, q; \theta_1, \underline{\theta}_1\right)},$$

$$\bar{z}_2\left(\mathrm{w}|\widetilde{\theta}_1, \theta_2\right) = \ln \frac{\frac{\sigma_2}{\widetilde{\sigma}_1} f_{e|x,q}\left(\left.\frac{\sigma_2 e + x'(\beta_2 - \widetilde{\beta}_1)}{\widetilde{\sigma}_1}\right| x, q; \widetilde{\alpha}\right)}{f_{e|x,q}\left(e|x, q; \alpha\right)} = \ln \frac{f_{e|x,q}\left(e|x, q; \widetilde{\theta}_1, \underline{\theta}_2\right)}{f_{e|x,q}\left(e|x, q; \theta_2, \underline{\theta}_2\right)},$$

$$\bar{z}_{1i} = \bar{z}_1\left(\mathrm{w}_i|\theta_{20}, \theta_{10}\right) = \ln \frac{f_{e|x,q}\left(e|x, q; \theta_{20}, \underline{\theta}_{10}\right)}{f_{e|x,q}\left(e|x, q; \theta_{10}\right)},$$

$$\bar{z}_{2i} = \bar{z}_2\left(\mathrm{w}_i|\theta_{10}, \theta_{20}\right) = \ln \frac{f_{e|x,q}\left(e|x, q; \theta_{10}, \underline{\theta}_{20}\right)}{f_{e|x,q}\left(e|x, q; \theta_{20}\right)},$$

where

$$f_{e|x,q}\left(e|x, q; \widetilde{\theta}_{\ell'}, \underline{\theta}_\ell\right) = \frac{1}{\widetilde{\sigma}_{\ell'}} f_{e|x,q}\left(\left.\frac{\sigma_\ell e + x'\left(\beta_\ell - \widetilde{\beta}_{\ell'}\right)}{\widetilde{\sigma}_{\ell'}}\right| x, q; \widetilde{\alpha}\right),$$

$$f_{e|x,q}\left(e|x, q; \theta_\ell\right) = f_{e|x,q}\left(e|x, q; \theta_\ell, \underline{\theta}_{\ell 0}\right),$$

$$\theta_\ell = \left(\beta_\ell', \sigma_\ell, \alpha\right)', \quad \underline{\theta}_\ell = \left(\beta_\ell', \sigma_\ell\right)' \quad \text{and} \quad \mathrm{w} = \left(e, x', q\right)'.$$

θ_ℓ and $\widetilde{\theta}_\ell$ are two different θ_ℓ's while $\underline{\theta}_\ell$ is a subset of θ_ℓ. $f_{e|x,q}\left(e|x, q; \widetilde{\theta}_1, \underline{\theta}_1\right)$ is the conditional density of y in the left regime with the true parameter value being $\underline{\theta}_1$ but taken for $\widetilde{\theta}_1$, $f_{e|x,q}\left(e|x, q; \widetilde{\theta}_2, \underline{\theta}_1\right)$ is the conditional density of y when the threshold point is displaced on the left of the true value, and $f_{e|x,q}\left(e|x, q; \widetilde{\theta}_2, \underline{\theta}_2\right)$ and $f_{e|x,q}\left(e|x, q; \widetilde{\theta}_1, \underline{\theta}_2\right)$ are similarly defined. $\bar{z}_{1i}$ represents the effect on the log likelihood ratio when the threshold point is displaced on the left of γ_0, and $\bar{z}_{2i}$ represents the converse case. z_{1i} is defined as the limiting conditional distribution of $\bar{z}_{1i}$ given $\gamma_0 + \Delta < q_i \le \gamma_0$, $\Delta < 0$ with $\Delta \uparrow 0$, and z_{2i} is defined as the limiting conditional distribution of $\bar{z}_{2i}$ given $\gamma_0 < q_i \le \gamma_0 + \Delta$, $\Delta > 0$ with $\Delta \downarrow 0$. There is no mystery in the definition of $z_{\ell i}$. If $\bar{z}_{\ell i}$ and q_i have a joint density $f_{\bar{z}_\ell, q}(\bar{z}_\ell, q)$ which is continuous, then the density of $z_{\ell i}$ is $f_{\bar{z}_\ell, q}(z_\ell, \gamma_0)/f_q(\gamma_0)$; that is, the conditional density of $\bar{z}_\ell$ given $q = \gamma_0$.

Theorem 1. *Under Assumptions D0–D7, for every finite $I \subset H$,*

$$\left(Z_n(h)\right)_{h\in I} \xrightarrow{d} \left(\exp\left\{-\frac{1}{2}u'\mathcal{J}u + u'\mathcal{J}W + D\left(v\right)\right\}\right)_{h\in I} \equiv Z_\infty(h),$$

where

$$W \sim N\left(0, \mathcal{J}^{-1}\right);$$

$$D(v) = \begin{cases} \sum_{i=1}^{N_1(|v|)} z_{1i}, & \text{if } v \le 0; \\ \sum_{i=1}^{N_2(v)} z_{2i}, & \text{if } v > 0; \end{cases}$$

$\mathcal{J}$ is the information matrix of regular parameters and defined in (8) of Appendix A, $N_\ell(\cdot)$ is a Poisson process with intensity $f_q(\gamma_0)$, and all z_{1i}, z_{2i}, $i = 1, 2, \ldots, N_1(\cdot)$ and $N_2(\cdot)$ are mutually independent of each other. Furthermore, W and $\{D(v)\}_{h\in I}$ are independent of each other, and $D(v)$ is cadlag with $D(0) = 0$ almost surely.

The limit likelihood ratio process takes a separable form of regular local parameters and the nonregular local parameter. In Bayesian language, the posterior distributions of regular parameters and the nonregular parameter are asymptotically independent. From the frequentist perspective, W, which represents the randomnesses about regular parameters, and $D(v)$, which represents the randomness about the nonregular parameter, are independent. Such a double independence also appears in the boundary literature such as Chernozhukov and Hong (2004).

The most interesting component in this limit process is $D(v)$ which is a compound Poisson process. By the strict Jensen's inequality, $E[z_{1i}] < 0$ and $E[z_{2i}] < 0$, so $D(v)$ will diverge to $-\infty$ when $|v|$ goes to infinity. The jumps z_{1i} and z_{2i} depend on two other factors besides the conditional density $f_{e|x,q}$: the ratio of mean difference to variance $\frac{\beta_1-\beta_2}{\sigma_1}\left(\frac{\beta_1-\beta_2}{\sigma_2}\right)$, and variance ratio $\frac{\sigma_1}{\sigma_2}$. The effects of these two factors will be illustrated using a concrete example in Section 4. Similar to the simple example in Section 2, $D(v)$ only depends on the local data around γ_0. For example, z_{1i} (z_{2i}) follows a distribution conditional on q in the neighborhood of γ_0, and the intensity of $N_1(\cdot)$ $(N_2(\cdot))$ is $f_q(\gamma_0)$. More details on such a local information dependence can be found in Yu (2008a). Note that $D(v)$ cannot be expressed as the Poisson integral in Theorem 3.1 of Chernozhukov and Hong (2004) because the jumps $z_{\ell i}$ include extra randomness than the underlying Poisson random measure. This is in turn due to the fact that the observed dependent variable y can contain information different from the threshold crossing variable q, which makes the threshold model more general.

The limit likelihood ratio process in Theorem 1 hinges on the correct specification of $f_{e|x,q}(e|x,q;\alpha)$, while the asymptotic theory for the misspecified model is not necessary. The reason is that we can always recover the true distribution of the data as long as a semiparametric restriction is satisfied as shown by Yu (2008a). Of course, this is mainly due to the special structure of threshold regression; in regular parametric models with a general nonlinear structure in parameters as in White (1982), this is not the case. As to the local misspecification, it seems that Remark 3.3 of Chernozhukov and Hong (2004) can still be applied here.

The above analysis can also apply to more general models such as

$$y = \begin{cases} g_1(x,\beta_1,\delta) + e_1, & q \le \gamma; \\ g_2(x,\beta_2,\delta) + e_2, & q > \gamma; \end{cases}$$

where δ is the parameters that remain the same in the two regimes, and g_1 and g_2 are smooth functions which are not necessarily the same. e_1 and e_2 need not take the form $\sigma_1 e$ and $\sigma_2 e$; for example, the conditional distribution of e_ℓ can be $f_{e|x,q}(e|x,q;\alpha_\ell)$, and there can be some overlap between α_1 and α_2. For such models,

$$\bar{z}_{1i} = \ln \frac{f_{e_2|x,q}\left(e_{1i} + g_1(x_i,\beta_{10},\delta_0) - g_2(x_i,\beta_{20},\delta_0)\,|x_i,q_i;\alpha_{20}\right)}{f_{e_1|x,q}\left(e_{1i}|x_i,q_i;\alpha_{10}\right)}$$

and

$$\bar{z}_{2i} = \ln \frac{f_{e_1|x,q}\left(e_{2i} + g_2(x_i,\beta_{20},\delta_0) - g_1(x_i,\beta_{10},\delta_0)\,|x_i,q_i;\alpha_{10}\right)}{f_{e_2|x,q}\left(e_{2i}|x_i,q_i;\alpha_{20}\right)},$$

in $D(v)$, and all other components of $D(v)$ are the same as in Theorem 1.

3.2. Maximum likelihood estimation

The MLE is defined as the maximizer of the log likelihood function:

$$\widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta} Q_n(\theta) \equiv \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ln f_{y|x,q}(w_i|\theta).$$

As in Section 2, two MLEs of γ are defined: $\widehat{\gamma}_{\mathrm{LMLE}}$ and $\widehat{\gamma}_{\mathrm{MMLE}}$, which are the left endpoint and middle point of the maximizing interval, respectively. Following the literature (e.g., Hansen, 2000), concentration is used to find $\widehat{\gamma}_{\mathrm{LMLE}}$ and $\widehat{\gamma}_{\mathrm{MMLE}}$, but we need a two-step concentration under our general setup. First, conditional on γ and α, $\widehat{\underline{\theta}}_1(\alpha,\gamma) \equiv \left(\widehat{\beta}_1(\alpha,\gamma), \widehat{\sigma}_1(\alpha,\gamma)\right)$ is the MLE using the data with $q_i \le \gamma$, and $\widehat{\underline{\theta}}_2(\alpha,\gamma) \equiv \left(\widehat{\beta}_2(\alpha,\gamma), \widehat{\sigma}_2(\alpha,\gamma)\right)$ is the MLE using the data with $q_i > \gamma$. Denote the concentrated objective function as $Q_n(\alpha,\gamma)$. Second, concentrate $Q_n(\alpha,\gamma)$ further on γ. For any γ on the interval $I_i = [q_{(i-1)}, q_{(i)})$, $i = 2,\ldots,n$, $Q_n(\alpha,\gamma)$ as a function of α is the same, so $\widehat{\alpha}(\gamma)$ is essentially a function of I_i and can only take $n-1$ values. Denote the corresponding concentrated objective function as $Q_n(\gamma)$. At last, the MLEs of γ are uniquely defined as

$$\widehat{\gamma}_{\mathrm{LMLE}} = \arg\max_{\gamma\in\Gamma} L\, Q_n(\gamma), \qquad \widehat{\gamma}_{\mathrm{MMLE}} = \arg\max_{\gamma\in\Gamma} M\, Q_n(\gamma).$$

Here, $\arg\max L$ means that the left endpoint of the maximizing interval is taken as the estimator of γ; that is, the maximizer of γ is searched only among $\Gamma \cap \{q_{(1)},\ldots,q_{(n)}\}$. $\arg\max M$ means that the middle point of the maximizing interval is taken as the estimator of γ; that is, the maximizer of γ is searched only among $\Gamma \cap \left\{\frac{q_{(1)}+q_{(2)}}{2},\ldots,\frac{q_{(n-1)}+q_{(n)}}{2}\right\}$. The above analysis suggests that the MLE of $\underline{\theta}$ is numerically invariant regardless of which point on I_i is taken as our MLE of γ. So we define

$$\widehat{\underline{\theta}}_{\mathrm{MLE}} = \Big(\widehat{\beta}_1'\left(\widehat{\alpha}(\widehat{\gamma}),\widehat{\gamma}\right), \widehat{\beta}_2'\left(\widehat{\alpha}(\widehat{\gamma}),\widehat{\gamma}\right), \widehat{\sigma}_1\left(\widehat{\alpha}(\widehat{\gamma}),\widehat{\gamma}\right), \widehat{\sigma}_2\left(\widehat{\alpha}(\widehat{\gamma}),\widehat{\gamma}\right), \widehat{\alpha}(\widehat{\gamma})\Big),$$

where $\widehat{\gamma}$ can be either $\widehat{\gamma}_{\mathrm{LMLE}}$ or $\widehat{\gamma}_{\mathrm{MMLE}}$. Now, the localized MLEs

$$\frac{1}{\varphi_n}\left(\widehat{\theta}_{\mathrm{LMLE}} - \theta_0\right) = \arg\max_{h\in H_n} L\, Q_n(h),$$
$$\frac{1}{\varphi_n}\left(\widehat{\theta}_{\mathrm{MMLE}} - \theta_0\right) = \arg\max_{h\in H_n} M\, Q_n(h),$$

where $Q_n(h) = \ln Z_n(h)$, $\widehat{\theta}_{\mathrm{LMLE}} = \left(\widehat{\underline{\theta}}_{\mathrm{MLE}}, \widehat{\gamma}_{\mathrm{LMLE}}\right)$ and $\widehat{\theta}_{\mathrm{MMLE}} = \left(\widehat{\underline{\theta}}_{\mathrm{MLE}}, \widehat{\gamma}_{\mathrm{MMLE}}\right)$. By the argmax continuous mapping theorem and the limit likelihood ratio process in Section 3.1, the following theorem follows.

Theorem 2 (*Asymptotic Distributions of the LMLE and MMLE*). *Under Assumptions D0–D10,*

$$\widehat{\theta}_{\mathrm{LMLE}} - \theta_0 = O_p(\varphi_n), \qquad \widehat{\theta}_{\mathrm{MMLE}} - \theta_0 = O_p(\varphi_n),$$

and

$$\sqrt{n}\left(\widehat{\underline{\theta}}_{\mathrm{MLE}} - \underline{\theta}_0\right) \xrightarrow{d} W \equiv Z_{\underline{\theta},\mathrm{MLE}}$$
$$n\left(\widehat{\gamma}_{\mathrm{LMLE}} - \gamma_0\right) \xrightarrow{d} M_- \equiv Z_{\gamma,\mathrm{LMLE}}$$
$$n\left(\widehat{\gamma}_{\mathrm{MMLE}} - \gamma_0\right) \xrightarrow{d} \frac{M_- + M_+}{2} \equiv Z_{\gamma,\mathrm{MMLE}}$$

where W is defined in Theorem 1, and $[M_-, M_+) = \arg\max_{v\in\mathbb{R}} D(v)$, and $Z_{\underline{\theta},\mathrm{MLE}}$ is independent of $Z_{\gamma,\mathrm{LMLE}}$ and $Z_{\gamma,\mathrm{MMLE}}$. Furthermore, the asymptotic distribution of $\widehat{\underline{\theta}}_{\mathrm{MLE}}$ is the same as that in the case when γ_0 is known, and the asymptotic distribution of $\widehat{\gamma}_{\mathrm{LMLE}}$ $(\widehat{\gamma}_{\mathrm{MMLE}})$ is the same as that in the case when $\underline{\theta}_0$ is known.

$Z_{\gamma,\mathrm{LMLE}}$ and $Z_{\gamma,\mathrm{MMLE}}$ are well-defined random variables that cannot take $\pm\infty$. Since there are no ties on the sample path of $D(v)$, $[M_-, M_+)$ is uniquely identified. Also, $[M_-, M_+)$ cannot go to infinity as $D(\pm\infty) = -\infty$ and $D(0) = 0$. An interesting result is that the asymptotic distributions of regular parameters and the nonregular parameter do not affect each other. Such an informational independence contrasts the dependence among the regular parameters; i.e., $\mathcal{J}^{-1}$ may not be diagonal.

Theorem 2 actually covers the asymptotic distributions of many popular estimators. In structural change models, q essentially follows the uniform distribution on $[0,1]$ which is independent of (x,e), although its support is only a set of discrete points $\left\{0,\frac{1}{n},\ldots,\frac{n-1}{n},1\right\}$. As a result, $z_{\ell i} = \bar{z}_{\ell i}$, and the jumping locations of $N_\ell(\cdot)$ are fixed at $\{1,2,\ldots\}$. Now, $D(v)$ becomes a random walk instead of a compound Poisson process. Define $T_{\ell i}$'s as the interarrival times of $N_\ell(\cdot)$, and they follow i.i.d. exponential distributions with mean $1/f_q(\gamma_0)$. When q follows $U[0,1]$, the mean interarrival time $E[T_{\ell i}] = 1$, which is the same as the fixed interarrival time of the random walk. If we further assume $\sigma_1 = \sigma_2 = s$ and e is independent of x, then $D(v)$ can be further simplified. For example, when e is normally distributed, $\widehat{\gamma}_{\text{LMLE}}$ corresponds to the LSE in Theorem 2 of Chan (1993) and Proposition 2 of Bai (1997). If x includes only a constant 1, then z_{1i} and z_{2i} follow the same distribution $N\left(-2\Delta^2, 4\Delta^2\right)$ with $\Delta = \frac{|\beta_1-\beta_2|}{2s}$ defined in Section 3 of Hinkley (1970). If e follows the Laplace distribution, then $\widehat{\gamma}_{\text{LMLE}}$ corresponds to the LAD estimator in Theorem 3 of Bai (1995).

The algorithms to calculate the distribution and risk of $Z_{\gamma,\text{LMLE}}$ and $Z_{\gamma,\text{MMLE}}$ are given in Appendix D. When z_{1i} and z_{2i} follow the same distribution, the distribution of $Z_{\gamma,\text{MMLE}}$ is symmetric, but this is not true for $Z_{\gamma,\text{LMLE}}$.[5] In this case, $\widehat{\gamma}_{\text{MMLE}}$ is more efficient than $\widehat{\gamma}_{\text{LMLE}}$ under polynomial losses. In general, however, there is no ordering between the risks of $\widehat{\gamma}_{\text{LMLE}}$ and $\widehat{\gamma}_{\text{MMLE}}$ even under the square error loss or the absolute deviation error loss.

3.3. Bayes estimation

Define the posterior and the localized posterior as

$$p_n(\theta) = \frac{L_n(\theta)\pi(\theta)}{\int_\Theta L_n(\widetilde{\theta})\pi(\widetilde{\theta})d\widetilde{\theta}},$$
$$p_n^*(h) = |\varphi_n|\, p_n(\theta_0+\varphi_n h) = \frac{Z_n(u,v)\,\pi\left(\underline{\theta}_0+\frac{u}{\sqrt{n}}, \gamma_0+\frac{v}{n}\right)}{\int_{U_n\times V_n} Z_n(\widetilde{u},\widetilde{v})\,\pi\left(\underline{\theta}_0+\frac{\widetilde{u}}{\sqrt{n}}, \gamma_0+\frac{\widetilde{v}}{n}\right)d\widetilde{u}d\widetilde{v}},$$

where $|\varphi_n|$ is the determinant of the matrix φ_n. Accordingly, the localized BE

$$\frac{1}{\varphi_n}\left(\widehat{\theta}_{\text{BE}}-\theta_0\right) = \arg\min_{s,t}\psi_n(s,t),$$

where

$$\psi_n(s,t) \equiv \int_{U_n\times V_n} l(s-u, t-v)\, p_n^*(h)dudv,$$

and $l(\cdot,\cdot)$ is the loss function. By Theorem 1, $p_n^*(h)$ will converge weakly to $p^*(h) = p_1^*(u)\cdot p_2^*(v)$ since the prior will be dominated by the data, where

$$p_1^*(u) = \frac{\exp\left\{-\frac{1}{2}u'\mathcal{J}u + u'\mathcal{J}W\right\}}{\int_{\mathbb{R}^{2k+2+d_\alpha}}\exp\left\{-\frac{1}{2}\widetilde{u}'\mathcal{J}\widetilde{u}+\widetilde{u}'\mathcal{J}W\right\}d\widetilde{u}} = \sqrt{\frac{|\mathcal{J}|}{(2\pi)^{2k+2+d_\alpha}}}\exp\left(-\frac{1}{2}\left(u-Z_{\underline{\theta},\text{MLE}}\right)'\mathcal{J}\left(u-Z_{\underline{\theta},\text{MLE}}\right)\right)$$
$$p_2^*(v) = \frac{\exp\{D(v)\}}{\int_{\mathbb{R}}\exp\{D(\widetilde{v})\}d\widetilde{v}}$$

are normalized asymptotic posteriors. By the argmax continuous mapping theorem, the asymptotic distribution of the BE can be obtained, which is given in the next theorem. Furthermore, the BE is proved to be asymptotically efficient in the locally asymptotically minimax (LAM) sense.

[5] When e, x and q are independent of each other, and there is no threshold effect in variance, z_{1i} and z_{2i} having the same distribution implies $P(z_{1i}\le C) = P(z_{2i}\le C)$; that is, $P\left(\frac{f_e\left(e_i+x_i'\frac{\beta_1-\beta_2}{s}|\alpha\right)}{f_e(e_i|\alpha)}\le C\right) = P\left(\frac{f_e\left(e_i-x_i'\frac{\beta_1-\beta_2}{s}|\alpha\right)}{f_e(e_i|\alpha)}\le C\right)$ for any $C\ge 0$. If the distribution of e or any one to one transformation of e is symmetric, this condition is satisfied. For example, e can follow a normal, log normal, Laplace, logistic or t distribution. See Bai (1995) and Hinkley (1970) for more discussions.

Theorem 3. *Under Assumptions D0–D12, P and L,*

(i) $\widehat{\theta}_{\text{BE}}-\theta_0 = O_p(\varphi_n)$.

(ii) *If*

$$\psi(s,t) \equiv \int_{\mathbb{R}^{2k+3+d_\alpha}} l(s-u,t-v)\cdot p_1^*(u)p_2^*(v)dudv$$

reaches its minimum at a unique point $Z_{\theta,\text{BE}} = \left(Z_{\underline{\theta},\text{BE}}, Z_{\gamma,\text{BE}}\right)$, *then*

$$\varphi_n^{-1}\left(\widehat{\theta}_{\text{BE}}-\theta_0\right)\xrightarrow{d} Z_{\theta,\text{BE}}.$$

(iii) *Assume* $l(h) = l_1(u)+l_2(v)$; *that is, the loss function is separable in regular parameters and the nonregular parameter. Further assume*

$$\psi_{\underline{\theta}}(s) \equiv \int_{\mathbb{R}^{2k+2+d_\alpha}} l_1(s-u)\,p_1^*(u)du,$$

and

$$\psi_\gamma(t) \equiv \int_{\mathbb{R}} l_2(t-v)\,p_2^*(v)dv,$$

reach their minima at a unique point $Z_{\underline{\theta},\text{BE}}$ *and* $Z_{\gamma,\text{BE}}$, *respectively, then*

$$\sqrt{n}\left(\widehat{\underline{\theta}}_{\text{BE}}-\underline{\theta}_0\right)\xrightarrow{d} Z_{\underline{\theta},\text{BE}},$$
$$n\left(\widehat{\gamma}_{\text{BE}}-\gamma_0\right)\xrightarrow{d} Z_{\gamma,\text{BE}},$$

and $Z_{\underline{\theta},\text{BE}}$ *and* $Z_{\gamma,\text{BE}}$ *are independent.*

(iv) $\widehat{\theta}_{\text{BE}}$ *is asymptotically efficient at* θ_0 *with respect to the loss function* l *in the LAM sense:*

$$\lim_{\delta\to 0}\lim_{n\to\infty}\left[\inf_{T_n}\sup_{|\theta-\theta_0|<\delta}E_\theta\left[l\left(\frac{1}{\varphi_n}(T_n-\theta)\right)\right] - \sup_{|\theta-\theta_0|<\delta}E_\theta\left[l\left(\frac{1}{\varphi_n}\left(\widehat{\theta}_{\text{BE}}-\theta\right)\right)\right]\right] = 0, \quad (6)$$

where $E_\theta[\cdot]$ *is the expectation under* θ, *and* T_n *can be any estimator of* θ. *If* $l(h) = l_1(u)+l_2(v)$, *then* $\widehat{\underline{\theta}}_{\text{BE}}$ *and* $\widehat{\gamma}_{\text{BE}}$ *are asymptotically efficient at* $\underline{\theta}_0$ *and* γ_0 *with respect to the loss function* l_1 *and* l_2 *in the LAM sense, respectively:*

$$\lim_{\delta\to 0}\lim_{n\to\infty}\left[\inf_{T_{n,\underline{\theta}}}\sup_{|\gamma-\gamma_0|<\delta}\sup_{|\underline{\theta}-\underline{\theta}_0|<\delta}E_\theta\left[l_1\left(\sqrt{n}\left(T_{n,\underline{\theta}}-\underline{\theta}\right)\right)\right] - \sup_{|\gamma-\gamma_0|<\delta}\sup_{|\underline{\theta}-\underline{\theta}_0|<\delta}E_\theta\left[l_1\left(\sqrt{n}\left(\widehat{\underline{\theta}}_{\text{BE}}-\underline{\theta}\right)\right)\right]\right] = 0,$$

$$\lim_{\delta\to 0}\lim_{n\to\infty}\left[\inf_{T_{n,\gamma}}\sup_{|\underline{\theta}-\underline{\theta}_0|<\delta}\sup_{|\gamma-\gamma_0|<\delta}E_\theta\left[l_2\left(n\left(T_{n,\gamma}-\gamma\right)\right)\right] - \sup_{|\underline{\theta}-\underline{\theta}_0|<\delta}\sup_{|\gamma-\gamma_0|<\delta}E_\theta\left[l_2\left(n\left(\widehat{\gamma}_{\text{BE}}-\gamma\right)\right)\right]\right] = 0,$$

where $T_{n,\underline{\theta}}$ *can be any estimator of* $\underline{\theta}$, *and* $T_{n,\gamma}$ *can be any estimator of* γ.

In Theorem 3, $Z_{\underline{\theta},\text{BE}}$ and $Z_{\gamma,\text{BE}}$ can be correlated when the loss function is not separable. When the loss function is separable and $l_1(\cdot)$ is bowl-shaped, $Z_{\underline{\theta},\text{BE}} = Z_{\underline{\theta},\text{MLE}} \sim N\left(0,\mathcal{J}^{-1}\right)$. Therefore, the MLE and the BE are asymptotically equivalent for regular parameters. But for different loss functions $l_2(\cdot)$, unlike the regular parameter case, the most efficient estimator of γ_0 is different even if these loss functions are bowl-shaped. For a given $l_2(\cdot)$, the BE of γ is strictly more efficient than the MLEs under the same loss function except in some extreme cases, since the MLEs can be viewed as a BE under any loss function that approximates the delta function, e.g., 0–1 loss $\mathbf{1}(|v| > \epsilon)/\epsilon$.

To further appreciate why the BE of γ is more efficient than the MLEs, note that the regular component of $Z_\infty(h)$ is determined by W which has the same dimension as $\underline{\theta}$, but the nonregular component $D(v)$ is determined by $\left\{\{T_{1i}, T_{2i}\}_{i=0}^{\infty}, \{z_{1i}, z_{2i}\}_{i=1}^{\infty}\right\}$ which is infinite-dimensional. From the well-known Rao–Blackwell argument, the limit optimal (Bayes) estimators of $\underline{\theta}$ are all some shift transformations of W since the information about $\underline{\theta}$ in the data is completely controlled by W asymptotically. When l_1 is bowl-shaped, they are just W. For γ, however, a one-dimensional estimator cannot cover the information in infinite-dimensional statistics. How to use the information efficiently and which part of the information to use depend on the loss function. For example, suppose $Z_{\gamma,\text{LMLE}}$ is obtained at the Mth jump on $v \le 0$, then $Z_{\gamma,\text{LMLE}} = -\sum_{i=0}^{M} T_{1i}$, indicating that $Z_{\gamma,\text{LMLE}}$ only uses the information in $\{T_{1i}\}_{i=0}^{M}$ and $\{z_{1i}, z_{2i}\}_{i=1}^{\infty}$. In contrast, the posterior mean is a function of all $\left\{\{T_{1i}, T_{2i}\}_{i=0}^{\infty}, \{z_{1i}, z_{2i}\}_{i=1}^{\infty}\right\}$, which uses the information more efficiently than $Z_{\gamma,\text{LMLE}}$ under the square error loss.

The most popular specifications of loss function $l_2(\cdot)$ include: (a) $l_2(v) = v^2$, then $Z_{\gamma,\text{BE}}$ is the mean of $p_2^*(v)$; (b) $l_2(v) = |v|$, then $Z_{\gamma,\text{BE}}$ is the median of $p_2^*(v)$; and (c) $l_2(-v) = (\tau - \mathbf{1}(v \le 0))\, v$ for $\tau \in (0, 1)$, then $Z_{\gamma,\text{BE}}$ is the τ'th percentile of $p_2^*(v)$, denoted as $Z_{\tau,\text{BE}}$. The algorithms to calculate the distribution and risk of $Z_{\gamma,\text{BE}}$ for the above loss functions are given in Appendix D. In applications, the MCMC method is used to compute the BE. For example, if the goal is to find the BE of γ, first draw a Markov chain $S = \left(\theta^{(1)}, \ldots, \theta^{(B)}\right)$, whose marginal density is approximately $p_n(\theta)$. Then choose the sequence for γ, denoted as S_γ:

$$S_\gamma = \left(\gamma^{(1)}, \ldots, \gamma^{(B)}\right). \tag{7}$$

$\widehat{\gamma}_{\text{BE}}$ is some function of S_γ depending on the loss function l_2. In cases of (a)–(c), $\widehat{\gamma}_{\text{BE}}$ is approximated by the mean, median, and the τ'th percentile of S_γ, respectively.

3.4. Confidence interval construction of γ

Although Appendix D provides the algorithms for simulating the asymptotic distributions of the estimators, Wald-type CIs are hard to construct for two main reasons. First, while the conditional distribution $f_{e|x,q}(e|x, q; \alpha)$ is usually specified in a parametric model, $f_{x,q}(x, q)$ is rarely known in practice, and as a result, the intensity of $\underline{N}_\ell(\cdot)$ must be estimated by some nonparametric method, e.g., $\widehat{f}_q(\widehat{\gamma})$, where $\widehat{f}_q$ is a nonparametric density estimator such as the kernel smoother, and $\widehat{\gamma}$ can be any consistent estimator. But such a nonparametric estimator will make the algorithms imprecise due to the low convergence rate of nonparametric methods. Second, infinite independent copies of z_{1i} and z_{2i} are also needed to simulate sample paths of $D(v)$, but they involve an unknown generator of conditional random variables unless q is assumed to be independent of (x, e). See Section 4.1 of Yu (2008a) for more discussions about the difficulties in constructing Wald-type CIs in threshold regression.

Three valid methods for CI construction in the parametric case are as follows. (i) Hansen's method (2000), which inverts the acceptance region of the likelihood ratio test. Hansen (2000) derives an elegant asymptotic distribution in the framework of asymptotically vanishing threshold effect, but his framework does not allow the threshold effect in variance. His Theorem 3 shows that in our framework with i.i.d. normal errors, his CI is valid but conservative. (ii) The parametric bootstrap method. The validity of this method for the LMLE can be shown using Proposition 1.1 in Beran (1997). The critical step is to check the condition (a) there, which requires that $n(\widehat{\gamma}_{\text{LMLE}} - \gamma_n) \xrightarrow{d} Z_{\gamma,\text{LMLE}}$ for any sequence of θ_n converging to θ_0. In the proof of Theorem 1, we show that the weak convergence of the log likelihood ratio process is uniform for θ in a neighborhood of θ_0, so this condition is automatically satisfied. Similar arguments can apply to the MMLE and BE.[6] Since $f_{x,q}(x, q)$ is unknown in practice, the parametric wild bootstrap would be suggested. In the parametric wild bootstrap, we condition on $\{x_i, q_i\}_{i=1}^{n}$, and only utilize the randomness from $f_{e|x,q}(e|x, q; \widehat{\alpha})$. But as argued in Section 4.3 of Yu (2008b), the parametric wild bootstrap is not valid. (iii) The subsampling method, which is proposed by Politis and Romano (1994) and summarized in Politis et al. (1999). The only assumption for the validity of this method is that there is a continuous asymptotic distribution, which is proved in Appendix D.[7] A similar question as in the parametric bootstrap is that whether the parametric wild subsampling is valid. Given the validity of both the parametric subsampling and the nonparametric subsampling, the parametric wild subsampling, which can be treated as an in-between procedure, should be valid.[8] But a formal development of this result is beyond the scope of this paper.

The Bayes method can lead to a straightforward construction of the CI for γ. The Bayesian credible set (corresponding to confidence interval in the frequentist language) for γ can be constructed as follows. Define

$$F_n(x) = \int_{\theta\in\Theta:\gamma\le x} p_n(\theta)d\theta \quad \text{and} \quad c_n(\tau) = \inf\{x : F_n(x) \ge \tau\},$$

then the $(1-\tau)$ credible set for γ is given by $[c_n(\tau/2), c_n(1-\tau/2)]$. Note that the Bayesian credible set does not rely on the specific form of $f_{x,q}(x, q)$. $f_{x,q}(x, q)$ is absorbed in the constant term of the posterior since it does not include any parameter of interest. In applications, choose the $\tau/2$ and $1-\tau/2$ percentile of the marginal MCMC sequence S_γ in (7). For regular parameters, similar steps can be followed. The following theorem makes sure that this credible set is an asymptotically valid confidence set for γ.

Theorem 4. *Under Assumptions D0–D12 and P, if $Z_{\tau/2,\text{BE}}$ and $Z_{1-\tau/2,\text{BE}}$, the $\tau/2$'th and $(1-\tau/2)$'th percentile of $p_2^*(v)$, have positive densities over an open neighborhood of* 0*, then*

$$\lim_{n\to\infty} P\left(c_n(\tau/2) \le \gamma_0 \le c_n(1-\tau/2)\right) = 1 - \tau.$$

Proof. This result is from the optimality of the Bayes estimator; see the proof of Theorem 3.3 in Chernozhukov and Hong (2004) for more details. □

We can use the algorithms in Appendix D to check the condition that $Z_{\tau/2,\text{BE}}$ and $Z_{1-\tau/2,\text{BE}}$ have positive densities over an open neighborhood of 0. In fact, for a confidence interval, asymptotic validity is only the basic requirement. Given the efficiency of the Bayes procedure, we also expect that the credible set has advantages in length, and this is confirmed by simulations in Section 4.2.

6 In the conventional boundary case, Remark 3.6 of Chernozhukov and Hong (2004) provides a similar result.

7 In proving the validity of the nonparametric subsampling, Gonzalo and Wolf (2005) assume the asymptotic distribution of the test statistic to be continuous without proof.

8 To see why the parametric wild subsampling is valid, let us go back to the simple example in Section 2. Now, the parametric wild subsampling is equivalent to the nonparametric subsampling because the only randomness is from q. Given that the asymptotic distribution, which is exponential or double exponential, is continuous, the parametric wild subsampling is valid. Note that given $f_{x,q}$ is unknown in reality, the only practical parametric resampling method is the parametric wild subsampling.

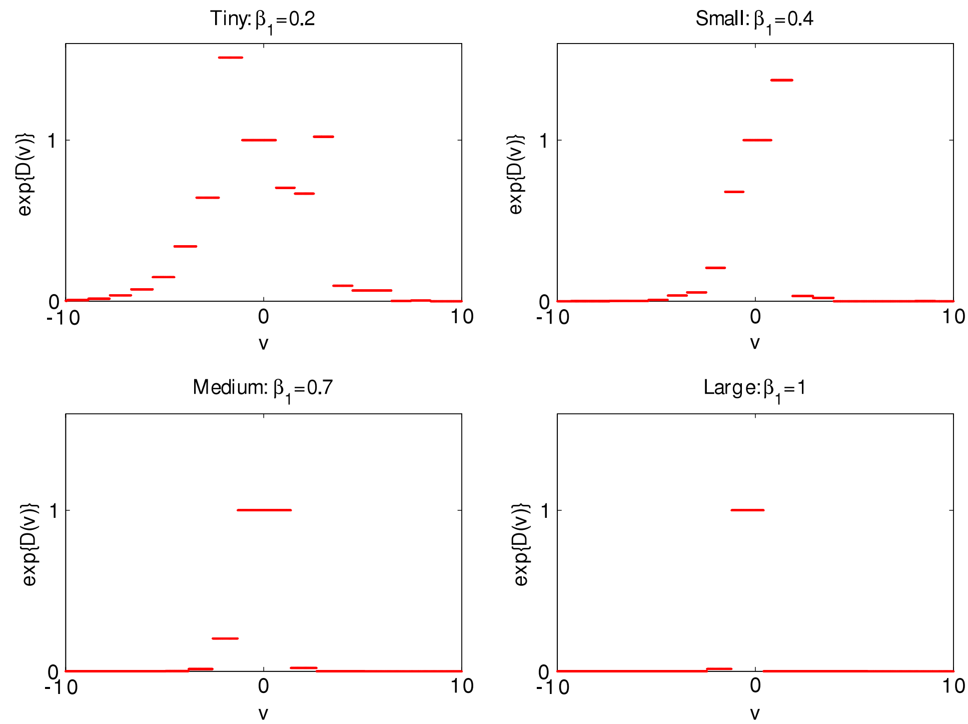

Fig. 2. Sample paths of $\exp\{D(v)\}$ for four values of β_1.

4. Simulations

Two simulations are presented in this section. For convenience, the following simple setup is used:

$$y = \begin{cases} \beta_1 + \sigma_1 e, & q \le \gamma; \\ \sigma_2 e, & q > \gamma; \end{cases}$$
$$q \sim U[0, 1], e \sim N(0, 1), \text{ and } q \text{ is independent of } e,$$

where σ_1, σ_2 and β_1 are known, and γ is the only parameter of interest. $\gamma_0 = 1/2$. $\sigma_{10} = 0.2$, $\sigma_{20} = 0.4$, and four values of β_1 are used: 0.2, 0.4, 0.7 and 1, corresponding to tiny, small, medium and large threshold effects, respectively.

The first simulation is to compare the risks of the BE and the MLEs, while the second is to compare coverage and length properties of methods for CI construction, including the asymptotic method, parametric bootstrap, parametric wild subsampling and posterior interval. The results for parametric subsampling are omitted since their performance is between parametric bootstrap and parametric wild subsampling.[9] The performance of Hansen's method (2000) will be reported in Yu (2008a), since it uses a different asymptotic framework and is a semiparametric method.

It is useful to study asymptotic properties of the estimators before examining their finite-sample performance. From Section 3.1,

$$z_{1i} = \ln\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{1}{2}\left(e_i^2 - \frac{(\sigma_1 e_i + \beta_1)^2}{\sigma_2^2}\right),$$
$$z_{2i} = \ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(e_i^2 - \frac{(\sigma_2 e_i - \beta_1)^2}{\sigma_1^2}\right),$$

and $N_\ell(\cdot)$ is a standard Poisson process. Note that $E[z_{1i}] = \ln\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{1}{2}\left(1 - \left(\frac{\sigma_1}{\sigma_2}\right)^2\right) - \frac{1}{2}\left(\frac{\beta_1}{\sigma_2}\right)^2 \le -\frac{1}{2}\left(\frac{\beta_1}{\sigma_2}\right)^2 < 0$. Since $E[z_{1i}]$ reaches its maximum at $\frac{\sigma_1}{\sigma_2} = 1$, $\frac{\sigma_1}{\sigma_2}$ indeed provides information for γ. This point is not clear in the least squares estimation. When $\beta_1 = 0$, γ cannot be identified by the least squares criterion regardless of the value of $\frac{\sigma_1}{\sigma_2}$; that is, the least squares criterion can only identify the threshold effect in mean, not in variance. In contrast, using the likelihood principle, even when $\beta_1 = 0$, γ can still be identified as long as $\frac{\sigma_1}{\sigma_2} \ne 1$. $\frac{\beta_1}{\sigma_2}$ affects $E[z_{1i}]$ in an obvious way. Similar arguments apply to z_{2i}. Typical sample paths of $\exp\{D(v)\}$ for the four β_1 values are shown in Fig. 2. Comparing Fig. 2 to Fig. 1 in Section 2, we can see that $\exp\{D(v)\}$ decreases to 0 gradually rather than jumps from 1 to 0 in only one step. When β_1 gets larger, the sample path of $\exp\{D(v)\}$ gets more similar to that in Fig. 1.

Note that z_{1i} and z_{2i} have different distributions, so the asymptotic distribution of the MMLE is not symmetric. Fig. 3 shows the asymptotic densities of the LMLE and MMLE based on the algorithms in Appendix D. Interestingly, while the asymptotic distribution of the LMLE is continuous, its density is not. The asymptotic densities of the posterior mean and median are visibly indistinguishable from that of the MMLE, so they are omitted here. When β_1 gets larger, the densities in Fig. 3 get more concentrated and more similar to those in Fig. 1. For example, the asymptotic density of the MMLE gets more symmetric, and the asymptotic density of the LMLE on $v > 0$ gets vanishing.

4.1. Simulation 1: risk comparison

In this simulation, the key task is to simulate from the posterior $p_n(\theta)$. The function slicesample in Matlab is used to carry out this task. The MMLE is used as the starting value, and 10 000 samples are drawn from the posterior after discarding the first 200 "burn-in" draws. The prior of γ is assumed to be uniform on $(q_{\min}, q_{\max})$. For details of the slice sampler, see Neal (2003).

[9] In the conventional boundary estimation, Chernozhukov and Hong (2004) mention two motivations for parametric subsampling in their Remarks 3.5 and 3.6: first, subsampling is less demanding in terms of computation; second, it is more robust to local misspecifications of the parametric model. In our simulation, the model is correctly specified, so there is no theoretical reason to report the results for parametric subsampling.

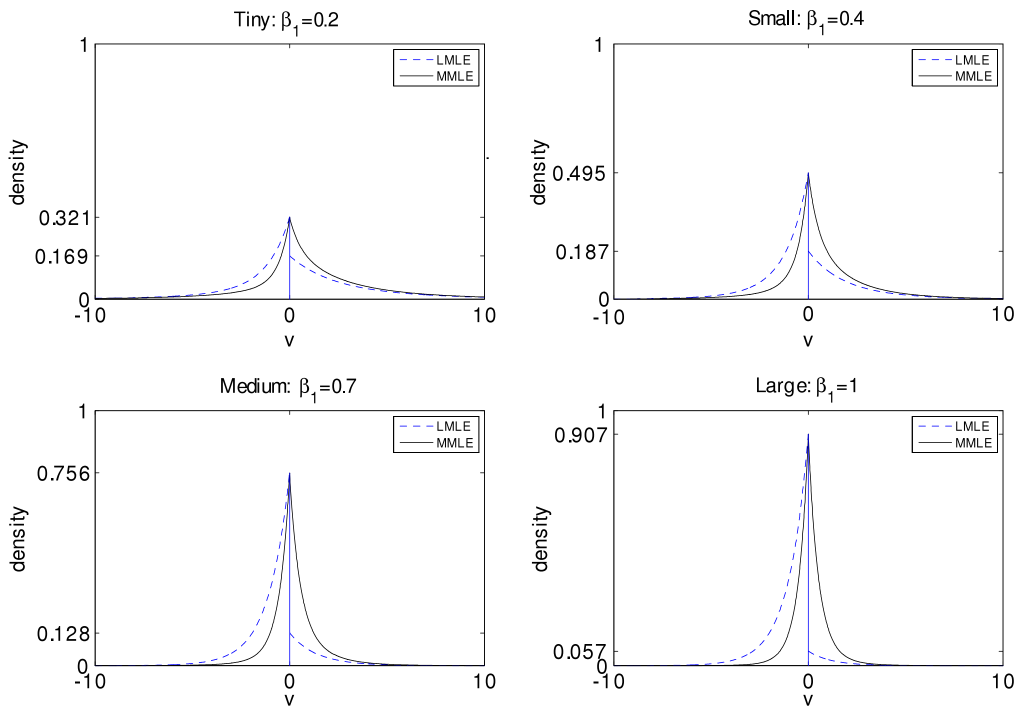

Fig. 3. Asymptotic densities of the LMLE and MMLE for four values of β_1.

The simulation results are summarized in Table 1. This table also reports the asymptotic risk of the left-endpoint LSE (LLSE) and the middle-point LSE (MLSE) which are similarly defined as the LMLE and MMLE. For the finite-sample performance of the LSEs, see Yu (2008a). The following conclusions can be drawn from Table 1. (i) The BE is more efficient than the MLEs even in finite samples. The difference between the BE and the MLEs gets larger when β_1 gets smaller. (ii) The asymptotic and finite-sample risk of the MMLE is between the LMLE and the BE when β_1 is not too small. When $\beta_1 = 0.2$, the LMLE can be more efficient than the MMLE. (iii) The asymptotic risk of the LSEs is larger than the MLEs. This is because the MLEs incorporate the information in the threshold effect of variance.[10] Such a result is particularly true when β_1 is small, since the threshold effect in variance provides significant information when γ is hard to identify by the threshold effect in mean. (iv) Under the mean square error loss, the posterior mean is most efficient both asymptotically and in finite samples. Under the absolute deviation error loss, the posterior median is most efficient both asymptotically and in finite samples. (v) When β_1 is not too small, the risk when $n = 400$ is roughly a quarter of the risk when $n = 100$, and the risk of $n = 100$ is roughly $\frac{1}{100}$ of the asymptotic risk. This justifies the n consistency of all estimators and suggests that the finite-sample distributions of these estimators are close to their asymptotic distributions even when n is as small as 100. (vi) When β_1 is small, the risk when $n = 400$ is less than a quarter of the risk when $n = 100$, and is about $\frac{1}{400}$ of the asymptotic risk. This suggests that a larger sample size is needed to approximate the asymptotic distribution when β_1 is small. The finite-sample distribution when β_1 is small and $n = 100$ may be better approximated in the framework of Hansen (2000). (vii) The speed of efficiency gain from a larger β_1 becomes slower when β_1 gets larger. The limit risk when $\beta_1 = \infty$ corresponds to the risk in the extreme case in Section 2. So the suggestions based on this simulation are (a) use the Bayes method to estimate γ; (b) if the maximum likelihood method is used, the MMLE is preferable to the LMLE.

4.2. Simulation 2: comparison of confidence intervals

First we will briefly describe the parametric wild subsampling method for constructing CIs of γ in the general model (1). Supposing the MMLE is used in the subsampling procedure, the algorithms are as follows:

Algorithm 1 (*Generating the Parametric Wild Subsampling Sample* $\{y_i^*, x_i^*, q_i^*\}_{i=1}^m$).

1. Get the MMLE of θ, denoted as $\widehat{\theta}$.
2. Generate a sequence $\{x_i^*, q_i^*\}_{i=1}^m$ by sampling without replacement from $\{x_i, q_i\}_{i=1}^n$.
3. Generate e_i^* from the conditional distribution $f_{e|x,q}\left(e|x_i^*, q_i^*; \widehat{\alpha}\right)$, $i = 1, \ldots, m$.
4. Generate a sequence $\{y_i^*\}_{i=1}^m$ by

$$y_i^* = \begin{cases} x_i^{*\prime}\widehat{\beta}_1 + \widehat{\sigma}_1 e_i^*, & \text{if } q_i^* \le \widehat{\gamma}, \\ x_i^{*\prime}\widehat{\beta}_2 + \widehat{\sigma}_2 e_i^*, & \text{if } q_i^* > \widehat{\gamma}. \end{cases}$$

Algorithm 2 (*Constructing the Parametric Wild Subsampling CI for a Fixed Block Size m*).

1. Generate $\{y_i^*, x_i^*, q_i^*\}_{i=1}^m$ by Algorithm 1.
2. Calculate the MMLE $\widehat{\theta}^*$ based on $\{y_i^*, x_i^*, q_i^*\}_{i=1}^m$.
3. Repeat step 1 and 2 B times to get a sequence of estimators $\{\widehat{\theta}_b^*\}_{b=1}^B$.
4. Find the $\frac{\tau}{2}$ and $\left(1 - \frac{\tau}{2}\right)$ percentiles of $\{m\left(\widehat{\gamma}_b^* - \widehat{\gamma}\right)\}_{b=1}^B$, denoted as $c_{nm}\left(\frac{\tau}{2}\right)$ and $c_{nm}\left(1 - \frac{\tau}{2}\right)$, then the equal-tailed subsampling CI for γ is $\left[\widehat{\gamma} - n^{-1}c_{nm}\left(1 - \frac{\tau}{2}\right), \widehat{\gamma} - n^{-1}c_{nm}\left(\frac{\tau}{2}\right)\right]$. Similarly, the symmetric subsampling CI for γ is constructed by finding the $1 - \tau$ percentile of $\{m\left|\widehat{\gamma}_b^* - \widehat{\gamma}\right|\}_{b=1}^B$, denoted as $c_{nm}(1 - \tau)$, and then constructing the CI as $[\widehat{\gamma} - n^{-1}c_{nm}(1 - \tau), \widehat{\gamma} + n^{-1}c_{nm}(1 - \tau)]$.

10 When there is no threshold effect in variance, the MLEs and the LSEs are equivalent in this setup.

Table 1
Estimator performance for γ (based on 1000 repetitions).

Estimators	β_1 Values							
	$\beta_1 = 0.2$		$\beta_1 = 0.4$		$\beta_1 = 0.7$		$\beta_1 = 1$	
	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD
$n = 100$ Risk ($\times 10^{-2}$)								
LMLE	5.032	3.107	2.899	1.810	1.470	1.042	1.548	1.062
MMLE	5.149	3.107	2.817	1.632	1.192	0.755	0.960	0.616
Posterior mean	4.033	2.702	2.310	1.519	1.066	0.747	0.859	0.594
Posterior median	4.284	2.672	2.468	1.475	1.085	0.718	0.861	0.588
$n = 400$								
LMLE	1.137	0.694	0.726	0.477	0.420	0.284	0.384	0.267
MMLE	1.140	0.686	0.724	0.440	0.345	0.210	0.233	0.160
Posterior mean	0.974	0.642	0.578	0.383	0.302	0.199	0.218	0.156
Posterior median	1.008	0.598	0.600	0.373	0.313	0.198	0.219	0.154
$n = \infty$ Risk								
LMLE	4.486	2.808	2.599	1.701	1.673	1.163	1.475	1.041
MMLE	4.562	2.815	2.561	1.577	1.346	0.840	0.923	0.607
LLSE	14.627	8.300	4.288	2.491	1.892	1.258	1.499	1.051
MLSE	14.828	8.430	4.395	2.444	1.733	1.005	1.076	0.674
Posterior mean	3.721	2.535	2.082	1.434	1.149	0.793	0.846	0.591
Posterior median	3.806	2.413	2.182	1.388	1.175	0.775	0.854	0.584

If the parametric subsampling is used, then step 2 of Algorithm 1 changes to "Generate a sequence $\{x_i^*, q_i^*\}_{i=1}^m$ by sampling from $f_{x,q}(x, q)$". If the parametric bootstrap is used, the only difference from the parametric subsampling is that n instead of m is used in Algorithm 2. Also, the following Algorithm 3 is not needed.

Algorithm 3 (*Selecting the Block Size m*)**.**

1. Fix a selection of reasonable block size m between m_{low} and m_{up}.
2. Generate K pseudo sequences $\{y_{ki}^*, x_{ki}^*, q_{ki}^*\}_{i=1}^n$, $k = 1, \ldots, K$ by Algorithm 1. For each k and m, compute a subsampling confidence interval $\text{CI}_{k,m}$ for γ by Algorithm 2.
3. Compute $\widehat{g}(m) = \#\{\widehat{\gamma} \in \text{CI}_{k,m}\}/K$.
4. Find the value $\widehat{m}$ that minimizes $|\widehat{g}(m) - (1-\tau)|$.

The asymptotic confidence interval is constructed by the following algorithm:

Algorithm 4 (*Constructing the Asymptotic CI*)**.**

1. Get the MMLE of θ, denoted as $\widehat{\theta}$.
2. Find the $\frac{\tau}{2}$ and $1-\frac{\tau}{2}$ quantiles of the asymptotic distribution of $\widehat{\gamma}$ using algorithms in Appendix D, denoted as $c(\tau/2)$ and $c(1-\tau/2)$. Similarly, find the $1-\tau$ quantile for the asymptotic distribution of $|\widehat{\gamma}|$, denoted as $c(1-\tau)$.
3. The asymptotic equal-tailed CI is constructed as $[\widehat{\gamma} - n^{-1}c(1-\frac{\tau}{2}),$ $\widehat{\gamma} - n^{-1}c(\frac{\tau}{2})]$, and the symmetric CI is constructed as $[\widehat{\gamma} - n^{-1}c(1-\tau), \widehat{\gamma} + n^{-1}c(1-\tau)]$.

The simulation results are summarized in Table 2. As discussed in Section 3.4, the asymptotic method and parametric bootstrap are not practical because $f_{x,q}(x, q)$ is unknown, they are reported here only for comparison. Among all methods for CI construction except the posterior interval, only the results associated with the MMLE are reported, since the MMLE works better than the LMLE in most cases. $n/4$ is used as the block size in the subsampling method. It is time-consuming to use Algorithm 3 to select the block size adaptively. The following selection of the parameters in Algorithm 3 is suggested: for $n = 100$, $m_{\text{low}} = 15$, $m_{\text{up}} = 40$; for $n = 400$, $m_{\text{low}} = 50$, $m_{\text{up}} = 150$. $K = 1000$. The number of both the bootstrap and subsampling replication is 1000.

A few results of interest from Table 2 are summarized as follows. (i) The posterior interval works the best in terms of both coverage and length, as it has a high coverage with a short length. (ii) The asymptotic method has a good coverage property, which reproduces the result in Simulation 1 that the finite-sample distribution is close to the asymptotic distribution due to the superconsistency of the MMLE. Its length property is comparable to the bootstrap method, but worse than the posterior interval. When $\beta_1 = 0.2$ and $n = 100$, there seems to be an undercover problem, which indicates that the framework of Hansen (2000) may be suitable in this case. (iii) In the asymptotic or bootstrap inference, the equal-tailed interval works generally better than the symmetric interval especially when β_1 is small, which indicates that the asymptotic distribution of the MMLE is not symmetric as shown in Fig. 3. (iv) The performance of the wild subsampling is worse than the bootstrap. This is essentially because only a "subsample" is used for inference. For regular parameters, the subsampling method has a worse coverage refinement than the bootstrap as shown in Politis and Romano (1994). There is no result about finite-sample refinement for nonregular parameters. From this simulation, it seems that such a result can also apply to nonregular parameters. (v) The length when $n = 400$ is roughly a quarter of the length when $n = 100$, which justifies the n consistency of γ estimators. So the suggestion based on this simulation is to use the posterior interval as the set estimator of γ.

5. Conclusion

This paper discusses likelihood-based estimation and inference in general parametric threshold regression models. By connecting threshold regression with the boundary literature, we find that the Bayes estimator is most efficient, and especially, strictly more efficient than the MLEs. Also, the posterior interval is proved to be an asymptotically valid confidence interval and is attractive in both length and coverage in finite samples. Algorithms are developed to calculate asymptotic distributions and risk for the estimators of the threshold point.

Acknowledgments

I am grateful to my committee members: Bruce Hansen, Jack Porter and Kenneth West for their help in producing this paper. This paper also benefits from useful suggestions of Bilgehan Karabay and seminar participants at University of Göttingen and

Table 2
Comparison of inference methods: coverage and average length of nominal 95% confidence intervals for γ (based on 1000 repetitions).

CIs	β_1 Values							
	$\beta_1 = 0.2$		$\beta_1 = 0.4$		$\beta_1 = 0.7$		$\beta_1 = 1$	
	Cov and leng ($\times 10^{-2}$)							
	Cov	Length	Cov	Length	Cov	Length	Cov	Length
$n = 100$								
Asymptotic MMLE (ET)	0.939	18.655	0.942	10.558	0.968	5.587	0.938	3.802
Asymptotic MMLE (S)	0.926	19.814	0.938	11.248	0.969	5.758	0.945	3.806
Bootstrap MMLE (ET)	0.933	18.281	0.938	10.395	0.966	5.519	0.940	3.745
Bootstrap MMLE (S)	0.923	19.321	0.937	11.056	0.965	5.671	0.945	3.731
Wild subsampling MMLE (ET)	0.883	15.453	0.920	9.719	0.945	5.036	0.913	3.334
Wild subsampling MMLE (S)	0.899	16.601	0.932	10.401	0.959	5.293	0.936	3.444
Posterior interval	0.939	12.244	0.947	7.000	0.950	3.468	0.947	2.446
$n = 400$								
Asymptotic MMLE (ET)	0.941	4.664	0.934	2.640	0.957	1.397	0.938	0.951
Asymptotic MMLE (S)	0.943	4.954	0.933	2.812	0.948	1.440	0.942	0.952
Bootstrap MMLE (ET)	0.939	4.615	0.936	2.623	0.951	1.389	0.936	0.943
Bootstrap MMLE (S)	0.941	4.882	0.932	2.787	0.945	1.424	0.938	0.941
Wild subsampling MMLE (ET)	0.935	4.493	0.924	2.537	0.939	1.327	0.910	0.877
Wild subsampling MMLE (S)	0.941	4.775	0.935	2.734	0.947	1.389	0.925	0.904
Posterior interval	0.940	3.134	0.944	1.715	0.942	0.871	0.950	0.615

Note: "ET" for equal-tailed CI and "S" for symmetric CI.

UW-Madison for which I want to express my gratitude. Special thanks go to the co-editor John Geweke, four referees, and especially, one associate editor for constructive comments.

Appendix A. Regularity conditions

First, notations are collected for reference in all assumptions, lemmas and proofs. In all statements about a general θ_n, only the case for $\theta_n = \theta_0$ is proved. The general case only complicates notations without changing the essential idea.

$\mathcal{N}$ is an open neighborhood of θ_0, $\mathcal{N}_\ell$ is an open neighborhood of $\theta_{\ell 0}$, and $\mathcal{N}_\gamma$ is an open neighborhood of γ_0. $\mathcal{N}_0$ is the closure of an η-ball around θ_0 such that $B(\theta_0, 2\eta) \subset \mathcal{N}$, where $B(\theta_0, 2\eta)$ is a ball with center θ_0 and radius 2η.

$$Z_n^d(v) = \exp\left\{\sum_{i=1}^n \bar{z}_{1i}\mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \le \gamma_0\right)\right\} + \exp\left\{\sum_{i=1}^n \bar{z}_{2i}\mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{n}\right)\right\}$$

$$S(\mathrm{w}|\theta) = \begin{pmatrix} -\frac{\partial \ln f_{e|x,q}}{\partial e}(e|x,q;\alpha)\frac{x'}{\sigma_1}\mathbf{1}(q \le \gamma) \\ -\frac{\partial \ln f_{e|x,q}}{\partial e}(e|x,q;\alpha)\frac{x'}{\sigma_2}\mathbf{1}(q > \gamma) \\ -\frac{1}{\sigma_1}\left(1 + \frac{\partial \ln f_{e|x,q}}{\partial e}(e|x,q;\alpha)e\right)\mathbf{1}(q \le \gamma) \\ -\frac{1}{\sigma_2}\left(1 + \frac{\partial \ln f_{e|x,q}}{\partial e}(e|x,q;\alpha)e\right)\mathbf{1}(q > \gamma) \\ \frac{\partial \ln f_{e|x,q}}{\partial \alpha}(e|x,q;\alpha) \end{pmatrix} \equiv \begin{pmatrix} S_{\beta_1}(\theta) \\ S_{\beta_2}(\theta) \\ S_{\sigma_1}(\theta) \\ S_{\sigma_2}(\theta) \\ S_\alpha(\theta) \end{pmatrix}$$

is the score function of $\underline{\theta}$, and $S_i = S(\mathrm{w}_i|\theta_0)$.

$$\mathcal{J}(\theta) \equiv E\left[S(\mathrm{w}|\theta)S'(\mathrm{w}|\theta)\right] \quad \text{and}$$
$$\mathcal{J} = \mathcal{J}(\theta_0) = E\left[S_i S_i'\right] \tag{8}$$

are the information matrices of regular parameters, and $\bar{z} = \frac{\mathcal{J}^{-1}}{\sqrt{n}}\sum_{i=1}^n S_i$.

$$LR_n\left(\bar{z}, Z_n^d, h\right) = -\frac{1}{2}u'\mathcal{J}u + u'\mathcal{J}\bar{z} + \ln Z_n^d(v)$$

is an approximation of the log likelihood ratio statistic. $\bar{z}$ is the asymptotically sufficient statistic for regular parameters, and Z_n^d is the asymptotically sufficient statistic for the nonregular parameter.

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ln f_{y|x,q}(w_i|\theta), \qquad Q(\theta) = E\left[\ln f_{y|x,q}(w|\theta)\right],$$
$$\mathbb{G}_n \ln f_{y|x,q}(w|\theta) = \sqrt{n}\left(Q_n(\theta) - Q(\theta)\right).$$

The following formula is used repetitively in Appendices B and C:

$$\begin{aligned}
&\ln f_{y|x,q}(w|\theta) - \ln f_{y|x,q}(w|\theta_0) \\
&= \ln \frac{f_{e|x,q}(e|x,q;\theta_1)}{f_{e|x,q}(e|x,q;\theta_{10})}\mathbf{1}(q \le \gamma \wedge \gamma_0) \\
&\quad + \ln \frac{f_{e|x,q}(e|x,q;\theta_2)}{f_{e|x,q}(e|x,q;\theta_{20})}\mathbf{1}(q > \gamma \vee \gamma_0) \\
&\quad + \bar{z}_1(\mathrm{w}|\theta_2, \theta_{10})\,\mathbf{1}(\gamma \wedge \gamma_0 < q \le \gamma_0) \\
&\quad + \bar{z}_2(\mathrm{w}|\theta_1, \theta_{20})\,\mathbf{1}(\gamma_0 < q \le \gamma \vee \gamma_0) \\
&\equiv A(\mathrm{w}|\theta) + B(\mathrm{w}|\theta) + C(\mathrm{w}|\theta) + D(\mathrm{w}|\theta),
\end{aligned} \tag{9}$$

where $e|x,q \sim f_{e|x,q}(e|x,q;\alpha_0)$, and $f_{e|x,q}(e|x,q;\theta_{\ell 0}) = \frac{1}{\sigma_{\ell 0}}f_{e|x,q}(e|x,q;\alpha_0)$.

Assumptions on the data generating process

Assumption D1. (x, q) has a marginal density $f_{x,q}$, and e has a conditional density $f_{e|x,q}(e|x,q;\alpha)$ which is continuously differentiable in both e and α for almost every w.

Assumption D2. $f_q(\cdot)$ is continuous, $0 < \underline{f}_q \le f_q(q) \le \overline{f}_q < \infty$ for $q \in \Gamma$, $P(q < \underline{\gamma}) > 0$ and $P(q > \overline{\gamma}) > 0$.

Assumption D3. $\mathcal{J}(\theta)$ is continuous, nonsingular and finite for $\theta \in \mathcal{N}$.

Assumption D4. For every $(\mu_1, \sigma_1, \alpha_1)$ and $(\mu_2, \sigma_2, \alpha_2)$ with σ_1 and σ_2 in a bounded set and α_1 and α_2 in Λ, there exists a slope function $m(\mathrm{w})$ such that $E\left[m(\mathrm{w})^4\right] < \infty$,

$$\left|\ln f_{e|x,q}\left(\mu_1 + \sigma_1 e|x, q; \alpha_1\right) - \ln f_{e|x,q}\left(\mu_2 + \sigma_2 e|x, q; \alpha_2\right)\right| \leq m(\mathrm{w}) \left(|\mu_1 - \mu_2| + |\sigma_1 - \sigma_2| + \|\alpha_1 - \alpha_2\|\right).$$

Assumption D5. $E[\|x\|^4] < \infty$.

Assumption D6. Uniformly for $\theta_1 \in \mathcal{N}_1, \theta_2 \in \mathcal{N}_2$ and $\gamma \in \mathcal{N}_\gamma$,

$$E\left[\sup_{\widetilde{\theta}_2 \in \mathcal{N}_2} |\overline{z}_1(\mathrm{w}|\widetilde{\theta}_2, \theta_1)| \middle| q = \gamma\right] < \infty, \quad \text{and}$$
$$E\left[\sup_{\widetilde{\theta}_1 \in \mathcal{N}_1} |\overline{z}_2(\mathrm{w}|\widetilde{\theta}_1, \theta_2)| \middle| q = \gamma\right] < \infty.$$

Assumption D7. Both z_{1i} and z_{2i} have absolutely continuous distributions.

Assumption D8. Uniformly for $\theta_1 \in \mathcal{N}_1, \theta_2 \in \mathcal{N}_2$ and $\gamma \in \mathcal{N}_\gamma$,

$$E\left[\overline{z}_1(\mathrm{w}|\theta_2, \theta_1)|q = \gamma\right] < 0, \quad \text{and} \quad E\left[\overline{z}_2(\mathrm{w}|\theta_1, \theta_2)|q = \gamma\right] < 0.$$

Assumption D9. $P\left(f_{e|x,q}(e|x, q; \theta_\ell) \neq f_{e|x,q}\left(e|x, q; \widetilde{\theta}_\ell\right)|q\right) > 0$ for any $\theta_\ell, \widetilde{\theta}_\ell \in B_\ell \times \Omega_\ell \times \Lambda$, $\theta_\ell \neq \widetilde{\theta}_\ell$, and $q \in \mathbb{Q}$.

Assumption D10. $E\left[|\ln f_{e|x,q}(e|x, q; \alpha_0)|\right] < \infty$.

Assumption D11.

$$E\left[\frac{\partial}{\partial \widetilde{\theta}_1} \ln f_{e|x,q}\left(e|x, q; \widetilde{\theta}_1, \underline{\theta}_1\right) \frac{\partial}{\partial \widetilde{\theta}_1'} \ln f_{e|x,q} \times \left(e|x, q; \widetilde{\theta}_1, \underline{\theta}_1\right) \mathbf{1}(q \leq \gamma)\right]$$

and

$$E\left[\frac{\partial}{\partial \widetilde{\theta}_2} \ln f_{e|x,q}\left(e|x, q; \widetilde{\theta}_2, \underline{\theta}_2\right) \frac{\partial}{\partial \widetilde{\theta}_2'} \ln f_{e|x,q} \times \left(e|x, q; \widetilde{\theta}_2, \underline{\theta}_2\right) \mathbf{1}(q > \gamma)\right]$$

are nonsingular and finite for $\widetilde{\theta}_\ell, \underline{\theta}_\ell$, and γ in an open neighborhood of the true value.

Assumption D12. For $\omega \in [0, 1]$,

$$E\left[\frac{\partial}{\partial \widetilde{\theta}_1} \ln f_{e|x,q}\left(e|x, q; \omega\theta_1 + (1-\omega)\theta_2, \underline{\theta}_1\right) \frac{\partial}{\partial \widetilde{\theta}_1'} \ln f_{e|x,q} \times \left(e|x, q; \omega\theta_1 + (1-\omega)\theta_2, \underline{\theta}_1\right) \middle| q = \gamma\right]$$

and

$$E\left[\frac{\partial}{\partial \widetilde{\theta}_2} \ln f_{e|x,q}\left(e|x, q; \omega\theta_2 + (1-\omega)\theta_1, \underline{\theta}_2\right) \frac{\partial}{\partial \widetilde{\theta}_2'} \ln f_{e|x,q} \times \left(e|x, q; \omega\theta_2 + (1-\omega)\theta_1, \underline{\theta}_2\right) \middle| q = \gamma\right]$$

are nonsingular and finite for θ_1, θ_2 and γ in an open neighborhood of the true value.

Remark 1. All assumptions are standard in nonlinear parametric estimation. To appreciate the validity of these assumptions, we can assume that e is independent of (x, q) and follows a standard normal distribution, then all assumptions can be easily checked. For example, $E\left[m(\mathrm{w})^4\right] < \infty$ in D4 corresponds to $E[e^4] < \infty$. Assumption D8 corresponds to Condition 4 in Chan (1993) in least squares case. It is an essential assumption needed for the discontinuous threshold regression model; that is, we do not require every component of $\underline{\theta}_1$ to be different from the corresponding component of $\underline{\theta}_2$ as long as this assumption holds. Assumptions D9 and D10 are used in Theorem 2 for identification and proving consistency of the MLEs, while in Bayes estimation, Assumption L is used for identification. D11 and D12 limit the information matrix for $f_{e|x,q}\left(e|x, q; \widetilde{\theta}_\ell, \underline{\theta}_\ell\right)$ locally. They are used for bounding the tail behavior and small variations of the likelihood ratio process.

Assumptions on the prior and the loss function

Assumption P. The prior $\pi(\theta) : \mathbb{R}^{2k+3+d_\alpha} \to [0, \infty)$ is continuous and positive at θ_0 with a polynomial majorant.

Assumption L. $l : \mathbb{R}^{2k+3+d_\alpha} \to [0, \infty)$ satisfies the following four conditions:

(i) l is continuous and not identically 0;
(ii) The sets $\{x : l(x) < C\}$ are convex for all $C > 0$, and bounded for all $C > 0$ sufficiently small;
(iii) l has a polynomial majorant;
(iv) There exist numbers $\zeta > 0$ and $H_0 \geq 0$ such that for $H \geq H_0$,
$\sup\left\{l(x) : |x| \leq H^\zeta\right\} - \inf\{l(x) : |x| \geq H\} \leq 0.$[11]

Remark 2. Assumption P is standard, e.g., the noninformative prior satisfies this assumption. Assumption L is fairly weak, and the most popular loss functions satisfy this assumption. For example, l can be a convex function with a unique minimum at 0 such as $l(x) = \|x\|^r$, $r \geq 1$. Under such a loss function, the uniqueness of $Z_{\theta,\mathrm{BE}}$ is guaranteed. l is not necessarily symmetric, so it can be the check function of quantile regression as in Section 3.3. Condition (iv) in Assumption L limits the amount that the loss function can decrease in the tails.

Appendix B. Proofs

Proof of Theorem 1. From Lemma 2,

$$\ln Z_n(h) = LR_n\left(\overline{z}, Z_n^d, h\right) + o_p(1),$$

where the $o_p(1)$ is uniform for h on any compact set in $\mathbb{R}^{2k+3+d_\alpha}$. So the regular component in the weak limit of $Z_n(h)$ is straightforward, and the following proof focuses on the nonregular component $Z_n^d(\cdot)$. Suppose $v_{1J_1} < \cdots < v_{11} < v_{10} = 0$ and $0 = v_{20} < v_{21} < \cdots < v_{2J_2}$, where J_1 and J_2 are positive integers. It is sufficient to show that the weak limit of $\left\{\ln Z_n^d(v_{\ell j_\ell}) - \ln Z_n^d(v_{\ell,j_\ell-1})\right\}_{j_\ell=1}^{J_\ell}$ matches the distribution of $\left\{D(v_{\ell j_\ell}) - D\left(v_{\ell,j_\ell-1}\right)\right\}_{j_\ell=1}^{J_\ell}$ because $\left\{\ln Z_n^d(v_{\ell j_\ell})\right\}_{j_\ell=1}^{J_\ell}$ is a linear transformation of $\left\{\ln Z_n^d(v_{\ell j_\ell}) - \ln Z_n^d(v_{\ell,j_\ell-1})\right\}_{j_\ell=1}^{J_\ell}$.

[11] This implies $H^\zeta \leq H$.

The characteristic function is used to find the weak limit of $\left\{\ln Z_n^d(v_{\ell j_\ell}) - \ln Z_n^d(v_{\ell,j_\ell-1})\right\}_{j_\ell=1}^{J_\ell}$ and prove their asymptotic independence with $\bar{z}$. First, define terms as follows:

$$T_{1j_1i} = \bar{z}_{1i}\mathbf{1}\left(\gamma_0 + \frac{v_{1j_1}}{n} < q_i \le \gamma_0 + \frac{v_{1,j_1-1}}{n}\right),$$
$$T_{2j_2i} = \bar{z}_{2i}\mathbf{1}\left(\gamma_0 + \frac{v_{2,j_2-1}}{n} < q_i \le \gamma_0 + \frac{v_{2,j_2}}{n}\right),$$
$$T_{3i} = \frac{S_i}{\sqrt{n}},$$

for $j_\ell = 1, \ldots, J_\ell$. Since

$$\exp\left\{\sqrt{-1}t_{1j_1}T_{1j_1i}\right\} = 1 + \mathbf{1}\left(\gamma_0 + \frac{v_{1j_1}}{n} < q_i \le \gamma_0 + \frac{v_{1,j_1-1}}{n}\right) \times \left[\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\} - 1\right],$$
$$\exp\left\{\sqrt{-1}t_{2j_2}T_{2j_2i}\right\} = 1 + \mathbf{1}\left(\gamma_0 + \frac{v_{2,j_2-1}}{n} < q_i \le \gamma_0 + \frac{v_{2,j_2}}{n}\right) \times \left[\exp\left\{\sqrt{-1}t_{2j_2}\bar{z}_{2i}\right\} - 1\right],$$

it follows

$$\begin{aligned}
&E\left[\exp\left\{\sqrt{-1}\left(\sum_{j_1=1}^{J_1} t_{1j_1}T_{1j_1i} + \sum_{j_2=1}^{J_2} t_{2j_2}T_{2j_2i} + t_3'T_{3i}\right)\right\}\right] \\
&= E\left[\prod_{j_1=1}^{J_1}\exp\left\{\sqrt{-1}t_{1j_1}T_{1j_1i}\right\}\prod_{j_2=1}^{J_2}\exp\left\{\sqrt{-1}t_{2j_2}T_{2j_2i}\right\} \times \exp\left\{\sqrt{-1}t_3'T_{3i}\right\}\right] \\
&= E\left[\exp\left\{\sqrt{-1}t_3'T_{3i}\right\}\right] \\
&\quad + \sum_{j_1=1}^{J_1}\frac{|v_{1j_1} - v_{1,j_1-1}|}{n} f_q(\gamma_0)\, E\left[\exp\left\{\sqrt{-1}t_3'T_{3i}\right\} \times \left\{\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\} - 1\right\} | q_i = \gamma_0-\right] \\
&\quad + \sum_{j_2=1}^{J_2}\frac{v_{2j_2} - v_{2,j_2-1}}{n} f_q(\gamma_0)\, E\left[\exp\left\{\sqrt{-1}t_3'T_{3i}\right\} \times \left\{\exp\left\{\sqrt{-1}t_{2j_2}\bar{z}_{2i}\right\} - 1\right\} \Big| q_i = \gamma_0+\right] + o\left(\frac{1}{n}\right) \\
&= 1 + \frac{1}{n}\left[-\frac{1}{2}t_3'\mathcal{J}t_3 + f_q(\gamma_0)\sum_{j_1=1}^{J_1}|v_{1j_1} - v_{1,j_1-1}| \left(E\left[\left\{\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\}\right\} \Big| q_i = \gamma_0-\right] - 1\right)\right. \\
&\quad \left. + f_q(\gamma_0)\sum_{j_2=1}^{J_2}\left(v_{2j_2} - v_{2,j_2-1}\right) \times \left(E\left[\left\{\exp\left\{\sqrt{-1}t_{2j_2}\bar{z}_{2i}\right\}\right\} \Big| q_i = \gamma_0+\right] - 1\right)\right] + o\left(\frac{1}{n}\right)
\end{aligned}$$

where the last equality is from the Taylor expansion of $\exp\{\sqrt{-1}t_3' T_{3i}\}$. From Assumption D2, $o(1)$ in the second equality is a quantity going to zero uniformly over $i = 1, \ldots, n$. So

$$\begin{aligned}
&E\left[\exp\left\{\sqrt{-1}\left(\sum_{j_1=1}^{J_1} t_{1j_1}\left(\ln Z_n^d(v_{1j_1}) - \ln Z_n^d(v_{1,j_1-1})\right) + \sum_{j_2=1}^{J_2} t_{2j_2}\left(\left(\ln Z_n^d(v_{2j_2}) - \ln Z_n^d(v_{2,j_2-1})\right)\right) + t_3'\mathcal{J}\bar{z}\right)\right\}\right] \\
&= E\left[\exp\left\{\sqrt{-1}\sum_{i=1}^{n}\left(\sum_{j_1=1}^{J_1} t_{1j_1}T_{1j_1i} + \sum_{j_2=1}^{J_2} t_{2j_2}T_{2j_2i} + t_3'T_{3i}\right)\right\}\right] \\
&= \prod_{i=1}^{n} E\left[\exp\left\{\sqrt{-1}\sum_{j_1=1}^{J_1} t_{1j_1}T_{1j_1i} + \sum_{j_2=1}^{J_2} t_{2j_2}T_{2j_2i} + t_3'T_{3i}\right\}\right] \\
&\to \exp\left\{-\frac{1}{2}t_3'\mathcal{J}t_3 + f_q(\gamma_0)\sum_{j_1=1}^{J_1}|v_{1j_1} - v_{1,j_1-1}| \times \left(E\left[\left\{\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\}\right\} | q_i = \gamma_0-\right] - 1\right)\right. \\
&\quad \left. + f_q(\gamma_0)\sum_{j_2=1}^{J_2}\left(v_{2j_2} - v_{2,j_2-1}\right) \times \left(E\left[\left\{\exp\left\{\sqrt{-1}t_{2j_2}\bar{z}_{2i}\right\}\right\} \Big| q_i = \gamma_0+\right] - 1\right)\right\}.
\end{aligned}$$

It is well known that the characteristic function of $\mathcal{J}W$ is $\exp\left\{-\frac{1}{2}t_3'\mathcal{J}t_3\right\}$. By the definition of $D(v)$, the characteristic function of $\left\{D(v_{1j_1}) - D\left(v_{1,j_1-1}\right)\right\}_{j_1=1}^{J_1}$ is

$$\begin{aligned}
&E\left[\exp\left\{\sqrt{-1}\sum_{j_1=1}^{J_1} t_{1j_1}\left(D(v_{1j_1}) - D\left(v_{1,j_1-1}\right)\right)\right\}\right] \\
&\overset{(1)}{=} E\left[\prod_{j_1=1}^{J_1}\prod_{i=1}^{N_1\left(|v_{1j_1}|\right)-N_1\left(|v_{1,j_1-1}|\right)}\exp\left\{\sqrt{-1}t_{1j_1}z_{1i}\right\}\right] \\
&\overset{(2)}{=} E\left[\prod_{j_1=1}^{J_1}\prod_{i=1}^{N_1\left(|v_{1j_1}|\right)-N_1\left(|v_{1,j_1-1}|\right)} E\left[\exp\left\{\sqrt{-1}t_{1j_1}z_{1i}\right\}\right]\right] \\
&\overset{(3)}{=} E\left[\prod_{j_1=1}^{J_1}\prod_{i=1}^{N_1\left(|v_{1j_1}|\right)-N_1\left(|v_{1,j_1-1}|\right)} E\left[\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\} \Big| q_i = \gamma_0-\right]\right] \\
&\overset{(4)}{=} E\left[\exp\left\{\sum_{j_1=1}^{J_1}\left(N_1\left(|v_{1j_1}|\right) - N_1\left(|v_{1,j_1-1}|\right)\right)\ln \times \left(E\left[\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\} \Big| q_i = \gamma_0-\right]\right)\right\}\right] \\
&\overset{(5)}{=} \exp\left\{f_q(\gamma_0)\sum_{j_1=1}^{J_1}\left|v_{1j_1} - v_{1,j_1-1}\right| [\exp \times \left\{\ln\left(E\left[\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\} \Big| q_i = \gamma_0-\right]\right)\right\} - 1]\right\} \\
&\overset{(6)}{=} \exp\left\{f_q(\gamma_0)\sum_{j_1=1}^{J_1}\left|v_{1j_1} - v_{1,j_1-1}\right| \times \left[E\left[\exp\left\{\sqrt{-1}t_{1j_1}\bar{z}_{1i}\right\} \Big| q_i = \gamma_0-\right] - 1\right]\right\},
\end{aligned}$$

where (1), (4), (6) are obvious, (2) is from the law of iterated expectation and the fact that $N_1\left(|v_{1j_1}|\right) - N_1\left(|v_{1,j_1-1}|\right)$ is independent of z_{1i}, (3) is from the definition of z_{1i}, and (5) is from the moment generating function of Poisson random variables. A similar calculation can be applied to $\left\{D(v_{2j_2}) - D\left(v_{2,j_2-1}\right)\right\}_{j_2=1}^{J_2}$.

So by the independence among $\mathcal{J}W$, $\left\{D(v_{1j_1}) - D\left(v_{1,j_1-1}\right)\right\}_{j_1=1}^{J_1}$ and $\left\{D(v_{2j_2}) - D\left(v_{2,j_2-1}\right)\right\}_{j_2=1}^{J_2}$, the characteristic function of $Z_\infty(h)$ matches the limit of that of $Z_n(h)$, and thus the result of interest follows. Assumption D6 guarantees the existence of z_{1i} and z_{2i} in the theorem.[12] By Theorem 7 in Appendix I of Ibragimov and Has'minskii (1981) and Assumptions D3 and D6, the weak convergence can be shown to be uniformly valid for $\theta \in \mathcal{N}$ instead of a fixed point θ_0. □

Proof of Theorem 2. Only the asymptotic results for the LMLE is proved here, since the proof for the MMLE is similar. The consistency is proved in Lemma 4, and the convergence rate is proved in Lemma 5. As to the asymptotic distribution, a modified version of the argmax continuous mapping theorem (Theorem 3.2.2 in Van der Vaart and Wellner (1996)) is used.

First define some basic topological structures for this problem. Let $\mathcal{D}_K$ be the space of functions $g = g^1(u)+g^2(v) : K \subset H \longmapsto \mathbb{R}$, where $g^1(\cdot)$ is continuous, $g^2(\cdot)$ is right-continuous and piecewise constant, and K is a compact subset of H. For each $g \in \mathcal{D}_K$, let $v \longmapsto J_g(v)$ be the cadlag counting process with $J_g(0-) = 0$ and jumps of size -1 at each jump point in $g^2(v)$ for $v \geq 0$, and let $v \longmapsto J_g(-v)$ be the similar (but caglad) process also on $v \geq 0$ (the left-continuity comes from the reversed time scale).[13] In other words, $J_g(v)$ is increasing for $v < 0$ and decreasing for $v \geq 0$. For $g_1, g_2 \in \mathcal{D}_K$, define the distance $d_K(g_1, g_2)$ to be the sum of the uniform distance $\|g_1 - g_2\|_K$ and the Skorohod distance between J_{g_1} and J_{g_2}. Now, the smallest argmax function is continuous on $\mathcal{D}_K$ with respect to d_K. If we can prove

(1) $\ln Z_n(h) \xrightarrow{d} \ln Z_\infty(h)$ in $(\mathcal{D}_K, d_K)$;

then by the continuous mapping theorem, the smallest argmax of the restriction of $\ln Z_n(h)$ to K will converge weakly to the smallest argmax of the restriction of $\ln Z_\infty(h)$ to K. If we can further prove that

(2) $\frac{1}{\varphi_n}\left(\widehat{\theta}_{LMLE} - \theta_0\right) = O_p(1)$;
(3) $\arg\max L_h \ln Z_\infty(h) = O_p(1)$;
(4) $\arg\max L_h \ln Z_\infty(h)$ is unique;

then by checking the proof of Theorem 3.2.2 of Van der Vaart and Wellner (1996), the proof is complete. (1) is proved in Lemma 6, (2) is proved in Lemma 5, (3) can be seen from the algorithms in Appendix D, and Assumption D7 guarantees (4), so the proof is complete. □

Proof of Theorem 3. The proof of (ii) and (iii) follows from Theorem I.10.2 of Ibragimov and Has'minskii (1981), and the convergence rate of $\widehat{\theta}_{BE}$ is an intermediate result of the proof. Since (iii) is a special case of (ii), we will focus on (ii) here. The following three conditions verify the corresponding conditions of Theorem I.10.2 in Ibragimov and Has'minskii (1981). Note that only the case that $K = \mathcal{N}_0$ is of interest.

(1) Both Hölder continuity of $Z_{n,\theta}^{1/2}(h)$ in the mean square and the exponential bound on the expected likelihood tail are proved in Lemma 8.
(2) The finite-dimensional convergence of $Z_{n,\theta}(h)$ is established in Theorem 1.
(3) The uniqueness of the minimizer of $\psi(s,t)$ is an assumption of this theorem.

The assumptions on the prior and the loss function are summarized in Assumptions P and L of Appendix A.

Now, applying Theorem I.10.2 in Ibragimov and Has'minskii (1981), we have uniformly for $\theta \in K$,

$$\varphi_n^{-1}\left(\widehat{\theta}_{BE} - \theta\right) \xrightarrow{d} Z_{\theta,BE},$$

and

$$\lim_{n\to\infty} E_\theta\left[l\left(\varphi_n^{-1}\left(\widehat{\theta}_{BE} - \theta\right)\right)\right] = E_\theta\left[l\left(Z_{\theta,BE}\right)\right] < \infty. \tag{10}$$

Part (iv) follows from Theorem I.9.1 of Ibragimov and Has'minskii (1981). The only condition we need to check is that $E_\theta\left[l\left(Z_{\theta,BE}\right)\right]$ as a function of θ is continuous and bounded on K. The boundedness is established above. To show the continuity, note that $\lim_{n\to\infty,\widetilde{\theta}\to\theta} E_{\widetilde{\theta}}\left[l\left(\varphi_n^{-1}\left(\widehat{\theta}_{BE} - \widetilde{\theta}\right)\right)\right] \to E_\theta\left[l\left(Z_{\theta,BE}\right)\right]$ by repeating the proof for (10). So $E_\theta\left[l\left(Z_{\theta,BE}\right)\right]$ is continuous on K. The optimality of the BE when the loss function is separable can be proved by a simple contradiction. □

Appendix C. Lemmas

Lemma 1 (*Lipschitz Continuity*). *Under Assumptions D0–D1 and D4–D5,*

$$\ln f_{e|x,q}(e|x,q;\theta_1), \quad \ln f_{e|x,q}(e|x,q;\theta_2),$$
$$\ln f_{e|x,q}\left(e|x,q;\theta_1,\underline{\theta}_{20}\right), \quad \ln f_{e|x,q}\left(e|x,q;\theta_2,\underline{\theta}_{10}\right)$$

are all Lipschitz continuous in $\underline{\theta}$ with the slope function in L^2 space.

Proof. Only the results for $\ln f_{e|x,q}\left(e|x,q;\theta_1\right)$ and $\ln f_{e|x,q}\left(e|x,q;\theta_1,\underline{\theta}_{20}\right)$ are proved, since the others are similar.

By Assumptions D0 and D4, we have the expressions in Box I. By the Cauchy–Schwarz inequality, $1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|$ is in L^2 based on Assumptions D4 and D5. □

Lemma 2. *Suppose Assumptions D0–D5 hold, then*

$$\ln Z_n(h) = LR_n\left(\bar{z}, Z_n^d, h\right) + o_p(1),$$

where the $o_p(1)$ is uniform for h on any compact set in $\mathbb{R}^{2k+3+d_\alpha}$.

Proof.

$$\begin{aligned}
\ln Z_n(h) &\overset{(1)}{=} \sum_{i=1}^n A\left(\mathrm{w}_i|\theta_0+\varphi_n h\right) + \sum_{i=1}^n B\left(\mathrm{w}_i|\theta_0+\varphi_n h\right) \\
&\quad + \sum_{i=1}^n C\left(\mathrm{w}_i|\theta_0+\varphi_n h\right) + \sum_{i=1}^n D\left(\mathrm{w}_i|\theta_0+\varphi_n h\right), \\
&\overset{(2)}{=} \sum_{i=1}^n \ln \\
&\quad \times \frac{\frac{1}{\sigma_{10}+\frac{u_{\sigma_1}}{\sqrt{n}}} f_{e|x,q}\left(\frac{\sigma_{10}e_i - x_i'\frac{u_{\beta_1}}{\sqrt{n}}}{\sigma_{10}+\frac{u_{\sigma_1}}{\sqrt{n}}}\middle| x_i, q_i; \alpha_0 + \frac{u_\alpha}{\sqrt{n}}\right)}{\frac{1}{\sigma_{10}} f_{e|x,q}\left(e_i|x_i,q_i;\alpha_0\right)} \\
&\quad \times \mathbf{1}\left(q_i \leq \gamma_0\right) + \sum_{i=1}^n \ln \\
&\quad \times \frac{\frac{1}{\sigma_{20}+\frac{u_{\sigma_2}}{\sqrt{n}}} f_{e|x,q}\left(\frac{\sigma_{20}e_i - x_i'\frac{u_{\beta_2}}{\sqrt{n}}}{\sigma_{20}+\frac{u_{\sigma_2}}{\sqrt{n}}}\middle| x_i, q_i; \alpha_0 + \frac{u_\alpha}{\sqrt{n}}\right)}{\frac{1}{\sigma_{20}} f_{e|x,q}\left(e_i|x_i,q_i;\alpha_0\right)} \\
&\quad \times \mathbf{1}\left(q_i > \gamma_0\right) + \ln Z_n^d(v) + o_p(1) \\
&\overset{(3)}{=} -\frac{1}{2}u'\mathcal{J}u + u'\mathcal{J}\bar{z} + \ln Z_n^d(v) + o_p(1),
\end{aligned}$$

[12] Such a result is from IV.3.2 on Page 124 of Neveu (1965).

[13] The class of functions of g such that $g^2(v)$ is the same is treated as the same element in the mapping $g \longmapsto J_g(v)$.

$$\left|\ln \frac{1}{\widetilde{\sigma}_1} f_{e|x,q}\left(\frac{\sigma_{10}e + x'\left(\beta_{10}-\widetilde{\beta}_1\right)}{\widetilde{\sigma}_1}\middle| x, q; \widetilde{\alpha}\right) - \ln \frac{1}{\sigma_1} f_{e|x,q}\left(\frac{\sigma_{10}e + x'\left(\beta_{10}-\beta_1\right)}{\sigma_1}\middle| x, q; \alpha\right)\right|$$
$$\le \frac{|\widetilde{\sigma}_1-\sigma_1|}{\min\{\sigma_1,\widetilde{\sigma}_1\}} + m(\mathrm{w})\left(\frac{|\widetilde{\sigma}_1-\sigma_1|\left(\|x\|\|\beta_{10}\|+\sigma_{10}\right)+\|x\|\left(\sigma_1\left\|\widetilde{\beta}_1-\beta_1\right\|+\|\beta_1\||\widetilde{\sigma}_1-\sigma_1|\right)}{\sigma_1\widetilde{\sigma}_1} + \|\widetilde{\alpha}-\alpha\|\right)$$
$$\le C\left(1+m(\mathrm{w})+m(\mathrm{w})\|x\|\right)\left(|\widetilde{\sigma}_1-\sigma_1| + \left\|\widetilde{\beta}_1-\beta_1\right\| + \|\widetilde{\alpha}-\alpha\|\right),$$

and

$$\ln \frac{1}{\widetilde{\sigma}_2} f_{e|x,q}\left(\frac{\sigma_{10}e + x'\left(\beta_{10}-\widetilde{\beta}_2\right)}{\widetilde{\sigma}_2}\middle| x, q; \widetilde{\alpha}\right)$$
$$- \ln \frac{1}{\sigma_2} f_{e|x,q}\left(\frac{\sigma_{10}e + x'\left(\beta_{10}-\beta_2\right)}{\sigma_2}\middle| x, q; \alpha\right)$$
$$\le m(\mathrm{w})\left(\frac{\sigma_{10}}{\sigma_2\widetilde{\sigma}_2}|\widetilde{\sigma}_2-\sigma_2| + \frac{\|x\|}{\widetilde{\sigma}_2}\left\|\widetilde{\beta}_2-\beta_2\right\|\right.$$
$$\left. + \frac{|\widetilde{\sigma}_2-\sigma_2|}{\sigma_2\widetilde{\sigma}_2}\|x\|\left(\|\beta_{10}\|+\|\beta_2\|\right) + \|\widetilde{\alpha}-\alpha\|\right) + \frac{|\widetilde{\sigma}_2-\sigma_2|}{\min\{\sigma_2,\widetilde{\sigma}_2\}}$$
$$\le C\left(1+m(\mathrm{w})+m(\mathrm{w})\|x\|\right)$$
$$\times\left(|\widetilde{\sigma}_2-\sigma_2| + \left\|\widetilde{\beta}_2-\beta_2\right\| + \|\widetilde{\alpha}-\alpha\|\right).$$

Box I.

where (1) is from (9), (2) is from Lemma 1 (Lipschitz Continuity) and Assumption D2, and $o_p(1)$ here is uniform for h on any compact set in $\mathbb{R}^{2k+3+d_\alpha}$. Some explanation of (3) is given as follows. From Assumptions D1 and D3, the model is differentiable in quadratic mean at $\underline{\theta}_0$ when γ_0 is known by Lemma 7.6 of Van der Vaart (1998). Theorem 7.2 of Van der Vaart (1998) shows that

$$\sum_{i=1}^n \ln \frac{\frac{1}{\sigma_{10}+\frac{u_{\sigma_1}}{\sqrt{n}}} f_{e|x,q}\left(\frac{y_i - x_i'\left(\beta_{10}+\frac{u_{\beta_1}}{\sqrt{n}}\right)}{\sigma_{10}+\frac{u_{\sigma_1}}{\sqrt{n}}}\middle| x_i, q_i; \alpha_0 + \frac{u_\alpha}{\sqrt{n}}\right)}{\frac{1}{\sigma_{10}} f_{e|x,q}\left(e_i|x_i,q_i;\alpha_0\right)}$$
$$\times \mathbf{1}(q_i\le\gamma_0) + \sum_{i=1}^n \ln$$
$$\times \frac{\frac{1}{\sigma_{20}+\frac{u_{\sigma_2}}{\sqrt{n}}} f_{e|x,q}\left(\frac{y_i - x_i'\left(\beta_{20}+\frac{u_{\beta_2}}{\sqrt{n}}\right)}{\sigma_{20}+\frac{u_{\sigma_2}}{\sqrt{n}}}\middle| x_i, q_i; \alpha_0 + \frac{u_\alpha}{\sqrt{n}}\right)}{\frac{1}{\sigma_{20}} f_{e|x,q}\left(e_i|x_i,q_i;\alpha_0\right)}$$
$$\times \mathbf{1}(q_i>\gamma_0)$$
$$= -\frac{1}{2}u'\mathcal{J}u + u'\mathcal{J}\mathcal{Z} + o_p(1),$$

where the residual is $o_p(1)$ under θ_0. But from Lemma 19.31 of Van der Vaart (1998) and Lemma 1, this $o_p(1)$ can be strengthened to be uniform for u on any compact set. □

Lemma 3 (*Identification*)**.** *Under Assumptions D0–D2, θ is identified in the sense that*

$$P\left(f_{y|x,q}(w|\theta) \ne f_{y|x,q}\left(w|\widetilde{\theta}\right)\right) > 0$$

for any θ and $\widetilde{\theta}\in\Theta$ and $\theta\ne\widetilde{\theta}$.

Proof. This is equivalent to the statement that $P(f_{y|x,q}(w|\theta) = f_{y|x,q}(w|\widetilde{\theta})) = 1$ implies $\theta = \widetilde{\theta}$. If not, then there are two cases: $\widetilde{\gamma}\ne\gamma$ and $\widetilde{\gamma}=\gamma$. For $\widetilde{\gamma}\ne\gamma$, there are three subcases:

Case(i): $\widetilde{\gamma}\ne\gamma, \underline{\theta} = \widetilde{\underline{\theta}}$. By D2, $f_q(\gamma) > 0$. So D0 and D9 imply $P\left(f_{y|x,q}(w_i|\theta) = f_{y|x,q}\left(w_i|\widetilde{\theta}\right)\right) < 1$, and thus $\widetilde{\gamma}=\gamma$;

Case (ii): $\widetilde{\gamma} > \gamma, \underline{\theta} \ne \widetilde{\underline{\theta}}, \widetilde{\theta}_1 = \theta_2$; or $\widetilde{\gamma} < \gamma, \underline{\theta} \ne \widetilde{\underline{\theta}}, \widetilde{\theta}_2 = \theta_1$. From Assumption D2, $P(q<\gamma) > 0$. Then by Assumptions D0 and D9, $P\left(f_{y|x,q}(w_i|\theta) = f_{y|x,q}\left(w_i|\widetilde{\theta}\right)\right) < 1$, so $\widetilde{\gamma}=\gamma$;

Case (iii): $\widetilde{\gamma} > \gamma, \widetilde{\theta}_1 \ne \theta_2$; or $\overline{\gamma} < \gamma, \widetilde{\theta}_2 \ne \theta_1$. Similar arguments as in Case (i) lead to $\widetilde{\gamma}=\gamma$.

So $\widetilde{\gamma}=\gamma$. With $\widetilde{\gamma}=\gamma$, by Assumption D9, $\widetilde{\theta}_\ell = \theta_\ell$. In summary, $P\left(f_{y|x,q}(w_i|\theta) = f_{y|x,q}\left(w_i|\widetilde{\theta}\right)\right) = 1$ implies $\theta=\widetilde{\theta}$. □

Lemma 4 (*Consistency*)**.** *Under Assumptions D0–D2, D4–D5 and D9–D10, $\widehat{\theta}_{LMLE} \xrightarrow{p} \theta_0$, and $\widehat{\theta}_{MMLE} \xrightarrow{p} \theta_0$.*

Proof. Theorem 2.1 of Newey and McFadden (1994) is used in this proof. First, since the density of q is bounded on Γ, $Q(\theta)$ is continuous. Second, the uniqueness of the maximizer of $Q(\theta)$ follows from Lemma 3 by the Kullback–Leibler information inequality. It remains to show that $Q_n(\theta)$ converges uniformly in probability to $Q(\theta)$, which will be proved by applying Lemma 2.8 of Pakes and Pollard (1989). So we need to check the class of functions $\left\{\ln f_{y|x,q}(w|\theta) : \theta\in\Theta\right\}$ is Euclidean with an envelope that has a finite first moment.

$$\ln f_{y|x,q}(w|\theta) = \ln f_{e|x,q}(e|x,q;\theta_1)\,\mathbf{1}(q\le\gamma\wedge\gamma_0)$$
$$+ \ln f_{e|x,q}(e|x,q;\theta_2)\,\mathbf{1}(q>\gamma\vee\gamma_0)$$
$$+ \ln f_{e|x,q}\left(e|x,q;\theta_2,\underline{\theta}_{10}\right)\mathbf{1}(\gamma\wedge\gamma_0<q\le\gamma_0)$$
$$+ \ln f_{e|x,q}\left(e|x,q;\theta_1,\underline{\theta}_{20}\right)\mathbf{1}(\gamma_0<q\le\gamma\vee\gamma_0).$$

From Lemma 1, the class of functions $\left\{\ln f_{e|x,q}(e|x,q;\theta_1) : \underline{\theta}\in\underline{\Theta}\right\}$ is Lipschitz continuous. By Lemma 2.13 of Pakes and Pollard (1989), it is Euclidean with envelope $|\ln f_{e|x,q}(e|x,q;\theta_{10})| + C(1+m(\mathrm{w})+m(\mathrm{w})\|x\|)$. $\{\mathbf{1}(q\le\gamma\wedge\gamma_0):\gamma\in\Gamma\}$ is Euclidean with envelope 1 by Lemma 2.4 of Pakes and Pollard (1989). So $\left\{\ln f_{e|x,q}(e|x,q;\theta_1)\,\mathbf{1}(q\le\gamma\wedge\gamma_0):\theta\in\Theta\right\}$ is Euclidean with envelope $\left|\ln f_{e|x,q}(e|x,q;\theta_{10})\right| + C(1+m(\mathrm{w})+m(\mathrm{w})\|x\|)$ by Lemma 2.14(ii) of Pakes and Pollard (1989). Similarly, $\{\ln f_{e|x,q}(e|x,q;\theta_2)\mathbf{1}(q>\gamma\vee\gamma_0):\theta\in\Theta\}$ is Euclidean with envelope $|\ln f_{e|x,q}(e|x,q;\theta_{20})| + C(1+m(\mathrm{w})+m(\mathrm{w})\|x\|)$, $\{\ln f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{10})\mathbf{1}(\gamma\wedge\gamma_0<q\le\gamma_0):\theta\in\Theta\}$ is Euclidean with envelope $\left|\ln f_{e|x,q}\left(e|x,q;\theta_{20},\underline{\theta}_{10}\right)\right|$

$+ C(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|)$, and $\{\ln f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{20})\mathbf{1}(\gamma_0 < q \le \gamma \vee \gamma_0) : \theta \in \Theta\}$ is Euclidean with envelope $\left|\ln f_{e|x,q}(e|x,q;\theta_{10},\underline{\theta}_{20})\right| + C(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|)$. By Assumption D4,

$$\left|\ln f_{e|x,q}(e|x,q;\theta_{20},\underline{\theta}_{10}) - \ln f_{e|x,q}(e|x,q;\alpha_0)\right| \le |\ln\sigma_{20}| + m(\mathrm{w})\left(\left|\frac{\sigma_{10}}{\sigma_{20}} - 1\right| + \frac{\|x\|\,\|\beta_{10}-\beta_{20}\|}{\sigma_{20}}\right).$$

A similar result applies to $\left|\ln f_{e|x,q}(e|x,q;\theta_{10},\underline{\theta}_{20})\right|$. So by Lemma 2.4(i) of Pakes and Pollard (1989), $\left\{\ln f_{y|x,q}(w|\theta) : \theta \in \Theta\right\}$ is Euclidean with envelope

$$C\left(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\| + \left|\ln f_{e|x,q}(e|x,q,\alpha_0)\right|\right),$$

which has a finite first moment by Assumption D10 and Lemma 1. □

Lemma 5 (*Rate of Convergence*). *Under Assumptions D0–D5, D8–D10, $\frac{1}{\varphi_n}\left(\widehat{\theta}_{LMLE} - \theta_0\right) = O_p(1)$, and $\frac{1}{\varphi_n}\left(\widehat{\theta}_{MMLE} - \theta_0\right) = O_p(1)$.*

Proof. This proof uses Corollary 3.2.6 of Van der Vaart and Wellner (1996).

First, $Q(\theta) - Q(\theta_0) \le -Cd^2(\theta,\theta_0)$ with $d(\theta,\theta_0) = \|\underline{\theta} - \underline{\theta}_0\| + \sqrt{|\gamma - \gamma_0|}$ for $\theta \in \mathcal{N}$.

$$\begin{aligned}
&Q(\theta) - Q(\theta_0)\\
&\overset{(1)}{=} E\left[A(\mathrm{w}|\theta) + B(\mathrm{w}|\theta) + C(\mathrm{w}|\theta) + D(\mathrm{w}|\theta)\right]\\
&\overset{(2)}{=} E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_1)}{f_{e|x,q}(e|x,q;\theta_{10})}\left(\mathbf{1}(q \le \gamma\wedge\gamma_0) + \mathbf{1}(\gamma\wedge\gamma_0 < q \le \gamma_0)\right)\right]\\
&\quad + E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_2)}{f_{e|x,q}(e|x,q;\theta_{20})}\left(\mathbf{1}(q > \gamma\vee\gamma_0) + \mathbf{1}(\gamma_0 < q \le \gamma\vee\gamma_0)\right)\right] + E\left[\left(\ln\frac{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{10})}{f_{e|x,q}(e|x,q;\theta_{10})}\right.\right.\\
&\quad \left.\left. - \ln\frac{f_{e|x,q}(e|x,q;\theta_1)}{f_{e|x,q}(e|x,q;\theta_{10})}\right)\mathbf{1}(\gamma\wedge\gamma_0 < q \le \gamma_0)\right]\\
&\quad + E\left[\left(\ln\frac{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{20})}{f_{e|x,q}(e|x,q;\theta_{20})} - \ln\frac{f_{e|x,q}(e|x,q;\theta_2)}{f_{e|x,q}(e|x,q;\theta_{20})}\right)\right.\\
&\quad \left.\times\mathbf{1}(\gamma_0 < q \le \gamma\vee\gamma_0)\right]\\
&\overset{(3)}{=} E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_1)}{f_{e|x,q}(e|x,q;\theta_{10})}\mathbf{1}(q\le\gamma_0)\right]\\
&\quad + E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_2)}{f_{e|x,q}(e|x,q;\theta_{20})}\mathbf{1}(q>\gamma_0)\right]\\
&\quad + E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{10})}{f_{e|x,q}(e|x,q;\theta_1)}\mathbf{1}(\gamma\wedge\gamma_0 < q \le \gamma_0)\right]\\
&\quad + E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{20})}{f_{e|x,q}(e|x,q;\theta_2)}\mathbf{1}(\gamma_0 < q \le \gamma\vee\gamma_0)\right]\\
&\overset{(4)}{\le} -C\left(\underline{\theta}-\underline{\theta}_0\right)'\mathcal{I}\left(\underline{\theta}-\underline{\theta}_0\right) + \underline{f}_q|\gamma-\gamma_0|\\
&\quad \times\left(\sup_{\gamma\in\mathcal{N}_\gamma} E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{10})}{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{10})}\middle| q=\gamma\right]\right.\\
&\quad \left. + \sup_{\gamma\in\mathcal{N}_\gamma} E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{20})}{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{20})}\middle| q=\gamma\right]\right)\\
&\overset{(5)}{\le} -C\left(\|\underline{\theta}-\underline{\theta}_0\|^2 + |\gamma-\gamma_0|\right),
\end{aligned}$$

where (1)–(3) are straightforward. The first part of (4) is from a similar analysis in the proof of Theorem 5.39 in Van der Vaart (1998) since the model is regular when γ_0 is known, and the second part is from Assumption D2. (5) relies on $E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{10})}{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{10})}\middle| q=\gamma\right] < 0$ and $E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{20})}{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{20})}\middle| q=\gamma\right] < 0$ for $\theta \in \mathcal{N}$, which in turn follows from the following two facts: First, from Assumption D8,

$$E\left[\ln\frac{\frac{\sigma_{10}}{\sigma_2}f_{e|x,q}\left(\frac{\sigma_{10}e+x'(\beta_{10}-\beta_2)}{\sigma_2}\middle| x,q;\alpha_0\right)}{f_{e|x,q}(e|x,q;\alpha_0)}\middle| q=\gamma\right] < 0,$$

$$E\left[\ln\frac{\frac{\sigma_{20}}{\sigma_1}f_{e|x,q}\left(\frac{\sigma_{20}e+x'(\beta_{20}-\beta_1)}{\sigma_1}\middle| x,q;\alpha_0\right)}{f_{e|x,q}(e|x,q;\alpha_0)}\middle| q=\gamma\right] < 0.$$

Second,

$$E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{10})}{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{10})}\middle| q=\gamma\right] = E\left[\ln\frac{\frac{\sigma_1}{\sigma_2}f_{e|x,q}\left(\frac{\sigma_{10}e+x'(\beta_{10}-\beta_2)}{\sigma_2}\middle| x,q;\alpha\right)}{f_{e|x,q}\left(\frac{\sigma_{10}e+x'(\beta_{10}-\beta_1)}{\sigma_1}\middle| x,q;\alpha\right)}\middle| q=\gamma\right]$$

is continuous in θ_1, and

$$E\left[\ln\frac{f_{e|x,q}(e|x,q;\theta_1,\underline{\theta}_{20})}{f_{e|x,q}(e|x,q;\theta_2,\underline{\theta}_{20})}\middle| q=\gamma\right] = E\left[\ln\frac{\frac{\sigma_2}{\sigma_1}f_{e|x,q}\left(\frac{\sigma_{20}e+x'(\beta_{20}-\beta_1)}{\sigma_1}\middle| x,q;\alpha\right)}{f_{e|x,q}\left(\frac{\sigma_{20}e+x'(\beta_{20}-\beta_2)}{\sigma_2}\middle| x,q;\alpha\right)}\middle| q=\gamma\right]$$

is continuous in θ_2.

Second, $E^*\left[\sup_{d(\theta,\theta_0)<\delta}\left|\mathbb{G}_n\left(\ln f_{y|x,q}(w|\theta) - \ln f_{y|x,q}(w|\theta_0)\right)\right|\right] \le C\delta$. To obtain this result, we need to analyze the four terms in (9). From the proof of Lemma 4, $\{A(\mathrm{w}|\theta) : d(\theta,\theta_0) < \delta\}$ is a VC subgraph class of functions with envelope $C(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|)\|\underline{\theta}-\underline{\theta}_0\|$, and $\{C(\mathrm{w}|\theta) : d(\theta,\theta_0) < \delta\}$ is VC subgraph with envelope $C(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|)\mathbf{1}(\gamma\wedge\gamma_0 < q \le \gamma_0)$. Similar results apply to $B(\mathrm{w}|\theta)$ and $D(\mathrm{w}|\theta)$. So by the preservation theorem, $\left\{\ln f_{y|x,q}(w|\theta) - \ln f_{y|x,q}(w|\theta_0) : d(\theta,\theta_0) < \delta\right\}$ is VC subgraph with envelope

$$F \equiv C(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|)\left(\|\underline{\theta}-\underline{\theta}_0\| + \mathbf{1}(\gamma\wedge\gamma_0 < q \le \gamma_0) + \mathbf{1}(\gamma_0 < q \le \gamma\vee\gamma_0)\right).$$

From Theorem 2.14.2 of Van der Vaart and Wellner (1996),

$$E^*\left[\sup_{d(\theta,\theta_0)<\delta}\left|\mathbb{G}_n\left(\ln f_{y|x,q}(w|\theta) - \ln f_{y|x,q}(w|\theta_0)\right)\right|\right] \le C\|F\|_2.$$

By Assumption D2 and Lemma 1,

$$\|F\|_2 \le C\sqrt{E\left[(1 + m(\mathrm{w}) + m(\mathrm{w})\|x\|)^2\right]\delta} = C\delta.$$

So $\phi(\delta) = \delta$ in Corollary 3.2.6 of Van der Vaart and Wellner (1996), and $\frac{\delta}{\delta^\alpha}$ is decreasing for all $1 < \alpha < 2$. Since $r_n^2\phi\left(\frac{1}{r_n}\right) =$

r_n, $\sqrt{n}d\left(\widehat{\theta}_{LMLE}-\theta_0\right)=O_p(1)$ and $\sqrt{n}d\left(\widehat{\theta}_{MMLE}-\theta_0\right)=O_p(1)$. By the definition of d, the result follows. □

Lemma 6 (*Weak Convergence on a Compact Set*). *Under Assumptions D0–D7, $\ln Z_n(h)\xrightarrow{d}\ln Z_\infty(h)$ in $(\mathcal{D}_K,d_K)$, where $(\mathcal{D}_K,d_K)$ is defined in the proof of Theorem 2.*

Proof. From Theorem 1, any finite-dimensional marginal distribution of $\ln Z_n(h)$ on K converges weakly to that of $\ln Z_\infty(h)$. It remains to show that $\ln Z_n(h)$ is asymptotically tight on K. The proof for the regular part is straightforward by the analysis in Lemma 2. For the nonregular part $\ln Z_n^d(v)$, a condition called Aldous's (1978) condition is sufficient; see Theorem 16 on Page 134 of Pollard (1984). Here, we only prove such a result for

$$\ln Z_{2n}^d(v)\equiv\sum_{i=1}^n\bar{z}_{2i}\mathbf{1}\left(\gamma_0<q_i\le\gamma_0+\frac{v}{n}\right),$$

since the proof for the other part is similar. Suppose v_1 and v_2, $0<v_1<v_2$, are stopping times in K_v which is the projection of K on v coordinate, then for any $\epsilon>0$,

$$\begin{aligned}
&P\left(\sup_{|v_2-v_1|<\delta}\left|\ln Z_{2n}^d(v_2)-\ln Z_{2n}^d(v_1)\right|>\epsilon\right)\\
&\overset{(1)}{\le}P\left(\sum_{i=1}^n|\bar{z}_{2i}|\cdot\sup_{|v_2-v_1|<\delta}\mathbf{1}\left(\gamma_0+\frac{v_1}{n}<q_i\le\gamma_0+\frac{v_2}{n}\right)>\epsilon\right)\\
&\overset{(2)}{\le}\sum_{i=1}^n E\left[|\bar{z}_{2i}|\sup_{|v_2-v_1|<\delta}\mathbf{1}\left(\gamma_0+\frac{v_1}{n}<q_i\le\gamma_0+\frac{v_2}{n}\right)\right]\Big/\epsilon\\
&\overset{(3)}{\le}\frac{C\delta}{\epsilon},
\end{aligned}$$

where (1) is obvious, (2) is from Markov's inequality, and C in (3) can take $\overline{f}_q\sup_{\gamma\in\mathcal{N}_\gamma}E\left[|\bar{z}_{2i}|\,|q_i=\gamma\right]<\infty$ according to Assumptions D2 and D6. □

The following two lemmas are used to check the conditions of Theorem I.10.2 in Ibragimov and Has'minskii (1981). Lemma 8 is based on Lemma 7. The first property in Lemma 8 bounds small variations of the likelihood ratio process, and the second property bounds the tail behavior of the likelihood ratio process. The second property is used in approximating the expected posterior loss by integrals over a large bounded region. The first property is used for proving the weak convergence of the expected posterior loss on a compact set, which is the critical step when applying the argmax theorem. Note that in Bayes estimation, we do not need to check the stochastic equicontinuity of the likelihood ratio process. Instead, we need only check the stochastic equicontinuity of the expected posterior loss which is the objective function of the Bayes estimation. The expected posterior loss smoothes the jumps of the likelihood ratio process out, so its stochastic equicontinuity is easier to check. In this sense, the BE is less stringent than the MLEs on the data generating process. The first property in Lemma 8 is essentially playing the role of stochastic equicontinuity in Bayes estimation.

Define the Hellinger distance as

$$r_2(\theta,\theta+h)^2=\int\left|f^{1/2}(w|\theta+h)-f^{1/2}(w|\theta)\right|^2dw.$$

Lemma 7 (*Hellinger Distance Properties*). *Under Assumptions D0–D2 and D9–D12,*

(i) $r_2(\theta,\theta+h)^2\le C\left(\|u\|^2+|v|\right)$ *for all* $\theta\in\mathcal{N}_0$, $\theta+h\in\mathcal{N}$;

(ii) $r_2(\theta,\theta+h)^2\ge C\frac{\|u\|^2+|v|}{1+(\|u\|^2+|v|)}$ *for all* $\theta\in\mathcal{N}_0$, $\theta+h\in\Theta$.

Proof. Without loss of generality, suppose $v>0$.

(i)

$$\begin{aligned}
&r_2(\theta,\theta+h)^2\\
&=\int\left|f^{1/2}(w|\theta+h)-f^{1/2}(w|\theta)\right|^2dw\\
&=\int_{q\le\gamma}\iint\left|\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'(\beta_1+u_{\beta_1})}{\sigma_1+u_{\sigma_1}}\Big|x,q;\alpha+u_\alpha\right)}{\sqrt{\sigma_1+u_{\sigma_1}}}-\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'\beta_1}{\sigma_1}\Big|x,q;\alpha\right)}{\sqrt{\sigma_1}}\right|^2dy\cdot f_{x|q}dx\cdot f_qdq\\
&+\int_{q>\gamma+v}\iint\left|\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'(\beta_2+u_{\beta_2})}{\sigma_2+u_{\sigma_2}}\Big|x,q;\alpha+u_\alpha\right)}{\sqrt{\sigma_2+u_{\sigma_2}}}-\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'\beta_2}{\sigma_2}\Big|x,q;\alpha\right)}{\sqrt{\sigma_2}}\right|^2dy\cdot f_{x|q}dx\cdot f_qdq\\
&+\int_\gamma^{\gamma+v}\iint\left|\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'(\beta_1+u_{\beta_1})}{\sigma_1+u_{\sigma_1}}\Big|x,q;\alpha+u_\alpha\right)}{\sqrt{\sigma_1+u_{\sigma_1}}}-\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'\beta_2}{\sigma_2}\Big|x,q;\alpha\right)}{\sqrt{\sigma_2}}\right|^2dy\cdot f_{x|q}dx\cdot f_qdq\\
&=\text{Term 1}+\text{Term 2}+\text{Term 3}.
\end{aligned}$$

We will analyze these three terms one by one. First define $u_\ell=\left(u'_{\beta_\ell},u_{\sigma_\ell},u'_\alpha\right)'$.

Term 1:

$$\begin{aligned}
&\int_{q\le\gamma}\iint\left|\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'(\beta_1+u_{\beta_1})}{\sigma_1+u_{\sigma_1}}\Big|x,q;\alpha+u_\alpha\right)}{\sqrt{\sigma_1+u_{\sigma_1}}}-\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'\beta_1}{\sigma_1}\Big|x,q;\alpha\right)}{\sqrt{\sigma_1}}\right|^2dy\cdot f_{x|q}dx\cdot f_qdq\\
&\overset{(1)}{=}\int_{q\le\gamma}\iint\left|\frac{f_{e|x,q}^{1/2}\left(\frac{\sigma_1e+x'\beta_1-x'(\beta_1+u_{\beta_1})}{\sigma_1+u_{\sigma_1}}\Big|x,q;\alpha+u_\alpha\right)}{\sqrt{\sigma_1+u_{\sigma_1}}}-\frac{f_{e|x,q}^{1/2}(e|x,q;\alpha)}{\sqrt{\sigma_1}}\right|^2\sigma_1de\cdot f_{x|q}dx\cdot f_qdq\\
&\overset{(2)}{=}\frac{1}{4}\int_0^1\int_{q\le\gamma}\iint u_1'\frac{\partial}{\partial\widetilde{\theta}_1}\ln f_{e|x,q}\left(e|x,q;\widetilde{\theta}_1,\underline{\theta}_1\right)\\
&\quad\times\frac{\partial}{\partial\widetilde{\theta}_1'}\ln f_{e|x,q}\left(e|x,q;\widetilde{\theta}_1,\underline{\theta}_1\right)|_{\widetilde{\theta}_1=\theta_1+\omega u_1}u_1
\end{aligned}$$

$$\times f_{e|x,q}\left(e|x,q;\widetilde{\theta}_1,\underline{\theta}_1\right)|_{\widetilde{\theta}_1=\theta_1+\omega u_1}\sigma_1 de\cdot f_{x|q}dx\cdot f_q dq\cdot d\omega$$
$$\stackrel{(3)}{=} O\left(\|u_1\|^2\right),$$

where (1) is obtained by changing variable $y = x'\beta_1 + \sigma_1 e$, (2) is from the mean value theorem and Fubini's theorem, and (3) is from Assumption D11.

Similarly, it can be shown that Term 2 $= O\left(\|u_2\|^2\right)$.

Term 3: Similar analysis as in Term 1 shows that

$$\int_\gamma^{\gamma+v}\iint\left|\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'\left(\beta_1+u_{\beta_1}\right)}{\sigma_1+u_{\sigma_1}}|x,q;\alpha+u_\alpha\right)}{\sqrt{\sigma_1+u_{\sigma_1}}}\right.$$
$$\left.-\frac{f_{e|x,q}^{1/2}\left(\frac{y-x'\beta_2}{\sigma_2}|x,q;\alpha\right)}{\sqrt{\sigma_2}}\right|^2 dy\cdot f_{x|q}dx\cdot f_q dq$$
$$=\frac{1}{4}\int_0^1\int_\gamma^{\gamma+v}\iint\left(\theta_1+u_1-\theta_2\right)'\frac{\partial}{\partial\widetilde{\theta}_1}\ln f_{e|x,q}$$
$$\times\left(e|x,q;\overline{\theta}_{12}(\omega),\underline{\theta}_1\right)\frac{\partial}{\partial\widetilde{\theta}_1'}\ln f_{e|x,q}\left(e|x,q;\overline{\theta}_{12}(\omega),\underline{\theta}_1\right)$$
$$\times\left(\theta_1+u_1-\theta_2\right)f_{e|x,q}\left(e|x,q;\overline{\theta}_{12}(\omega),\underline{\theta}_1\right)\sigma_1 de$$
$$\times f_{x|q}dx\cdot f_q dq\cdot d\omega,$$

which is $O(|v|)$ by Assumptions D12 and D2, where $\overline{\theta}_{12}(\omega) = \theta_2+\omega\left(\theta_1+u_1-\theta_2\right)$.

(ii) This lower bound can be established separately for $\|h\| > \delta$ and $\|h\|\le\delta$. For $\|h\| > \delta$, by Lemma 3,

$$r_2(\theta,\theta+h)^2\ge\epsilon_\delta > 0.$$

If it can be shown that for $\|h\|\le\delta$,

$$r_2(\theta,\theta+h)^2\ge C\left(\|u\|^2+|v|\right),$$

then there will always exist some constant such that this lower bound is satisfied. For $\|h\|\le\delta$, there are still three terms as in (i). Similar analysis as in (i) shows this result by Assumptions D2, D11 and D12. □

Lemma 8 (*Hölder Continuity and Exponential Tails*). *Under Assumptions D0–D2 and D9–D12, uniformly for $\theta\in\mathcal{N}_0$,*

(i) for any given $R > 0$, $E_\theta\left[\left|Z_{n,\theta}^{1/2}(h_2)-Z_{n,\theta}^{1/2}(h_1)\right|^2\right]\le C\|h_2-h_1\|(1+2R)$ for all $h_1,h_2\in H_n$, and $\|h_1\|\le R$, $\|h_2\|\le R$.

(ii) $E_\theta\left[Z_{n,\theta}^{1/2}(h)\right]\le\exp\left[-C\left(\|h\|-1\right)\right]$ for all $h\in H_n$.

Proof. This lemma is a corollary of Lemma 7. The proof is similar to that of Lemma B.2 in Chernozhukov and Hong (2004), so omitted here. □

Appendix D. Algorithms

This appendix presents the algorithms for calculating asymptotic distributions and risk of the γ estimators in the main text. To simplify the notation, Z is used to denote the respective weak limits of γ estimators in each subsection. For example, Z in the LMLE subsection represents $Z_{\gamma,\text{LMLE}}$, etc. The loss function is assumed to be separable in all calculations, and only loss functions $l_2(v)=|v|^r$ with $r = 1, 2, \ldots$ are considered. To meet space constraints, we will only detail the algorithms for the LMLE since they provide the basic insight of our analysis. For other estimators, we only provide the results, while the derivations are available upon request. A corollary of these algorithms is that asymptotic distributions of all γ estimators are continuous.

The LMLE.

For $t\le 0$,

$$P(Z\le t)=\sum_{k=0}^{\infty}P(Z\le t|\text{Max}=k)\,P(\text{Max}=k) \tag{11}$$

where Max is the number of jumps before reaching the maximum of $D(v)$ on $v\le 0$. Since $P(Z\le t|\text{Max}=k) = P\left(N_1(-t)\le k\right) = \sum_{j=0}^{k}\frac{e^{f_q(\gamma_0)t}\left(-f_q(\gamma_0)t\right)^j}{j!}$, it remains to calculate $P(\text{Max}=k)\equiv p_{1k}$. The event $E^{(k)}\equiv\{\text{Max}=k\}$ is equivalent to $\left\{\sum_{i=1}^{k}z_{1i}\ge\sum_{i=1}^{j}z_{1i}\text{ for } j\in\mathbb{Z}_+, \sum_{i=1}^{k}z_{1i}\ge\sum_{i=1}^{j}z_{2i}\text{ for } j\in\mathbb{N}\right\}$, where $\sum_{i=1}^{0}\cdot\equiv 0$, $\mathbb{Z}_+$ is the set of nonnegative integers, and $\mathbb{N}=\mathbb{Z}_+\backslash\{0\}$. The question is how to calculate the probability of this event.

Note that the event $E^{(k)}$ is the intersection of three events:

$$E_1^{(k)}=\left\{\sum_{i=j}^{k}z_{1i}\ge 0, j=1,\ldots,k\right\},$$
$$E_2^{(k)}=\left\{\sum_{i=k+1}^{j}z_{1i}\le 0, j=k+1,\ldots\right\},$$
$$E_3^{(k)}=\left\{\sum_{i=1}^{k}z_{1i}\ge\sum_{i=1}^{j}z_{2i}\text{ for } j\in\mathbb{N}\right\},$$

and the event $E_2^{(k)}$ is independent of $E_1^{(k)}\cap E_3^{(k)}$, so calculate the probability of $E_2^{(k)}$ first. Define this probability as $F_1(0)$, then

$$F_1(x)\equiv P\left\{\sum_{i=k+1}^{j}z_{1i}\le x, j=k+1,\ldots\right\}$$
$$=\int_{-\infty}^{x}\phi_1\left(z_{1,k+1}\right)$$
$$\times P\left(\sum_{i=k+2}^{j}z_{1i}\le x-z_{1,k+1}, j=k+2,\ldots\right)dz_{1,k+1}$$
$$=\int_{-\infty}^{x}\phi_1\left(z_{1,k+1}\right)F_1(x-z_{1,k+1})dz_{1,k+1}$$
$$=\int_0^{\infty}\phi_1(x-t)\,F_1(t)dt \tag{12}$$

where $\phi_1(\cdot)$ is the density function of z_{1i}. This is an integral equation called the homogeneous Wiener–Hopf equation of the second kind with boundary condition $F_1(-\infty)=0, F_1(\infty)=1$, where $F_1(\cdot)$ is the cdf of $\max\left\{\sum_{i=k+1}^{j}z_{1i}, j=k+1,\ldots\right\}$ and does not depend on k. For $E_1^{(k)}\cap E_3^{(k)}$, notice that

$$E_1^{(k)}\cap E_3^{(k)}=\left\{\sum_{i=j}^{k}z_{1i}\ge 0, j=2,\ldots,k,\right.$$
$$\left.\sum_{i=1}^{k}z_{1i}\ge\max\left\{0,\sum_{i=1}^{j}z_{2i}\text{ for all } j\in\mathbb{N}\right\}\right\},$$

where $\sum_{i=j}^{k}\cdot = 0$ if $k < j$. Suppose the random variable $\max\left\{\sum_{i=1}^{j} z_{2i} \text{ for all } j \in \mathbb{N}\right\}$ has the cdf $F_2(\cdot)$, then

$$\begin{aligned}
&P\left(E_1^{(k)}(x) \cap E_3^{(k)}(x)\right) \\
&\quad \equiv P\left(\sum_{i=j}^{k} z_{1i} \geq x, j = 2, \ldots, k, \sum_{i=1}^{k} z_{1i} \geq x \right.\\
&\qquad \left. + \max\left\{0, \sum_{i=1}^{j} z_{2i} \text{ for all } j \in \mathbb{N}\right\}\right) \\
&\quad = \int_x^\infty \int_{x-z_{1k}}^\infty \cdots \int_{x-\sum_{j=3}^{k} z_{1j}}^\infty \\
&\qquad \times \left[F_2(0) \int_{x-\sum_{j=2}^{k} z_{1j}}^\infty \phi_1(z_{11})\, dz_{11} \right.\\
&\qquad \left. + \int_0^\infty \int_{x-\sum_{j=2}^{k} z_{1j}+t}^\infty \phi_1(z_{11})\, dz_{11} dF_2(t) \right] \\
&\qquad \times \phi_1(z_{12}) \cdots \phi_1\left(z_{1,k-1}\right) \phi_1\left(z_{1,k}\right) dz_{12} \cdots dz_{1k-1} dz_{1k}.
\end{aligned}$$

While this formula is complicated, there is a recursive solution: for $x \leq 0$,

$$P\left(E_1^{(0)}(x) \cap E_3^{(0)}(x)\right) = F_2(-x) \tag{13}$$

and for $k \geq 0$ and $x \leq 0$,

$$\begin{aligned}
&P\left(E_1^{(k+1)}(x) \cap E_3^{(k+1)}(x)\right) \\
&\quad = \int_x^\infty P\left(E_1^{(k)}(x - z_{1,k+1}) \cap E_3^{(k)}(x - z_{1,k+1})\right) \\
&\qquad \times \phi_1\left(z_{1,k+1}\right) dz_{1,k+1} \\
&\quad = \int_{-\infty}^0 P\left(E_1^{(k)}(t) \cap E_3^{(k)}(t)\right) \phi_1(x-t)\, dt.
\end{aligned} \tag{14}$$

This is the left-side convolution of $P\left(E_1^{(k)}(t) \cap E_3^{(k)}(t)\right)$ and $\phi_1(x)$.

In summary, for $t \leq 0$,

$$\begin{aligned}
P(Z \leq t) &= \sum_{k=0}^{\infty} P(Z \leq t | \text{Max} = k)\, P(\text{Max} = k) \\
&= \sum_{k=0}^{\infty} p_{1k} \sum_{j=0}^{k} \frac{e^{f_q(\gamma_0)t}\left(-f_q(\gamma_0)t\right)^j}{j!} \\
&= \sum_{k=0}^{\infty} F_1(0) P\left(E_1^{(k)} \cap E_3^{(k)}\right) \sum_{j=0}^{k} \frac{e^{f_q(\gamma_0)t}\left(-f_q(\gamma_0)t\right)^j}{j!},
\end{aligned}$$

where $F_1(0)$ is the solution of (12) evaluated at 0, and $P\left(E_1^{(k)} \cap E_3^{(k)}\right)$ is recursively solved by (13) and (14).[14] p_{1k} is a decreasing function of k and converges to zero when k goes to infinity since z_{1i} has a negative mean. Note that t only appears in $P(Z \leq t|\text{Max} = k)$, while $P(\text{Max} = k)$ is determined by the distribution of z_{1i} and z_{2i}, and independent of t. $F_1(0)$ is only determined by the distribution of z_{1i}, while $P\left(E_1^{(k)} \cap E_3^{(k)}\right)$ depends on the distributions of both z_{1i} and z_{2i}.

For $t > 0$, the derivation is similar:

$$\begin{aligned}
P(0 < Z \leq t) &= \sum_{k=1}^{\infty} P(0 < Z \leq t | \text{Max} = k)\, P(\text{Max} = k) \\
&= \sum_{k=1}^{\infty} P(N_2(t) \geq k)\, P(\text{Max} = k) \\
&= \sum_{k=1}^{\infty} p_{2k} \sum_{j=k}^{\infty} \frac{e^{-f_q(\gamma_0)t}\left(f_q(\gamma_0)t\right)^j}{j!},
\end{aligned}$$

where Max is the number of jumps before reaching the maximum of $D(v)$ on $v > 0$. p_{2k} is the same as p_{1k} except that the subscript 1 is replaced by 2 in all expressions above, and the recursion of (13) and (14) starts from 1 instead of 0. It should be emphasized that p_{1k} and p_{2k} are determined by the whole distributions of z_{1i} and z_{2i} instead of their means. When $E[z_{1i}] > E[z_{2i}]$, p_{1k} is not necessarily greater than p_{2k}.

Note that $\{p_1; p_2\} \equiv \{p_{10}, p_{11}, \ldots; p_{21}, p_{22}, \ldots\}$ are the point masses of a discrete distribution which represents the probability that a jump reaches the maximum of $D(v)$. Furthermore, $P(Z \leq 0) = \sum_{k=0}^{\infty} p_{1k}$ because $\lim_{t\to 0} \sum_{j=0}^{k} \frac{e^{f_q(\gamma_0)t}(-f_q(\gamma_0)t)^j}{j!} = 1$ for any $k \in \mathbb{Z}_+$, and $P(0 < Z < \infty) = \sum_{k=1}^{\infty} p_{2k}$ because $\lim_{t\to\infty} \sum_{j=k}^{\infty} \frac{e^{-f_q(\gamma_0)t}(f_q(\gamma_0)t)^j}{j!} = 1 - \lim_{t\to\infty} \sum_{j=1}^{k-1} \frac{e^{-f_q(\gamma_0)t}(f_q(\gamma_0)t)^j}{j!} = 1$ for any $k \in \mathbb{N}$. This is not difficult to understand, since $P(Z \leq 0) = P(D(v)$ achieves maximum at $v \leq 0)$, and $P(0 < Z < \infty) = P(D(v)$ achieves maximum at $v > 0)$. So $P(Z \leq 0) + P(0 < Z < \infty) = 1$, and the derivation above was indeed calculating the cdf of Z. In summary, the cdf of Z is

$$F_Z(t) = \begin{cases} \sum_{k=0}^{\infty} p_{1k} \sum_{j=0}^{k} \frac{e^{f_q(\gamma_0)t}\left(-f_q(\gamma_0)t\right)^j}{j!}, & \text{if } t \leq 0; \\ \sum_{k=0}^{\infty} p_{1k} + \sum_{k=1}^{\infty} p_{2k} \sum_{j=k}^{\infty} \frac{e^{-f_q(\gamma_0)t}\left(f_q(\gamma_0)t\right)^j}{j!}, & \text{if } t > 0. \end{cases}$$

So the pdf of Z is

$$f_Z(t) = \begin{cases} f_q(\gamma_0) \sum_{k=0}^{\infty} \frac{e^{f_q(\gamma_0)t}\left(-f_q(\gamma_0)t\right)^k}{k!} p_{1,k} = f_q(\gamma_0) \\ \quad \cdot \left(\text{Poisson}(-f_q(\gamma_0)t) \circ p_1\right), \quad \text{if } t \leq 0; \\ f_q(\gamma_0) \sum_{k=0}^{\infty} \frac{e^{-f_q(\gamma_0)t}\left(f_q(\gamma_0)t\right)^k}{k!} p_{2,k+1} = f_q(\gamma_0) \\ \quad \cdot \left(\text{Poisson}(f_q(\gamma_0)t) \circ p_2\right), \quad \text{if } t > 0 \end{cases}$$

where $\circ$ means the inner product of two vectors in $\mathbb{R}^\infty$. $f_Z(t)$ is more concentrated when $f_q(\gamma_0)$ gets larger since more data are sampled in the neighborhood of γ_0, and the threshold point is easier to identify. Z has an exponential decay tail, and $\frac{f_Z(0)}{f_q(\gamma_0)} = p_{10} = F_1(0)F_2(0) \in (0, 1)$. When $p_{10} = 1$, $f_Z(\cdot)$ reduces to the negative exponential density in Section 2. For any $t > 0$, if $\{p_{1i}\}_{i=1}^{\infty} = \{p_{2i}\}_{i=1}^{\infty}$, then

$$\begin{aligned}
f_Z(-t) - f_Z(t) &= f_q(\gamma_0) \sum_{k=0}^{\infty} \frac{e^{-f_q(\gamma_0)t}\left(f_q(\gamma_0)t\right)^k}{k!} \left(p_{1,k} - p_{2,k+1}\right) \\
&= f_q(\gamma_0) \cdot \left(\text{Poisson}(f_q(\gamma_0)t) \circ (p_1 - p_2)\right) > 0,
\end{aligned}$$

so the density of Z on $v < 0$ is thicker than that on $v > 0$. This is because the left end point of the minimizing interval is taken as the estimate.

[14] To appreciate the correctness of the formula for p_{1k}, p_{10} is checked. Note that $p_{10} = P(Z_1 < 0, Z_2 \leq 0) = F_1(0)F_2(0)$, where $Z_1 \equiv \max\left\{\sum_{i=1}^{k} z_{1i},\ k = 1, 2, \ldots\right\}$ and $Z_2 = \max\left\{\sum_{i=1}^{k} z_{2i},\ k = 1, 2, \ldots\right\}$.

Although an elegant form of the density function of Z is developed, $\{p_1; p_2\}$ still must be calculated numerically, which involves solving an integral equation. Hinkley (1970) develops an algorithm to calculate this discrete probability distribution, but it seems only reliable in some special cases.[15] A different expression of $\{p_1; p_2\}$ is provided here, which can have theoretical value in the future research. In practice, when the distributions of z_{1i} and z_{2i} are known, the simulation method is suggested to find $\{p_1; p_2\}$.

It can be shown that

$$R^{(r)}_{\mathrm{LMLE}} \equiv E\left[|Z|^r\right] = \frac{1}{f_q(\gamma_0)^r}\sum_{k=0}^{\infty}\frac{(k+r)!}{k!}\left(p_{2,k+1}+p_{1,k}\right).$$

This risk inversely depends on $f_q(\gamma_0)$, which matches the density function form above.

The MMLE.

The cdf of Z is

$$F_Z(t) = \begin{cases} p_{10}\dfrac{e^{2f_q(\gamma_0)t}}{2} + \displaystyle\sum_{k=1}^{\infty} p_{1k}\left[2\sum_{i=0}^{k-1}(-1)^i \times \left(\sum_{j=0}^{k-i-1}\frac{\left(-f_q(\gamma_0)t\right)^j}{j!}\right)e^{f_q(\gamma_0)t} + (-1)^k e^{2f_q(\gamma_0)t}\right], & \text{if } t \le 0; \\ \displaystyle\sum_{k=1}^{\infty} p_{1k} + p_{10}\left(1-\frac{e^{-2f_q(\gamma_0)t}}{2}\right) + \sum_{k=1}^{\infty} p_{2k}\left[2\sum_{i=0}^{k-1}(-1)^i \times \left(\sum_{j=k-i}^{\infty}\frac{\left(f_q(\gamma_0)t\right)^j}{j!}\right)e^{-f_q(\gamma_0)t} + (-1)^k\left(1-e^{-2f_q(\gamma_0)t}\right)\right], & \text{if } t > 0; \end{cases}$$

and the pdf is

$$f_Z(t) = \begin{cases} p_{10}f_q(\gamma_0)e^{2f_q(\gamma_0)t} + 2f_q(\gamma_0)\displaystyle\sum_{k=1}^{\infty} p_{1,k} \times \left[\sum_{i=0}^{k-1}(-1)^i\frac{e^{f_q(\gamma_0)t}\left(-f_q(\gamma_0)t\right)^{k-1-i}}{(k-1-i)!} + (-1)^k e^{2f_q(\gamma_0)t}\right], & \text{if } t \le 0; \\ p_{10}f_q(\gamma_0)e^{-2f_q(\gamma_0)t} + 2f_q(\gamma_0)\displaystyle\sum_{k=1}^{\infty} p_{2k} \times \left[\sum_{i=0}^{k-1}(-1)^i\frac{e^{-f_q(\gamma_0)t}\left(f_q(\gamma_0)t\right)^{k-1-i}}{(k-1-i)!} + (-1)^k e^{-2f_q(\gamma_0)t}\right], & \text{if } t > 0. \end{cases}$$

When $p_{10} = 1, f_Z(\cdot)$ reduces to the double exponential density in Section 2. If $p_{1k} = p_{2k}$ for $k = 1, 2, \ldots$, then $f_Z(t)$ is symmetric.

It can be shown that

$$R^{(r)}_{\mathrm{MMLE}} \equiv E\left[|Z|^r\right] = \frac{1}{f_q(\gamma_0)^r}\left\{2\sum_{k=1}^{\infty}\left(p_{1k}+p_{2k}\right)\sum_{i=0}^{k-1}(-1)^i\frac{(k+r-i-1)!}{(k-i-1)!} + \frac{r!}{2^r}\left[\sum_{k=1}^{\infty}(-1)^k\left(p_{1k}+p_{2k}\right)+p_{10}\right]\right\}.$$

It is hard to get a general ordering between the risks of the LMLE and MMLE. The ordering depends on r and the allocation of $\{p_1; p_2\}$, but not on $f_q(\gamma_0)$. Note that

$$\begin{aligned} f_q(\gamma_0)^r&\left(R^{(r)}_{\mathrm{LMLE}} - R^{(r)}_{\mathrm{MMLE}}\right) \\ &= \sum_{k=0}^{\infty}\frac{(k+r)!}{k!}\left(p_{2,k+1}+p_{1,k}\right) - 2\sum_{k=1}^{\infty}\left(p_{1k}+p_{2k}\right)\sum_{i=0}^{k-1}(-1)^i\frac{(k+r-i-1)!}{(k-i-1)!} - \frac{r!}{2^r}\left[\sum_{k=1}^{\infty}(-1)^k\left(p_{1k}+p_{2k}\right)+p_{10}\right] \\ &= \left(1-\frac{1}{2^r}\right)r!\cdot p_{10} + \sum_{k=1}^{\infty}\left(\frac{(k+r)!}{k!} - \frac{r!}{2^r}(-1)^k - 2\sum_{i=0}^{k-1}(-1)^i\frac{(k+r-i-1)!}{(k-i-1)!}\right)p_{1,k} \\ &\quad + \sum_{k=1}^{\infty}\left(\frac{(k-1+r)!}{(k-1)!} - \frac{r!}{2^r}(-1)^k - 2\sum_{i=0}^{k-1}(-1)^i\frac{(k+r-i-1)!}{(k-i-1)!}\right)p_{2,k} \\ &\equiv \left(1-\frac{1}{2^r}\right)r!\cdot p_{10} + \sum_{k=1}^{\infty}A(k;r)p_{1,k} + \sum_{k=1}^{\infty}B(k;r)p_{2,k}, \end{aligned}$$

where $A(k; r) > 0$ and $B(k; r) < 0$ from the definition of the LMLE and MMLE. If $p_{1k} = p_{2k}$ for $k = 1, 2, \ldots$, then

$$A(k;r) + B(k;r) = \frac{(k+r)!}{k!} + \frac{(k-1+r)!}{(k-1)!} - \frac{r!}{2^{r-1}}(-1)^k - 4\sum_{i=0}^{k-1}(-1)^i\frac{(k+r-i-1)!}{(k-i-1)!} \ge 0$$

with equality being achieved when $r = 1$, so $R^{(r)}_{\mathrm{LMLE}} > R^{(r)}_{\mathrm{MMLE}}$ in this case. Otherwise, if p_{10} dominates $\{p_{1k}, p_{2k}\}_{k=1}^{\infty}$, or p_{2k} is not much larger than $p_{1,k}$, $R^{(r)}_{\mathrm{LMLE}}$ is still greater than $R^{(r)}_{\mathrm{MMLE}}$. The above conclusions can also apply to a convex and symmetric loss function, but not to bowl-shaped loss functions in general.

The posterior mean and quantile

The key insight in this section is that $D(v)$ can be approximated by its truncated version $D^{(k)}(v)$:

$$D^{(k)}(v) = \begin{cases} \exp\left\{\sum_{i=1}^{k} z_{1i}\right\}, & \text{if } -\sum_{j=0}^{k} T_{1j} \le v \le -\sum_{j=0}^{k-1} T_{1j}; \\ \vdots & \\ 1, & \text{if } -T_{10} \le v \le 0; \\ 1, & \text{if } 0 < v \le T_{20}; \\ \vdots & \\ \exp\left\{\sum_{i=1}^{k} z_{2i}\right\}, & \text{if } \sum_{j=0}^{k-1} T_{2j} \le v \le \sum_{j=0}^{k} T_{2j}; \\ 0, & \text{otherwise,} \end{cases}$$

15 See also Atkinson (1974) for a closed-form solution.

where $\{T_{\ell i}\}_{i=0}^{\infty}$ are interarrival times of $N_\ell(\cdot)$. When the posterior mean is considered,

$$P(Z \le t) = P\left(\int v \frac{\exp(D(v))}{\int \exp(D(\widetilde{v}))\, d\widetilde{v}} dv \le t\right) = P\left(\int (v - t)\exp(D(v))\, dv \le 0\right),$$

where the $D(v)$ in $\int v \exp(D(v))\, dv$ and $\int \exp(D(v))\, dv$ can be approximated by $D^{(k)}(v)$. In the case of the τ'th posterior quantile,

$$P(Z \le t) = P\left(\int_{-\infty}^{t} \frac{\exp(D(v))}{\int_{-\infty}^{\infty} \exp(D(\widetilde{v}))\, d\widetilde{v}} dv \ge \tau\right) = P\left(\frac{\int_{-\infty}^{t} \exp(D(v))\, dv}{\int_{t}^{\infty} \exp(D(v))\, dv} \ge \frac{\tau}{1-\tau}\right)$$

where the $D(v)$ in $\int_{-\infty}^{t} \exp(D(v))\, dv$ and $\int_{t}^{\infty} \exp(D(v))\, dv$ can be approximated by $D^{(k)}(v)$. In practice, a large number of sums are used for approximation until the algorithm is stable. Note that the algorithm will stabilize eventually, since $\exp(D(v))$ is exponentially decaying. As to the asymptotic risk, the Riemann–Stieltjes integral $\int |t|^r\, dP(Z \le t)$ is suggested, where $P(Z \le t)$ is calculated numerically as above.[16]

References

Aldous, D., 1978. Stopping times and tightness. The Annals of Probability 6, 335–340.
Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 821–856.
Atkinson, C., 1974. A Wiener–Hopf integral equation arising in some inference and queueing problems. Biometrika 61, 277–283.
Bai, J., 1995. Least absolute deviation estimation of a shift. Econometric Theory 11, 403–436.
Bai, J., 1997. Estimation of a change point in multiple regression models. Review of Economics and Statistics 79, 551–563.
Bajari, P., et al. 2010. Regression discontinuity designs with an endogenous forcing variable and an application to contracting in health care. Manuscript. University of Minnesota (Unpublished).
Beran, R., 1997. Diagnosing bootstrap success. Annals of the Institute of Statistical Mathematics 49, 1–24.
Boetel, B.L., et al., 2007. Estimating investment rigidity within a threshold regression framework: the case of US hog production sector. American Journal of Agricultural Economics 89, 36–51.
Chan, K.S., 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. The Annals of Statistics 21, 520–533.
Chan, N.H., Kutoyants, Y.A., 2010. On parameter estimation of threshold autoregressive models. Manuscript. Department of Statistics, Chinese University of Hong Kong (Unpublished).
Chernozhukov, V., Hong, H., 2004. Likelihood estimation and inference in a class of nonregular econometric models. Econometrica 1445–1480.
Durlauf, S.N., Johnson, P.A., 1995. Multiple regimes and cross-country growth behavior. Journal of Applied Econometrics 10, 365–384.
Geweke, J., Terui, N., 1993. Bayesian threshold autoregressive models for nonlinear time series. Journal of Time Series Analysis 14, 441–454.
Gijbels, I., et al., 1999. On the estimation of jump points in smooth curves. Annals of the Institute of Statistical Mathematics 51, 231–251.
Gonzalo, J., Wolf, M., 2005. Subsampling inference in threshold autoregressive models. Journal of Econometrics 127, 201–224.
Hansen, B.E., 2000. Sample splitting and threshold estimation. Econometrica 575–603.
Hinkley, D.V., 1970. Inference about the change-point in a sequence of random variables. Biometrika 57, 1–17.
Hirano, K., Porter, J.R., 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. Econometrica 71, 1307–1338.
Howell, T., 2007. Birth of the threshold time series model. Statistica Sinica 17, 8–14.
Huang, B.N., Yang, C.W., 2006. Demand for cigarettes revisited: an application of the threshold regression model. Agricultural Economics 34, 81–86.
Ibragimov, I., Has'minskii, R., 1981. Statistical Estimation: Asymptotic Theory. Springer-Verlag, New York.
Lee, S., Seo, M.H., 2008. Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. Journal of Econometrics 144, 492–499.
Neal, R.M., 2003. Slice sampling. The Annals of Statistics 31, 705–767.
Neveu, J., 1965. Mathematical Foundations of the Calculus of Probability. Holden-Day, San Francisco.
Newey, W.K., McFadden, D.L., 1994. Large sample estimation and hypothesis testing. In: Eagle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. Elsevier Science B.V., pp. 2113–2245 (Chapter 36).
Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057.
Politis, D., Romano, J., 1994. Large sample confidence regions based on subsamples under minimal assumptions. The Annals of Statistics 22, 2031–2050.
Politis, D., et al., 1999. Subsampling. Springer-Verlag, New York.
Pollard, D., 1984. Convergence of Stochastic Processes. Springer-Verlag, New York.
Potter, S.M., 1995. A nonlinear approach to US GNP. Journal of Applied Econometrics 2, 109–125.
Savvides, A., Stengos, T., 2000. Income inequality and economic development: evidence from the threshold regression model. Economics Letters 69, 207–212.
Trochim, W., 1984. Research Design for Program Evaluation: The Regression Discontinuity Approach. Sage Publications, Beverly Hills.
Van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, New York.
Van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–26.
Yu, P., 2008a. Adaptive estimation of the threshold point in threshold regression. Manuscript. Department of Economics, University of Auckland (Unpublished).
Yu, P., 2008b. The bootstrap in threshold regression. Manuscript. Department of Economics, University of Auckland (Unpublished).

[16] The posterior variance is not a consistent estimate of the asymptotic variance of either the posterior mean or median; e.g., it is easy to check this statement in the simple example in Section 2. This is different from the regular case, where the posterior variance is a consistent estimate of the asymptotic variance of both the posterior mean and median.