Journal of Econometrics 189 (2015) 83–100

Contents lists available at ScienceDirect

Journal of Econometrics

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

Adaptive estimation of the threshold point in threshold regression

Ping Yu

School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 17 June 2011
Received in revised form
4 June 2012
Accepted 11 September 2013
Available online 15 July 2015

JEL classification:
C11
C14
C21

Keywords:
Nonregular model
Threshold regression
Semiparametric efficiency
Adaptive estimation
Semiparametric empirical Bayes
Middle-point LSE
Nonparametric posterior interval
Curse of dimensionality
Additively separable loss function
Compound Poisson process

ABSTRACT

This paper studies semiparametric efficient estimation of the threshold point in threshold regression. The classical literature of semiparametric efficient estimation rests on the fact that the maximum likelihood estimator is efficient in any parametric submodel for a large class of loss functions. However, in threshold regression, the maximum likelihood estimator is not efficient, while the Bayes estimators are efficient and different loss functions induce different efficient estimators. For an additively separable loss function that separates the efficiency problem of the threshold point from that of other parameters, we show that the semiparametric and parametric efficiency risk bounds coincide. Then we design a semiparametric empirical Bayes estimator to achieve this bound. In consequence, the threshold point can be adaptively estimated even under conditional moment restrictions. We also provide a valid confidence interval called the nonparametric posterior interval for the threshold point. Simulation studies show that the semiparametric empirical Bayes approach is substantially better than existing methods. To illustrate our procedure in practice, we apply it to an economic growth model for detecting different growth patterns.

1. Introduction

The linear regression model has played a prominent role in econometric analysis. One important limitation of the linear regression model is that different groups of entities may have different behaviors in a specific economic problem. For example, Durlauf and Johnson (1995) show that rich countries and poor countries have different growth patterns. The question is how to separate these two groups of countries and estimate their respective growth paths. The threshold regression (TR) model introduced by Tong (1978, 1983) and Tong and Lim (1980) is designed to answer such a question; see Tong (1990, 2011) for a summary of the TR literature in statistics and Hansen (2011) in econometrics. The typical setup of TR models is as follows:

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

where q is the threshold variable used to split the sample with pdf $f_q(\cdot)$ and cdf $F_q(\cdot)$, γ is the unknown threshold point, $x \in \mathbb{R}^k$ includes characteristics, $\beta \equiv (\beta_1', \beta_2')' \in \mathbb{R}^{2k}$ and $\sigma \equiv (\sigma_1, \sigma_2)'$ are parameters in the mean and variance of the two groups. We assume that x does not include the intercept, discrete regressors or q; these cases can be easily adapted in the following discussion. We also set $E[e^2] = 1$ as a normalization of the error variance and allow for conditional heteroskedasticity. The usual conditional moment restriction is

$$E[e|x, q] = 0. \tag{2}$$

There are two asymptotic frameworks for statistical inferences 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 follows the discontinuous framework of Chan (1993) with i.i.d. data.

Both Chan (1993) and Hansen (2000) use the least squares estimator (LSE) to estimate γ, but the problem of semiparametric efficient estimation of γ has not been studied. The difficulty lies in the fact that parameters considered in most existing literature of semiparametric efficiency are regular, while γ is not a regular parameter. For a regular parameter, the maximum likelihood

E-mail address: pingyu@hku.hk.

http://dx.doi.org/10.1016/j.jeconom.2013.09.002
0304-4076/

estimator (MLE) is asymptotically normal and efficient in any parametric submodel for a large class of loss functions. As a result, the efficiency of an estimator is indicated by its asymptotic variance. This is the starting point of finding semiparametric efficiency bounds. As the semiparametric problem is not easier than any parametric subproblem, the semiparametric efficiency bound is defined as the supremum of asymptotic variances of the MLEs among all submodels. In threshold regression, Yu (2012), inspired by the literature on boundary estimation such as Hirano and Porter (2003) and Chernozhukov and Hong (2004), shows that the MLE is not efficient for γ, while the Bayes estimators are efficient. Furthermore, different loss functions induce different efficient estimators. This makes the techniques for finding semiparametric efficient estimators in the existing literature not applicable.

In this paper, we solve the semiparametric efficiency problem of γ in two steps. First, in Section 2, we separate the efficiency problem of γ from that of other regular parameters by using an additively separable loss function. Given any such loss function, we show that the risk of the Bayes estimator of γ in any parametric submodel is the same as that when the true conditional density $f_{e|x,q}$ is known. Therefore, the semiparametric efficiency risk bound of γ for a given loss function is the risk in the parametric model, and the conditional moment restriction (2) does not lose any information from the completely known $f_{e|x,q}$ case. Second, in Section 3, we use a semiparametric empirical Bayes (SEB) approach to find an estimator of γ that achieves the efficiency risk bound. The SEB estimator (SEBE) is adaptive in the sense that the risk in the parametric case can be reached even if $f_{e|x,q}$ is not exactly known. It should be pointed out that all Bayes procedures in this paper are evaluated by classical efficiency criteria; in other words, the randomness is confined to the data and does not include parameters.

Although the SEBE has the same asymptotic risk as the parametric Bayes estimator, it is less susceptible to misspecification because no parametric specification is needed for the distribution of e. γ can be identified as in the correctly specified parametric model as long as (2) is imposed. Also, the SEBE avoids the Diaconis and Freedman (1986a,b)'s inconsistency problem by estimating the nuisance density $f_{e|x,q}$ rather than imposing a Dirichlet prior on it. The literature discussing how to avoid the Diaconis and Freedman's problem in the Bayesian framework all concentrates on regular models.

A corollary of the SEB method is to provide a valid confidence interval (CI) for γ. The CI construction for γ is unsolved in Chan (1993) and reconsidered in Hansen (2000). In Section 4, we discuss the difficulties in the previous papers and propose an alternative valid CI for γ—the nonparametric posterior interval (NPI). Section 5 includes some simplification and extension of the SEB approach to increase its applicability and to improve its finite-sample performance. The simulation results in Section 6 show that the SEBE has a lower risk and the NPI has better coverage and length properties than the existing methods. Section 7 applies the SEB method to an economic growth model and Section 8 concludes. All regularity conditions, proofs and tables in simulations and the application are given in Appendices A–C, respectively. Certain technical materials of the paper are collected in supplementary materials (see Appendix D).[1] Notations: the Euclidean norm of a vector $x \in \mathbb{R}^k$ is denoted as $\|x\|$, and C or C with a subscript is used as a generic positive constant, which need not be the same in each occurrence.

2. Semiparametric efficiency risk bound

We first recall the parametric results of Yu (2012) in Section 2.1. We then show that the semiparametric bound is the same as the parametric bound using a simple example and provide some intuition for this adaptive result in Section 2.2. At the end of Section 2.2, we also discuss a technical assumption on the loss function in the semiparametric efficiency risk bound derivation.

2.1. Parametric efficient estimation

The main results of Yu (2012) are that the Bayes estimator (BE) is more efficient than the MLE for estimating γ, the threshold point. Suppose $f_{e|x,q}$ is known as $f_{e|x,q}(e|x, q; \alpha)$, where $\alpha \in \mathbb{R}^{d_\alpha}$ is some nuisance parameter affecting the shape of the error distribution. Assume further that the loss function is additively separable on regular parameters and the nonregular parameter γ; that is, $l(\theta) = l(\underline{\theta}, \gamma) = l_1(\underline{\theta}) + l_2(\gamma)$, where $\theta = (\underline{\theta}', \gamma)'$, $\underline{\theta} = (\beta', \sigma', \alpha')'$, and l_1 is bowl-shaped.[2] This assumption is important for the semiparametric efficient estimation of γ, as it separates the efficiency problem of γ from that of the regular parameters. Such an assumption is motivated by the sequential estimation of γ and $\underline{\theta}$. Usually, a profiled procedure is used to estimate γ first, and then estimate $\underline{\theta}$ as if γ were known; see, e.g., Hansen (2000) and Yu (2012). It is reasonable to impose a loss function on each of these two steps without interactions.

Under regularity conditions specified in Yu (2012), the BE $\left(\widehat{\underline{\theta}}_{BE}', \widehat{\gamma}_{BE}\right)'$ based on l is most efficient in the locally asymptotically minimax (LAM) sense, and the asymptotic distribution is

$$\begin{aligned} &\sqrt{n}\left(\widehat{\underline{\theta}}_{BE} - \underline{\theta}_0\right) \xrightarrow{d} Z_{\underline{\theta}} \sim N\left(0, \mathcal{I}_{\underline{\theta}_0}^{-1}\right), \\ &n\left(\widehat{\gamma}_{BE} - \gamma_0\right) \xrightarrow{d} Z_\gamma = \arg\min_t \int_{\mathbb{R}} l_2(t - v)\, p_2^*(v) dv, \end{aligned} \tag{3}$$

where $\mathcal{I}_{\underline{\theta}_0}$ is the information matrix of $\underline{\theta}$, $p_2^*(v) = \frac{\exp\{D(v)\}}{\int_{\mathbb{R}} \exp\{D(\overline{v})\} d\overline{v}}$ is the normalized asymptotic posterior of γ, and these two asymptotic distributions are independent. Note that $\widehat{\underline{\theta}}_{BE}$ has the same asymptotic distribution as the MLE. The $D(v)$ in $p_2^*(v)$ is a compound Poisson process defined as

$$D(v) = \begin{cases} \sum_{i=1}^{N_1(|v|)} z_{1i}, & \text{if } v \le 0; \\ \sum_{i=1}^{N_2(v)} z_{2i}, & \text{if } v > 0; \end{cases} \tag{4}$$

which is cadlag with $D(0) = 0$, where all z_{1i}, z_{2i}, $i = 1, 2, \ldots, N_1(\cdot)$ and $N_2(\cdot)$ are mutually independent of each other, $N_\ell(\cdot), \ell = 1, 2$, is a Poisson process with intensity $f_q(\gamma_0)$, z_{1i} follows the limiting conditional distribution of

$$\overline{z}_{1i} \equiv \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f_{e|x,q}\left(\frac{\sigma_{10} e_i + x_i'(\beta_{10} - \beta_{20})}{\sigma_{20}} \middle| x_i, q_i; \alpha_0\right)}{f_{e|x,q}(e_i|x_i, q_i; \alpha_0)}$$

given $\gamma_0 + \Delta < q_i \le \gamma_0$, $\Delta < 0$ with $\Delta \uparrow 0$ and is denoted as $\overline{z}_{1i}|(q_i = \gamma_0-)$, and z_{2i} follows the limiting conditional distribution of

$$\overline{z}_{2i} \equiv \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f_{e|x,q}\left(\frac{\sigma_{20} e_i - x_i'(\beta_{10} - \beta_{20})}{\sigma_{10}} \middle| x_i, q_i; \alpha_0\right)}{f_{e|x,q}(e_i|x_i, q_i; \alpha_0)}$$

[1] The code for simulations and application and the supplementary materials are available at http://homes.eco.auckland.ac.nz/pyu013/research.html.

[2] A function is defined to be bowl-shaped if the sublevel sets $\{x : l(x) \le C\}$ are convex and symmetric about the origin.

given $\gamma_0 < q_i \leq \gamma_0 + \Delta$, $\Delta > 0$ with $\Delta \downarrow 0$ and is denoted as $\bar{z}_{2i}|\ (q_i = \gamma_0+)$. Because γ is essentially a boundary of q, only the local information in the neighborhood of γ_0 is relevant to the estimation of γ_0. For example, z_{1i} (z_{2i}) follows a distribution conditional on q in the neighborhood of γ_0, and the intensity of $N_1(\cdot)$ and $N_2(\cdot)$ is $f_q(\gamma_0)$.

2.2. Semiparametric efficiency risk bound of γ

Now, we explore the semiparametric efficiency risk bound for γ using a simple threshold regression model. The arguments can be easily extended to the general case. Suppose the semiparametric threshold regression model is

$$y = \beta_1 \mathbf{1}(q \leq \gamma) + e, \quad \text{where } E[e|q] = 0. \tag{5}$$

That is, $x = 1$, $\beta_{20} = 0$ and $\sigma_{10} = \sigma_{20} = 1$ are known in (1). In this simple model, there is a threshold γ in q such that the mean of y is β_1 when $q \leq \gamma$ and 0 when $q > \gamma$, and there is no threshold effect in variance.

For any one-dimensional submodel $f_{e|q}(e|q;\alpha)$ with $f_{e|q}(e|q;\alpha_0) = f_{e|q}(e|q)$ and the loss function $l(\beta_1, \alpha, \gamma) = l_1(\beta_1, \alpha) + l_2(\gamma)$ with $l_1(\cdot)$ bowl-shaped, the asymptotic distribution of the BE for $(\beta_1, \alpha, \gamma)$ is

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_1 - \beta_{10} \\ \widehat{\alpha} - \alpha_0\end{pmatrix} \xrightarrow{d} N\left(0, \mathcal{I}_{\underline{\theta}_0}^{-1}\right)$$

$$n(\widehat{\gamma}_{BE} - \gamma_0) \xrightarrow{d} \arg\min_t \int_{\mathbb{R}} l_2(t-v)\left(\frac{\exp(D(v))}{\int_{\mathbb{R}}\exp(D(\bar{v}))d\bar{v}}\right)dv$$

where

$$\mathcal{I}_{\underline{\theta}_0} = E\begin{bmatrix} I_{e|q}\mathbf{1}(q \leq \gamma_0) & -E[S_e S_\alpha|q]\mathbf{1}(q \leq \gamma_0) \\ -E[S_e S_\alpha|q]\mathbf{1}(q \leq \gamma_0) & E\left[S_\alpha^2|q\right] \end{bmatrix}$$

with $I_{e|q} = E\left[S_e^2|q\right]$, $S_e = \frac{\partial \ln f_{e|q}(e|q)}{\partial e}$, $S_\alpha = \frac{\partial \ln f_{e|q}(e|q;\alpha_0)}{\partial \alpha}$, and $D(v)$ is defined in (4). It is clear that the nuisance parameter α affects the efficiency of the regular parameter β_1 due to the correlation between S_e and S_α (see Newey (1990) for more discussions on how α affects the efficiency of β_1), but does not affect the efficiency of the nonregular parameter γ. This result is true for any nuisance parameter α, so the risk upper bound for γ is the same as the risk when the conditional density function $f_{e|q}(e|q)$ is completely known. In other words, the semiparametric efficiency risk bound for γ is the same as the parametric bound.

This adaptiveness of the semiparametric efficiency risk bound for γ is due to the fact that γ uses different information in the data from that used in the regular parameter β_1. First, note that the regular parameter β_1 uses global information or information related to moments of the data. The optimal asymptotic variance for β_1 under $E[e|q] = 0$ is $E\left[\mathbf{1}(q \leq \gamma_0)/\sigma^2(q)\right]^{-1}$ (see, e.g., Chamberlain (1987)), which is achieved by the feasible generalized least squares estimator (GLSE) of Robinson (1987), where $\sigma^2(q) = E[e^2|q]$ is the conditional variance. Only moments of e and q (e.g., $\sigma^2(q), q \leq \gamma_0$) are relevant to this efficiency bound, which implies that each data point has the same importance to the estimation of β_1. In contrast, the data points in the neighborhood of γ_0 are more important to the estimation of γ: a data point (y_i, q_i) with q_i close to γ_0 provides more information than a data point with q_i far from γ_0. This can be seen from the form of $D(v)$; see also Section 2 of Yu (2012) for an illustration of this point. From Yu (2012), γ is essentially a "middle" boundary of q. Lemma 21.19 of Van der Vaart (1998) shows that the asymptotic distribution of a boundary estimator is independent of that of a regular parameter; that is, the information used to estimate a boundary and the information used to estimate a regular parameter are independent. Based on this lemma, the above argument is not surprising. Actually, Yu (2013a) shows that even under the nonparametric setup of $E[y|x, q]$ in the two regimes, γ can still be adaptively estimated.

The above discussion assumes that l is additively separable. When l is not additively separable, the efficiency problem for γ is intractable. Consider a standard non-separable loss function $l(\underline{\theta}, \gamma) = \left(|\underline{\theta}| + |\gamma|\right)^2$ with $\dim(\underline{\theta}) = 1$, then from Yu (2012),

$$\begin{pmatrix}\sqrt{n}\left(\widehat{\underline{\theta}}_{BE} - \underline{\theta}_0\right) \\ n\left(\widehat{\gamma}_{BE} - \gamma_0\right)\end{pmatrix} \xrightarrow{d} \arg\min_{s,t}\int_{\mathbb{R}^2} l(s-u, t-v)\, p_1^*(u)p_2^*(v)dudv,$$

where

$$p_1^*(u) = \sqrt{\frac{\left|\mathcal{I}_{\underline{\theta}_0}\right|}{2\pi}}\exp\left\{-\frac{\mathcal{I}_{\underline{\theta}_0}}{2}\left(u - Z_{\underline{\theta}}\right)^2\right\}$$

is the normalized asymptotic posterior of $\underline{\theta}$, and $Z_{\underline{\theta}}$ is defined in (3). The first-order conditions for this minimization problem are

$$\int_{\mathbb{R}}(s-u)\, p_1^*(u)du + \int_{\mathbb{R}}\text{sign}(s-u)\, p_1^*(u)du\int_{\mathbb{R}}|t-v|\, p_2^*(v)dv = 0,$$

$$\int_{\mathbb{R}}(t-v)\, p_2^*(v)du + \int_{\mathbb{R}}\text{sign}(t-v)\, p_2^*(v)dv\int_{\mathbb{R}}|s-u|\, p_1^*(u)du = 0,$$

and the solutions are denoted as $\left(\widehat{s}, \widehat{t}\right)$. Obviously, $\widehat{s} = \int_{\mathbb{R}} up_1^*(u)du$, which is the posterior mean. This is because the posterior mean and posterior median are the same for a normal posterior so $\int_{\mathbb{R}}\text{sign}(\widehat{s}-u)\, p_1^*(u)du = 0$. But $\widehat{t} \neq \int_{\mathbb{R}} vp_2^*(v)du$, as the posterior mean and posterior median are not the same for $p_2^*(v)$ which is not normal and contains infinite-dimensional sufficient statistics, so $\int_{\mathbb{R}}\text{sign}\left(\widehat{t}-v\right)p_2^*(v)dv \neq 0$.[3] In consequence, $\widehat{s}$ affects $\widehat{t}$ in a complicated way and that is intractable in our analysis. If $p_2^*(v)$ is also normal, then $\widehat{t}$ is the posterior mean of $p_2^*(v)$ for any bowl-shaped loss; see Lemma 8.5 (Anderson's Lemma) in Van der Vaart (1998) for a rigorous statement. This is essentially why we allow for interaction between elements of $\underline{\theta}$ in the loss function, but not between $\underline{\theta}$ and γ.

In summary, the semiparametric efficiency problem for the nonregular parameter γ is very different from that for regular parameters as discussed in Newey (1990) and Bickel et al. (1998). For any separable loss function, the conditional moment conditions do not lose any information from the parametric model for γ. After the semiparametric efficiency risk bound is found, we develop the semiparametric empirical Bayes estimator that can achieve this bound in the next section.

3. Feasible efficient estimation

In this section, we propose an estimator of γ that is adaptive to the unknown $f_{e|x,q}$. It begins with the construction of the estimator, followed by its asymptotic distribution, and concludes with a discussion about why orthogonality moment conditions are not enough to identify γ. The key result in this section is Theorem 1, that the nonparametric posterior converges to the true posterior in the total variation of moments norm. Based on this result, we prove that our estimator of γ is adaptive and is semiparametric efficient in a suitable sense. As to the efficient estimation of β, we relegate the discussion to supplementary materials as such results are standard in the literature (see Appendix D).

[3] There is some loss function with cross terms such that the first-order condition with respect to t is separable in s and t; for example $l(\underline{\theta}, \gamma) = \underline{\theta}^2 + \gamma^2 + |\underline{\theta}|\,\gamma$. But usually, the loss function should be a function of $|\gamma|$ rather than γ, so we do not pursue such general losses further.

3.1. Feasible efficient estimation of γ

Suppose there is an n-consistent estimator of γ in (1); for example, the LSE of Chan (1993) can serve this purpose. We then refine this estimator in its $1/n$ neighborhood to obtain an efficient estimator. Such a two-step estimator is also popular in regular models, such as the GLSE of Robinson (1987). Since $f_{e|x,q}$ is unknown, we will use a nonparametric method to estimate it, and then plug this nonparametric estimator in the Bayes formula to estimate γ. Such an estimator is called the semiparametric empirical Bayes estimator (SEBE).[4] Note that we only need to estimate the joint density $f(e, x, q)$ instead of $f_{e|x,q}$ since there is no parameter of interest in $f(x, q)$. To simplify notations, define $\mathsf{w} = \left(e, x', q\right)'$.

In this section, we consider the general case of unrestricted $f(\mathsf{w})$. In Section 5.1, we will consider a simpler setting where e is independent of $(x', q)'$. In the general case, the following algorithm is used to estimate γ.

Algorithm G. Step G1: Get an n-consistent estimator of γ and a $\sqrt{n}$-consistent estimator of $\left(\beta', \sigma'\right)'$, denoted as $\left(\widetilde{\beta}', \widetilde{\sigma}', \widetilde{\gamma}\right)'$. The corresponding residuals are denoted as $\{\widetilde{e}_i\}_{i=1}^n$.

Step G2: Estimate the joint density of w by kernel smoothing,

$$\widehat{f}(\mathsf{w}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\widetilde{\mathsf{w}}_i - \mathsf{w})$$

with $\widetilde{\mathsf{w}}_i = \left(\widetilde{e}_i, x_i', q_i\right)'$ and $K_h(\cdot) = \frac{1}{h^{k+2}} K\left(\frac{\cdot}{h}\right)$, where h is the bandwidth, and $K(\cdot) : \mathbb{R}^{k+2} \to \mathbb{R}$ is a kernel density function.

Step G3: Define the SEBE as

$$\widehat{\gamma} = \arg\min_t \int_\Gamma l_{2n}(t-\gamma)\widehat{\mathcal{L}}_n(\gamma)\pi_2(\gamma)\, d\gamma \tag{6}$$

where $l_{2n}(t-\gamma) = l_2(n(t-\gamma))$ is the loss function of γ, $\pi_2(\gamma)$ is the prior of γ, and

$$\begin{aligned}\widehat{\mathcal{L}}_n(\gamma) &= \prod_{i=1}^{n}\left[\frac{1}{\widetilde{\sigma}_1}\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}, x_i, q_i\right)\mathbf{1}(q_i \le \gamma)\right.\\ &\quad\left. + \frac{1}{\widetilde{\sigma}_2}\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}, x_i, q_i\right)\mathbf{1}(q_i > \gamma)\right]\\ &= \exp\left\{\sum_{i=1}^{n}\mathbf{1}(q_i \le \gamma)\ln\left(\frac{1}{\widetilde{\sigma}_1}\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}, x_i, q_i\right)\right)\right.\\ &\quad\left. + \sum_{i=1}^{n}\mathbf{1}(q_i > \gamma)\ln\left(\frac{1}{\widetilde{\sigma}_2}\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}, x_i, q_i\right)\right)\right\},\end{aligned}$$

denoted as $\exp\left\{\widehat{L}_n(\gamma)\right\}$, is the estimated likelihood function.

Based on Algorithm G, $\widehat{\gamma}$ is a Bayes estimator with the posterior distribution

$$\widehat{p}_n(\gamma) = \frac{\exp\left\{\widehat{L}_n(\gamma)\right\}\pi_2(\gamma)}{\int_\Gamma \exp\left\{\widehat{L}_n(\overline{\gamma})\right\}\pi_2(\overline{\gamma})\, d\overline{\gamma}};$$

that is,

$$\widehat{\gamma} = \arg\min_t \int_\Gamma l_{2n}(t-\gamma)\widehat{p}_n(\gamma)d\gamma.$$

As will be shown in the next subsection, the nonparametric estimation of the density function does not affect the efficiency of $\widehat{\gamma}$; in other words, $\widehat{\gamma}$ is adaptive to the estimation of $f(\mathsf{w})$.

To implement Step G3, $\pi_2(\gamma)$ must be specified. Following Yu (2012), we use the uniform distribution on $(q_{\min}, q_{\max})$ as the prior, where $q_{\min}$ ($q_{\max}$) is the minimum (maximum) of $\{q_i\}_{i=1}^n$. For a sampling scheme, the Metropolis–Hastings sampler, slice sampler or other MCMC methods can be used. Specifically, we first draw a Markov chain

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

whose marginal density is approximately $\widehat{p}_n(\gamma)$. Then the estimator $\widehat{\gamma}$ is approximated by a function of S depending on the loss function l_2. Popular loss functions and the corresponding $\widehat{\gamma}$ include: (a) $l_2(v) = v^2$, then $\widehat{\gamma}$ can be approximated by the mean of S; (b) $l_2(v) = |v|$, then $\widehat{\gamma}$ can be approximated by the median of S; and (c) $l_2(-v) = (\tau - \mathbf{1}(v \le 0))v$ for $\tau \in (0, 1)$, then $\widehat{\gamma}$ can be approximated by the τth percentile of S.

The estimation procedure above can also be applied to more general setups of the model. For example, the model with known but nonlinear regression functions can be estimated using this procedure. It can also be applied when there are some parameters common to the two regimes. The setup considering both cases is

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

where δ is the vector of parameters that remain the same in the two regimes, g_1 and g_2 are smooth functions which are not necessarily the same, and e_1 and e_2 need not take the form $\sigma_1 e$ and $\sigma_2 e$.

3.2. Asymptotic theory of $\widehat{\gamma}$

To obtain the asymptotic distribution of $\widehat{\gamma}$, we need regularity conditions on the loss function, prior, kernel, bandwidth and $f(\mathsf{w})$. These conditions are summarized in Appendix A. Here, we only specify the data and parameters, and collect notations. Throughout this paper, we maintain the assumptions that the data are i.i.d. and that the mean independence assumption $E[e|x, q] = 0$ holds. Define $w = \left(y, x', q\right)'$, $w_i = \left(y_i, x_i', q_i\right)'$.

Assumption 0. $w_i \in \mathbb{W} \equiv \mathbb{R} \times \mathbb{X} \times \mathbb{Q} \subset \mathbb{R}^{k+2}$, $\beta_1 \in B_1 \subset \mathbb{R}^k$, $\beta_2 \in B_2 \subset \mathbb{R}^k$, $0 < \sigma_1 \in \Omega_1 \subset \mathbb{R}$, $0 < \sigma_2 \in \Omega_2 \subset \mathbb{R}$, $\gamma \in \Gamma = [\underline{\gamma}, \overline{\gamma}] \subset \mathbb{R}$, $\Theta = B_1 \times B_2 \times \Omega_1 \times \Omega_2 \times \Gamma$ is compact, $\theta_0 \in \Theta_0$, where Θ_0 is in the interior of Θ, $\left(\beta_{10}', \sigma_{10}\right)' \ne \left(\beta_{20}', \sigma_{20}\right)'$, where $\ne$ means all corresponding coordinates of two vectors are not the same.

We introduce the following notations:

$$\begin{aligned}&\beta = \left(\beta_1', \beta_2'\right)', \quad \sigma = (\sigma_1, \sigma_2)', \quad \underline{\theta} = \left(\beta', \sigma'\right),\\ &\theta = \left(\underline{\theta}', \gamma\right)',\\ &\beta_0 = \left(\beta_{10}', \beta_{20}'\right)', \quad \sigma_0 = (\sigma_{10}, \sigma_{20})', \quad \underline{\theta}_0 = \left(\beta_0', \sigma_0'\right),\\ &\theta_0 = \left(\underline{\theta}_0', \gamma_0\right)\\ &B = B_1 \times B_2, \quad \Omega = \Omega_1 \times \Omega_2, \quad \underline{\Theta} = B \times \Omega\\ &u = \sqrt{n}\left(\underline{\theta} - \underline{\theta}_0\right), \quad v = n(\gamma - \gamma_0)\\ &U_n = \sqrt{n}\left(\underline{\Theta} - \underline{\theta}_0\right), \quad V_n = n(\Gamma - \gamma_0), \quad H_n = U_n \times V_n,\end{aligned}$$

where $\left(u', v\right)$ is the local parameter for θ, and H_n is the local parameter space which converges to $\mathbb{R}^{2k+3}$ by Assumption 0.

[4] Empirical Bayes methods obtain a prior from the data. See Section 4.5 of Berger (1985) for an introduction. Our method is labeled empirical Bayes due to replacing an unknown density by its estimate. Despite the similarity in terminology, note that the SEBE proposed here is distinct from nonparametric Bayes and semiparametric Bayes methods in the literature. See Hirano (2002) and Ghosal (2010) for a summary of the literature on this topic in statistics and econometrics, respectively.

Before studying the asymptotic distribution of $\widehat{\gamma}$, we first have a close look at the nonparametric posterior distribution which is the basic building block of $\widehat{\gamma}$. Our conclusion is that $\widehat{p}_n(\gamma)$ concentrates around γ_0 at rate $1/n$ as measured by the total variation of moments norm used in Chernozhukov and Hong (2003). Define the localized posterior as

$$\widehat{p}_n^*(v) = \frac{1}{n}\widehat{p}_n\left(\gamma_0 + \frac{v}{n}\right),$$

and the total variation of moments norm for $\widehat{p}_n^*(v)$ as

$$\left\|\widehat{p}_n^*\right\|_{TVM(\kappa)} = \int_{V_n} (1 + |v|^\kappa)\left|\widehat{p}_n^*(v)\right| dv.$$

Theorem 1 (*Adaptiveness of $\widehat{p}_n^*$*). *Under Assumptions 0, B, D, E, K and P in Appendix A, for any $0 \leq \kappa < \infty$,*

$$\left\|\widehat{p}_n^* - p_n^*\right\|_{TVM(\kappa)} \equiv \int_{V_n} (1 + |v|^\kappa)\left|\widehat{p}_n^*(v) - p_n^*(v)\right| dv \xrightarrow{p} 0, \quad (8)$$

where

$$p_n^*(v) = \frac{1}{n}p_n\left(\gamma_0 + \frac{v}{n}\right), \qquad p_n(\gamma) = \frac{\exp\{L_n(\gamma)\}\pi_2(\gamma)}{\int_\Gamma \exp\{L(\overline{\gamma})\}\pi_2(\overline{\gamma})\, d\overline{\gamma}},$$

with

$$L_n(\gamma) = \sum_{i=1}^n \mathbf{1}(q_i \leq \gamma)\ln\left(\frac{1}{\sigma_{10}} f\left(\frac{y_i - x_i\beta_{10}}{\sigma_{10}}, x_i, q_i\right)\right) + \sum_{i=1}^n \mathbf{1}(q_i > \gamma)\ln\left(\frac{1}{\sigma_{20}} f\left(\frac{y_i - x_i\beta_{20}}{\sigma_{20}}, x_i, q_i\right)\right)$$

being the log likelihood function with $f(\mathrm{w})$ and $\underline{\theta}_0$ known.

By Theorem 1, the sampling error introduced by the nonparametric estimation of $f(\mathrm{w})$ in $\widehat{p}_n(\gamma)$ can be neglected asymptotically. This is the basis for all remaining inferences about γ. Because $n(\widehat{\gamma} - \gamma_0) = \arg\min_t \int_{V_n} l_2(t - v)\widehat{p}_n^*(v)dv$, the argmax continuous mapping theorem implies the following theorem which can be viewed as a corollary of Theorem 1.

Theorem 2 (*Large Sample Inference of $\widehat{\gamma}$*). *Under Assumptions 0, B, D, E, K, L and P in Appendix A,*

$$n(\widehat{\gamma} - \gamma_0) - Z_n \xrightarrow{p} 0,$$

where

$$Z_n = \arg\min_t \int_{\mathbb{R}} l_2(t - v)p_n^*(v)dv.$$

This implies that

$$n(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \arg\min_t \int_{\mathbb{R}} l_2(t - v)\, p_2^*(v)dv \equiv Z,$$

where $p_2^*(v)$ is the same as that in (3) except that $f_{e|x,q}(\cdot|x_i, q_i, \alpha_0)$ in $D(v)$ is replaced by $f_{e|x,q}(\cdot|x_i, q_i)$, the true conditional density.

By Theorem 2, the asymptotic distribution of $\widehat{\gamma}$ is the same as that in the parametric case where $f(\mathrm{w})$ is known, so γ_0 can be adaptively estimated. The following theorem gives a precise sense in which $\widehat{\gamma}$ is semiparametric efficient.

Theorem 3 (*Semiparametric Efficiency of $\widehat{\gamma}$*). *Under Assumptions 0, B, D, E, K, L and P in Appendix A, if $\widetilde{\gamma}$ is uniformly n-consistent, that is, $\sup_{|\gamma-\gamma_0|<\delta} n|\widetilde{\gamma} - \gamma| = O_p(1)$ for some $\delta > 0$, then $\widehat{\gamma}$ is semiparametric efficient in the pointwisely locally asymptotically minimax (PLAM) sense:*

$$\lim_{\delta\to 0}\varliminf_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma\left[l_2\left(n\left(\widehat{\gamma} - \gamma\right)\right)\right]\right] \geq \lim_{\delta\to 0}\varliminf_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma\left[l_2\left(n\left(T_n - \gamma\right)\right)\right]\right], \quad (9)$$

where T_n is any estimator of γ, and $E_\gamma[\cdot]$ is the expectation under $\left(\beta_0', \sigma_0', \gamma, f_0(\mathrm{w})\right)'$ with $f_0(\mathrm{w})$ being the true density of w.

The LSE of γ satisfies the uniform n-consistency condition in Theorem 3 when more stringent regularity conditions are imposed.[5] The "pointwise" in the PLAM criterion refers to "pointwise in global parameters $\left(\beta_0', \sigma_0'\right)'$ and the true density $f_0(\mathrm{w})$". Actually, the supremum in (9) can be strengthened to be taken over θ rather than only over γ when a uniformly $\sqrt{n}$-consistent estimator (e.g., the LSE) of $\left(\beta', \sigma'\right)'$ is available. Nevertheless, the PLAM criterion is weaker than the genuine LAM criterion which requires that

$$\lim_{\delta\to 0}\varliminf_{n\to\infty}\left[\sup_{\|\vartheta-\vartheta_0\|<\delta} E_\vartheta\left[l_2\left(n\left(\widehat{\gamma} - \gamma\right)\right)\right]\right] \geq \lim_{\delta\to 0}\varliminf_{n\to\infty}\left[\sup_{\|\vartheta-\vartheta_0\|<\delta} E_\vartheta\left[l_2\left(n\left(T_n - \gamma\right)\right)\right]\right],$$

where the supremum is taken over $\vartheta = \left(\theta', f(\mathrm{w})\right)'$ in some neighborhood of $\vartheta_0 = \left(\theta_0', f_0(\mathrm{w})\right)'$. Note that a norm is required to define a neighborhood of $f_0(\mathrm{w})$, while the usual norms such as that in Chamberlain (1987) are not suitable in this nonregular environment. This is because the neighborhoods defined by this kind of norm include discrete distributions, while the distributions of q and e are required to be continuous in threshold regression. Even if such a norm (such as the Hölder norm) is well defined, the proof under the LAM criterion is difficult since it requires a serious extension of the parametric arguments in Ibragimov and Has'minskii (1981) to the nonparametric scenario. Even in the regular case, such an extension needs caution; see, e.g., Ritov and Bickel (1990). In spite of its restrictiveness, the PLAM criterion is not rare in the literature; e.g., Fan (1993) uses it in nonparametric regression and Chen and Reiss (2011) use it in nonparametric instrumental regression. In supplementary materials (see Appendix D), we discuss two other efficiency criteria and argue that they are not suitable in semiparametric threshold regression.

Note further that the assumptions on the prior in Appendix A are very weak since the only unknown parameter is γ in Algorithm G. An alternative estimation method in semiparametric threshold regression is to use the semiparametric Bayesian approach where the specification of priors is critical. In such a method, the parameter space is infinite-dimensional, as it also includes the function space where $f(\mathrm{w})$ stays. Some priors can induce inconsistent Bayes estimates in this case; see, e.g., Diaconis and Freedman (1986a,b). Fortunately, this problem is avoided by a frequentist estimation of $f(\mathrm{w})$ in Algorithm G. The semiparametric Bayesian method is still an open question even in regular models.

[5] Basically, we just strengthen the regularity conditions from the pointwise version to the uniform version; see, e.g., Ibragimov and Has'minskii (1981) for such uniform convergence results in the maximum likelihood method.

3.3. Orthogonality conditions vs. conditional moment restrictions

As is well known, the conditional moment restriction (2) adds information above the orthogonality conditions in estimating regular parameters such as β by exploring the information in conditional variance. As will be shown below, (2) provides very different information for γ from that for β. First, note that (2) is indeed required in our SEB procedure. For example, if the LSE is used as the n-consistent estimator in Step G1 of Algorithm G, then (2) is needed in identifying γ_0 and proving the n-consistency of the LSE.

To see the role of (2) in estimating γ, we will return to the simple case (5) in the following discussion. Since γ is of main interest, we assume that $\beta_{10} = 1$ is known. Note that the orthogonality conditions $E[e\mathbf{1}(q \le \gamma)] = 0$ and $E[e\mathbf{1}(q > \gamma)] = 0$ are not enough to identify γ, but $E[e|q] = 0$ is sufficient. To see why, let us consider a parametric submodel. Suppose the joint distribution of q and e is as follows:

$$f_{q,e}(q,e) = \begin{cases} \phi(e-1/2), & \text{if } 0 \le q < \frac{1}{4}, \\ \phi(e+1/2), & \text{if } \frac{1}{4} \le q \le \frac{1}{2}, \\ \phi(e-1/2), & \text{if } \frac{1}{2} < q \le \frac{3}{4}, \\ \phi(e+1/2), & \text{if } \frac{3}{4} < q \le 1, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $\phi(\cdot)$ is the density of the standard normal distribution. Suppose $\gamma_0 = \frac{1}{2}$, then it is easy to check that $E[e\mathbf{1}(q \le \gamma_0)] = 0$ and $E[e\mathbf{1}(q > \gamma_0)] = 0$, but γ_0 cannot be identified. This is because the joint distribution of y and q is

$$f_{y,q}(y,q) = \begin{cases} \phi(y-3/2), & \text{if } 0 \le q < \frac{1}{4}, \\ \phi(y-1/2), & \text{if } \frac{1}{4} \le q \le \frac{3}{4}, \\ \phi(y+1/2), & \text{if } \frac{3}{4} < q \le 1, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

so γ_0 is identified by the least squares criterion as $\left[\frac{1}{4}, \frac{3}{4}\right]$, not $\frac{1}{2}$; that is, this is a partially identified model.[6] To point-identify γ_0 as $\frac{1}{2}$, the mean of y in the left and right neighborhoods of γ_0 must be different. $E[e|q] = 0$ guarantees this condition, while the joint distribution of $(q', e)'$ in (10) does not satisfy this condition. When γ_0 is identified as a boundary of q, the n-consistency is straightforward according to the discussion in Section 2 of Yu (2012).[7]

Similarly, in quantile threshold regression, any moment condition $E[\tau - \mathbf{1}(e \le 0)|x, q] = 0$ indexed by a fixed $\tau \in (0,1)$ essentially provides sufficient identification information for γ in the SEB procedure as (2). After γ is confined in an n^{-1} neighborhood of its true value, e.g., using the procedures in Bai (1995), Caner (2002), Oka and Qu (2011) and Yu (2013b), Algorithm G can be employed to get an adaptive estimator of γ.

4. Confidence interval construction of γ

A common method to form CIs is through inverting Wald or t-statistics. We call such CIs Wald-type CIs. Due to the nonregularity of γ, Wald-type CIs of γ are difficult to construct as shown in Section 4.1. A natural antidote to inference when an estimator's distribution is nonstandard is to turn to the resampling methods. In Section 4.2, we review the existing resampling methods used in threshold regression. Finally, we provide in Section 4.3 an alternative CI for γ, called the nonparametric posterior interval (NPI), and prove its validity. A thorough finite-sample comparison of competing CIs is conducted in Section 6.2.

4.1. The difficulty of constructing Wald-type confidence intervals

Suppose the LSE is used as in Chan (1993). The LSE of γ is usually defined by a profiled procedure:

$$\widetilde{\gamma} = \arg\min_{\gamma} M_n(\gamma),$$

where

$$M_n(\gamma) \equiv \min_{\beta_1,\beta_2} \frac{1}{n}\sum_{i=1}^{n}\left(y - x'\beta_1\mathbf{1}(q \le \gamma) - x'\beta_2\mathbf{1}(q > \gamma)\right)^2.$$

Its asymptotic distribution is

$$\begin{aligned} n(\widetilde{\gamma}_l - \gamma_0) &\xrightarrow{d} M_-, \\ n(\widetilde{\gamma}_m - \gamma_0) &\xrightarrow{d} \frac{M_- + M_+}{2}, \end{aligned} \tag{12}$$

where $\widetilde{\gamma}_l$, called the left-endpoint LSE (LLSE), uses the left endpoint of the minimizing interval in least squares estimation, and $\widetilde{\gamma}_m$, called the middle-point LSE (MLSE), uses the middle point of the minimizing interval. In addition, $[M_-, M_+) = \arg\min_v D_{LS}(v)$ is the minimizing interval of $D_{LS}(v)$. $D_{LS}(v)$ is similarly defined as $D(v)$ in (4) except that

$$\begin{aligned} z_{1i} &= \left\{2x_i'(\beta_{10} - \beta_{20})\sigma_{10}e_i \right. \\ &\quad \left. + (\beta_{10} - \beta_{20})' x_i x_i' (\beta_{10} - \beta_{20})\right\} | (q_i = \gamma_0-); \\ z_{2i} &= \left\{-2x_i'(\beta_{10} - \beta_{20})\sigma_{20}e_i \right. \\ &\quad \left. + (\beta_{10} - \beta_{20})' x_i x_i' (\beta_{10} - \beta_{20})\right\} | (q_i = \gamma_0+). \end{aligned}$$

The MLSE is motivated similarly as the MMLE in Yu (2012).

It is hard to construct a Wald-type CI for γ by searching for percentiles of M_- and $\frac{M_- + M_+}{2}$. First, $f_q(\gamma_0)$ is required to find the distribution of M_- and $\frac{M_- + M_+}{2}$. If $f_q(\cdot)$ is known, then $f_q(\gamma_0)$ can be estimated by $f_q(\widetilde{\gamma})$, where $\widetilde{\gamma}$ can be $\widetilde{\gamma}_l$ or $\widetilde{\gamma}_m$. In reality, however, even in parametric models, the density of q is seldom specified. A nonparametric estimator of $f_q(\cdot)$ might be suggested, but such an estimator will suffer from the low convergence rate of nonparametric estimation. Second, infinite independent copies of z_{1i} and z_{2i} are also needed to simulate the sample path of $D_{LS}(\cdot)$, but they involve an unknown generator of conditional random variables.[8] Third, such a generator is not needed if q

[6] The objective function of the least squares estimator is $E\left[(y - \mathbf{1}(q \le \gamma))^2\right] = E\left[y^2 - 2y\mathbf{1}(q \le \gamma) + \mathbf{1}(q \le \gamma)\right] = \frac{7}{4} - 2$

$$\cdot \begin{cases} \frac{3}{2}\gamma, & \text{if } 0 \le \gamma < \frac{1}{4}, \\ \frac{3}{8} + \frac{1}{2}\left(\gamma - \frac{1}{4}\right), & \text{if } \frac{1}{4} \le \gamma \le \frac{3}{4}, \\ \frac{5}{8} - \frac{1}{2}\left(\gamma - \frac{3}{4}\right), & \text{if } \frac{3}{4} < \gamma \le 1, \end{cases} + \gamma = \frac{7}{4} - \begin{cases} 2\gamma, & \text{if } 0 \le \gamma < \frac{1}{4}, \\ \frac{1}{2}, & \text{if } \frac{1}{4} \le \gamma \le \frac{3}{4}, \\ 2 - 2\gamma, & \text{if } \frac{3}{4} < \gamma \le 1, \end{cases}$$

which is minimized on $\left[\frac{1}{4}, \frac{3}{4}\right]$. Hall et al. (2012) independently find a similar unidentification result in the GMM framework of structural change.

[7] By checking the proof of n-consistency of the LSE, we can see that the superconsistency is from the unsmoothness of the limit objective function at γ_0. In this simple example, the left derivative of the limit objective function is $-f_q(\gamma_0)$, while the right derivative is $f_q(\gamma_0)$.

[8] Kosorok and Song (2007) indeed put forward such a generator in a class of parametric threshold models arising in survival analysis, but the generator depends on some tuning parameters which are hard to select and also only on a subsample so that the generator cannot be precise. See also Li and Ling (2012) for an alternative generator in threshold autoregressive models.

is assumed to be independent of $(x', e)'$, but the distribution of $(x_i', e_i)'$ and $(\beta_0', \sigma_0')'$ are still needed to simulate z_{1i} and z_{2i}. The empirical distribution of $(x_i', \widetilde{e}_i)'$ can potentially substitute for the distribution of $(x_i', e_i)'$, and the least squares estimate substitutes for $(\beta_0', \sigma_0')'$, but this is very burdensome when the dimension of x is large. If all these above are specified, the algorithms in Appendix D of Yu (2012) can be used to construct a confidence interval for γ.

Due to the above difficulties, three alternative methods are put forward to avoid using $D_{LS}(\cdot)$. First, Seo and Linton (2007) propose a Wald-type CI for γ based on the smoothed least squares estimation (SLSE) in our framework. Because the asymptotic distribution of the SLSE is normal, standard CI construction methods such as inverting the t-statistic or the bootstrap can be used. But the convergence rate is less than n, which implies that the CI would be relatively wide. Also, a key smoothing parameter needs to be specified in practice. Second, Hansen (2000) employs a different framework as mentioned in the introduction and constructs a CI for γ by inverting the likelihood ratio statistic. But such a CI may be a union of disjoint intervals and is conservative in the present framework.[9] Third, Bai (1997) constructs a Wald-type CI for γ in structural change models using a similar framework as Hansen (2000) except allowing for threshold effects also in variance. The asymptotic distribution of $\widetilde{\gamma}_l$ and $\widetilde{\gamma}_m$ are the same in his framework and related to a two-sided Wiener process. However, the convergence rate is less than n, and as argued in Section 4.1 of Hansen (2000), this Wald-type CI performs poorly because the parameter has a region where identification fails.

4.2. Resampling methods

A few resampling methods were proposed in the semiparametric threshold regression literature. In a simple threshold regression where no covariates exist, only threshold effects in mean are present, and e is independent of q, Seijo and Sen (2011) show that Efron's nonparametric bootstrap is invalid. In the general framework (1), Yu (2014) further shows that the asymptotic bootstrap distribution does not even exist. He also shows that other bootstrap schemes such as the wild bootstrap or bootstrapping residuals when e is independent of $(x', q)'$ are not valid either.

Although the usual multiplier bootstrap mentioned above fails, there are two other resampling schemes that are valid in threshold regression. The first scheme is the smoothed bootstrap of Gijbels et al. (2004) and Seijo and Sen (2011). As argued in Yu (2014), this scheme is not practical since we need to estimate the joint density of $(x', q)'$ even if e is assumed to be independent of $(x', q)'$. Such an estimation obviously suffers from the curse of dimensionality when the dimension of $(x', q)'$ is large. The second scheme is the subsampling method of Gonzalo and Wolf (2005) (see also the related m out of n bootstrap in Seijo and Sen (2011)). This scheme is valid under very weak assumptions. Let $\{\widetilde{\gamma}_b^*\}_{b=1}^B$ denote the subsampling estimators, and m denote the subsample size. Politis and Romano (1994) state that the validity of the subsampling inference is based on two conditions: (a) The cdf of $\{m(\widetilde{\gamma}_b^* - \gamma_0)\}_{b=1}^B$ is a good approximation of the asymptotic distribution of $n(\widetilde{\gamma} - \gamma_0)$, which is guaranteed by the assumption that $m \to \infty$; (b) $m(\widetilde{\gamma} - \gamma_0)$ converges to zero in probability, which is insured by the assumption that $m/n \to 0$. Because we need both $m \to \infty$ and $m/n \to 0$ to guarantee the validity of subsampling, the sample size n is expected to be huge to make the subsampling work satisfactorily. In threshold regression, since the convergence rate of $\widetilde{\gamma}$ is n (rather than the usual $\sqrt{n}$), we find in the simulation of Section 6.2 that condition (a) is approximately satisfied even when n is as small as 400. Satisfaction of condition (b) is more subtle. When n is not so big, m might be treated as cn for some $c \in (0, 1)$ instead of $o(n)$. For example, in Seijo and Sen (2011), $m = n^{4/5}$, then when $n = 400$, $m \approx 0.3n$. If m/n is not approximately zero, the satisfaction of condition (b) relies on the small bias of $\widetilde{\gamma}$ in finite samples. In the simulation of Section 6.1, we find that $\widetilde{\gamma}_l$ has a larger bias than $\widetilde{\gamma}_m$, so we expect that the performance of the subsampling CIs based on $\widetilde{\gamma}_l$ is worse than those based on $\widetilde{\gamma}_m$.

4.3. Nonparametric posterior interval

Define

$$\widehat{F}_n(x) = \int_{\gamma \in \Gamma : \gamma \le x} \widehat{p}_n(\gamma) d\gamma \quad \text{and} \quad c_n(\tau) = \inf\left\{x : \widehat{F}_n(x) \ge \tau\right\},$$

then the $100(1-\tau)\%$ NPI for γ is given by $[c_n(\tau/2), c_n(1-\tau/2)]$, where $\tau \in (0, 1)$. In practice, such a confidence interval can be constructed by picking out the $\tau/2$ and $1-\tau/2$ quantile of the MCMC sequence S in (7). To make sure such a confidence interval is asymptotically valid, we need the following theorem.

Theorem 4 (*Asymptotic Validity of NPI*). *Under Assumptions 0, B, D, E, K and P in Appendix A, for any $\tau \in (0, 1)$, as long as $H_\infty^{-1}(\tau/2)$ and $H_\infty^{-1}(1-\tau/2)$ have positive density in an open neighborhood of 0, where $H_\infty^{-1}(\tau) \equiv \inf\left\{t : \int_{v \in \mathbb{R}: v \le t} p_2^*(v) dv \ge \tau\right\}$, then*

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

Theorem 4 states that the quantile of the posterior is asymptotically consistent. Note that $H_\infty^{-1}(\tau) = \arg\min_t \int (\tau - \mathbf{1}(v - t \le 0))(v - t) \cdot p_2^*(v) dv$ is the asymptotic distribution of $n(c_n(\tau) - \gamma_0)$, so the theorem requires that the posterior quantiles have positive density in an open neighborhood of 0. This assumption is needed in constructing the parametric posterior interval; see Theorem 3.3 of Chernozhukov and Hong (2004) and Theorem 4 of Yu (2012). For a CI, asymptotic validity is only the basic requirement. From the adaptiveness of our SEB procedure, we expect that the NPI also has advantages in length, and this is confirmed by simulations in Section 6.2.

5. Simplification and extension

This section provides some simplification and extension of our SEB approach. Some terminology will be used hereinafter: "the LSE" means the LLSE and the MLSE, while "the LSEs" also includes the SLSE.

5.1. Circumvention of the curse of dimensionality

Although the proposed method is theoretically appealing, it suffers from the curse of dimensionality when the dimension of $(x', q)'$ is high. One simplification is to assume that e is independent of $(x', q)'$. Such an assumption is standard in the literature; e.g., Chan (1993). In this case, only the marginal distribution of e, instead of the joint distribution of w, needs to be estimated, so the dimensionality of estimation is greatly reduced. Accordingly, Algorithm G can be simplified; especially, Step G2 is simplified to

[9] Theorem 3 of Hansen (2000) only proves the conservativeness of his CI in a special case of the framework used in this paper. The simulation studies and application there indicate that his CI is indeed conservative in our framework; see Sections 6.2 and 7 for more evidences on this point.

StepG2′: Estimate the density of e by kernel smoothing,

$$\widehat{f}(e) = \frac{1}{n}\sum_{i=1}^{n} K_h(\widetilde{e}_i - e)$$

with $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$, where h is the bandwidth, and $K(\cdot): \mathbb{R} \to \mathbb{R}$ is a kernel density.

Also, the $\widehat{f}(\mathrm{w})$ in Step G3 will be substituted by $\widehat{f}(e)$. The asymptotic distribution in Section 3.2 is also simplified, as the conditional density $f_{e|x,q}$ in $D(v)$ of Theorem 2 is changed to f_e, the marginal density of e. The procedures in Section 4 should also be adjusted correspondingly to construct CIs for γ. Note that if e is not independent of $(x', q)'$, that is, the model is misspecified, then the asymptotic distribution of this pseudo-SEBE is the same as that in Theorem 2 except that $f_{e|x,q}$ is replaced by f_e. Since e_i and $(x_i', q_i)'$ are dependent in the definition of $z_{\ell i}$, $\ell = 1, 2$, this estimator is not efficient and is similar to the quasi-MLE in Qu and Perron (2007). Also note that the NPI is not valid in this misspecified model as the validity of the posterior interval critically relies on the assumption that the model is correctly specified, see Theorem 4 of Yu (2012).

Under the assumption that e is independent of $(x', q)'$, even the regular parameters β can be adaptively estimated; see Bickel (1982) and Bickel et al. (1998) for the details. It is noteworthy that the reasons for adaptive estimation of γ and β are different: β can be adaptively estimated because the independence provides full information about the distribution of e, while for γ, it is because no local information around γ is lost from the assumption of independence as discussed in Section 2.2.

5.2. Recursive semiparametric empirical Bayes method

In Algorithm G, the performance of $\widehat{\gamma}$ critically depends on the estimation of $f(\mathrm{w})$. There are two types of finite-sample errors in this estimation. The first is from the slow convergence rate of the kernel smoother. The second is from the impreciseness of the initial estimator $\widetilde{\gamma}$ in Step G1. When the sample splitting of $\widetilde{\gamma}$ is far from the splitting of the true value γ_0, $\{\widetilde{e}_i\}_{i=1}^n$ will contain some bias.

The first error cannot be controlled for a fixed dataset, but the second error can. Specifically, in Algorithm G, we can use the SEBE $\widehat{\gamma}$ as the new preliminary estimate of γ in Step G1, and generate new $\left(\widetilde{\beta}', \widetilde{\sigma}'\right)'$ and $\{\widetilde{e}_i\}_{i=1}^n$ using $\widehat{\gamma}$ as the splitting point. Step G2 and G3 then follow to find a new estimate of γ. Repeat the above procedure until convergence. Note that if two consecutive SEBEs split the sample in the same way, i.e., both fall into the same interval $[q_{(i)}, q_{(i+1)})$ with $q_{(i)}$ being the ith order statistic of $\{q_i\}_{i=1}^n$, then the algorithm converges. The estimate of γ in the last iteration will be taken as our ultimate point estimate of γ, and the corresponding NPI as our ultimate set estimate of γ. Such a procedure is called the recursive semiparametric empirical Bayes (RSEB) method. It should be emphasized that the convergence properties of the RSEB method have not been theoretically proved yet. Nevertheless, from the simulation studies in Section 6.2 and the application in Section 7, for most data sets, the efficiency gain is achieved in the first SEBE and the algorithm will converge in the second iteration.

The RSEB method will improve the performance of γ estimation especially when $\beta_2 - \beta_1$ is small, since the identification power for γ is weak and the LSE of γ is not precise in this case. Note that the RSEB estimator has the same asymptotic distribution as the SEB estimator because in Algorithm G, we only require the starting estimator of γ to be n-consistent.

6. Simulations

In this section, we will report two simulations. For comparison with the parametric case, the same setup as in Section 4 of Yu (2012) 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. This simple setting is done to focus attention on the threshold estimation and to save simulation time. In the simulations, $\gamma_0 = 1/2$, $\sigma_{10} = 0.2$, $\sigma_{20} = 0.4$, and four β_1 values are used: 0.2, 0.4, 0.7 and 1, corresponding to tiny, small, medium and large threshold effects, respectively. Because e is independent of $(x', q)'$, the simplified procedures in Section 5.1 can be used.

The first simulation compares the risk of the SEBE and the RSEB estimator with the LSEs, and the second compares the coverage and length properties of the valid CIs discussed in Sections 4.1 and 4.2 with our NPI. To save space, all other procedures except the SEB approach are described in supplementary materials (see Appendix D).

6.1. Simulation 1: Risk comparison with the LSEs

Two key procedures of the SEB method are the estimation of f_e and the simulation from the nonparametric posterior $\widehat{p}_n(\gamma)$. In this simulation, we use two functions in Matlab, ksdensity and slicesample, to carry out these two procedures. ksdensity uses by default the optimal bandwidth when the true density is normal to get a smoothing density; see Section 3.4.2 of Silverman (1986) for details of this bandwidth selection. In the function slicesample, the arguments are some initial value and a posterior function form that is not necessarily normalized as a density; unlike in the Metropolis–Hastings sampler, the proposal distribution (or transition distribution) density is not required. We use the MLSE as the starting value, and draw 2000 samples from the posterior after discarding the first 200 "burn-in" draws. The prior in $\widehat{p}_n(\gamma)$ is specified in Section 3.1, but other specifications should not introduce any significant change in the performance. Refer to Neal (2003) for a concrete description of the slice sampling.

The performance of the estimators is summarized in Tables 1 and 3 of Appendix C. To explore the finite-sample information loss in the nonparametric density estimation of the SEB procedure, we also report the parametric results where the density of e is known. The asymptotic risk of all estimators is reported as well for comparison. Note that the asymptotic risk of the SLSE is not comparable with others since the SLSE has a slower convergence rate than n.

From Tables 1 and 3, the following conclusions can be drawn. (i) The performance of the posterior mean and median is robust to all β_1 values, and dominates the LSEs in all cases. Even in this semiparametric case, the posterior mean has the smallest MSE and the posterior median has the smallest MAD among all semiparametric estimators in most cases. (ii) Compared with the parametric results, the SEBE does not lose much efficiency in the nonparametric estimation of the error density when β_1 is not too small. When β_1 is small, the SEBE indeed loses a lot from the parametric case, although it is still better than the LSEs. This is partially due to the poor performance of the MLSE in the SEBE construction when β_1 is small. (iii) The asymptotic and finite-sample risk of the MLSE is between that of the LLSE and the SEBE in most cases. However, when β_1 is small, the ordering might reverse. This is not surprising from the risk calculation in Appendix D of Yu (2012). (iv) The SLSE performs worst in almost all cases, which validates

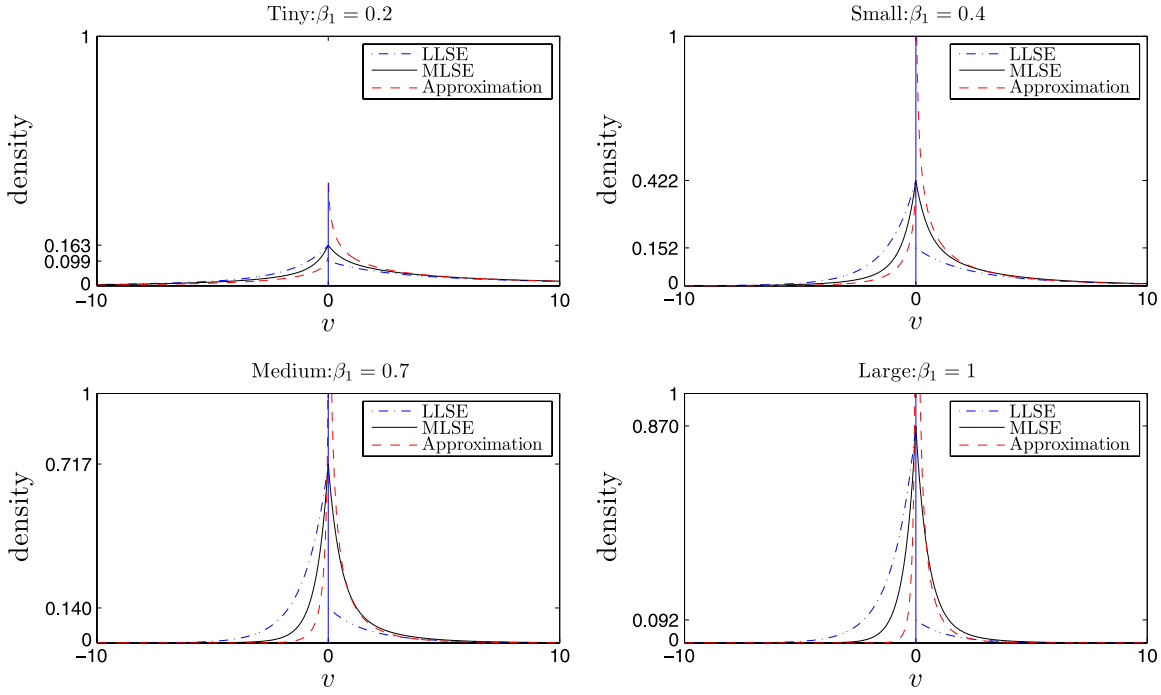

Fig. 1. Asymptotic density of the LSE compared with the approximation in Bai (1997).

its low convergence rate. Comparing the asymptotic and finite-sample risk of the SLSE, its convergence rate is faster than $\sqrt{n}$ but much slower than n. (v) The risk refinement in the RSEB method is secondary relative to that in the SEBE. The refinement gets smaller when n is larger, the recursion number is larger, or β_1 is larger (just as expected in Section 5.2). (vi) Table 3 shows the convergence property in the recursion. In most cases, the estimation will stop in two iterations. The recursion stops earlier when n or β_1 gets larger. Also, since it is more likely for the LSE, the SEBE and the RSEBE to match the sample splitting of the true threshold point when β_1 or n is larger, there is a smaller bias in $\{\widetilde{e}_i\}_{i=1}^n$ such that the refinement of the risk gets smaller in these cases. (vii) The finite-sample risk of all estimators except the SLSE matches their asymptotic risk when β_1 is not too small. This indicates that the finite-sample distribution is close to the asymptotic distribution, which is partially due to the superconsistency of the estimators. The n-consistency is also hinted at the risk comparison between the case $n = 100$ and $n = 400$: the risk of $n = 400$ is about a quarter of $n = 100$. When β_1 is small, such a result does not hold. In such a case, the small-threshold-effect asymptotic structure of Hansen (2000) may be suitable, and the convergence rate is slower than n.

6.2. Simulation 2: Comparison of confidence intervals

In this simulation, Hansen's method (2000) is applied to a misspecified model, since his method does not allow for threshold effects in error variance. Bai's method (1997) is therefore used to compare with Hansen's method. From Bai (1997), we can roughly state that $n(\widetilde{\gamma} - \gamma_0) \xrightarrow{d} \omega T(\psi)$, where $\omega = \sigma_{10}^2/\beta_{10}^2$, and $T(\psi)$ is the minimizer of a two-sided Brownian motion indexed by $\psi = \sigma_{20}^2/\sigma_{10}^2$. Fig. 1 plots the distribution of M_- and $\frac{M_- + M_+}{2}$ in (12) and $\omega T(\psi)$ for the four β_1 values. From Fig. 1, this approximation is accurate for the MLSE when β_1 is small, but not for the LLSE or when β_1 is large. For Seo and Linton's method (2007), we only report the CIs based on the bootstrap-t method to gain the finite-sample refinement, where the number of bootstrap repetitions is set to be 1000. In subsampling, $m = n/4$.

The simulation results are summarized in Table 2 of Appendix C. A few results of interest from Table 2 are as follows. (i) When β_1 is not too small, the NPI is the best in both coverage and length. Such a result is consistent with the efficiency results in Simulation 1. When β_1 is small, the NPI has an undercoverage problem as in almost all other methods.[10] (ii) When β_1 is not too small, the NPI does not lose much in coverage and length compared with the parametric confidence interval. The difference in the performance of the NPI and the parametric posterior interval gets larger when β_1 gets smaller. (iii) The coverage of Hansen's method (2000) depends on the value of β_1, which is also observed in Hansen (2000). When β_1 is large, this method is very conservative. Bai's method (1997) has a good coverage when β_1 is small, but has an undercoverage problem when β_1 is large. The CIs centered at the MLSE perform better than those centered at the LLSE. These results are consistent with the intuition in Fig. 1 that Bai's approximation is closer to the MLSE than the LLSE and is better when β_1 is smaller. In general, we can say that Hansen's method performs better than Bai's method.[11] (iv) The smoothed bootstrap CIs are very wide although the coverage property is acceptable. Among the four smoothed bootstrap CIs, it seems that the symmetric intervals based on the MLSE perform the best considering both coverage and length; see Section 4.2 of Yu (2014) for similar simulation results in more general setups. Nevertheless, it is fair to claim that this method is dominated by Hansen's method. (v) The CIs based on the SLSE performs worst in both coverage and length, which matches the efficiency results in Simulation 1. The symmetric CIs have a very different coverage from the equal-tailed CIs, which indicates that the finite-sample distribution of the SLSE, unlike the asymptotic distribution, is not symmetric.

[10] In practice, when β_1 is small, the data will not pass the specification test such as in Hansen (1996), so it is appropriate to only consider the data that pass the test.

[11] In another simulation unreported here, we specify $\sigma_{10} = \sigma_{20} = 0.3$, so both Hansen's method and Bai's method can be applied without misspecification. The main results do not change except that Hansen's method is conservative even when $\beta_{10} = 0.2$ and $n = 100$.

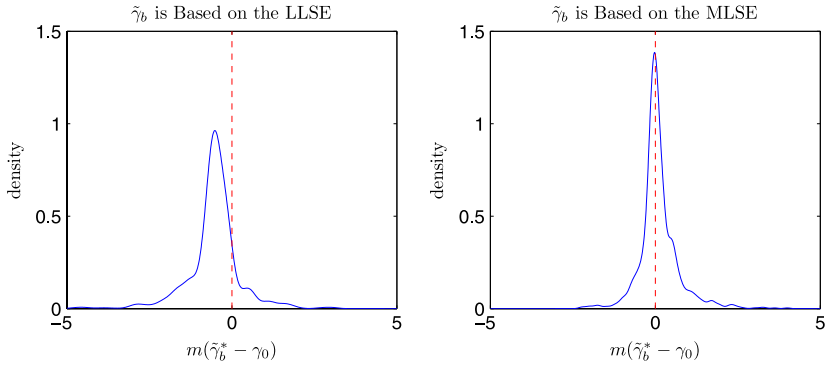

Fig. 2. Typical subsampling densities when $\beta_1 = 1$ and $n = 400$.

(vi) The subsampling method suffers from the undercoverage problem. This is especially true for the equal-tailed interval and the interval based on the LLSE or when β_1 is small. Furthermore, the subsampling CIs are much longer than the NPI. The undercoverage problem also happens in the setup of Gonzalo and Wolf (2005). It may come from the fixed block size. As a solution, the adaptive selection of block size can match the coverage, although it cannot solve the length problem. The worse length property of the subsampling method is related to the worse efficiency property of the LSE. Comparing the subsampling CIs based on the MLSE and those based on the LLSE, the former works better than the latter in both coverage and length when β_1 is not too small. First, the subsampling intervals based on the MLSE is shorter than those based on the LLSE. This is not surprising from Simulation 1 since the MLSE has a smaller variance than the LLSE. Second, the subsampling intervals based on the MLSE have a better coverage than those based on the LLSE. As explained in Section 4.2, this is from the fact that $\widetilde{\gamma}_l$ has a larger bias than $\widetilde{\gamma}_m$ in finite samples. Fig. 2 shows typical subsampling densities based on the LLSE and MLSE when $\beta_1 = 1$ and $n = 400$. Comparing Fig. 2 to Fig. 1, condition (a) in Section 4.2 is roughly satisfied for both $\widetilde{\gamma}_l$ and $\widetilde{\gamma}_m$. However, $m/n = 0.25$ in this simulation, so condition (b) in Section 4.2 is hardly satisfied. As a result, the bias of $\widetilde{\gamma}_l$ will affect $m(\widetilde{\gamma}_l - \gamma_0)$ in an unneglectable way. Gonzalo and Wolf (2005) only report the performance of the symmetric subsampling interval based on the LLSE, so it is unclear whether such a phenomenon also happens in their setups.

(vii) There is a trade-off between coverage and length in symmetric intervals and equal-tailed intervals. Although symmetric intervals have a better coverage, they are longer than equal-tailed intervals. (viii) When β_1 is not too small, the length of $n = 400$ is roughly a quarter of the length of $n = 100$ except the intervals based on the SLSE, which validates the n-consistency of γ estimators. (ix) The coverage property of the recursive NPIs is not improved by recursion, while the length property is improved although not as much as the improvement in the original NPI (relative to other CIs).

Based on these simulations, our suggestions are (a) use the SEB or RSEB method for inference on γ in both point and set estimation; (b) if the LSEs are used, the MLSE is preferable to the LLSE and the SLSE.

7. Application

In this section, we apply the SEB method to the growth data used in Durlauf and Johnson (1995) and reanalyzed in Hansen (2000). Here, the concern is whether there is a threshold effect in the GDP growth. The growth theory with multiple equilibria motivates the following threshold regression model:

$$\ln\left(\frac{Y}{L}\right)_{i,1985} - \ln\left(\frac{Y}{L}\right)_{i,1960} = \begin{cases} \beta_{10} + \beta_{11}\ln\left(\frac{Y}{L}\right)_{i,1960} + \beta_{12}\ln\left(\frac{I}{Y}\right)_i \\ \quad +\beta_{13}\ln(n_i + g + \delta) + \beta_{14}\ln S_i \\ \quad +\sigma_1 e_i, \quad \text{if } \left(\frac{Y}{L}\right)_{i,1960} \le \gamma; \\ \beta_{20} + \beta_{21}\ln\left(\frac{Y}{L}\right)_{i,1960} + \beta_{22}\ln\left(\frac{I}{Y}\right)_i \\ \quad +\beta_{23}\ln(n_i + g + \delta) + \beta_{24}\ln S_i \\ \quad +\sigma_2 e_i, \quad \text{if } \left(\frac{Y}{L}\right)_{i,1960} > \gamma. \end{cases}$$

For each country i, $\left(\frac{Y}{L}\right)_{i,t}$ is the real GDP per member of the population aged 15–64 in year t, $\left(\frac{I}{Y}\right)_i$ is the investment to GDP ratio, n_i is the growth rate of the working-age population, and S_i is the fraction of working-age population enrolled in secondary schools. The variables not indexed by t are annual averages over the period 1960–1985. Following Durlauf and Johnson (1995), we set $g + \delta = 0.05$. The data are assumed to be i.i.d. sampled. This assumption is approximately true, since there are not many interactions, such as trade, international capital flows, etc., between any two countries during this period. Furthermore, e_i is assumed to be independent of the regressors and the threshold variable. The objective is to check whether the growth rate depends on the starting point.

The LSE of β_ℓ and σ_ℓ^2, $\ell = 1, 2$, suggested in Chan (1993) are

$$\widetilde{\beta}_1 = (4.312, -0.657, 0.228, -0.295, 0.018)', \qquad \widetilde{\sigma}_1^2 = 0.0375,$$
$$\widetilde{\beta}_2 = (3.663, -0.323, 0.496, -0.488, 0.356)', \qquad \widetilde{\sigma}_2^2 = 0.0942.$$

Clearly, the difference between $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$ is significant relative to $\widetilde{\sigma}_\ell^2$, so the SEB method is suitable. When the RSEB method is used, the estimates of β_ℓ and σ_ℓ^2, $\ell = 1, 2$ are

$$\widehat{\beta}_1 = (4.778, -0.792, 0.314, -0.429, -0.029)',$$
$$\widehat{\sigma}_1^2 = 0.0261,$$
$$\widehat{\beta}_2 = (3.509, -0.313, 0.436, -0.506, 0.386)', \qquad \widehat{\sigma}_2^2 = 0.0939.$$

Notice that the effect of schooling is negative for the poor countries, and a simple test shows that this negative effect is not significantly different from zero. The effect of investment is significantly greater than zero. For rich countries, both effects are statistically positive. This result shows that the direct investment has a more significant effect on the economic growth of poor countries than schooling.

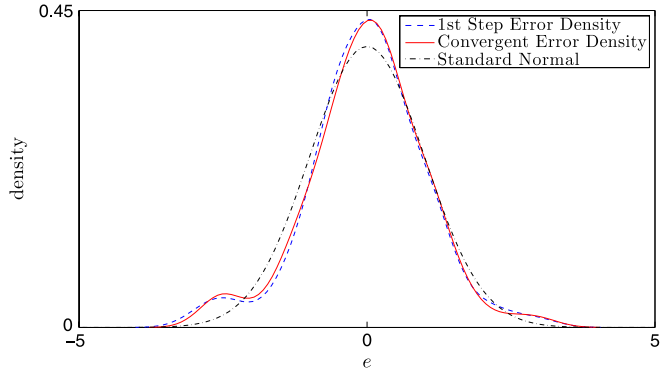

Fig. 3. Error density compared with standard normal.

The estimates of σ_1^2 and σ_2^2 are not statistically equal,[12] so there is a threshold effect in variance which Hansen (2000) does not assume. The SLSE of β is

$$\widehat{\beta}_1^s = (4.337, -0.660, 0.227, -0.288, 0.017)\,,$$
$$\widehat{\beta}_2^s = (3.661, -0.322, 0.496, -0.485, 0.357)\,,$$

which is not significantly different from the LSE.

The results on statistical inference of γ are summarized in Table 4 of Appendix C. The items under Hansen (2000) assume $\sigma_1 = \sigma_2$. The block size of the subsampling interval is 25. The smoothed bootstrap cannot be applied here since $\dim(x) = 4$ and n is only 96. The prior for the SEB estimation is uniform on $\left(\left(\left(\frac{Y}{L}\right)_{i,1960}\right)_{\min}, \left(\left(\frac{Y}{L}\right)_{i,1960}\right)_{\max}\right)$. Table 4 shows that the results in this practical example match the intuitions provided by the simulations in Section 6, especially the cases when the threshold effects are large. First, the CI suggested in Hansen (2000) is very conservative. Because there is a misspecification problem in Hansen's method (2000), we also report the CI of Bai (1997) which allows for the threshold effect in variance. This CI is short, but it suffers from the undercoverage problem when the threshold effects are large as suggested in the simulations. Second, the CIs based on the SLSE are extremely conservative due to the low convergence rate. They are not reasonable in this application since they even cover some negative values that $\left(\frac{Y}{L}\right)_{i,1960}$ cannot take. These features are also shared by the subsampling method. Third, our recursive NPI is the shortest among all CIs, and has a good coverage as suggested by the simulations. The first-step and the convergent error densities in our RSEB method are shown in Fig. 3. They are far from the standard normal distribution. Another observation is that the LSEs are out of the convergent NPI, which indicates a large bias in the LSEs. The five countries covered by the convergent NPI are Central African Republic, Mauritania, Togo, Burundi and Nepal. The countries with $\left(\frac{Y}{L}\right)_{i,1960}$ below this interval are mostly poor African countries. So the estimation shows that there are different growth patterns between poor countries and rich countries, and only a few countries are uncertain about the growth pattern.

8. Conclusion

In this paper, we propose a semiparametric empirical Bayes estimator of the threshold point under standard conditional moment restrictions in threshold regression. A critical result is that the threshold point can be adaptively estimated even when the error term is not independent of the regressors and the threshold variable. We also propose a valid confidence interval, the nonparametric posterior interval, for the threshold point. Simulation studies show that the semiparametric empirical Bayes estimator has a smaller risk than the least squares estimators, and the nonparametric posterior interval has better coverage and length properties than the existing methods. The extension of our method to the time series case considered in the existing literature such as Chan (1993) and Hansen (2000), however, requires nontrivial efforts. Furthermore, the case with small threshold effects deserves further explorations.

Acknowledgments

I am grateful to my committee members: Bruce Hansen, Jack Porter and Kenneth West for their helpful comments that significantly improved the manuscript. Special thanks are given to my advisor Bruce Hansen for his constant encouragement and valuable suggestions to this paper. Thanks are also given to Steven Durlauf and Stephen Walker for their help with the nonparametric Bayes method. This paper also benefits from the insightful comments from Songxi Chen, Kjell Doksum, Nikolay Gospodinov, Christian Hansen, Guido Kuersteiner, Ryo Okui, Peter Phillips, Donggyu Sul, Hansheng Wang, Jonathan Wright, Chunming Zhang, the editor, two referees and seminar participants at Auckland, Concordia, GSM/PKU, HKUST and UW-Madison, for which I want to express my thanks.

Appendix A. Regularity conditions

Assumptions on the prior, loss function, kernel and bandwidth

Assumption P. The prior $\pi_2 : \Gamma \to \mathbb{R}_+ \equiv \{x \in \mathbb{R} | x \geq 0\}$ is a continuous density function with $\pi_2(\gamma_0) > 0$.

Assumption L. The loss function $l_n : \mathbb{R}^{2k+3} \to \mathbb{R}_+$ satisfies:

(i) $l_n(u, v) = l_{1n}(u) + l_{2n}(v)$; that is, l_n is separable.
(ii) $l_{1n}(u) = l_1(\sqrt{n}u)$, and $l_{2n}(v) = l_2(nv)$, where $l_i(x) \geq 0$ and $l_i(x) = 0$ iff $x = 0$ for $i = 1, 2$.
(iii) $l_i(x)$ is convex, and $l_i(x) \leq 1 + \|x\|^p$ for some $p \geq 1$, $i = 1, 2$. Furthermore, $l_1(x)$ is bowl-shaped.

Remark 1. The assumptions on the loss function are standard. The separability of l_n is discussed in the main text. The convexity of l_i is not only for computational purpose, but for the proof convenience, since the convexity lemma can be applied under this assumption. The bowl-shapedness of l_1 is to assure that the MLE and the BE of $\underline{\theta}$ are asymptotically equivalent.

Assumption K. $K(x)$ is a known symmetric $C^{(1)}$ density with a compact support, and $\int x_i K(x)\, dx = 0$, $i = 1, \ldots, k+2$.

Remark 2. We only assume $K(x)$ to be a second order kernel to avoid a negative density estimator. Higher order kernels can be used to reduce bias. Compactness of the support is a standard assumption in the literature to simplify the proof. Many kernels, e.g., the product quartic kernel, satisfy Assumption K.

Assumption B. $h \to 0$, $n^{(1-\eta)/2}h^{(k+2)/2} \to \infty$, $n^{(1-\eta)/2}h^{k/2+3} \to 0$, and $n^{\eta/2-1/4}h^{k/2+2} \to \infty$, for some $1/2 < \eta < 1$.

Remark 3. The set of feasible h in Assumption B is not empty, e.g., η can take $\frac{3}{4}$, and $h = Cn^{-\frac{1}{4(k+5)}}$. There is a trade-off between the order of the kernel and the freedom of h. h can get more freedom when a higher order kernel is used, and this is standard in the discussion about the bias of kernel density estimation.

[12] $\sigma_2^2 > \sigma_1^2$ means that the rich countries are more easily impacted by the disturbance than the poor countries. This is understandable if we examine how the financial crisis starting from 2008 affected the US and the North Korea.

Assumptions on the data generating process

Assumption D. (**D1**) $f(\mathsf{w})$ is a $C^{(2)}$ bounded function on $\mathbb{R}^{k+2}$ with the second order derivatives uniformly bounded for all w.

Remark 4. The continuity of $f(\mathsf{w})$ on $\mathbb{R}^{k+2}$ implies that it is zero on its boundary of support, which guarantees the uniform consistency of the nonparametric density estimator in Lemma 1 of supplementary materials (see Appendix D). This assumption is standard in the literature; see, e.g., Assumption 3 of Stoker (1986) or Assumption 2 of Powell et al. (1989).

(**D2**) There exists compact $\mathbb{W}_n$ such that $n \cdot P(\mathsf{w}_i \notin \mathbb{W}_n) \to 0$, $\mathbb{W}_n \to \mathbb{W}$, and the diameter of $\mathbb{W}_n$ is less than n^{ν} for some $\nu \in \left(0, \frac{1}{2}\right)$. Moreover, $f(\mathsf{w}) > Cn^{-(1-\eta)/2}h^{-(k+2)/2} + Cn^{\nu-1/2}$ on $\mathbb{W}_n$.

Remark 5. The first part of this assumption puts some constraints on the tail of $f(\mathsf{w})$. The second part bounds $f(\mathsf{w})$ from below on $\mathbb{W}_n$ to avoid trimming the density estimator. This assumption can be relaxed with a messier proof as in Ai (1997). When x has a compact support, we only need $\nu > 0$ and $f(\mathsf{w}) > Cn^{-(1-\eta)/2}h^{-(k+2)/2}$ on $\mathbb{W}_n$. Depending on the relative magnitude of $n^{-(1-\eta)/2}h^{-(k+2)/2}$ and $n^{\nu-1/2}$, the lower bound of $f(\mathsf{w})$ on $\mathbb{W}_n$ can be absorbed into one term. From the proof below, $Cn^{-(1-\eta)/2}h^{-(k+2)/2}$ and $Cn^{\nu-1/2}$ represent two kinds of errors in Step G3 of Algorithm G: the former is from the kernel density estimation, and the latter is from substituting $\left(\widetilde{\beta}', \widetilde{\sigma}'\right)'$ for $\left(\beta_0', \sigma_0'\right)'$.

(**D3**) There exists a positive number C such that we have the equation given in Box I.

Remark 6. This assumption puts some restrictions on the tail of $f(\mathsf{w})$. It is standard, see, e.g., Assumption D6 of Yu (2012), to assume that

$$E\left[\left|\ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i(\beta_{10}-\beta_{20})}{\sigma_{10}}, x_i, q_i\right)}{f(\mathsf{w}_i)}\right|\right] < \infty,$$

$$E\left[\left|\ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i(\beta_{10}-\beta_{20})}{\sigma_{20}}, x_i, q_i\right)}{f(\mathsf{w}_i)}\right|\right] < \infty.$$

Since $n^{-(1-\eta)/2}h^{-(k+2)/2}$ and $n^{\nu-1/2}$ converge to 0 based on Assumptions B and D2, the expectation in D3 is assumed to exist when the arguments of the ln function are increased a little bit. The indicator function $\mathbf{1}(\mathsf{w}_i \in \mathbb{W}_n)$ assures the denominator greater than zero. When x has a compact support, the term $Cn^{\nu-1/2}$ can be omitted from the numerator and denominator.

(**D4**) $0 < \underline{f}_q \le f_q(q) \le \bar{f}_q < \infty$ for $q \in \Gamma$.

(**D5**) $E\left[\|x\|^4\right] < \infty$, and $E\left[e^4\right] < \infty$.

Remark 7. Assumption D includes only regularity conditions about the data generating process, and does not include any information related to the conditional moment restriction (2). Since Assumption D is the main assumption in proving the adaptiveness of $\widehat{p}_n^*$, the SEB procedure does not use any information related to (2) which provides information only through identification of γ as discussed in Section 3.3.

Assumption E. All other assumptions in Assumption D of Yu (2012) that are not covered by Assumption D are satisfied. The only difference is that the nuisance parameter α there is assumed to be known. To save space, they are summarized in supplementary materials (see Appendix D).

Remark 8. The division of Assumptions D and E is to separate the conditions needed in the SEB procedure from those required in the parametric estimation.

Appendix B. Proofs

Proof of Theorem 1. This proof is inspired by Theorem 1 of Chernozhukov and Hong (2003). From Lemma 1 in supplementary materials (see Appendix D), $n^{(1-\eta)/2}h^{(k+2)/2}\sup_{\mathsf{w}\in\mathbb{W}_n}\left|\widehat{f}(\mathsf{w}) - f(\mathsf{w})\right| \xrightarrow{p} 0$ with $\widetilde{f}(\mathsf{w}) = \frac{1}{n}\sum_{i=1}^n K_h(\mathsf{w}_i - \mathsf{w})$. Note that

$$\begin{aligned}
\widehat{f}(\mathsf{w}) &= \frac{1}{n}\sum_{i=1}^n K_h(\widetilde{\mathsf{w}}_i - \mathsf{w}) \\
&= \frac{1}{n}\sum_{i=1}^n K_h(\mathsf{w}_i - \mathsf{w}) + \frac{1}{n}\sum_{i=1}^n \{K_h(\widetilde{\mathsf{w}}_i - \mathsf{w}) - K_h(\mathsf{w}_i - \mathsf{w})\} \\
&= \frac{1}{n}\sum_{i=1}^n K_h(\mathsf{w}_i - \mathsf{w}) + \frac{1}{n}\sum_{i=1}^n K'_{1h}(\xi_i, x_i - x, q_i - q)(\widetilde{e}_i - e_i),
\end{aligned}$$

where the last equality is from the mean value theorem, ξ_i is some point between $\widetilde{e}_i - e$ and $e_i - e$, and K'_{1h} is the partial derivative of K_h with respect to the first argument. Notice that

$$\begin{aligned}
\widetilde{e}_i - e_i &\overset{(1)}{=} \frac{y_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}\mathbf{1}(q_i \le \widetilde{\gamma}) + \frac{y_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}\mathbf{1}(q_i > \widetilde{\gamma}) - e_i \\
&\overset{(2)}{=} \frac{x_i'\beta_{10} + \sigma_{10}e_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}\mathbf{1}(q_i \le \gamma_0 \wedge \widetilde{\gamma}) \\
&\quad + \frac{x_i'\beta_{10} + \sigma_{10}e_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}\mathbf{1}(\widetilde{\gamma} < q_i \le \gamma_0) \\
&\quad + \frac{x_i'\beta_{20} + \sigma_{20}e_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}\mathbf{1}(\gamma_0 < q_i \le \widetilde{\gamma}) \\
&\quad + \frac{x_i'\beta_{20} + \sigma_{20}e_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}\mathbf{1}(q_i > \gamma_0 \vee \widetilde{\gamma}) - e_i \\
&\overset{(3)}{=} x_i'\frac{\beta_{10} - \widetilde{\beta}_1}{\widetilde{\sigma}_1}\mathbf{1}(q_i \le \gamma_0 \wedge \widetilde{\gamma}) \\
&\quad + x_i'\frac{\beta_{20} - \widetilde{\beta}_2}{\widetilde{\sigma}_2}\mathbf{1}(q_i > \gamma_0 \vee \widetilde{\gamma}) \\
&\quad + \left[\frac{\sigma_{10}}{\widetilde{\sigma}_1}\mathbf{1}(q_i \le \gamma_0 \wedge \widetilde{\gamma}) + \frac{\sigma_{20}}{\widetilde{\sigma}_2}\mathbf{1}(q_i > \gamma_0 \vee \widetilde{\gamma}) - 1\right]e_i \\
&\quad + O_p\left(n^{-3/4}\right) \\
&\overset{(4)}{=} O_p\left(n^{-1/4}\right) \quad \text{uniformly for } i, \qquad (13)
\end{aligned}$$

where (1) is from the definition of $\widetilde{e}_i$, (2) is decomposing the support of q_i into four pieces, and substituting y_i on each piece, (3) is from $\widetilde{\gamma} - \gamma_0 = O_p(n^{-1})$ and Assumptions D4 and D5,[13] and (4) is from the assumptions on $\left(\widetilde{\beta}', \widetilde{\sigma}', \widetilde{\gamma}\right)'$ in Step G1. So $\frac{1}{n}\sum_{i=1}^n K'_{1h}(\xi_i, x_i - x, q_i - q)(\widetilde{e}_i - e_i) = O_p\left(\frac{1}{n^{1/4}h^{k+3}}\right)$. From Assumption B, $n^{1/4-\eta/2}h^{-k/2-2} \to 0$, so $n^{-1/4}h^{-(k+3)} = o\left(n^{-(1-\eta)/2}h^{-(k+2)/2}\right)$, and

$$n^{(1-\eta)/2}h^{(k+2)/2}\sup_{\mathsf{w}\in\mathbb{W}_n}\left|\widehat{f}(\mathsf{w}) - f(\mathsf{w})\right| \xrightarrow{p} 0. \qquad (14)$$

[13] Assumption D5 implies $\max_{1\le i\le n}\|x_i\| < n^{1/4}$, and $\max_{1\le i\le n}|e_i| < n^{1/4}$ with probability one when n is large enough by the Borel–Cantelli lemma. So $\max_{1\le i\le n}\left|\frac{x_i'\beta_{10} + \sigma_{10}e_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}\right| = O_p(n^{1/4})$ given Assumption 0 and the consistency of $\widetilde{\beta}_2$ and $\widetilde{\sigma}_2$. From Assumption D4, for any i, $\mathbf{1}(\widetilde{\gamma} < q_i \le \gamma_0) = O_p(n^{-1})$, so $\frac{x_i'\beta_{10} + \sigma_{10}e_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}\mathbf{1}(\widetilde{\gamma} < q_i \le \gamma_0) = O_p(n^{-3/4})$ for any i.

$$E\left[\sup_n \left|\mathbf{1}(w_i \in \mathbb{W}_n) \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i'(\beta_{10}-\beta_{20})}{\sigma_{10}}, x_i, q_i\right) + Cn^{-(1-\eta)/2}h^{-(k+2)/2} + Cn^{\nu-1/2}}{f(w_i) - Cn^{-(1-\eta)/2}h^{-(k+2)/2} - Cn^{\nu-1/2}}\right|\right] < \infty,$$

$$E\left[\sup_n \left|\mathbf{1}(w_i \in \mathbb{W}_n) \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\beta_{20})}{\sigma_{20}}, x_i, q_i\right) + Cn^{-(1-\eta)/2}h^{-(k+2)/2} + Cn^{\nu-1/2}}{f(w_i) - Cn^{-(1-\eta)/2}h^{-(k+2)/2} - Cn^{\nu-1/2}}\right|\right] < \infty$$

Box I.

From the definition of $\widehat{p}_n^*(v)$,

$$\widehat{p}_n^*(v) = \frac{\pi_2\left(\gamma_0 + \frac{v}{n}\right)\exp\{\widehat{\omega}(v)\}}{C_n}$$

where $\widehat{\omega}(v) = \widehat{L}_n\left(\gamma_0 + \frac{v}{n}\right) - \widehat{L}_n(\gamma_0)$, and $C_n = \int_{V_n} \pi_2\left(\gamma_0 + \frac{v}{n}\right)\exp\{\widehat{\omega}(v)\}\,dv$. If it can be proved that for each $\kappa \geq 0$,

$$A_{1n} = \int_{V_n} |v|^\kappa \pi_2\left(\gamma_0 + \frac{v}{n}\right)|\exp\{\widehat{\omega}(v)\} - \exp\{\omega(v)\}|\,dv \xrightarrow{p} 0, \tag{15}$$

where $\omega(v) = L_n\left(\gamma_0 + \frac{v}{n}\right) - L_n(\gamma_0)$, then taking $\kappa = 0$,

$$\left|C_n - \int_{V_n} \pi_2\left(\gamma_0 + \frac{v}{n}\right)\exp\{\omega(v)\}\,dv\right| \xrightarrow{p} 0, \tag{16}$$

which implies that C_n is bounded from 0 and is equal to $O_p(1)$ by Yu (2012). Note that it suffices to show the following to prove (8):

$$\int_{V_n} |v|^\kappa \left|\widehat{p}_n^*(v) - p_n^*(v)\right| dv \xrightarrow{p} 0, \tag{17}$$

where the left hand side is equal to A_n/C_n with

$$A_n = \int_{V_n} |v|^\kappa \pi_2\left(\gamma_0 + \frac{v}{n}\right) \times \left|\exp\{\widehat{\omega}(v)\} - \frac{\exp(\omega(v))}{\int_{V_n} \pi_2\left(\gamma_0 + \frac{\overline{v}}{n}\right)\exp(\omega(\overline{v}))d\overline{v}} C_n\right| dv.$$

Since C_n is positive and bounded from 0, it suffices to show that $A_n \xrightarrow{p} 0$. But $A_n \leq A_{1n} + A_{2n}$ with

$$A_{2n} = \int_{V_n} |v|^\kappa \pi_2\left(\gamma_0 + \frac{v}{n}\right) \times \left|\frac{\exp(\omega(v))}{\int_{V_n} \pi_2\left(\gamma_0 + \frac{\overline{v}}{n}\right)\exp(\omega(\overline{v}))d\overline{v}} C_n - \exp\{\omega(v)\}\right| dv.$$

Note that

$$A_{2n} = \left|\frac{C_n}{\int_{V_n} \pi_2\left(\gamma_0 + \frac{\overline{v}}{n}\right)\exp(\omega(\overline{v}))d\overline{v}} - 1\right| \times \int_{V_n} |v|^\kappa \pi_2\left(\gamma_0 + \frac{v}{n}\right)\exp(\omega(v))dv \xrightarrow{p} 0$$

by (16) and Yu (2012), so it remains only to show (15).

Note that

$$\widehat{\omega}(v) = \sum_{i=1}^n \ln \frac{\frac{\widetilde{\sigma}_2}{\widetilde{\sigma}_1}\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}, x_i, q_i\right)}{\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}, x_i, q_i\right)} \mathbf{1}\left(\gamma_0 < q_i \leq \gamma_0 + \frac{v}{n}\right) + \sum_{i=1}^n \ln \frac{\frac{\widetilde{\sigma}_1}{\widetilde{\sigma}_2}\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_2}{\widetilde{\sigma}_2}, x_i, q_i\right)}{\widehat{f}\left(\frac{y_i - x_i'\widetilde{\beta}_1}{\widetilde{\sigma}_1}, x_i, q_i\right)} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \leq \gamma_0\right),$$

$$\omega(v) = \sum_{i=1}^n \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{y_i - x_i'\beta_{10}}{\sigma_{10}}, x_i, q_i\right)}{f\left(\frac{y_i - x_i'\beta_{20}}{\sigma_{20}}, x_i, q_i\right)} \mathbf{1}\left(\gamma_0 < q_i \leq \gamma_0 + \frac{v}{n}\right) + \sum_{i=1}^n \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{y_i - x_i'\beta_{20}}{\sigma_{20}}, x_i, q_i\right)}{f\left(\frac{y_i - x_i'\beta_{10}}{\sigma_{10}}, x_i, q_i\right)} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \leq \gamma_0\right).$$

Split the integral in A_{1n} into two separate areas: Area (i): $|v| \leq M$ and Area (ii): $|v| > M$, where these two areas are understood as the intersection with V_n.

Area (i) The objective is to show that for each $0 < M < \infty$, and each $\epsilon > 0$,

$$\lim_{n\to\infty} P_*\left\{\int_{|v|\leq M} |v|^\kappa \pi_2\left(\gamma_0 + \frac{v}{n}\right)|\exp\{\widehat{\omega}(v)\} - \exp\{\omega(v)\}|\,dv < \epsilon\right\} > 1 - \epsilon, \tag{18}$$

but this follows from

$$\sup_{|v|\leq M} |\widehat{\omega}(v) - \omega(v)| \xrightarrow{p} 0,$$

where P_* is the inner probability which avoids the technicality of measurability. From Assumption D4, this is equivalent to show

$$\left|\frac{1}{n}\sum_{i=1}^n \ln \frac{\frac{\widetilde{\sigma}_2}{\widetilde{\sigma}_1}\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'(\widetilde{\beta}_1 - \beta_{20})}{\widetilde{\sigma}_1}, x_i, q_i\right)}{\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'(\widetilde{\beta}_2 - \beta_{20})}{\widetilde{\sigma}_2}, x_i, q_i\right)} - \frac{1}{n}\sum_{i=1}^n \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i'(\beta_{10} - \beta_{20})}{\sigma_{10}}, x_i, q_i\right)}{f(w_i)}\right| \xrightarrow{p} 0,$$

$$\left|\frac{1}{n}\sum_{i=1}^n \ln \frac{\frac{\widetilde{\sigma}_1}{\widetilde{\sigma}_2}\widehat{f}\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10} - \widetilde{\beta}_2)}{\widetilde{\sigma}_2}, x_i, q_i\right)}{\widehat{f}\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10} - \widetilde{\beta}_1)}{\widetilde{\sigma}_1}, x_i, q_i\right)} - \frac{1}{n}\sum_{i=1}^n \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10} - \beta_{20})}{\sigma_{20}}, x_i, q_i\right)}{f(w_i)}\right| \xrightarrow{p} 0.$$

These convergence results follow from

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'\left(\widetilde{\beta}_1-\beta_{20}\right)}{\widetilde{\sigma}_1}, x_i, q_i\right) - \ln f\left(\frac{\sigma_{20}e_i - x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{10}}, x_i, q_i\right)\right] \xrightarrow{p} 0,$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'\left(\widetilde{\beta}_2-\beta_{20}\right)}{\widetilde{\sigma}_2}, x_i, q_i\right) - \ln f\left(\mathbf{w}_i\right)\right] \xrightarrow{p} 0,$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln\widehat{f}\left(\frac{\sigma_{10}e_i + x_i'\left(\beta_{10}-\widetilde{\beta}_2\right)}{\widetilde{\sigma}_2}, x_i, q_i\right) - \ln f\left[\frac{\sigma_{10}e_i + x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{20}}, x_i, q_i\right]\right] \xrightarrow{p} 0,$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln\widehat{f}\left(\frac{\sigma_{10}e_i + x_i'\left(\beta_{10}-\widetilde{\beta}_1\right)}{\widetilde{\sigma}_1}, x_i, q_i\right) - \ln f\left(\mathbf{w}_i\right)\right] \xrightarrow{p} 0.$$

We only prove the first convergence result, and all the others can be proved similarly. From the assumptions on $\left(\widetilde{\beta}', \widetilde{\sigma}'\right)'$ in Step G1 of Algorithm G, Lemma 1 and the discussion at the beginning of this proof can be modified to show that

$$\sup_{\mathbf{w}_i\in\mathbb{W}_n}\left|\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'\left(\widetilde{\beta}_1-\beta_{20}\right)}{\widetilde{\sigma}_1}, x_i, q_i\right) - f\left(\frac{\sigma_{20}e_i - x_i'\left(\widetilde{\beta}_1-\beta_{20}\right)}{\widetilde{\sigma}_1}, x_i, q_i\right)\right| \xrightarrow{p} 0.$$

Moreover, since the first partial derivative of $f(\mathbf{w})$ is bounded,

$$\sup_{\mathbf{w}_i\in\mathbb{W}_n}\left|f\left(\frac{\sigma_{20}e_i - x_i'\left(\widetilde{\beta}_1-\beta_{20}\right)}{\widetilde{\sigma}_1}, x_i, q_i\right) - f\left(\frac{\sigma_{20}e_i - x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{10}}, x_i, q_i\right)\right| = O_p\left(n^{\nu-1/2}\right) \xrightarrow{p} 0. \quad (19)$$

Combining these two convergence results with the assumption that $n\cdot P\left(\mathbf{w}_i \notin \mathbb{W}_n\right) \to 0$, the result follows.

Area (ii) The goal is to show that for each $\epsilon > 0$, there exists a large M such that

$$\lim_{n\to\infty} P_*\left\{\int_{|v|>M} |v|^{\kappa}\,\pi_2\left(\gamma_0+\frac{v}{n}\right)\left|\exp\left\{\widehat{\omega}(v)\right\} - \exp\left\{\omega(v)\right\}\right| dv < \epsilon\right\} > 1-\epsilon. \quad (20)$$

Since the second term can be made arbitrarily small by setting M large according to Yu (2012), it suffices to show that for each $\epsilon > 0$, there exists a large M such that

$$\lim_{n\to\infty} P_*\left\{\int_{|v|>M} |v|^{\kappa}\,\pi_2\left(\gamma_0+\frac{v}{n}\right)\exp\left\{\widehat{\omega}(v)\right\} dv < \epsilon\right\} > 1-\epsilon.$$

In order to do so, it suffices to show that $\exp\left\{\widehat{\omega}(v)\right\} \le \exp\left\{C\omega(v)\right\}$ for some $0 < C < 1$ for all $|v| > M$. From (14) and the analysis for area (i), there exists a $C > 0$ such that with probability approaching 1, we have the equation in Box II. From Assumptions B, D2, D3 and the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\frac{\sigma_{20}}{\sigma_{10}}f\left(\frac{\sigma_{20}e_i - x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{10}}, x_i, q_i\right) + Cn^{\nu-1/2} + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f\left(\mathbf{w}_i\right) - Cn^{\nu-1/2} - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \times \mathbf{1}\left(\mathbf{w}_i\in\mathbb{W}_n\right) \xrightarrow{p} E\left[\ln\frac{\frac{\sigma_{20}}{\sigma_{10}}f\left(\frac{\sigma_{20}e_i - x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{10}}, x_i, q_i\right)}{f\left(\mathbf{w}_i\right)}\right],$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\frac{\sigma_{10}}{\sigma_{20}}f\left(\frac{\sigma_{10}e_i + x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{20}}, x_i, q_i\right) + Cn^{\nu-1/2} + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f\left(\mathbf{w}_i\right) - Cn^{\nu-1/2} - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \times \mathbf{1}\left(\mathbf{w}_i\in\mathbb{W}_n\right) \xrightarrow{p} E\left[\ln\frac{\frac{\sigma_{10}}{\sigma_{20}}f\left(\frac{\sigma_{10}e_i + x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{20}}, x_i, q_i\right)}{f\left(\mathbf{w}_i\right)}\right].$$

From the strict Jensen's inequality, $E\left[\ln\frac{\frac{\sigma_{20}}{\sigma_{10}}f\left(\frac{\sigma_{20}e_i - x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{10}}, x_i, q_i\right)}{f\left(\mathbf{w}_i\right)}\right] < 0$, and $E\left[\ln\frac{\frac{\sigma_{10}}{\sigma_{20}}f\left(\frac{\sigma_{10}e_i + x_i'\left(\beta_{10}-\beta_{20}\right)}{\sigma_{20}}, x_i, q_i\right)}{f\left(\mathbf{w}_i\right)}\right] < 0$, so there exists $0 < C < 1$ such that we have the Equation in Box III. In summary, $\exp\left\{\widehat{\omega}(v)\right\} \le \exp\left\{C\omega(v)\right\}$ for some $0 < C < 1$ for all $|v| > M$ with probability approaching 1. Combining (18) and (20), the proof is complete. ■

Remark 9. Checking the proof of Theorem 1, the n-consistency of $\widetilde{\gamma}$ is only used in (13). If $\widetilde{\gamma}$ has a slower convergence rate such as the SLSE in Seo and Linton (2007), the SEB procedure will improve the convergence rate of $\widetilde{\gamma}$ to n. Of course, Assumption B should be adjusted correspondingly. Also, the $\sqrt{n}$-consistency of $\widehat{\beta}$ is only used to show $\max_{1\le i\le n} x_i'\left(\widehat{\beta}_\ell - \beta_{\ell 0}\right) = O_p\left(n^{-1/4}\right)$, $\ell = 1, 2$. If the conditional mean of y in each regime takes a nonparametric form $g_\ell(x)$, and we can find an estimator of $g_\ell(x)$, say $\widehat{g}_\ell(x)$, such that $\sup_{x\in\mathbb{W}_n^x}\left|\widehat{g}_\ell(x) - g_\ell(x)\right| = O_p\left(n^{-\zeta}\right)$ for some $\zeta > 0$, where $\mathbb{W}_n^x = \{x|\mathbf{w}\in\mathbb{W}_n\}$, then by adjusting Assumption B, our proof can still go through.

Proof of Theorem 2. This proof is inspired by Theorem 2 of Chernozhukov and Hong (2003). Consider the objective function

$$\widehat{Q}_n(t) = \int_{V_n} l_2(t-v)\widehat{p}_n^*(v)dv,$$

which is minimized at $n\left(\widehat{\gamma}-\gamma_0\right)$. Define

$$Q_n(t) = \int_{\mathbb{R}} l_2(t-v)p_n^*(v)dv$$

which is uniquely minimized at Z_n by Assumption L. We first prove for each fixed t,

$$\widehat{Q}_n(t) - Q_n(t) \xrightarrow{p} 0.$$

From Assumption L, $l_2(v) \le 1 + |v|^p$ and by $|a+b|^p \le 2^{p-1}|a|^p + 2^{p-1}|b|^p$ for $p \ge 1$:

$$\left|\widehat{Q}_n(t) - Q_n(t)\right| \le \int_{V_n}\left(1+|t-v|^p\right)\left|\widehat{p}_n^*(v) - p_n^*(v)\right| dv \le \int_{V_n}\left(1+2^{p-1}|t|^p + 2^{p-1}|v|^p\right)\left|\widehat{p}_n^*(v) - p_n^*(v)\right| dv = o_p(1)$$

where the last equality is from Theorem 1.

Now, note that $\widehat{Q}_n(t)$ and $Q_n(t)$ are convex and finite, and $Z_n = \arg\min_{t\in\mathbb{R}}\int_{\mathbb{R}} l_2(t-v)p_n^*(v)dv = O_p(1)$ from Yu (2012). By the convexity lemma of Pollard (1991), pointwise convergence entails the uniform convergence over compact sets K:

$$\sup_{t\in K}\left|\widehat{Q}_n(t) - Q_n(t)\right| \xrightarrow{p} 0.$$

By a similar argument on page 331 of Chernozhukov and Hong (2003), we can conclude that

$$n\left(\widehat{\gamma}-\gamma_0\right) - Z_n = o_p(1),$$

$$
\begin{aligned}
\widehat{\omega}(v) &= \sum_{i=1}^{n} \ln \frac{\frac{\widetilde{\sigma}_2}{\widetilde{\sigma}_1}\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'(\widetilde{\beta}_1-\beta_{20})}{\widetilde{\sigma}_1}, x_i, q_i\right)}{\widehat{f}\left(\frac{\sigma_{20}e_i - x_i'(\widetilde{\beta}_2-\beta_{20})}{\widetilde{\sigma}_2}, x_i, q_i\right)} \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{n}\right) + \sum_{i=1}^{n} \ln \frac{\frac{\widetilde{\sigma}_1}{\widetilde{\sigma}_2}\widehat{f}\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\widetilde{\beta}_2)}{\widetilde{\sigma}_2}, x_i, q_i\right)}{\widehat{f}\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\widetilde{\beta}_1)}{\widetilde{\sigma}_1}, x_i, q_i\right)} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \le \gamma_0\right) \\
&\overset{(1)}{\le} \sum_{i=1}^{n} \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i'(\widetilde{\beta}_1-\beta_{20})}{\widetilde{\sigma}_1}, x_i, q_i\right) + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f\left(\frac{\sigma_{20}e_i - x_i'(\widetilde{\beta}_2-\beta_{20})}{\widetilde{\sigma}_2}, x_i, q_i\right) - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{n}\right) \mathbf{1}(\mathbf{w}_i \in \mathbb{W}_n) \\
&\quad + \sum_{i=1}^{n} \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\widetilde{\beta}_2)}{\widetilde{\sigma}_2}, x_i, q_i\right) + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\widetilde{\beta}_1)}{\widetilde{\sigma}_1}, x_i, q_i\right) - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \le \gamma_0\right) \mathbf{1}(\mathbf{w}_i \in \mathbb{W}_n) \\
&\overset{(2)}{\le} \sum_{i=1}^{n} \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i'(\beta_{10}-\beta_{20})}{\sigma_{10}}, x_i, q_i\right) + Cn^{\nu-1/2} + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f(\mathbf{w}_i) - Cn^{\nu-1/2} - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{n}\right) \mathbf{1}(\mathbf{w}_i \in \mathbb{W}_n) \\
&\quad + \sum_{i=1}^{n} \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\beta_{20})}{\sigma_{20}}, x_i, q_i\right) + Cn^{\nu-1/2} + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f(\mathbf{w}_i) - Cn^{\nu-1/2} - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \le \gamma_0\right) \mathbf{1}(\mathbf{w}_i \in \mathbb{W}_n). \qquad (21)
\end{aligned}
$$

Here, (1) is from the uniform convergence in (14) and the assumption that $n \cdot P(\mathbf{w}_i \notin \mathbb{W}_n) \to 0$, and (2) is from a similar analysis as in (19).

Box II.

$$
\begin{aligned}
&P\left(\sum_{i=1}^{n} \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i'(\beta_{10}-\beta_{20})}{\sigma_{10}}, x_i, q_i\right) + Cn^{\nu-1/2} + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f(\mathbf{w}_i) - Cn^{\nu-1/2} - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{n}\right) \mathbf{1}(\mathbf{w}_i \in \mathbb{W}_n)\right. \\
&\qquad \left. \le C \sum_{i=1}^{n} \ln \frac{\frac{\sigma_{20}}{\sigma_{10}} f\left(\frac{\sigma_{20}e_i - x_i'(\beta_{10}-\beta_{20})}{\sigma_{10}}, x_i, q_i\right)}{f(\mathbf{w}_i)} \mathbf{1}(\gamma_0 < q_i \le \gamma_0 + \frac{v}{n})\right) \longrightarrow 1, \\
&P\left(\sum_{i=1}^{n} \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\beta_{20})}{\sigma_{20}}, x_i, q_i\right) + Cn^{\nu-1/2} + Cn^{-(1-\eta)/2}h^{-(k+2)/2}}{f(\mathbf{w}_i) - Cn^{\nu-1/2} - Cn^{-(1-\eta)/2}h^{-(k+2)/2}} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \le \gamma_0\right) \mathbf{1}(\mathbf{w}_i \in \mathbb{W}_n)\right. \\
&\qquad \left. \le C \sum_{i=1}^{n} \ln \frac{\frac{\sigma_{10}}{\sigma_{20}} f\left(\frac{\sigma_{10}e_i + x_i'(\beta_{10}-\beta_{20})}{\sigma_{20}}, x_i, q_i\right)}{f(\mathbf{w}_i)} \mathbf{1}\left(\gamma_0 + \frac{v}{n} < q_i \le \gamma_0\right)\right) \longrightarrow 1
\end{aligned}
$$

by Assumption D4.

Box III.

which in turn implies that $n(\widehat{\gamma} - \gamma_0) \xrightarrow{d} Z$ since $Z_n \xrightarrow{d} Z$ from Yu (2012) when Assumption E holds. ■

Proof of Theorem 3. Under the assumption that $\sup_{|\gamma-\gamma_0|<\delta} n|\widetilde{\gamma} - \gamma| = O_p(1)$ for some $\delta > 0$, we can show that

$$\sup_{|\gamma-\gamma_0|<\delta} \left\|\widehat{p}_n^* - p_n^*\right\|_{TVM(\kappa)} \xrightarrow{p} 0$$

by following the lines of Theorem 1, where $\widehat{p}_n^*(v)$ is redefined as $\frac{1}{n}\widehat{p}_n\left(\gamma + \frac{v}{n}\right)$, and $p_n^*(v)$ is redefined as $\frac{1}{n}p_n\left(\gamma + \frac{v}{n}\right)$. Similarly, from the proof of Theorem 2, we can show that

$$\sup_{|\gamma-\gamma_0|<\delta} |n(\widehat{\gamma} - \gamma) - Z_n| \xrightarrow{p} 0. \qquad (22)$$

From the proof of Theorem 1, the tail of $\widehat{p}_n^*(v)$ is exponentially decaying. Using the usual "shelling" approach (see, e.g., Lemma 9 of Hirano and Porter (2003)), we can show that for any N, uniformly in γ such that $|\gamma - \gamma_0| < \delta$,

$$\lim_{n,H\to\infty} H^N P_\gamma(|n(\widehat{\gamma} - \gamma)| > H) = 0,$$

where $P_\gamma(\cdot)$ is similarly defined as $E_\gamma[\cdot]$. Since l_2 has a polynomial majorant, $\underline{\lim}_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma[l_2(n(\widehat{\gamma} - \gamma))]\right] < \infty$. By the dominating arguments, (22) implies that

$$
\begin{aligned}
&\underline{\lim}_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma[l_2(n(\widehat{\gamma} - \gamma))]\right] \\
&= \underline{\lim}_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma[l_2(Z_n)]\right]. \qquad (23)
\end{aligned}
$$

Now, from Theorem 3(iv) of Yu (2012),

$$
\begin{aligned}
&\underline{\lim}_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma[l_2(Z_n)]\right] \\
&\le \underline{\lim}_{n\to\infty}\left[\sup_{|\gamma-\gamma_0|<\delta} E_\gamma[l_2(n(T_n - \gamma))]\right] \qquad (24)
\end{aligned}
$$

with T_n being any estimator of γ. Combining (23) and (24) and letting δ shrink to zero, the proof is complete. ■

Proof of Theorem 4. This proof is inspired by Theorem 3 of Chernozhukov and Hong (2003). Evaluate $\widehat{F}_n(x)$ at $x = \gamma_0 + t/n$ and change the variable of integration to get:

$$\widehat{H}_n(t) \equiv \widehat{F}_n(\gamma_0 + t/n) = \int_{v\in V_n : v\le t} \widehat{p}_n^*(v)dv.$$

Table 1
Estimator performance for γ with four β_1 values (based on 1000 repetitions).

β_1 Values →	$\beta_1 = 0.2$		$\beta_1 = 0.4$		$\beta_1 = 0.7$		$\beta_1 = 1$	
Estimators ↓ Risk($\times 10^{-2}$) →	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD
$n = 100$								
LLSE	12.502	7.757	4.083	2.401	1.960	1.205	1.536	1.050
MLSE	12.565	7.825	4.177	2.348	1.890	0.995	1.059	0.676
SLSE	12.184	9.343	5.865	4.634	3.874	3.086	2.930	2.373
SEB Post Mean	10.945	6.095	3.361	1.884	1.330	0.825	0.897	0.617
Recursive SPMean_1	10.705	5.838	3.338	1.874	1.222	0.803	0.894	0.620
Recursive SPMean_2	10.555	5.689	3.333	1.869	1.195	0.798	0.892	0.619
SEB Post Median	11.352	6.242	3.489	1.865	1.420	0.798	0.913	0.612
Recursive SPMedian_1	11.258	6.083	3.432	1.848	1.387	0.786	0.907	0.609
Recursive SPMedian_2	11.228	6.015	3.426	1.844	1.289	0.773	0.907	0.609
Par Post Mean	4.033	2.702	2.310	1.519	1.066	0.747	0.859	0.594
Par Post Median	4.284	2.672	2.468	1.475	1.085	0.718	0.861	0.588
$n = 400$								
LLSE	3.936	2.156	1.080	0.628	0.476	0.310	0.385	0.267
MLSE	3.982	2.186	1.110	0.606	0.439	0.253	0.261	0.172
SLSE	5.160	4.075	2.526	2.012	1.636	1.301	1.302	1.042
SEB Post Mean	1.765	0.891	0.631	0.415	0.313	0.207	0.219	0.157
Recursive SPMean_1	1.414	0.772	0.613	0.408	0.312	0.206	0.220	0.157
Recursive SPMean_2	1.367	0.752	0.608	0.406	0.311	0.206	0.220	0.157
SEB Post Median	1.803	0.850	0.661	0.406	0.324	0.203	0.219	0.155
Recursive SPMedian_1	1.617	0.770	0.646	0.399	0.323	0.201	0.219	0.155
Recursive SPMedian_2	1.558	0.754	0.646	0.398	0.323	0.201	0.219	0.155
Par Post Mean	0.974	0.642	0.578	0.383	0.302	0.199	0.218	0.156
Par Post Median	1.008	0.598	0.600	0.373	0.313	0.198	0.219	0.154
$n = \infty$ Risk →	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD
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
SLSE	1.774	1.416	0.929	0.741	0.591	0.472	0.472	0.376
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

Table 2
Comparison of inference methods: coverage and average length of the nominal 95% confidence intervals for γ with four β_1 values (based on 1000 repetitions).

β_1 Values →	$\beta_1 = 0.2$		$\beta_1 = 0.4$		$\beta_1 = 0.7$		$\beta_1 = 1$	
CIs ↓ Cov and Leng($\times 10^{-2}$) →	Cov	Length	Cov	Length	Cov	Length	Cov	Length
$n = 100$								
Hansen (2000)	0.923	36.422	0.940	11.473	0.957	4.355	0.972	2.792
Bai (1997) LLSE	0.959	55.266	0.759	13.020	0.581	4.512	0.356	2.211
Bai (1997) MLSE	0.961		0.876		0.854		0.717	
Bootstrap SLSE (ET)	0.477	36.061	0.705	26.579	0.595	13.073	0.661	13.357
Bootstrap SLSE (S)	0.892	47.547	1.000	33.486	0.999	18.870	1.000	17.845
Smoothed Bootstrap LLSE (ET)	0.911	54.073	0.903	17.652	0.886	7.959	0.851	5.683
Smoothed Bootstrap LLSE (S)	0.969	66.975	0.955	20.428	0.971	8.478	0.953	6.029
Smoothed Bootstrap MLSE (ET)	0.908	53.044	0.936	17.089	0.957	6.661	0.934	4.317
Smoothed Bootstrap MLSE (S)	0.958	65.877	0.950	20.042	0.959	7.388	0.943	4.395
Subsampling LLSE (ET)	0.716	19.735	0.882	12.440	0.911	6.864	0.879	4.965
Subsampling LLSE (S)	0.762	22.539	0.920	14.233	0.944	7.065	0.913	5.406
Subsampling MLSE (ET)	0.693	18.605	0.888	11.676	0.935	5.827	0.907	3.814
Subsampling MLSE (S)	0.745	21.407	0.908	14.181	0.943	6.654	0.929	4.034
Nonpar Post Interval	0.808	13.136	0.926	7.104	0.950	3.701	0.944	2.577
Recursive NPI _1_Mean	0.803	12.197	0.926	6.991	0.949	3.658	0.944	2.573
Recursive NPI _2_Mean	0.805	11.841	0.926	6.948	0.949	3.630	0.944	2.573
Recursive NPI _1_Med	0.800	11.983	0.922	6.916	0.950	3.618	0.943	2.568
Recursive NPI _2_Med	0.804	11.766	0.922	6.879	0.950	3.604	0.943	2.568
Par Post Interval	0.939	12.244	0.947	7.000	0.950	3.468	0.947	2.446
$n = 400$								
Hansen (2000)	0.931	10.510	0.939	2.868	0.963	1.116	0.984	0.699
Bai (1997) LLSE	0.933	13.817	0.832	3.454	0.585	1.128	0.346	0.553
Bai (1997) MLSE	0.936		0.903		0.844		0.694	
Bootstrap SLSE (ET)	0.514	16.067	0.599	9.097	0.414	6.061	0.636	5.815
Bootstrap SLSE (S)	0.954	24.389	0.996	11.785	0.987	12.062	1.000	8.541
Smoothed Bootstrap LLSE (ET)	0.926	15.761	0.902	4.264	0.877	1.989	0.856	1.443
Smoothed Bootstrap LLSE (S)	0.942	19.074	0.956	4.916	0.967	2.112	0.948	1.527
Smoothed Bootstrap MLSE (ET)	0.935	15.731	0.946	4.115	0.959	1.662	0.938	1.088
Smoothed Bootstrap MLSE (S)	0.941	19.044	0.952	4.800	0.942	1.835	0.953	1.099
Subsampling LLSE (ET)	0.893	11.244	0.910	3.980	0.914	1.843	0.905	1.331
Subsampling LLSE (S)	0.914	13.958	0.950	4.448	0.948	1.860	0.920	1.425
Subsampling MLSE (ET)	0.895	11.150	0.930	3.871	0.936	1.568	0.911	1.015

(continued on next page)

Table 2 (*continued*)

β_1 Values →	$\beta_1 = 0.2$		$\beta_1 = 0.4$		$\beta_1 = 0.7$		$\beta_1 = 1$	
CIs ↓ Cov and Leng($\times 10^{-2}$) →	Cov	Length	Cov	Length	Cov	Length	Cov	Length
Subsampling MLSE (S)	0.913	14.003	0.946	4.567	0.939	1.768	0.941	1.065
Nonpar Post Interval	0.919	3.535	0.940	1.789	0.950	0.918	0.952	0.641
Recursive NPI _1_Mean	0.914	3.141	0.940	1.745	0.950	0.908	0.954	0.639
Recursive NPI _2_Mean	0.911	3.056	0.940	1.731	0.950	0.904	0.954	0.639
Recursive NPI _1_Med	0.912	3.110	0.942	1.732	0.950	0.900	0.952	0.638
Recursive NPI _2_Med	0.905	3.057	0.942	1.725	0.950	0.898	0.952	0.638
Par Post Interval	0.940	3.134	0.944	1.715	0.942	0.871	0.950	0.615

Table 3
Frequency of sample splitting coincidence of recursive SEB method with previous recursion and the true value (based on 1000 repetitions).

β_1 Values →	$\beta_1 = 0.2$				$\beta_1 = 0.4$			
n →	100		400		100		400	
Frequency (%) ↓	Mean	Med	Mean	Med	Mean	Med	Mean	Med
LSE	16.6		15.9		43.9		46.0	
SEB	17.3 / 25.1	25.7 / 28.8	14.6 / 28.1	21.5 / 37.7	39.9 / 45.9	55.0 / 52.9	40.6 / 47.7	56.3 / 56.2
Recursive SEB_1	70.7 / 26.0	77.3 / 31.2	71.8 / 29.3	78.3 / 39.3	91.3 / 46.9	93.5 / 54.3	92.8 / 49.1	95.2 / 56.4
Recursive SEB_2	89.1 / 27.3	93.6 / 32.4	89.8 / 30.0	92.0 / 40.1	97.2 / 47.3	97.9 / 54.5	97.0 / 49.1	98.4 / 56.8
β_1 Values →	$\beta_1 = 0.7$				$\beta_1 = 1$			
n →	100		400		100		400	
Frequency (%) ↓	Mean	Med	Mean	Med	Mean	Med	Mean	Med
LSE	73.0		71.2		87.3		86.3	
SEB	70.8 / 73.2	83.1 / 78.4	70.3 / 75.4	79.8 / 80.0	86.1 / 88.0	91.5 / 91.3	85.5 / 89.8	89.5 / 92.1
Recursive SEB_1	97.6 / 73.7	98.3 / 79.4	98.1 / 76.3	99.4 / 80.1	99.7 / 88.1	99.8 / 91.3	99.8 / 90.0	99.9 / 92.0
Recursive SEB_2	99.4 / 74.1	99.7 / 79.2	99.4 / 75.8	99.8 / 80.3	100 / 88.1	100 / 91.3	99.9 / 89.9	100 / 92.0

Table 4
Comparison of estimators and confidence intervals in the economic growth model.

Estimators	$\widehat{\gamma}$	95% CIs	Length of CIs	Ratio of countries covered by CIs
Hansen (2000)	863	[594, 1794]	1 200	40/96
Bai (1997) MLSE	871	[759, 890]	131	8/96
Bootstrap SLSE (ET)	878	[−36 801, 8 251]	45 052	88/96
Bootstrap SLSE (S)		[−23 095, 24 851]	47 947	96/96
Subsampling LLSE (ET)	863	[−459, 885]	1 344	19/96
Subsampling LLSE (S)		[−245, 1971]	2 216	51/96
Subsampling MLSE (ET)	871	[−780, 879]	1 659	19/96
Subsampling MLSE (S)		[−517, 2259]	2 776	56/96
SEB Post Mean	831	[757, 878]	121	6/96
SEB Post Med	836			
Recursive SPMean_1	801	[747, 843]	96	5/96
Recursive SPMedian_1	812	[718, 875]	157	8/96
Recursive SPMedian_2	805	[747, 843]	96	5/96

Define

$$H_n(t) = \int_{v\in V_n: v\le t} p_n^*(v)dv, \qquad H_\infty(t) = \int_{v\in\mathbb{R}: v\le t} p_2^*(v)dv,$$

then by Yu (2012), $\sup_t |H_n(t) - H_\infty(t)| \xrightarrow{p} 0$. By the definition of total variation of moments norm and Theorem 1, $\sup_t \left|\widehat{H}_n(t) - H_n(t)\right| \xrightarrow{p} 0$, so $\sup_t \left|\widehat{H}_n(t) - H_\infty(t)\right| \xrightarrow{p} 0$, where the sup is taken over the support of $\widehat{H}_n(t)$. By some elementary analysis, we can show that for any sequence of cdfs F_n, if $\sup_t |F_n(t) - F(t)| \to 0$ as $n \to \infty$, and $F'(t) > 0$ in a neighborhood of $F^{-1}(\tau)$, then $\left|F_n^{-1}(\tau) - F^{-1}(\tau)\right| \to 0$ as $n \to \infty$. From Yu (2012), $H'_\infty(t) > 0$ for any $t \in \mathbb{R}$ almost surely, so combining with $\sup_t \left|\widehat{H}_n(t) - H_\infty(t)\right| \xrightarrow{p} 0$, we have

$$\left|\widehat{H}_n^{-1}(\tau) - H_\infty^{-1}(\tau)\right| \xrightarrow{p} 0.$$

By the equivariance of quantile with respect to monotone transformations, $\widehat{H}_n^{-1}(\tau) = n\,(c_n(\tau) - \gamma_0)$. By Theorem 1 and the definition of weak convergence,

$$\begin{aligned}\lim_{n\to\infty} P\left(c_n(\tau) \le \gamma_0\right) &= \lim_{n\to\infty} P\left(\widehat{H}_n^{-1}(\tau) \le 0\right)\\ &= P\left(H_\infty^{-1}(\tau) \le 0\right) = \tau.\end{aligned}$$

Here, the second equality is because 0 is assumed to be a continuity point of the distribution of $H_\infty^{-1}(\tau)$, and the last equality is from the optimality of the Bayes estimator (see the proof of Theorem 3.3 in Chernozhukov and Hong (2004)). ■

Appendix C. Tables in simulations and application

First, the notations used in the following tables are collected below for reference.

LLSE/MLSE: Least squares estimator using the left endpoint/middle point of the minimizing interval for γ.

SLSE: Smoothed least squares estimator

SEB Post Mean/Median: Posterior mean/median in the SEB estimation of γ.

Par Post Mean/Median: Posterior mean/median in the parametric estimation of γ.

(ET)/(S): Equal-tailed/Symmetric confidence interval.

Nonpar Post Interval: Nonparametric posterior interval.

Par Post Interval: Parametric posterior interval.

Frequency: The numerator is the frequency of the sample splitting coincidence with the last recursion, compared with the LSE in the first recursion, and the denominator is the frequency of the sample splitting coincidence with the true value.

Recursive SPMean_k: Recursive SEB posterior mean in the kth recursion, $k = 1, 2$.

Recursive SPMedian_k: Recursive SEB posterior median in the kth recursion, $k = 1, 2$.

Recursive NPI_k: Recursive nonparametric posterior interval in the kth recursion, $k = 1, 2$.

Appendix D. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2013.09.002.

References

Ai, C., 1997. A semiparametric maximum likelihood estimator. Econometrica 933–963.

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. Rev. Econom. Statist. 79, 551–563.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, second ed. Springer-Verlag, New York.

Bickel, P.J., 1982. On adaptive estimation. Ann. Statist. 10, 647–671.

Bickel, P.J., et al., 1998. Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York.

Caner, M., 2002. A note on least absolute deviation estimation of a threshold model. Econometric Theory 18, 800–814.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. J. Econometrics. 34, 305–334.

Chan, K.S., 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. Ann. Statist. 21, 520–533.

Chen, X., Reiss, M., 2011. On rate optimality for ill-posed inverse problems in econometrics. Econometric Theory 27, 497–521.

Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. J. Econometrics. 115, 293–346.

Chernozhukov, V., Hong, H., 2004. Likelihood estimation and inference in a class of nonregular econometric models. Econometrica 72, 1445–1480.

Diaconis, P., Freedman, D., 1986a. On the consistency of Bayes estimates. Ann. Statist. 14, 1–26.

Diaconis, P., Freedman, D., 1986b. On inconsistent Bayes estimates of location. Ann. Statist. 14, 68–87.

Durlauf, S.N., Johnson, P.A., 1995. Multiple regimes and cross-country growth behavior. J. Appl. Econometrics. 10, 365–384.

Fan, J., 1993. Local linear regression smoothers and their minimax efficiency. Ann. Statist. 21, 196–216.

Ghosal, S., 2010. In: Hjort, N.L., et al. (Eds.), Dirichlet Process, Related Priors and Posterior Asymptotics, in Bayesian Nonparametrics. Cambridge University Press, New York, pp. 35–79. (Chapter 2).

Gijbels, I., et al., 2004. Interval and band estimation for curves with jumps. J. Appl. Probab. 41A, 65–79. Special issue "Stochastic Methods and Their Applications", in honor of Chris Heyde.

Gonzalo, J., Wolf, M., 2005. Subsampling inference in threshold autoregressive models. J. Econometrics. 127, 201–224.

Hall, A.R., et al., 2012. Inference regarding multiple structural changes in linear models with endogenous regressors. J. Econometrics. 170, 281–302.

Hansen, B.E., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64, 413–430.

Hansen, B.E., 2000. Sample splitting and threshold estimation. Econometrica 68, 575–603.

Hansen, B.E., 2011. Threshold autoregression in economics. Stat. Interface 4, 123–127.

Hirano, K., 2002. Semiparametric Bayesian inference in autoregressive panel data models. Econometrica 781–799.

Hirano, K., Porter, J.R., 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. Econometrica 71, 1307–1338.

Ibragimov, I., Has'minskii, R., 1981. Statistical Estimation: Asymptotic Theory. Springer-Verlag, New York.

Kosorok, M.R., Song, R., 2007. Inference under right censoring for transformation models with a change-point based on a covariate threshold. Ann. Statist. 35, 957–989.

Li, D., Ling, S.Q., 2012. On the least squares estimation of multiple-regime threshold autoregressive models. J. Econometrics. 167, 240–253.

Neal, R.M., 2003. Slice sampling. Ann. Statist. 31, 705–767.

Newey, W.K., 1990. Semiparametric efficiency bounds. J. Appl. Econometrics. 5, 99–135.

Oka, T., Qu, Z., 2011. Estimating structural changes in regression quantiles. J. Econometrics. 162, 248–267.

Politis, D., Romano, J., 1994. Large sample confidence regions based on subsamples under minimal assumptions. Ann. Statist. 22, 2031–2050.

Pollard, D., 1991. Asymptotic for least absolute deviations regression estimator. Econometric Theory 7, 186–199.

Powell, J.L., et al., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

Qu, Z., Perron, P., 2007. Estimating and testing structural changes in multivariate regressions. Econometrica 459–502.

Ritov, Y., Bickel, P.J., 1990. Achieving information bounds in non and semiparametric models. Ann. Statist. 18, 925–938.

Robinson, P.M., 1987. Asymptoticlly efficient estimation in the presence of heteroskedasticity of unknown form. Econometrica 55, 875–891.

Seijo, E., Sen, B., 2011. Change-point in stochastic design regression and the bootstrap. Ann. Statist. 39, 1580–1607.

Seo, M.H., Linton, O., 2007. A smoothed least squares estimator for threshold regression models. J. Econometrics. 141, 704–735.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Stoker, T.M., 1986. Consistent estimation of scaled coefficients. Econometrica 1461–1481.

Tong, H., 1978. On a threshold model. In: Chen, C.H. (Ed.), Pattern Recognition and Signal Processing. Sijthoff & Noordhoff, Amsterdam, pp. 575–586.

Tong, H., 1983. Threshold Models in Nonlinear Time Series analysis. In: Lecture Notes in Statistics. Springer-Verlag, New York.

Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Clarendon Press, Oxford.

Tong, H., 2011. Threshold models in time series analysis - 30 years on. Stat. Interface. 4, 107–118.

Tong, H., Lim, K.S., 1980. Threshold autoregression, limit cycles and cyclical data. J. R. Stat. Soc. Ser. B 42, 245–292.

Van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, New York.

Yu, P., 2012. Likelihood estimation and inference in threshold regression. J. Econometrics. 167, 274–294.

Yu, P., 2013a. Threshold Regression with Endogeneity, Mimeo, Department of Economics, University of Auckland.

Yu, P., 2013b. Integrated Quantile Threshold Regression and Distributional Threshold Effects, Mimeo, Department of Economics, University of Auckland.

Yu, P., 2014. The bootstrap in threshold regression. Econometric Theory 30, 676–714.