Submitted to the Annals of Statistics

FACTOR-DRIVEN TWO-REGIME REGRESSION*

By Sokbae Lee, Yuan Liao, Myung Hwan Seo and Youngki Shin

Columbia University, Rutgers University, Seoul National University and McMaster University

We propose a novel two-regime regression model where regime switching is driven by a vector of possibly unobservable factors. When the factors are latent, we estimate them by the principal component analysis of a panel data set. We show that the optimization problem can be reformulated as mixed integer optimization, and we present two alternative computational algorithms. We derive the asymptotic distribution of the resulting estimator under the scheme that the threshold effect shrinks to zero. In particular, we establish a phase transition that describes the effect of first-stage factor estimation as the cross-sectional dimension of panel data increases relative to the time-series dimension. Moreover, we develop bootstrap inference and illustrate our methods via numerical studies.

1. Introduction.

Suppose that y_t is generated from

$$y_t = x_t'\beta_0 + x_t'\delta_0 1\{f_t'\gamma_0 > 0\} + \varepsilon_t, \tag{1.1}$$

$$\mathbb{E}\left(\varepsilon_t | \mathcal{F}_{t-1}\right) = 0,\ t = 1, \ldots, T, \tag{1.2}$$

where x_t and f_t are adapted to the filtration $\mathcal{F}_{t-1}$, $(\beta_0, \delta_0, \gamma_0)$ is a vector of unknown parameters, and the unobserved random variable ε_t satisfies the conditional mean restriction in (1.2). We interpret f_t to be a vector of factors determining regime switching. When $f_t'\gamma_0 > 0$, the regression function becomes $x_t'(\beta_0 + \delta_0)$; if $f_t'\gamma_0 \leq 0$, it reduces to $x_t'\beta_0$. We allow for either observable or unobservable factors. For the latter, we assume that they can be recovered from a panel data set. In light of this feature, we call the model in (1.1) and (1.2) a *factor-driven two-regime regression model*.

*We would like to thank an associate editor and two anonymous referees for helpful comments. We would like to thank the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01033487), the Social Sciences and Humanities Research Council of Canada (SSHRC-435-2018-0275), the European Research Council for financial support (ERC-2014-CoG-646917-ROMIA) and the UK Economic and Social Research Council for research grant (ES/P008909/1) to the CeMMAP.

MSC 2010 subject classifications: Primary 62H25; secondary 62F12

Keywords and phrases: threshold regression, principal component analysis, mixed integer optimization, phase transition, oracle properties

Our paper is closely related to the literature on threshold models with unknown change points (see, e.g., [9], [16], [20], [25], [26], and [28], among many others). In the conventional threshold regression model, an intercept term and a scalar observed random variable constitute f_t. For instance, Chan [9] and Hansen [16] studied the model in which $1\{f_t'\gamma_0 > 0\}$ in (1.1) is replaced by $1\{q_t > \widetilde{\gamma}_0\}$ for some observable scalar variable q_t with a scalar unknown parameter $\widetilde{\gamma}_0$. In practice, it might be controversial to choose which observed variable plays the role of q_t. For example, if the two different regimes represent the status of two environments of the population, arguably it is difficult to assume that the change of the environment is governed by just a single variable. On the contrary, our proposed model introduces a regime change due to a single index of factors that can be "learned" from a potentially much larger dataset. Specifically, we consider the framework of latent approximate factor models in order to model a regime switch based on a potentially large number of covariates.

In view of the conditional mean restriction in (1.2), a natural strategy to estimate $(\beta_0, \delta_0, \gamma_0)$ is to rely on least squares. A least-squares estimator for our model brings new challenges in terms of both computation and asymptotic theory. First of all, when the dimension of f_t is larger than 2, it is computationally demanding to estimate $(\beta_0, \delta_0, \gamma_0)$. We overcome this difficulty by developing new computational algorithms based on the method of mixed integer optimization (MIO). See, for example, section 2.1 in Bertsimas et al. [8] for a discussion on computational advances in solving the MIO problems.

Second, we establish asymptotic properties of our proposed estimator by adopting a diminishing thresholding effect. That is, we assume that $\delta_0 = T^{-\varphi}d_0$ for some unknown $\varphi \in (0, 1/2)$ and unknown non-diminishing vector d_0. The diminishing threshold has been one of the standard frameworks in the change point literature (e.g., [2, 17, 18]). The unknown parameter φ reflects the difficulty of estimating γ_0 and affects the identification and estimation of the change-point γ_0. Both the rate of convergence and the asymptotic distribution depend on φ. This is a widely employed tool to allow for flexible signal strengths of the parameters in the nonlinear model. For instance, McKeague and Sen [22] studied a "point impact" linear model, where the identification and estimation of γ_0 are affected by an unknown slope δ_0. While specifically assuming $\delta_0 \neq 0$, they encountered a similar parameter φ, reflecting the difficulty of estimating γ_0. The asymptotic theory for the estimated δ_0 under the diminishing jump setting is fundamentally different from the fixed jump setting: the former is determined by a Gaussian process (e.g., [16]), and the latter by a compound poison process (e.g., [9]).

While both settings lead to important asymptotic implications, we focus on the diminishing setting because when the factors are estimated, there is a new and interesting *phase transition* phenomenon that smoothly appears in the "bias" term of the Gaussian process. The phase transition characterizes the continuous change of the asymptotic distribution as the precision of the estimated factors increases relative to the size of the jump, which we shall detail below.

When the factor f_t is latent, we estimate it using principal component analysis (PCA) from a potentially much larger dataset, whose dimension is N. It turns out that the asymptotic distribution for the estimator of $\alpha_0 \equiv (\beta_0', \delta_0')'$ is identical to that when γ_0 were known, regardless of whether factors are directly observable or not; therefore, the estimator of α_0 enjoys an oracle property.

The issue is more sophisticated for the distribution of the estimator of γ_0. When factors are directly observable, we prove that

$$T^{1-2\varphi}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \operatorname*{argmin}_{g \in \mathcal{G}} B(g) + 2W(g),$$

where $B(g)$ represents a "drift function" of the criterion function, which is linear with a kink at zero, $W(g)$ is a mean-zero Gaussian process and $\mathcal{G}$ is a rescaled parameter space. However, when factors are not directly observable, the estimation error from the PCA plays an essential role and may slow down the rates of convergence, depending on the relation between N and T. Specifically, we show that

$$\left(\left(NT^{1-2\varphi}\right)^{1/3} \wedge T^{1-2\varphi}\right)(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \operatorname*{argmin}_{g \in \mathcal{G}} A(\omega, g) + 2W(g),$$

with a new drift function $A(\omega, g)$ that depends on $\omega = \lim \sqrt{N}T^{-(1-2\varphi)} \in [0, \infty]$. On one hand, when $\omega = \infty$, we find that $A(\omega, g) = B(g)$, so the limiting distribution becomes the same as if the factors were observable. This case corresponds to the *super-consistency rate* (e.g., [16]). On the other hand, when $\omega = 0$, it turns out that $A(\omega, g)$ is quadratic in g, corresponding to a *cube root rate* similar to the maximum score estimator (e.g., [19, 27]). Furthermore, both the drift function and the resulting rates of convergence have continuous transitions as ω changes between 0 and ∞. Therefore, one of our key findings for the estimator of γ_0 is the occurrence of a phase transition from a *weak-oracle* limiting distribution to a *semi-strong* oracle one, and then to a *strong* oracle one as ω increases.

As the asymptotic distribution of $\widehat{\gamma}$ is non-pivotal, we propose a wild bootstrap for inference of γ_0. Importantly, we construct bootstrap confi-

dence intervals for γ_0 that do not require knowledge of φ. This facilitates applications in which the jump diminishing speed is not known in advance.

The remainder of the paper is organized as follows. In Section 2, we propose the least-squares estimator and algorithms to compute the proposed estimator. In Section 3, we establish asymptotic theory when f_t is directly observed. In Section 4, we consider estimation when f_t is a vector of latent factors, we propose a two-step estimator via PCA, and we analyze asymptotic properties of our proposed estimator. In Section 5, we develop bootstrap inference, and in Section 6 we give the results of Monte Carlo experiments. In Section 7, we illustrate our methods by applying them to threshold autoregressive models of unemployment. We conclude in Section 8. The online appendices provides details that are omitted from the main text.

The notation used in the paper is as follows. The sample size is denoted by T and the transpose of a matrix is denoted by a prime. The true parameter is denoted by the subscript 0, whereas a generic element has no subscript. The Euclidean norm is denoted by $|\cdot|_2$, the Frobenius norm of a matrix is $|\cdot|_F$, the spectral norm of a matrix is $|\cdot|_2$, and the ℓ_0-norm is $|\cdot|_0$. For a generic random variable or vector z_t, let its density function be denoted by p_{z_t}. Similarly, let $p_{y_t|x_t}(y)$ denote the conditional density of y_t given x_t for the random vectors y_t and x_t. The abbreviation *a.s.* means almost surely.

2. Least-Squares Estimator via Mixed Integer Optimization.

2.1. *Identifiability.* We use the convention that the constant 1 is the first element of x_t and -1 is the last element of f_t. Define $\alpha := (\beta', \delta')'$ and $Z_t(\gamma) := (x_t', x_t' 1\{f_t'\gamma > 0\})'$. Then, we can rewrite the model as

$$y_t = Z_t(\gamma_0)'\alpha_0 + \varepsilon_t.$$

Because only the sign of the index $f_t'\gamma_0$ determines the regime switching, the scale of γ_0 is not identifiable. We assume that the first element of γ_0 equals 1. Let d_x and d_f denote the dimensions of x_t and f_t, respectively.

ASSUMPTION 2.1. *$\alpha_0 \in \mathbb{R}^{2d_x}$ and $\gamma_0 \in \Gamma := \{(1, \gamma_2')' : \gamma_2 \in \Gamma_2\}$, where $\Gamma_2 \subset \mathbb{R}^{d_f - 1}$ is a compact set.*

We decompose f_t into a scalar random variable f_{1t} and other variables f_{2t}, so that $f_t'\gamma \equiv f_{1t} + f_{2t}'\gamma_2$. In view of the conditional mean zero restriction in (1.2), it is natural to impose conditions under which both α_0 and γ_0 are identified by the L_2-loss. Introduce the excess loss

$$R(\alpha, \gamma) := \mathbb{E}(y_t - x_t'\beta - x_t'\delta 1\{f_t'\gamma > 0\})^2 - \mathbb{E}(\varepsilon_t^2). \tag{2.1}$$

In order to establish that $R(\alpha,\gamma) > R(\alpha_0,\gamma_0) = 0$ whenever $(\alpha,\gamma) \neq (\alpha_0,\gamma_0)$, we make the following regularity conditions.

ASSUMPTION 2.2. *For any $\varepsilon > 0$, (α_0,γ_0) satisfies*

$$\inf_{\{(\alpha',\gamma')' \in \mathbb{R}^{2d_x} \times \Gamma : |(\alpha',\gamma') - (\alpha_0',\gamma_0')|_2 > \varepsilon\}} R(\alpha,\gamma) > 0.$$

Online Appendix A provides sufficient conditions for Assumption 2.2.

2.2. *Estimator.* We now propose the least-squares estimator and two alternative algorithms to compute the proposed estimator. For computational purposes, we assume that $\alpha \in \mathcal{A} \subset \mathbb{R}^{2d_x}$ for some known compact set $\mathcal{A}$. In practice, we can take a large $2d_x$-dimensional hyper-rectangle so that the resulting estimator is not on the boundary of $\mathcal{A}$. The unknown parameters can be estimated by least squares: $(\widehat{\alpha},\widehat{\gamma})$ solves

$$\min_{(\alpha',\gamma')' \in \mathcal{A} \times \Gamma} \mathbb{S}_T(\alpha,\gamma) \equiv \frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\beta - x_t'\delta 1\{f_t'\gamma > 0\})^2 \tag{2.2}$$

$$\text{subject to:} \quad \tau_1 \leq \frac{1}{T}\sum_{t=1}^{T} 1\{f_t'\gamma > 0\} \leq \tau_2. \tag{2.3}$$

We assume that the restriction (2.3) is satisfied when $\gamma = \gamma_0$ *a.s.* Here, $0 < \tau_1 < \tau_2 < 1$ for some predetermined τ_1 and τ_2 (e.g., $\tau_1 = 0.05$ and $\tau_2 = 0.95$). In the special case that $1\{f_t'\gamma_0 > 0\} = 1\{q_t > \widetilde{\gamma}_0\}$ with a scalar variable q_t and a parameter $\widetilde{\gamma}_0$, it is standard to assume that the parameter space for $\widetilde{\gamma}_0$ is between the τ and $(1-\tau)$ quantiles of q_t for some known $0 < \tau < 1$. We can interpret (2.3) as a natural generalization of this restriction so that the proportion of one regime is never too close to 0 or 1.

When γ is of high dimension, the naive grid search would not work well. Dynamic programming (e.g., [7]) or smooth global optimization (e.g., [24]) might be considered but are not readily available. We overcome this computational difficulty by replacing the naive grid search with MIO. We present two alternative algorithms based on MIO below.

2.3. *Mixed Integer Quadratic Programming.* Our first algorithm is based on mixed integer quadratic programming (MIQP), which jointly estimates (α,γ). It is guaranteed to obtain a global solution once it is found. To write the original least-squares problem in MIQP, we introduce $d_t := 1\{f_t'\gamma > 0\}$ and $\ell_t := \delta d_t$ for $t = 1,\ldots,T$. Then, rewrite the objective function as

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\beta - x_t'\ell_t)^2, \tag{2.4}$$

which is a quadratic function of β and ℓ_t. The goal is to introduce only linear constraints with respect to variables of optimization, and to construct an MIQP that is equivalent to the original least-squares problem. Then, we can apply modern MIO packages (e.g., Gurobi) to solve MIQP. The assumption $\alpha \in \mathcal{A}$ implies that there exist known upper and lower bounds for δ_j: $L_j \leq \delta_j \leq U_j$, where δ_j denotes the jth element of δ for $j = 1, \ldots, d_x$. In addition, to make sure that $\ell_{j,t} = \delta_j d_t$ for each j and t, we impose two additional restrictions:

$$d_t L_j \leq \ell_{j,t} \leq d_t U_j \quad \text{and} \quad L_j(1 - d_t) \leq \delta_j - \ell_{j,t} \leq U_j(1 - d_t). \tag{2.5}$$

It is then straightforward to check that these constraints imply $\ell_{j,t} = \delta_j d_t$. To introduce another key constraint, we define $M_t \equiv \max_{\gamma \in \Gamma} |f_t'\gamma|$ for each $t = 1, \ldots, T$, where Γ is the parameter space for γ_0. We can compute M_t easily for each t using linear programming. We store them as inputs to our algorithm. The following new constraints along with (2.3) and (2.5) ensure that the reformulated problem (2.4) is the same as the original problem:

$$(d_t - 1)(M_t + \epsilon) < f_t'\gamma \leq d_t M_t,$$

where $\epsilon > 0$ is a small predetermined constant (e.g., $\epsilon = 10^{-6}$). The following defines an algorithm for the MIQP algorithm.

Algorithm 1: Mixed Integer Quadratic Programming (MIQP)

Input: $\{(y_t, x_t, f_t, M_t) : t = 1, \ldots, T\}$

Output: $(\widehat{\alpha}, \widehat{\gamma})$

1 Let $\boldsymbol{d} = (d_1, \ldots, d_T)'$ and $\boldsymbol{\ell} = \{\ell_{j,t} : j = 1, \ldots, d_x, t = 1, \ldots, T\}$, where $\ell_{j,t}$ is a real-valued variable. Solve the following problem:

$$\min_{\beta, \delta, \gamma, \boldsymbol{d}, \boldsymbol{\ell}} \mathbb{Q}_T(\beta, \boldsymbol{\ell}) \equiv \frac{1}{T} \sum_{t=1}^{T} (y_t - x_t'\beta - \sum_{j=1}^{d_x} x_{j,t}\ell_{j,t})^2 \tag{2.6}$$

subject to

$$\begin{aligned} &(\beta, \delta) \in \mathcal{A},\ \gamma \in \Gamma,\ d_t \in \{0, 1\},\ L_j \leq \delta_j \leq U_j, \\ &(d_t - 1)(M_t + \epsilon) < f_t'\gamma \leq d_t M_t, \\ &d_t L_j \leq \ell_{j,t} \leq d_t U_j, \\ &L_j(1 - d_t) \leq \delta_j - \ell_{j,t} \leq U_j(1 - d_t), \\ &\tau_1 \leq T^{-1} \sum_{t=1}^{T} d_t \leq \tau_2 \end{aligned} \tag{2.7}$$

for each $t = 1, \ldots, T$ and each $j = 1, \ldots, d_x$, where $0 < \tau_1 < \tau_2 < 1$.

Our proposed algorithm is mathematically equivalent to the original least-squares problem (2.2) subject to (2.3) in terms of values of objective functions. Formally, we state it as the following theorem.

THEOREM 2.1. *Let* $(\bar{\alpha}, \bar{\gamma})$ *denote a solution using MIQP as described above. Then,* $\mathbb{S}_T(\widehat{\alpha}, \widehat{\gamma}) = \mathbb{S}_T(\bar{\alpha}, \bar{\gamma})$*, where* $(\widehat{\alpha}, \widehat{\gamma})$ *is defined in* (2.2)*.*

The proposed algorithm in Section 2.3 may run slowly when the dimension d_x of x_t is large. To mitigate this problem, we reformulate MIQP in Appendix B.2 and use the alternative formulation in our numerical work; however, we present a simpler form here to help readers follow our basic ideas more easily.

2.4. *Block Coordinate Descent.* While the MIQP jointly estimates (α, γ) and aims at obtaining a global solution, it might not compute as fast as necessary in large-scale problems. To mitigate the issue of scalability, we introduce a faster alternative approach based on mixed integer linear programming (MILP), whose objective function is linear in d_t. The algorithm solves for α and γ iteratively, which we call a block coordinate descent (BCD) algorithm, starting with an initial value that can be obtained through MIQP with an early stopping rule. At step k, given $\widehat{\alpha}^{k-1}$, which is obtained in the previous step, we estimate γ by solving

$$\min_{\gamma \in \Gamma, d_1, \ldots, d_T} \frac{1}{T} \sum_{t=1}^{T} \left(y_t - x_t' \widehat{\beta}^{k-1} - x_t' \widehat{\delta}^{k-1} d_t \right)^2 \tag{2.10}$$

subject to similar constraints as in MIQP. Note that the least-squares problem (2.10) is linear in d_t as $d_t^2 = d_t$. The BCD algorithm is defined in Algorithm 2. Intuitively speaking, it runs the MIQP algorithm for the amount of time `MaxTime_1`, then switches to the MILP for the amount of time `MaxTime_2`. The BCD approach is a descent algorithm in the sense that the least-squares objective function is a non-increasing function of k. In other words, BCD in Steps 2 and 3 can provide a higher-quality solution than MIQP with an early stopping rule `MaxTime_1`. The time limit `MaxTime_2` in Step 2 can be smaller than `MaxTime_1` as it is easier to solve an MILP problem than to solve an MIQP problem. Furthermore, the alternative minimization approach efficiently solves for $\widehat{\alpha}^k$ because it has an explicit solution.

Figure 1 illustrates the performance of MIQP and BCD in one simulation draw. After spending `MaxTime_1` (600 seconds) in Step 1, BCD switches into Step 2 and it converges to the solution quickly just in one iteration. Meanwhile, MIQP achieves a similar objective function value after spending the whole time budget of 1800 seconds. In Monte Carlo experiments,

Algorithm 2: Block Coordinate Descent (BCD)

Input: $\{(y_t, x_t, f_t, M_t) : t = 1, \ldots, T\}$, `MaxTime_1`, `MaxTime_2`
Output: $(\widehat{\alpha}, \widehat{\gamma})$

Set $k = 1$;
Step 1. Obtain an initial estimate $\left(\widehat{\alpha}^0, \widehat{\gamma}^0\right)$ using MIQP with the pre-specified time limit `MaxTime_1`;
if *a solution is found before reaching* `MaxTime_1`, **then**
  set the initial estimate as the final estimate and terminate;
end
while *elapsed time is no greater than* `MaxTime_2` **do**
  Step 2. For the given $\widehat{\alpha}^{k-1}$, obtain an estimate $\widehat{\gamma}^k$ via MILP:

$$\min_{\gamma \in \Gamma, d_1, \ldots, d_T} \frac{1}{T} \sum_{t=1}^{T} \left\{ (x_t' \widehat{\delta}^{k-1})^2 - 2(y_t - x_t' \widehat{\beta}^{k-1}) x_t' \widehat{\delta}^{k-1} \right\} d_t \tag{2.8}$$

  subject to

$$\begin{aligned} &(d_t - 1)(M_t + \epsilon) < f_t' \gamma \leq d_t M_t, \\ &d_t \in \{0, 1\} \quad \text{for each } t = 1, \ldots, T, \\ &\tau_1 \leq \frac{1}{T} \sum_{t=1}^{T} d_t \leq \tau_2; \end{aligned} \tag{2.9}$$

  if $\mathbb{S}_T\left(\widehat{\alpha}^{k-1}, \widehat{\gamma}^k\right) \geq \mathbb{S}_T\left(\widehat{\alpha}^{k-1}, \widehat{\gamma}^{k-1}\right)$, **then**
    terminate;
  end
  Step 3. For the given $\widehat{\gamma}^k$, obtain

$$\widehat{\alpha}^k = \left[\frac{1}{T} \sum_{t=1}^{T} Z_t\left(\widehat{\gamma}^k\right) Z_t\left(\widehat{\gamma}^k\right)' \right]^{-1} \frac{1}{T} \sum_{t=1}^{T} Z_t\left(\widehat{\gamma}^k\right) y_t;$$

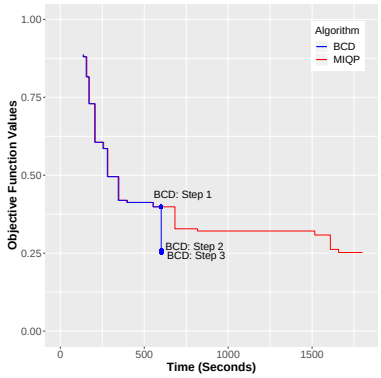
  Let $k = k + 1$;
end

we compare MIQP with BCD more thoroughly, subject to the same total computing time restrictions, and we demonstrate the efficiency of BCD.

3. Asymptotic Properties with Known Factors. We split the asymptotic properties of the estimator into two cases: known and unknown factors. In this section, we consider the former.

ASSUMPTION 3.1. *(i) $\{x_t, f_t, \varepsilon_t\}$ is a sequence of strictly stationary, ergodic, and ρ-mixing random vectors with $\sum_{m=1}^{\infty} \rho_m^{1/2} < \infty$, $\mathbb{E}\,|x_t|_2^4 < \infty$, and there exists a constant $C < \infty$ such that $\mathbb{E}(|x_t|_2^8 \,|\, f_t'\gamma = 0) < C$ and $\mathbb{E}(\varepsilon_t^8 \,|\, f_t'\gamma = 0) < C$ for all $\gamma \in \Gamma$.*

FIG 1. *Computation Example of MIQP and BCD*



(ii) $\{\varepsilon_t\}$ *is a martingale difference sequence, that is,* $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$*, where* x_t *and* f_t *are adapted to the filtration* $\mathcal{F}_{t-1}$*.*

(iii) The smallest eigenvalue of $\mathbb{E}[Z_t(\gamma) Z_t(\gamma)']$ *is bounded away from zero for all* $\gamma \in \Gamma$*.*

We decompose f_t into a scalar random variable f_{1t} and the other variables f_{2t}, so that $f_t'\gamma \equiv f_{1t} + f_{2t}'\gamma_2$. Define $u_t := f_t'\gamma_0$.

ASSUMPTION 3.2. *(i) For some* $0 < \varphi < 1/2$ *and* $d_0 \neq 0$*,* $\delta_0 = d_0 T^{-\varphi}$*.*

(ii) $p_{u_t|f_{2t}}(u)$*,* $\mathbb{E}[(x_t'd_0)^2|f_{2t}, u_t = u]$ *and* $\mathbb{E}[(\varepsilon_t x_t'd_0)^2|f_{2t}, u_t = u]$ *are continuous and bounded away from zero at* $u = 0$ *a.s.*

(iii) For some $M < \infty$*,* $\inf_{|r|_2=1} \mathbb{E}(|f_{2t}'r| 1\{|f_{2t}|_2 \leq M\}) > 0$*.*

Most of the conditions in Assumptions 3.1 and 3.2 are a natural extension of the scalar case in the literature, when $f_t = (q_t, -1)'$ for a scalar random variable (e.g., [16]). Assumption 3.2(iii) is a rank condition on f_{2t} due to the vector of threshold parameter to be estimated and it is in terms of the first moment because of the asymptotic linear approximation of criterion function near γ_0. It also allows for discrete variables in f_{2t}. Assumption 3.2(ii) ensures the presence of a jump, not just a kink at the change point.

THEOREM 3.1. *Let* $\mathcal{G} := \{g \in \mathbb{R}^{d_f} : g_1 = 0\}$*. Let Assumptions 2.1, 2.2, 3.1, and 3.2 hold. Assume further that* α_0 *is in the interior of* $\mathcal{A}$ *and that* γ_0 *is in the interior of* Γ*. In addition, let* W *denote a mean-zero Gaussian process whose covariance kernel is given by*

$$H(s,g) := \frac{1}{2}\mathbb{E}\left[(\varepsilon_t x_t'd_0)^2 \left(|f_t'g| + |f_t's| - |f_t'(g-s)|\right) p_{u_t|f_{2t}}(0)\right]. \tag{3.1}$$

Then, as $T \to \infty$, *we have*

$$\sqrt{T}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}(0, (\mathbb{E}Z_t(\gamma_0)Z_t(\gamma_0)')^{-1}\mathrm{var}(Z_t(\gamma_0)\varepsilon_t)(\mathbb{E}Z_t(\gamma_0)Z_t(\gamma_0)')^{-1}),$$
$$T^{1-2\varphi}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \operatorname*{argmin}_{g \in \mathcal{G}} \left\{ \mathbb{E}\left[\left(x_t' d_0\right)^2 \left| f_t' g \right| p_{u_t|f_{2t}}(0) \right] + 2W(g) \right\},$$

where $\sqrt{T}(\widehat{\alpha} - \alpha_0)$ *and* $T^{1-2\varphi}(\widehat{\gamma} - \gamma_0)$ *are asymptotically independent.*

The normalization scheme is embedded in the asymptotic distribution. Because $\gamma_1 = 1$, the minimum in the limit is taken after fixing the first element of g at zero (recall that $\mathcal{G} = \{g \in \mathbb{R}^{d_f} : g_1 = 0\}$). Also note that, in the scalar threshold case, $f_t = (q_t, -1)'$ and $\gamma_0 = (1, \widetilde{\gamma}_0)'$,

$$H(s,g) = \frac{1}{2}\mathbb{E}\left[\left(\varepsilon_t x_t' d_0\right)^2 (2\min(|g_2|, |s_2|) 1\{\mathrm{sgn}(g_2) = \mathrm{sgn}(s_2)\}) p_{u_t|f_{2t}}(0) \right],$$

which becomes the two-sided Brownian motion, as in Hansen [16].

4. Estimation with Unobserved Factors. In this section, we consider the case in which the factors are estimated.

4.1. *The Model.* Consider the following factor model,

$$\mathcal{Y}_t = \Lambda g_{1t} + e_t, \ t = 1, \ldots, T, \tag{4.1}$$

where $\mathcal{Y}_t$ is an $N \times 1$ vector of time series, Λ is an $N \times K$ matrix of factor loadings, g_{1t} is a $K \times 1$ vector of common factors, and e_t is an $N \times 1$ vector of idiosyncratic components. Throughout this section, we make it explicit that there is a constant term in the factors, and we replace the regression model in (1.1) with

$$y_t = x_t'\beta_0 + x_t'\delta_0 1\{g_t'\phi_0 > 0\} + \varepsilon_t, \tag{4.2}$$

where $g_t = (g_{1t}', -1)'$ is a vector of unknown factors in (4.1) plus a constant term (-1), and ϕ_0 is a vector of unknown parameters. In addition, we allow g_{1t} to contain lagged (dynamic) factors, but we treat them as static factors and estimate them using the PCA without losing the validity of the estimated factors.

It is well known that g_t is identifiable and estimable by the PCA up to an invertible matrix transformation (i.e., $H_T' g_t$), whose exact form will be given in Section 4.5. Therefore, it is customary in the literature (see, e.g., [3, 4]) to treat $H_T' g_t$ as a centering object in the limiting distribution of estimated factors. Following this convention, in this section, let

$$f_t := H_T' g_t \quad \text{and} \quad \gamma_0 := H_T^{-1}\phi_0. \tag{4.3}$$

Using the fact that $g_t'\phi_0 = f_t'\gamma_0$, we can rewrite (4.2) as the original formulation in (1.1):

$$y_t = x_t'\beta_0 + x_t'\delta_0 1\{f_t'\gamma_0 > 0\} + \varepsilon_t.$$

Hence, γ_0 depends on the sample in this section but we suppress dependence on T for the sake of notational simplicity.

Our estimation procedure now consists of two steps. In the first step, a $(K+1)\times 1$ vector of estimated factors and the constant term (i.e., $\widetilde{f}_t := (\widetilde{f}_{1t}', 1)'$) are obtained by the method of principal components. To describe estimated factors, let $\mathcal{Y}$ be the $T \times N$ matrix whose t-th row is $\mathcal{Y}_t'$. Let $(\widetilde{f}_{11}, \ldots, \widetilde{f}_{1T})$ be the $K \times T$ matrix, whose rows are K eigenvectors (multiplied by $\sqrt{T}$) associated with the largest K eigenvalues of $\mathcal{Y}\mathcal{Y}'/NT$ in decreasing order. In the second step, unknown parameters (α_0, γ_0) are estimated by the same algorithm in Section 2 with $\widetilde{f}_t$ as inputs.

4.2. *Regularity Conditions.* We introduce assumptions needed for asymptotic results with estimated factors. We first replace Assumptions 2.1–3.2 with the following assumption. Define

$$\Phi_T := \{\phi : \phi = H_T\gamma \text{ for some } \gamma \in \Gamma_\epsilon\}, \tag{4.4}$$

where Γ_ϵ is an ϵ-enlargement of Γ. Note that ϕ cannot be a vector whose first K elements are zeros due to the normalization on γ and the block diagonal structure of H_T that will be defined in (4.7). The space Φ_T for ϕ is defined through H_T and excludes the case that $g_t'\phi$ is degenerate. The ϵ-enlargement of Γ is needed because the factors are latent.

ASSUMPTION 4.1. *(i) Assumptions 2.1, 2.2, and 3.2(i) hold after replacing f_t and γ_0 with g_t and ϕ_0, respectively.*

(ii) $\{x_t, g_t, e_t, \varepsilon_t\}$ is a sequence of strictly stationary, ergodic, and ρ-mixing random vectors with $\sum_{m=1}^{\infty} \rho_m^{1/2} < \infty$, and there exists a constant $C < \infty$ such that $\mathbb{E}(|x_t|_2^8 \,|g_t, e_t) < C$, $\mathbb{E}(\varepsilon_t^8 | g_t, e_t) < C$ a.s., and $g_t'\phi$ has a density that is continuous and bounded by C for all $\phi \in \Phi_T$.

Recall that by the normalization in Assumption 2.1, the first element of γ is fixed at 1. One caveat of this normalization scheme is that the sign of the first element of f_t might not be the same as that of the first element of g_t due to random rotation H_T; however, if we assume that $\delta_0 \neq 0$ and we also know the sign of one of the non-zero coefficients of δ_0, then we can determine the sign of the first element of f_t after estimating the model. This is a "labeling" problem that is common in models with hidden regimes. For simplicity, we assume that the first element of γ_0 is 1.

The following assumption is standard in the literature. In particular, we allow weak serial correlation among e_t.

ASSUMPTION 4.2. *(i)* $\lim_{N\to\infty} \frac{1}{N}\Lambda'\Lambda = \Sigma_\Lambda$ *for some* $K \times K$ *matrix* Σ_Λ*, whose eigenvalues are bounded away from both zero and infinity.*
(ii) The eigenvalues of $\Sigma_\Lambda^{1/2}\mathbb{E}(g_{1t}g_{1t}')\Sigma_\Lambda^{1/2}$ *are distinct.*
(iii) All the eigenvalues of the $N \times N$ *covariance* $\mathrm{var}(e_t)$ *are bounded away from both zero and infinity.*
(iv) For any t, $\frac{1}{N}\sum_{s=1}^{T}\sum_{i=1}^{N}|\mathbb{E}e_{it}e_{is}| < C$ *for some* $C > 0$.

Define λ_i' to be the ith row of Λ, so that $\Lambda = (\lambda_1, \ldots, \lambda_N)'$. Further, let

$$\xi_{s,t} := N^{-1/2}\sum_{i=1}^{N}(e_{is}e_{it} - \mathbb{E}e_{is}e_{it}), \quad \psi := (TN)^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{N} g_t e_{it}\lambda_i',$$
$$\eta_t := (TN)^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{N} g_{1s}(e_{is}e_{it} - \mathbb{E}e_{is}e_{it}), \quad \zeta_t := N^{-1/2}\sum_{i=1}^{N}\lambda_{it}e_{it}.$$

We require the following additional exponential-tail conditions.

ASSUMPTION 4.3. *There exist finite, positive constants* C, C_1 *and* c_1 *such that for any* $x > 0$ *and for any* $\varpi \in \Xi := \{e_{it}, g_{1t}, \xi_{s,t}, \zeta_t, vec(\psi), \eta_t\}$,

$$\mathbb{P}(|\varpi|_2 > x) \le C\exp(-C_1 x^{c_1}).$$

These conditions impose exponential tail conditions on various terms. First, it requires weak cross-sectional correlations among e_{it}. This assumption can be verified under some low-level conditions such as the α-mixing condition of the type of Merlevède et al. [23] across both (i,t) and individual exponential-tailed distributions on $\{e_{it}, g_t\}$. While the quantities in Ξ are often assumed to have finite moments in the high-dimensional factor model literature, these moment bounds would no longer be sufficient in the current context. Instead, exponential-type probability bounds are more useful for us to characterize the effect of the estimated factors. To see the point, note that we have the following asymptotic expansion:

$$\widetilde{f}_t = \widehat{f}_t + r_t, \quad \widehat{f}_t := H_T'(g_t + N^{-1/2}h_t). \tag{4.5}$$

Here, r_t is a remainder term,

$$H_T' := \begin{pmatrix} \widetilde{H}_T' & 0 \\ 0 & 1 \end{pmatrix}, \quad h_t := \begin{pmatrix} h_{1t} \\ 0 \end{pmatrix}, \quad h_{1t} := (\frac{1}{N}\Lambda'\Lambda)^{-1}\frac{1}{\sqrt{N}}\Lambda' e_t, \tag{4.6}$$

and the exact form of $\widetilde{H}_T$ is given in (4.7). The diagonality in H_T and the zero element in h_t reflect the inclusion of the constant in g_t. We establish the following uniform approximation result: uniformly for γ over a compact set,

$$\max_{t\le T}\left|\mathbb{P}(\widetilde{f}_t'\gamma>0)-\mathbb{P}(\widehat{f}_t'\gamma>0)\right|\le O\left(\frac{(\log T)^c}{T}\right)+\max_{t\le T}\mathbb{P}\left(|r_t|>C\frac{(\log T)^c}{T}\right)$$

for some constants $C, c > 0$. The above exponential-tail assumption then enables us to derive a sharp bound so that $\max_{t\le T}\mathbb{P}(|r_t| > C(\log T)^c T^{-1})$ is asymptotically negligible.

Next, we state important technical conditions to facilitate the local asymptotic expansion of the least-squares criterion function. A technical challenge in the analysis is that even the expected criterion function is non-smooth with respect to the factors. As such, we introduce some conditional density conditions to study the effect of estimating factors $H_T'h_t = \sqrt{N}(\widehat{f}_t - f_t)$.

ASSUMPTION 4.4. *(i)* $\sup_{x_t,g_t}|\mathbb{P}(h_t'\phi_0<0|x_t,g_t)-(1/2)| = O(N^{-1/2})$.
(ii) Let $\sigma^2_{h,x_t,g_t} := \operatorname{plim}_{N\to\infty}\mathbb{E}[(h_t'\phi_0)^2|x_t,g_t]$ *and let* $\mathcal{Z}_t$ *be a sequence of Gaussian random variables whose conditional distribution, given* x_t *and* g_t*, is* $\mathcal{N}(0,\sigma^2_{h,x_t,g_t})$*. Then, there are positive constants* c*,* c_0*, and* C *such that* $\sigma^2_{h,x_t,g_t} > c_0$ *a.s.,* $\sup_{x_t,g_t}\sup_{|z|<c} p_{h_t'\phi_0|g_t,x_t}(z) < C$*, and*

$$\sup_{x_t,g_t}\sup_{|z|<c}|p_{h_t'\phi_0|g_t,x_t}(z)-p_{\mathcal{Z}_t|g_t,x_t}(z)| = o(1).$$

Assumption 4.4 is concerned with the asymptotic behavior of the distribution of h_t as $N\to\infty$. The rate $N^{-1/2}$ in Assumption 4.4(i) is a reminiscent of the Berry–Essen theorem. The Edgeworth expansion of the sample means at zero implies that the approximation error is $CN^{-1/2}$, where the universal constant C depends on the moments of the summand up to the third order [14]. Thus, condition (i) holds for a broad range of setups including heteroskedastic errors e_{it}. For instance, if the idiosyncratic error has the form $e_{it} = \sigma(g_t)\xi_{it}$, where g_t and ξ_{it} are two independent sequences and $\{\xi_{it}\}$ is an independent and identically distributed (i.i.d.) sequence across i, then the condition is satisfied as long as both $\sigma(g_t)^3$ and $\mathbb{E}|\xi_{it}|^3$ are bounded. Furthermore, it holds trivially if the conditional distribution of $h_t'\phi_0$ given x_t and g_t is symmetric around zero or more generally if its median is zero. Assumption 4.4 ensures, among other things, that for some function $\Psi(\cdot)$ such that $\mathbb{E}|\Psi(x_t,g_t)|<\infty$,

$$\mathbb{E}\left[\Psi(x_t,g_t)\left(1\{h_t'\phi_0\le 0\}-1\{\mathcal{Z}_t\le 0\}\right)\middle|x_t,g_t\right] = O(N^{-1/2}).$$

Above all, because h_t is a cross-sectional average multiplied by $\sqrt{N}$, this assumption can be verified by a cross-sectional central limit theorem (CLT), if $\{e_{it} : i \leq N\}$ satisfies some cross-sectional mixing condition.

In the next assumption, recall that, by the identification condition, we can write $\gamma = (1, \gamma_2)$, where 1 is the first element of γ. Correspondingly, let f_{2t} and $\widehat{f}_{2t}$ be the subvectors of f_t and $\widehat{f}_t$, excluding their first elements. Also, let $u_t := g_t'\phi_0 = f_t'\gamma_0$ and $\breve{g}_t := g_t + h_t/\sqrt{N}$.

ASSUMPTION 4.5. *There exist positive constants c, c_0, M_0, and M such that the following hold* a.s.*.*

(i) $\inf_{|u|<c} p_{\widehat{f}_t'\gamma_0|\widehat{f}_{2t},x_t}(u) \geq c_0$ *and* $\sup_{|f|_2<M_0} p_{f_{2t}|h_t}(f) < M$.

(ii) $\inf_{|u|<c} p_{u_t|f_{2t},h_t,x_t}(u) \geq c_0$. *For all* $|u_1| < c, |u_2| < c$,

$$|p_{u_t|h_t'\phi_0,f_{2t},x_t}(u_1) - p_{u_t|h_t'\phi_0,f_{2t},x_t}(u_2)| \leq M|u_1 - u_2|.$$

(iii) $\inf_{|r|_2=1} \mathbb{E}\left[|f_{2t}'r|^k 1\{|f_{2t}|_2 < M_0\}\right] \geq c_0$ *for* $k = 1, 2$.

(iv) $\sup_{|r|_2=1} \sup_{|u|<c} p_{g_t'r|h_t}(u) \leq M$.

(v) Each of $\inf_{\phi\in\Phi_T} |g_t'\phi|$, $\inf_{\phi\in\Phi_T} |\breve{g}_t'\phi|$, $\sup_{\phi\in\Phi_T} |h_t'\phi|$, *and* $\breve{g}_t'\phi_0$ *has a density function bounded and continuous at zero, with* Φ_T *given in* (4.4).

(vi) $\mathbb{E}[(x_t'd_0)^2|g_t, h_t]$ *is bounded above by* M_0 *and below by* c_0.

(vii) For any s *and* w *that are linearly independent of* ϕ_0, $p_{\breve{g}_t'\phi_0|\breve{g}_t's,\breve{g}_t'w}(u)$ *and* $\mathbb{E}((\varepsilon_t x_t'd_0)^2|\breve{g}_t'\phi_0 = u, \breve{g}_t's, \breve{g}_t'w)$ *are continuously differentiable at* $u = 0$ *with bounded derivatives. Furthermore,* $\mathbb{E}((\varepsilon_t x_t'd_0)^4 |\breve{g}_t|_2^2 |\breve{g}_t'\phi_0) \leq M$.

These conditions control the local characteristics of the centered least-squares criterion function near the true parameter value. As the model is perturbed by the error in the estimated factors, the centered criterion is a drifting sequence $\widehat{f}_t$. Its leading term changes depending on whether $N = O(T^{2-4\varphi})$ or not. The lower bounds in the above assumption are part of rank conditions that ensure that the leading terms are well defined. As a result, it entails a phase transition on the distribution of $\widehat{\gamma}$. Because they are rather technical, we provide a more detailed discussion on Assumption 4.5 in Online Appendix G.2.

4.3. *Rates of Convergence.* The following theorem presents the rates of convergence for the estimators.

THEOREM 4.1. *Let Assumptions 4.1–4.5 hold. Suppose* $T = O(N)$. *Then*

$$|\widehat{\alpha} - \alpha_0|_2 = O_P\left(\frac{1}{\sqrt{T}}\right) \quad \text{and} \quad |\widehat{\gamma} - \gamma_0|_2 = O_P\left(\frac{1}{T^{1-2\varphi}} + \frac{1}{(NT^{1-2\varphi})^{1/3}}\right).$$

While the convergence rate for $\widehat{\alpha}$ is standard, the convergence rate of $\widehat{\gamma}$ merits further explanation. First of all, when N is relatively large so that $T^{2-4\varphi} = o(N)$, $\widehat{\gamma} - \gamma_0$ converges at a super-consistent rate of $T^{-(1-2\varphi)}$. Contrary to this case, when $N = o(T^{2-4\varphi})$, the estimated threshold parameter has a cube root rate, which is similar to that of the maximum score type estimators [19]. Therefore, as $\sqrt{N}/T^{1-2\varphi}$ varies in $[0, \infty]$, the rate of convergence varies between the super-consistency rate of the usual threshold models to the cube root rate of the maximum score type estimators.

The convergence rates exhibit a continuous transition from one to the other. To explain this transition phenomenon, we can show that uniformly in (α, γ), the objective function has the following expansion: there are functions $R_1(\cdot)$ and $R_2(\cdot, \cdot)$ such that

$$\mathbb{S}_T(\alpha, \gamma) - \mathbb{S}_T(\alpha_0, \gamma_0) = R_1(\gamma) + R_2(\alpha, \gamma),$$

where $\gamma \mapsto R_1(\gamma)$ is a non-stochastic function, representing the "mean" of the loss function, but is also highly non-smooth with respect to γ, and $R_2(\alpha, \gamma)$ is the remaining stochastic part. A key step is to derive a sharp lower bound for $R_1(\gamma)$. When N is relatively large, the effect of estimating latent factors is negligible, and $R_1(\gamma)$ has a high degree of non-smoothness. Similar to the usual threshold model, we have

$$R_1(\gamma) \geq CT^{-2\varphi}|\gamma - \gamma_0|_2 - O_P(T^{-1}).$$

This lower bound leads to a super-consistency rate. On the other hand, when N is relatively small, there are extra noises arising from the cross-sectional idiosyncratic errors when estimating the latent factors, which we call "cross-sectional noises." A remarkable feature of our model is that the cross-sectional noises help smooth the objective function in this case. As a result, the behavior of $R_1(\gamma)$ is similar to that of the maximum score type estimators, where a quadratic lower bound can be derived:

$$R_1(\gamma) \geq CT^{-2\varphi}\sqrt{N}|\gamma - \gamma_0|_2^2 - O_P(T^{-2\varphi}N^{-5/6}).$$

The quadratic lower bound, together with a larger error rate, then leads to a cube root rate type of convergence. See Online Appendix G.1 for a detailed description of the roadmap of the proof.

4.4. *Consistency of Regime-Classification.* We introduce an error rate in (in-sample) regime-classification,

$$\widehat{R}_T = \frac{1}{T}\sum_{t=1}^{T}\left|1\left\{\widetilde{f}_t'\widehat{\gamma} > 0\right\} - 1\left\{f_t'\gamma_0 > 0\right\}\right|.$$

The uncertainty about the regime classification comes from either $\widetilde{f}_t$ or $\widehat{\gamma}$ or both. We establish its convergence rate in the following theorem.

THEOREM 4.2. *Let Assumptions 4.1–4.5 hold. Suppose $T = O(N)$. Then*

$$\widehat{R}_T = O_P\left(\left(NT^{1-2\varphi}\right)^{-1/3} + T^{-1+2\varphi} + N^{-1/2}\right).$$

This is a useful corollary of the derivation of the rates of convergence for the threshold estimator. We expect a good performance of our regime classification rule even with a moderate size of T.

4.5. *Asymptotic Distribution.* To describe the asymptotic distribution, we introduce additional notation. Let V_T denote the $K \times K$ diagonal matrix whose elements are the K largest eigenvalues of $\mathcal{Y}\mathcal{Y}'/NT$. Define

$$\widetilde{H}_T' := V_T^{-1} \frac{1}{T} \sum_{t=1}^{T} \widetilde{f}_{1t} g_{1t}' \frac{1}{N} \Lambda'\Lambda, \quad H_T := \operatorname{diag}(\widetilde{H}_T, 1), \tag{4.7}$$

and $H := \operatorname{plim}_{T,N\to\infty} H_T$, which is well defined, following Bai [3]. Let

$$\omega := \lim_{N,T\to\infty} \frac{\sqrt{N}}{T^{1-2\varphi}} \in [0,\infty], \quad \zeta_\omega := \max\{\omega, \omega^{1/3}\}, \quad \text{and} \quad M_\omega := \max\{1, \omega^{-1/3}\}.$$

Define, for $u_t = f_t'\gamma_0$,

$$A(\omega, g) := M_\omega \mathbb{E}\left[(x_t d_0)^2 \left(\left|f_t' g + \zeta_\omega^{-1}\mathcal{Z}_t\right| - \left|\zeta_\omega^{-1}\mathcal{Z}_t\right|\right) \middle| u_t = 0\right] p_{u_t}(0)$$

for $\omega \in (0,\infty]$, with the convention that $1/\omega = 0$ for $\omega = \infty$, and

$$A(0, g) := \mathbb{E}\left[(x_t' d_0)^2 (f_t' g)^2 \middle| u_t = 0, \mathcal{Z}_t = 0\right] p_{u_t,\mathcal{Z}_t}(0,0)$$

for $\omega = 0$. Recall $Z_t(\gamma) := (x_t', x_t' 1\{f_t'\gamma > 0\})'$.

THEOREM 4.3. *Let Assumptions 4.1–4.5 hold. Suppose $T = O(N)$. Let $\mathcal{G} := \{0\} \times \mathbb{R}^K$. In addition, let W denote the same Gaussian process as in Theorem 3.1. Then, as $N, T \to \infty$, we have*

$$\sqrt{T}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}\left(0, \left(\mathbb{E}Z_t(\gamma_0)Z_t(\gamma_0)'\right)^{-1} \mathbb{E}\left(Z_t(\gamma_0)Z_t(\gamma_0)'\varepsilon_t^2\right)\left(\mathbb{E}Z_t(\gamma_0)Z_t(\gamma_0)'\right)^{-1}\right),$$

$$\left(\left(NT^{1-2\varphi}\right)^{1/3} \wedge T^{1-2\varphi}\right)(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \operatorname*{argmin}_{g\in\mathcal{G}} A(\omega, g) + 2W(g),$$

and $\sqrt{T}(\widehat{\alpha} - \alpha_0)$ and $((NT^{1-2\varphi})^{1/3} \wedge T^{1-2\varphi})(\widehat{\gamma} - \gamma_0)$ are asymptotically independent. Moreover, $A(0, g) = \lim_{w\to 0} A(w, g)$.

It is worth noting that $A(\omega, g)$ is continuous everywhere, which implies that the distribution of the argmin of the limit processes $A(\omega, g) + 2W(g)$ is also continuous in ω in virtue of the argmax continuous mapping theorem [see e.g.,[29]]. Furthermore, the asymptotic distribution of $\widehat{\gamma}$ is well defined for any ω due to Lemma 2.6 of Kim and Pollard [19]. Specifically, the argmin of the limit Gaussian process is $O_P(1)$ since $A(\omega, g)$ is a deterministic function of order at least $|g|$ for any ω while the variance of $W(g)$ grows at the rate of $|g|$ as $g \to \infty$. It also possesses a unique minimizer almost surely.

In the literature, Bai and Ng [4, 5] have shown that the oracle property (with regard to the estimation of the factors) holds for the linear regression if $T^{1/2} = o(N)$ and for the extremum estimation if $T^{5/8} = o(N)$, in the presence of estimated factors. Thus, it appears that the oracle property demands a larger N as the nonlinearity of the estimating equation rises. In view of this, we regard our condition, $T = O(N)$, as not too stringent because we need to deal with estimated factors inside the indicator functions.

4.6. *Phase Transition.* To demonstrate that our asymptotic results are sharp, we consider a special case that $N = T^{\kappa}$ for $\kappa \geq 1$. In this case, the asymptotic results can be depicted on the (κ, φ)-space.

We categorize the results of Theorem 4.3 into three groups. In all three cases, the estimators enjoy certain oracle properties.

- Strong oracle: $T^{2-4\varphi} = o(N)$ or $\omega = \infty$. This is equivalent to $\kappa > 2 - 4\varphi$. The drift function $A(\infty, g)$ has a kink at $g = 0$. Intuitively, a bigger N makes the estimated factors more precise. This yields the oracle result for both $\widehat{\gamma}$ and $\widehat{\alpha}$, and the same asymptotic distribution as in the known factor case.
- Weak oracle: $N = o\left(T^{2-4\varphi}\right)$ or $\omega = 0$. This is equivalent to $\kappa < 2-4\varphi$. The drift function $A(0, g)$ is approximately quadratic in g near the origin. Because it is harder to identify the minimum when the function is smooth than when it has a kink at the minimum, this results in a non-oracle asymptotic distribution as well as a slower rate of convergence for $\widehat{\gamma}$ to $\left(NT^{1-2\varphi}\right)^{-1/3}$. However, the asymptotic distribution for $\widehat{\alpha}$ are still the same as those when the unknown factors are observed. So the oracle property for $\widehat{\alpha}$ is preserved.
- Semi-strong oracle: $N \asymp T^{2-4\varphi}$ or $\omega \in (0, \infty)$. This is equivalent to $\kappa = 2 - 4\varphi$. In this case, $A(\omega, g)$ has a continuous transition between the two polar cases discussed above. The effect of estimating factors is non-negligible for $\widehat{\gamma}$ and yet the estimator enjoys the same rate of convergence. The estimator $\widehat{\alpha}$ continues to achieve the oracle efficiency.

The phase transition occurs when $\kappa = 2 - 4\varphi$, which is the *semi-strong*

oracle case and the *critical boundary* of the phase transition. Changes in the convergence rates and asymptotic distributions are continuous along the critical boundary.

FIG 2. *Phase Diagram*

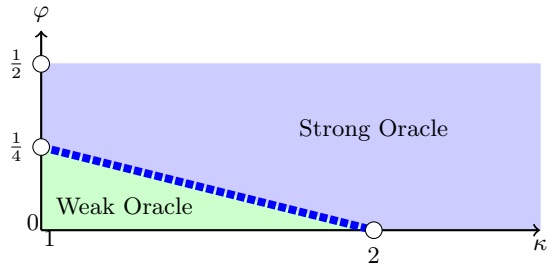
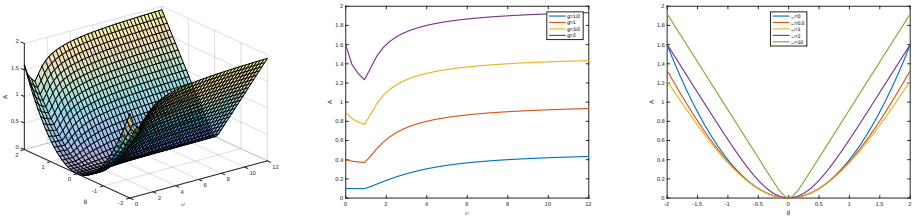


Figure 2 depicts a phase transition from the strong oracle phase to the weak oracle phase. The critical boundary $\kappa = 2 - 4\varphi$ is shown by closely dotted points in the figure. On one hand, as φ moves from 0 to 1/2, the strong oracle region for κ increases. That is, as the convergence rate for $\widehat{\gamma}$ becomes slower, the requirement for the minimal sample size N for factor estimation becomes less stringent. On the other hand, as κ becomes larger, the strong oracle region for φ increases. In other words, as N becomes larger, the range of attainable oracle rates of convergence for $\widehat{\gamma}$ becomes wider. In this way, we provide a thorough characterization of the effect of estimated factors.

FIG 3. *An Example of* $A(\omega, g)$



4.7. *Graphical Representation of* $A(\omega, g)$. To plot $A(\omega, g)$, we consider the simple case that $g_t = (q_t, -1)'$, $g = (0, g_2)'$, $x_t = 1$, $d_0 = 1$, and h_t and q_t are independent of each other. We write $g_2 = g$ for simplicity. The left panel of Figure 3 shows the three-dimensional graph of $A(\omega, g)$, the middle panel

depicts the profile of $A(\omega, g)$ as a function of ω for several values of g, and the right panel exhibits that of $A(\omega, g)$ as a function of g for given values of ω. First of all, it can be seen that $A(\omega, g)$ is continuous everywhere but has a kink at $\omega = 1$. As ω approaches zero, the shape of $A(\omega, g)$ is clearly quadratic in g; whereas, as ω becomes larger, it becomes almost linear in g. Also, $A(\omega, g)$ is quite flat around its minimum at $g = 0$ when ω is close to zero; however, $A(\omega, g)$ has a sharp minimum at zero for a larger value of ω. This reflects the fact that the rate of convergence increases as ω becomes larger.

5. Inference. In this section, we consider inference. Regarding α_0, Theorems 3.1 and 4.3 imply that inference for α_0 can be carried out as if γ_0 were known. Therefore, the standard inference method based on the asymptotic normality can be carried out for α_0 for both observed and estimated f_t.

We now focus on the inference issue regarding γ_0. Let $\theta_0 = h(\gamma_0)$ denote the parameter of interest for some known linear transformation $h(\cdot)$. For instance, this can be a particular element of γ_0 or a linear combination of the elements of γ_0. We use a quasi-likelihood ratio statistic:

$$LR(\theta) := \frac{\mathbb{S}_T(\widehat{\alpha}_h, \widehat{\gamma}_h) - \mathbb{S}_T(\widehat{\alpha}, \widehat{\gamma})}{\mathbb{S}_T(\widehat{\alpha}, \widehat{\gamma})},$$
$$(\widehat{\alpha}_h, \widehat{\gamma}_h) := \arg \min_{\alpha, h(\gamma)=\theta} \mathbb{S}_T(\alpha, \gamma), \quad (\widehat{\alpha}, \widehat{\gamma}) := \arg \min_{\alpha, \gamma} \mathbb{S}_T(\alpha, \gamma),$$

where $\mathbb{S}_T$ denotes the least-squares loss function, using f_t when factors are observable, and $\widetilde{f}_t$ when factors are estimated. Then, the $100(1-a)\%$-level confidence set for θ_0 is $\{\theta : LR(\theta) \leq \mathtt{cv}_a\}$, where $\mathtt{cv}_a$ denotes a critical value. As Theorem 5.1 shows, the asymptotic distribution is non-pivotal, so the critical value is computed based on the bootstrap.

5.1. *The Bootstrap with Estimated Factors.* We focus on the case of estimated factors, where we use $\widetilde{f}_t$ as the "true" factors, and denote by f_t^* as the *estimated factors* in the bootstrap world. To preserve the phase transition brought by the effect of PCA factor estimators, f_t^* should be a "perturbed" version of $\widetilde{f}_t$. Specifically, let f_t^* be re-estimated factors in the bootstrap sample via PCA. This is given by Gonçalves and Perron [13]. To maintain the cross-sectional dependence among the idiosyncratic components in the bootstrap factor models, we generate bootstrap data by

$$\mathcal{Y}_t^* := \widehat{\Lambda}\widetilde{f}_t + \widehat{\mathrm{var}}(e_t)^{1/2}\mathcal{W}_t^*,$$

where $\{\mathcal{W}_t^* : t \leq T\}$ is a sequence of independent $N \times 1$ multivariate standard normal random vectors and $\widehat{\mathrm{var}}(e_t)$ is the estimated covariance matrix

of e_t. If the covariance is a sparse matrix, we apply the thresholding covariance estimator of Fan, Liao, and Mincheva [11]. Then, we apply PCA to estimate factors to obtain $\widetilde{F}_t^*$. However, $\widetilde{F}_t^*$ estimates $\widetilde{f}_t$, the "true factors" in the bootstrap sample, up to a new rotation matrix H_T^*. Fortunately, such a rotation indeterminacy can be removed because H_T^* is known in the bootstrap world. Following Gonçalves and Perron [12, 13], we define

$$f_t^* := H_T^{*\prime -1} \widetilde{F}_t^* \tag{5.1}$$

as the final "estimated factors" in the bootstrap sample. The bootstrap distribution of $f_t^* - \widetilde{f}_t$ mimics well the asymptotic sampling distribution of $\widetilde{f}_t - H_T' g_t$, that is $\mathcal{N}(0, \Sigma_h)$. We give more details of this method, the definition of H_T^*, and an alternative method based on Gaussian perturbation in Online Appendix H.2.

5.2. *The k-Step Bootstrap Algorithm.* We now describe the bootstrap algorithm in detail. Define

$$\widetilde{Z}_t(\gamma) := (x_t', x_t' 1\{\widetilde{f}_t'\gamma > 0\})' \text{ and } Z_t^*(\gamma) := (x_t', x_t' 1\{f_t^{*\prime}\gamma > 0\})'. \tag{5.2}$$

For each $t = 1, \ldots, T$, construct $\{y_t^*\}_{t \le T}$ by

$$y_t^* := \widetilde{Z}_t(\widehat{\gamma})' \widehat{\alpha} + \eta_t \widehat{\varepsilon}_t \text{ with } \widehat{\varepsilon}_t := y_t - \widetilde{Z}_t(\widehat{\gamma})' \widehat{\alpha}, \tag{5.3}$$

where η_t is an i.i.d. sequence whose mean is zero and whose variance is one. For example, $\eta_t \sim \mathcal{N}(0,1)$ or it can be simulated from a discrete distribution (e.g., the Rademacher distribution). The bootstrap least-squares loss is given by

$$\mathbb{S}_T^*(\alpha, \gamma) := \frac{1}{T} \sum_{t=1}^{T} [y_t^* - Z_t^*(\gamma)' \alpha]^2. \tag{5.4}$$

In principle, the bootstrap analog of the original constraint is $h(\gamma) = h(\widehat{\gamma})$ and the bootstrap analogous LR is defined as

$$\widetilde{LR}^* := \frac{\min_{\alpha, h(\gamma)=h(\widehat{\gamma})} \mathbb{S}_T^*(\alpha, \gamma) - \min_{\alpha,\gamma} \mathbb{S}_T^*(\alpha, \gamma)}{\min_{\alpha,\gamma} \mathbb{S}_T^*(\alpha, \gamma)}.$$

A potential computational problem for $\widetilde{LR}^*$ is that it is necessary to fully solve two joint MIO problems: $\min_{\alpha,\gamma} \mathbb{S}_T^*(\alpha, \gamma)$ and $\min_{\alpha, h(\gamma)=h(\widehat{\gamma})} \mathbb{S}_T^*(\alpha, \gamma)$ in each of the bootstrap repetitions. To circumvent this problem, we adopt the approach of Andrews [1]. Because a solution based on the original data

should be close to a solution based on the bootstrapped data, within each bootstrap replication, we can employ the MILP algorithm, with $(\widehat{\alpha}, \widehat{\gamma})$ as the initial value, and iteratively update the algorithm for k steps rather than computing the full bootstrap solutions. A computationally convenient k-step LR statistic (LR_k^*) and its computational details are given in Algorithm 3.

Algorithm 3: Bootstrap for Estimated Factors

Input: $\{(y_t, x_t, \widetilde{f}_t, M_t, \widehat{\varepsilon}_t) : t = 1, \ldots, T\}$, $\widehat{\text{var}}(e_t)$, $\widehat{\Lambda}$, $\widehat{\alpha}$, $\widehat{\gamma}$, $\widehat{\gamma}_h$, B
Output: bootstrap critical value $\texttt{cv}_a^*$

Set $b = 1$;
while $b \leq B$ **do**
  Generate an i.i.d. sequence $\{\eta_t\}_{t \leq T}$ whose mean is zero and variance is one and an i.i.d. sequence of multivariate vectors $\{\mathcal{W}_t^*\}_{t \leq T}$ from $\mathcal{N}(0, I)$;
  Generate $\mathcal{Y}_t^* = \widehat{\Lambda}\widetilde{f}_t + \widehat{\text{var}}(e_t)^{1/2}\mathcal{W}_t^*, t = 1, ..., T$;
  Apply PCA to $\{\mathcal{Y}_t^*\}$ and obtain $\widetilde{F}_t^*$ as the PCA factor estimates;
  Compute H_T^* and $f_t^* = H_T^{*'-1}\widetilde{F}_t$, $t = 1, ..., T$;
  Construct $y_t^* = \widetilde{Z}_t(\widehat{\gamma})'\widehat{\alpha} + \eta_t\widehat{\varepsilon}_t$, $t = 1, ..., T$, where $\widetilde{Z}_t(\gamma) = (x_t', x_t' 1\{\widetilde{f}_t'\gamma > 0\})'$;
  Initialize at $\widehat{\gamma}^{*,0} = \widehat{\gamma}$, $\widehat{\gamma}_h^{*,0} = \widehat{\gamma}_h$;
  Set $l = 1$;
  while $l \leq k$ **do**
    Compute $\widehat{\alpha}^{*,l} = \alpha^*(\widehat{\gamma}^{*,l-1})$ and $\widehat{\alpha}_h^{*,l} = \alpha^*(\widehat{\gamma}_h^{*,l-1})$, where

$$\alpha^*(\gamma) = \left[\frac{1}{T}\sum_{t=1}^{T} Z_t^*(\gamma) Z_t^*(\gamma)'\right]^{-1} \frac{1}{T}\sum_{t=1}^{T} Z_t^*(\gamma) y_t^*;$$

    For the given $(\widehat{\alpha}^{*l}, \widehat{\alpha}_h^{*l})$, compute the following by MILP:

$$\widehat{\gamma}^{*,l} = \arg\min_{\gamma} \mathbb{S}_T^*(\widehat{\alpha}^{*,l}, \gamma),$$
$$\widehat{\gamma}_h^{*,l} = \arg\min_{h(\gamma)=h(\widehat{\gamma})} \mathbb{S}_T^*(\widehat{\alpha}_h^{*,l}, \gamma);$$

    Let $l = l + 1$;
  end
  Compute

$$LR_k^* := \frac{\mathbb{S}_T^*\left(\widehat{\alpha}_h^{*,k}, \widehat{\gamma}_h^{*,k}\right) - \mathbb{S}_T^*\left(\widehat{\alpha}^*, \widehat{\gamma}^*\right)}{\mathbb{S}_T^*\left(\widehat{\alpha}^*, \widehat{\gamma}^*\right)};$$

  Let $b = b + 1$;
end
Obtain $\texttt{cv}_a^*$ by the $(1 - a)$ th quantile of the empirical distribution of LR_k^*.

5.3. *Asymptotic Distribution.* To describe the asymptotic distribution of the quasi-likelihood ratio statistic, let σ_ε^2 be the variance of ε_t. In addition,

recall the asymptotic distributions of $\widehat{\gamma}$, the minimizer of

$$\mathbb{Q}(\omega, g) := A(\omega, g) + 2W(g),$$

and, as we discussed for Theorem 4.3, $\omega = \infty$ also corresponds to the case of known factors.

Note that $A(\omega, g)$ depends on the true value ϕ_0, the rotation matrix H, and the covariance matrix Σ_h. For the bootstrap sampling distribution, we consider drifting sequences around these values. For this, define

$$\begin{aligned}&\mathbb{A}(\omega, g, \Sigma, \bar{H}, \phi)\\ &:= M_\omega \mathbb{E}\left[(x_t d_0)^2 \left(\left|g_t' H g + \zeta_\omega^{-1} \mathcal{W}_t^{*\prime} \Sigma^{1/2} \bar{H}^{-1} \phi\right| - \left|\zeta_\omega^{-1} \mathcal{W}_t^{*\prime} \Sigma^{1/2} \bar{H}^{-1} \phi\right|\right) \middle| g_t'\phi = 0\right] p_{g_t'\phi}(0)\end{aligned}$$

for $\omega \in (0, \infty]$, and

$$\begin{aligned}&\mathbb{A}(0, g, \Sigma, \bar{H}, \phi)\\ &:= \mathbb{E}\left[(x_t' d_0)^2 (g_t' H g)^2 \middle| g_t'\phi = 0, \mathcal{W}_t^{*\prime} \Sigma^{1/2} \bar{H}^{-1} \phi = 0\right] p_{g_t'\phi, \mathcal{W}_t^{*\prime} \Sigma^{1/2} \bar{H}^{-1} \phi}(0, 0).\end{aligned}$$

Note that $A(\omega, g) = \mathbb{A}(\omega, g, H'\Sigma_h H, H, \phi_0)$.

ASSUMPTION 5.1. *(i) Uniformly for ϕ inside a neighborhood of ϕ_0, $\sup_{x_t, f_{2t}} |p_{\breve{g}_t'\phi | x_t, f_{2t}}(0) - p_{g_t'\phi_1 | x_t, f_{2t}}(0)| = o(1)$.*

(ii) For each fixed $\omega \in [0, \infty]$ and g, $\mathbb{A}(\omega, g, S)$ is continuous with respect to $S = (\Sigma, \bar{H}, \phi)$.

(iii) The factor idiosyncratic component e_t is independent of (x_t, g_t), and $|\widehat{\mathrm{var}}(e_t) - \mathrm{var}(e_t)|_2 = o_P(1)$ under the matrix spectral norm.

(iv) $\inf_\gamma |\widehat{f}_t^{*\prime}\gamma|$ has a density (jointly with respect to $(e_t, g_t, \mathcal{W}_t^*)$) bounded and continuous at zero, where $\widehat{f}_t^* = \widehat{f}_t + N^{-1/2}\widehat{\Sigma}_h^{1/2}\mathcal{W}_t^*$.

Fan, Liao, and Mincheva [11] showed that under mild sparsity assumptions, for the matrix spectral norm, $|\widehat{\mathrm{var}}(e_t) - \mathrm{var}(e_t)|_2 = o_P(1)$, given that $\log N$ does not grow too fast relative to T. The following theorem presents the asymptotic distribution of LR, and the validity of the k-step bootstrap procedure.

THEOREM 5.1. *Suppose that assumptions of Theorem 3.1 (for the known factor case) or assumptions of Theorem 4.3 (for the estimated factor case) and Assumption 5.1 hold. Let $h(\cdot)$ be a $\mathbb{R}^m$-valued linear function with a*

fixed m *and let* $r_{NT} := \left(NT^{1-2\varphi}\right)^{1/3} \wedge T^{1-2\varphi}$, *where we set* $N = T^2$ *in case of the known factor. Then, under* $\mathcal{H}_0 : h(\gamma_0) = \theta$, *we have*

$$\sqrt{r_{NT}T^{1+2\varphi}} \cdot LR \to^d \sigma_\varepsilon^{-2} \min_{g_h'\nabla h=0} \mathbb{Q}(\omega, g_h) - \sigma_\varepsilon^{-2} \min_g \mathbb{Q}(\omega, g),$$

and for any $k \geq 1$ *as the number of iterations in the* k*-step bootstrap,*

$$\sqrt{r_{NT}T^{1+2\varphi}} \cdot LR_k^* \to^{d^*} \sigma_\varepsilon^{-2} \min_{g_h'\nabla h=0} \mathbb{Q}(\omega, g_h) - \sigma_\varepsilon^{-2} \min_g \mathbb{Q}(\omega, g).$$

In the above, $\to^{d^*}$ *represents the convergence in distribution with respect to the conditional distribution of* $\{\eta_t, \mathcal{W}_t^*\}_{t\leq T}$ *given the original data. Also,* ∇h *denotes the gradient of* $h(\cdot)$, *which is independent of* γ_0 *as* h *is linear.*

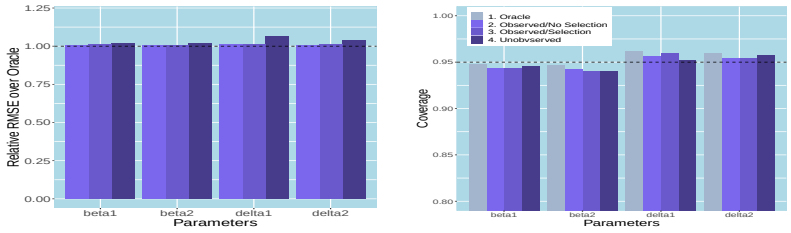
6. Monte Carlo Experiments. In this section, we study the finite sample properties of the proposed method via Monte Carlo experiments. The data are generated from the following design:

$$y_t = x_t'\beta_0 + x_t'\delta_0 1\{g_t'\phi_0 > 0\} + \varepsilon_t \quad \text{for} \quad t = 1, \ldots, T,$$

where $\varepsilon_t \sim N(0, 0.5^2)$, $x_t \equiv (1, x_{2,t}')'$, and $g_t \equiv (g_{1,t}', -1)'$. Both $x_{2,t}$ and $g_{1,t}$ follow the vector autoregressive model of order 1: $x_{2,t} = \rho_x x_{2,t-1} + \nu_t, g_{1,t} = \rho_g g_{1,t-1} + u_t$, where $\nu_t \sim N(0, I_{d_x-1})$ and $u_t \sim N(0, I_K)$. When the factor g_t is not observable, we instead observe $\mathcal{Y}_t$ that is generated from $\mathcal{Y}_t = \Lambda g_{1,t} + \sqrt{K}e_t, e_t = \rho_e e_{t-1} + \omega_t$, where $\mathcal{Y}_t$ is an $N \times 1$ vector and ω_t is an i.i.d. innovation generated from $N(0, I_N)$. The terms ε_t, ν_t, u_t, and ω_t are mutually independent.

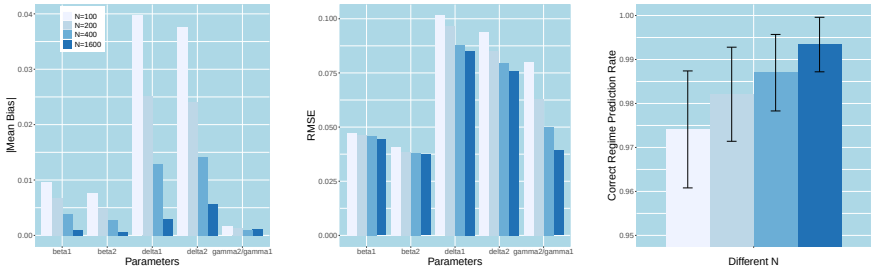
In the baseline model, we set $T = N = 200$, $d_x = 2$, and $K = 3$, and apply the MIQP algorithm. The additional parameter values are set as follows: $\beta_0 = \delta_0 = (1, 1)$; $\phi_0 = (1, 2/3, 0, 2/3)$; $\rho_x = \text{diag}(0.5, \ldots, 0.5)$; $\rho_g = \text{diag}(\rho_{g,1}, \ldots, \rho_{g,K})$, where $\rho_{g,k} \sim U(0.2, 0.8)$ for $k = 1, \ldots, K$, the ith row of Λ, $\lambda_i' \sim N(0', K \cdot I_K)$; and $\rho_e = \text{diag}(\rho_{e,1}, \ldots, \rho_{e,N})$, where $\rho_{e,i} \sim U(0.3, 0.5)$ for $i = 1, \ldots, N$. The values of ρ_g and ρ_e are drawn only once and kept for the whole replications. The factor model design is similar to Bai and Ng [6] and Cheng and Hansen [10]. All simulation results are based on 1,000 replications unless otherwise mentioned. We use a desktop computer equipped with an AMD RYZEN Threadripper 1950X CPU (16 cores with 3.4 GHz) and 64 GB RAM. The replication R codes for both the Monte Carlo experiments and empirical applications are available at `https://github.com/yshin12/fadtwo`. Also, the full simulation results can be found in Tables A-2–A-8 in Online Appendix J.

FIG 4. *Simulation Results: Baseline Model*



First, we study the baseline model under four scenarios: (i) when we know the correct regime, i.e. ϕ_0, (Oracle); (ii) when we observe g_t and know that the third factor is irrelevant (Observed Factors/No Selection); (iii) when we observe g_t and have to select the relevant factors (Observed Factors/Selection); and (iv) when we do not observe g_t but estimate factors from $\mathcal{Y}_t$ by PCA. We set the dimension of γ to be 4 in (iv). Figure 4 reports the relative size of the root-mean-square errors (RMSEs) for β, δ as well as the coverage rate for the 95% confidence intervals. As predicted by the asymptotic theory in the previous sections, the relative RMSEs over Oracle are close to 1 in all scenarios. The coverage rates for the 95% confidence intervals are also close to the nominal value. Not surprisingly, these results on α are based on the good estimation performance of ϕ (or γ).

FIG 5. *Unobserved Factors with Different N*



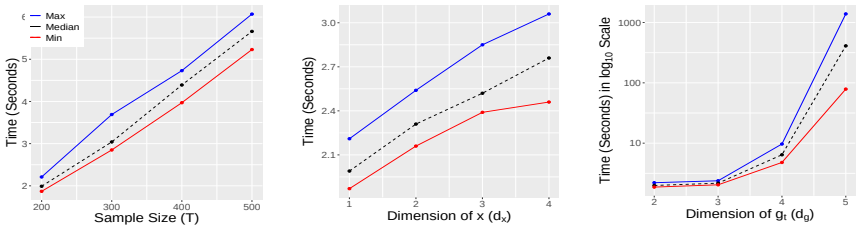
Note. The whisker plot in the panel on the right denotes one standard deviation computed over replication draws.

Second, we focus on the unobserved factor model and investigate the performance as N increases. For each simulated sample of $\{y_t, x_t, g_t\}$, we generate $\mathcal{Y}_t$ with $N = 100, 200, 400, 1600$. We use the same baseline design with $T = 200$, $d_x = 2$, but $K = 1$ to speed up computations. Figure 5 summarizes the results. The regimes are predicted more precisely as N increases and the performance of the estimator improves. We observe relatively more

TABLE 1
Size of Bootstrap Test

Null hypothesis	Scenarios	Significance level	
		5%	1%
$H_0 : \gamma_{02} = 0$	Estimated factor	3.8%	0.7%
$H_0 : \phi_{02} = 0$	Known factor	4.3%	0.5%
$H_0 : \phi_{02} = 0$ and $\phi_{03} = 0$	Many known factors	7.5%	1.1%

FIG 6. *Computation Time over T, d_x, and d_g*



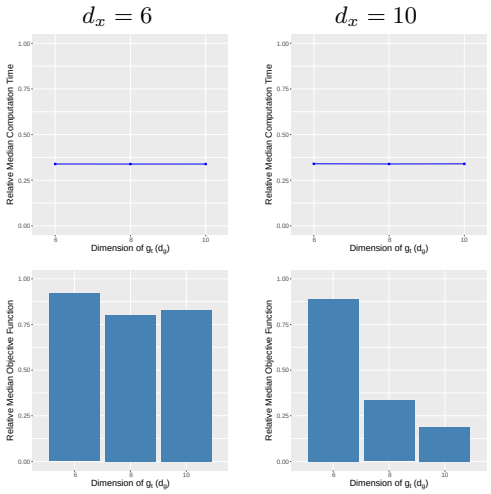
improvements in γ rather than α. This is because $\widehat{\alpha}$ already enjoys the oracle property, provided that $T = O(N)$.

Third, we investigate the performance of the bootstrap test under three scenarios: (i) an estimated factor; (ii) a known factor; (iii) many known factors. The parameters are set as follows: $T = 200$, $N = 400$, $B = 499$, $\varepsilon_t \sim N(0,1)$, $\eta_t \sim N(0,1)$, $\beta_0 = (1,1)$, $\delta_0 = (0.5, 0.5)$, γ_0(or ϕ_0) $= (1,0)$ in (i) and (ii), and $\phi_0 = (1,0,0,0)$ in (iii). We test a simple null hypothesis of $H_0 : \gamma_{02}$(or ϕ_{02}) $= 0$ in (i) and (ii) and a joint hypothesis of $H_0 : \phi_{02} = \phi_{03} = 0$ in (iii). There is no serial correlation in the model ($\rho_x = \rho_g = \rho_e = 0$). Table 1 reports the size of the bootstrap test in each scenario and it is satisfactory but we observe over-rejection in the joint hypothesis case.

We next investigate the computation time. We start from a set of simple models and extend to large dimensional models. We simplify the baseline model by considering scenario (ii) (i.e., Observed/No Selection), and by setting $\rho_x = \rho_g = 0$. The results are based on 100 replications. We set $T = 200$, $d_x = 1$, and $d_g = 2$, initially and increase each dimension as follows: $T = \{200, 300, 400, 500\}$, $d_x = \{1, 2, 3, 4\}$ while keeping $T = 200$ and $d_g = 2$; $d_g = \{2, 3, 4, 5\}$ while keeping $T = 200$ and $d_x = 1$. Figure 6 reports the computation time of MIQP. The results indicate that the computation time stays in a reasonable bound and increases linearly as T and d_x increase. However, it increases exponentially as d_g increases.

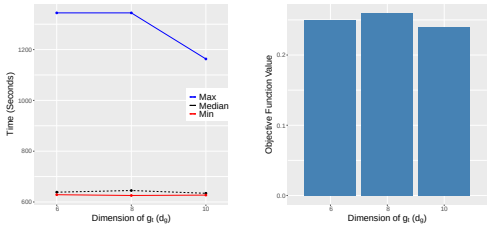
We now consider large dimensional models and handle the computational challenge by implementing the BCD algorithm in addition to MIQP. We extend the dimension of the models as $T = \{500, 1000\}$, $d_x = \{6, 8, 10\}$,

FIG 7. *Large Dimensional Models ($T = 500$)*



Note. The relative measures are calculated by dividing the outcome of BCD by that of MIQP.

FIG 8. *Larger Dimensional Models using BCD ($T = 1000$ and $d_x = 6$)*



and $d_g = \{6, 8, 10\}$. Note that $d_g = 10$ would be quite challenging and the standard grid search method would be infeasible in practice with $T =$ 1,000. The results are based on 10 iterations of each model. We set the total time budget as 1,800 seconds for both MIQP and BCD so that each estimation terminates after that even if it does not converge. In BCD, we set `MaxTime_1`=600 (seconds) and `MaxTime_2`=60 (seconds). Figure 7 reports the ratio of the median computation time and median objective function values between BCD and MIQP when $T = 500$. BCD spends a third of the computation time, whereas MIQP spends the total time budget. BCD achieves better objective function values in all cases and the performance of MIQP deteriorates quickly as d_g increases when $d_x = 10$. Figure 8 reports the summary statistics of computation time and the median objective function values of BCD when $T = 1{,}000$. As the computation is more challenging, we

observe that the maximum computation time is higher for all d_g. However, the median computation time is still around 600 seconds and the achieved objective function values are quite stable.

Based on our simulation studies, we propose to use the BCD algorithm by assigning 1/3 of the total time budget into the maximum time (`MaxTime_1`) for Step 1 (MIQP). When the global solution is not attainable within `MaxTime_1`, the BCD algorithm would switch into Steps 2–3 (MILP) automatically. We recommend assigning 1/30 of the total time budget into the maximum time (`MaxTime_2`) for each cycle of Step 2.

In summary, the simulation studies reveal that the proposed method achieves the properties predicted by the asymptotic theory, especially the oracle property of α and the inference based on the bootstrap method. The BCD algorithm also shows quite satisfactory results in a large dimensional change-point model whose computation is infeasible with grid search.

7. Classifying the Regimes of US Unemployment. We revisit the empirical application of Hansen [15], who considered threshold autoregressive models for the US unemployment rate. Specifically, Hansen [15] used monthly unemployment rates (i.e., u_t) for males age 20 and over, and set $y_t = \Delta u_t$ in (1.1). The lag length in the autoregressive model was $p = 12$ and the preferred threshold variable was $q_{t-1} = u_{t-1} - u_{t-12}$. In this section, we investigate the usefulness of using unknown but estimated factors. We use the first principal component (i.e., F_t) of Ludvigson and Ng [21] that is estimated from 132 macroeconomic variables. This factor not only explains the largest fraction of the total variation in their panel data set but also loads heavily on employment, production, and so on. Ludvigson and Ng call it a *real factor* and thus it is a legitimate candidate for explaining the unemployment rate. We consider three different specifications for f_t: (1) $f_{1t} = (q_{t-1}, -1)$, (2) $f_{2t} = (F_{t-1}, -1)$, and (3) $f_{3t} = (q_{t-1}, F_{t-1}, -1)$. We combined the updated estimates of the real factor, which are available on Ludvigson's web page at `https://www.sydneyludvigson.com`, with Hansen's data, yielding a monthly sample from March 1960 to July 1996.

Table 2 reports estimation results that are obtained by the MIQP algorithm. We show the goodness of fit by reporting the average squared residuals and also the results of regime misclassification relative to the NBER business cycle dates. The latter is obtained by

$$\frac{1}{T}\sum_{t=1}^{T}\left|1\left\{f_{jt}'\widehat{\gamma}_j > 0\right\} - 1_{\mathrm{NBER},t}\right| \quad \text{for each } j = 1, 2, 3, \tag{7.1}$$

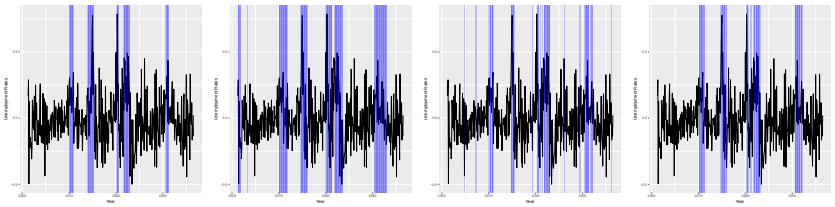
where $1_{\mathrm{NBER},t}$ is the indicator function that has value 1 if and only if the

TABLE 2
Estimation Results

Specification	(1) $f_{1t} = (q_{t-1}, -1)$	(2) $f_{2t} = (F_{t-1}, -1)$	(3) $f_{3t} = (q_{t-1}, F_{t-1}, -1)$
Regime 1 ("Expansion")	$q_{t-1} \leq 0.302$	$F_{t-1} \leq -0.28$	$q_{t-1} + 3.55F_{t-1} \leq -1.60$
Prediction error	0.0264	0.0272	0.0252
Classification error	0.193	0.106	0.104

Note. See Table A-1 in the Online Appendix for estimated coefficients and their heteroskedasticity-robust standard errors. Regime 2 ("Contraction") is the complement of regime 1. "Prediction Error" refers to the average of squared residuals ($T^{-1}\sum_{i=1}^{T}\widehat{\varepsilon}_t^2$). "Classification Error" corresponds to the proportion of misclassification defined in (7.1).

FIG 9. *Regime Classification*



Note. The leftmost panel shows NBER recession dates in the shaded area, and the other three panels display those with specifications (1), (2) and (3), respectively.

economy is in contraction according to the NBER dates. Accordingly, we label regime 1 "expansion" and regime 2 "contraction", respectively. Figure 9 gives the graphical representation of regime classification. Specification (1) suffers from the highest level of misclassification and tends to classify recessions more often than NBER. Specification (2) mitigates the misclassification risk but at the expense of a worse goodness of fit. On one hand, the threshold autoregressive model solely by q_{t-1} fittingly explains the unemployment rate but is short of classifying the overall economic conditions satisfactorily. On the other hand, the model based only on F_{t-1} is adequate at describing the underlying overall economy but does not explain the unemployment rate well. It turns out that specification (3) has the lowest misclassification error and best explains unemployment. Thus, we have shown the real benefits of using a vector of possibly unobserved factors to explain the unemployment dynamics.

As an additional check, we tested the null hypothesis of no threshold effect. The resulting p-value is 0.002 based on 500 bootstrap replications, thus providing strong evidence for the existence of two regimes. See Table

A-1 in the Online Appendix for details and additional results.

8. Conclusions. We have proposed a new method for estimating a two-regime regression model where regime switching is driven by a vector of possibly unobservable factors. We show that our optimization problem can be reformulated as MIO and have presented two alternative computational algorithms. We have also derived the asymptotic distribution of the resulting estimator under the scheme that the threshold effect shrinks to zero as the sample size tends to infinity. As a possible interesting extension, we can consider nonparametric regime switching, where the switching indicator is replaced by $1\{F(w_t) > 0\}$ with a vector of observables w_t and a nonparametric function $F(\cdot)$. We intend to study this in the future.

SUPPLEMENTARY MATERIAL

Factor-Driven Two-Regime Regression: Online Appendix: (). The online appendix contains additional results and all the proofs. We also propose an ℓ_0-penalized factor selection procedure to select the active factors, as well as testing the linearity of the model in (1.1), $\mathcal{H}_0 : \delta_0 = 0$.

References.

[1] Andrews, D. W. (2002): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators," *Econometrica*, 70(1), 119–162.

[2] Bai, J. (1994): "Least squares estimation of a shift in linear processes," *Journal of Time Series Analysis*, 15(5), 453–472.

[3] Bai, J. (2003): "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135–171.

[4] Bai, J., and S. Ng (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74(4), 1133–1150.

[5] ——— (2008): "Extremum Estimation when the Predictors are Estimated from Large Panels," *Annals of Economics and Fiance*, 9(2), 201–222.

[6] ——— (2009): "Boosting diffusion indices," *J. Appl. Econom.*, 24(4), 607–629.

[7] Bai, J., and P. Perron (2003): "Computation and analysis of multiple structural change models," *J. Appl. Econom.*, 18(1), 1–22.

[8] Bertsimas, D., A. King, and R. Mazumder (2016): "Best subset selection via a modern optimization lens," *Annals of Statistics*, 44(2), 813–852.

[9] Chan, K.-S. (1993): "Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model," *Annals of Statistics*, 21(1), 520–533.

[10] Cheng, X., and B. E. Hansen (2015): "Forecasting with factor-augmented regression: A frequentist model averaging approach," *J. Econom*, 186(2), 280–293.

[11] Fan, J., Y. Liao, and M. Mincheva (2013): "Large covariance estimation by thresholding principal orthogonal complements (with discussion)," *Journal of the Royal Statistical Society, Series B*, 75, 603–680.

[12] Gonçalves, S., and B. Perron (2014): "Bootstrapping factor-augmented regression models," *J. Econom*, 182(1), 156–173.

[13] ——— (2019): "Bootstrapping factor models with cross sectional dependence," *J. Econom*, forthcoming.

[14] Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*. Springer.
[15] Hansen, B. E. (1997): "Inference in TAR Models," *Studies in Nonlinear Dynamics and Econometrics*, 2(1), 1–14.
[16] ——— (2000): "Sample splitting and threshold estimation," *Econometrica*, 68(3), 575–603.
[17] Hawkins, D., A. Gallant, and W. Fuller (1986): "A simple least squares method for estimating a change in mean," *Communications in Statistics-Simulation and Computation*, 15(3), 523–530.
[18] Horváth, L., and P. Kokoszka (1997): "The effect of long-range dependence on change-point estimators," *Journal of Statistical Planning and Inference*, 64(1), 57–81.
[19] Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18(1), 191–219.
[20] Ling, S. (1999): "On the probabilistic properties of a double threshold ARMA conditional heteroskedastic model," *Journal of Applied Probability*, 36(3), 688–705.
[21] Ludvigson, S. C., and S. Ng (2009): "Macro Factors in Bond Risk Premia," *Review of Financial Studies*, 22(12), 5027–5067.
[22] McKeague, I. W., and B. Sen (2010): "Fractals with point impact in functional linear regression," *Annals of statistics*, 38(4), 2559–2586.
[23] Merlevède, F., M. Peligrad, and E. Rio (2011): "A Bernstein type inequality and moderate deviations for weakly dependent sequences," *Probability Theory and Related Fields*, 151(3), 435–474.
[24] Qu, Z., and D. Tkachenko (2017): "Global Identification in DSGE Models Allowing for Indeterminacy," *Review of Economic Studies*, 84(3), 1306–1345.
[25] Seijo, E., and B. Sen (2011): "Change-point in stochastic design regression and the bootstrap," *Annals of Statistics*, 39(3), 1580–1607.
[26] Seo, M. H., and O. Linton (2007): "A smoothed least squares estimator for threshold regression models," *J. Econom*, 141(2), 704–735.
[27] Seo, M. H., and T. Otsu (2018): "Local M-estimation with discontinuous criterion for dependent and limited observations," *The Annals of Statistics*, 46(1), 344–369.
[28] Tong, H. (1990): *Non-linear time series: a dynamical system approach*. Oxford University Press.
[29] van der Vaart, A., and J. Wellner (1996): *Weak convergence and empirical processes*. Springer, New York.

Address:420 West 118th Street,
New York, NY 10027, USA
E-mail: sl3841@columbia.edu

Address: 1 Gwanak-ro, Gwanak-gu,
Seoul 08826, Korea
E-mail: myunghseo@snu.ac.kr

Address: 75 Hamilton St.,
New Brunswick, NJ 08901, USA
E-mail: yuan.liao@rutgers.edu

Address: 1280 Main St.W.,
Hamiloton, ON L8S 4L8, Canada.
E-mail: shiny11@mcmaster.ca