Econometric Theory, 2018, Page 1 of 45.
doi:10.1017/S0266466618000245

THE FACTOR-LASSO AND *K*-STEP BOOTSTRAP APPROACH FOR INFERENCE IN HIGH-DIMENSIONAL ECONOMIC APPLICATIONS

CHRISTIAN HANSEN
University of Chicago

YUAN LIAO
Rutgers University

We consider inference about coefficients on a small number of variables of interest in a linear panel data model with additive unobserved individual and time specific effects and a large number of additional time-varying confounding variables. We suppose that, in addition to unrestricted time and individual specific effects, these confounding variables are generated by a small number of common factors and high-dimensional weakly dependent disturbances. We allow that both the factors and the disturbances are related to the outcome variable and other variables of interest. To make informative inference feasible, we impose that the contribution of the part of the confounding variables not captured by time specific effects, individual specific effects, or the common factors can be captured by a relatively small number of terms whose identities are unknown. Within this framework, we provide a convenient inferential procedure based on factor extraction followed by lasso regression and show that the procedure has good asymptotic properties. We also provide a simple *k*-step bootstrap procedure that may be used to construct inferential statements about the low-dimensional parameters of interest and prove its asymptotic validity. We provide simulation evidence about the performance of our procedure and illustrate its use in an empirical application.

1. INTRODUCTION

The availability of rich, high-dimensional data for use in empirical analyses is rapidly increasing. High-dimensional data offer many opportunities, but informative data analysis in high-dimensional data requires the imposition of dimension reducing structure. Two distinct structures which are common in the econometrics

The authors are grateful to Shakheeb Khan, Roger Moon, and seminar participants at the Australasian Meetings of the Econometric Society, Inference in Large Econometric Models at Montréal, University of Chile, National University of Singapore, Xiamen University, University of Toronto, and Stevens Institute of Technology for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 1558636 and the University of Chicago Booth School of Business. First version: June 2016. This version: July 23, 2018. Address correspondence to Christian Hansen, Booth School of Business, University of Chicago, Chicago, IL 60637, USA; e-mail: Christian.Hansen@chicagobooth.edu and Yuan Liao, Department of Economics, Rutgers University, New Brunswick, NJ 08901, USA; e-mail: yuan.liao@rutgers.edu.

literature are sparse structures and factor structures. In this article, we consider estimation and inference on a low-dimensional parameter of interest within a panel data model that accommodates both sparse and factor structures.

Specifically, we consider a linear panel model defined by

$$y_{it} = \alpha d_{it} + \xi_t' f_i + U_{it}'\theta + g_i + \nu_t + \epsilon_{it}, \tag{1.1}$$

$$d_{it} = \delta_{dt}' f_i + U_{it}'\gamma_d + \zeta_i + \mu_t + \eta_{it}, \tag{1.2}$$

$$X_{it} = \Lambda_t f_i + w_i + \rho_t + U_{it}, \tag{1.3}$$

where $i \leq n$ indexes cross-sectional observations, $t \leq T$ indexes time series observations, X_{it} are observed confounding variables, and d_{it} is an a priori specified "treatment" variable whose coefficient α is the parameter of interest.[1] In (1.1)–(1.3), f_i is a $K \times 1$ vector of individual-specific unobservables; and ξ_t, δ_{dt}, and Λ_t are, respectively, $K \times 1$, $K \times 1$, and $p \times K$ dimensional time-specific unobservables.[2] In each equation, we also allow for unrestricted additive unobserved individual effects, (g_i, ζ_i, w_i'), and time specific effects, (ν_t, μ_t, ρ_t'), where g_i, ζ_i, ν_t, and μ_t are scalars and w_i and ρ_t are $p \times 1$ vectors. The term U_{it} represents the part of the observed X_{it} that is orthogonal to the individual-specific unobservable f_i and additive unobserved time and individual specific heterogeneity. We allow U_{it} to be correlated with both the outcome and variable of interest. Because U_{it} is high-dimensional, we impose that θ and γ_d are sparse to facilitate informative estimation and inference for α. Following Hahn, Mukeherjee, and Carvalho (2013), we refer to the model (1.1)–(1.3) as the "panel partial factor model" (PPFM).[3] We note that the only observed variables in (1.1)–(1.3) are (y_{it}, d_{it}, X_{it}).

Deviating from much of the literature on factor models, we use subscript i to denote the common factors, with the understanding that the common factors are individual-specific. This treatment is motivated by microeconometric, "short T," applications where a major concern when trying to learn structural effects is confounding due to unobserved, individual specific attributes. For example, in a state-level panel, one may believe that unobserved features that are potentially confounded with policies of interest are largely captured by a few state specific factors such as state laws or policies or state-level social preferences that may reasonably be taken as time invariant over moderate time horizons but may have time-varying associations with the variables of interest. The structure in (1.3) then posits that such factors are associated with many time-varying state-level observables. Finally, the inclusion of the factor residuals in the equations (1.1)–(1.2) allows for the confounding between the variable of interest and time varying observables that are uncaptured by the latent factors, whose presence is reasonable in many applications. We note that the inclusion of these factor residuals is analogous to the conventional inclusion of time-varying observables in additive linear fixed effects models. Of course, the PPFM framework also accommodates many other environments.

The first contribution of the present article is offering a practical estimation and inference procedure that is appropriate for inference for α in the PPFM and providing a formal treatment of the procedure's theoretical properties. Specifically, we proceed by first running a factor extraction step and taking residuals from regressing each observed variable on the estimated factors. Using these residuals, we then follow the lasso-based estimation and inference procedures of Belloni, Chernozhukov, Hansen, and Kozbur (2016). We show that the resulting estimator of α is asymptotically normal with readily estimated asymptotic variance under sensible conditions. These conditions allow for errors in selection of the elements of U_{it} that load after controlling for the factors but maintain sufficiently strong conditions to allow oracle selection of the number of factors. The theoretical analysis is substantially complicated by the fact that factors and factor-residuals are not observed and must be extracted from the data. The estimation error in this extraction then enters the second step nonlinear and nonsmooth lasso problem. Due to this complication, the theoretical results in this article make use of arguments that, to our knowledge, are not implied by results existing in the current factor modeling literature or the current lasso literature which may be of interest outside of the present article.

By addressing estimation and inference in an interesting high-dimensional factor augmented regression model appropriate for panel data, our article complements the large factor model literature and the rapidly growing literature dealing with obtaining valid inferential statements following regularized estimation. See, for example, Bai (2003), Bai and Ng (2002), Stock and Watson (2002), and Fan, Xue, and Yao (2017) for fundamental references on factor models in econometrics and Bai and Ng (2006) and Bernanke, Boivin, and Eliasz (2005) for factor augmented regression. For approaches to obtaining valid inferential statements in a variety of different high-dimensional settings, see, for example, Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, Fernández-Val, and Hansen (2017), Belloni, Chernozhukov, and Hansen (2014), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2016), Dezeure, Bühlmann, and Zhang (2017), Fan and Li (2001), van de Geer, Bühlmann, Ritov, and Dezeure (2014), Wager and Athey (2017), and Zhang and Zhang (2014).

As a second contribution, we offer a new, computationally convenient bootstrap method for inference. Specifically, we consider a bootstrap where we apply our main procedure, including extraction of factors and lasso estimation steps, within each bootstrap replication. As computation of the lasso estimator within each bootstrap sample may be demanding, we use a *k*-step bootstrap following Andrews (2002) where we start at the lasso solution from the full sample and then iterate a numeric solution algorithm for the lasso estimator for *k*-steps. We make use of solution algorithms for which the updates are available in closed form which leads to fast computation. The *k*-step bootstrap we propose complements other bootstrap procedures that have been proposed for lasso-based inference, for example, Belloni et al. (2017), Chatterjee and Lahiri (2011), Chernozhukov, Chetverikov, and Kato (2013), and Dezeure et al. (2017). The approach we take

is something of a middle ground between Chernozhukov et al. (2013), which uses resampling of model scores to avoid recomputation of the lasso estimator, and Dezeure et al. (2017) which fully recomputes the lasso solution within each bootstrap replication. The former approach is computationally convenient and asymptotically valid but does not capture any finite sample uncertainty introduced in the lasso selection, while the latter may be computationally cumbersome due to fully recomputing the lasso solution within each iteration. We note that the bootstrap procedure could be easily applied outside of the specific model considered in this article and that the technical analysis here is new and may be of interest in other contexts.

The remainder of this article is organized as follows. In Section 2, we provide further motivation of the PPFM and outline the basic algorithm we will employ for inference. We present formal results for the proposed procedure in Section 3. Section 4 describes the k-step bootstrap approach in detail and provides a formal analysis establishing the validity of the resulting bootstrap inference. We then provide simulation and empirical examples in Section 5. Key proofs are collected in an appendix with additional results provided online in supplementary material associated with this article, available at Cambridge Journals Online (`journals.cambridge.org/ect`).

1.1. Notation and Asymptotic Sequence

Throughout the article, we use $\|\beta\|_1$ and $\|\beta\|_2$ to, respectively, denote the ℓ_1- and ℓ_2- norms of a vector β; and we use $\|A\|$ and $\|A\|_F$ to, respectively, denote the spectral and Frobenius norms of a matrix A. In addition, we denote the cardinality of a finite set J as $|J|_0$. Finally, for two positive sequences a_n, b_n, we write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$.

We will take asymptotics where $\dim(X_{it}) = p \to \infty$, $n \to \infty$, and T is either fixed or growing slowly relative to n and p when stating our formal results, and we explicitly allow for scenarios where $p \gg nT$. Having T fixed or increasing slowly captures microeconometric applications where T is typically similar to or smaller than n. The number of factors K is assumed fixed throughout the article. For simplicity in the formal development, we also assume that K is known a priori, though our procedure admits data-dependent methods (e.g., Bai and Ng, 2002; or Ahn and Horenstein, 2013) for selecting K so long as the estimated K is consistent; and we strongly recommend that such methods be employed in practice.

2. DISCUSSION OF THE PANEL PARTIAL FACTOR MODEL AND THE FACTOR-LASSO ALGORITHM

In this section, we first discuss the panel partial factor model with an emphasis on relating it to high-dimensional sparse linear models and conventional factor augmented regression models. We then outline our procedure for estimating and doing inference for the treatment parameter of interest, α in (1.1).

2.1. Panel Partial Factor Model

The PPFM defined in (1.1)–(1.3) offers a simple generalization of the high-dimensional sparse linear fixed effects model and a factor augmented regression model. This generalization allows us to capture features that may be missed in either of these useful baseline models.

Specifically, the PPFM generalizes the high-dimensional sparse fixed effects model examined in Belloni et al. (2016). (1.1)–(1.3) clearly reduces to the high-dimensional sparse fixed effects model when $\xi_t' f_i$, $\delta_{dt}' f_i$, and all elements of $\Lambda_t f_i$ are 0 for all i and t. Relative to the high-dimensional sparse fixed effects model, the PPFM accommodates an important case of strong dependence among the columns of the observed control variable X via the latent factor structure $\Lambda_t f_i$. Importantly, the PPFM also allows all of the X variables to be confounded with the treatment, captured by f_i. In this case, trying to estimate α via the sparse high-dimensional fixed effects model as in Belloni et al. (2016) could fail as the confounding may not be captured via controlling directly for a small number of the observed X variables.

The PPFM also shares many features with factor augmented regression models; e.g., Bai and Ng (2006) and Bernanke et al. (2005). The key difference between the standard factor augmented regression model and the PPFM is the presence of the unobserved high-dimensional vector U_{it} in (1.1) and (1.2). Adding U_{it} to (1.1) and (1.2) can be justified by noting that the U_{it} contain any explanatory power remaining in X_{it} after controlling for common factors. In many settings, it seems reasonable to believe that the factor structure may fail to capture all sources of confounding but that any confounding not captured by the latent factor structure is concentrated among only a few variables. We choose to include U_{it} instead of X_{it} as control variables because the components of U_{it} are pairwise weakly correlated and are orthogonal to f_i, which facilitates the identification and estimation of (γ_d, θ).[4]

Another potential approach to accounting for the possibility that a low-dimensional factor structure may not account for all the confounding variation in X_{it} would be to consider a growing number of factors ($K \to \infty$) and penalizing the coefficients on the factors. The major difficulty of this approach is that K must grow very slowly compared to p; otherwise, there may be insufficient information available to even consistently estimate the individual factors. Using a slowly growing K without augmenting with factor residuals as in the PPFM then effectively reduces to a pure low-dimensional factor model, which potentially leaves a significant portion of the information in X_{it} unexploited.

The PPFM is also related to, but distinct from, interactive fixed effects models as in Bai (2009), Bai and Li (2014), Moon and Weidner (2017, 2015), Pesaran (2006), and Su and Chen (2013).[5] A simple version of the interactive fixed effects model analogous to (1.1) is

$$y_{it} = \alpha d_{it} + z_{it}'\beta + \lambda_t f_i + \epsilon_{it}.$$

In this model, z_{it} represents a known, low-dimensional set of variables that must be controlled for in addition to the factors in f_i. There appear to be two key distinctions between the high-dimensional PPFM and interactive fixed effects approaches. First, we relax the assumption that one knows the exact identity of the variables that should appear in the model, z_{it}, by allowing for a high-dimensional set of observed potential confounds in X_{it}. Second, we directly extract estimates of the factors and U from X which can proceed even when T is small, whereas most approaches in the interactive fixed effects model take T to be large. We thus view the PPFM and interactive fixed effects approaches as complementary where one may prefer one or the other depending on the nature of the data at hand.

Finally, our article is related to the interesting work of Hsiao, Ching, and Wan (2012) and Li and Bell (2017) but differs in a few key regards. First, the treatment variable of interest in Hsiao et al. (2012) and Li and Bell (2017) appears on only one or finitely many individuals after a specific time period $t > T_0$, and these articles use a factor model to predict the counterfactual outcomes for the periods $t > T_0$. In our model, we consider inference for the coefficient of a generic time-varying variable of interest that may change continuously across all individuals and time periods. Second, Hsiao et al. (2012) and Li and Bell (2017) do not estimate the unknown factors but instead use the factor structure to show that the outcome variable can be written as a linear combination of the other observed outcome variables. Third, while Li and Bell (2017) suggest using lasso in a high-dimensional setting, they provide formal results only in a low-dimensional setting. We thus again view the approaches as complementary where the preference would depend on the specifics in a given empirical setting.

We conclude this section by noting that it is possible to check whether the high-dimensional regressors $\{X_{it}\}$ admit a factor structure. In practice, researchers should employ a consistent estimator for the number of factors (K) as in, e.g., Ahn and Horenstein (2013). If $K > 0$ is estimated, one may apply our full procedure including factor extraction, and a pure sparsity-based approach may be applied otherwise. Of course, our full procedure also reduces to the pure sparsity-based approach when $K = 0$ is estimated.

2.2. Estimation Algorithm

We take the following steps to estimate α. The estimation algorithm adapts the approach from Belloni et al. (2016) to allow for the estimation of factors.

Step 1: Remove the unobserved heterogeneity. We begin by taking the within transformation of all observed variables to remove the additive fixed effects. To this end, let $\tilde{M}_t = M_t - \bar{M}$ for any collection of matrices M_t indexed only by t and $\bar{M} = \frac{1}{T}\sum_{t=1}^{T} M_t$, and let $\tilde{f}_i = f_i - \frac{1}{n}\sum_{i=1}^{n} f_i$. Also, let $\tilde{z}_{it} = z_{it} - \bar{z}_{\cdot t} - \bar{z}_{i\cdot} + \bar{\bar{z}}$ for any variable z_{it} where $\bar{z}_{\cdot t} = \frac{1}{n}\sum_{i=1}^{n} z_{it}$, $\bar{z}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} z_{it}$, and $\bar{\bar{z}} = \frac{1}{nT}\sum_{i=1,t=1}^{n,T} z_{it}$. We can then define a demeaned model as

$$\tilde{y}_{it} = \alpha \tilde{d}_{it} + \tilde{\xi}_t' \tilde{f}_i + \tilde{U}_{it}'\theta + \tilde{\epsilon}_{it}, \tag{2.1}$$

$$\tilde{d}_{it} = \tilde{\delta}_{dt}' \tilde{f}_i + \tilde{U}_{it}'\gamma_d + \tilde{\eta}_{it}, \tag{2.2}$$

$$\tilde{X}_{it} = \tilde{\Lambda}_t \tilde{f}_i + \tilde{U}_{it}. \tag{2.3}$$

Step 2: Estimate $\tilde{f}_i, \tilde{U}_{it}$ and $\tilde{\Lambda}_t$. We estimate the (demeaned) latent factors as well as the (demeaned) idiosyncratic components from the model $\tilde{X}_{it} = \tilde{\Lambda}_t \tilde{f}_i + \tilde{U}_{it}$.[6] Let $\widehat{F} = (\widehat{f}_1, \ldots, \widehat{f}_n)'$ be the $n \times K$ matrix of estimated factors. We discuss estimation of $\widehat{F}$ via principal components analysis in Supplementary Appendix D. Given $\widehat{F}$, we estimate $\tilde{\Lambda}_t$ and $\tilde{U}_{it}$ by least squares:

$$\widehat{\Lambda}_t = \sum_{i=1}^{n} \tilde{X}_{it} \widehat{f}_i' (\widehat{F}'\widehat{F})^{-1}, \quad \widehat{U}_{it} = \tilde{X}_{it} - \widehat{\Lambda}_t \widehat{f}_i, \quad i \le n,\ t \le T. \tag{2.4}$$

Step 3: Estimate coefficients on $\tilde{f}_i$. Substituting (2.2) to (2.1), we obtain

$$\begin{aligned} \tilde{y}_{it} &= \alpha(\tilde{\delta}_{dt}' \tilde{f}_i + \tilde{U}_{it}'\gamma_d + \tilde{\eta}_{it}) + \tilde{\xi}_t' \tilde{f}_i + \tilde{U}_{it}'\theta + \tilde{\epsilon}_{it} \\ &:= \tilde{\delta}_{yt}' \tilde{f}_i + \tilde{U}_{it}'\gamma_y + \tilde{e}_{it}, \end{aligned}$$

where $\tilde{e}_{it} = \alpha\tilde{\eta}_{it} + \tilde{\epsilon}_{it}$, $\tilde{\delta}_{yt} = \alpha\tilde{\delta}_{dt} + \tilde{\xi}_t$, and $\gamma_y = \alpha\gamma_d + \theta$. Now let $\tilde{Y}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{nt})'$ and $\tilde{D}_t = (\tilde{d}_{1t}, \ldots, \tilde{d}_{nt})'$ denote the vectors of outcome and treatment variable within each time period t. From the models $\tilde{d}_{it} = \tilde{\delta}_{dt}' \tilde{f}_i + \tilde{U}_{it}'\gamma_d + \tilde{\eta}_{it}$ and $\tilde{y}_{it} = \tilde{\delta}_{yt}' \tilde{f}_i + \tilde{U}_{it}'\gamma_y + \tilde{e}_{it}$, we regress $\tilde{Y}_t$ and $\tilde{D}_t$ onto the extracted factors $\widehat{F}$ to estimate $\{\tilde{\delta}_{yt}\}_{t=1}^T$ and $\{\tilde{\delta}_{dt}\}_{t=1}^T$:

$$\widehat{\delta}_{yt} = (\widehat{F}'\widehat{F})^{-1}\widehat{F}'\tilde{Y}_t \text{ and } \widehat{\delta}_{dt} = (\widehat{F}'\widehat{F})^{-1}\widehat{F}'\tilde{D}_t. \tag{2.5}$$

The above two regressions make use of (2.4) which implies $\sum_{i=1}^N \widehat{f}_i \widehat{U}_{it} = 0$.

Step 4: Estimate coefficients on $\tilde{U}_{it}$ via lasso. Let

$$\tilde{\gamma}_y = \arg\min_{\gamma \in \mathbb{R}^p} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} (\tilde{y}_{it} - \widehat{\delta}_{yt}' \widehat{f}_i - \widehat{U}_{it}'\gamma)^2 + \kappa_n \| \widehat{\Psi}^y \gamma \|_1 \text{ and} \tag{2.6}$$

$$\tilde{\gamma}_d = \arg\min_{\gamma \in \mathbb{R}^p} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} (\tilde{d}_{it} - \widehat{\delta}_{dt}' \widehat{f}_i - \widehat{U}_{it}'\gamma)^2 + \kappa_n \| \widehat{\Psi}^d \gamma \|_1, \tag{2.7}$$

where the tuning parameter κ_n is chosen as

$$\kappa_n = \frac{2c_0}{\sqrt{nT}} \Phi^{-1}(1 - q_n/(2p)), \quad \log(q_n^{-1}) = O(\log p), \tag{2.8}$$

for some $c_0 > 1$ and $q_n \to 0$,[7] and $\widehat{\Psi}^y$ and $\widehat{\Psi}^d$ are diagonal penalty loading matrices. Given the fixed effects panel structure, we use the clustered penalty loadings

of Belloni et al. (2016) which have diagonal elements defined as

$$[\widehat{\Psi}^{y}]_{j,j} = \sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\widehat{U}_{it,j}\widehat{U}_{it',j}\widehat{e}_{it}\widehat{e}_{it'}} \tag{2.9}$$

$$[\widehat{\Psi}^{d}]_{j,j} = \sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T}\widehat{U}_{it,j}\widehat{U}_{it',j}\widehat{\eta}_{it}\widehat{\eta}_{it'}}, \tag{2.10}$$

where $\widehat{e}_{it}$ is an estimator of $\tilde{e}_{it} = \tilde{y}_{it} - \tilde{\delta}_{yt}'\tilde{f}_i - \tilde{U}_{it}'\gamma_y$ and $\widehat{\eta}_{it}$ is an estimator of $\tilde{\eta}_{it} = \tilde{d}_{it} - \tilde{\delta}_{dt}'\tilde{f}_i - \tilde{U}_{it}'\gamma_d$.[8]

Final Step: Residual regression using post-lasso-selection. We adopt the post-double-selection procedure of Belloni et al. (2014). Let $\widehat{J} = \{j \le p : \tilde{\gamma}_{y,j} \neq 0\} \cup \{j \le p : \tilde{\gamma}_{d,j} \neq 0\}$, and let $\widehat{U}_{it,\widehat{J}}$ be a subvector of $\widehat{U}_{it}$ whose elements are $\{\widehat{U}_{it,j} : j \in \widehat{J}\}$. We then run the regression of $\tilde{y}_{it} - \widehat{\delta}_{yt}'\widehat{f}_i$ on $\widehat{U}_{it,\widehat{J}}$ and $\tilde{d}_{it} - \widehat{\delta}_{dt}'\widehat{f}_i$ on $\widehat{U}_{it,\widehat{J}}$ and obtain

$$\widehat{\gamma}_y = \Big(\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{U}_{it,\widehat{J}}\widehat{U}_{it,\widehat{J}}'\Big)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{U}_{it,\widehat{J}}(\tilde{y}_{it} - \widehat{\delta}_{yt}'\widehat{f}_i), \tag{2.11}$$

$$\widehat{\gamma}_d = \Big(\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{U}_{it,\widehat{J}}\widehat{U}_{it,\widehat{J}}'\Big)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{U}_{it,\widehat{J}}(\tilde{d}_{it} - \widehat{\delta}_{dt}'\widehat{f}_i). \tag{2.12}$$

Note that (2.11) and (2.12) are estimating subvectors, indexed by $\widehat{J}$, of the high-dimensional γ. To make this indexing explicit, one could denote these subvectors as $\widehat{\gamma}_{y,\widehat{J}}$ and $\widehat{\gamma}_{d,\widehat{J}}$, respectively. However, we keep this indexing implicit and denote these vectors as $\widehat{\gamma}_y$ and $\widehat{\gamma}_y$ for notational simplicity.

The final estimator of α is then given by

$$\widehat{\alpha} = \Big(\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{\eta}_{it}^2\Big)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{\eta}_{it}\widehat{e}_{it}, \tag{2.13}$$

where $\widehat{e}_{it} = \tilde{y}_{it} - \widehat{\delta}_{yt}'\widehat{f}_i - \widehat{U}_{it,\widehat{J}}'\widehat{\gamma}_y$ and $\widehat{\eta}_{it} = \tilde{d}_{it} - \widehat{\delta}_{dt}'\widehat{f}_i - \widehat{U}_{it,\widehat{J}}'\widehat{\gamma}_d$ are the residuals from the regressions specified in (2.11) and (2.12).

Note that the estimator $\widehat{\alpha}$ is numerically equivalent to the coefficient on $\tilde{d}_{it}$ in the regression of $\tilde{y}_{it}$ on $\tilde{d}_{it}$, $\widehat{f}_i$ interacted with time dummy variables, and $\widehat{U}_{it,\widehat{J}}$. In Theorem 3.1 of the next section, we verify that inference for $\widehat{\alpha}$ can proceed using the output from this OLS regression as long as clustered standard errors (e.g., Liang and Zeger, 1986; and Arellano, 1987) are used.

The following algorithm summarizes the estimation strategy detailed above.

ALGORITHM (Factor-lasso estimation of α).

(1) Obtain $\{\widehat{f}_i, \widehat{U}_{it}\}_{i\le n, t\le T}$ by extracting factors from the model $\tilde{X}_{it} = \tilde{\Lambda}_t \tilde{f}_i + \tilde{U}_{it}$.
(2) For $\widehat{\delta}_{yt}$ and $\widehat{\delta}_{dt}$ defined in (2.5), run the cluster-lasso programs (2.6) and (2.7) to obtain $\tilde{\gamma}_y$ and $\tilde{\gamma}_d$.
(3) Obtain the estimator $\widehat{\alpha}$ and corresponding estimated standard error as the coefficient on $\tilde{d}_{it} - \widehat{\delta}'_{dt}\widehat{f}_i - \widehat{U}'_{it,\widehat{J}}\widehat{\gamma}_d$ and associated clustered standard error from the regression of $\tilde{y}_{it} - \widehat{\delta}'_{yt}\widehat{f}_i - \widehat{U}'_{it,\widehat{J}}\widehat{\gamma}_y$ on $\tilde{d}_{it} - \widehat{\delta}'_{dt}\widehat{f}_i - \widehat{U}'_{it,\widehat{J}}\widehat{\gamma}_d$ where $\widehat{U}_{it,\widehat{J}}$ is the subvector of $\widehat{U}_{it}$ whose elements are $\{\widehat{U}_{it,j} : j \in \widehat{J}\}$.

3. ASSUMPTIONS AND ASYMPTOTIC THEORY

In this section, we present a set of sufficient conditions under which we establish asymptotic normality of $\widehat{\alpha}$ and provide a consistent estimator of its asymptotic variance. Throughout we consider sequences of data generating processes (DGPs) where p increases as n and T increase and where model parameters are allowed to depend on n and T. We suppress this dependence for notational simplicity. We use the term "absolute constants" to mean given constants that do not depend on the DGP.

3.1. Regularity Conditions

Write $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})'$, $\eta_t = (\eta_{1t}, \ldots, \eta_{nt})'$, and $U_t = (U'_{1t}, \ldots, U'_{nt})'$. Similarly, let $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$, $\eta_i = (\eta_{i1}, \ldots, \eta_{iT})'$, and $U_i = (U'_{i1}, \ldots, U'_{iT})'$.

We assume there are positive absolute constants C_1, C_2, and C_3 such that the following assumption holds.

Assumption 3.1 (DGP)**.** (i) $\{f_i, \eta_i, \epsilon_i, U_i\}_{i\le n}$ are independent and identically distributed across $i = 1, 2, \ldots, n$ and satisfy

$$E(\eta_i|\epsilon_i, U_i, f_i) = 0, \quad E(\epsilon_i|\eta_i, U_i, f_i) = 0, \quad E(U_i|\eta_i, \epsilon_i, f_i) = 0.$$

In addition, given $\{f_i\}_{i\le n}$, the sequence $\{U_i, \eta_i, \epsilon_i\}_{i\le n}$ is also conditionally independent across i.

(ii) Given $\{f_i\}_{i\le n}$, the sequence $\{U_t, \eta_t, \epsilon_t\}_{t\le T}$ is stationary across t, and satisfies a conditional strong-mixing condition. That is, there exists an absolute constant $r > 0$ such that for all $T \in \mathbb{R}^+$,

$$\sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_T} |P(A|\mathcal{F}_F)P(B|\mathcal{F}_F) - P(A\cap B|\mathcal{F}_F)| \le \exp(-C_1 T^r),$$

where $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras generated by $\{(U_t, \eta_t, \epsilon_t) : -\infty \leq t \leq 0\}$ and $\{(U_t, \eta_t, \epsilon_t) : T \leq t \leq \infty\}$, respectively, and $\mathcal{F}_F$ denotes the σ-algebra generated by $\{f_i : i \leq n\}$.

(iii) Almost surely,

$$\max_{i\leq n, m\leq p, t\leq T} \sum_{k=1}^{p}\sum_{s=1}^{T} |E(U_{it,k}\, U_{is,m}|f_i, \epsilon_i, \eta_i)| < C_2.$$

(iv) There is $C_3 > 0$ so that almost surely (in f_i) and for any $s > 0$, $i \leq n$, $j \leq p$ and $k \leq K$,

$$P(|U_{it,j}| > s|f_i) \leq \exp(-C_3 s^2), \quad P(|f_{ik}| > s) \leq \exp(-C_3 s^2),$$
$$P(|\eta_{it}| > s|f_i) \leq \exp(-C_3 s^2), \quad P(|\epsilon_{it}| > s|f_i) \leq \exp(-C_3 s^2).$$

(v) Let θ_m and $\gamma_{d,m}$ be the m^{th} entries of θ and γ_d, and let λ'_{tm} be the m^{th} row of Λ_t. We have

$$|\alpha| + \max_{t\leq T}(\|\xi_t\| + \|\delta_{dt}\|) + \max_{m\leq p}(|\theta_m| + |\gamma_{d,m}|) + \max_{m\leq p, t\leq T}\|\lambda_{tm}\| < C_2.$$

Assumption 3.1 collects reasonably standard regularity conditions that restrict the dependence across observations and tail behavior of random variables. Condition (ii) imposes a conditional strong-mixing condition as in Prakasa Rao (2009) and Su and Chen (2013). Condition (iii) imposes weak conditional dependence in the factor residuals, U_{it}. In the simple case where U_{it} is independent of f_i, η_i, and ϵ_i for all t, this condition reduces to weak intertemporal correlation and no strong dependence among the columns of U_{it}. Importantly, it allows correlation among the observed X_{it} that is not explained by the factors, allowing a rich covariance structure across elements in U_{it}. Condition (iv) is somewhat strong. We use this condition to establish uniform convergence of many sequences, such as $\max_{k\leq p, t\leq K}\|\frac{1}{n}\sum_{i=1}^{n} U_{it,k} f_i\|_2$, using concentration inequalities for sub-Gaussian random variables when p is potentially very large. The need to establish uniform convergence of such sequences is absent in the conventional factor model literature. Finally, condition (v) requires that all low-dimensional parameters are well-bounded.

Let $e_{it} = \alpha\eta_{it} + \epsilon_{it}$.

Assumption 3.2 (Moment bounds)**.** For $m \leq p, i \leq n, t \leq T$, define

$$W_{im} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(U_{it,m} - \bar{U}_{i\cdot,m})(e_{it} - \bar{e}_{i\cdot}).$$

There are absolute constants $c, C > 0$, such that
(i) $\max_{i\leq n, m\leq p} E|W_{im}|^3 \leq C$, $c < \min_{i\leq n, m\leq p} EW_{im}^2 \leq \max_{i\leq n, m\leq p} EW_{im}^2 < C$, and

$$\text{Var}\left(\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}(\eta_{it} - \bar{\eta}_{i\cdot})(\epsilon_{it} - \bar{\epsilon}_{i\cdot})\right) > c.$$

$$\max_{i\le n} E\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right|^{2+\delta} < C, \quad \text{for some } \delta > 0.$$

(ii) Almost surely in $F = (f_1, \ldots, f_n)'$,

$$\max_{m\le p, t\le T}\frac{1}{n}\sum_{i=1}^{n} E(U_{it,m}^8|F) < C, \quad \max_{t\le T}\frac{1}{n}\sum_{i=1}^{n} E(e_{it}^8|F) < C.$$

Assumption 3.2 collects additional high-level moment bounds. The bounds on moments of normalized sums in Condition (i) could be established under a variety of sufficient lower level conditions. Condition (ii) places restrictions on the dependence between $\{U_{it}, e_{it}\}_{i=1,t=1}^{n,T}$ and $\{f_i\}_{i=1}^{n}$.[9]

Before stating the next assumption, we decompose the high dimensional coefficients as

$$\gamma_y = \underbrace{\gamma_y^0}_{\text{exactly sparse}} + \underbrace{R_y}_{\text{remainder}} \quad \text{and} \quad \gamma_d = \underbrace{\gamma_d^0}_{\text{exactly sparse}} + \underbrace{R_d}_{\text{remainder}},$$

where γ_y^0 and γ_d^0 are sparse vectors that approximate the potentially dense true coefficient vectors γ_y and γ_d and R_y and R_d represent approximation errors. Let $J = \{j \le p : \gamma_{y,j}^0 \neq 0\} \cup \{j \le p : \gamma_{d,j}^0 \neq 0\}$ be the union of the support of the exactly sparse components.

Assumption 3.3 (Rate conditions)**.** (i) $\|R_d\|_1 + \|R_y\|_1 = o\left(\sqrt{\frac{\log p}{nT}}\right)$.

(ii) $|J|_0^2 \log^3(p) = O(n)$ and $\log^\gamma(p) = o(n)$ for some $\gamma > 2/r$ where r is defined in Assumption 3.1(ii).

(iii) $|J|_0^2 T = o(n)$. In addition, the number of factors, K, is constant.[10]

Condition (i) requires that a sparse approximation provides a high-enough quality approximation to γ_y and γ_d. This condition is similar to the approximate sparsity condition imposed, for example, in Belloni et al. (2014), though in Belloni et al. (2014) the condition is imposed on errors in approximating a general nonparametric function with a sparse linear model. As we are maintaining a linear factor structure, we impose the restriction directly on the coefficients. Condition (ii) imposes restrictions on both the rate of growth of the dimension p and the decay rate of the strong-mixing coefficient. Condition (iii) imposes that T be much smaller than n. The need for this condition arises from the fact that we need to obtain high-quality estimates of U_{it} in the factor equation, which depends on accurately estimating both the unknown factors and the loadings. Estimating the loading matrix Λ_t well for any given t requires a relatively large n, and we thus require T to be smaller than n as the number of unknown loading matrices $\{\Lambda_t\}_{t\le T}$ is $O(T)$.

Our next assumption restricts the covariance matrix of the within-transformed factor residuals $\tilde{U}_{it}$.

Assumption 3.4 (Sparse eigenvalue)**.** For any $\delta \in \mathbb{R}^p \backslash \{0\}$, write

$$\mathcal{R}(\delta) = \frac{\delta' \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{U}_{it} \tilde{U}_{it}' \delta}{\delta' \delta}.$$

Define the sparse eigenvalue constants:

$$\phi_{\min}(m) = \inf_{\delta \in \mathbb{R}^p : 1 \leq \|\delta\|_0 \leq m} \mathcal{R}(\delta),$$
$$\phi_{\max}(m) = \sup_{\delta \in \mathbb{R}^p : 1 \leq \|\delta\|_0 \leq m} \mathcal{R}(\delta).$$

There is a sequence of absolute constants $l_T \to \infty$ and $c_1, c_2 > 0$ so that with probability approaching one,

$$c_1 < \phi_{\min}(l_T |J|_0) \leq \phi_{\max}(l_T |J|_0) < c_2.$$

Maintaining Assumptions 3.1–3.3, a simple sufficient condition for Assumption 3.4 is that all the eigenvalues of $\frac{1}{nT} \sum_i \sum_t E(U_{it} - \bar{U}_{i,\cdot})(U_{it} - \bar{U}_{i\cdot})'$ are well bounded (see Lemma 4.1 below). Maintaining this condition is standard in high-dimensional approximate factor models (e.g., Bai, 2003; Stock and Watson, 2002). It ensures that the idiosyncratic components are weakly dependent and therefore the decomposition $\tilde{X}_{it} = \tilde{\Lambda}_t \tilde{f}_i + \tilde{U}_{it}$ is asymptotically identified (as $p \to \infty$).

Finally, we require a high-level condition on the accuracy of $\widehat{F}$, given in Assumption D.4 in the Supplemental Appendix. The high-level conditions potentially allow for many estimators of the factors, and we verify that these conditions hold under more primitive assumptions for the case of estimating the factors using PCA in the Supplemental Appendix.

3.2. Main Results

The asymptotic variance of $\widehat{\alpha}$ will depend on the quantities

$$\sigma_{\eta\epsilon} = \mathrm{Var}\left(\frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} (\eta_{it} - \bar{\eta}_{i\cdot})(\epsilon_{it} - \bar{\epsilon}_{i\cdot}) \right) \quad \text{and} \quad \sigma_{\eta}^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathrm{Var}(\eta_{it} - \bar{\eta}_{i\cdot})$$

for which

$$\widehat{\sigma}_{\eta\epsilon} = \frac{1}{nT} \sum_{i=1}^{n} \left(\sum_{t=1}^{T} \widehat{\eta}_{it} \widehat{\epsilon}_{it} \right)^2 \quad \text{and} \quad \widehat{\sigma}_{\eta}^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \widehat{\eta}_{it}^2$$

are natural estimators. Note that $\widehat{\sigma}_{\eta\epsilon}$ is just the usual clustered covariance estimator with clustering at the individual level.

THEOREM 3.1. *Suppose $n, p \to \infty$, and T is either fixed or growing. Under Assumptions 3.1–3.4 and Assumption D.4 in the Supplementary Appendix,*

$$\sqrt{nT}\sigma_{\eta\epsilon}^{-1/2}\sigma_{\eta}^{2}(\widehat{\alpha}-\alpha) \to^{d} \mathcal{N}(0,1).$$

In addition,

$$\sqrt{nT}\widehat{\sigma}_{\eta\epsilon}^{-1/2}\widehat{\sigma}_{\eta}^{2}(\widehat{\alpha}-\alpha) \to^{d} \mathcal{N}(0,1).$$

COROLLARY 3.1. *Let $\mathcal{P}$ be a collection of all DGP's such that the assumptions of Theorem 3.1 hold uniformly over all the DGP's in $\mathcal{P}$. Let $\zeta_\tau = \Phi^{-1}(1-\tau/2)$. Then as $n, p \to \infty$, and T is either fixed or growing with n, uniformly over $P \in \mathcal{P}$,*

$$\lim_{n,p\to\infty} P\left(\alpha \in [\widehat{\alpha} \pm \frac{\zeta_\tau}{\sqrt{nT}}\widehat{\sigma}_{\eta\epsilon}^{1/2}\widehat{\sigma}_{\eta}^{-2}]\right) = 1-\tau.$$

The main implication of Theorem 3.1 and Corollary 3.1 is that $\widehat{\alpha}$ converges at a $\sqrt{nT}$ rate and that inference may proceed using standard asymptotic confidence intervals and hypothesis tests. Importantly, the inferential results hold uniformly across a large class of approximately sparse models which includes cases where perfect selection over which elements of $\tilde{U}_{it}$ enter the model is impossible even in the limit. It is also important to highlight that the conditions on estimation of the factors rule out the presence of weak factors, and the inferential results do not hold uniformly over sequences of models in which perfect selection of the number of factors and fast convergence of the factors and factor loadings do not hold. The difficulty with handling weak factors arises due to the entry of the estimation errors of the factors in the cluster-lasso problems (2.6) and (2.7) and the nonsmooth and highly nonlinear nature of this problem. Extending the results to accommodate the presence of weak factors and imperfect selection of the number of factors would be an interesting direction for further research.

4. K-STEP BOOTSTRAP

We now present a computationally tractable bootstrap procedure that can be used in lieu of the plug-in asymptotic inference formally presented in Theorem 3.1 and Corollary 3.1. In the following, we introduce a bootstrap procedure which only approximately solves the cluster-lasso problem within each bootstrap replication and thus may remain computationally convenient while also intuitively capturing the sampling variation introduced in the lasso selection.

4.1. The *k*-step Bootstrap

Let $D^* = \{\tilde{y}_{it}^*, \tilde{d}_{it}^*, \tilde{X}_{it}^*\}_{i\le n, t\le T}$ denote a sample of bootstrap data obtained through application of Algorithm (***k*-Step Wild Bootstrap**) below, and let $\widehat{\alpha}^*$ be the estimator obtained by applying the factor-lasso estimator with data D^*. Let B denote the number of bootstrap repetitions.

Most algorithms that solve the lasso problem rely on iterations. A potential computational problem with bootstrap procedures for lasso estimation is that one needs to solve B lasso problems where B will typically be fairly large. To circumvent this problem, we adopt the approach of Andrews (2002) by using the fact that the complete lasso estimator based on the original data, denoted by $\tilde{\gamma}_{lasso}$, should be close to the complete lasso estimator based on bootstrapped data D^*, denoted by $\tilde{\gamma}^*_{lasso}$. Hence, within each bootstrap replication, we can use $\tilde{\gamma}_{lasso}$ as the initial value for solving the lasso problem and iteratively update the lasso algorithm for k steps rather than computing the full solution on the bootstrap data, $\tilde{\gamma}^*_{lasso}$. Denote the resulting k-step bootstrap lasso estimator by $\tilde{\gamma}^*$. We simply use $\tilde{\gamma}^*$ in place of $\tilde{\gamma}^*_{lasso}$ wherever the solution to a lasso problem shows up in the factor-lasso problem. The main result of this section is showing that the k-step bootstrap procedure is first-order valid for statistical inference about α as long as the minimization error after k steps is less than the statistical error (i.e., $o_{P^*}((nT)^{-1/2})$).

The substantive difference between the present context and Andrews (2002) is that Andrews (2002) makes use of Newton–Raphson updates for the k-steps while we face a regularized optimization problem at each iteration. Tractability relies on the fact that there are a variety of procedures for updating within the lasso problem that are available in closed form. Using these analytic updates greatly reduces the overall computational task and makes a k-step bootstrap procedure attractive within the lasso context.

Specifically, consider the following lasso problems on the bootstrap data. Let

$$\begin{aligned}\tilde{\gamma}^*_{y,lasso} &= \arg\min_{\gamma\in\mathbb{R}^p} \mathcal{L}^*_y(\gamma) + \kappa_n \|\widehat{\Psi}^y \gamma\|_1,\\ \tilde{\gamma}^*_{d,lasso} &= \arg\min_{\gamma\in\mathbb{R}^p} \mathcal{L}^*_d(\gamma) + \kappa_n \|\widehat{\Psi}^d \gamma\|_1,\end{aligned} \tag{4.1}$$

where

$$\mathcal{L}^*_y(\gamma) = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(\tilde{y}^*_{it} - \widehat{\delta}^{*\prime}_{yt}\widehat{f}^*_i - \widehat{U}^{*\prime}_{it}\gamma)^2 \text{ and } \mathcal{L}^*_d(\gamma) = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(\tilde{d}^*_{it} - \widehat{\delta}^{*\prime}_{dt}\widehat{f}^*_i - \widehat{U}^{*\prime}_{it}\gamma)^2.$$

The definitions of $\{\tilde{y}^*_{it}, \tilde{d}^*_{it}, \widehat{\delta}^*_{yt}, \widehat{\delta}^*_{dt}, \widehat{f}^*_i, \widehat{U}^*_{it}\}_{i\le n, t\le T}$ will be formally given below. Let $\tilde{\gamma}_y$ and $\tilde{\gamma}_d$ be the lasso solutions obtained from the original data. Also, note that we fix the value of κ_n and of the penalty loadings $\widehat{\Psi}^y$ and $\widehat{\Psi}^d$ to the same values as used to obtain the solutions $\tilde{\gamma}_y$ and $\tilde{\gamma}_d$ in the original data.

Within each bootstrap replication, we then approximately solve the lasso problems (4.1) by applying the following procedure. We note that the maximum number of steps k to be taken within each bootstrap replication should be determined on a case-by-case basis according to the available computational capacity.[11]

ALGORITHM (k-Step lasso iteration).

Set k to be a predetermined number of iterations.

(A1) Set $l = 0$ and initialize at $\gamma_{y,0} = \tilde{\gamma}_y$, $\gamma_{d,0} = \tilde{\gamma}_d$.

(A2) Determine one-step iteration mappings $\mathcal{S}_y, \mathcal{S}_d : \mathbb{R}^p \to \mathbb{R}^p$. Let

$$\gamma_{y,l+1} = \mathcal{S}_y(\gamma_{y,l}), \quad \gamma_{d,l+1} = \mathcal{S}_d(\gamma_{d,l}). \tag{4.2}$$

Set $l = l + 1$.

(A3) Repeat (A2) until $l = k$. Let the k-step lasso estimators be

$$\tilde{\gamma}_y^* = \gamma_{y,k}, \quad \tilde{\gamma}_d^* = \gamma_{d,k}.$$

There are a variety of iteration mappings that can be used in Step (A2) of the k-step lasso problem. A commonly used and simple mapping is the "coordinate descent method," also known as the "shooting method," studied by Fu (1998).[12] For solving problem (4.1), write the solution after the l^{th} iteration as $\gamma_{y,l} = (\gamma_{y,l,1}, \ldots, \gamma_{y,l,p})'$. The coordinate descent method updates $\gamma_{y,l+1}$ by iteratively cycling through all coordinates. Specifically, we solve the following one-dimensional optimization problem for $m = 1, \ldots, p$,

$$\gamma_{y,l+1,m} = \arg\min_{g \in \mathbb{R}} \frac{1}{nT} \sum_{i,t} (\tilde{y}_{it}^* - \widehat{\delta}_{yt}^{*\prime} \widehat{f}_i^* - \widehat{U}_{it,m^-}^{*\prime} \gamma_{y,l+1,m^-} - \widehat{U}_{it,m^+}^{*\prime} \gamma_{y,l,m^+} - \widehat{U}_{it,m}^* g)^2 + \kappa_n |\widehat{\Psi}_m^y g|. \tag{4.3}$$

Here $m^- = \{j : j < m\}$; and $\gamma_{y,l+1,m^-}$ and $\widehat{U}_{it,m^-}^*$ are $\mathbb{R}^{m-1}$ dimensional vectors whose components are, respectively, those of $\{\gamma_{y,l+1,j} : j < m\}$ and $\{\widehat{U}_{it,j}^* : j < m\}$. Similarly, $m^+ = \{j : j > m\}$; and γ_{y,l,m^+} and $\widehat{U}_{it,m^+}^*$ are $\mathbb{R}^{p-m}$ dimensional vectors whose components are, respectively, those of $\{\gamma_{y,l,j} : j > m\}$ and $\{\widehat{U}_{it,j}^* : j > m\}$. When $m = 1$, m^- is empty; and when $m = p$, m^+ is empty. In these cases, the corresponding subvectors, $\gamma_{y,l+1,m^-}$ and $\widehat{U}_{it,m^-}^*$ or γ_{y,l,m^+}, and $\widehat{U}_{it,m^+}^*$, are defined as zero. Note that when $\gamma_{y,l+1,m}$ is being updated the previous $m-1$ elements have already been updated, while the remaining $p-m$ elements are yet to be updated. Thus, $\gamma_{y,l+1,m^-}$ is a subvector of $\gamma_{y,l+1}$, but γ_{y,l,m^+} is still a subvector of $\gamma_{y,l}$ in the l^{th} update. Denote by $\gamma_{y,l+1}^{(m)} := (\gamma_{y,l+1,m^-}, \gamma_{y,l+1,m}, \gamma_{y,l,m^+})'$ the vector that results immediately after the m^{th} coordinate has been updated during the $(l+1)^{\text{th}}$ iteration. When $m = p$, all the components have been updated; and we obtain $\gamma_{y,l+1} := \gamma_{y,l+1}^{(p)}$.

Importantly, (4.3) is a one-dimensional ℓ_1-penalized quadratic problem which has an analytical solution given by the soft thresholding operation:

$$\gamma_{y,l+1,m} = \left[sgn\left(\frac{1}{nT} \sum_{i=1}^{T} \sum_{t=1}^{T} Z^*_{it,l,m} \widehat{U}^*_{it,m} \right) \right] \times \left(\left| \frac{1}{nT} \sum_{i=1}^{T} \sum_{t=1}^{T} Z^*_{it,l,m} \widehat{U}^*_{it,m} \right| - \frac{1}{2} \kappa_n \widehat{\Psi}^y_m \right)_+ \left(\frac{1}{nT} \sum_{i=1}^{T} \sum_{t=1}^{T} \widehat{U}^{*2}_{it,m} \right)^{-1}, \tag{4.4}$$

where $Z^*_{it,l,m} := \tilde{y}^*_{it} - \widehat{\delta}^{*'}_{yt} \widehat{f}^*_i - \widehat{U}^{*'}_{it,m^-} \gamma_{y,l+1,m^-} - \widehat{U}^{*'}_{it,m^+} \gamma_{y,l,m^+}$, $(x)_+ = \max\{x, 0\}$, and $sgn(x)$ takes the sign of x. Therefore, the mappings in (4.2) are given by

$$\mathcal{S}_y(\gamma_{y,l}) = (\gamma_{y,l+1,1}, \ldots, \gamma_{y,l+1,p})', \quad \text{where each } \gamma_{y,l+1,m} \text{ is given in (4.4).}$$

$\mathcal{S}_d(\gamma_{d,l})$ is obviously defined similarly.

With the k-step lasso program defined, we now state the complete algorithm for the proposed k-step bootstrap procedure. We make use of a wild residual bootstrap to generate the data at each bootstrap replication.

ALGORITHM (k-Step wild bootstrap).

Let $\{\widehat{f}_i, \widehat{U}_{it}, \widehat{\Lambda}_t\}_{i \le n, t \le T}$ denote the estimates of the features of the factor model using the original data. Let $\widehat{\alpha}, \widehat{\delta}_{dt}, \widehat{\delta}_{yt}, \widehat{\gamma}_d, \widehat{\gamma}_y$ be the estimated coefficients from the original data, defined in (2.4) through (2.13). Also, let

$$\widehat{\xi}_t = \widehat{\delta}_{yt} - \widehat{\alpha}\widehat{\delta}_{dt}, \quad t = 1, \ldots, T, \text{ and}$$
$$\widehat{\theta} = \widehat{\gamma}_y - \widehat{\alpha}\widehat{\gamma}_d.$$

(1) For each $i = 1, \ldots, n$, let w^x_i $(x = U, Y, D)$ be mutually independent random variables, where $\{w^x_i\}_{i \le n}$ are i.i.d. with mean zero and variance one. Let

$$\tilde{U}^*_{it} = w^U_i \widehat{U}_{it}, \quad \tilde{\eta}^*_{it} = w^D_i \widehat{\eta}_{it}, \quad \tilde{\epsilon}^*_{it} = w^Y_i \widehat{\epsilon}_{it}, \quad t = 1, \ldots, T.$$

Define $\{\tilde{y}^*_{it}, \tilde{d}^*_{it}, \tilde{X}^*_{it}\}_{t \le T}$ as

$$\tilde{y}^*_{it} = \widehat{\alpha}\tilde{d}^*_{it} + \widehat{\xi}'_t \widehat{f}_i + \tilde{U}^{*'}_{it}\widehat{\theta} + \tilde{\epsilon}^*_{it},$$
$$\tilde{d}^*_{it} = \widehat{\delta}'_{dt} \widehat{f}_i + \tilde{U}^{*'}_{it}\widehat{\gamma}_d + \tilde{\eta}^*_{it},$$
$$\tilde{X}^*_{it} = \widehat{\Lambda}_t \widehat{f}_i + \tilde{U}^*_{it}.$$

(2) Apply the Factor-Lasso Algorithm to the bootstrap data $\{\tilde{y}^*_{it}, \tilde{d}^*_{it}, \tilde{X}^*_{it}\}_{i \le n, t \le T}$ to obtain an estimated alpha $\widehat{\alpha}^*$ replacing the lasso estimation in Step (2) of the Factor-Lasso Algorithm with steps (A1)–(A3) from the k-Step Lasso Iteration defined above.

(3) Repeat the above steps (1)–(2) B times to obtain $\{\widehat{\alpha}_b^*\}_{b\le B}$.

Let q_τ^* be the τ^{th} upper quantile of $\{\sqrt{nT}|\widehat{\alpha}_b^* - \widehat{\alpha}|\}_{b\le B}$ so that

$$P^*(\sqrt{nT}|\widehat{\alpha}_b^* - \widehat{\alpha}| \le q_\tau^*) = 1 - \tau. \tag{4.5}$$

Construct the bootstrap confidence interval:

$$\left[\widehat{\alpha} \pm \frac{q_\tau^*}{\sqrt{nT}}\right].$$

In (4.5), P^* denotes the bootstrap probability measure induced by the bootstrap resampling. More specifically, it is induced by the conditional distribution of the bootstrap weights $\{\{w_i^x\}_{i\le n}^b\}_{b\le B}$, $x = U, Y, D$, given the original sample.

4.2. Validity of the *k*-step Bootstrap Confidence Interval

We now present conditions under which we verify that the bootstrap confidence intervals are asymptotically valid:

$$P\left(\alpha \in \left[\widehat{\alpha} \pm \frac{q_\tau^*}{\sqrt{nT}}\right]\right) \to 1 - \tau.$$

The first assumption imposes high-level conditions that will admit the use of general updating rules in (4.2) of the k-Step Lasso Iteration. Recall that $\tilde{\gamma}_d^* = \gamma_{d,k}$ is the k-step bootstrap solution, and $\tilde{\gamma}_{x,lasso}$ is the complete lasso solution using the bootstrap data given by (4.1), $x \in \{y, d\}$.

Assumption 4.1. The following conditions hold for $x \in \{y, d\}$:
(i) Minimization Error: There is a deterministic sequence a_n such that $a_n\sqrt{nT} = o(1)$, and $K_0 > 0$, such that when $k > K_0$,

$$\mathcal{L}_x^*(\tilde{\gamma}_x^*) + \kappa_n\|\widehat{\Psi}^x\tilde{\gamma}_x^*\|_1 \le \mathcal{L}_x^*(\tilde{\gamma}_{x,lasso}^*) + \kappa_n\|\widehat{\Psi}^x\tilde{\gamma}_{x,lasso}^*\|_1 + O_{P^*}(a_n).$$

(ii) Sparsity: $|\widehat{J}^*| = O_{P^*}(|J|_0)$, where $\widehat{J}^* = \{j \le p : \tilde{\gamma}_{dj}^* \ne 0\} \cup \{j \le p : \tilde{\gamma}_{yj}^* \ne 0\}$.

Condition (i) requires that the minimization error should be negligible compared to the statistical error after k iteration steps. Condition (ii) guarantees the sparsity of the iterated solutions. As a concrete example, we verify both conditions for the coordinate descent method. We note that, to the best of our knowledge, showing the $|J|_0$-sparsity of the k-step iterated coordinate descent estimator has not been done previously when p is potentially much larger than n and may be of some independent interest.

PROPOSITION 4.1. *The coordinate descent iteration as given in (4.4) satisfies Assumption 4.1.*

Remark 4.1. We would like to point out that the original lasso estimators also satisfy Assumption 4.1(i). In fact, the k-step bootstrap lasso is asymptotically valid for any $k \geq 0$, where $k = 0$ corresponds to the case that we do not repeat the lasso step in the bootstrap procedure, treating the index set $\widehat{J}$ as fixed. While asymptotically, the k-step bootstrap is equivalent for any $k \geq 0$, the major difference of using $k \geq 1$ lies in the finite-sample performance. With a larger k, the bootstrap may also mimic some of the sampling error from the lasso-estimation step. Though this estimation error is asymptotically negligible, it does affect the ultimate estimator's finite-sample behavior. Therefore, we recommend the use of k-step bootstrap with k chosen as large as possible within one's computational budget.

We next impose a fairly standard notion of regularity on the high-dimensional component $\tilde{U}_{it}$.

Assumption 4.2 (Restricted strong convexity)**.** There is a constant $c > 0$, and a sequence $\tau_n = o(|J|_0^{-1})$ so that for all $\delta \in \mathbb{R}^p$,

$$\delta' \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{U}_{it} \tilde{U}_{it}' \delta \geq \frac{c}{2} \|\delta\|_2^2 - O_P(\tau_n) \|\delta\|_1^2.$$

This assumption has been discussed by many authors, and various sufficient conditions have been provided (e.g., Raskutti, Wainwright, and Yu, 2010; and Loh and Wainwright, 2015). The following lemma provides a simple sufficient condition for both Assumption 4.2 and the sparse eigenvalue assumption (Assumption 3.4).

LEMMA 4.1. *Suppose Assumption 3.1 holds. Let* $\lambda_1 \leq \cdots \leq \lambda_p$ *be the eigenvalues of* $\frac{1}{nT} \sum_i \sum_t E\left[(U_{it} - \bar{U}_{i,\cdot})(U_{it} - \bar{U}_{i\cdot})'\right]$. *Suppose for some* $0 < c < C$,

$$c < \lambda_1 \leq \lambda_p < C.$$

Then Assumptions 3.4 and 4.2 are satisfied.

The following conditions are imposed on the bootstrap weights.

Assumption 4.3. For $x = U, Y, D$, $E[w_i^x] = 0$ and $\mathrm{Var}(w_i^x) = 1$. In addition, there exist $L, r > 0$, such that for any $s > 0$, $i \leq n$,

$$P(|w_i^x| > s) \leq \exp(-Ls^r).$$

The subexponential condition for the bootstrap weights enables us to bound many stochastic processes uniformly in $m \leq p$ and $t \leq T$. In our numerical

studies, we follow Mammen (1993) and use $w_i^x = \zeta_{1,i}^x/\sqrt{2} + ((\zeta_{2,i}^x)^2 - 1)/2$ for $x \in \{U, Y, D\}$ where $\zeta_{1,i}^x$ and $\zeta_{2,i}^x$ are independent standard normals.

Finally, we impose a high-level assumption on the quality of estimation of the factors in the bootstrap data, given in Assumption D.5 in the Supplementary Appendix.

Under these additional conditions, we are able to verify that the confidence interval resulting from application of the k-step bootstrap procedure has asymptotically correct coverage.

THEOREM 4.1. *Suppose $n, p \to \infty$, and T is either fixed or growing. Under Assumptions 3.1–3.4, 4.1–4.3, and Assumptions D.4 and D.5 in the Supplementary Appendix,*

$$\sqrt{nT}\sigma_{\eta\epsilon}^{-1/2}\sigma_\eta^2(\widehat{\alpha}^* - \widehat{\alpha}) \to^{d^*} \mathcal{N}(0,1),$$

which means for any $x \in \mathbb{R}$, $\left|P^*(\sqrt{nT}\sigma_{\eta\epsilon}^{-1/2}\sigma_\eta^2(\widehat{\alpha}^* - \widehat{\alpha}) < x) - \Phi(x)\right| \to^P 0$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. In addition,

$$P(\sqrt{nT}|\widehat{\alpha} - \alpha| \le q_\tau^*) \to 1 - \tau.$$

5. NUMERICAL STUDIES AND EXAMPLES

We now present simulation and empirical results in support of the formal analysis presented in the previous sections.

5.1. Simulation Example

We report results for estimation and inference on α with data generated according to

$$\begin{aligned} y_{it} &= \alpha d_{it} + (c_\xi \xi_t)' f_i + U_{it}'(c_\theta \theta) + g_i + v_t + \epsilon_{it}, \\ d_{it} &= (c_\delta \delta_{dt})' f_i + U_{it}'(c_\gamma \gamma_d) + \zeta_i + \mu_t + \eta_{it}, \\ X_{it} &= (c_\Lambda \Lambda_t) f_i + w_i + \rho_t + U_{it}, \end{aligned}$$

with $n = 100$, $T = 10$, $K = 3$, and $p = 100$.[13] We take $\epsilon_{it} \sim N(0,1)$, $\eta_{it} \sim N(0,1)$, and $U_{it} \sim N(0_p, \Sigma_U)$ where 0_p is a $p \times 1$ vector of zeros, Σ_U has (r,s) element given by $[\Sigma_U]_{[r,s]} = .7^{|r-s|}$, and ϵ_{it}, η_{it}, and U_{it} are i.i.d. over i and t and jointly independent of each other. We generate unobserved individual-specific and time-specific heterogeneity by taking n i.i.d. draws, one for each individual, $(g_i, \zeta_i, w_i) \sim N(0_{p+2}, I_{p+2})$ where I_{p+2} is a $(p+2) \times (p+2)$ identity matrix and taking T i.i.d. draws, one for each time period, $(v_t, \mu_t, \rho_t) \sim N(0_{p+2}, I_{p+2})$. The latent factors, f_i, are generated as i.i.d. draws from $N(0_K, I_K)$. The factor loading vectors ξ_t and δ_{dt} and factor loading matrix Λ_t are drawn independently over time with each entry generated as an independent draw from a standard normal

random variable. The individual-specific and time-specific heterogeneity terms and the factor loadings are drawn once, and the same values are used in each simulation replication.

We set the j^{th} entry of θ and γ_d as $\theta_j = \gamma_{d,j} = \frac{1}{j^2}$. c_Λ, c_δ, c_γ, c_ξ, and c_θ are scalars that are set to alter the relative strength of f_i and U_{it} in each equation. We choose c_Λ so that the average R^2 from the p regressions of $X_{it,j}$ on f_i is 0.5. We choose (c_δ, c_γ) so that the R^2 of the infeasible regression of $d_{it} - \zeta_i - \mu_t$ on $(c_\delta \delta_{dt})' f_i + U_{it}'(c_\gamma \gamma_d)$ is 0.7 and the factors account for 0%, 25%, 50%, 75%, or 100% of the explanatory power in this regression. We similarly choose (c_ξ, c_θ) so that the R^2 of the infeasible regression of $y_{it} - \alpha d_{it} - g_i - \nu_t$ on $(c_\xi \xi_t)' f_i + U_{it}'(c_\theta \theta)$ is 0.7 and the factors account for 0%, 25%, 50%, 75%, or 100% of the explanatory power in this regression. Finally, we set $\alpha = 1$.

We compare the performance of the procedure developed in this article to several benchmarks. Because we consider a design with $p < nT$, ordinary least squares of y_{it} on d_{it}, X_{it} and a full set of individual and time dummy variables is feasible (OLS). We also consider estimating α based on the assumption that confounding is entirely captured by latent factors. To implement this procedure, we extract factors, $\widehat{f}_i$, from $\tilde{X}_{it}$ by PCA as discussed in Section D in the supplement. We then regress y_{it} on d_{it}, $\widehat{f}_i$ interacted with a complete set of time dummy variables, and a full set of individual and time dummy variables to obtain the estimator for α (Factor). For our third procedure, we directly apply the fixed effects double-selection procedure of Belloni et al. (2016) which is appropriate for a sparse high-dimensional model with fixed effects (Double Selection). We then consider two *ad hoc* variants of the double-selection approach. In the first, we extract the first 20 principal components and interact these with a full set of time dummies. We then apply the fixed effects double-selection procedure of Belloni et al. (2016) to the data (Y, D, X^*) where X^* denotes the original X variables augmented to include the interactions of principal components with time dummies (Double Selection F). The second *ad hoc* procedure extracts factors from $\tilde{X}_{it}$ by PCA. We then obtain estimates $\widehat{U}_{it}$ as in (2.4) and apply the fixed effects double-selection procedure of Belloni et al. (2016) to the data $(Y, D, \hat{U}^*)$ where $\hat{U}^*$ denotes the matrix formed by combining $\widehat{U}$ with the interactions of principal components with time dummies (Double Selection U). Finally, we directly apply the factor-lasso approach outlined in this article (Factor Lasso). We use the Ahn and Horenstein (2013) procedure to select the number of factors to use in obtaining the Factor, Double Selection U, and Factor Lasso results.

Figure 1 gives simulation RMSEs for the estimator of α resulting from applying each procedure. The RMSEs are truncated at 0.1 for readability of the figure. The most striking feature of Figure 1 is that only the proposed factor lasso procedure delivers uniformly good performance regardless of the relative strength of the factors and factor residuals in this simulation design. Each of the other procedures exhibits behavior that depends strongly on the exact strength of the factors in the different equations. In terms of RMSE, the factor-lasso procedure

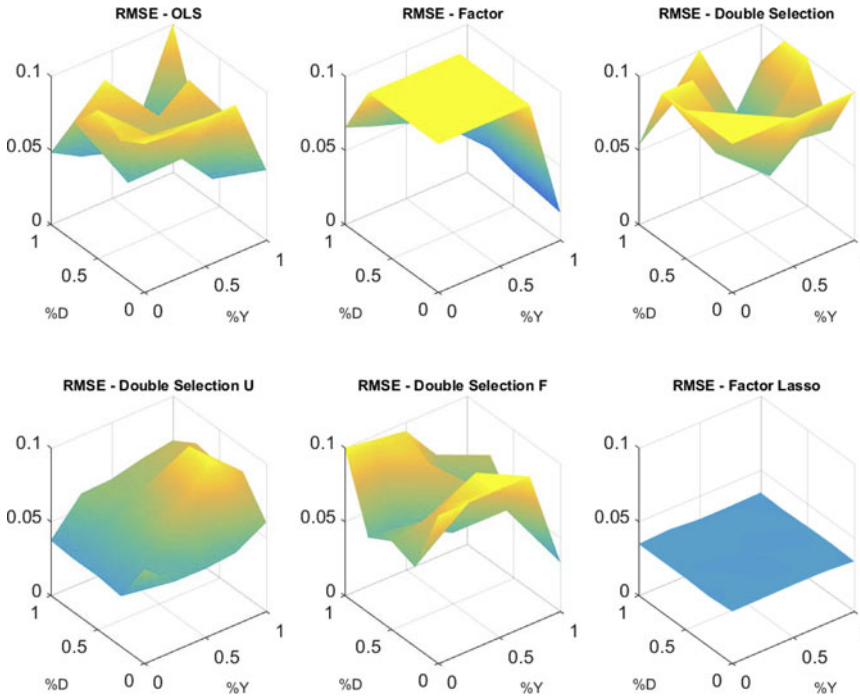

FIGURE 1. This figure shows the simulation RMSE of each of the estimators described in the text for estimating the coefficient of interest in a panel partial factor model. RMSE (truncated at 0.1) is shown in the vertical axis. The horizontal axes give the fraction of the explanatory power in an infeasible regression of Y on factors and factor residuals, "%Y," and the fraction of the explanatory power in an infeasible regression of D on factors and factor residuals, "%D," where the infeasible regressions are described in the text.

uniformly dominates regular OLS, Double Selection ignoring the factor structure, and the *ad hoc* procedure Double Selection F within the design considered. The factor-lasso estimator of α is outperformed by the pure factor model in the case where all of the explanatory power in the outcome equation is contained in the factors, which corresponds to the case where the pure factor model is correctly specified and there is no additional confounding based on the factor residuals, and the Double Selection U procedure when the factors have no explanatory power in the treatment (D) equation but all explanatory power in the Y equation. It is also important to note that the performance loss is small in these few cases where the factor lasso is outperformed. A final interesting point to note is that the conventional lasso-based double selection procedure is outperformed by the factor lasso even when the factors do not load in either the treatment or outcome equation.

We report size of 5% level tests based on standard asymptotic approximations for each of the six procedures considered in Figure 2 where the sizes are truncated

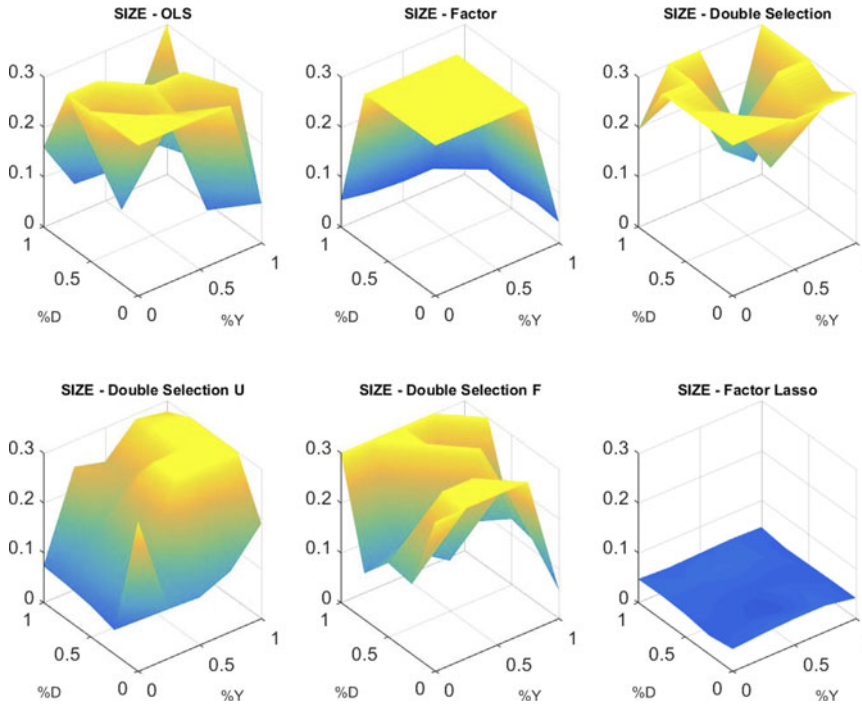

FIGURE 2. This figure shows the simulation size of 5% level tests based on each of the estimators described in the text for the PPFM. Size (truncated at 0.3) is shown in the vertical axis. The horizontal axes give the fraction of the explanatory power in an infeasible regression of Y on factors and factor residuals, "%Y," and the fraction of the explanatory power in an infeasible regression of D on factors and factor residuals, "%D," where the infeasible regressions are described in the text.

at 0.3 for readability of the figure. In each panel, we report the rejection frequency of the standard t-test of the null hypothesis that $\alpha = 1$ with standard errors clustered at the individual level. The most striking feature of the figure is again the uniformly good performance of tests based on the proposed factor lasso procedure. Tests based on the factor-lasso procedure effectively control size, with size ranging between 3.3% and 5.3% across the design parameters considered in the simulation. This behavior is in sharp contrast to the other procedures considered which may have large size distortions depending upon exactly how large the relative contribution of the factors is in the D and Y equations. Importantly, this good behavior does not come at the cost of using an inferior estimator as evidenced by the RMSE results.

We conclude this discussion by looking at the performance of the k-step bootstrap. In Figure 3, we report size of 5% level tests using the factor-lasso estimator and the asymptotic approximation provided in Theorem 3.1, the k-step bootstrap, and a score bootstrap based on Belloni et al. (2017). The k-step bootstrap and

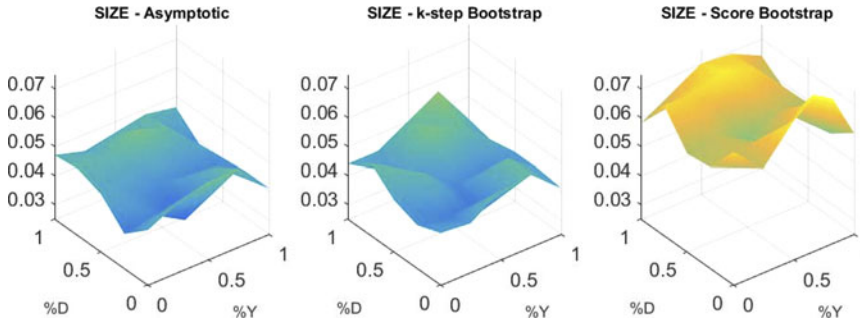

FIGURE 3. This figure shows the simulation size of 5% level tests based on the factor-lasso estimator in the PPFM and the asymptotic Gaussian approximation, the *k*-step bootstrap, and a score based bootstrap. Size is shown in the vertical axis. The horizontal axes give the fraction of the explanatory power in an infeasible regression of Y on factors and factor residuals, "%Y," and the fraction of the explanatory power in an infeasible regression of D on factors and factor residuals, "%D," where the infeasible regressions are described in the text.

asymptotic approximation have similar performance that keeps size close to the promised level. Interestingly, the score-based bootstrap that does not reestimate the factors or the lasso parts of the model exhibits mild size distortions across all of the design settings in this example.

5.2. Estimating the Effects of Gun Prevalence on Crime

In this example, we follow Belloni et al. (2016) who build upon the work of Cook and Ludwig (2006) and attempt to estimate the effect of gun prevalence on crime in a setting with a high-dimensional set of potential controls. As in Belloni et al. (2016), we focus exclusively on trying to measure the effect of gun prevalence on homicide rates. An important difficulty with estimating the effect of gun prevalence in the United States is that exact gun-ownership numbers are difficult to obtain. Due to this difficulty, Cook and Ludwig (2006) use the fraction of suicides committed with a firearm (abbreviated FSS) within a county to proxy for county-level gun ownership rates.

Both Cook and Ludwig (2006) and Belloni et al. (2016) estimate linear fixed effects models of the form

$$\log Y_{it} = \alpha \log \mathrm{FSS}_{it-1} + X_{it}'\beta + g_i + \nu_t + \epsilon_{it}, \tag{5.1}$$

where g_i and ν_t are treated as parameters to be estimated, X_{it} are control variables, and Y_{it} is one of three dependent variables: the overall homicide rate within county i in year t, the firearm homicide rate within county i in year t, or the non-firearm homicide rate within county i in year t. Cook and Ludwig (2006) use the four variables percent African American, percent of households with female head, nonviolent crime rates, and percent of the population that lived in the same

house five years earlier as their set of controls X_{it}. Belloni et al. (2016) maintain the assumption of approximate sparsity and employ their variable selection approach using a much larger set of potential controls generated by taking variables compiled by the US Census Bureau as X_{it}. Their variables include county-level measures of demographics, the age distribution, the income distribution, crime rates, federal spending, home ownership rates, house prices, educational attainment, voting patterns, employment statistics, and migration rates along with interactions of the initial (1980) values of all control variables with a linear, quadratic, and cubic term in time.

We employ the PPFM, (1.1)–(1.3), and factor-lasso approach to estimate α using 909 variables in X_{it} constructed as in Belloni et al. (2016).[14] The PPFM model seems very appropriate for this data as it directly incorporates a mechanism to accommodate the concern that there are features of counties that are not directly observed, the f_i, but are related to the evolution of the outcome and treatment variable of interest, which is captured by the time-varying factor loadings. Obviously, exclusion of these factors would then lead to omitted variables bias in any estimator of α that fails to capture them.

The key assumption that we leverage to allow us to simply accommodate these latent factors is that the same correlated unobserved factors that lead to confounding are related to the evolution of other observed county-level aggregates and that we have access to a large number of these auxiliary aggregates. While this key assumption is strong, the PPFM also naturally provides some robustness to the presence of shocks (U_{it}) that are related to movements of the observed X_{it} series as well as movements in the variable of interest and outcome. Such shocks may be motivated, for example, by the factor structure being misspecified, by the presence of variables that are not strongly related to factors but are confounded with the treatment and outcome, and simply by the presence of local shocks not captured by the factors that are related to the observed series.

We present estimation results in Table 1 with results for each dependent variable presented across the columns and rows corresponding to different estimation approaches. As a baseline, we report numbers taken directly from the first row of Table 3 in Cook and Ludwig (2006) in the first row of Table 1 ("Cook and Ludwig (2006) Baseline"). We report results obtained from our data in the remaining rows.[15] For these results, we first report the point estimate and estimate of the asymptotic standard error obtained by clustering by county. Immediately below these results, we report the 95% confidence interval obtained from applying the k-step bootstrap procedure in brackets. The rows labeled "Post Double Selection" apply the procedure of Belloni et al. (2016). The rows labeled "Factor" are based on a pure factor model; the rows labeled "Factor-Lasso" use the proposed factor-lasso procedure. All factors are estimated using PCA and the number of factors is selected using Ahn and Horenstein (2013).

We see that the estimates and inferential statements produced for the firearm homicide rate ("Gun") and the nonfirearm homicide rate ("non-Gun") are broadly consistent with each other. In all cases, there is a fairly large positive point

TABLE 1. Estimates of the effect of gun prevalence on homicide rates

	Overall	Gun	Non-Gun
Cook and Ludwig (2006) baseline	0.086 (0.038)	0.173 (0.049)	−0.033 (0.040)
post double selection	0.062 (0.042)	0.138 (0.059)	−0.055 (0.042)
	[−0.019,0.143]	[0.036,0.240]	[−0.139,0.029]
Factor	0.104 (0.043)	0.210 (0.064)	−0.022 (0.040)
	[0.019,0.189]	[0.097,0.323]	[−0.099,0.055]
Factor-Lasso	0.069 (0.036)	0.167 (0.046)	−0.048 (0.040)
	[0.000,0.138]	[0.078,0.256]	[−0.128,0.032]

This table presents estimates of the effect of gun ownership on homicide rates for a panel of 195 US Counties over the years 1980–1999. The columns "Overall", "Gun", and "non-Gun", respectively, report the estimated effect of gun prevalence on the log of the overall homicide rate, the log of the gun homicide rate, and the log of the nongun homicide rate. Each row corresponds to a different specification as described in the text. In each specification, the outcome corresponding to the column label is regressed on lagged log(FSS) (a proxy for gun ownership) and additional covariates as described in the text. Each specification includes a full set of year and county fixed effects. Standard errors clustered by county are provided in parentheses. *k*-step bootstrap 95% confidence intervals are given in brackets.

estimate for the effect on the firearm homicide rate with corresponding 95% confidence intervals that exclude zero, suggesting positive association between the used measure of gun prevalence and gun homicides. For the nonfirearm homicide rate, all point estimates are negative and confidence intervals include both positive and negative values. The broad results for the overall homicide rate ("Overall") are slightly more mixed. The baseline results for Cook and Ludwig (2006) and results from a pure factor model suggest a strongly significant, positive effect of gun prevalence on the overall homicide rate. Assuming sparsity and applying Belloni et al. (2016) yields a positive estimate of the effect which is statistically insignificant at the 5% level. Finally, the factor-lasso estimator is similar in magnitude to the sparsity-based estimator but borderline significant at the 5% level using the bootstrap confidence interval.

A more interesting comparison can be made by looking more closely and considering the variable and factor selection results. The "Post Double Selection" procedure ends up selecting three variables for estimating the effect on overall homicide rates, three variables for gun homicide rates, and two variables for nongun homicide rates. The pure factor model uses one factor. The factor-lasso approach then uses one factor in all cases but selects eight additional variables for estimating the effect on the overall homicide rate, eight additional variables for the gun homicide rate, and five additional variables for the nongun homicide rate. These results suggest that the "Post Double Selection" and "Factor" results may be based on models that fail to adequately capture the effect of potential confounds. We also see that the "Factor" estimates are substantially shifted away from the "Factor Lasso" estimates relative to standard errors and that the factor-lasso estimates are the most precise in the sense of having the

shortest confidence intervals. Both findings are consistent with the asymptotic theory and with the simulation results.

NOTES

1. Our results will immediately apply to the case where d_{it} is an $r \times 1$ vector with r fixed. We also note that our results clearly apply to models without additive fixed effects or to a single cross-section, though we treat only the PPFM defined in (1.1)–(1.3) in the formal analysis. We also consider only the case where d_{it} is exogenous, but note that it would be straightforward to extend the results to accommodate endogenous d_{it} when a low-dimensional set of excluded instruments is available. We provide results for a cross-sectional instrumental variables version of the model in both a simulation and an empirical example in the Supplemental Appendix.

2. Though the same set of f_i appears in both the outcome and treatment equations, components of ξ_t and δ_{dt} in (1.1) and (1.2) may be zero and these zeros may occur at different positions. Thus, the outcome and the treatment are allowed to depend on different elements of f_i.

3. Hahn et al. (2013) consider a similar structure to (1.1)–(1.3) which excludes the individual and time effects and imposes that the ϵ_{it} are i.i.d. Gaussian innovations. They refer to this model as a partial factor model.

4. Note that (1.1)–(1.2) are observationally equivalent to models that replace U_{it} with X_{it} or that include X_{it} in addition to U_{it}, under (1.3), after suitable redefinition of all parameters except α, which is unchanged by the substitution.

5. See also Bonhomme and Manresa (2015) for a distinct but related approach based on a grouped fixed effects model.

6. We note that recovering the untransformed f_i and U_{it} would only be possible with large n and T due to the presence of the unrestricted fixed effects. Fortunately, recovering these quantities is unnecessary within the model with common coefficients θ, γ_d, and α as only $\tilde{f}_i$ and $\tilde{U}_{it}$ appear in the equations of interest. This simplification would not generally occur if we allowed heterogeneity in θ, γ_d, or α over time or across individuals, and we would need to consider incidental parameters bias introduced by removing the additive fixed effects. We leave exploration of this issue to future research.

7. We use $c_0 = 1.1$ and $q_n = 0.1/\log(n)$ in the simulation and empirical examples.

8. We obtain $\widehat{e}_{it}$ and $\widehat{\eta}_{it}$ through an iterative algorithm similar to that of Belloni et al. (2014). First, we start from preliminary estimates $[\widehat{\Psi}^y]^0_{j,j} := \sqrt{\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \sum_{t'=1}^T \widehat{U}_{it,j}\widehat{U}_{it',j}\tilde{y}_{it}\tilde{y}_{it'}}$ and $[\widehat{\Psi}^d]^0_{j,j} := \sqrt{\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \sum_{t'=1}^T \widehat{U}_{it,j}\widehat{U}_{it',j}\tilde{d}_{it}\tilde{d}_{it'}}$. We then run (2.6) and (2.7) with the diagonal entries of $\widehat{\Psi}^y$ and $\widehat{\Psi}^d$ replaced with $[\widehat{\Psi}^y]^0_{j,j}$ and $[\widehat{\Psi}^d]^0_{j,j}$. This procedure provides an initial estimator $(\tilde{\gamma}^0_y, \tilde{\gamma}^0_d)$ which then provides $\widehat{e}_{it} = \tilde{y}_{it} - \widehat{\delta}'_{yt}\widehat{f}_i - \widehat{U}'_{it}\tilde{\gamma}^0_y$ and $\widehat{\eta}_{it} = \tilde{d}_{it} - \widehat{\delta}'_{dt}\widehat{f}_i - \widehat{U}'_{it}\tilde{\gamma}^0_d$.

9. It is straightforward to check that Assumption 3.2 holds in the simple case that $\{U_t, \epsilon_t, \eta_t, \}_{t\leq T}$ are independent across t, and $\{U_{it}, \eta_{it}, \epsilon_{it}\}_{i\leq n, t\leq T}$ are independent of $\{f_i\}_{i\leq n}$.

10. We follow the standard approach taken in the high-dimensional factor models literature by assuming K to be fixed. It would be straightforward to allow for K to grow at a slow rate as in Li, Li, and Shi (2017) at the cost of further technical and notational complication. Importantly, K would need to grow much more slowly than p for the purpose of dimension reduction using common factors; see discussion in Section 2.

11. In applications where obtaining the full lasso solution is not too burdensome, one may simply iterate to convergence.

12. Another commonly used iterative scheme that could readily be applied in the present setting is the "composite gradient method" (e.g., Nesterov, 2007; and Agarwal, Negahban, and Wainwright, 2012). We choose to focus on the coordinate descent method as our concrete example as it does not rely on additional tuning parameters and performed well numerically in preliminary simulation experiments. In addition, coordinate descent requires weaker regularity conditions than the composite gradient method for our theoretical analysis.

13. We have also experimented with $p = 10$ and $p = 50$ though we do not report the results to conserve space. The results with $p = 50$ are similar to those reported here. In line with the theory, no

procedure works well with $p = 10$ as factor extraction is difficult but models which do not attempt to extract the factors are misspecified.

14. The exact identities of the variables are available upon request. The data is from the U.S. Census Bureau USA Counties Database, http://www.census.gov/support/USACdataDownloads.html.

15. All results are based on weighted regression where we weight by the within-county average population over 1980–1999.

REFERENCES

Agarwal, A., S. Negahban, & M.J. Wainwright et al. (2012) Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* 40, 2452–2482.

Ahn, S.C. & A.R. Horenstein (2013) Eigenvalue ratio test for the number of factors. *Econometrica* 81, 1203–1227.

Andrews, D.W. (2002) Higher-order improvements of a computationally attractive *k*-step bootstrap for extremum estimators. *Econometrica* 70, 119–162.

Arellano, M. (1987) Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics* 49, 431–434.

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

Bai, J. (2009) Panel data models with interactive fixed effects. *Econometrica* 77, 1229–1279.

Bai, J. & K. Li (2014) Theory and methods of panel data models with interactive effects. *The Annals of Statistics* 42, 142–170.

Bai, J. & S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.

Bai, J. & S. Ng (2006) Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74, 1133–1150.

Belloni, A., D. Chen, V. Chernozhukov, & C. Hansen (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80, 2369–2429.

Belloni, A., V. Chernozhukov, I. Fernández-Val, & C. Hansen (2017) Program evaluation with high-dimensional data. *Econometrica* 85, 233–298.

Belloni, A., V. Chernozhukov, & C. Hansen (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81, 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, & D. Kozbur (2016) Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics* 34, 590–605.

Bernanke, B.S., J. Boivin, & P. Eliasz (2005) Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics* 120, 387–422.

Bonhomme, S. & E. Manresa (2015) Grouped patterns of heterogeneity in panel data. *Econometrica* 83, 1147–1184.

Chatterjee, A. & S.N. Lahiri (2011) Bootstrapping lasso estimators. *Journal of the American Statistical Association* 106, 608–625.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, & W. Newey (2016) Double machine learning for treatment and causal parameters. ArXiv e-prints 1608.00060.

Chernozhukov, V., D. Chetverikov, & K. Kato (2013) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics* 41, 2786–2819.

Cook, P.J. & J. Ludwig (2006) The social costs of gun ownership. *Journal of Public Economics* 90, 379–391.

Dezeure, R., P. Bühlmann, & C.-H. Zhang (2017) High-dimensional simultaneous inference with the bootstrap. *Test* 26, 685–719.

Fan, J. & R. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

Fan, J., L. Xue, & J. Yao (2017) Sufficient forecasting using factor models. *Journal of Econometrics* 201, 292–306.

Fu, W.J. (1998) Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7, 397–416.
Hahn, P.R., S. Mukeherjee, & C. Carvalho (2013) Partial factor modeling: Predictor dependent shrinkage for linear regression. *Journal of the American Statistical Association* 108, 999–1008.
Hsiao, C., H.S. Ching, & S. Wan (2012) A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with Mainland China. *Journal of Applied Econometrics* 27, 705–740.
Kadkhodaie, M., M. Sanjabi, & Z.-Q. Luo (2014) On the linear convergence of the approximate proximal splitting method for non-smooth convex optimization. *Journal of the Operations Research Society of China* 2, 123–141.
Li, H., Q. Li, & Y. Shi (2017) Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics* 197, 76–86.
Li, K.T. & D.R. Bell (2017) Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics* 197, 65–75.
Liang, K.-Y. & S. Zeger (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
Loh, P.-L. & M.J. Wainwright (2015) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* 16, 559–616.
Mammen, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* 21, 255–285.
Moon, H.R. & M. Weidner (2017) Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory* 33, 158–195.
Moon, R. & M. Weidner (2015) Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* 83, 1543–1579.
Nesterov, Y. (2007) Gradient Methods for Minimizing Composite Objective Function. Technical report, University College London.
Pesaran, H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74, 967–1012.
Prakasa Rao, B. (2009) Conditional independence, conditional mixing and conditional association. *Annals of the Institute of Statistical Mathematics* 61, 441–460.
Raskutti, G., M.J. Wainwright, & B. Yu (2010) Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research* 99, 2241–2259.
Stock, J. & M. Watson (2002) Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.
Su, L. & Q. Chen (2013) Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory* 29, 1079–1135.
van de Geer, S., P. Bühlmann, Y. Ritov, & R. Dezeure (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* 42, 1166–1202.
Wager, S. & S. Athey (2017) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*. Forthcoming 2018.
Zhang, C.-H. & S.S. Zhang (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* 76, 217–242.

APPENDIX A: Proofs of Theorem 3.1 and Corollary 3.1

Recall

$$\tilde{y}_{it} = \alpha \tilde{d}_{it} + \tilde{\xi}_t' \tilde{f}_i + \tilde{U}_{it}'\theta + \tilde{\epsilon}_{it}, \tag{A.1}$$

$$\tilde{d}_{it} = \tilde{\delta}_{dt}' \tilde{f}_i + \tilde{U}_{it}'\gamma_d + \tilde{\eta}_{it}, \tag{A.2}$$

$$\tilde{X}_{it} = \tilde{\Lambda}_t \tilde{f}_i + \tilde{U}_{it}. \tag{A.3}$$

Let $\tilde{U}$ be the $(pT) \times n$ matrix of $\tilde{U}_{it}$ defined in Section D. Let $(KT) \times 1$ matrices $\tilde{\Xi} = (\tilde{\xi}_1', \ldots, \tilde{\xi}_T')'$ and $\tilde{\Delta}_d = (\tilde{\delta}_{d1}', \ldots, \tilde{\delta}_{dT}')'$. Also define the $n \times K$ matrix $\tilde{F} = (\tilde{f}_1, \ldots, \tilde{f}_n)'$. Then (A.1) and (A.2) can be written in the matrix form:

$$\tilde{Y} = \tilde{D}\alpha + (I_T \otimes \tilde{F})\tilde{\Xi} + \tilde{U}\theta + \tilde{\epsilon}$$
$$\tilde{D} = (I_T \otimes \tilde{F})\tilde{\Delta}_d + \tilde{U}\gamma_d + \tilde{\eta}.$$

Note that $\widehat{\eta} = M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{D}$. Hence,

$$\begin{aligned}\widehat{\alpha} &= (\widehat{\eta}'\widehat{\eta})^{-1}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{Y} \\ &= \alpha + (\widehat{\eta}'\widehat{\eta})^{-1}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})[(I_T \otimes \tilde{F})\tilde{\Xi} + \tilde{U}\theta + \tilde{\epsilon}] \\ &= \alpha + (\widehat{\eta}'\widehat{\eta})^{-1}(\widehat{\eta} - \tilde{\eta})' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{\epsilon} + (\widehat{\eta}'\widehat{\eta})^{-1}\tilde{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{\epsilon} \\ &\quad + (\widehat{\eta}'\widehat{\eta})^{-1}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}\tilde{F})\tilde{\Xi} + (\widehat{\eta}'\widehat{\eta})^{-1}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\theta.\end{aligned}$$

Note that $\tilde{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{\epsilon} = \tilde{\eta}'\tilde{\epsilon} - \tilde{\eta}'(I_T \otimes P_{\widehat{F}})\tilde{\epsilon} - \tilde{\eta}' P_{\widehat{U}_{\widehat{J}}}\tilde{\epsilon} + \tilde{\eta}' P_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{\epsilon}$. Hence,

$$\sqrt{nT}\left(\frac{1}{nT}\widehat{\eta}'\widehat{\eta}\right)(\widehat{\alpha} - \alpha) = \frac{1}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} + \sum_{i=1}^{6} A_i, \tag{A.4}$$

where

$$A_1 = \frac{1}{\sqrt{nT}}(\widehat{\eta} - \tilde{\eta})' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{\epsilon}, \qquad A_2 = \frac{1}{\sqrt{nT}}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}\tilde{F})\tilde{\Xi},$$
$$A_3 = -\frac{1}{\sqrt{nT}}\tilde{\eta}'(I_T \otimes P_{\widehat{F}})\tilde{\epsilon}, \qquad A_4 = \frac{1}{\sqrt{nT}}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\theta,$$
$$A_5 = -\frac{1}{\sqrt{nT}}\tilde{\eta}' P_{\widehat{U}_{\widehat{J}}}\tilde{\epsilon}, \qquad A_6 = \frac{1}{\sqrt{nT}}\tilde{\eta}' P_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{\epsilon} = 0.$$

We shall prove that $A_i = o_P(1)$ for $i = 1, \ldots, 6$ and $\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\tilde{\eta}'\tilde{\eta} = o_P(1)$. Note

$$\begin{aligned}\widehat{\eta} &= M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{D} = M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})((I_T \otimes \tilde{F})\tilde{\Delta}_d + \tilde{U}\gamma_d + \tilde{\eta}) \\ &= M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}\tilde{F})\tilde{\Delta}_d + M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\gamma_d + M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{\eta}.\end{aligned}$$

Using the fact that $M_{\widehat{F}}\widehat{F} = 0$, it can be proven that

$$\begin{aligned}\frac{1}{\sqrt{nT}}(\widehat{\eta} - \tilde{\eta}) &= \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Delta}_d \\ &\quad + \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\gamma_d - \frac{1}{\sqrt{nT}} P_{\widehat{U}_{\widehat{J}}}\tilde{\eta} - \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{\eta}.\end{aligned} \tag{A.5}$$

In the subsequent subsections, we provide bounds for A_i for $i = 1, \ldots, 6$ and for $\frac{1}{\sqrt{nT}}\|\widehat{\eta} - \tilde{\eta}\|_2$.

A.1. Bounding $\widehat{\eta}-\tilde{\eta}$

Write

$$\psi_n := \kappa_n |J|_0^{1/2} + \|R_y\|_1 + \Delta_F |J|_0 + \sqrt{\frac{|J|_0}{n}}. \tag{A.6}$$

PROPOSITION A.1. $\frac{1}{\sqrt{nT}}\|\widehat{\eta}-\tilde{\eta}\|_2 = O_P(\psi_n)$.

Proof. Note that $\|\tilde{\Delta}_d\|_2 = O(\sqrt{T})$. Hence by Lemma H.1,

$$\|\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Delta}_d\|_2 \leq O_P(1)\frac{1}{\sqrt{nT}}\|\tilde{F}H - \widehat{F}\|_F \|\tilde{\Delta}_d\|_2 = O_P(\Delta_F)$$

$$\|\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}\tilde{U}\gamma_d\|_2 = O_P\left(\kappa_n |J|_0^{1/2} + \|R_y\|_1 + \Delta_F |J|_0 + \sqrt{\frac{|J|_0}{n}}\right),$$

$$\|\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{U}\gamma_d\|_2 \leq \|\frac{1}{\sqrt{nT}}(I_T \otimes P_{\widehat{F}})\tilde{U}\gamma_d\|_2 = O_P\left(\sqrt{\frac{|J|_0}{n}} + \Delta_F |J|_0\right),$$

$$\|\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{\eta}\|_2 \leq \|\frac{1}{\sqrt{nT}}(I_T \otimes P_{\widehat{F}})\tilde{\eta}\|_2 = O_P\left(\frac{1}{\sqrt{n}} + \Delta_F\right),$$

$$\|\frac{1}{\sqrt{nT}} P_{\widehat{U}_{\widehat{J}}}\tilde{\eta}\|_2 = O_P\left(\sqrt{|J|_0 \frac{\log p}{nT}}\right).$$

Hence, equation (A.5) implies $\frac{1}{\sqrt{nT}}\|\widehat{\eta}-\tilde{\eta}\|_2 = O_P(\psi_n)$. ■

A.2. Showing $A_1, A_3, A_5, A_6 = o_P(1)$

By equation (A.5), Lemma H.10, $P_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}}) = 0$, and noting that $P_{\widehat{U}_{\widehat{J}}} M_{\widehat{U}_{\widehat{J}}} = 0$, we can show that

$$\begin{aligned} A_1 &= \frac{1}{\sqrt{nT}}(\widehat{\eta}-\tilde{\eta})' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{\epsilon} \\ &= \tilde{\epsilon}' \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\gamma_d + \tilde{\epsilon}' \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Delta}_d. \end{aligned}$$

It then follows from Lemma H.3(i)(v) that $A_1 = o_P(1)$.

We can also immediately apply Lemma H.3(iii) to establish that $A_3 = o_P(1)$.

Also, it follows from Lemma H.1(iv) that, since $|J|_0^2 \log^2 p = o(nT)$,

$$|A_5| = \left|\frac{1}{\sqrt{nT}}\tilde{\eta}' P_{\widehat{U}_{\widehat{J}}}\tilde{\epsilon}\right| \leq \sqrt{nT}\left\|\frac{1}{\sqrt{nT}} P_{\widehat{U}_{\widehat{J}}}\tilde{\epsilon}\right\|_2 \left\|\frac{1}{\sqrt{nT}} P_{\widehat{U}_{\widehat{J}}}\tilde{\eta}\right\|_2 = O_P\left(\frac{|J|_0 \log p}{\sqrt{nT}}\right) = o_P(1).$$

Finally, it follows immediately from Lemma H.10 that $A_6 = 0$.

A.3. Showing $A_2 = o_P(1)$

By (A.5),

$$A_2 = \frac{1}{\sqrt{nT}}\widehat{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Xi}$$

$$= \frac{1}{\sqrt{nT}}\tilde{\eta}' M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Xi} \quad \textbf{(A.7)}$$

$$+ \tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Delta}_d \quad \textbf{(A.8)}$$

$$- \tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{\eta} \quad \textbf{(A.9)}$$

$$+ \tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\gamma_d. \quad \textbf{(A.10)}$$

It follows from Lemma H.3(i) that (A.7) is $o_P(1)$. By the Cauchy–Schwarz inequality and under the assumption that $\sqrt{nT}\,\Delta_F^2 = o(1)$, (A.8) is bounded by

$$|\tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}}(\tilde{F}H - \widehat{F})H^{-1})\tilde{\Delta}_d|$$

$$\leq \frac{1}{\sqrt{nT}} \max_{G=\tilde{\Xi},\tilde{\Delta}_d} \|G'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}}) M_{\widehat{U}_{\widehat{J}}}\|_2^2$$

$$\leq \frac{1}{\sqrt{nT}} \max_{G=\tilde{\Xi},\tilde{\Delta}_d} \|G'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\|_2^2$$

$$\leq \frac{1}{\sqrt{nT}} \max_{g_t=\tilde{\xi}_t,\tilde{\delta}_{dt}} \sum_t \|g_t' H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}}\|_2^2$$

$$\leq O_P\left(\frac{\sqrt{T}}{\sqrt{n}}\right)\|\tilde{F}H - \widehat{F}\|_F^2 = O_P(\sqrt{nT}\,\Delta_F^2) = o_P(1).$$

Term (A.9) equals

$$-\frac{1}{\sqrt{nT}}\tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}}) M_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}})\tilde{\eta}$$

$$= -\frac{1}{\sqrt{nT}}\tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})(I_T \otimes P_{\widehat{F}})\tilde{\eta} = 0,$$

where the first equality is due to $P_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}}) = 0$ and the second equality is due to $M_{\widehat{F}} P_{\widehat{F}} = 0$ and the fact that the Kronecker product satisfies $(A \otimes B)(C \otimes D) = AC \otimes BD$.

Finally, using $M_{\widehat{F}} P_{\widehat{F}} = 0$ and $P_{\widehat{U}_{\widehat{J}}}(I_T \otimes P_{\widehat{F}}) = 0$, (A.10) equals

$$\tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}(I_T \otimes M_{\widehat{F}})\tilde{U}\gamma_d \quad \textbf{(A.11)}$$

$$= \tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}})\frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}}\tilde{U}\gamma_d$$

$$
\begin{aligned}
&- \tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}}) \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}} (I_T \otimes P_{\widehat{F}}) \tilde{U}\gamma_d \\
&= \tilde{\Xi}'(I_T \otimes H^{'-1}(\tilde{F}H - \widehat{F})' M_{\widehat{F}}) \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}} \tilde{U}\gamma_d = o_P(1),
\end{aligned}
$$

where the last equality follows from Lemma H.3 (vi).

Hence, $A_2 = o_P(1)$.

A.4. Showing $A_4 = o_P(1)$

$$
\begin{aligned}
A_4 &= \frac{1}{\sqrt{nT}} \widehat{\eta}' M_{\widehat{U}_{\widehat{J}}} (I_T \otimes M_{\widehat{F}}) \tilde{U}\theta \\
&= \frac{1}{\sqrt{nT}} \tilde{\eta}' M_{\widehat{U}_{\widehat{J}}} (I_T \otimes M_{\widehat{F}}) \tilde{U}\theta && \textbf{(A.12)} \\
&\quad + \frac{1}{\sqrt{nT}} \theta' \tilde{U}' (I_T \otimes M_{\widehat{F}}) M_{\widehat{U}_{\widehat{J}}} (I_T \otimes M_{\widehat{F}}) \tilde{U}\gamma_d && \textbf{(A.13)} \\
&\quad + \frac{1}{\sqrt{nT}} \theta' \tilde{U}' (I_T \otimes M_{\widehat{F}}) M_{\widehat{U}_{\widehat{J}}} (I_T \otimes M_{\widehat{F}} (\tilde{F}H - \widehat{F}) H^{-1}) \tilde{\Delta}_d. && \textbf{(A.14)}
\end{aligned}
$$

It follows from Lemma H.3(iii) that term (A.12) is $o_P(1)$.

By Lemma H.1(i) and (ii), we can bound term (A.13) by

$$
\begin{aligned}
&\frac{1}{\sqrt{nT}} \theta' \tilde{U}' (I_T \otimes M_{\widehat{F}}) M_{\widehat{U}_{\widehat{J}}} (I_T \otimes M_{\widehat{F}}) \tilde{U}\gamma_d \\
&\quad \le \sqrt{nT} \max_{g=\theta,\gamma_d} \left\| \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}} (I_T \otimes M_{\widehat{F}}) \tilde{U} g \right\|_2^2 \\
&\quad \le 2\sqrt{nT} \max_{g=\theta,\gamma_d} \left\| \frac{1}{\sqrt{nT}} M_{\widehat{U}_{\widehat{J}}} \tilde{U} g \right\|_2^2 + 2\sqrt{nT} \max_{g=\theta,\gamma_d} \left\| \frac{1}{\sqrt{nT}} (I_T \otimes P_{\widehat{F}}) \tilde{U} g \right\|_2^2 \\
&\quad \le \sqrt{nT} O_P \left(\|R_y\|_1^2 + \kappa_n^2 |J|_0 + |J|_0^2 \Delta_F^2 + \frac{|J|_0}{n} \right) = o_P(1),
\end{aligned}
$$

under the assumption $\left(\kappa_n^2 |J|_0 + \|R_y\|_1^2 + \Delta_F^2 |J|_0^2 + \frac{|J|_0}{n}\right)\sqrt{nT} = o(1)$.

The same argument as that employed in the bound given by equation (A.11) yields that term (A.14) is $o_P(1)$.

A.5. Proof of Theorem 3.1

(i) Write $\iota_{it} := (\eta_{it} - \bar{\eta}_{i\cdot})^2$.

Step 1: Show $|\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\sum_{it} E\iota_{it}| = o_P(1)$.

It follows from Proposition A.1 that $|\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\tilde{\eta}'\tilde{\eta}| = o_P(1)$. Also,

$$
\frac{1}{nT}\tilde{\eta}'\tilde{\eta} = \frac{1}{nT}\sum_{i,t} \tilde{\eta}_{it}^2 = \frac{1}{nT}\sum_{i,t} (\eta_{it} - \bar{\eta}_{i\cdot})^2 - \frac{1}{T}\sum_t \bar{\eta}_{\cdot t}^2 + \bar{\bar{\eta}}^2.
$$

We have that $E\left[\frac{1}{T}\sum_t \bar{\eta}_{\cdot t}^2\right] = \frac{1}{T}\sum_t \frac{1}{n^2}\sum_i E\left[\eta_{it}^2\right] = O(1/n)$ and that $\bar{\bar{\eta}}^2 = o_P(1)$. Hence,

$$\frac{1}{nT}\tilde{\eta}'\tilde{\eta} = \frac{1}{nT}\sum_{it}(\eta_{it} - \bar{\eta}_{i\cdot})^2 + o_P(1) = \frac{1}{nT}\sum_{it}\iota_{it} + o_P(1).$$

Note that

$$\mathrm{Var}\left(\frac{1}{nT}\sum_{it}\iota_{it}\right) = \frac{1}{n^2T^2}\sum_i \mathrm{Var}\left(\sum_t \iota_{it}\right) = O(1/n).$$

Hence, $\left|\frac{1}{nT}\sum_{it}\iota_{it} - \frac{1}{nT}\sum_{it}E\iota_{it}\right| = o_P(1)$. We then have

$$\left|\frac{1}{nT}\tilde{\eta}'\tilde{\eta} - \frac{1}{nT}\sum_{it}E\iota_{it}\right| = o_P(1), \quad \left|\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\sum_{it}E\iota_{it}\right| = o_P(1), \tag{A.15}$$

and $\frac{1}{nT}\widehat{\eta}'\widehat{\eta}$ is bounded away from zero.

Let $m_{n,i} = \frac{1}{\sqrt{T}}\sum_t(\eta_{it} - \bar{\eta}_{i\cdot})(\epsilon_{it} - \bar{\epsilon}_{i\cdot})$, $b_n = \left[\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_i m_{n,i}\right)\right]^{-1/2}$, and $g_{n,i} = b_n m_{n,i}$. In addition, let $s_n^2 = \sum_i \mathrm{Var}(g_{n,i}) = \sum_i \mathrm{Var}(m_{n,i})b_n^2 = n$.

Step 2: Show $\frac{b_n}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} = \frac{1}{s_n}\sum_i g_{n,i} + o_P(1)$.

By Assumption 3.1, $\{\eta_i, \epsilon_i\}_{i\le n}$ are independent over i. Hence

$$\begin{aligned}
E\left(\frac{1}{T}\sum_t \bar{\eta}_{\cdot t}\bar{\epsilon}_{\cdot t}\right)^2 &= \frac{1}{T}\sum_t \frac{1}{n}\sum_i \frac{1}{n}\sum_j \frac{1}{T}\sum_{s=1}^T \frac{1}{n}\sum_{l=1}^n \frac{1}{n}\sum_{v=1}^n E\eta_{ls}\epsilon_{vs}\eta_{it}\epsilon_{jt} \\
&= \sum_t \frac{1}{n^4}\sum_i \frac{1}{T^2}\sum_{s=1}^T E\eta_{is}\epsilon_{is}\eta_{it}\epsilon_{it} \\
&\quad + \frac{1}{T}\sum_t \frac{1}{n^4}\sum_i\sum_{j\neq i}\frac{1}{T}\sum_{s=1}^T E\eta_{js}\epsilon_{jt}E\epsilon_{is}\eta_{it} \\
&\quad + \frac{1}{T}\sum_t \frac{1}{n^4}\sum_i\sum_{j\neq i}\frac{1}{T}\sum_{s=1}^T E\eta_{is}\eta_{it}E\epsilon_{js}\epsilon_{jt} = O\left(\frac{1}{n^2}\right).
\end{aligned}$$

In the above, by Assumption 3.1, there is $C > 0$, $|E\eta_{js}\epsilon_{js}\eta_{it}\epsilon_{it}| + |E\eta_{is}\eta_{it}| + |E\epsilon_{js}\epsilon_{jt}| < C$ for all i, j, s, t. Thus the right hand side of the second equality equals $O(\frac{1}{n^3} + \frac{1}{n^2} + \frac{1}{n^2}) = O\left(\frac{1}{n^2}\right)$.

We have $E[m_{n,i}] = 0$. We also have, under our assumptions, $\mathrm{Var}(m_{n,i})$ and b_n bounded away from both zero and infinity uniformly in i. Then

$$\begin{aligned}
\frac{b_n}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} &= \frac{b_n}{\sqrt{nT}}\sum_{it}(\eta_{it} - \bar{\eta}_{i\cdot})(\epsilon_{it} - \bar{\epsilon}_{i\cdot}) - \frac{b_n\sqrt{nT}}{T}\sum_t \bar{\eta}_{\cdot t}\bar{\epsilon}_{\cdot t} + b_n\sqrt{nT}\bar{\bar{\eta}}\bar{\bar{\epsilon}} \\
&= \frac{1}{s_n}\sum_i g_{n,i} + o_P(1).
\end{aligned}$$

Step 3: Apply the CLT.

We now verify the Lindeberg condition for the triangular array $\{g_{n,i}\}$. By Assumption 3.2 there is $\delta > 0$ so that $\max_i E|m_{n,i}|^{2+\delta} < C$ and that $\lambda_{\min}(\frac{1}{n}\sum_i \mathrm{Var}(m_{n,i})) > c_0$, we have $b_n < C$, and $E|g_{n,i}|^{2+\delta} \le CE|m_{n,i}|^{2+\delta} < C$. For any $\varepsilon > 0$, and by Holder's inequality: $E|ab| \le (E|a|^{1+\delta/2})^{1/(1+\delta/2)}(E|b|^{\delta'})^{1/\delta'}$, for $\delta' = (2+\delta)/\delta$,

$$s_n^{-2}\sum_i E\left[g_{n,i}^2 1\{|g_{n,i}| > \varepsilon s_n\}\right] \le \frac{1}{n}\sum_i (E|g_{n,i}|^{2+\delta})^{1/(1+\delta/2)} P(|g_{n,i}| > \varepsilon s_n)^{1/\delta'}$$
$$\le \max_i (E|g_{n,i}|^{2+\delta})^{1/(1+\delta/2)}(E|g_{n,i}|)^{1/\delta'}(\varepsilon\sqrt{n})^{-1/\delta'} \to 0.$$

This implies, by the Lindeberg central limit theorem, $\frac{b_n}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} \to^d \mathcal{N}(0,1)$.

In the previous subsections, we have proven $A_i = o_P(1)$ for $i = 1,\ldots,6$. Hence, it follows from (A.4) that $\sqrt{nT}(\frac{1}{nT}\widehat{\eta}'\widehat{\eta})(\widehat{\alpha}-\alpha) = \frac{1}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} + o_P(1)$. In addition, recall $\iota_{it} := (\eta_{it} - \bar{\eta}_{i\cdot})^2$,

$$b_n|(\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\sum_{i,t} E[\iota_{it}])\sqrt{nT}(\widehat{\alpha}-\alpha)|$$
$$\le b_n|\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\sum_{i,t} E[\iota_{it}]|(\frac{1}{nT}\widehat{\eta}'\widehat{\eta})^{-1}|\frac{1}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} + o_P(1)| = o_P(1).$$

Therefore,

$$\sigma_{\eta\epsilon}^{-1/2}\sigma_\eta^2\sqrt{nT}(\widehat{\alpha}-\alpha) = b_n\frac{1}{nT}\sum_{i,t} E[\iota_{it}]\sqrt{nT}(\widehat{\alpha}-\alpha) = b_n\frac{1}{nT}\widehat{\eta}'\widehat{\eta}\sqrt{nT}(\widehat{\alpha}-\alpha) + o_P(1)$$
$$= \frac{b_n}{\sqrt{nT}}\tilde{\eta}'\tilde{\epsilon} + o_P(1) \to^d \mathcal{N}(0,1).$$

(ii) To verify normality with the estimated asymptotic variance, we need to prove consistency of $\widehat{\sigma}_{\eta\epsilon}$ and $\widehat{\sigma}_\eta^2$. We have previously shown $|\frac{1}{nT}\widehat{\eta}'\widehat{\eta} - \frac{1}{nT}\sum_{i,t} E[\iota_{it}]| = o_P(1)$ which establishes consistency of $\sigma_\eta^2 = \frac{1}{nT}\sum_{i,t} E[\iota_{it}]$. Hence, it remains to prove $\widehat{\sigma}_{\eta\epsilon} - \sigma_{\eta\epsilon} = o_P(1)$. Recall that

$$\sigma_{\eta\epsilon} = \mathrm{Var}\left(\frac{1}{\sqrt{nT}}\sum_{i=1}^n\sum_{t=1}^T(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right).$$

Step 1: Bound $\Delta_1 := \widehat{\sigma}_{\eta\epsilon} - \frac{1}{nT}\sum_{i=1}^n(\sum_{t=1}^T \tilde{\eta}_{it}\tilde{\epsilon}_{it})^2$. We have

$$\Delta_1 = \frac{1}{nT}\sum_{i=1}^n\left[\left(\sum_{t=1}^T\widehat{\eta}_{it}\widehat{\epsilon}_{it}\right)^2 - \left(\sum_{t=1}^T\tilde{\eta}_{it}\tilde{\epsilon}_{it}\right)^2\right]$$
$$= \frac{1}{nT}\sum_{i=1}^n\left[\sum_{t=1}^T(\widehat{\eta}_{it}\widehat{\epsilon}_{it} + \tilde{\eta}_{it}\tilde{\epsilon}_{it})\right]\left[\sum_{t=1}^T(\widehat{\eta}_{it}\widehat{\epsilon}_{it} - \tilde{\eta}_{it}\tilde{\epsilon}_{it})\right]$$

$$= \frac{1}{nT}\sum_{i=1}^{n}\left[\sum_{t=1}^{T}(\widehat{\eta}_{it}\widehat{\epsilon}_{it} - \tilde{\eta}_{it}\tilde{\epsilon}_{it})\right]^2 \tag{A.16}$$

$$+ \frac{2}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it}\right)\sum_{s=1}^{T}(\widehat{\eta}_{is}\widehat{\epsilon}_{is} - \tilde{\eta}_{is}\tilde{\epsilon}_{is}). \tag{A.17}$$

By Lemma H.4, term (A.16) is $o_P(1)$.

For term (A.17), we have

$$E\left[\frac{1}{n}\sum_{i}|\frac{1}{T}\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it}|^2\right] = E\left[\frac{1}{n}\sum_{i}|\frac{1}{T}\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{\cdot t})(\epsilon_{it}-\bar{\epsilon}_{\cdot t}) - (\bar{\eta}_{i\cdot}-\bar{\bar{\eta}})(\bar{\epsilon}_{i\cdot}-\bar{\bar{\epsilon}})|^2\right] = O\left(\frac{1}{T}\right).$$

since the process $\{(\eta_t, \epsilon_t)\}_{t=-\infty}^{+\infty}$ satisfies the strong mixing condition with exponential tails. Thus by Cauchy–Schwarz,

$$\left[\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{T}\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it})\sum_{s=1}^{T}(\widehat{\eta}_{is}\widehat{\epsilon}_{is} - \tilde{\eta}_{is}\tilde{\epsilon}_{is})\right]^2 \le \left[\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{T}\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(\sum_{s=1}^{T}(\widehat{\eta}_{is}\widehat{\epsilon}_{is} - \tilde{\eta}_{is}\tilde{\epsilon}_{is}))^2\right]$$

$$\le O_P\left(\frac{1}{T}\right)\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{s=1}^{T}(\widehat{\eta}_{is}\widehat{\epsilon}_{is} - \tilde{\eta}_{is}\tilde{\epsilon}_{is})\right)^2 = o_P(1).$$

This result then implies $\Delta_1 = o_P(1)$.

Step 2: Bound $\Delta_2 := \frac{1}{nT}\sum_{i=1}^{n}(\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it})^2 - \frac{1}{nT}\sum_{i=1}^{n}(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}))^2$.

Note that

$$\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it} = \sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}) \underbrace{- \sum_{t=1}^{T}\bar{\eta}_{\cdot t}(\epsilon_{it}-\bar{\epsilon}_{\cdot t}) - \sum_{t=1}^{T}\eta_{it}\bar{\epsilon}_{\cdot t} + T\bar{\bar{\eta}}(\bar{\epsilon}_{i\cdot}-\bar{\bar{\epsilon}}) + T\bar{\eta}_{i\cdot}\bar{\bar{\epsilon}}}_{B_i}$$

and that $\frac{1}{T}\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}) = \frac{1}{T}\sum_{t=1}^{T}\eta_{it}\epsilon_{it} - \bar{\eta}_{i\cdot}\bar{\epsilon}_{i\cdot}$. Hence,

$$\frac{1}{nT}\sum_{i=1}^{n}(\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it})^2 = \frac{1}{nT}\sum_{i=1}^{n}(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}) + B_i)^2$$

$$= \frac{1}{nT}\sum_{i=1}^{n}(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}))^2 + \frac{1}{nT}\sum_{i=1}^{n}B_i^2$$

$$+ \frac{2}{nT}\sum_{i=1}^{n}B_i\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}).$$

Note that

$$\left[\frac{1}{nT}\sum_{i=1}^{n}B_i\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right]^2$$

$$\le \frac{1}{nT}\sum_{i}B_i^2\frac{1}{n}\sum_{i}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right)^2$$

$$= \frac{1}{nT}\sum_i B_i^2 O_P\left(\frac{1}{n}\sum_i \mathrm{Var}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}))\right)$$
$$= \frac{1}{nT}\sum_i B_i^2 O_P(\sigma_{\eta\epsilon}).$$

Therefore, $|\Delta_2| \le \frac{1}{nT}\sum_i B_i^2 + (\frac{1}{nT}\sum_i B_i^2)^{1/2} O_P(1)$. It suffices to prove $\frac{1}{nT}\sum_i B_i^2 = o_P(1)$. In fact, $\frac{1}{nT}\sum_i B_i^2 \le C\sum_{l=1}^{4}\bar{A}_l$ for a constant $C > 0$ and

$$\bar{A}_1 = \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\eta_{it}\bar{\epsilon}_{\cdot t}\right)^2, \quad \bar{A}_2 = \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\bar{\eta}_{\cdot t}(\epsilon_{it}-\bar{\epsilon}_{\cdot t})\right)^2,$$
$$\bar{A}_3 = \frac{T}{n}\bar{\bar{\epsilon}}^2\sum_{i=1}^{n}\bar{\eta}_{i\cdot}^2, \quad \bar{A}_4 = \frac{T}{n}\bar{\bar{\eta}}^2\sum_{i=1}^{n}(\bar{\epsilon}_{i\cdot}-\bar{\bar{\epsilon}})^2,$$

where each $\bar{A}_l = O_P(E\bar{A}_l)$. We then have

$$E\bar{A}_1 = \frac{1}{n^3T}\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{s=1}^{T}\sum_{t=1}^{T}E\eta_{it}\epsilon_{jt}\eta_{is}\epsilon_{ms} = \frac{1}{n^3T}\sum_{i=1}^{n}\sum_{s=1}^{T}\sum_{t=1}^{T}E\eta_{it}\epsilon_{it}\eta_{is}\epsilon_{is}$$
$$+\frac{1}{n^3T}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{s=1}^{T}\sum_{t=1}^{T}E\eta_{it}\eta_{is}E\epsilon_{js}\epsilon_{jt} = O\left(\frac{T}{n}\right) = o(1).$$

Similarly, $E\bar{A}_2 = o(1)$. In addition, $\bar{\bar{\epsilon}}^2 = O_P(n^{-1})$ and $\bar{\bar{\eta}}^2 = O_P(n^{-1})$, so $\bar{A}_3$ and $\bar{A}_4$ are each $O_P(T/n) = o_P(1)$. Combining verifies that

$$\Delta_2 := \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\tilde{\eta}_{it}\tilde{\epsilon}_{it}\right)^2 - \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right)^2 = o_P(1). \qquad \textbf{(A.18)}$$

Step 3: Bound $\Delta_3 := \frac{1}{nT}\sum_{i=1}^{n}(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}))^2 - \sigma_{\eta\epsilon}$.

Note that $\sigma_{\eta\epsilon} = E\left[\frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right)^2\right]$, and let

$$m_{n,i} = \frac{1}{\sqrt{T}}\sum_t(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot}).$$

Then $\Delta_3 = \frac{1}{n}\sum_{i=1}^{n}(m_{n,i}^2 - Em_{n,i}^2)$. Because $\frac{1}{n}\sum_i \mathrm{Var}(m_{n,i}^2) = O(1)$, we have

$$E\Delta_3^2 = \mathrm{Var}(\Delta_3) = \mathrm{Var}\left(\frac{1}{n}\sum_i m_{n,i}^2\right) = \frac{1}{n^2}\sum_i \mathrm{Var}(m_{n,i}^2) = o(1), \qquad \textbf{(A.19)}$$

which implies

$$\Delta_3 := \frac{1}{nT}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}(\eta_{it}-\bar{\eta}_{i\cdot})(\epsilon_{it}-\bar{\epsilon}_{i\cdot})\right)^2 - \sigma_{\eta\epsilon} = o_P(1). \qquad \textbf{(A.20)}$$

Combining the above three steps, we reach $|\widehat{\sigma}_{\eta\epsilon} - \sigma_{\eta\epsilon}| = o_P(1)$.

Proof of Corollary 3.1. Given Theorem 3.1, the corollary follows from the same argument as that of Corollary 1(i) of Belloni et al. (2014). We thus refer to Belloni et al. (2014) for details. ■

Appendix B: Convergence of the k-step Bootstrap Lasso

In this section, we obtain the statistical convergence rate (in O_{P^*}) of the k-step bootstrap lasso estimators $\tilde{\gamma}_d^*$ and $\tilde{\gamma}_y^*$. We focus on $\tilde{\gamma}_y^*$, as the proof of $\tilde{\gamma}_d^*$ is similar. Recall that

$$\mathcal{L}_y^*(\gamma) = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(\tilde{y}_{it}^* - \widehat{\delta}_{yt}^{*\prime}\widehat{f}_i^* - \widehat{U}_{it}^{*\prime}\gamma)^2,$$
$$\tilde{\gamma}_{y,lasso}^* = \arg\min_{\gamma\in\mathbb{R}^p} \mathcal{L}_y^*(\gamma) + \kappa_n\|\widehat{\Psi}^y\gamma\|_1, \tag{B.1}$$

and that

$$\begin{aligned}\widehat{\gamma}_y &= \text{the post-lasso estimator based on the original data}\\ \tilde{\gamma}_y^* &= \text{the } k\text{-step lasso estimator based on the bootstrap data}\\ \tilde{\gamma}_{y,lasso}^* &= \text{the lasso estimator based on the bootstrap data}\\ &\quad\ \text{if a complete lasso program is carried out.}\end{aligned}$$

In particular, $\widehat{\gamma}_y$ is used as the coefficient when generating the bootstrap data.

We divide the proof into three subsections. Section B.1 proves the statistical convergence of $\|\tilde{\gamma}_{y,lasso}^* - \widehat{\gamma}_y\|_1$ in the bootstrap sampling space. Section B.2 quantifies the computational error $\|\tilde{\gamma}_y^* - \tilde{\gamma}_{y,lasso}^*\|_1$ and shows that the computational error of the k-step lasso is negligible using the assumed high-level conditions on the iterative scheme $\mathcal{S}_y(\cdot)$. Section B.3 verifies the high-level conditions for the coordinate descent, or "shooting", method (Fu (1998); Kadkhodaie, Sanjabi, and Luo (2014)).

Let P^* denote the probability measure in the bootstrap sampling space. More specifically, it is the probability measure generated by the conditional distribution of the bootstrap weights $\{\{w_i^x\}_{i\le n}^b\}_{b\le B}$, $x = U, Y, D$, given the data. We employ the usual definition of $o_{P^*}(1)$ and $O_{P^*}(1)$. We say that a sequence X_n^* in the bootstrap sampling space is $o_{P^*}(1)$ if for any $\varepsilon, \delta > 0$,

$$P\{P^*(|X_n^*| > \varepsilon) > \delta\} \to 0,$$

and that $X_n^* = O_{P^*}(1)$ if for any $\delta > 0$, there is $M > 0$, such that

$$P\{P^*(|X_n^*| > M) > \delta\} \to 0.$$

B.1. The Convergence of Lasso on Bootstrap Data

The main result in this subsection is the following proposition.

PROPOSITION B.1. $\|\tilde{\gamma}_{y,lasso}^* - \widehat{\gamma}_y\|_1 = O_{P^*}(\kappa_n|J|_0)$.

Proof. Recall that $\widehat{F}$, $\widehat{\delta}_{yt}$, and $\widehat{\gamma}_y$, respectively, denote the estimated factors, the estimator of δ_{yt}, and the post-lasso estimator of γ_y using the original data. We also have that $\tilde{U}_t^*$ denotes the wild bootstrapped idiosyncratic term in the factor equation and that the following relations hold:

$$\tilde{Y}_t^* = \widehat{F}\widehat{\delta}_{yt} + \tilde{U}_t^*\widehat{\gamma}_y + \tilde{e}_t^*, \quad \tilde{e}_t^* = \tilde{\epsilon}_t^* + \tilde{\eta}_t^*\widehat{\alpha}. \tag{B.2}$$

In addition, recall that $\widehat{\delta}_{yt}^*$ and $\widehat{U}_t^*$ denote the estimates obtained from the bootstrap data. Define

$$M_t^* = \widehat{F}\widehat{\delta}_{yt} - \widehat{F}^*\widehat{\delta}_{yt}^* + (\tilde{U}_t^* - \widehat{U}_t^*)\widehat{\gamma}_y, \quad \Delta_\gamma^* = \widehat{\gamma}_y - \tilde{\gamma}_{y,lasso}^*. \tag{B.3}$$

By definition, $\mathcal{L}_y^*(\tilde{\gamma}_{y,lasso}^*) + \kappa_n\|\widehat{\Psi}^y\tilde{\gamma}_{y,lasso}^*\|_1 \le \mathcal{L}_y^*(\widehat{\gamma}_y) + \kappa_n\|\widehat{\Psi}^y\widehat{\gamma}_y\|_1$, which implies

$$\frac{1}{nT}\sum_{t=1}^{T}\left(\|\widehat{U}_t^*\Delta_\gamma^*\|_2^2 + 2(\tilde{e}_t^{*'} + M_t^{*'})\widehat{U}_t^*\Delta_\gamma^*\right) + \kappa_n\|\widehat{\Psi}^y\tilde{\gamma}_{y,lasso}^*\|_1 \le \kappa_n\|\widehat{\Psi}^y\widehat{\gamma}_y\|_1. \tag{B.4}$$

By Lemma H.17 and $\kappa_n = \frac{2c_0}{\sqrt{nT}}\Phi^{-1}(1 - q_n/(2p))$ for some $c_0 > 1$,

$$\begin{aligned}
&\left|\frac{1}{nT}\sum_{t=1}^{T}2(\tilde{e}_t^{*'} + M_t^{*'})\widehat{U}_t^*\Delta_\gamma^*\right| \\
&\quad\le \left\|\frac{1}{nT}\sum_{t=1}^{T}2\tilde{e}_t^{*'}\tilde{U}_t^*\widehat{\Psi}^{y-1}\right\|_\infty\|\widehat{\Psi}^y\Delta_\gamma^*\|_1 + \left(\left\|\frac{1}{nT}\sum_{t=1}^{T}2\tilde{e}_t^{*'}(\widehat{U}_t^* - \tilde{U}_t^*)\right\|_\infty\right. \\
&\qquad\left. + \left\|\frac{1}{nT}\sum_{t=1}^{T}2M_t^{*'}\widehat{U}_t^*\Delta_\gamma^*\right\|_\infty\right)\|\widehat{\Psi}^y\Delta_\gamma^*\|_1\max_m(\widehat{\Psi}_m^y)^{-1} \\
&\quad\le \left[\frac{2}{\sqrt{nT}}\Phi^{-1}(1 - \frac{q_n}{2p})(1 + o_{P^*}(1)) + o_{P^*}\left(\sqrt{\frac{\log p}{nT}}\right)\right]\|\widehat{\Psi}^y\Delta_\gamma^*\|_1 \\
&\quad\le \frac{(c_0+1)}{\sqrt{nT}}\Phi^{-1}\left(1 - \frac{q_n}{2p}\right)\|\widehat{\Psi}^y\Delta_\gamma^*\|_1 = \frac{c_0+1}{2c_0}\kappa_n\|\widehat{\Psi}^y\Delta_\gamma^*\|_1
\end{aligned}$$

with P^* approaching one. Equation (B.4) then implies, for the support set $\widehat{J}$ of $\widehat{\gamma}_y$,

$$\frac{1}{nT}\sum_{t=1}^{T}\|\widehat{U}_t^*\Delta_\gamma^*\|_2^2 + \frac{c_0-1}{2c_0}\kappa_n\|(\widehat{\Psi}^y\Delta_\gamma^*)_{\widehat{J}^c}\|_1 \le \kappa_n\|(\widehat{\Psi}^y\Delta_\gamma^*)_{\widehat{J}}\|_1\frac{3c_0+1}{2c_0}. \tag{B.5}$$

Hence, $\|(\Delta_\gamma^*)_{\widehat{J}^c}\|_1 \le c\|(\Delta_\gamma^*)_{\widehat{J}}\|_1$ for some $c > 0$. This also implies for some generic $C > 0$, $\|\Delta_\gamma^*\|_1^2 \le C\|\Delta_{\gamma,\widehat{J}}\|_1^2 \le C\|\Delta_\gamma^*\|_2^2 O_P(|J|_0)$ as $|\widehat{J}|_0 = O_P(|J|_0)$. (by Proposition F.1 to be proved in Appendix F, $|\widehat{J}|_0 = O_P(|J|_0)$).

We can now apply Lemma H.17 to obtain

$$\frac{1}{nT}\sum_{t=1}^{T}\|\widehat{U}_t^*\Delta_\gamma^*\|_2^2 \ge \frac{1}{nT}\sum_{t=1}^{T}\|\tilde{U}_t\Delta_\gamma^*\|_2^2 - \|\Delta_\gamma^*\|_2^2 o_{P^*}(1).$$

In addition, the sparse-eigenvalue condition implies the restricted eigenvalue condition: For any $m > 0$, there is $\underline{\phi} > 0$ so that, for $|\widehat{J}|_0 = O_P(|J|_0)$, with probability arbitrarily close to one,

$$\inf_{\delta \in \mathbb{R}^p : \|\delta_{\widehat{J}}\|_1 \le m \|\delta_{\widehat{J}^c}\|_1} \frac{\delta' \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \tilde{U}_{it} \tilde{U}_{it}' \delta}{\delta' \delta} \ge \underline{\phi}.$$

Hence $\frac{1}{nT} \sum_{t=1}^T \|\widehat{U}_t^* \Delta_\gamma^*\|_2^2 \ge \underline{\phi} \|\Delta_\gamma^*\|_2^2/2$. Then (B.5) and $\max_m |\widehat{\Psi}_m^y| = O_P(1)$ (where $\widehat{\Psi}^y$ is a diagonal matrix with $\widehat{\Psi}_m^y$ as its m th diagonal entry) imply, for some $C > 0$,

$$\|\Delta_\gamma^*\|_2^2 \le \kappa_n C \|(\Delta_\gamma^*)_{\widehat{J}}\|_1 \le \kappa_n C \|(\Delta_\gamma^*)_{\widehat{J}}\|_2 \sqrt{|\widehat{J}|_0}.$$

Together with $|\widehat{J}|_0 = O_P(|J|_0)$, we have $\|\Delta_\gamma^*\|_2 \le O_P(\kappa_n \sqrt{|J|_0})$ and the previously proved inequality implies $\|\Delta_\gamma^*\|_1^2 \le C \|\Delta_\gamma^*\|_2^2 O_P(|J|_0) = O_P(\kappa_n^2 |J|_0^2)$. Still by (B.5), $\frac{1}{nT} \sum_{t=1}^T \|\widehat{U}_t^* \Delta_\gamma^*\|_2^2 \le C \kappa_n \|\Delta_\gamma^*\|_1$. Thus,

$$\frac{1}{nT} \sum_{t=1}^T \|\widehat{U}_t^* \Delta_\gamma^*\|_2^2 = O_{P^*}(\kappa_n^2 |J|_0), \quad \|\Delta_\gamma^*\|_1 = O_{P^*}(\kappa_n |J|_0). \tag{B.6}$$

■

B.2. The Computational Error of the k-Step Lasso

The main result in this subsection is the following proposition.

PROPOSITION B.2. (*i*) $\|\tilde{\gamma}_y^* - \tilde{\gamma}_{y,lasso}^*\|_1 \le c \|\widehat{\gamma}_y - \tilde{\gamma}_{y,lasso}^*\|_1 + O_{P^*}\left(\frac{a_n}{\kappa_n} + \sqrt{a_n |J|_0}\right)$ *for some* $c > 0$.
(*ii*) $\|\widehat{\gamma}_y - \tilde{\gamma}_y^*\|_1 = O_{P^*}\left(\kappa_n |J|_0 + \frac{a_n}{\kappa_n} + \sqrt{a_n |J|_0}\right)$,
(*iii*) $\frac{1}{nT} \sum_{t=1}^T \|\widehat{U}_t^* (\widehat{\gamma}_y - \tilde{\gamma}_y^*)\|_2^2 = O_{P^*}(\kappa_n^2 |J|_0 + a_n)$.

Proof. (i) We apply Lemma B.1 below. Note that condition (B.7) in this lemma is satisfied under Assumption 4.1 with $b_n = O_{P^*}(a_n)$. Hence applying Lemma B.1 with $\gamma = \tilde{\gamma}_y^*$ immediately implies the result.

(ii) The conclusion follows immediately from part (i) of this proposition and Proposition B.1.

(iii) By equation (B.8) given below in the proof of Lemma B.1 with $b_n = O_{P^*}(a_n)$,

$$\frac{2}{nT} \sum_{it} \|\widehat{U}_{it}^* (\tilde{\gamma}_y^* - \tilde{\gamma}_{y,lasso}^*)\|_2^2 = O_{P^*}(a_n).$$

Hence by (B.6), $\frac{1}{nT} \sum_{t=1}^T \|\widehat{U}_t^* (\widehat{\gamma}_y - \tilde{\gamma}_y^*)\|_2^2 \le O_{P^*}(\kappa_n^2 |J|_0 + a_n)$. ■

Below, $\mathcal{L}^*$ denotes either $\mathcal{L}_y^*$ or $\mathcal{L}_d^*$. Correspondingly, $\widehat{\Psi}$ denotes either $\widehat{\Psi}^y$ or $\widehat{\Psi}^d$, and $\tilde{\gamma}_{lasso}^*$ denotes either $\tilde{\gamma}_{y,lasso}^*$ or $\tilde{\gamma}_{d,lasso}^*$.

LEMMA B.1. *For each γ, suppose for some b_n (either stochastic or deterministic),*

$$\mathcal{L}^*(\gamma)+\kappa_n\|\widehat{\Psi}\gamma\|_1 \le \mathcal{L}^*(\tilde{\gamma}^*_{lasso})+\kappa_n\|\widehat{\Psi}\tilde{\gamma}^*_{lasso}\|_1+b_n. \tag{B.7}$$

Then

$$\|\gamma-\tilde{\gamma}^*_{lasso}\|_1 \le C\|(\widehat{\gamma}_y-\tilde{\gamma}^*_{y,lasso})_{\widehat{J}}\|_1+\frac{b_n}{\kappa_n}+O_{P^*}(\sqrt{b_n|J|_0}),$$
$$\|\gamma-\tilde{\gamma}^*_{lasso}\|_2 \le b_n^{1/2}+o_{P^*}\left(|J|_0^{-1/2}\right)\left(C\|(\widehat{\gamma}_y-\tilde{\gamma}^*_{y,lasso})_{\widehat{J}}\|_1+\frac{b_n}{\kappa_n}\right).$$

Proof. We prove for $\mathcal{L}^*=\mathcal{L}^*_y$. The case with $\mathcal{L}^*=\mathcal{L}^*_d$ follows by the same argument.

Step 1: Show $\|\Delta\|_2^2 \le O_{P^*}(b_n)+\|\Delta\|_1^2 o_{P^*}(|J|_0^{-1})$. Here $\Delta=\gamma-\tilde{\gamma}^*_{y,lasso}$.

Since $\mathcal{L}^*_y(\gamma)$ is quadratic, for any γ_1,γ_2,

$$\mathcal{L}^*_y(\gamma_1)-\mathcal{L}^*_y(\gamma_2)=(\gamma_1-\gamma_2)'\nabla\mathcal{L}^*_y(\gamma_2)+(\gamma_1-\gamma_2)'\nabla^2\mathcal{L}^*_y(\gamma_2)(\gamma_1-\gamma_2),$$

where

$$\nabla\mathcal{L}^*_y(\gamma_2)=-\frac{2}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{U}^*_{it}(\tilde{y}^*_{it}-\widehat{\delta}^{*\prime}_{yt}\widehat{f}^*_i-\widehat{U}^{*\prime}_{it}\gamma_2),$$
$$\nabla^2\mathcal{L}^*_y(\gamma_2)=\frac{2}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{U}^*_{it}\widehat{U}^{*\prime}_{it}.$$

Now let $\gamma_1=\gamma$, and $\gamma_2=\tilde{\gamma}^*_{y,lasso}$. Condition (B.7) then implies

$$\begin{aligned}\Delta'\frac{2}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{U}^*_{it}\widehat{U}^{*\prime}_{it}\Delta &\le b_n+\kappa_n\|\widehat{\Psi}^y\tilde{\gamma}^*_{y,lasso}\|_1-\kappa_n\|\widehat{\Psi}^y\gamma\|_1-\Delta'\nabla\mathcal{L}^*_y(\tilde{\gamma}^*_{y,lasso})\\ &\le b_n\end{aligned} \tag{B.8}$$

where, to establish the last inequality, we used $\kappa_n\|\widehat{\Psi}^y\tilde{\gamma}^*_{y,lasso}\|_1-\kappa_n\|\widehat{\Psi}^y\gamma\|_1-\Delta'\nabla\mathcal{L}^*_y(\tilde{\gamma}^*_{y,lasso})\le 0$ which follows due to the first order condition of (B.1) and the convexity of $\|.\|_1$. (See the proof of Lemma 11 of Agarwal, Negahban, and Wainwright et al. (2012).)

We now establish a lower bound for the left hand side of (B.8).

$$\begin{aligned}\Delta'\frac{2}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{U}^*_{it}\widehat{U}^{*\prime}_{it}\Delta &= \frac{2}{nT}\sum_{t=1}^{T}\|\widehat{U}^*_t\Delta\|_2^2\\ &\ge^{(a)} \frac{2}{nT}\sum_{t=1}^{T}\|\tilde{U}_t\Delta\|_2^2-\|\Delta\|_1^2 o_{P^*}(|J|_0^{-1})\\ &\ge^{(b)} c\|\Delta\|_2^2-\|\Delta\|_1^2 o_{P^*}(|J|_0^{-1}),\end{aligned}$$

where (a) follows from the inequality (a.1) of (H.48) and (b) follows from Assumption 4.2. Substituting this lower bound in for the left-hand-side of (B.8) then yields

$$\|\Delta\|_2^2 \le b_n+\|\Delta\|_1^2 o_{P^*}(|J|_0^{-1}). \tag{B.9}$$

Step 2: Show $\|\Delta\|_1 \leq b_n/\kappa_n + \|(\widehat{\gamma}_y - \widehat{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \|\Delta\|_2 O_P(\sqrt{|J|_0})$. Recall $\Delta = \gamma - \tilde{\gamma}^*_{y,lasso}$. We revisit the proof of Proposition B.1. Note that (B.5) implies, for some $c > 0$,

$$\kappa_n \|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}^c\|_1 \leq \kappa_n \|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 c.$$

The same argument also applies using γ in place of $\tilde{\gamma}^*_{y,lasso}$ due to Condition (B.7), yielding

$$\kappa_n \|(\widehat{\gamma}_y - \gamma)\widehat{J}^c\|_1 \leq \kappa_n \|(\widehat{\gamma}_y - \gamma)\widehat{J}\|_1 c + b_n.$$

Adding these two inequalities and using the triangle inequality, we have

$$\begin{aligned}
\|(\Delta)\widehat{J}^c\|_1 = \|(\gamma - \tilde{\gamma}^*_{y,lasso})\widehat{J}^c\|_1 &\leq \|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}^c\|_1 + \|(\widehat{\gamma}_y - \gamma)\widehat{J}^c\|_1 \\
&\leq \|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 c + \|(\widehat{\gamma}_y - \gamma)\widehat{J}\|_1 c + \frac{b_n}{\kappa_n} \\
&\leq 2c\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \|(\Delta)\widehat{J}\|_1 c + \frac{b_n}{\kappa_n} \\
&\leq 2c\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \|(\Delta)\widehat{J}\|_2 c\sqrt{|\widehat{J}|_0} + \frac{b_n}{\kappa_n}.
\end{aligned}$$

Note that the first inequality in the above follows from the triangle inequality. We then obtain

$$\begin{aligned}
\|\Delta\|_1 &\leq \|(\Delta)\widehat{J}^c\|_1 + \|(\Delta)\widehat{J}\|_1 \\
&\leq 2c\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \|\Delta\|_2 O_{P^*}(\sqrt{|J|_0}) + \frac{b_n}{\kappa_n}.
\end{aligned} \tag{B.10}$$

Step 3: Complete the proof. Substituting (B.10) in for the right-hand-side of (B.9) gives

$$\|\Delta\|_2^2 \leq b_n + o_{P^*}(|J|_0^{-1})\left(C\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \frac{b_n}{\kappa_n}\right)^2 + \|\Delta\|_2^2 o_{P^*}(1),$$

yielding $\|\Delta\|_2^2 \leq b_n + o_{P^*}(|J|_0^{-1})\left(C\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \frac{b_n}{\kappa_n}\right)^2$, and thus we reach the second result of Lemma B.1:

$$\|\Delta\|_2 \leq b_n^{1/2} + o_{P^*}(|J|_0^{-1/2})\left(C\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \frac{b_n}{\kappa_n}\right).$$

Substituting this bound back in for $\|\Delta\|_2$ in (B.10) then yields the first result of Lemma B.1:

$$\|\Delta\|_1 \leq C\|(\widehat{\gamma}_y - \tilde{\gamma}^*_{y,lasso})\widehat{J}\|_1 + \frac{b_n}{\kappa_n} + O_{P^*}(\sqrt{b_n|J|_0}). \tag{B.11}$$

■

B.3. Verifying Assumption 4.1

We now prove Proposition 4.1, which states that the shooting method of Fu (1998) satisfies Assumption 4.1.

We make use of the following lemma in proving Lemma B.3. Below, $\mathcal{L}^*$ denotes either $\mathcal{L}_y^*$ or $\mathcal{L}_d^*$. Correspondingly, $\widehat{\Psi}$ denotes either $\widehat{\Psi}^y$ or $\widehat{\Psi}^d$, and $\tilde{\gamma}_{lasso}^*$ denotes either $\tilde{\gamma}_{y,lasso}^*$ or $\tilde{\gamma}_{d,lasso}^*$.

LEMMA B.2. *Recall that $\widehat{\gamma}$ denotes the post-lasso estimator using the original data and γ_{lasso}^* denotes the lasso estimator using the bootstrap data. We have that*

$$0 \leq \mathcal{L}^*(\widehat{\gamma}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}\|_1 - (\mathcal{L}^*(\widehat{\gamma}_{lasso}^*) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}_{lasso}^*\|_1) = O_{P^*}(\kappa_n^2 |J|_0). \tag{B.12}$$

Proof. The first inequality follows from the definition of $\widehat{\gamma}_{lasso}^*$.

We now show the equality. Note that for each γ,

$$\mathcal{L}^*(\gamma) = \frac{1}{nT}\sum_{t=1}^{T}\left(\|\widehat{U}_t^*(\widehat{\gamma}-\gamma)\|_2^2 + \|M_t^* + \tilde{e}_t^*\|_2^2 + 2(\tilde{e}_t^{*'} + M_t^{*'})\widehat{U}_t^*(\widehat{\gamma}-\gamma)\right) + \kappa_n\|\widehat{\Psi}\gamma\|_1,$$

where M_t^* and $\tilde{e}_t^*$ are defined in the proof of Proposition B.1. Hence by Proposition B.1 and Lemma H.17,

$$\begin{aligned}
&\mathcal{L}^*(\widehat{\gamma}) + \kappa_n\|\widehat{\Psi}\widehat{\gamma}\|_1 - (\mathcal{L}^*(\widehat{\gamma}_{lasso}^*) + \kappa_n\|\widehat{\Psi}\widehat{\gamma}_{lasso}^*\|_1) \\
&\quad = \kappa_n\|\widehat{\Psi}^y\widehat{\gamma}\|_1 - \kappa_n\|\widehat{\Psi}\widehat{\gamma}_{lasso}^*\|_1 - \frac{1}{nT}\sum_{t=1}^{T}\left(\|\widehat{U}_t^*(\widehat{\gamma}-\widehat{\gamma}_{lasso}^*)\|_2^2 + 2(\tilde{e}_t^{*'} + M_t^{*'})\widehat{U}_t^*(\widehat{\gamma}-\widehat{\gamma}_{lasso}^*)\right) \\
&\quad \leq \kappa_n\|\widehat{\Psi}^y(\widehat{\gamma}-\widehat{\gamma}_{lasso}^*)\|_1 + \|\frac{2}{nT}\sum_{t=1}^{T}(\tilde{e}_t^{*'} + M_t^{*'})\widehat{U}_t^*\|_\infty\|\widehat{\gamma}-\widehat{\gamma}_{lasso}^*\|_1 \\
&\quad \leq O_{P^*}(\kappa_n)\|\widehat{\gamma}-\widehat{\gamma}_{lasso}^*\|_1 = O_{P^*}(\kappa_n^2|J|_0).
\end{aligned}$$

■

Below, $\tilde{\gamma}^*$ denotes the k-step lasso estimator based on the bootstrap data, which equals either $\tilde{\gamma}_y^*$ (for the y equation) or $\tilde{\gamma}_d^*$ (for the d equation).

LEMMA B.3. *For the shooting method, (i) $\mathcal{L}^*(\gamma_l) + \kappa_n\|\widehat{\Psi}\gamma_l\|_1 \leq \mathcal{L}^*(\gamma_{l-1}) + \kappa_n\|\widehat{\Psi}\gamma_{l-1}\|_1$.*
(ii) $\mathcal{L}^*(\tilde{\gamma}^*) + \kappa_n\|\widehat{\Psi}\tilde{\gamma}^*\|_1 \leq \mathcal{L}^*(\widehat{\gamma}_{lasso}^*) + \kappa_n\|\widehat{\Psi}\widehat{\gamma}_{lasso}^*\|_1 + O_{P^*}(\kappa_n^2|J|_0)$.
(iii) $|\widehat{J}^*|_0 = O_{P^*}(|J|_0)$.

Proof. Write $\gamma_l = (\gamma_{l,1}, \ldots, \gamma_{l,p})'$, where γ_l can be either $\gamma_{d,l}$ or $\gamma_{y,l}$, to denote the solution after the l^{th} iteration. Note that $\gamma_k = \tilde{\gamma}^*$ is the k-step lasso estimator.

(i) For the shooting method, each $\gamma_{l,m}$ for $m \leq p$ is defined as

$$\gamma_{l,m} = \arg\min_g \frac{1}{nT}\sum_{i,t}(\tilde{y}_{it}^* - \widehat{\delta}_{yt}^{*'}\widehat{f}_i^* - \widehat{U}_{it,m^-}^{*'}\gamma_{l,m^-} - \widehat{U}_{it,m^+}^{*'}\gamma_{l-1,m^+} - \widehat{U}_{it,m}^* g)^2 + \kappa_n|\widehat{\Psi}_m g|.$$

As discussed in Section 4.1, after the m^{th} element is updated in the l^{th} iteration, the vector becomes $\gamma_l^{(m)} := (\gamma_{l,m^-}, \gamma_{l,m}, \gamma_{l-1,m^+})'$, where $m^- = \{j : j < m\}$, $m^+ = \{j :$

$j > m\}$, γ_{l,m^-} represents the subvector of γ_l whose indices are in m^-, and γ_{l-1,m^+} represents the subvector of γ_{l-1} whose indices are in m^+. With this notation, after the $(m-1)^{\text{th}}$ element is updated in the l^{th} iteration, the current solution vector is $\gamma_l^{(m-1)} = (\gamma_{l,(m-1)^-}, \gamma_{l,m-1}, \gamma_{l-1,(m-1)^+})'$. This vector can be rearranged as

$$(\gamma_{l,(m-1)^-}, \gamma_{l,m-1}, \gamma_{l-1,(m-1)^+})' = (\gamma_{l,m^-}, \gamma_{l-1,m}, \gamma_{l-1,m^+})'.$$

It can be seen that the loss function is nonincreasing after the m^{th} element is updated:

$$\begin{aligned}
&\mathcal{L}^*(\gamma_l^{(m)}) + \kappa_n \|\widehat{\Psi}\gamma_l^{(m)}\|_1 \\
&\quad = \mathcal{L}^*((\gamma_{l,m^-}, \gamma_{l,m}, \gamma_{l-1,m^+})) \\
&\qquad + \kappa_n \|\widehat{\Psi}_{m^-}\gamma_{l,m^-}\|_1 + \kappa_n |\widehat{\Psi}_m \gamma_{lm}| + \kappa_n \|\widehat{\Psi}_{m^+}\gamma_{l-1,m^+}\|_1 \\
&\quad \le \mathcal{L}^*((\gamma_{l,m^-}, \gamma_{l-1,m}, \gamma_{l-1,m^+})) \\
&\qquad + \kappa_n \|\widehat{\Psi}_{m^-}\gamma_{l,m^-}\|_1 + \kappa_n |\widehat{\Psi}_m \gamma_{l-1,m}| + \kappa_n \|\widehat{\Psi}_{m^+}\gamma_{l-1,m^+}\|_1 \\
&\quad = \mathcal{L}^*((\gamma_{l,(m-1)^-}, \gamma_{l,m-1}, \gamma_{l-1,(m-1)^+})) \\
&\qquad + \kappa_n \|\widehat{\Psi}_{(m-1)^-}\gamma_{l,(m-1)^-}\|_1 + \kappa_n |\widehat{\Psi}_{m-1}\gamma_{lm-1}| + \kappa_n \|\widehat{\Psi}_{(m-1)^+}\gamma_{l-1,(m-1)^+}\|_1 \\
&\quad = \mathcal{L}^*(\gamma_l^{(m-1)}) + \kappa_n \|\widehat{\Psi}\gamma_l^{(m-1)}\|_1.
\end{aligned}$$

Note that $\gamma_l^{(p)} = \gamma_l$. Hence by (B.12) in Lemma B.2,

$$\begin{aligned}
\mathcal{L}^*(\gamma_l) + \kappa_n \|\widehat{\Psi}\gamma_l\|_1 &\le \mathcal{L}^*(\gamma_l^{(1)}) + \kappa_n \|\widehat{\Psi}\gamma_l^{(1)}\|_1 \le \mathcal{L}^*(\gamma_{l-1}^{(p)}) + \kappa_n \|\widehat{\Psi}\gamma_{l-1}^{(p)}\|_1 \\
&= \mathcal{L}^*(\gamma_{l-1}) + \kappa_n \|\widehat{\Psi}\gamma_{l-1}\|_1.
\end{aligned}$$

(ii) From (i), $\mathcal{L}^*(\gamma_k) + \kappa_n \|\widehat{\Psi}\gamma_k\|_1 \le \mathcal{L}^*(\gamma_0) + \kappa_n \|\widehat{\Psi}\gamma_0\|_1 = \mathcal{L}^*(\widehat{\gamma}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}\|_1$. In addition, by (B.12) in Lemma B.2,

$$\mathcal{L}^*(\widehat{\gamma}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}\|_1 - (\mathcal{L}^*(\widehat{\gamma}^*_{lasso}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}^*_{lasso}\|_1) = O_{P^*}(\kappa_n^2 |J|_0)$$

for $\widehat{\gamma}$ and γ^*_{lasso}, respectively, denoting the completed lasso estimator (as opposed to the k-step lasso solution) using the original data and the bootstrap data. Note that $\kappa_n^2 |J|_0 \sqrt{nT} = o(1)$ and $\gamma_k = \tilde{\gamma}^*$,

$$\begin{aligned}
\mathcal{L}^*(\tilde{\gamma}^*) + \kappa_n \|\widehat{\Psi}\tilde{\gamma}^*\|_1 &\le \mathcal{L}^*(\widehat{\gamma}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}\|_1 \\
&\le \mathcal{L}^*(\widehat{\gamma}^*_{lasso}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}^*_{lasso}\|_1 + o_{P^*}((nT)^{-1/2}),
\end{aligned}$$

which verifies Assumption 4.1(i).

(iii) We now focus on the k-step lasso estimator $\gamma_k = \tilde{\gamma}^*$ and let $\gamma_{k,m}$ denote its m^{th} element. By the KKT condition, if $\gamma_{k,m} \ne 0$, then

$$\begin{aligned}
-\kappa_n \widehat{\Psi}_m sgn(\gamma_{k,m}) &= \frac{2}{nT} \sum_{it} \widehat{U}^*_{it,m} (\tilde{y}^*_{it} - \widehat{\delta}^{*\prime}_{yt} \widehat{f}^*_i - \widehat{U}^{*\prime}_{it,m^-} \gamma_{k,m^-} - \widehat{U}^{*\prime}_{it,m^+} \gamma_{k-1,m^+} - \widehat{U}^*_{it,m} \gamma_{k,m}) \\
&= \frac{2}{nT} \sum_{it} \widehat{U}^*_{it,m} (\tilde{y}^*_{it} - \widehat{\delta}^{*\prime}_{yt} \widehat{f}^*_i - \widehat{U}^{*\prime}_{it} \gamma_k^{(m)}) \\
&= \frac{2}{nT} \sum_{it} \widehat{U}^*_{it,m} (M^*_{it} + \tilde{e}^*_{it} + \widehat{U}^{*\prime}_{it} (\widehat{\gamma} - \gamma_k^{(m)})),
\end{aligned} \tag{B.13}$$

where $\gamma_k^{(m)} := (\gamma_{k,m^-}, \gamma_{k,m}, \gamma_{k-1,m^+})'$, and M_{it}^*, $\tilde{e}_{it}^*$ are, respectively, defined in (B.2)–(B.3). Let $\widehat{U}_{it,\widehat{J}^*}^*$ denote the subvector of $\widehat{U}_{it}^*$, consisting of $\{\widehat{U}_{it,m}^* : \gamma_{k,m} \neq 0, m \leq p\} = \{\widehat{U}_{it,m}^* : m \in \widehat{J}^*\}$. Let $\widehat{\Psi}(\widehat{J}^*)$ be a $|\widehat{J}^*|_0 \times 1$ subvector consisting of the diagonal entries of $\widehat{\Psi}$ (which equals either $\widehat{\Psi}^y$ or $\widehat{\Psi}^d$), with elements in $\widehat{J}^*$. Then the vector form of (B.13) is

$$-\kappa_n \widehat{\Psi}(\widehat{J}^*) sgn(\gamma_{k,m} : m \in \widehat{J}^*) = \frac{2}{nT} \sum_{it} \widehat{U}_{it,\widehat{J}^*}^* (M_{it}^* + \tilde{e}_{it}^*) + A,$$

where (without loss of generality, we assume $\{m \leq p : \gamma_{k,m} \neq 0\} = \{1, \ldots, |\widehat{J}^*|_0\}$)

$$A = \begin{pmatrix} \frac{2}{nT} \sum_{it} \widehat{U}_{it,1}^* \widehat{U}_{it}^{*\prime} (\widehat{\gamma} - \gamma_k^{(1)}) \\ \vdots \\ \frac{2}{nT} \sum_{it} \widehat{U}_{it,|\widehat{J}^*|_0}^* \widehat{U}_{it}^{*\prime} (\widehat{\gamma} - \gamma_k^{(|\widehat{J}^*|_0)}) \end{pmatrix} = \begin{pmatrix} \frac{2}{nT} \widehat{U}_1^{*\prime} \widehat{U}^* (\widehat{\gamma} - \gamma_k^{(1)}) \\ \vdots \\ \frac{2}{nT} \widehat{U}_{|\widehat{J}^*|_0}^{*\prime} \widehat{U}^* (\widehat{\gamma} - \gamma_k^{(|\widehat{J}^*|_0)}) \end{pmatrix}$$
$$= \begin{pmatrix} (\widehat{\gamma} - \gamma_k^{(1)})' \frac{2}{nT} \widehat{U}^{*\prime} & 0 \cdots & 0 \\ 0 & \ddots & \\ 0 & \cdots & (\widehat{\gamma} - \gamma_k^{(|\widehat{J}^*|_0)})' \frac{2}{nT} \widehat{U}^{*\prime} \end{pmatrix} \begin{pmatrix} \widehat{U}_1^* \\ \vdots \\ \widehat{U}_{|\widehat{J}^*|_0}^* \end{pmatrix} := \frac{2}{nT} B \widehat{U}_{\widehat{J}^*}.$$

Therefore

$$\kappa_n \|\widehat{\Psi}(\widehat{J}^*)\|_2 \leq \max_j |\frac{2}{nT} \sum_{it} \widehat{U}_{it,j}^* (M_{it}^* + \tilde{e}_{it}^*)| \sqrt{|\widehat{J}^*|_0} + \|\frac{2}{\sqrt{nT}} B\| \|\frac{1}{\sqrt{nT}} \widehat{U}_{\widehat{J}^*}\|. \quad \textbf{(B.14)}$$

Note that here the norm in both $\|\frac{2}{\sqrt{nT}} B\|$ and $\|\frac{1}{\sqrt{nT}} \widehat{U}_{\widehat{J}^*}\|$ is the operator norm and we have used the inequality $\|B\widehat{U}_{\widehat{J}^*}\|_2 \leq \|B\| \|\widehat{U}_{\widehat{J}^*}\|$.

Now by (H.46),

$$\begin{aligned} \frac{1}{nT} \|\widehat{U}_{\widehat{J}^*}^*\|^2 &\leq \frac{2}{nT} \|\tilde{U}_{\widehat{J}^*}^*\|^2 + \frac{2}{nT} \|\widehat{U}_{\widehat{J}^*}^* - \tilde{U}_{\widehat{J}^*}^*\|^2 \\ &\leq 2\phi_{\max}(|\widehat{J}^*|_0) + \frac{2}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{m \in \widehat{J}^*} (\widehat{U}_{it,m}^* - \tilde{U}_{it,m}^*)^2 \\ &\leq 2\phi_{\max}(|\widehat{J}^*|_0) + O_{P^*}(\Delta_F^{*2} + \frac{\log(pT)}{n}) |\widehat{J}^*|_0, \text{ and} \\ \frac{1}{nT} \|B\|^2 &= \max_{m \in \widehat{J}^*} \frac{4}{nT} \|\widehat{U}^* (\widehat{\gamma} - \gamma_k^{(m)})\|^2 \\ &\leq \frac{8}{nT} \|\widehat{U}^* (\widehat{\gamma} - \tilde{\gamma}_{lasso}^*)\|_2^2 + \max_{m \in \widehat{J}^*} \frac{8}{nT} \|\widehat{U}^* (\tilde{\gamma}_{lasso}^* - \gamma_k^{(m)})\|_2^2 \\ &\leq O_{P^*}(\kappa_n^2 |J|_0), \end{aligned} \quad \textbf{(B.15)}$$

where $\frac{1}{nT} \|\widehat{U}^* (\widehat{\gamma} - \tilde{\gamma}_{lasso}^*)\|_2^2 = O_{P^*}(\kappa_n^2 |J|_0)$ follows from (B.6). To show

$$\max_{m \in \widehat{J}^*} \frac{8}{nT} \|\widehat{U}^* (\tilde{\gamma}_{lasso}^* - \gamma_k^{(m)})\|_2^2 = O_{P^*}(\kappa_n^2 |J|_0),$$

we note that part (i) and (B.12) demonstrate

$$\mathcal{L}^*(\gamma_k^{(m)}) + \kappa_n \|\widehat{\Psi}\gamma_k^{(m)}\|_1 \leq \mathcal{L}^*(\gamma_0) + \kappa_n \|\widehat{\Psi}\gamma_0\|_1$$
$$= \mathcal{L}^*(\widehat{\gamma}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}\|_1 \qquad \leq \mathcal{L}^*(\widehat{\gamma}^*_{lasso}) + \kappa_n \|\widehat{\Psi}\widehat{\gamma}^*_{lasso}\|_1 + O_{P^*}(\kappa_n^2 |J|_0).$$

Thus, the same argument as used in equation (B.8) leads to

$$\max_{m \in \widehat{J}^*} \frac{8}{nT} \|\widehat{U}^*(\tilde{\gamma}^*_{lasso} - \gamma_k^{(m)})\|_2^2 \leq O_{P^*}(\kappa_n^2 |J|_0).$$

By Lemma H.17, $\max_j |\frac{2}{nT} \sum_{it} \widehat{U}^*_{it,j}(M^*_{it} + \tilde{e}^*_{it})| = o_{P^*}(\kappa_n)$ and $\kappa_n \|\widehat{\Psi}(\widehat{J}^*)\|_2 \geq c\kappa_n \sqrt{|\widehat{J}^*|_0}$. Hence, (B.14) and (B.15) imply

$$\kappa_n^2 |\widehat{J}^*|_0 \leq (C\phi_{\max}(|\widehat{J}^*|_0) + O_{P^*}\left(\Delta_F^{*2} + \frac{\log(pT)}{n}\right)|\widehat{J}^*|_0)\kappa_n^2 |J|_0.$$

We thus obtain exactly the same inequality as given (F.4). The rest of the proof then follows from the same argument as used to show Proposition F.1(ii). We conclude $|\widehat{J}^*|_0 = O_{P^*}(|J|_0)$ which verifies Assumption 4.1(ii). ■