# LASSO for Stochastic Frontier Models with Many Efficient Firms*

William C. Horrace† Hyunseok Jung‡ Yoonseok Lee§

June 2021

**Abstract**

We apply the adaptive LASSO (Zou, 2006) to select a set of maximally efficient firms in the panel fixed-effect stochastic frontier model. The adaptively weighted $L_1$ penalty for firm-level inefficiencies allows simultaneous estimation of the maximal efficiency and firm-level inefficiency parameters, which results in a faster rate of convergence of the corresponding estimators than the least-squares dummy variable approach. We show that the estimator possesses the oracle property and propose an efficient optimization algorithm, based on coordinate descent, which dramatically reduces computational costs. We apply the method to estimate a set of maximally efficient Indonesian rice farms.

Keywords: Panel Data, Fixed Effect Stochastic Frontier Model, Adaptive LASSO, $L_1$ Regularization, Zero Inefficiency.

*We are grateful to Badi Baltagi, Christopher Parmeter and the participants at the 15th European Workshop on Efficiency and Productivity Analysis, the 28th annual meeting of the Midwest Econometrics Group and the International Association for Applied Econometric 2019 Annual Conference for their valuable comments and suggestions. All errors are our own.

†Department of Economics, Syracuse University, Syracuse, NY, 13244. whorrace@syr.edu

‡Corresponding author: Department of Economics, University of Arkansas, Fayetteville, AR 72701. hj020@uark.edu

§Department of Economics, Syracuse University, Syracuse, NY, 13244. ylee41@syr.edu

# 1 Introduction

Stochastic frontier (SF) models for panel data typically estimate firm-level efficiency from firm fixed-effects and rank them to identify a single firm in the sample as the most efficient firm. That is, SF estimators do not identify efficiency ties in general, yet there may be several firms in the sample tied for most efficient, particularly in competitive markets.

There exist some methodologies to identify multiple efficient firms in the literature, but they are based on two-step procedures and rely on strong parametric assumptions. In the first step, firm-level efficiencies (or equivalent measures) are estimated, and in the second step a separate inference technique or selection criterion is used to determine membership in a subset of most efficient firms. For example, in the parametric SF model of Aigner, Lovell and Schmidt (1977), there have been several papers to construct parametric prediction intervals for the conditional mean efficiency estimates based on Jondrow, Lovell, Materov and Schmidt (JLMS, 1982). Horrace and Schmidt (1996), Simar and Wilson (2009), and Wheat, Greene and Smith (2014) estimate JLMS efficiency and then construct univariate intervals that imply statistical indistinguishability of firms from the largest estimates. Horrace (2005) and Flores-Lagunes, Horrace and Schnier (2007) extend this to multivariate intervals that account for the multiplicity inherent in the ranked estimates, and Horrace and Schmidt (2000) develop multivariate intervals for the fixed-effect SF model of Schmidt and Sickles (1984) for panel data. Despite the semi-parametric nature of the fixed-effect model, these inference techniques still rely on a parametric assumption on the distribution of estimated efficiencies (i.e., that they are normally distributed or asymptotically so). More recently, Kumbhakar, Parmeter and Tsionas (2013) propose a zero inefficiency stochastic frontier (ZISF) model for cross sectional data that produces a subset of firms in the sample that are fully efficient. They estimate the probability of a firm falling into the zero inefficiency regime

using a latent class model, then use the probability to determine efficient firms. However, the ZISF model suffers from the same issues as the aforementioned techniques; it is parametric and it is a two-step procedure.[1]

To address these issues, in this paper we explicitly assume that some fraction of firms in the panel are fully efficient and develop a one-step, semi-parametric procedure for identifying a subset of efficient firms using the adaptive LASSO (Zou, 2006). Specifically, the proposed approach proceeds as the existing least squared dummy variable (LSDV) estimation, but the objective function is augmented with the adaptively weighted shrinkage $L_1$ penalty for the firm-level inefficiencies. The estimation procedure identifies a subset of firm-level inefficiencies as *exactly zero*, which is an interesting feature of our model compared to the conventional LASSO where identification of *non-zero* coefficients is of primary interest. The LASSO has been applied to various selection problems, but our paper is the first to consider its application to the stochastic frontier models for the identification of efficient firms. We also propose an efficient optimization algorithm based on the coordinate descent method, which dramatically reduces the computation cost.

We analyze the asymptotic properties of the proposed estimator for the case $(N, T) \to \infty$, where $N$ is the number of firms and $T$ is the number of time periods in the sample. We allow for time-series dependence and heteroskedasticity in errors and covariates across firms, which is new for the analysis of panel SF models in the literature.[2] Also, in our approach, $N$ can grow faster than $T$, which may be important for our case since the existence of many efficient firms is more reasonable when markets are large and competitive. We show that the proposed estimator consistently identifies a set of true zero inefficiencies when the two groups, efficient and inefficient firms, are well separated and errors and covariates satisfy

[1] Rho and Schmidt (2015) raise an identification issue for this model.

[2] Park, Sickles and Simar (1998) study the asymptotic properties of the LSDV estimators under $i.i.d$ data. In this paper, we derive the rates of convergence of the LSDV estimators under our setup as well.

proper dependence and tail conditions. The LASSO estimator for within-sample maximal efficiency (the maximum of the individual fixed effects) shows $\sqrt{\delta NT}$ consistency, where $\delta$ is the proportion of fully efficient firms in the sample, while the LSDV estimator exhibits $\sqrt{T/(\log N)^2}$ consistency. Also, the LASSO estimator for an individual firm inefficiency achieves $\sqrt{T}$ consistency, which is faster than that of LSDV, which is $\sqrt{T/(\log N)^2}$. Consequently, the LASSO estimators outperform LSDV in many panels, including short panels. This is borne out in our simulation study.

We apply the LASSO to Indonesian rice farm data previously analyzed by Erwidodo (1990), and Horrace and Schmidt (1996, 2000) among others. The LASSO selects a set of maximally efficient rice farms that is comparable to the Gupta subset of Horrace and Schmidt (2000). However, the LASSO does so without multivariate inference on the efficiency estimates and the distributional assumptions that it entails.

The rest of this paper is organized as follows. The next section introduces the model and the adaptive LASSO estimator. Section 3 provides some technical assumptions and derives the oracle property of the estimator. Section 4 discusses tuning parameter selection and optimization algorithm. Section 5 and 6 provide simulation and empirical application results, and section 7 concludes. All the proofs are given in the Appendix.

# 2 LASSO for Identifying Efficient Firms

## 2.1 Production function

We consider the panel SF model with time-invariant technical inefficiency (e.g. Schmidt and Sickles, 1984) given as

$$y_{it} = \alpha_0 + x_{it}'\beta_0 + v_{it} - u_{0,i} \tag{1}$$

for $i = 1, ..., N$ and $t = 1, ..., T$, where $y_{it}$ is the logarithm of scalar output of the $i$th firm in the $t$th period, $\alpha_0$ is a common intercept, $x_{it}$ is the logarithm of a $p \times 1$ input vector, and $\beta_0$ is a $p \times 1$ corresponding parameter vector of marginal effects. The regression equation has two error terms: the first error term $v_{it}$ is a two sided noise with $E[v_{it}|x_{it}, u_{0,i}] = 0$ and the second error term $u_{0,i}$ is time-invariant firm-specific inefficiencies, which can be arbitrarily correlated with $x_{it}$ (i.e., fixed effect). We suppose no cross-sectional dependence, but allow time-series dependence over errors and covariates. Unlike the standard fixed-effect panel regression models, we restrict $u_{0,i} \geq 0$ for all $i$ but we do not impose a distributional assumption on this inefficiency.

Existing studies estimate (1) using the standard least squared dummy variables (LSDV) method. More precisely, we rewrite (1) as

$$y_{it} = \alpha_{0,i} + x_{it}'\beta_0 + v_{it}, \tag{2}$$

where $\alpha_{0,i} = \alpha_0 - u_{0,i}$ is the firm-specific fixed effect. We can consistently estimate $\alpha_{0,i}$ (as $T \to \infty$) and $\beta_0$ (as $N$ or $T \to \infty$) by the standard within estimation, denoting each estimator $\hat{\alpha}_i$ and $\hat{\beta}$, respectively, provided $x_{it}$ does not include any time-invariant variables.[3] We let the frontier parameter estimator as

$$\hat{\alpha} = \max_{1 \leq i \leq N} \hat{\alpha}_i, \tag{3}$$

which can be verified to be consistent for $\alpha_0$ with $(N, T) \to \infty$ under the assumption that the density of $u_{0,i}$ is nonzero in the neighborhood of zero, so $\min_{1 \leq i \leq N} u_{0,i} \to 0$ as $N \to \infty$ with probability approaching to one (w.p.a.1) and consequently $\max_{1 \leq i \leq N} \alpha_{0,i} \to \alpha_0$ as $N \to \infty$

[3]Feng and Horrace (2007) consider the case where $x_{it}$ includes categorical time-invariant variables and propose a within-a-category comparison approach to characterize individual firm efficiency.

(e.g. Greene, 1980; Schmidt and Sickles, 1984). The individual firm inefficiency $u_{0,i}$ is then consistently estimated as

$$\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i.$$

In this case, $\hat{\alpha}$ represents the maximal efficiency in the sample, and we interpret $\hat{u}_i$ as the relative inefficiency to the most efficient firm.

In practice, it is very unlikely that there are ties in the estimates $\hat{u}_i$. For this reason, all the firms have strictly positive $\hat{u}_i$ values except for the most efficient firm in the sample. Therefore, standard approaches have the limitation that they can identify only one (relatively the most) efficient firm, even when there are multiple efficient firms with $u_{0,i} = 0$.

To overcome such a limitation, we instead estimate (1) using the adaptive least absolute shrinkage and selection operator (adaptive LASSO) method, from which we can identify multiple efficient firms (i.e., all the firms with the true $u_{0,i}$ are zero) by shrinking small values of $\hat{u}_i$ toward zero. To this end, we first assume the following sparsity condition. We let $\mathcal{S} = \{i : u_{0,i} = 0\}$ (i.e. the index set of efficient firms) and $|C|$ be the cardinality of a set $C$.

**Assumption 1** $\delta = |\mathcal{S}|/N \to \delta_0 \in (0, 1)$ as $N \to \infty$.

This sparsity assumption implies that $|\mathcal{S}|$ firms are efficient in the sample and the fraction of efficient firms doesn't vanish as $N \to \infty$, which plays an important role in the asymptotic analysis below. Note that the model (1) becomes the standard fixed-effect SF model when $|\mathcal{S}| = 1$ and it becomes the neoclassic production model when $|\mathcal{S}| = N$ (i.e. every firm in the sample is efficient). Although we suppose $p = \dim(\beta_0)$ is fixed in this paper, we can also allow $p$ to increase with $N$ and assume sparsity on $\beta_0$, under which we can identify nonzero elements

of $\beta_0$ as well. However, this result is already well-studied (e.g. Belloni, Chernozhukov, Hansen and Kozbur, 2016; Caner, Han and Lee, 2018), so we focus on shrinkage estimators of $u_{0,i}$ in this paper. Even when we apply LASSO on $\beta_0$, the rest of the discussion holds as $\hat{\beta} - \beta_0 = O_p((NT)^{-1/2})$.

## 2.2 Adaptive LASSO estimation

We let $\hat{\beta}$ be a (first-stage) consistent estimator of $\beta_0$ from (2), such as the standard LSDV estimator:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \tilde{x}_{it}'\beta\right)^2, \tag{4}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$ with $\bar{y}_i = (1/T)\sum_{s=1}^{T} y_{is}$ and similarly for $\tilde{x}_{it}$. After concentrating out $\beta_0$ in (1), the adaptive LASSO estimator for $\theta_0 = (\alpha_0, u_{0,1}, ..., u_{0,N})'$ is then defined as (e.g. Zou, 2006)[4]

$$\begin{aligned} \hat{\theta}(\lambda) &= (\hat{\alpha}(\lambda), \hat{u}_1(\lambda), ..., \hat{u}_N(\lambda))' \\ &= \underset{\alpha, u_1, ..., u_N;\ u_i \geq 0\ \forall i}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}'\hat{\beta} - \alpha + u_i \right)^2 + \lambda \sum_{i=1}^{N} \hat{\pi}_i u_i \right\}, \end{aligned} \tag{5}$$

where $\lambda > 0$ is a tuning parameter. $\{\hat{\pi}_i\}_{i=1}^{N}$ are some data-dependent weights, which are obtained from some consistent initial estimates of $u_{0,i}$. In particular, we let $\hat{\pi}_i = \hat{u}_i^{-\gamma}$ for some $\gamma > 1$, where $\hat{u}_i$ is the LSDV estimator described in the previous section.[5] Unlike the original LASSO by Tibshirani (1996), the adaptive LASSO allows for unequal shrinkage for each parameter depending on the data-dependent weight $\hat{\pi}_i$, which results in the oracle

[4]Our adaptive LASSO is based on a concentrated least square loss function with sign restrictions whereas Zou(2006) is based on a standard least squares.

[5]Note that in the LSDV estimation, the firm with the largest firm fixed-effect estimate has a zero inefficiency estimate. For $\hat{u}_i = 0$, we use an arbitrarily small value (e.g. $1/N$) to construct the weight.

property (see Fan and Li, 2001; Zou, 2006).

One important remark on (5) is that we estimate $\alpha_0$ and $(u_{0,1}, ..., u_{0,N})'$ together in one step. This is not feasible in the standard fixed effect SF model because of the perfect multicollinearity between the constant term and the individual dummies. In contrast, this is feasible in this model due to the sparsity assumption and $L_1$ penalty term, which allows for elimination of some of the individual dummies.

The main goal of this estimation is to identify two groups: efficient firms and inefficient firms. Therefore, this approach seems similar to Bonhomme and Manresa (2015), who also consider a latent group structure problem determined by group-specific fixed effects. However, their methodology relies on minimization of a least squares criterion with respect to all possible groupings, whereas we use the LASSO technique to identify the latent groups (efficient firms vs. inefficient firms) under sign-restrictions on the fixed effects.

The adaptive LASSO problem in (5) is also related to the latent group structure model by Su, Shi and Phillips (2016), or the fused LASSO by Tibshirani, Saunders, Rosset, Zhu and Knight (2005). They penalize over pairwise-differences among the coefficient values and hence produce group identification. From (2), since $\alpha_{0,i} = \alpha_0 - u_{0,i}$ and $\alpha_0 \geq \alpha_{0,i}$ for all $i$, the adaptive LASSO problem in (5) can be rewritten as

$$\underset{\alpha,\alpha_1,...,\alpha_N;\ \alpha \geq \alpha_i\ \forall i}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}^{'} \hat{\beta} - \alpha_i \right)^2 + \lambda \sum_{i=1}^{N} \hat{\pi}_i (\alpha - \alpha_i) \right\},$$

which resembles Su, Shi and Phillips (2016), given $\hat{\beta}$. However, our problem is different from theirs because we impose sign restriction on $u_{0,i} = \alpha_0 - \alpha_{0,i}$.[6]

[6] As we shall see, we allow that the size of the smallest (non-zero) inefficiency shrinks to zero with an appropriate rate condition whereas parameter values are fixed in Su, Shi and Phillips (2016).

# 3 Oracle Properties

Unlike the original LASSO, the adaptive LASSO allows for unequal shrinkage for each parameter depending on the data-dependent weight, which results in the oracle property (i.e. consistent model selection and estimation). Such oracle property extends to our case under $(N,T)$-asymptotics. Particularly, in our approach, $N$ can grow faster than $T$ when errors and covariates satisfy proper dependence and tail conditions. This result is important for our case since the assumption of time-invariant inefficiency and existence of many efficient firms is more reasonable when $N$ (market size) is large and $T$ is small.

We assume the following conditions in our asymptotic analysis. Recall $\hat{\beta}$ is the (first-stage) standard LSDV estimator in (4), and we let $\eta = \min_{i \in \mathcal{S}^c} u_{0,i}$.

**Assumption 2:** (1) (i) $E[v_{it}|x_{it}, u_{0,i}] = 0$ for all $i$ and $t$, and $\{(x_{it}, v_{it}) : t = 1, ..., T\}$ are independent over $i$; (ii) For each $i$, $\{(x_{it}, v_{it}) : t = 1, ..., T\}$ is strong mixing with mixing coefficients $\alpha[t] \leq c_\alpha \rho^t$ for some $c_\alpha > 0$ and $\rho \in (0,1)$; (iii) $\sup_{i \geq 1} \sup_{t \geq 1} E||x_{it}||^q < \infty$ and $\sup_{i \geq 1} \sup_{t \geq 1} E|v_{it}|^q < \infty$ for some $q \geq 4$.

(2) As $(N,T) \to \infty$, (i) $\hat{\beta} - \beta_0 = O_p((NT)^{-1/2})$; (ii) $NT^{1-q/2}(\log T)^{2q} \to 0$; (iii) $\lambda T^{-1/2} N^{1/2} \eta^{-\gamma} \to 0$ and $\lambda T^{(\gamma-1)/2} (\log N)^{-2\gamma-1} \to \infty$ for some $\gamma > 1$.

In Assumption 2-(1), we rule out cross-sectional dependence, but allow for time-series dependence in the errors and covariates. In Assumption 2-(1)-(ii) and (iii), we require $(x_{it}, v_{it})$ be a strong mixing process over $t$ with geometric decay rate, and further restrict the moments of $||x_{it}||$ and $|v_{it}|$ to be finite up to a certain order. The tail restrictions and finite moment condition allow us to use exponential inequalities for strong mixing processes (e.g. Merlevède, Peligrad and Rio, 2009) to bound misclassification probabilities and achieve selection consistency.[7]

[7] Alternatively, exponential moment conditions can be employed as in Bonhomme and Manresa (2015).

Assumption 2-(2)-(i) is the standard for the LSDV estimator under $(N,T) \to \infty$. Assumption 2-(2)-(ii) and (iii) impose rate conditions on $N, T, \eta$ and $\lambda$ to ensure both selection and estimation consistency. Assumption 2-(2)-(ii) implies $N$ can grow faster than $T$ depending on $q$ in the finite moment condition. As $q$ increases (i.e. as the distribution of the error decays faster), $N$ can grow substantially faster than $T$. Therefore it covers many panel structures including short panels. The rate conditions also control for the magnitude of the tuning parameter $\lambda$, so the LASSO procedure can select the zero coefficients correctly without yielding bias in the nonzero coefficient estimators in the limit. The assumption allows the nonzero inefficiencies to be close to zero, but it shrinks sufficiently slow enough to be distinguished from the zero coefficients and also not affected by shrinkage estimation.

Recall that $\mathcal{S} = \{i : u_{0,i} = 0\}$, and we let $\hat{\mathcal{S}} = \{i : \hat{u}_i(\lambda) = 0\}$. We first derive the convergence rates of the LSDV estimators under Assumption 1 and 2, which will primarily serve as a technical lemma to prove some theorems later,[8] but also allows us to compare the convergence rate of $\hat{\alpha}$, the LSDV estimator, with that of $\hat{\alpha}(\lambda)$, the LASSO estimator.

**Lemma** *Recall that $\hat{\alpha}$ is the LSDV estimators for $\alpha_0$ where $\hat{\alpha} = \max_{1 \le i \le N} \hat{\alpha}_i$ and $\hat{\alpha}_i$ is the LSDV estimator for $\alpha_{0,i}$. Then, under Assumption 1 and 2, as $(N,T) \to \infty$,*

$$
\begin{aligned}
&(i) \quad \Pr\left(\hat{\alpha} = \max_{i \in \mathcal{S}} \hat{\alpha}_i\right) \to 1 \\
&(ii) \quad \hat{\alpha} - \alpha_0 = O_p\left((\log N)/T^{1/2}\right).
\end{aligned}
$$

The proofs are in Appendix. The lemma implies $\hat{\alpha}$ is estimated from one of the efficient firms in the sample w.p.a.1, and $\hat{\alpha}$ has a convergence rate of $(\log N)/T^{1/2}$ as $(N,T) \to \infty$. The rate is identical to that derived in Park, Sickles and Simar (1998), but their result is derived under *i.i.d* data with exponential moment conditions imposed on errors and covariates.[9] So

[8] This is because the LASSO estimator uses the LSDV estimators, $\hat{u}_i$, to construct the adaptive weights.

[9] Recall that we impose only finite moment conditions for the errors and covariates and allow for time-

the lemma can be seen as a generalization of their result.[10]

Now we turn to the LASSO results. We first establish the selection consistency of the LASSO procedure.

**Theorem 1** *Suppose Assumptions 1 and 2 hold. Then,* $\Pr(\hat{\mathcal{S}} = \mathcal{S}) \to 1$ *as* $(N, T) \to \infty$.

The theorem implies that the LASSO consistently identifies two latent groups provided the rate conditions on $N, T, \eta$ and $\lambda$ are satisfied. This, in turn, implies that in the limit, the latent groups can be treated as known (i.e. the oracle information) and used for the estimation of $\alpha_0$ and inefficiencies to improve their convergence rates.

We introduce the following assumptions and notations to simplify the limiting distributions of the LASSO estimators.

**Assumption 3** (i) There exist positive constants $\sigma^2_{\mathcal{S}_1}$, $\sigma^2_{\mathcal{S}_2}$, $\sigma_{\mathcal{S}_1\mathcal{S}_2}$, $\sigma^2_{\mathcal{S}^c}$ and $\sigma^2_i$ for each $i \in \mathcal{S}^c$ such that

$$\begin{aligned}
\sigma^2_{\mathcal{S}_1} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{\delta NT} \sum_{i\in\mathcal{S}} \sum_{t=1}^{T} \sum_{k=1}^{T} v_{it} v_{ik} \\
\sigma^2_{\mathcal{S}_2} &= \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \operatorname*{plim}_{N,T\to\infty} \frac{1}{\delta NT} \sum_{i\in\mathcal{S}} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \tilde{x}'_{it} \right\} H_0^{-1} \Upsilon_{\mathcal{S}} \\
\sigma_{\mathcal{S}_1\mathcal{S}_2} &= \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \operatorname*{plim}_{N,T\to\infty} \frac{1}{\delta NT} \sum_{i\in\mathcal{S}} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \right\} \\
\sigma^2_{\mathcal{S}^c} &= \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \operatorname*{plim}_{N,T\to\infty} \frac{1}{(1-\delta) NT} \sum_{i\in\mathcal{S}^c} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \tilde{x}'_{it} \right\} H_0^{-1} \Upsilon_{\mathcal{S}} \\
\sigma^2_i &= \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{T} v_{it} v_{ik}.
\end{aligned}$$

series dependence.

[10] A direct consequence of (ii) of the lemma is $\hat{u}_i - u_{0,i} = O_p\left((\log N)/T^{1/2}\right)$ for all $i$. See the proof of this lemma in Appendix for more detail.

where $\Upsilon_{\mathcal{S}} = \underset{N,T\to\infty}{\text{plim}} \frac{1}{\delta NT}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T} x_{it}$, and $H_0 = \underset{N,T\to\infty}{\text{plim}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}' > 0$; (ii) As $(N,T)\to\infty$, $\frac{1}{\sqrt{\delta NT}}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T} v_{it} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{\mathcal{S}_1})$, $\frac{1}{\sqrt{\delta NT}}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T}\Upsilon_{\mathcal{S}}' H_0^{-1}\tilde{x}_{it}v_{it} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{\mathcal{S}_2})$, $\frac{1}{\sqrt{(1-\delta)NT}}\sum_{i\in\mathcal{S}^c}\sum_{t=1}^{T}\Upsilon_{\mathcal{S}}' H_0^{-1}\tilde{x}_{it}v_{it} \xrightarrow{d} \mathcal{N}(0,\sigma^2_{\mathcal{S}^c})$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it} \xrightarrow{d} \mathcal{N}(0,\sigma_i^2)$ for each $i\in\mathcal{S}^c$.

**Theorem 2** *Suppose Assumptions 1, 2 and 3 hold. Then, as* $(N,T)\to\infty$,

$$(i)\quad \sqrt{\delta NT}(\hat{\alpha}(\lambda)-\alpha_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{S}_1} + \delta^2\sigma^2_{\mathcal{S}_2} - 2\delta\sigma^2_{\mathcal{S}_1\mathcal{S}_2} + \delta(1-\delta)\sigma^2_{\mathcal{S}^c}\right)$$

$$(ii)\quad \sqrt{T}(\hat{u}_i(\lambda)-u_{0,i}) \xrightarrow{d} \mathcal{N}(0,\sigma_i^2) \text{ for each } i\in\mathcal{S}^c.$$

This theorem combined with Theorem 1 establishes the selection consistency and asymptotic normality of the adaptive LASSO estimators. In the limit, we can identify a set of efficient firms without loss of estimation efficiency in this procedure. It is worthy to note that $\hat{\alpha}(\lambda)$ shows a faster convergence rate than the LSDV estimator, $\hat{\alpha}$. This is because the LSDV estimator uses only a single best firm's observations, but $\hat{\alpha}(\lambda)$ uses $|\hat{\mathcal{S}}|\cdot T$ observations of the firms identified as efficient by the LASSO. As long as $\delta$ doesn't vanish as $N\to\infty$, which is a reasonable assumption for competitive markets, the LASSO estimator will be preferred.

# 4 Computation

## 4.1 Optimization algorithm

The $L_1$ penalty term in the LASSO object function has no second derivative at the origin, so we can't directly apply the standard quadratic optimization algorithms (e.g. Newton-Raphson). Many alternative optimization algorithms have been developed: local quadratic approximation (Fan and Li, 2001), least angle regression (Efron, Hastie, Johnstone and

Tibshirani, 2004), coordinate descent algorithm (Friedman, Hastie and Tibshirani, 2010), among others. For our problem, however, such conventional computation methods may not be the optimal, which is because we impose sign restrictions.

When we conduct optimization with a small number of sign restrictions, we can use sequential quadratic programming algorithm in general. However, when the number of sign restrictions is very large, like $u_{0,i} \geq 0$ for all $i = 1, ..., N$ in our case, its computation cost becomes very high or the computation can be even infeasible. For this reason, we propose an alternative algorithm based on the coordinate descent method, which significantly reduces the computational cost. Using preliminary inefficiency ranking information among the firms from the initial LSDV estimation, this algorithm allows us to skip a large number of irrelevant optimization steps. The algorithm is summarized as follows.

1. Using $\hat{\beta}$ from the initial estimation, let

$$\hat{\alpha}_i^{(0)} = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}'\hat{\beta}), \quad \hat{\alpha}^{(0)} = \max_{1\leq i\leq N} \hat{\alpha}_i^{(0)}, \quad \text{and} \quad \hat{u}_i^{(0)} = \hat{\alpha}^{(0)} - \hat{\alpha}_i^{(0)}$$

   for each $i$. Define order statistics $\hat{\alpha}_{[1]}^{(0)} \leq \hat{\alpha}_{[2]}^{(0)} \leq ... \leq \hat{\alpha}_{[N]}^{(0)}$ and $\hat{u}_{[1]}^{(0)} \geq \hat{u}_{[2]}^{(0)} \geq ... \geq \hat{u}_{[N]}^{(0)}$ so that $\hat{\alpha}_{[N]}^{(0)} = \max_{1\leq i\leq N} \hat{\alpha}_i^{(0)}$ and $\hat{u}_{[N]}^{(0)} = \min_{1\leq i\leq N} \hat{u}_i^{(0)}$. In this step, we have only one fully efficient firm with $\hat{u}_{[N]} = \hat{u}_{[N]}^{(0)} = 0$.

2. For a given $\lambda$, check the Karush-Kuhn-Tucker (KKT) condition[11] for the second best firm based on the sign of

$$\Delta_{[N-1]} = \hat{u}_{[N-1]}^{(0)} - \frac{\lambda}{2T}\hat{\pi}_{[N-1]}.$$

   In particular, if $\Delta_{[N-1]} \leq 0$, let $\hat{u}_{[N-1]}^{(1)} = 0$ and update $\hat{\alpha}^{(0)}$ as $\hat{\alpha}^{(1)} = (\hat{\alpha}_{[N]}^{(0)} + \hat{\alpha}_{[N-1]}^{(0)})/2$.

[11] See the proof of Theorem 1 in Appendix for more detail.

Using this new frontier parameter estimate $\hat{\alpha}^{(1)}$, update the rest of the inefficiencies as $\hat{u}^{(1)}_{[N-1-j]} = \hat{u}^{(0)}_{[N-1-j]} - (\hat{\alpha}^{(0)} - \hat{\alpha}^{(1)})$ for all $j \leq N-2$. If $\Delta_{[N-1]} > 0$, go to the Step 4 below.

3. Sequentially repeat Step 2 for each $\Delta_{[N-k]}$ for $k = 2, 3, ..., N-1$ as long as $\Delta_{[N-k]} \leq 0$ holds. For each $k$, we let $\hat{u}^{(k)}_{[N-k]} = 0$ and update $\hat{\alpha}^{(k-1)}$ as $\hat{\alpha}^{(k)} = (1/(k+1)) \sum_{j=0}^{k} \hat{\alpha}^{(k-1)}_{[N-j]}$. We also update $\hat{u}^{(k)}_{[N-1-j]} = \hat{u}^{(k-1)}_{[N-1-j]} - (\hat{\alpha}^{(k-1)} - \hat{\alpha}^{(k)})$ for all $k \leq j \leq N-2$.

4. If $\Delta_{[N-k]} > 0$ at some $k \geq 1$, we update the non-zero inefficiencies (i.e., $\hat{u}_{[N-j]} > 0$ for $k \leq j \leq N-1$) as $\hat{u}^{k}_{[N-j]} = \hat{u}^{(k-1)}_{[N-j]} - \Pi \frac{\hat{\pi}_k}{2T}$ for all $k \leq j \leq N-1$ and then report the results.[12]

This coordinate descent algorithm uses the convexity of the object function and the preliminary inefficiency ranking at the same time, which enables us to reach the minimum of the object function quickly. As an illustration, we compare the computation time between this algorithm and the standard sequential quadratic programming (SQP) algorithm in Matlab. We find that the two algorithms produce very similar estimation results, but the new algorithm runs much faster than SQP. For instance, in the simulation with a small sample size $(N, T) = (20, 10)$ in Section 5, this algorithm takes 1.5 seconds for each replication on average whereas the SQP algorithm takes 345 seconds. This gap gets pronounced as $N$ increases.

[12]In fact, minimizing the objective function (5) induces shrinkage effect not only on $\hat{u}_i(\lambda)$ but also on $\alpha(\lambda)$. Technically, the algorithm should include additional steps to reflect such shrinkage effect on $\alpha(\lambda)$. However, this is a undesirable shrinkage bias, which may slow down the convergence of $\alpha(\lambda)$, particularly when $N$ is large (Equation A.6 in Appendix includes the explicit form of this bias). Therefore, in the spirit of post-LASSO estimation (e.g. Belloni and Chernozhukov, 2013), in this algorithm, we skip the additional steps to achieve smaller finite sample bias. This omission doesn't alter any of the asymptotic results in Section 3. We use this algorithm for our simulations and empirical application.

### 4.2 Tuning parameter choice

The performance of the adaptive LASSO estimator relies on an appropriate choice of the tuning parameter, $\lambda$. Methods based on cross validations and AIC criteria are known to result in over-selection (i.e., too many nonzero estimates), which will result in under-selection of the efficient firms in this context. Wang, Li and Tsai (2007) instead propose tuning parameter choice based on the BIC-type criterion, which is shown to consistently estimate the correct model when it exists.

We also consider a BIC-type criterion in choosing $\lambda$, which is given by

$$\lambda^* = \arg\min_{\lambda} \log \hat{\sigma}^2(\lambda) + \frac{\log T}{NT}|\hat{\mathcal{S}}^c|, \tag{6}$$

where $\hat{\mathcal{S}}^c = \{i : \hat{u}_i(\lambda) > 0\}$ and

$$\hat{\sigma}^2(\lambda) = \frac{1}{nT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}^{'}\hat{\beta} - \hat{\theta}(\lambda)\right)^2$$

from (5) for a fixed $\lambda$. We also experimented other types of selection criteria in the simulation study, including ERIC by Hui, Warton and Foster (2015) and $IC_{p1}$ by Bai and Ng (2002), and find that (6) performs the best in this panel SF model. The latter two criteria yield similar results with tending to select more sparse models than (6).

## 5 Simulations

In this section, we study the finite sample performance of the LASSO estimator. We consider the model (1) with $\alpha_0 = 1$, $x_{it} \sim N(0, \Sigma)$, where the $(j_1, j_2)$th element of $\Sigma$ is $0.5^{|j_1 - j_2|}$ for $j_1, j_2 = 1, ..., 8$, and $v_{it} \sim \mathcal{N}(0, 1)$. We assume 30% of firms in the sample are fully efficient

(i.e., $\delta = 0.3$) and in every simulation each nonzero individual inefficiency is identically and independently generated from an exponential distribution, say $\max\{0.01, (1/\sigma_u)e^{-u_{0,i}/\sigma_u}\}$ for some $\sigma_u > 0$, where trimming is to ensure all draws are strictly positive. We experiment with $\sigma_u \in \{1, 2, 4\}$. Note that as $\sigma_u$ gets smaller, the probability of small inefficiency draws gets higher, making it more difficult for the LASSO to distinguish them from zero. This would be particularly difficult when the sample size is small.[13] Figure 1 shows the density of the inefficiency $u_{0,i}$ for each given $\sigma_u$ value (figure on the left) and an example of draws from each case (figure on the right). We can clearly see that inefficiencies have high density near zero when $\sigma_u = 1$. For the penalty function, we set $\gamma = 2$ and the tuning parameter is selected by (6) from a grid search over 250 evenly spaced points between $10^{-4}$ and $10T$.[14] We simulate each case 1000 times with twelve combinations of $N \in \{100, 200, 1000\}$ and $T \in \{10, 30, 50, 70\}$.

[=== Figure 1 is here ===]

First, Table 1 reports and compares the results from the adaptive LASSO estimation in (5) and the conventional LSDV approach described in Section 2.1. In particular, we report the root mean squared errors (RMSE) of $\hat{U}(\lambda) = (\hat{u}_1(\lambda), ..., \hat{u}_N(\lambda))'$ and $\hat{U} = (\hat{u}_1, ..., \hat{u}_N)'$; point estimates of $\alpha_0$ from $\hat{\alpha}(\lambda)$ and $\hat{\alpha}$; and the sample correlations between the ranking of $U_{0\mathcal{S}^c}$ (i.e. nonzero inefficiencies) and the ranking of their counterpart estimates $\hat{U}_{\mathcal{S}^c}(\lambda)$ and $\hat{U}_{\mathcal{S}^c}$ for given $\mathcal{S}$.[15]

[13] In this case, the rate condition on $\eta$ in Section 3 is likely to be violated.

[14] We are free to choose the value of $\gamma$ as long as it satisfies the rate conditions in Assumption 2. From the asymptotic analysis, we can see that setting a higher value for $\gamma$ ensures the LASSO estimates zero coefficients as zero, but also increases the probability of estimating (small) nonzero coefficients as zero. Therefore, in empirics $\gamma$ should be determined in light of this trade-off. Also the rate condition in Assumption implies the tuning parameter should get smaller as $N$ increases.

[15] More precisely, the entries in Table 1 (and also those in Table 2) are the average values for each measure over 1,000 replications and their corresponding standard deviations in parentheses. Rank correlations are computed only among the inefficiencies whose true values are nonzero, that is $corr(R(U_{0\mathcal{S}^c}), R(\hat{U}(\lambda)_{\mathcal{S}^c}))$ and Similarly for LSDV where $R(\cdot)$ is a mapping from estimates to rankings.

[=== Table 1 here ===]

As $T$ and $\sigma_u$ increase, the RMSE from the LASSO decreases, but the effect of $\sigma_u$ on RMSE is relatively small. Recall that $\sigma_u$ determines the frequency of near zero inefficiencies and hence, as we shall see below, small inefficiency draws tend to be misclassified as zero inefficiency when $\sigma_u$ is small. However, this misclassification doesn't affect the RSME much since small inefficiency draws are already near zero so estimating them as zero doesn't increase the RMSE much.

[=== Figure 2 here ===]

The LASSO outperforms LSDV in terms of RMSE. This is mainly because of more accurate estimation of $\alpha_0$ in the LASSO. Figure 2 presents the distribution of the estimates of $\alpha_0$ from the LASSO (solid line) and LSDV (dashed line). The distributions of the LASSO estimators are centered close to the true value ($\alpha_0 = 1$) even when $T$ and $\sigma_u$ are small, and the variation in the distribution decreases significantly as $N$ or $T$ increase. However, those from LSDV are consistently displaced away from the true value. This is related to the finding that the LASSO estimator of $\alpha_0$ converges faster than the LSDV estimator as shown in the asymptotic analysis. In addition, the max operator that the LSDV uses to estimate $\alpha_0$ tends to pick the most biased individual intercept estimates. Therefore, in the presence of a group of zero inefficiency firms, the max operator produces a biased estimate for $\alpha_0$, which, in turn, leads to a significant bias in the estimation of the inefficiencies $u_{0,i}$'s.[16]

The LASSO and the LSDV show similar rank correlation results. It appears that the LASSO preserves the original ranking better than LSDV when $T$ and $\sigma_u$ are small. This is when we have a large uncertainty in the inefficiency estimates and the LASSO improves the ranking accuracy by estimating statistically indistinguishable small inefficiencies as zero.

[16] Wang and Schmidt (2009) also document the "upward" bias of LSDV estimates using simulations.

Second, Table 2 presents the selection accuracy of the LASSO estimation. In particular, we report the probability of yielding a nonzero estimate for $i \in \mathcal{S}^c$, $P_{\mathcal{S}^c} = \Pr(i \in \hat{\mathcal{S}}^c | i \in \mathcal{S}^c)$; the probability of yielding a zero estimate for $i \in \mathcal{S}$, $P_{\mathcal{S}} = \Pr(i \in \hat{\mathcal{S}} | i \in \mathcal{S})$; the sample proportion of efficient firms $\hat{\delta}$; and the maximum value of $u_{0,i}$, whose true value is nonzero but estimated as zero, representing the degree of misclassification (i.e., $\max_{i \in \hat{\mathcal{S}}^c | i \in \mathcal{S}} u_{0,i}$; Max-miss). We also depict the selection probabilities in Figure 3 for $N = 100$ with $T \in \{10, 30, 50, 70\}$ and $\sigma_u \in \{1, 2, 4\}$.

[=== Table 2 here ===]

[=== Figure 3 here ===]

We can see that when $T$ and $\sigma_u$ are small, the LASSO incorrectly estimates many of nonzero inefficiencies as zeros. However, the misclassification improves as $T$ or $\sigma_u$ increases. Note that most of the firms incorrectly estimated as efficient firms (i.e., those with zero inefficiency estimates) would have near zero inefficiency. The small values of Max-miss in Table 2 imply that only the firms near zero inefficiency could be incorrectly categorized as fully efficient in the LASSO procedure.

More importantly, it is impressive that even when $T$ is small, including the $(N, T) = (1000, 10)$ case, $P_{\mathcal{S}}$ and $P_{\mathcal{S}^c}$ are close to 1 when $\sigma_u = 4$. This implies that the distribution of firm inefficiency affects more to the selection performance than the size of $T$. This gives us an important implication: our approach can be used even for short panels, as long as there are not too many near zero inefficient firms. Hence, in practice, information on the variance of $u_{0,i}$ would be important in the choice of the proposed LASSO approach. Cai, Horrace and Lee (2021) studies nonparametric identification of $\sigma_u$ in the panel setup, where it is allowed to be conditionally heteroskedastic.

# 6 Empirical Application: Rice Farms in Indonesia

In this section, we apply the adaptive LASSO estimation to the rice farm data previously analyzed by Erwidodo (1990), and Horrace and Schmidt (1996, 2000), among others. In our context, the LASSO is used to select a group of efficient farms. The idea of selecting a subset of best farms is related to the "ranking and selection" approach (R&S, hereafter) in the SF literature (e.g. Horrace and Schmidt, 1996, 2000). The R&S is summarized as follows: We estimate the LSDV model and obtain $\hat{\alpha}_i$ for $i = 1, ..., N$. Then, we select a subset of the population that contains individual(s) with the largest value of $\alpha_i$ based on an inferential decision rule with some pre-specified error rate. If in truth $\alpha_i = \alpha_0 - u_i$, this is equivalent to selecting individuals with $u_i$ closest to zero at the pre-specified error rate. In this sense, the basic idea of R&S is quite similar to that of the LASSO; we compare these two approach in this application.

We use data of 171 rice farms in Indonesia, observed for three wet and three dry seasons from six different villages ($N = 171$, $T = 6$). See Erwidodo (1990) for the complete description of the data. We estimate the standard Cobb-Douglas log-linear panel production function of rice. Output is measured in kilograms of rice. Inputs includes in the weights of the seed (kg), urea (kg), and trisodium phosphate (TSP) (kg); the hours of labor (hours); and the area of land (hectares). The regression model also includes the following dummy variables: $DP = 1$ if pesticides were used; $DV1 = 1$ if high yield varieties of rice were planted; $DV2 = 1$ if mixed varieties were planted;[17] $DSS = 1$ if it was a wet season. Since our focus is on the efficiency estimates, we do not report the first-stage estimation result (i.e., $\hat{\beta}$ in (1)) of the marginal effect here. See Horrace and Schmidt (2000) for these estimates. As in the simulations, we set $\gamma = 2$ and the tuning parameter is chosen from (6). The model is estimated after standardizing the input variables.

[17]The omitted category with $DV1 = DV2 = 0$ represents that traditional varieties were planted.

The LASSO estimates 69.6 % of farms (which is 119 out of 171 farms) as efficient. The distribution of the inefficiencies are reported in Figure 4. In Figure 4, the blue histogram represents the distribution of the inefficiencies from the conventional LSDV approach and the orange one represents that from the LASSO, where the 69.6 % of mass is concentrated at zero.

We also conduct the R&S procedure with error rate of 0.05, and find that 67% of rice farms (which is 115 out of 171 farms) are in the subset of the best farms. Figure 5 matches the firm IDs estimated as fully efficient by the LASSO (yellow) and those included in the best subset by the R&S (red). It shows that all of the farms in the best subset by the R&S are also estimated as efficient by the LASSO. Though the R&S approach gives more parsimonious selection result in this exercise, it could be reversed based on the choice of the tuning parameter of the LASSO and the error rate of the R&S. This result implies the two procedures are closely related and can be alternately used depending on the goals of research. The similarity between the results may be understood by noting that the two methods are both selecting subsets of the best firms after accounting for the impacts of the individual inefficiency estimates on the entire model, which is similar to the F-test-based procedure. However, note that the R&S procedure relies on normal approximation based on a central limit theorem whereas the LASSO achieve the selection consistency without such restriction. Also, the multivariate confidence intervals that R&S uses to determine the best subset are based on the $N(N-1)$ differences, $\hat{\alpha}_i - \hat{\alpha}_j$ for all $i \neq j$, which could be computationally more demanding than the LASSO when $N$ is large.

[=== Figure 4 is here ===]

[=== Figure 5 is here ===]

# 7 Conclusion

We have shown the proposed adaptive LASSO estimator has the oracle property under regularity conditions. Moreover, the finite sample simulations demonstrate that the estimator outperforms the conventional LSDV-based approach in many aspects. The empirical application reveals that our methodology is comparable with the R&S procedure, but it requires less parametric assumptions and computational cost.

The technique developed in this paper is broadly applicable. When we have a panel linear regression model with individual fixed effects, and the ranked fixed effects contain important information, this approach can identify a subset of the best (or worst) effects. For example, consider an education outcome function. After controlling for other factors (i.e. family background and teacher quality), we may want to estimate a group of the best or worst students based on their individual-specific outcomes. Mutual fund performance can be another example. Moreover, this type of "best and the rest" classification can be useful in big-data settings, which can be used as an adaptive sample splitting method.

# References


Aigner, D., Lovell, C. and Schmidt, P. (1977), 'Formulation and estimation of stochastic frontier production function models', *Journal of Econometrics* **6**, 21–37.

Bai, J. and Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Belloni, A. and Chernozhukov, V. (2013), 'Least squares after model selection in high-dimensional sparse models', *Bernoulli* **19**(2), 521–547.

Belloni, A., Chernozhukov, V., Hansen, C. and Kozbur, D. (2016), 'Inference in high-dimensional panel models with an application to gun control', *Journal of Business & Economic Statistics* **34**(4), 590–605.

Bonhomme, S. and Manresa, E. (2015), 'Grouped patterns of heterogeneity in panel data', *Econometrica* **83**(3), 1147–1184.

Cai, J., Horrace, W. C. and Lee, Y. (2021), Panel nonparametric conditional heteroskedastic frontiers with application to CO2 emissions. Working paper.

Caner, M., Han, X. and Lee, Y. (2018), 'Adaptive elastic net gmm estimation with many invalid moment conditions: Simultaneous model and moment selection', *Journal of Business & Economic Statistics* **36**(1), 24–36.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), 'Least angle regression', *The Annals of Statistics* **32**(2), 407–451.

Erwidodo (1990), Panel data analysis on farm-level efficiency, input demand and output supply of rice farming in west java, indonesia. Unpublished dissertation, Department of Agricultural Economics, Michigan State University, East Lansing, MI.

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

Feng, Q. and Horrace, W. C. (2007), 'Fixed-effect estimation of technical efficiency with time-invariant dummies', *Economics Letters* **95**(2), 247–252.

Flores-Lagunes, A., Horrace, W. C. and Schnier, K. E. (2007), 'Identifying technically efficient fishing vessels: a non-empty, minimal subset approach', *Journal of Applied Econometrics* **22**(4), 729–745.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent.', *Journal of statistical software* **33**(1), 1–22.

Greene, W. H. (1980), 'Maximum likelihood estimation of econometric frontier functions', *Journal of Econometrics* **13**(1), 27–56.

Horrace, W. C. (2005), 'On ranking and selection from independent truncated normal distributions', *Journal of Econometrics* **126**(2), 335–354.

Horrace, W. C. and Schmidt, P. (1996), 'Confidence statements for efficiency estimates from stochastic frontier models', *Journal of Productivity Analysis* **7**(2), 257–282.

Horrace, W. C. and Schmidt, P. (2000), 'Multiple comparisons with the best, with economic applications', *Journal of Applied Econometrics* **15**(1), 1–26.

Hui, F. K. C., Warton, D. I. and Foster, S. D. (2015), 'Tuning parameter selection for the adaptive lasso using eric', *Journal of the American Statistical Association* **110**(509), 262–269.

Jondrow, J., Lovell, C., Materov, I. S. and Schmidt, P. (1982), 'On the estimation of technical inefficiency in the stochastic frontier production function model', *Journal of Econometrics* **19**(2-3), 233–238.

Kumbhakar, S. C., Parmeter, C. F. and Tsionas, E. G. (2013), 'A zero inefficiency stochastic frontier model', *Journal of Econometrics* **172**(1), 66–76.

Merlevède, F., Peligrad, M. and Rio, E. (2009), 'Bernstein inequality and moderate deviations under strong mixing conditions', *Inst. Math. Stat. (IMS) Collect: High Dimensional Probability* **V**, 273–292.

Park, B., Sickles, R. and Simar, L. (1998), 'Stochastic panel frontiers: A semiparametric approach', *Journal of Econometrics* **84**(2), 273–301.

Qian, J. and Su, L. (2016), 'Shrinkage estimation of regression models with multiple structural changes', *Econometric Theory* **32**(6), 1376–1433.

Rho, S. and Schmidt, P. (2015), 'Are all firms inefficient?', *Journal of Productivity Analysis* **43**(3), 327–349.

Schmidt, P. and Sickles, R. C. (1984), 'Production frontiers and panel data', *Journal of Business and Economic Statistics* **2**(4), 367–374.

Simar, L. and Wilson, P. W. (2009), 'Inferences from cross-sectional, stochastic frontier models', *Econometric Reviews* **29**(1), 62–98.

Su, L., Shi, Z. and Phillips, P. C. B. (2016), 'Identifying latent structures in panel data', *Econometrica* **84**(6), 2215–2264.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.

Wang, H., Li, R. and Tsai, C.-L. (2007), 'Tuning parameter selectors for the smoothly clipped absolute deviation method.', *Biometrika* **94**(3), 553–568.

Wang, W. S. and Schmidt, P. (2009), 'On the distribution of estimated technical efficiency in stochastic frontier models', *Journal of Econometrics* **148**(1), 36–45.

Wheat, P., Greene, W. and Smith, A. (2014), 'Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models', *Journal of Productivity Analysis* **42**(1), 55–65.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

# Figures and Tables

Table 1: Estimation accuracy

| | | RMSE | | Point Estimate ($\alpha_0 = 1$) | | Ranking correlation | |
|---|---|---|---|---|---|---|---|
| (N, T) | $\sigma_u$ | $\hat{U}(\lambda)$ | $\hat{U}$ | $\hat{\alpha}(\lambda)$ | $\hat{\alpha}$ | LASSO | LSDV |
| (100,10) | 1 | 0.2980 | 0.7591 | 1.005 | 1.687 | 0.92 | 0.88 |
| | | (0.0370) | (0.1394) | (0.097) | (0.152) | (0.030) | (0.033) |
| (100,30) | 1 | 0.1840 | 0.4243 | 0.979 | 1.382 | 0.96 | 0.94 |
| | | (0.0263) | (0.0786) | (0.051) | (0.087) | (0.015) | (0.016) |
| (100,50) | 1 | 0.1456 | 0.3252 | 0.977 | 1.292 | 0.97 | 0.96 |
| | | (0.0214) | (0.0599) | (0.037) | (0.066) | (0.010) | (0.011) |
| (100,70) | 1 | 0.1223 | 0.2777 | 0.982 | 1.250 | 0.98 | 0.97 |
| | | (0.0188) | (0.0533) | (0.032) | (0.058) | (0.008) | (0.009) |
| (100,10) | 2 | 0.3020 | 0.7390 | 1.041 | 1.665 | 0.96 | 0.95 |
| | | (0.0416) | (0.1401) | (0.105) | (0.155) | (0.012) | (0.013) |
| (100,30) | 2 | 0.1762 | 0.4193 | 0.994 | 1.376 | 0.98 | 0.98 |
| | | (0.0218) | (0.0795) | (0.049) | (0.087) | (0.005) | (0.006) |
| (100,50) | 2 | 0.1365 | 0.3219 | 0.992 | 1.288 | 0.99 | 0.98 |
| | | (0.0162) | (0.0625) | (0.036) | (0.069) | (0.003) | (0.004) |
| (100,70) | 2 | 0.1143 | 0.2772 | 0.994 | 1.249 | 0.99 | 0.99 |
| | | (0.0132) | (0.0528) | (0.030) | (0.058) | (0.003) | (0.003) |
| (100,10) | 4 | 0.2997 | 0.7266 | 1.062 | 1.652 | 0.98 | 0.98 |
| | | (0.0429) | (0.1417) | (0.102) | (0.157) | (0.004) | (0.004) |
| (100,30) | 4 | 0.1699 | 0.4211 | 1.004 | 1.377 | 0.99 | 0.99 |
| | | (0.0175) | (0.0820) | (0.046) | (0.090) | (0.002) | (0.002) |
| (100,50) | 4 | 0.1298 | 0.3274 | 1.000 | 1.294 | 0.99 | 0.99 |
| | | (0.0133) | (0.0654) | (0.032) | (0.071) | (0.001) | (0.001) |
| (100,70) | 4 | 0.1102 | 0.2737 | 0.995 | 1.245 | 0.99 | 0.99 |
| | | (0.0120) | (0.0529) | (0.027) | (0.058) | (0.001) | (0.001) |
| (200,10) | 1 | 0.2923 | 0.8240 | 1.010 | 1.759 | 0.92 | 0.89 |
| | | (0.0264) | (0.1334) | (0.075) | (0.145) | (0.022) | (0.022) |
| (200,70) | 1 | 0.1216 | 0.3055 | 0.985 | 1.281 | 0.98 | 0.98 |
| | | (0.0142) | (0.0511) | (0.025) | (0.055) | (0.005) | (0.006) |
| (200,10) | 4 | 0.2939 | 0.7917 | 1.060 | 1.725 | 0.98 | 0.98 |
| | | (0.0263) | (0.1294) | (0.076) | (0.139) | (0.003) | (0.003) |
| (200,70) | 4 | 0.1093 | 0.3046 | 0.999 | 1.279 | 0.99 | 0.99 |
| | | (0.0082) | (0.0502) | (0.020) | (0.054) | (0.000) | (0.000) |
| (1000,10) | 1 | 0.2867 | 0.9762 | 1.017 | 1.923 | 0.92 | 0.89 |
| | | (0.0133) | (0.1130) | (0.041) | (0.118) | (0.011) | (0.010) |
| (1000,10) | 2 | 0.2880 | 0.9751 | 1.050 | 1.921 | 0.97 | 0.96 |
| | | (0.0103) | (0.1178) | (0.042) | (0.124) | (0.004) | (0.004) |
| (1000,10) | 4 | 0.2871 | 0.9696 | 1.068 | 1.916 | 0.99 | 0.99 |
| | | (0.0096) | (0.1209) | (0.039) | (0.127) | (0.001) | (0.001) |

Table 2: Selection accuracy

| | $\sigma_u = 1$ | | | | $\sigma_u = 2$ | | | | $\sigma_u = 4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (N, T) | $P_{\mathcal{S}^c}$ | $P_{\mathcal{S}}$ | $\hat{\delta}$ | Max-miss | $P_{\mathcal{S}^c}$ | $P_{\mathcal{S}}$ | $\hat{\delta}$ | Max-miss | $P_{\mathcal{S}^c}$ | $P_{\mathcal{S}}$ | $\hat{\delta}$ | Max-miss |
| (100,10) | 0.6842 | 0.8843 | 0.4864 | 0.7495 | 0.8321 | 0.8598 | 0.3754 | 0.6754 | 0.9117 | 0.8558 | 0.3186 | 0.5735 |
| | (0.1051) | (0.0966) | (0.0950) | (0.2177) | (0.0705) | (0.1139) | (0.0755) | (0.2317) | (0.0473) | (0.1149) | (0.0589) | (0.2606) |
| (100,30) | 0.7501 | 0.9440 | 0.4581 | 0.4749 | 0.8637 | 0.9375 | 0.3767 | 0.4275 | 0.9289 | 0.9372 | 0.3310 | 0.3662 |
| | (0.0853) | (0.0613) | (0.0713) | (0.1321) | (0.0568) | (0.0607) | (0.0514) | (0.1445) | (0.0376) | (0.0686) | (0.0396) | (0.1649) |
| (100,50) | 0.7788 | 0.9589 | 0.4425 | 0.3818 | 0.8868 | 0.9545 | 0.3656 | 0.3357 | 0.9381 | 0.9587 | 0.0.3310 | 0.2780 |
| | (0.0762) | (0.0487) | (0.0620) | (0.1074) | (0.0490) | (0.0516) | (0.0434) | (0.1173) | (0.0335) | (0.0503) | (0.0315) | (0.1276) |
| (100,70) | 0.8085 | 0.9594 | 0.4219 | 0.3205 | 0.9005 | 0.9602 | 0.3577 | 0.2773 | 0.9437 | 0.9673 | 0.3296 | 0.2379 |
| | (0.0704) | (0.0498) | (0.0580) | (0.0915) | (0.0453) | (0.0480) | (0.0398) | (0.1033) | (0.0329) | (0.0450) | (0.0302) | (0.1178) |
| (200,10) | 0.6957 | 0.8792 | 0.4768 | 0.8300 | 0.8400 | 0.8523 | 0.3677 | 0.7516 | 0.9149 | 0.8531 | 0.3155 | 0.6837 |
| | (0.0796) | (0.0789) | (0.0749) | (0.1917) | (0.0517) | (0.0844) | (0.0565) | (0.2028) | (0.0335) | (0.0946) | (0.0464) | (0.2245) |
| (200,70) | 0.8154 | 0.9575 | 0.4165 | 0.3574 | 0.9021 | 0.9559 | 0.3553 | 0.3268 | 0.9472 | 0.9626 | 0.3257 | 0.2912 |
| | (0.0526) | (0.0375) | (0.0441) | (0.0813) | (0.0339) | (0.0383) | (0.0311) | (0.0911) | (0.0224) | (0.0344) | (0.0221) | (0.0989) |
| (1000,10) | 0.7093 | 0.8735 | 0.4655 | 0.9991 | 0.8477 | 0.8473 | 0.3608 | 0.9562 | 0.9205 | 0.8415 | 0.3081 | 0.8852 |
| | (0.0461) | (0.0481) | (0.0452) | (0.1441) | (0.0291) | (0.0554) | (0.0353) | (0.1572) | (0.0176) | (0.0547) | (0.0269) | (0.1630) |

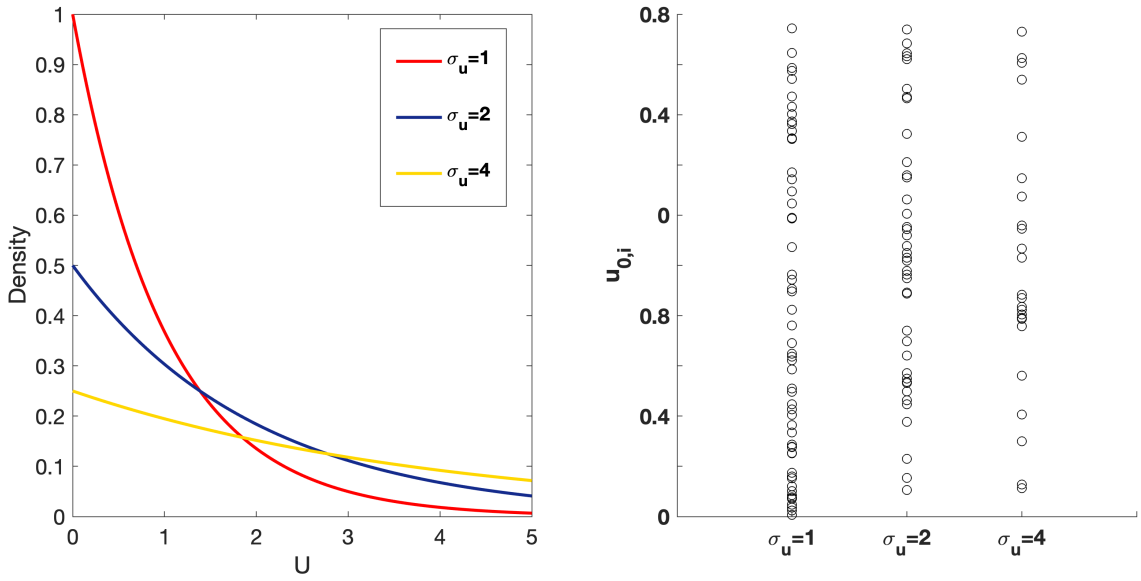


Figure 1: PDFs of inefficiency with different $\sigma_u$ values and an example of draws from each PDF.

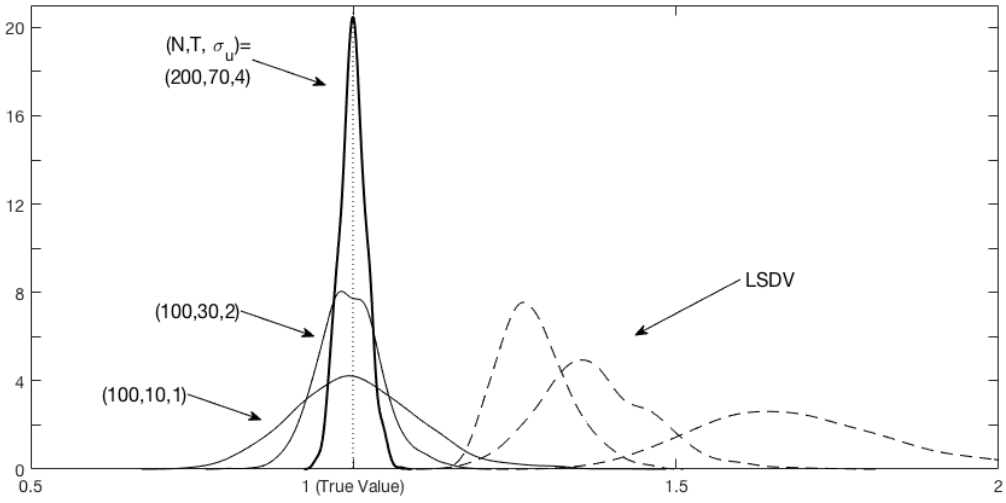


Figure 2: Distribution of estimates of $\alpha_0$ from the LASSO (solid) and LSDV (dashed)

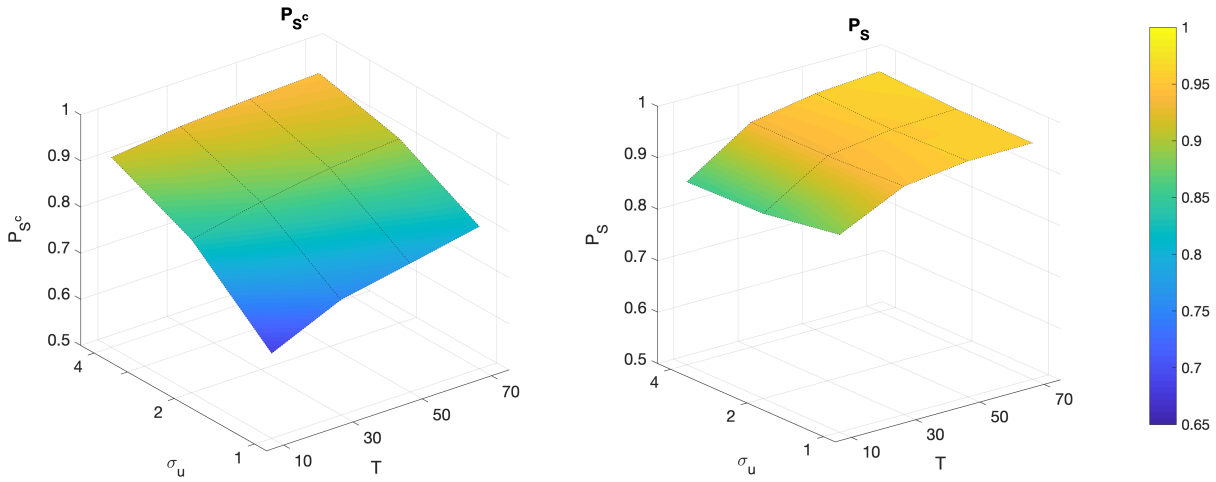


Figure 3: Visualization of the selection performance ($N = 100$)

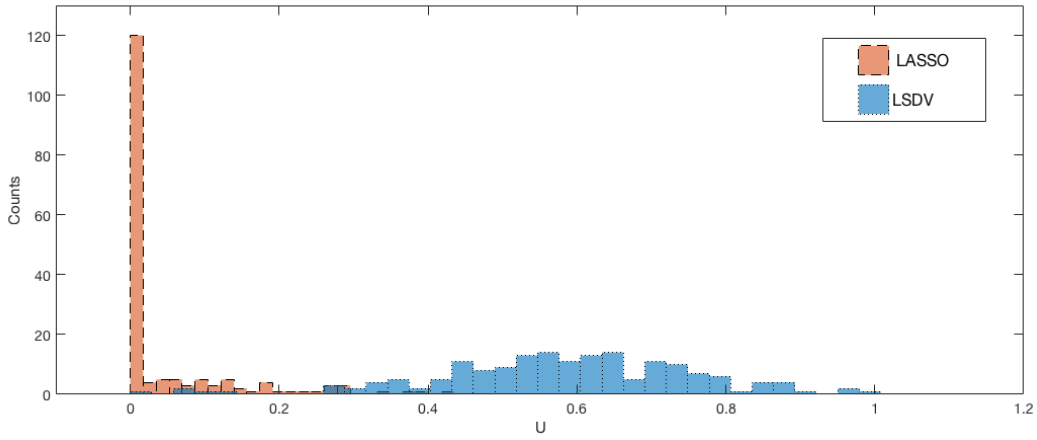


Figure 4: Distribution of the rice farm inefficiency estimates

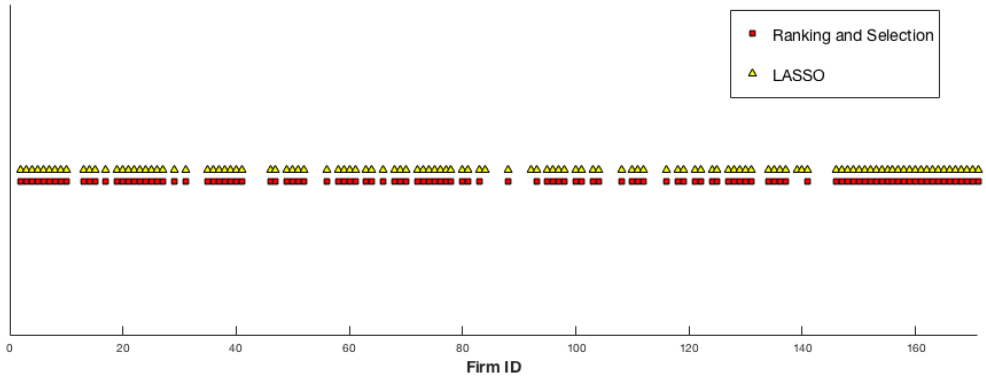


Figure 5: Firm IDs estimated as fully efficient by the LASSO and in the best subset by R&S

# A Appendix: Proofs

Let $\varkappa_{NT} = (\log N)/\sqrt{T}$. We first derive some technical lemmas.

**Lemma A.1** *Suppose Assumption 2-(1) and 2-(2)-(ii) hold. Then, for some $0 < C_x, C_v < \infty$, as $(N,T) \to \infty$, we have*

$$\begin{aligned}
&(a) \quad \max_{1\le i\le N} \Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| \ge C_x\varkappa_{NT}\right) = o\left(N^{-1}\right), \text{ and} \\
&\qquad \max_{1\le i\le N} \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} v_{it}\right| \ge C_v\varkappa_{NT}\right) = o\left(N^{-1}\right); \\
&(b) \quad \Pr\left(\max_{1\le i\le N}\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| \ge C_x\varkappa_{NT}\right) = o\left(1\right), \text{ and} \\
&\qquad \Pr\left(\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T} v_{it}\right| \ge C_v\varkappa_{NT}\right) = o\left(1\right).
\end{aligned}$$

**Proof of Lemma A.1** We only prove for the first part of (a) since the proof for the second part of (a) is analogous, and (a) imply (b) because

$$\begin{aligned}
\Pr\left(\max_{1\le i\le N}\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| \ge C_x\varkappa_{NT}\right) &\le \sum_{i=1}^{N}\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| \ge C_x\varkappa_{NT}\right) \\
&\le N\max_{1\le i\le N}\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| \ge C_x\varkappa_{NT}\right) \\
&= N\cdot o(N^{-1}) = o(1)
\end{aligned}$$

and similarly for the second part of (b), if (a) is true.

To prove the first result of (a), we let $M_T = \sqrt{T}/(\log T)^2$ and $\mathbf{1}_{it} = \mathbf{1}\{||x_{it}|| < M_T\}$. We define

$$\begin{aligned}
\xi_{1,it} &= x_{it}\mathbf{1}_{it} - E\left[x_{it}\mathbf{1}_{it}\right], \\
\xi_{2,it} &= x_{it}\left(1 - \mathbf{1}_{it}\right),
\end{aligned}$$

$$\xi_{3,it} \quad = \quad -E\left[x_{it}\left(1-\mathbf{1}_{it}\right)\right].$$

Then, $x_{it} - E[x_{it}] = \xi_{1,it} + \xi_{2,it} + \xi_{3,it}$ and thus we have

$$\begin{aligned}\max_{1\leq i\leq N}\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it}-E[x_{it}]\}\right\| \geq C_x\varkappa_{NT}\right) \quad &\leq \quad \max_{1\leq i\leq N}\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{1,it}\right\| + \left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{2,it}\right\|\right. \\ &\qquad \left. + \left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3,it}\right\| \geq C_x\varkappa_{NT}\right).\end{aligned}$$

We prove the first part of (a) by showing

$$\begin{aligned}&\text{(a1)} \quad N\cdot\max_{1\leq i\leq N}\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{1,it}\right\| \geq \frac{C_x}{2}\varkappa_{NT}\right) = o\left(1\right), \\ &\text{(a2)} \quad N\cdot\max_{1\leq i\leq N}\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{2,it}\right\| \geq \frac{C_x}{2}\varkappa_{NT}\right) = o\left(1\right), \text{ and} \\ &\text{(a3)} \quad \max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3,it}\right\| = o(\varkappa_{NT}).\end{aligned}$$

To prove (a1), we let $\xi_{1,it}^{\varphi} = \varphi'\xi_{1,it}$ for some constant $p\times 1$ vector $\varphi$ with $||\varphi|| = 1$. Then, by Assumption 2-(1)-(ii), $\xi_{1,it}^{\varphi}$ is a zero-mean strong mixing process, not necessarily stationary, with the mixing coefficients satisfying $\alpha[t] \leq c_\alpha\rho^t$ for some $c_\alpha > 0$ and $\rho \in (0,1)$. In addition, $\max_{1\leq t\leq T}|\xi_{1,it}^{\varphi}| \leq 2M_T$ almost surely by construction. We define $v_N^2 = \max_{1\leq i\leq N}\sup_{t\geq 1}\{var(\xi_{1,it}^{\varphi}) + 2\sum_{s=t+1}^{\infty}|cov(\xi_{1,it}^{\varphi},\xi_{1,is}^{\varphi})|$, which is bounded by Assumption 2-(1)-(ii) and (iii), and the Davydov inequality. Then, by Lemma S1.1 of Su, Shi and Phillips (2016), there exists a constant $C_0 > 0$ such that for any $T \geq 2$ and $C_x > 0$,

$$\begin{aligned}N\cdot\max_{1\leq i\leq N}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\xi_{1,it}^{\varphi}\right| \geq \frac{C_x}{2}\varkappa_{NT}\right) \quad &\leq \quad N\exp\left(-\frac{C_0C_x^2T^2\varkappa_{NT}^2/4}{v_N^2T + 4M_T^2 + 2C_xT\varkappa_{NT}M_T\left(\log T\right)^2/2}\right) \\ &= \quad \exp\left(-\left\{\frac{C_0C_x^2(\log N)^2/4}{v_N^2 + 4/(\log T)^4 + C_x(\log N)} - \log N\right\}\right).\end{aligned}$$

Thus, by choosing $C_x$ sufficiently large, it follows that

$$N \max_{1\leq i\leq N} \Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{1,it}\right\| \geq \frac{C_x}{2}\varkappa_{NT}\right) \to 0 \quad \text{as } (N,T)\to\infty.$$

Next, by Assumption 2-(1)-(iii) and 2-(2)-(ii), and the Boole and Markov inequalities, we have

$$\begin{aligned} N\cdot \max_{1\leq i\leq N} \Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{2,it}\right\| \geq \frac{C_x}{2}\varkappa_{NT}\right) &\leq N\cdot \max_{1\leq i\leq N} \Pr\left(\max_{1\leq t\leq T}\|x_{it}\| \geq M_T\right) \\ &\leq NT \max_{1\leq i\leq N}\max_{1\leq t\leq T} \Pr\left(\|x_{it}\| \geq M_T\right) \\ &\leq \frac{NT}{M_T^q}\max_{1\leq i\leq N}\max_{1\leq t\leq T} E\left||x_{it}\right||^q \\ &= o(1). \end{aligned}$$

Lastly, by Assumption 2-(1)-(iii), and the Hölder and Markov inequalities, for some $q>3$,

$$\begin{aligned} \max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3,it}\right\| &\leq \max_{1\leq i\leq N}\max_{1\leq t\leq T} E\left\|x_{it}\mathbf{1}\left\{\|x_{it}\|\geq M_T\right\}\right\| \\ &\leq \max_{1\leq i\leq N}\max_{1\leq t\leq T}\left(E\left\|x_{it}\right\|^{q/2}\right)^{2/q} \max_{1\leq i\leq N}\max_{1\leq t\leq T}\left\{\Pr\left(||x_{it}||\geq M_T\right)\right\}^{(q-2)/q} \\ &\leq \max_{1\leq i\leq N}\max_{1\leq t\leq T}\left(E\left\|x_{it}\right\|^{q/2}\right)^{2/q} \max_{1\leq i\leq N}\max_{1\leq t\leq T}\left(\frac{E\left\|x_{it}\right\|^q}{M_T^q}\right)^{(q-2)/q} \\ &= O\left(M_T^{-(q-2)}\right) = o(\varkappa_{NT}) \end{aligned}$$

where we use the fact that $M_T^{(q-2)}\varkappa_{NT} = T^{(q-3)/2}\log N/(\log T)^2 \to \infty$ for $q>3$ in the last step. Then, the desired result follows by combining (a1), (a2) and (a3). ■

**Proof of Lemma in the main text** First, note that Assumption 2 and Lemma A.1 yield

$$\begin{aligned} &\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{x_{it}'(\beta_0-\hat{\beta})+v_{it}\right\}\right| \\ \leq\ &\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it}-E[x_{it}]\}\right\| + \max_{1\leq i\leq N}E||x_{it}||\right)\left\|\hat{\beta}-\beta_0\right\| + \max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}v_{it}\right| \end{aligned}$$

$$
\begin{aligned}
&= (O_p(\varkappa_{NT}) + O(1))\, O_p((NT)^{-1/2}) + O_p(\varkappa_{NT}) \\
&= O_p(\varkappa_{NT}). \qquad \text{(A.1)}
\end{aligned}
$$

Recall $\eta = \min_{i\in\mathcal{S}^c} u_{0,i}$ and $\hat{\alpha}_i = T^{-1}\sum_{t=1}^{T}(y_{it} - x_{it}'\hat{\beta}) = T^{-1}\sum_{t=1}^{T}(\alpha_0 - u_{0,i} + x_{it}'(\beta_0 - \hat{\beta}) + v_{it})$ where $u_{0,i} = 0$ for all $i \in \mathcal{S}$. Thus, it follows that

$$
\begin{aligned}
&\min_{i\in\mathcal{S}} \hat{\alpha}_i - \max_{i\in\mathcal{S}^c} \hat{\alpha}_i \\
= &\min_{i\in\mathcal{S}} \left\{ \frac{1}{T}\sum_{t=1}^{T}(\alpha_0 + x_{it}'(\beta_0 - \hat{\beta}) + v_{it}) \right\} - \max_{i\in\mathcal{S}^c} \left\{ \frac{1}{T}\sum_{t=1}^{T}(\alpha_0 - u_{0,i} + x_{it}'(\beta_0 - \hat{\beta}) + v_{it}) \right\} \\
\geq &\min_{i\in\mathcal{S}^c} u_{0,i} + \left[ \min_{i\in\mathcal{S}} \left\{ \frac{1}{T}\sum_{t=1}^{T}(x_{it}'(\beta_0 - \hat{\beta}) + v_{it}) \right\} - \max_{i\in\mathcal{S}^c} \left\{ \frac{1}{T}\sum_{t=1}^{T}(x_{it}'(\beta_0 - \hat{\beta}) + v_{it}) \right\} \right] \\
\geq &\eta - 2\max_{1\leq i\leq N} \left| \frac{1}{T}\sum_{t=1}^{T}(x_{it}'(\beta_0 - \hat{\beta}) + v_{it}) \right| \\
> &\frac{\eta}{2} - O_p(\varkappa_{NT}),
\end{aligned}
$$

which implies

$$
\Pr\left( \min_{i\in\mathcal{S}} \hat{\alpha}_i - \max_{i\in\mathcal{S}^c} \hat{\alpha}_i > 0 \right) \to 1 \qquad \text{(A.2)}
$$

as $(N,T) \to \infty$ since $\eta > 0$ and $\eta/\varkappa_{NT} \to \infty$ by Assumption 2-(2)-(iii). (A.2), in turn, implies (i) of the Lemma because $\hat{\alpha}$ is defined as $\max_{1\leq i\leq N} \hat{\alpha}_i$.

By (A.2), we can let $\hat{\alpha} = \max_{i\in\mathcal{S}} \hat{\alpha}_i$ for sufficiently large $(N,T)$, instead of $\hat{\alpha} = \max_{1\leq i\leq N} \hat{\alpha}_i$. Hence, for sufficiently large $(N,T)$, we have

$$
\begin{aligned}
|\hat{\alpha} - \alpha_0| &= \left| \max_{i\in\mathcal{S}} \left\{ \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_0 + x_{it}'(\beta_0 - \hat{\beta}) + v_{it}\right) \right\} - \alpha_0 \right| \\
&\leq \max_{1\leq i\leq N} \left| \frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'(\beta_0 - \hat{\beta}) + v_{it}\right) \right| = O_p(\varkappa_{NT}),
\end{aligned}
$$

which proves (ii) of the Lemma. Since $\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i = (\hat{\alpha} - \alpha_0) + (\alpha_0 - \hat{\alpha}_i) = (\hat{\alpha} - \alpha_0) + (u_{0,i} + \alpha_{0,i} - \hat{\alpha}_i)$,

a direct consequence of (ii) of the Lemma is

$$\begin{aligned} |\hat{u}_i - u_{0,i}| &\leq |\hat{\alpha} - \alpha_0| + |\hat{\alpha}_i - \alpha_{0,i}| \\ &\leq |\hat{\alpha} - \alpha_0| + \max_{1\leq i\leq N} \left| \frac{1}{T}\sum_{t=1}^{T} x_{it}'(\beta_0 - \hat{\beta}) + v_{it} \right| \\ &= O_p(\varkappa_{NT}). \quad \blacksquare \end{aligned} \tag{A.3}$$

**Proof of Theorem 1** For Equation (5) in the main text, we form a Lagrangian as

$$\mathcal{L}\left(\alpha, \{u_i\}_{i=1}^N, \{\rho_i\}_{i=1}^N\right) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta} - \alpha + u_i\right)^2 + \lambda\sum_{i=1}^{N}\pi_i u_i - \sum_{i=1}^{N}\rho_i u_i,$$

where $\rho_i \geq 0$, $u_i \geq 0$, and $\rho_i u_i = 0$ (complementary slackness) for all $i$. From the Karush-Kuhn-Tucker (KKT) conditions, we have

$$\widehat{\alpha}(\lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta} + \widehat{u}_i(\lambda)\right) \tag{A.4}$$

$$\widehat{u}_i(\lambda) = \max\left\{0, \widehat{\alpha}(\lambda) - \frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta}\right) - \frac{\lambda}{2T}\hat{\pi}_i\right\}. \tag{A.5}$$

Recall $\delta = |\mathcal{S}|/N$ and let $\hat{\delta} = |\widehat{\mathcal{S}}|/N$. By plugging (A.5) into (A.4), we have

$$\begin{aligned} \widehat{\alpha}(\lambda) &= \frac{1}{NT}\sum_{i\in\widehat{\mathcal{S}}}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta}\right) + \frac{1}{NT}\sum_{i\in\widehat{\mathcal{S}}^c}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta} + \widehat{u}_i(\lambda)\right) \\ &= \frac{1}{NT}\sum_{i\in\widehat{\mathcal{S}}}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta}\right) + \frac{1}{N}\sum_{i\in\widehat{\mathcal{S}}^c}\left(\widehat{\alpha}(\lambda) - \frac{\lambda}{2T}\hat{\pi}_i\right) \\ &= \frac{1}{NT}\sum_{i\in\widehat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{it}'\left(\beta_0 - \widehat{\beta}\right) + \alpha_0 - u_{0,i} + v_{it}\right) + \left(1 - \widehat{\delta}\right)\widehat{\alpha}(\lambda) - \frac{\lambda}{2NT}\sum_{i\in\widehat{\mathcal{S}}^c}\hat{\pi}_i \end{aligned}$$

and hence

$$\widehat{\alpha}\left(\lambda\right)-\alpha_0=\frac{1}{\hat{\delta}NT}\sum_{i\in\widehat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{it}'\left(\beta_0-\widehat{\beta}\right)-u_{0,i}+v_{it}\right)-\frac{\lambda}{2\hat{\delta}NT}\sum_{i\in\widehat{\mathcal{S}}^c}\hat{\pi}_i. \tag{A.6}$$

This shows that $\hat{\alpha}(\lambda)$ is estimated as a common intercept for the firms classified as fully efficient by the LASSO and also contains bias due to the use of shrinkage on $\hat{u}_i(\lambda)$. From (A.5), it follows that, for $i\in\widehat{\mathcal{S}}^c$ (i.e. $\widehat{u}_i\left(\lambda\right)>0$),

$$\begin{aligned}\widehat{u}_i\left(\lambda\right) &= \widehat{\alpha}\left(\lambda\right)-\frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'\left(\beta_0-\widehat{\beta}\right)+\alpha_0-u_{0,i}+v_{it}\right)-\frac{\lambda}{2T}\hat{\pi}_i\\ &= \frac{1}{\hat{\delta}NT}\sum_{j\in\widehat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{jt}'\left(\beta_0-\widehat{\beta}\right)-u_{0,j}+v_{jt}\right)-\frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'\left(\beta_0-\widehat{\beta}\right)-u_{0,i}+v_{it}\right)\\ &\quad -\frac{\lambda}{2\hat{\delta}NT}\sum_{j\in\widehat{\mathcal{S}}^c}\hat{\pi}_j-\frac{\lambda}{2T}\hat{\pi}_i.\end{aligned}$$

We prove the theorem by showing $\mathcal{S}\subset\hat{\mathcal{S}}$ and $\mathcal{S}^c\subset\hat{\mathcal{S}}^c$ w.p.a.1.

(i) We first prove $\mathcal{S}\subset\hat{\mathcal{S}}$ w.p.a.1 by showing $\Pr\left(\max_{i\in\mathcal{S}}\hat{u}_i(\lambda)>0\right)\to 0$. Let $\hat{\tau}=\max_{i\in\mathcal{S}}\hat{u}_i$ and $\widetilde{\varkappa}_{NT}=\varkappa_{NT}\log N$. Then, from (A.5), we have

$$\begin{aligned}\Pr\left(\max_{i\in\mathcal{S}}\hat{u}_i(\lambda)>0\right) &= \Pr\left(\max_{i\in\mathcal{S}}\left\{\hat{\alpha}(\lambda)-\hat{\alpha}_i-\frac{\lambda}{T}\hat{\pi}_i\right\}>0\right)\\ &\leq \Pr\left(\max_{i\in\mathcal{S}}\left\{\hat{\alpha}(\lambda)-\hat{\alpha}_i-\frac{\lambda}{T}\hat{\pi}_i\right\}>0,\hat{\tau}\leq\widetilde{\varkappa}_{NT}\right)+\Pr(\hat{\tau}>\widetilde{\varkappa}_{NT})\\ &\leq \Pr\left(\max_{i\in\mathcal{S}}\left\{\frac{1}{\hat{\delta}NT}\sum_{j\in\widehat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{jt}'(\beta_0-\hat{\beta})-u_{0,j}+v_{jt}\right)-\frac{\lambda}{2\hat{\delta}NT}\sum_{j\in\widehat{\mathcal{S}}^c}\hat{\pi}_j\right.\right.\\ &\quad \left.\left.-\frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'(\beta_0-\hat{\beta})+v_{it}\right)-\frac{\lambda}{2T}\widetilde{\varkappa}_{NT}^{-\gamma}\right\}>0\right)+\Pr(\hat{\tau}>\widetilde{\varkappa}_{NT})\\ &\leq \Pr\left(2\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}x_{it}'(\beta_0-\hat{\beta})+v_{it}\right|-\frac{\lambda}{2T}\widetilde{\varkappa}_{NT}^{-\gamma}>0\right)+\Pr(\hat{\tau}>\widetilde{\varkappa}_{NT})\end{aligned} \tag{A.7}$$

where we use the fact that $u_{0,j} \geq 0$ and $\hat{\pi}_j \geq 0$ for all $j$ in the last step. Then, we can easily show that first term in (A.7) is $o(1)$ because $\max_{1\leq i\leq N}\left|T^{-1}\sum_{t=1}^{T} x_{it}'(\beta_0-\hat{\beta})+v_{it}\right| = O_p(\varkappa_{NT})$ from (A.1) and $((\lambda/T)\widetilde{\varkappa}_{NT}^{-\gamma})/\varkappa_{NT} \to \infty$ as $(N,T)\to\infty$ by Assumption 2-(2)-(iii). The second term in (A.7) is also $o(1)$ because

$$\hat{\tau} = \max_{i\in\mathcal{S}} \hat{u}_i = \max_{i\in\mathcal{S}} \{(\hat{\alpha}-\alpha_0)-(\hat{\alpha}_i-\alpha_{0,i})\} \leq 2 \max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T} x_{it}'(\beta_0-\hat{\beta})+v_{it}\right| = O_p(\varkappa_{NT})$$

where we use the fact $u_{0,i}=0$ for $i\in\mathcal{S}$.

(ii) Next, we prove $\mathcal{S}^c \subset \hat{\mathcal{S}}^c$ w.p.a.1. Define $\mathcal{D}_i \equiv \{\hat{u}_i(\lambda)=0\}$ and then,

$$\Pr\left(\text{there exists } i\in\mathcal{S}^c \text{ such that } \hat{u}_i(\lambda)=0\right) = \Pr\left(\bigcup_{i\in\mathcal{S}^c}\mathcal{D}_i\right).$$

Let $|\mathcal{S}^c|=J$. We arbitrarily list the firms in $\mathcal{S}^c$ and use an auxiliary index, $[j]$ for $j=1,...,J$, to denote the $j^{th}$ firm on the list. Then, using the probability algebra, we can partition $\bigcup_{i\in\mathcal{S}^c}\mathcal{D}_i$ into parts that are mutually exclusive and compose $\bigcup_{i\in\mathcal{S}^c}\mathcal{D}_i$ such that $\mathcal{D}_{[1]}\cap\left(\bigcup_{j=2}^{J}\mathcal{D}_{[j]}\right)^c$, $\mathcal{D}_{[2]}\cap\left(\bigcup_{j=3}^{J}\mathcal{D}_{[j]}\right)^c$, ..., and $\mathcal{D}_{[J]}$. Therefore, we have

$$\begin{aligned}
&\Pr\left(\bigcup_{i\in\mathcal{S}^c}\mathcal{D}_i\right)\\
=\;&\sum_{j=1}^{J}\Pr\left(\mathcal{D}_{[j]}\cap\left(\bigcup_{k=j+1}^{J}\mathcal{D}_{[k]}\right)^c\right)\\
=\;&\sum_{j=1}^{J}\Pr\left(\hat{u}_{[j]}(\lambda)=0,\hat{u}_{[j+1]}(\lambda)>0,\hat{u}_{[j+2]}(\lambda)>0,...,\hat{u}_{[J]}(\lambda)>0\right),
\end{aligned}$$

which is true regardless of the order of the firms on the list. So, we list the firms in $\mathcal{S}^c$ according to the size of inefficiency in ascending order so that $u_{0,[1]}\leq ...\leq u_{0,[j]}...\leq u_{0,[J]}$. Then, we have

$$\Pr\left(\text{there exists } i\in\mathcal{S}^c \text{ such that } \hat{u}_i(\lambda)=0\right)$$

$$
\begin{aligned}
&= \sum_{j=1}^{J} \Pr\left(\hat{u}_{[j]}(\lambda)=0, \hat{u}_{[j+1]}(\lambda)>0, \hat{u}_{[j+2]}(\lambda)>0, ..., \hat{u}_{[J]}(\lambda)>0\right) \\
&= \sum_{j=1}^{J} \Pr\left(\hat{u}_{[j]}(\lambda)=0 \,\middle|\, \hat{u}_{[j+1]}(\lambda)>0, ..., \hat{u}_{[J]}(\lambda)>0\right) \times \Pr\left(\hat{u}_{[j+1]}(\lambda)>0 \,\middle|\, \hat{u}_{[j+2]}(\lambda)>0, ...\right) \times ... \\
&\quad ... \times \Pr\left(\hat{u}_{[J-1]}(\lambda)>0 \,\middle|\, \hat{u}_{[J]}(\lambda)>0\right) \times \Pr\left(\hat{u}_{[J]}(\lambda)>0\right) \\
&\leq \sum_{j=1}^{J} \Pr\left(\hat{u}_{[j]}(\lambda)=0 \,\middle|\, \hat{u}_{[j+1]}(\lambda)>0, ..., \hat{u}_{[J]}(\lambda)>0\right) \\
&= \sum_{j=1}^{J} \Pr\left(\frac{1}{\hat{\delta}NT}\sum_{i\in\hat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{it}'(\beta_0-\hat{\beta})-u_{0,i}+v_{it}\right)-\frac{\lambda}{2\hat{\delta}NT}\sum_{i\in\hat{\mathcal{S}}^c}\hat{\pi}_i\right. \\
&\quad \left.-\frac{1}{T}\sum_{t=1}^{T}\left(x_{[j]t}'(\beta_0-\hat{\beta})-u_{0,[j]}+v_{[j]t}\right)-\frac{\lambda}{2T}\hat{\pi}_{[j]}<0 \,\middle|\, \hat{u}_{[j+1]}(\lambda)>0, ..., \hat{u}_{[J]}(\lambda)>0\right) \\
&= \sum_{j=1}^{J} \Pr\left(\underbrace{u_{0,[j]}-\frac{\sum_{i\in\hat{\mathcal{S}}}u_{0,i}}{\hat{\delta}N}}_{(*)}+\frac{1}{\hat{\delta}NT}\sum_{i\in\hat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{it}'(\beta_0-\hat{\beta})+v_{it}\right)-\frac{\lambda}{2\hat{\delta}NT}\sum_{i\in\hat{\mathcal{S}}^c}\hat{\pi}_i\right. \\
&\quad \left.-\frac{1}{T}\sum_{t=1}^{T}\left(x_{[j]t}'(\beta_0-\hat{\beta})+v_{[j]t}\right)-\frac{\lambda}{2T}\hat{\pi}_{[j]}<0 \,\middle|\, \hat{u}_{[j+1]}(\lambda)>0, ..., \hat{u}_{[J]}(\lambda)>0\right) \qquad \text{(A.8)}
\end{aligned}
$$

We let $\hat{\mathcal{S}}^* = \mathcal{S}^c \cap \hat{\mathcal{S}}$ and $\hat{\delta}^* = |\hat{\mathcal{S}}^*|/N$. Then, $(*)$ in the $j$th probability of (A.8) satisfies

$$
u_{0,[j]}-\frac{\sum_{i\in\hat{\mathcal{S}}}u_{0,i}}{\hat{\delta}N} \geq u_{0,[j]}-\frac{\hat{\delta}^* u_{0,[j]}}{\hat{\delta}}
$$

since $u_{0,i}=0$ for all $i\in\mathcal{S}$ and $u_{0,[j]}=\max_{i\in\hat{\mathcal{S}}^*}u_{0,i}$ in the $j$th event by construction, which, if $\mathcal{S}\subset\hat{\mathcal{S}}$, further gives us the results

$$
u_{0,[j]}-\frac{\hat{\delta}^*}{\hat{\delta}}u_{0,[j]}=\frac{\delta}{\hat{\delta}}u_{0,[j]} \geq \delta u_{0,[j]} \geq \delta\eta \qquad \text{(A.9)}
$$

since $\hat{\delta}-\hat{\delta}^*=\delta$ and $\delta\leq\hat{\delta}\leq 1$ if $\mathcal{S}\subset\hat{\mathcal{S}}$.

Let $\hat{\eta}=\min_{i\in\mathcal{S}^c}\hat{u}_i$, $\breve{\alpha}=\left|\frac{1}{\hat{\delta}NT}\sum_{i\in\hat{\mathcal{S}}}\sum_{t=1}^{T}\left(x_{it}'(\beta_0-\hat{\beta})+v_{it}\right)\right|$, and recall $\widetilde{\varkappa}_{NT}=\varkappa_{NT}\log N$.

Then, we finally have

$$
\begin{aligned}
&\Pr\left(\text{there exists } i \in \mathcal{S}^c \text{ such that } \hat{u}_i(\lambda) = 0\right) \\
\leq\ &\Pr\left(\text{there exists } i \in \mathcal{S}^c \text{ such that } \hat{u}_i(\lambda) = 0,\ ||\beta_0 - \hat{\beta}|| \leq \varkappa_{NT},\ \hat{\eta} \geq \eta - \widetilde{\varkappa}_{NT},\ \breve{\alpha} \leq \widetilde{\varkappa}_{NT},\ \mathcal{S} \subset \hat{\mathcal{S}}\right) \\
&+ \Pr\left(||\beta_0 - \hat{\beta}|| > \varkappa_{NT}\right) + \Pr\left(\breve{\alpha} > \widetilde{\varkappa}_{NT}\right) + \Pr\left(\hat{\eta} < \eta - \widetilde{\varkappa}_{NT}\right) + \Pr\left(\mathcal{S} \not\subset \hat{\mathcal{S}}\right) \qquad \text{(A.10)}
\end{aligned}
$$

where $\Pr\left(||\beta_0 - \hat{\beta}|| > \varkappa_{NT}\right) = o(1)$ by Assumption 2-(2)-(i), $\Pr\left(\mathcal{S} \not\subset \hat{\mathcal{S}}\right) = o(1)$ by the first part of this proof, $\Pr\left(\breve{\alpha} > \widetilde{\varkappa}_{NT}\right) = o(1)$ by the fact that $\breve{\alpha} \leq \max_{1 \leq i \leq N} \left|\frac{1}{T}\sum_{t=1}^{T} x_{it}'(\beta_0 - \hat{\beta}) + v_{it}\right| = O_p(\varkappa_{NT})$ due to (A.1), and $\Pr\left(\hat{\eta} < \eta - \widetilde{\varkappa}_{NT}\right) = o(1)$ since

$$
|\hat{\eta} - \eta| \leq |\hat{\eta} - u_\ell| + |\hat{u}_{\ell_0} - \eta| = O_p(\varkappa_{NT}) \qquad \text{(A.11)}
$$

by (A.3) where $\ell = \text{argmin}_{i \in \mathcal{S}^c} \hat{u}_i$ and $\ell_0 = \text{argmin}_{i \in \mathcal{S}^c} u_{0,i}$.[18] Furthermore, if $\mathcal{S} \subset \hat{\mathcal{S}}$, we have

$$
\frac{\lambda}{2\hat{\delta}NT}\sum_{i \in \hat{\mathcal{S}}^c} \hat{\pi}_i + \frac{\lambda}{2T}\hat{\pi}_{[j]} \leq \frac{\lambda}{2\hat{\delta}NT}(1 - \hat{\delta})N\hat{\eta}^{-\gamma} + \frac{\lambda}{2T}\hat{\eta}^{-\gamma} = \frac{\lambda}{2\hat{\delta}T}\hat{\eta}^{-\gamma} \leq \frac{\lambda}{\delta T}\hat{\eta}^{-\gamma}, \qquad \text{(A.12)}
$$

where we use the fact $\hat{\mathcal{S}}^c \subset \mathcal{S}^c$ and $\delta \leq \hat{\delta} \leq 1$ if $\mathcal{S} \subset \hat{\mathcal{S}}$. Then, for the first term in (A.10), by combining (A.8), (A.9) and (A.12), we have

$$
\begin{aligned}
&\Pr\left(\text{there exists } i \in \mathcal{S}^c \text{ such that } \hat{u}_i(\lambda) = 0,\ ||\beta_0 - \hat{\beta}|| \leq \varkappa_{NT},\ \hat{\eta} \geq \eta - \widetilde{\varkappa}_{NT}, \breve{\alpha} \leq \widetilde{\varkappa}_{NT}, \mathcal{S} \subset \hat{\mathcal{S}}\right) \\
\leq\ &\sum_{j=1}^{J} \Pr\left(\delta\eta - \widetilde{\varkappa}_{NT} - \left|\frac{1}{T}\sum_{t=1}^{T}\left\{x_{[j]t}'(\beta_0 - \hat{\beta}) + v_{[j]t}\right\}\right| - \frac{\lambda}{\delta T}\hat{\eta}^{-\gamma} < 0, ||\beta_0 - \hat{\beta}|| \leq \varkappa_{NT},\ \hat{\eta} \geq \eta - \widetilde{\varkappa}_{NT}\right) \\
\leq\ &\sum_{j=1}^{J} \Pr\left(\delta\eta - \widetilde{\varkappa}_{NT} - \left|\frac{1}{T}\sum_{t=1}^{T}\left\{x_{[j]t}'(\beta_0 - \hat{\beta}) + v_{[j]t}\right\}\right| - \frac{\lambda}{\delta T}(\eta - \widetilde{\varkappa}_{NT})^{-\gamma} < 0,\ ||\beta_0 - \hat{\beta}|| \leq \varkappa_{NT}\right) \\
\leq\ &\sum_{j=1}^{J} \Pr\left(\delta\eta - \widetilde{\varkappa}_{NT} - \varkappa_{NT}\left(\left|\left|\frac{1}{T}\sum_{t=1}^{T}\left\{x_{[j]t} - E[x_{[j]t}]\right\}\right|\right| + E\left\|x_{[j]t}\right\|\right) - \left|\frac{1}{T}\sum_{t=1}^{T} v_{[j]t}\right| - \frac{\lambda}{\delta T}(\eta - \widetilde{\varkappa}_{NT})^{-\gamma} < 0\right)
\end{aligned}
$$

[18] Note that $|\hat{\eta} - \eta| \leq |\hat{u}_{\ell_0} - \eta|$ if $\hat{\eta} > \eta$ and $|\hat{\eta} - \eta| \leq |\hat{\eta} - u_\ell|$ if $\hat{\eta} < \eta$.

$$
\begin{aligned}
&\leq \sum_{j=1}^{J} \Pr\left(\delta\eta - \widetilde{\varkappa}_{NT} - \varkappa_{NT}\left(\widetilde{\varkappa}_{NT} + E\left\|x_{[j]t}\right\|\right) - \left|\frac{1}{T}\sum_{t=1}^{T} v_{[j]t}\right| - \frac{\lambda}{\delta T}(\eta - \widetilde{\varkappa}_{NT})^{-\gamma} < 0\right) \\
&\quad + \sum_{j=1}^{J} \Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| > \widetilde{\varkappa}_{NT})\right) \\
&\leq N \max_{1\leq i\leq N} \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} v_{it}\right| > \Re_{NT}\right) + N \max_{1\leq i\leq N} \Pr\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\{x_{it} - E[x_{it}]\}\right\| > \widetilde{\varkappa}_{NT})\right) \qquad \text{(A.13)}
\end{aligned}
$$

where $\Re_{NT} = \delta\eta - \widetilde{\varkappa}_{NT} - \varkappa_{NT}\left(\widetilde{\varkappa}_{NT} + E||x_{it}||\right) - \frac{\lambda}{\delta T}(\eta - \widetilde{\varkappa}_{NT})^{-\gamma}$. Then we can easily show that the two terms in (A.13) are $o(1)$ by an application of Lemma A.1 and the fact that $\Re_{NT}/\widetilde{\varkappa}_{NT} = \frac{\delta\eta}{\widetilde{\varkappa}_{NT}} - 1 - \varkappa_{NT} - E||x_{it}||/\log N - \frac{\lambda}{\delta T}\eta^{-\gamma}\widetilde{\varkappa}_{NT}^{-1}(1 - \widetilde{\varkappa}_{NT}/\eta)^{-\gamma} \to \infty$ as $(N,T) \to \infty$ by Assumption 1 and 2. Thus, the proof is complete. ■

**Proof Theorem 2** By Theorem 1, w.p.a 1, we have

$$
\sqrt{\delta NT}(\hat{\alpha}(\lambda) - \alpha_0) = \frac{1}{\sqrt{\delta NT}}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T}\left(x_{it}'(\beta_0 - \hat{\beta}) + v_{it}\right) - \frac{\lambda}{2\sqrt{\delta NT}}\sum_{i\in\mathcal{S}^c}\hat{\pi}_i
$$

The second term is $o_p(1)$ since

$$
\frac{\lambda}{\sqrt{\delta NT}}\sum_{i\in\mathcal{S}^c}\hat{\pi}_i \leq \sqrt{\frac{(1-\delta)^2}{\delta}}\lambda\sqrt{\frac{N}{T}}\eta^{-\gamma}\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma} = o_p(1) \qquad \text{(A.14)}
$$

by Assumption 2-(2)-(iii) and the fact that

$$
\frac{\hat{\eta}}{\eta} \leq 1 + \frac{|\hat{\eta} - \eta|}{\eta} = 1 + o_p(1),
$$

by (A.11) and $\varkappa_{NT}/\eta \to 0$ as $(N,T) \to \infty$ due to Assumption 2-(2)-(iii).

Since $\hat{\beta} - \beta_0 = (\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}')^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{v}_{it}$, and $\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{v}_{it} =$

$\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}v_{it}$, we have

$$\begin{aligned}
& \sqrt{\delta NT}(\hat{\alpha}(\lambda)-\alpha_0) \\
= \; & \frac{1}{\sqrt{\delta NT}}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T} v_{it} \\
& -\sqrt{\delta}\left(\frac{1}{\delta NT}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T} x_{it}'\right)\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}v_{it}\right)+o_p(1).
\end{aligned}$$

We define

$$\begin{aligned}
\Upsilon_{\mathcal{S}} &= \operatorname*{plim}_{N,T\to\infty}\frac{1}{\delta NT}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T} x_{it} \\
H_0 &= \operatorname*{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'
\end{aligned}$$

where $H_0>0$ by Assumption 3. We split the sample into $\mathcal{S}$ and $\mathcal{S}^c$ and define two statistics as

$$\begin{aligned}
\Xi_{\mathcal{S},NT} &\equiv \frac{1}{\sqrt{\delta NT}}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T}\left\{v_{it}-\delta\Upsilon_{\mathcal{S}}'H_0^{-1}\tilde{x}_{it}v_{it}\right\} \\
\Xi_{\mathcal{S}^c,NT} &\equiv \frac{1}{\sqrt{(1-\delta)NT}}\sum_{i\in\mathcal{S}^c}\sum_{t=1}^{T}\sqrt{\delta(1-\delta)}\Upsilon_{\mathcal{S}}'H_0^{-1}\tilde{x}_{it}v_{it},
\end{aligned}$$

which are independent since the observations are cross-sectionally independent. By Assumption 3, we thus have

$$\begin{aligned}
\Xi_{\mathcal{S},NT} &\xrightarrow{d} \mathcal{N}\left(0,\sigma_{\mathcal{S}_1}^2+\delta^2\sigma_{\mathcal{S}_2}^2-2\delta\sigma_{\mathcal{S}_1\mathcal{S}_2}\right) \\
\Xi_{\mathcal{S}^c,NT} &\xrightarrow{d} \mathcal{N}\left(0,\delta(1-\delta)\sigma_{\mathcal{S}^c}^2\right)
\end{aligned}$$

as $(N,T)\to\infty$, where

$$\sigma_{\mathcal{S}_1}^2 = \operatorname*{plim}_{N,T\to\infty}\frac{1}{\delta NT}\sum_{i\in\mathcal{S}}\sum_{t=1}^{T}\sum_{k=1}^{T} v_{it}v_{ik}$$

$$
\begin{aligned}
\sigma^2_{\mathcal{S}_2} &= \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \operatorname*{plim}_{N,T\to\infty} \frac{1}{\delta NT} \sum_{i\in\mathcal{S}} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \tilde{x}'_{it} \right\} H_0^{-1} \Upsilon_{\mathcal{S}} \\
\sigma_{\mathcal{S}_1\mathcal{S}_2} &= \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \operatorname*{plim}_{N,T\to\infty} \frac{1}{\delta NT} \sum_{i\in\mathcal{S}} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \right\} \\
\sigma^2_{\mathcal{S}^c} &= \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \operatorname*{plim}_{N,T\to\infty} \frac{1}{(1-\delta) NT} \sum_{i\in\mathcal{S}^c} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \tilde{x}'_{it} \right\} H_0^{-1} \Upsilon_{\mathcal{S}}.
\end{aligned}
$$

Hence, $\sqrt{\delta NT}(\hat{\alpha}(\lambda) - \alpha_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{S}_1} + \delta^2 \sigma^2_{\mathcal{S}_2} - 2\delta^2 \sigma_{\mathcal{S}_1\mathcal{S}_2} + \delta(1-\delta)\sigma^2_{\mathcal{S}^c}\right)$ and the desired result follows.[19]

For the second result, for $i \in \mathcal{S}^c$, we have

$$
\begin{aligned}
\sqrt{T}(\hat{u}_i(\lambda) - u_{0,i}) &= \sqrt{T}(\hat{\alpha}(\lambda) - \alpha_0) - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x'_{it}(\beta_0 - \hat{\beta}) - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} v_{it} - \frac{\lambda}{2\sqrt{T}} \hat{\pi}_i \\
&\equiv \Psi_{1,NT} + \Psi_{2i,NT} + \Psi_{3i,T} + \Psi_{4i,NT},
\end{aligned}
$$

where $\Psi_{1,NT} = O_p(1/\sqrt{\delta N}) = o_p(1)$ from the first result, $\Psi_{2i,NT} = O_p(1/\sqrt{N}) = o_p(1)$ since $\hat{\beta} - \beta_0 = O_p(1/\sqrt{NT})$, and $\Psi_{4i,NT} = o_p(1)$ by a similar argument as in (A.14). Since $\Psi_{3i,T} \xrightarrow{d} \mathcal{N}(0, \sigma_i^2)$ as $T \to \infty$ by Assumption 3, where $\sigma_i^2 = \operatorname{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{T} v_{it} v_{ik}$ for each $i$, we have the desired result. ■

[19] When $v_{it}$ is conditionally i.i.d. across $i$, we have $\sigma^2_{\mathcal{S}_2} = \sigma^2_{\mathcal{S}^c} = \Upsilon'_{\mathcal{S}} H_0^{-1} \left\{ \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \sum_{k=1}^{T} \tilde{x}_{it} v_{it} v_{ik} \tilde{x}'_{it} \right\} H_0^{-1} \Upsilon_{\mathcal{S}}$ and the limiting expression simplies to $\mathcal{N}\left(0, \sigma^2_{\mathcal{S}_1} + \delta \sigma^2_{\mathcal{S}_2} - 2\delta \sigma_{\mathcal{S}_1\mathcal{S}_2}\right)$.