LASSO for Stochastic Frontier Models with Many Efficient Firms

Hyunseok Jung *

January, 2019

Abstract

The LASSO (Tibshirani, 1996) is applied to select a subset of maximally efficient firms in the fixed-effect stochastic frontier model for panel data of Schmidt and Sickles (1981). Asymptotic properties of the estimator are derived in this context. Under regularity conditions the LASSO estimator exhibits the oracle property, and simulations suggest that it outperforms the least squares dummy variable estimator in terms of the root mean squared error of the estimated firm-level inefficiencies. An application of the LASSO to rice farm data suggests that the resulting subset of maximally efficient firms is comparable to the efficiency subsets calculated from the same data in Horrace and Schmidt (2000).

Keywords Panel Data, Fixed Effect Stochastic Frontier Model, L_1 Regularization, Zero Inefficiency, Ranking and Selection.

*Department of Economics, University of Arkansas, Fayetteville, AR 72701. Email: hj020@uark.edu. I thank William Horrace, Badi Baltagi, Yoonseok Lee, Christopher Parmeter and the participants at the 15th European Workshop on Efficiency and Productivity Analysis and the 28th annual meeting of the Midwest Econometrics Group for their valuable comments and suggestions. All errors are my own.

1 Introduction

Current estimators of the stochastic frontier (SF) model yield point estimates of firm-level efficiency, which (when ranked) imply that a single firm in the sample is most efficient. That is, SF model estimators do not allow for efficiency ties, yet there may be several firms in the sample tied for most efficient, and we would like to develop techniques to allow for this scenario. Current approaches adopt two-step methodologies to identifying a subset of efficient firms. In the first step, firm-level efficiencies (or equivalent measures) are estimated, and in the second step an 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 (2010), and Wheat, Greene and Smith (2014) estimate JLMS efficiency and then construct univariate intervals that imply statistical indistinguishability of firms with the largest estimates. Horrace (2005a) and Flores-Lagunes, Horrace, and Schnier (2007) extend this to multivariate intervals that account for the multiplicity inherent in the ranked estimates. Using these intervals, they develop selection procedures that produce a subset of most efficient firms at a pre-specified error rate. Horrace and Schmidt (2000) develop multivariate intervals for the semi-parametric SF model of Schmidt and Sickles (1984) for panel data. Despite the semi-parametric model, their inference technique relies on a parametric assumption on the distribution of estimated efficiencies.

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 adjust (shrink) the individual inefficiency estimates to reflect the presence of both efficient and inefficient firms in the sample. Using the parameter estimates, they compute individual posterior estimates of the probability of being fully efficient, and then, with a pre-specified cut-off, they assign each firm to the fully efficient regime or the inefficient regime. Unfortunately, Rho and Schmidt (2015) discuss an identification issue in this model. They point out that if there is little inefficiency in the sampled firms, it is hard to distinguish whether this is due to small variance of individual efficiencies or due to a large proportion of fully efficient firms, which leads to an observational equivalence in the likelihood function and a lack of identification of the model. Moreover, the ZISF model suffers from the same issues as the previously mentioned techniques; 1) it is parametric and 2) it is a 2-step procedure. We would like to develop models that are semi-parametric and identify a subset of efficient firms in a single step.

We propose a new one-step, semi-parametric procedure for identifying latent membership in a subset of

efficient firms using LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996).[1] Specifically, we develop estimation procedures which identify a subset of marginal effects and firm-level inefficiencies as exactly zero. The proposed adaptive LASSO estimation proceeds as least squared dummy variable (LSDV) estimation, but the object function is augmented with two penalty terms: the adaptively weighted shrinkage L_1 penalties for the input coefficients and for the firm-level inefficiencies.[2] Here, we estimate and penalize the inefficiency terms directly. However, we show that this is equivalent to estimating the firm-level fixed effects and then penalizing their differences from the firm with the largest effect in the sample. We also show the object function can be equivalently solved by 'within' transformed versions of estimation. The within transformation eliminates the time-invariant inefficiencies (and their associate fixed-effects), leading to a two-step revision of the LASSO estimation.

Asymptotics are 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 show that the proposed estimator possess the oracle property under regularity conditions. More precisely, the LASSO consistently selects model coefficients as $(N,T) \to \infty$. Conditional on selection consistency, the estimates for the selected marginal effects and the individual firm inefficiencies are $\sqrt{NT}$- and $\sqrt{T}$-consistent (respectively) in our setting whereas the estimates from standard estimation procedures, like LSDV, for common estimates are $\sqrt{NT}$- and $\sqrt{T/(\log N)^2}$-consistent (respectively) in the same setting. The efficiency difference in the estimators is more pronounced in the estimation of the common intercept (the maximum of the individual fixed effects): The LASSO estimator shows $\sqrt{\delta_0 NT}$ consistency, where δ_0 is the fixed proportion of fully efficient firms, while the LSDV estimator exihbits $\sqrt{T/(\log N)^2}$ consistency.[3] Therefore, the LASSO estimator of the common intercept converges faster. Consequently, the LASSO may outperform LSDV (in terms of root mean squared error of the estimated common intercept) even when T is small, and this is borne out in our simulation study.

We apply the LASSO to Indonesian rice farm data previously analyzed by Erwidodo (1990), and Horrace

[1]The LASSO has received much attention in the statistics literature. For example, Wang et al. (2007a) consider its application to autoregression, Caner (2009) considers generalized methods of moments, Zhu et al. (2010) consider its application to spatial autoregression, and Lee et al. (2016) consider the selection of break points in time series. Recently in economics, the shrinkage technique is used in select valid and relevant moments: Belloni et al (2012), Cheng and Liao (2015), and Caner et al (2016) among others.

[2]Lu and Su (2016) use the adaptive group LASSO technique to select regressors and factors in the regression context. If our fixed effects model can be regarded as a special case of a factor error structure model, then our model and the model in Lu and Su (2016) are closely related. Cheng et al. (2016) use the LASSO to identify a break in the factor model where they selected a group of factor loadings generated due to the break in factors, but their model is a pure factor model, and they uses the group LASSO method to estimate the model.

[3]The LASSO estimate for the intercept is estimated as a common intercept of the firms categorized as fully efficient by LASSO technique, so if we know the true model, we can use the $\delta_0 NT$ observations to estimate the intercept (δ_0 is a fixed parameter so it is not relevant when calculating convergence rates. We intentionally leave it in the expression of the convergence rate for the common intercept to stress out the role of δ in our estimation. In this case, we may redefine $\delta_0 N = N_\delta$ and think N_δ grows proportionally to N.) However, the convergence rate of the estimate for the intercept in the standard fixed effect SF model can't be established clearly because it relies on the maximum value of the estimated fixed effects to estimate the intercept and the rate the maximum value converges to the true intercept should be determined by the assumption for the distribution of the inefficiency. Park et al (1998) studies the asymptotic properties of the estimator for the common intercept in LSDV and shows that it has the convergence rate of $O_p(\frac{\log N}{\sqrt{T}} + \frac{1}{N})$ if a inefficiency has a shifted half normal or exponential distribution.

and Schmidt (1996, 2000) among others. The LASSO selects a subset of maximally efficient rice farms that is comparable in size and composition to the Gupta subset of Horrace and Schmidt (2000). However, the LASSO does so without multivariate inference on the efficiency estimates of 171 rice farms and without the distributional assumptions that it entails.

The technique developed for SF models may be more broadly applicable. Whenever we have a linear regression model with individual fixed effects, and the ranked fixed effects contain important information, the LASSO may be applied to produce 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 calculate a group of the best or worst students based on their individual-specific outcomes. Mutual fund performance may be another example. Practitioners may conveniently use the LASSO to calculate subsets of the best funds, based on average returns. Moreover, this type of "best and the rest" classification may be more important as dataset sizes grow (big data), because group-level classification may be more useful than individual ranking, when the number of comparisons is large.

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 computational issues such as 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 Adaptive LASSO for many efficient firms

2.1 Production Function

The fixed effect SF panel data model with time-invariant technical inefficiency due to Schmidt and Sickles (1984) is,

$$y_{it} = \alpha_0 + x_{it}^{'}\beta_0 + v_{it} - u_{0,i} \quad \text{for } i = 1, ..., N \text{ and } t = 1, ..., T \tag{2.1}$$

where y_{it} is the logarithm of scalar output of the i^{th} firm in the t^{th} period, α_0 is a scalar common intercept (common to all firms i), x_{it} is a $p \times 1$ input vector, β_0 is a $p \times 1$ corresponding parameter vector of marginal effects, and v_{it} is a two sided noise with $v_{it} \sim iid(0, \sigma_{0,v}^2)$ and $v_{it} \perp\!\!\!\perp x_{it}$. The $u_{0,i}$ are time-invariant firm-specific inefficiencies, which are treated as fixed effects. We assume $u_i \geq 0$ and independent of v_{it} and the inputs,

but do not impose a distributional assumption on the inefficiency distribution. In matrix form the model is

$$Y = \alpha_0 l_{NT} + X_1\beta_0 - X_2 U_0 + v \tag{2.2}$$

where l_{NT} is $NT \times 1$ vector with ones, $X_1 = [x_{it}^{'}]$, and $X_2 = I_N \otimes l_T$.

Standard LSDV estimation (or equivalent within estimation) proceeds as follows. Rewrite equation (2.1) as $y_{it} = \alpha_{0,i} + x_{it}^{'}\beta_0 + v_{it}$, where $\alpha_{0,i} = \alpha_0 - u_{0,i}$ are firm-specific fixed-effects. If $[X_1, X_2]$ is full column rank, then all the parameters of the model are identified, and we regress Y on X_1 and X_2 to get ordinary least squares (OLS) estimates $\hat{\beta}_0$ and $\hat{\alpha}_{0,i}$, respectively. The OLS estimates are consistent for β_0 (as N or $T \to \infty$) and $\alpha_{0,i}$ (as $T \to \infty$), respectively. Also, $\hat{\alpha_0} = max_j \hat{\alpha}_{0,j}$ is consistent for the common intercept as N and $T \to \infty$ because $min_j u_{0,j} \to 0$ and $max_j \alpha_{0,j} \to \alpha$ as $N \to \infty$ as long as the data generating process for $u_{0,i}$ allows u arbitrarily close to zero with positive density (Greene, 1980; Schmidt and Sickles, 1984). The individual firm inefficiencies are accordingly consistently estimated by $\hat{u}_{0,i} = \hat{\alpha_0} - \hat{\alpha}_{0,i}$ (as N and $T \to \infty$). In this case, $\hat{\alpha_0}$ represents maximal output in the population, and the individual $\hat{u}_{0,i}$ are interpretable as absolute inefficiencies.

In practice, there are many reasons why $[X_1, X_2]$ may not be full column rank, but a leading case is when X_1 contains time-invariant regressors. If so, the marginal effects of time-invariant regressors and the individual fixed-effects (and also the common intercept) are indecomposable within the model, which leads to a fundamental identification problem. See Greene (2005) and Feng and Horrace (2007) for detailed discussions about this issue and potential solutions. Another interesting case to consider is when X_1 contains indicator or categorical variables that vary over t. If so, the point estimate of α_0 will vary with the omitted reference groups of the categorical variables, but the individual inefficiency estimates will not. In this case the estimated common intercept still can serve as an instrument to identify the individual firm-level inefficiencies,[4] but the $\hat{u}_{0,i}$ has to be interpreted as relative efficiencies. In general, it would take fortuitous circumstances for α_0 to be identified, so $\hat{u}_{0,i}$ is almost always interpreted as relative efficiency.

For the LASSO version of the fixed effect SF model in equation (2.1), we impose the following sparsity assumption on $\theta_0 = (\alpha_0, \beta_0^{'}, U_0^{'})^{'}$.

Sparsity Assumption (i) $\beta_0 = (\beta_{0,A}^{'}, \beta_{0,A^c}^{'})^{'}$ and $n(\beta_{0,A}) = p_0 < p = n(\beta_0)$ where $n(M)$ is the number of elements in M, $A = \{j : \theta_{0,j} \neq 0\}$ represents the index set for the nonzero coefficients in θ_0, and $A^c = \{j : \theta_{0,j} = 0\}$ is defined similarly so that $\beta_{0,A}$ represents the true non-zero input coefficients. (ii) Similarly, $U_0 = (U_{0,A}^{'}, U_{0,A^c}^{'})^{'}$ and $n(U_{0,A}) < N$. Denote $\delta_0 = \frac{n(U_{0,A^c})}{N}$, then the assumption is equivalently

[4]This is related to the identification issue in wage gap decomposition, discussed in Horrace and Oaxaca (2001). Their concern is identification of the wage gaps across various labor markets, while our concern is identification of efficiencies across firms.

stated by $\delta_0 > 0$.[5] By construction, $\min_{j\in A}|u_{0,j}| = \eta > 0$, however, it is allowed $\eta \to 0$ as $(N,T) \to \infty$.

The sparsity assumption is common in the LASSO literature and implies only a subset of the regressors are relevant to the true model, which justify the use of penalized technique to recover the true model. In addition to the sparsity in the regressors as in the literature, we assume there's a sparsity in inefficiency. For practical purposes, the assumption will be met when the population under analysis contains many highly efficient firms. This model becomes the standard fixed effect SF model if $p_0 = p$ and $\delta_0 = 0$.[6] It becomes the neoclassic production model, which assume every firm is efficient, if $p_0 = p$ and $\delta_0 = 1$. We allow $\eta \to 0$ as $(N,T) \to \infty$ as we do not restrict the lower bound of inefficiency in SF models.

2.2 Adaptive LASSO estimator

The LASSO version of the fixed effect SF model can be estimated by either the "within estimation" or the Least Squares Dummy Variable (LSDV) estimation. We start with the LSDV and then show it is equivalent to 'within' transformed version of estimation with a two-step procedures.

The adaptive LASSO estimator for θ_0 is defined as

$$\begin{aligned}
\hat{\theta}(\Lambda,\Pi) &= [\hat{\alpha}(\Pi), \hat{\beta}(\Lambda)^{'}, \hat{U}(\Pi)^{'}]^{'} \\
&= \text{argmin}_{\alpha,\beta,U} \left\{ \sum_T \sum_N \{y_{it} - \alpha - x_{it}^{'}\beta + u_i\}^2 + \Lambda \sum_{j=1}^{p} \hat{\lambda_j}|\beta_j| + \Pi \sum_{k=1}^{N} \hat{\pi_k}|u_k| \right\} \\
&= \text{argmin}_{\theta} \left\{ (Y - X\theta)^{'}(Y - X\theta) + \Lambda \sum_{j=1}^{p} \hat{\lambda_j}|\beta_j| + \Pi \sum_{k=1}^{N} \hat{\pi_k}|u_k| \right\} \quad \text{where } u_i \geq 0
\end{aligned} \tag{2.3}$$

where $X = [l_{NT}, X_1, -X_2]$, and Λ and Π are positive tuning parameters for β and U, respectively. $\{\hat{\lambda_j}\}_{j=1}^{p}$ are some data-dependent weights for β, which are usually obtained from the absolute value of some consistent estimate. In this paper, we set $\hat{\lambda_j} = |\hat{\beta}_j^{LSDV}|^{-\gamma_\beta}$ and $\hat{\pi_k} = |\hat{u}_k^{LSDV}|^{-\gamma_u}$ with some $\gamma_\beta > 0$ and $\gamma_u > 0$.[7]

There are two things to be noted in (2.3). First, we are estimating α and U in one step, which is not feasible in the standard fixed effect SF model because of identification problem (theoretically) or perfect multicollinearity (in practice). This is feasible in this model due to the presence of efficient firms, which allow us to identify α and U separately while avoiding the multicollinearity problem. Second, it is also notable that we are using two different tuning parameters, and $\{\hat{\beta}_j^{LSDV}\}_{j=1}^{p}$ and $\{\hat{u}_k^{LSDV}\}_{k=1}^{N}$ that have different convergence rates. That is, the former is a $\sqrt{NT}$-consistent estimate of $\beta_{0,j}$ while the latter is a

[5] For analytic simplicity, without loss of generality, it is assumed the last $\delta_0 \times N$ of firms are fully efficient.

[6] Note that $\hat{u}_{0,i}$ is estimated from $\hat{u}_{0,i} = \hat{\alpha_0} - \hat{\alpha}_{0,i}$ in the standard SF model where $\hat{\alpha_0} = \max_{j=1}^{N}\hat{\alpha}_{0,j}$. We always have one relatively 100 % efficient firm from the model, however, it does not mean that it is zero inefficiency based on a absolute standard. It only become absolute zero inefficiency when $N \to \infty$.

[7] It should be noted that $\{\hat{\beta}_j^{LSDV}\}_{j=1}^{p}$ are from a preliminary estimation but $\{\hat{u}_k^{LSDV}\}_{k=1}^{N}$ may be obtained from the residuals after the first step estimation of (2.5)

$\sqrt{T/(\log N)^2}$-consistent estimate of $u_{0,k}$, which will be formally proved in lemma 4.1. We use the two tuning parameters because of the difference in asymptotic behaviors of the two estimator, $\hat{\beta}(\Lambda)$ and $\hat{U}(\Pi)$, which will be discussed in details in the assumption 4.2.

Remark 2.1 Our model is related to latent group structure models, in particular, the model in Su et. al. (2016). In Su et. al. (2016), the regression coefficients are heterogeneous across groups but homogeneous within a group, and group membership is unknown. Their methodology forces some of individual coefficients to have the same value by penalizing their difference from a group-specific coefficient value which is simultaneously estimated within their model.[8] We can show our model has the same features by reparameterizing (2.1) as $y_{it} = \alpha_{0,i} + x_{it}^{'}\beta_0 + v_{it}$ where $\alpha_{0,i} = \alpha_0 - u_{0,i}$ and $\alpha_0 \geq \alpha_{0,i}$. Then, using a similar penalized technique, the reparameterized model can be estimated by

$$[\hat{\alpha}(\Pi), \hat{\alpha}_I(\Pi)^{'}, \hat{\beta}(\Lambda)^{'}] = \text{argmin}_{\alpha,\alpha_i,\beta} \left\{ \sum_T \sum_N \{y_{it} - \alpha_i - x_{it}^{'}\beta\}^2 + \Lambda \sum_{j=1}^{p} \hat{\lambda_j}|\beta_j| + \Pi \sum_{k=1}^{N} \hat{\pi_k}|\alpha - \alpha_i| \right\} \quad (2.4)$$

where $\alpha \geq \alpha_i$, $\hat{\alpha}_I(\Pi) = [\hat{\alpha}_1(\Pi), ...\hat{\alpha}_N(\Pi)]$. (2.4) is not much different from (2.3) conceptually and computationally, as the inefficiencies in the fixed effect SF model are nothing but the distances from the largest fixed effect in the sample.[9] We are penalizing the differences between the largest fixed effect and the firm fixed effects in (2.4) to identify the subset of best firms, which implies our model has a classification feature similar to the one in Su et. al. (2016)[10]. In the same spirit, we can modify the sparsity assumption on inefficiencies: For $\alpha_0 \geq \alpha_{i,0}$, $\alpha_{i,0} = \alpha_0$ if $i \in BS_0$ where BS_0 is the true set of efficient firms who have α_0 (the leading group specific fixed effect) as for their fixed effects and $BS_0 \neq \emptyset$. In this regard, our model can be viewed as a technique for simultaneous classification and estimation of the firm-level efficiency.

With some algebra,[11] we can concentrate out the α and U in (2.3), which implies the above problem can

[8] Their methodology (also, our methodology) is related to the fused LASSO proposed by Tibshirani, Saunders, Rosset, Zhu, and Knight (2005) where the parameters of interest have an order in some meaning way and some parameters take the same value with the neighboring parameters. The fused LASSO encourages sparsity of the differences between the neighboring parameters by penalizing the differences. This technique is in particular useful when there are multiple changes in the parameter values along the natural order of the parameters (e.g detecting multiple structural changes in the time series setting; Harchaoui and Lévy-Leduc (2010), Chan, Yau, and Zhang (2014), and Qian and Su (2015)).

[9] We may show they possess the same asymptotic properties using similar arguments in our asymptotic analysis. We verified that they produces similar estimation results in finite sample simulations (the differences in the estimation results from the two object functions were less than 10^{-3} in many cases).

[10] There are two distinct differences between our model and theirs. First, our group membership is determined by firm fixed effects (or firm-level inefficiencies) whereas it is structural parameters that determine membership in their paper. Bonhomme and Manresa (2015) consider a latent group structure problem, where group membership is determined by group specific fixed effects, but their methodology relies on minimization of a least squares criterion with respect to all possible groupings without the LASSO technique. Moreover, they don't require our constraint on the fixed effects. Second, we are not estimating an arbitrary latent group structure, but we are identifying a "best and the rest" group structure by imposing the constraints $\alpha \geq \alpha_i$.

[11] We prove the equivalence between (2.3), and (2.5) - (2.6) in Appendix B.

be equivalently solved in two steps: In the first step solve

$$\begin{aligned}\hat{\beta}(\Lambda) &= \text{argmin}_{\beta}\left\{\sum_{T}\sum_{N}\{y_{it}^{*}-x_{it}^{'*}\beta\}^{2}+\Lambda\sum_{j=1}^{p}\hat{\lambda_{j}}|\beta_{j}|\right\} \\ &= \text{argmin}_{\beta}\left\{(Y-X_{1}\beta)^{'}Q(Y-X_{1}\beta)+\Lambda\sum_{j=1}^{p}\hat{\lambda_{j}}|\beta_{j}|\right\}\end{aligned} \tag{2.5}$$

where $y_{it}^{*}=y_{it}-\bar{y_{i}}$, $x_{it}^{'*}=x_{it}^{'}-\bar{x_{i}}^{'}$, and Q is a within transformation matrix such that $I_{N}\otimes(I_{T}-\frac{1}{T}i_{T}i_{T}^{'})$. Then the second step is

$$\begin{aligned}&[\hat{\alpha}(\Pi|\hat{\beta}(\Lambda),\hat{U}(\Pi|\hat{\beta}(\Lambda))^{'}]^{'} \\ &= \text{argmin}_{\alpha,U}\left\{\sum_{T}\sum_{N}\left\{y_{it}-\alpha-x_{it}^{'}\hat{\beta}(\Lambda)+u_{i}\right\}^{2}+\Pi\sum_{k=1}^{N}\hat{\pi_{k}}|u_{k}|\right\} \quad \text{where } u_{i}\geq 0 \\ &= \text{argmin}_{\alpha,U}\left\{(Y-X\theta_{\hat{\beta}(\Lambda)})^{'}(Y-X\theta_{\hat{\beta}(\Lambda)})+\Pi\sum_{k=1}^{N}\hat{\pi_{k}}|u_{k}|\right\}\end{aligned} \tag{2.6}$$

where $\theta_{\hat{\beta}(\Lambda)}=[\alpha,\hat{\beta}(\Lambda)^{'},U^{'}]^{'}$.[12] For notational simplicity, we denote $\hat{\alpha}(\Pi|\hat{\beta}(\Lambda))\equiv\hat{\alpha}(\Pi)$ and $\hat{U}(\Pi|\hat{\beta}(\Lambda))\equiv\hat{U}(\Pi)$. Figure 1 in chapter 3 of Hastie et al (2009) visually explains how the L_1 penalty often leads to zero estimates for some parameters. The LASSO shrinks the estimates toward 0, as the tuning parameter increases, and leads to exactly zero values for some parameters, because of a singularity at the origin. The initial LASSO by Tibshirani (1996) has no weight on its penalty term, so it shrinks each estimate equally, which leads to a trade-off between consistent estimation and consistent variable selection (Fan and Li, 2001; Zou, 2006). To address this, Zou (2006) proposes the adaptive LASSO which puts different weights onto each penalty term for each parameter, enabling adjustment of the degree of shrinkage for each parameter, using the information from preliminary consistent estimation. He proves the adaptive LASSO possess the oracle property. In this paper, we set $\hat{\lambda_{j}}=|\hat{\beta}_{j}^{LSDV}|^{-\gamma_{\beta}}$. If the true parameter is zero, the $|\hat{\beta}_{j}^{LSDV}|$ would be close to zero as $N\to\infty$ or $T\to\infty$, which, in turn, leads to $\hat{\lambda_{j}}\to\infty$, so we would be more likely to have a zero estimate for the parameter in this case. However, it is impossible to completely remove the bias in nonzero parameter estimates, so the tuning parameter should be large enough to select the zero parameters but not too large as to induce bias in the nonzero parameter estimates. Selection of the tuning parameters and their asymptotic conditions to achieve oracle property will be discussed in next sections.

[12] As we consider a two step procedure, we may adopt the hybrid estimation approach of Efron et. al. (2004) and Hui et. al. (2015) to reduce the bias from using the shrinkage technique in the first step. That is, instead of $\hat{\beta}(\Lambda)$, we compute the "within" estimate for β_0 under the model selected in the first step and use the unpenalized estimate for the second step.

3 Computation

3.1 Optimization algorithm

The L_1 penalty term in the object function has no second derivative at the origin, so we can't directly apply standard quadratic optimization algorithms (e.g. Newton-Raphson). Many alternative optimization algorithms have been developed: Least Angle Regression (Efron et. al., 2004), Local quadratic approximation(Fan and Li, 2001), Coordinate Descent Algorithm (Friedman et al, 2008), among others. Optimization of (2.5) is a standard LASSO problem, so we may use one of the algorithms for implementation of the first step.

In the second step, we wish to obtain a result that retains $\hat{u}_{0,j}(\Pi) \geq 0$ in order to be consistent with the model assumption of $u_{0,j} \geq 0$. Note that, if we impose a positive constraint on U in optimization procedure, then (2.6) becomes a standard constrained quadratic optimization problem with no singularity. Therefore, we may use one of the standard constrained optimization algorithms (e.g. Sequential Quadratic Programming) for implementation of (2.6). However, this may be computationally costly because the number of constraints in our problem is N. Alternatively, we propose a coordinate decent algorithm which produces almost the same estimation results as the standard constrained optimization but without computational cost. Using preliminary inefficiency ranking information among the firms from the LSDV estimation, this algorithm allows us to skip a large number of irrelevant optimization steps. The algorithm is implemented as follows. For a simplicity, we suppress the (Π) notation.

1. Using $\hat{\beta(\Lambda)}$ from the first step,[13] compute $\hat{\alpha}_i = \frac{1}{T}\sum_i y_{it} - x_{it}^{'}\hat{\beta(\Lambda)}$ and $\hat{u}_i = \max_{j=1}^{N}\hat{\alpha}_j - \hat{\alpha}_i$ for all i. Let $\hat{\alpha}_{[1]} \leq \hat{\alpha}_{[2]} \leq ... \leq \hat{\alpha}_{[N]}$ be the rankings of the $\hat{\alpha}_i$, so $\hat{\alpha}_{[N]} = \max_{j=1}^{N}\hat{\alpha}_j$. Similarly, let $\hat{u}_{[N]} \leq \hat{u}_{[N-1]} \leq ... \leq \hat{u}_{[1]}$ be the rankings of the $\hat{u}_i$, so $\hat{u}_{[N]} = \min_{j=1}^{N}\hat{u}_j$. We set the initial value for α to $\hat{\alpha}_{[N]}$. Denote the current values for $\hat{u}_{[i]}$ and $\hat{\alpha}$ as $\hat{u}_{[i]}^{(0)}$ and $\hat{\alpha}^{(0)}$. Note that as we set $\hat{\alpha}^{(0)} = \hat{\alpha}_{[N]}$, we have one fully efficient firm, $\hat{u}_{[N]}^{(0)} = 0$, now.

2. For a given Π, sequentially check the KKT condition for the second best firm, the third best firm.... That is, check the sign of $\Delta_{[N-i]} = \hat{u}_{[N-i]}^{(0)} - \Pi\frac{\hat{\pi}_{[N-i]}}{2T}$ from $i = N-1$ to 1.

   (a) IF $\Delta_{[N-i]} \leq 0$, update $\hat{u}_{[N-i]}^{(0)}$ as $\hat{u}_{[N-i]} = 0$, and update $\hat{\alpha}^{(0)}$ as $\hat{\alpha} = \frac{1}{i+1}\sum_{k=0}^{i}\hat{\alpha}_{[N-i]}$. As we have new $\hat{\alpha}$ (the frontier parameter), we update the rest of the inefficiencies (from $[N-1-i]$ to $[1]$) as $\hat{u}_{[N-i-j]} = \hat{u}_{[N-i-j]}^{(0)} - (\hat{\alpha}^{(0)} - \hat{\alpha})$ for $j = 1, ..., N-i-1$. Then, go back to 2 and check next firm's KKT condition.

   (b) IF $\Delta_{[N-i]} > 0$, update $\{\hat{u}_{[N-i]}^{(0)}, ..., \hat{u}_{[1]}^{(0)}\}$ as $\hat{u}_{[N-k]} = \hat{u}_{[N-k]}^{(0)} - \Pi\frac{\hat{\pi}_k}{2T}$ for $k = i, ..., N-1$. Repeat

[13] As mentioned eariler, we may use the "within" estimator for β_0 from the model selected in the first step.

below **LOOP** until the absolute difference in the estimation results in two consecutive steps is smaller than a pre-specified threshold and then report the results.

LOOP

i. Update $\hat{\alpha}^{(0)}$ as $\hat{\alpha} = \hat{\alpha}^{(0)} - \frac{1}{N}\sum_{k=i}^{N-1}(\hat{u}^{(0)}_{[N-k]} - \hat{u}_{[N-k]})$

ii. Update $\{\hat{u}^{(0)}_{[N-i]}, ..., \hat{u}^{(0)}_{[1]}\}$ as $\hat{u}_{[N-k]} = \left(\hat{u}^{(0)}_{[N-k]} - (\hat{\alpha}^{(0)} - \hat{\alpha})\right)_+$ for $k = i, ..., N-1$ where $(x)_+ = x$ if $x \geq 0$ and $= 0$ otherwise.

We provide a figure in Appendix D illustrating the above procedure. This coordinate decent algorithm uses the convexity of the object function and the preliminary inefficiency ranking at the same time, enabling us to reach the minimum of the object function quickly. We compare the series of estimation results between this algorithm and the Sequential Quadratic Programming (SQP) algorithm in Matlab in Appendix C. We find that the two algorithms generally produce similar results, but the new algorithm is much faster than SQP.[14]

In the algorithm, the LOOP is necessary to optimize the object function, however, we can see that it shrinks $\hat{\alpha}(\Pi)$ as well. This is not desirable because it may slow down the convergence rate of $\hat{\alpha}(\Pi)$ and, in turn, it may induce bias on $\hat{U}(\Pi)$ when T is small. In order to prevent this, we intentionally skip the LOOP in the implementation of our algorithm.[15] The asymptotic results are derived from this modified algorithm and we use it for the simulation study and empirical exercises.

3.2 Tuning parameter

The performance of the adaptive LASSO estimator relies on an appropriate selection of the tuning parameters, and the CV and AIC criteria have been used in the LASSO literature. However, they lead to inconsistent model selection; too many nonzero estimates. Wang et al (2007b) shows that the tuning parameters based on a BIC-type criterion can identify the true model consistently. Therefore, in this paper, we consider the BIC type criteria for the selection of the two tuning parameters such that[16]

$$(\Lambda^*, \Pi^*) = \text{argmin}_{\Lambda,\Pi} \log \hat{\sigma}^2(\hat{\theta}(\Lambda, \Pi)) + \frac{|\hat{\beta}(\Lambda)| \log(NT) + |\hat{U}(\Pi)| \log(T)}{NT} \tag{3.1}$$

where $\hat{\sigma}^2(\hat{\theta}(\Lambda, \Pi))$ is the mean squared error based on Λ and Π. The criterion of Wang et al (2007b) is a special case with $T = 1$. Equation (3.1) can be implemented using a two-dimensional grid search: 1) for

[14] For one replication with a sample size $(N, T) = (20, 10)$, the new algorithm took on average 1.5 second whereas the standard algorithm took 345 seconds. This gap will be pronounced as N increases.

[15] This modification can be viewed as a bias correct procedure. The modified algorithm helps to achieve better asymptotic and simulation results, which we shall see in subsequent sections.

[16] We also experimented various types of selection criterions in the simulation study (e.g ERIC: Hui et al (2015) and IC_{p1}: Bai and Ng (2002)) and found (3.1) worked best in various panel structure. The two other criterions tended to select more sparse model than (3.1), however, the difference in the model selection between the criterions were not big.

every possible Λ, implement the first step and compute $\hat{\beta}(\Lambda)$; 2) with the $\hat{\beta}(\Lambda)$s, implement the second step and compute $\hat{U}(\Pi|\hat{\beta}(\Lambda))$; 3) choose Π^* based on (3.1) for every Λ; and 4) choose Λ^* based on (3.1) using the results from the previous step.[17]

4 Asymptotic Theory

Asymptotics are for the case $(N,T) \to \infty$. Our analysis below builds on Zou (2006), and Zou and Zhang (2009), among others. Let LSDV estimates be denoted as $\hat{\theta}(0) = [\hat{\alpha}(0), \hat{\beta}(0)', \hat{U}(0)']'$, where $\hat{\alpha}(0) = \max_i \hat{\alpha}_i(0)$, $\hat{\alpha}_i(0)$ are the LSDV estimates for individual intercepts, which is $(X_2'X_2)^{-1}X_2(Y - X_1\hat{\beta}(0)))$, and $\hat{u}_i(0) = \hat{\alpha}(0) - \hat{\alpha}_i(0)$. We now discuss consistency and the rate of convergence of the LSDV estimator in our setup. For this, we assume

Assumption 4.1 For all $i = 1, ..., N, t = 1, ..., T, j = 1, ..., P$ in the model matrix X, $\max_{it,j}|x_{it,j}| < \infty$ w.p.a 1. And, for any model ω identified in the interval $[\Lambda_{min}, \Lambda_{max}]$ and $[\Pi_{min}, \Pi_{max}]$, we have $c_1 \leq eig_{min}(\frac{1}{T}X_\omega'X_\omega) < eig_{max}(\frac{1}{NT}X_\omega'X_\omega) \leq c_2$ where c_1 and c_2 are some positive constants, and $eig_{min}(\cdot)$ and $eig_{max}(\cdot)$ denote minimum and maximum eigenvalues of some positive definite matrices, respectively.

Assumption 4.1 requires the regressor matrix to be well behaved. That is, the minimum eigenvalue grows by T whereas the maximum eigenvalue grows by NT. The assumption implies $c_1 \leq eig_{min}\left(\frac{1}{NT}X_{1,\omega}{}'QX_{1,\omega}\right) \leq eig_{max}\left(\frac{1}{NT}X'_{1,\omega}QX_{1,\omega}\right) \leq c_2$ because $X'_{1,\omega}QX_{1,\omega}$ is a submatrix of $X_\omega'X_\omega$. Usually, the upper bounds of the estimation efficiency of the LASSO estimator is that of the baseline estimator based on the correct model. However, the adaptive LASSO estimator and the LSDV estimator in our setup exhibits different convergence rates for the common intercept and the inefficiency estimates, as will be shown below. First, we derive the convergence rate for LSDV estimator.

Lemma 4.1 Under Assumption 4.1, $E((\hat{\alpha}(0) - \alpha_0)^2) = O(\frac{(\log N)^2}{T})$, $E(||\hat{\beta}(0) - \beta_0||_2^2) = O(\frac{1}{NT})$, and $E((\hat{u}_i(0) - u_{0,i})^2) = O(\frac{(\log N)^2}{T})$ for $\forall i$ as $(N,T) \to \infty$.

Proof of this lemma and other lemmas and theorem is contained in Appendix A. Lemma 4.1 shows that each element in $\hat{U}(0)$ is $\sqrt{T/(\log N)^2}$ consistent. In the standard fixed effect SF model, it is not assumed

[17] One may simplify the computation by using the criteria sequentially such that: For the selection of Λ,

$$\Lambda^* = \text{argmin}_\Lambda \log \hat{\sigma}^2(\hat{\beta}(\Lambda)) + |\hat{\beta}(\Lambda)|\frac{\log NT}{NT} \tag{3.2}$$

where $\hat{\sigma}^2(\hat{\beta}(\Lambda))$'s degree of freedom is $N(T-1)$. And for the selection of Π,

$$\Pi^* = \text{argmin}_\Pi \log \hat{\sigma}^2(\hat{U}(\Pi)|\hat{\beta}) + \frac{|\hat{\beta}(\Lambda)|\log(NT) + |\hat{U}(\Pi)|\log(T)}{NT} \tag{3.3}$$

There may be a loss of precision from this method but it will reduce the computational cost significantly.

that at least one firm in the sample is efficient, so it requires that $N \to \infty$, to ensure that we sample the first efficient firm. In our setup, we assume that we have at least one efficient firm, so practically speaking, only $T \to \infty$ is necessary. Next, for the oracle proof, we define

$$\hat{\theta}_A(\Pi, \Lambda) = \text{argmin}_\theta \left\{ (Y - X_A \cdot \theta)'(Y - X_A \cdot \theta) + \Lambda \sum_{j \in A} \hat{\lambda_j} |\beta_j| + \Pi \sum_{k \in A} \hat{\pi_k} |u_k| \right\} \tag{4.1}$$

where X_A consists of columns of X that correspond to the elements in A.[18] We derive the consistency and rate of convergence of $\hat{\theta}_A(\Lambda, \Pi)$ below. We require following assumptions.

Assumption 4.2 Denote $T^* = \frac{T}{(\log N)^2}$. (i)$\lim_{T,N\to\infty} \frac{\Lambda}{\sqrt{NT}} = 0$ and $\lim_{T,N\to\infty} \frac{\Lambda}{\sqrt{NT}} (NT)^{(\gamma_\beta)/2} = \infty$ with $\gamma_\beta > 0$ (ii)$\lim_{T,N\to\infty} \frac{\Pi}{\sqrt{T^*}} = 0$ and $\lim_{T,N\to\infty} \frac{\Pi}{\sqrt{T}} \cdot T^{*\gamma_u/2} = \infty$ with $\gamma_u > 0$ (iii)$\beta_{0,A}$ is bound from below by c $>>$ 0, and $U_{0,A}$ and $\lim_{T,N\to\infty} \left(\frac{\sqrt{T^*}}{\Pi} \right)^{1/\gamma_u} \cdot \eta = \infty$ where $\eta = \min(|u_{0,k}|)$

Similar assumptions can be found in Zou (2006), and Zou and Zhang (2009). Assumption 4.2 is crucial for the oracle property, because it controls the behavior of the tuning parameters, Λ and Π, so they can select the zero coefficients properly without producing asymptotic bias in the nonzero coefficient estimates. Assumption 4.2 (ii) implies a condtion that T has to grow faster than $(\log N)^2$ so that $\frac{(\log N)^2}{T} \to 0$, however, the restriction is not strong as it covers many panel structure. Assumption 4.2 (iii) restricts the convergence speed of the nonzero coefficients to zero so that they can be distinguished from the zero coefficients by the estimation procedure. This is important in our context as we do not impose any lower bound for inefficiencies in the SF literature.

Lemma 4.2 Under Assumption 4.1 and 4.2, $E((\hat{\alpha}(\Pi) - \alpha_0)^2) = O(\frac{1}{\delta_0 NT})$, $E(||\hat{\beta}_A(\Lambda) - \beta_{0,A}||_2^2) = O(\frac{1}{NT})$ and $E((\hat{u}_{A,i}(\Pi) - u_{0,A,i})^2) = O(\frac{1}{T})$ for $\forall i$ as $(N, T) \to \infty$.

This Lemma shows that if we select the correct model, the fixed effect LASSO estimator is an consistent estimator for θ_0. The proof of this lemma in Appendix A shows that the mean square error of the LASSO estimator can be decomposed into two parts: the first part due to the penalty terms, and the second part due to the two-sided random error. Under these assumptions, the first part vanishes faster than the second, so estimation consistency is achieved.

From this lemma, we observe that $\hat{\alpha}(0)$ and $\hat{\alpha}(\Pi)$ have different convergence rates. LSDV uses only T observation to estimate α_0 and the max operator further slows down the rate (Park et. al., 1998). The $\hat{\alpha}(\Pi)$ is estimated as a common intercept of the firms categorized as efficient by the LASSO, so if we knew the true model, we could use the $\delta_0 NT$ observations to estimate α_0. This result, in turn, leads to the convergence

[18] The two steps implementation can also be applied to this problem (this time, the $Q = Q_A$ where Q_A is a within transformation matrix corresponding $X_{2,A}$)

rate difference between $\hat{U}(0)$ and $\hat{U}_A(\Pi)$. However, we need to note that the difference in the convergence rates can only be observed after achieving selection consistency. That is, only when we can identify the true group of efficient firms will $\hat{\alpha}(\Pi)$ show a faster convergence rate. Otherwise, estimation error in $\hat{U}_A(\Pi)$ will be transfered into the estimation of α_0, and the optimal convergence rate will not be achieved.

We will proceed to the oracle proof. First, we derive a useful lemma for the oracle of $\hat{\theta}(\Lambda, \Pi)$. This lemma shows that $\hat{\theta}(\Lambda, \Pi)$ estimates all the elements in A^c as zero w.p.a 1. as $(N, T) \to \infty$.

Lemma 4.3 Under Assumption 4.1 and 4.2, $[\hat{\alpha}(\Pi), (\hat{\beta}_A(\Lambda)', 0'), (\hat{U}_A(\Pi)', 0')]'$ is the solution to the minimization problem of (2.3) w.p.a 1 as $(N, T) \to \infty$.

This lemma tells us that asymptotically $\hat{\theta}(\Lambda, \Pi)$ works as if it knows the true model by estimating the zero coefficients exactly as zero. As in the proof of this lemma in the Appendix, the assumptions on Λ, Π, and η are crucial for this asymptotic characteristic of $\hat{\theta}(\Lambda, \Pi)$.

Assumption 4.3 $\frac{X_{1,A}' Q X_{1,A}}{NT} \to_p \Sigma_{A,1}$, $\delta_0 \frac{l_\delta' X_{1,A}}{\delta_0 NT} (\frac{X_{1,A}' Q X_{1,A}}{NT})^{-1} \frac{X_{1,A}' l_\delta}{\delta_0 NT} \to_p \Sigma_{\alpha *}$, and the error terms and regressors satisfy the regularity conditions for Central Limit Theorem.

The next theorem confirms the oracle property of the adaptive LASSO estimator.

Theorem (Oracle Property) Under Assumption 4.1, 4.2 and 4.3, the LASSO estimator for θ_0, $\hat{\theta}(\Lambda, \Pi)$, has the oracle property. That is, the estimator satisfies:

1. Consistency in selection: $Pr(\{j : \hat{\theta}(\Lambda, \Pi)_j \neq 0\} = A) \to 1$ as $(N, T) \to \infty$
2. Asymptotic normality
   1) $\sqrt{NT}(\hat{\beta}_A(\Lambda) - \beta_{0,A}) \to_d N(0, \sigma_{v,0}^2 \Sigma_{A,1}^{-1})$
   2) $\sqrt{\delta_0 NT}(\hat{\alpha}(\Pi) - \alpha_0) \to_d N(0, \sigma_{v,0}^2 \Sigma_\alpha)$
   3) $\sqrt{T}(\hat{u}_{i,A}(\Pi) - u_{i,0}) \to_d N(0, \sigma_{v,0}^2)$

w.p.a 1 as $(N, T) \to \infty$ where Σ_α is from $1 + \delta_0 \frac{l_\delta' X_{1,A}}{\delta_0 NT} (\frac{X_{1,A}' Q X_{1,A}}{NT})^{-1} \frac{X_{1,A}' l_\delta}{\delta_0 NT} \to_p \Sigma_\alpha$.

This theorem establishes the selection consistency and asymptotic normality of the adaptive LASSO estimator. In other words, asymptotically speaking, we can select the right model and identify a set of efficient firms without loss of estimation efficiency in this procedure. One interesting thing is that the asymptotic distribution for the estimate of inefficiency is somewhat different from the one derived in Park et al (1998). Their results are based on the standard SF model and imply that the asymptotic variance of each inefficiency estimate is at least $2 \times \sigma_{v,0}^2$ and there will be more variation due to the uncertainty over

whether one of the firm in the sample achieves full efficiency or not. However, in our model, the assumption of at least one fully efficient firm [19] and the faster convergence rate of $\hat{\alpha}(\Pi)$ than that of $\hat{u}_i(\Pi)$ significantly reduce uncertainty in the estimation of $\hat{u}_i(\Pi)$.

5 Simulations

5.1 Setup

In this section, we study the finite sample performance of the estimator. We set up the model by defining $\alpha_0 = 1$, $\beta_0 = [\beta, \beta, \beta, 0, 0, 0, 0, 0]'$ with $\beta = 1$, $x_{it} \sim N(0, \Sigma)$ with the $(i,j)-$th element of Σ set to $0.5^{|i-j|}$, and $v_{it} \sim N(0,1)$. We assume 30% of firms in the sample are fully efficient firms ($\delta_0 = 0.3$), and in every simulation each nonzero individual inefficiency is identically and independently generated from an exponential distribution $\frac{1}{\sigma_u} e^{-u_{it}/\sigma_u}$.[20] We experiment with $\sigma_u \in \{1, 2, 4\}$. As σ_u gets smaller, the selection problem becomes more difficult because the probability of small inefficiency draws will be high, making it more difficult for the LASSO to distinguish them from zero. This would be particularly difficult when the sample size is small as it is likely that Assumption 4.2 (ii) in the asymptotic analysis is violated. Figure 1 shows the PDFs of inefficiency for each σ_u value (left) and an example of draws from each PDF (right). We can clearly see that inefficiencies are densely packed near zero when $\sigma_u = 1$.

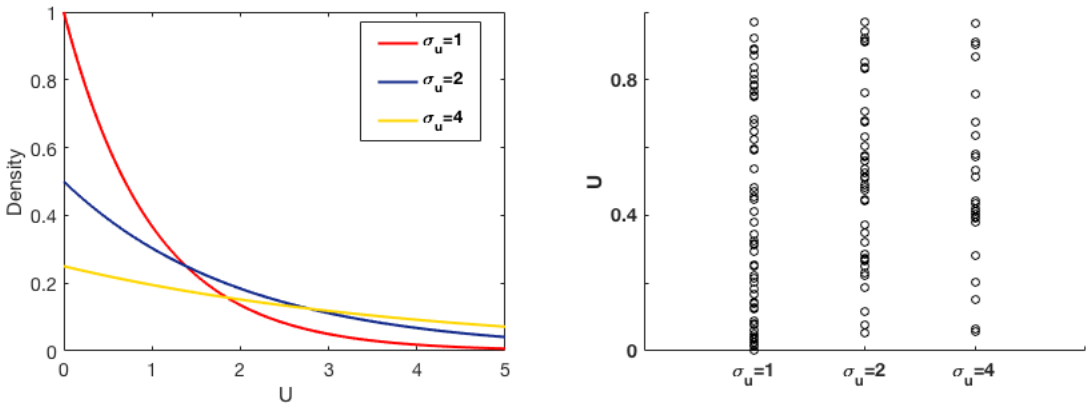


Figure 1: PDFs of inefficiency with different σ_u values and an example of draws from each PDF.

We set $\gamma_\beta = 1$ and $\gamma_U = 2$ [21], and the optimal tuning parameters are selected by (3.1) from a two dimensional grid search over $10^{linspace(\log 10(10^{-4}/NT):1:50)} \times NT$ for Λ and $10^{linspace(\log 10(10^{-4}/T):1:250)} \times T$ for Π where $linspace(a:b:c)$ is a row vector of c evenly spaced points between a and b.[22] We simulate each

[19] This also allow us to derive a deterministic form of asymptotic distribution for the estimate of α_0

[20] We add an arbitrary small number (e.g. 0.01) to each inefficiency draw to ensure they are not zero.

[21] We set $\gamma_U = 2$. However, we are free to choose the value of γ_U as long as it is positive. From the asymptotic analysis, we can see that setting a higher value for γ_U ensures the LASSO estimates zero coefficients as zero, but also increases the probability of estimating (small) nonzero coefficients as zero. Therefore, in empirics γ_U should be determined in light of this trade-off.

[22] We use a denser grid for Π than that for Λ because there are many inefficiency draws close to zero.

model 1,000 times with twelve combinations of $N \in \{100, 200, 1000\}$ and $T \in \{10, 30, 50, 70\}$.

5.2 Results

In what follows we only discuss results on $\hat{U}(\Pi)$, because it is our focus. Results on $\hat{\beta}(\Lambda)$ are in Appendix E. We report two types of statistics:

1. Estimation accuracy: a) square root of mean squared error (RMSE : $\sqrt{E(\hat{U}(\Pi) - U_0)^2}$), b) $\hat{\alpha}(\Pi)$ and $\hat{\alpha}(0)$, and c) Rank correlation between $U_{0,A}$ and $\hat{U}(\Pi)_A$.

2. Selection accuracy: a) Pr_{U_A}, b) $Pr_{U_{A^c}}$, c) $\hat{\delta}_0$, and d) Max $\tilde{U_A}$,

where Pr_{U_A} is the probability of yielding nonzero estimates for U_A[23]; $Pr_{U_{A^c}}$ is the probability of yielding zero estimates for U_{A^c}; $\hat{\delta}_0$ is the proportion of the firms estimated as efficient; Max $\tilde{U_A}$ is the maximum of U_A that are estimated as zero, that represents the worst case selection error when a model is underfitted.[24] Table 1 and Figure 2 present the estimation accuracy results and Table 2 and Figure 3 present the selection accuracy results.

5.2.1 Estimation accuracy of $\hat{U}(\Pi)$

Table 1 reports and compares the three types of estimation accuracy results from LASSO and LSDV. The results can be summarized as follows:

- As T and σ_u increase, the RMSE from the LASSO decreases, but the effect of σ_u on RMSE is small. This is due to the fact that σ_u, which determines the frequency of near zero inefficiencies, significantly affects the selection performance as shown in Table 2. However, it may not be case for the RSME, because small inefficiency draws are already near zero, so that they do not contribute much to the RMSE.

- The LASSO outperforms LSDV in terms of RMSE. Differences in the estimation error among the zero inefficiency draws may be one explanation for the difference. However, the main explanation is the persistent overestimation of α_0 in LSDV. Figure 2 presents the distribution of the estimates of α_0 from the LASSO (solid line) and LSDV (dashed line). The distributions from the LASSO over the replications are centered close to the true value even when T and σ_u are small, and the variation in the distribution is decreasing significantly as N or T increase. However, those from LSDV are consistently displaced away from the true value. This is due to the fact that the LASSO estimator for α_0 converges

[23] Pr_{U_A} is computed from averaging the percentage of nonzero estimates for U_A in each replication. $Pr_{U_{A^c}}$ is computed from the same manner.

[24] In the literature, underfitting means the case when we estimate one of the nonzero coefficients as zero.

Table 1: Estimation accuracy for $\hat{U}(\Pi)$

		RMSE		$\hat{\alpha}$ ($\alpha_0 = 1$)		Ranking correlation	
(N, T)	σ_u	LASSO	LSDV	LASSO	LSDV	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)

NOTE: The entries are the average values for each measure over 1,000 replications and their corresponding standard deviations in next row in parentheses. Rank correlations are computed only among the inefficiencies whose true values are nonzero. That is, $Rcorr = corr(R(U_{0,A}), R(\hat{U}(\Pi)_A))$ where $R(\cdot)$ is a mapping from estimates to rankings. Similarly for LSDV.

faster than the LSDV estimator as shown in the asymptotic analysis, and the max operator that LSDV uses to estimate α_0 tends to pick up the most biased individual intercept estimate (α_i) of the zero inefficiency firms. In short, in the presence of a group of zero inefficiency firms, the max operator

produces a biased estimate for α_0, which, in turn, leads to a significant bias in the estimation of the inefficiencies in LSDV.

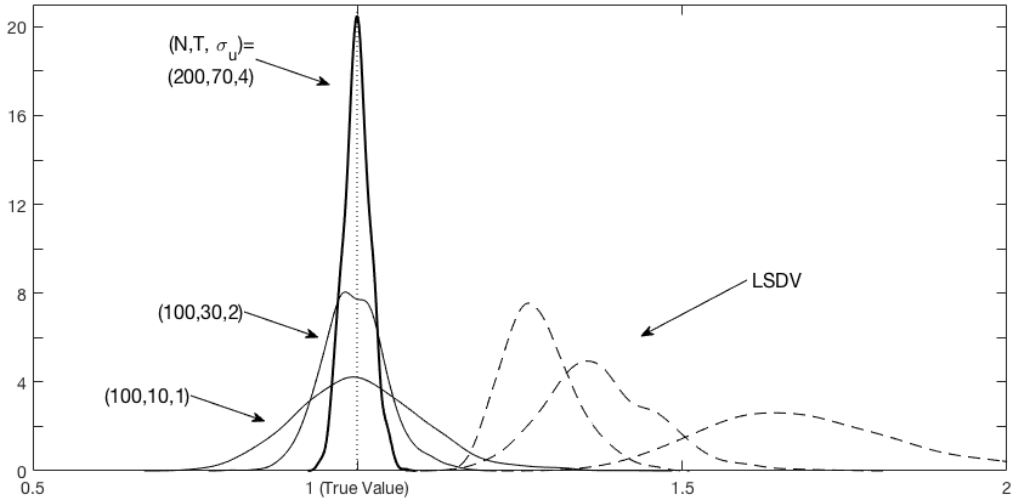


Figure 2: Distribution of estimates for α_0 from LASSO and LSDV

- The LASSO and LSDV show similar rank correlation results.[25] We need to note that these results are computed only among the nonzero inefficient firms, and the LASSO achieves this even after selecting fully efficient firms, which implies the LASSO preserve the original ranking well when compared to LSDV.

5.2.2 Selection accuracy of $\hat{U}(\Pi)$

Table 2 present the results for Pr_{U_A}, $Pr_{U_{A^c}}$ and $\hat{\delta}_0$. For the reader's convenience, we visualize the selection results in Figure 3 using the results from 12 cases with $N = 100$. The left panel is for Pr_{U_A}, and the right panel is for $Pr_{U_{A^c}}$, and the variables on each axises are given by $(x, y, z) = (\sigma_u, T,$ probability). Note that the graph is drawn continuously but the results are not. The continuous planes are generated from connecting 12 coordinates and we add a color to each probability level to show the general trend of the selection performance depending on (σ_u, T).

We can see that when T and σ_u are small, the LASSO incorrectly estimates many nonzero inefficiencies as zeros. However, the under fitting problem gets improved as T or σ_u increases. The under fitting or over fitting problem shouldn't be exaggerated because most of the firms incorrectly estimated as zero inefficiency

[25] Moreover, the LASSO shows slightly better results in many cases. The rank correlations are computed only from the nonzero inefficient firms, so the differences only come from the penalized effects. The difference is the biggest when T and σ_u are small, which is when we have a large uncertainty in the inefficiency estimates. This result may imply that the penalized technique may improve estimation of the nonzero inefficiencies as well, when LSDV is not reliable.

Table 2: Selection accuracy for $\hat{U}(\Pi)$

	$\sigma_u = 1$				$\sigma_u = 2$				$\sigma_u = 4$			
(N, T)	Pr_{U_A}	$Pr_{U_{A^c}}$	$\hat{\delta}$	Max $\tilde{U_A}$	Pr_{U_A}	$Pr_{U_{A^c}}$	$\hat{\delta}_0$	Max $\tilde{U_A}$	Pr_{U_A}	$Pr_{U_{A^c}}$	$\hat{\delta}$	Max $\tilde{U_A}$
(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)

NOTE: Pr_{U_A} represents the probability of yielding nonzero estimates for U_A and Pr_{U_A} is computed from averaging the percentage of nonzero estimates for U_A in each replication; $Pr_{U_{A^c}}$ represents the probability of yielding zero estimates for U_{A^c}; $\hat{\delta}^*$ represents the average proportion of the firms estimated as efficient; Max $\tilde{U_A}$ represents the maximum of U_A that is estimated as zero.

Table 3: My caption

would have near zero inefficiency. The small values of Max $\tilde{U}_A$ in Table 2 imply only the firms near the threshold of zero inefficiency may be incorrectly categorized as fully efficient. This may not be a serious problem in practice.

More importantly, it is impressive that even when T is small, including the $(N, T) = (1000, 10)$ case, the Pr_{U_A} and $Pr_{U_{A^c}}$ are close to 1 if $\sigma_u = 4$, which implies selection performance is more dependent on the distribution of the firms' inefficiency than the size of T in finite samples. This gives us an important implication, that our model can be used in various panel structures, not limited to the case where T is large, as long as there are not too many near zero inefficiencies, and our primary interest lies in identification of fully efficient firms rather than individual inefficiency estimates.

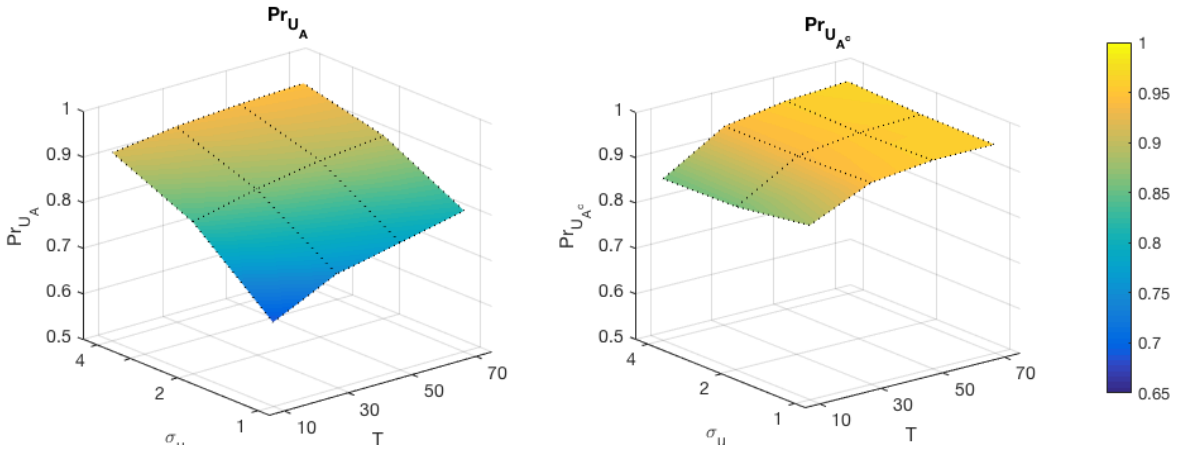


Figure 3: Visualization of the selection performance ($N = 100$)

6 Empirical application

6.1 Comparison with "Ranking and Selection" (R&S) procedure

In this section, we apply our LASSO model to the rice farm data previously analyzed by Erwidodo (1990), Horrace and Schmidt (1996) and Horrace and Schmidt (2000), among others. In our context, the LASSO is designed to select a group of efficient firms, BS_0. The idea of selecting a subset of best firms is related to the R&S literature. R&S proceeds as follows. Suppose we estimate the LSDV model, and it yields $\tilde{\alpha}_i$ for $i = 1, ..., N$. R&S is an inferential decision rule that selects some subset of the populations that contain the population with largest (best) value of α_i with some pre-specified error rate. If in truth $\alpha_i = \alpha_0 - u_i$, this is equivalent to selecting populations with u_i closest to zero at the pre-specified error rate. If so, the

connection between the LASSO and the R&S should be clear. The LASSO selects some subset of u_i to be zero in our model, while R&S select some subset of u_i closest to zero in a statistical sense. For more details of R&S procedure and its applications to economic problems, see Horrace and Schmidt (1996) and Horrace and Schmidt (2000).

To compare the two methodologies, we use data on 171 rice farms in Indonesia, observed for three wet and three dry seasons from six different villages. For a complete discussion of the data see Erwidodo (1990, unpublished manuscript). The empirical model is a standard Cobb-Douglas (loglinear) production function. Inputs to the production of rice included in the data set are seed (kg), urea (kg), trisodium phosphate (TSP) (kg), labor (labor-hours), and land (hectares). Output is measured in kilograms of rice. The data also include dummy variables. DP equals 1 if pesticides were used and 0 otherwise. DV 1 equals 1 if high yield varieties of rice were planted, and DV2 equals 1 if mixed varieties were planted; the omitted category represents that traditional varieties were planted. DSS equals 1 if it was a wet season. Since our focus is on the efficiency estimates, regression results are not presented here (see Horrace and Schmidt, 2000).[26]

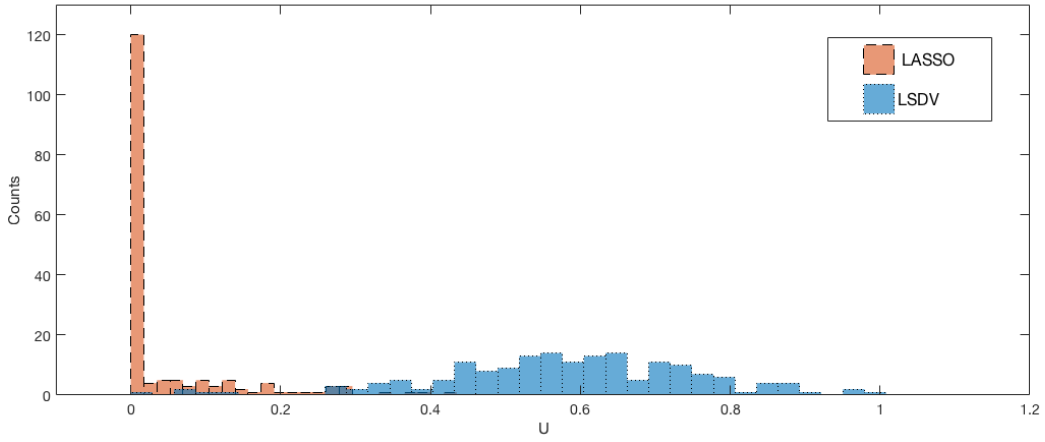


Figure 4: Distribution of the rice farm inefficiencies

6.2 Results

The LASSO estimates 69.6 % of farms (119 out of 171 farms) as efficient.[27] The distribution of the inefficiencies are reported in Figure 4. In Figure 4, the blue histogram represents the distribution of the inefficiencies from LSDV and the orange one represents that from LASSO where the 69.6 % of mass is concentrated at 0. We performed R&S procedure with error rate of 0.05, and found that 67% of rice farms (115 out of 171 farms) are in the subset of the best farms. Figure 5 matches the firm IDs estimated as fully efficient

[26] As in the simulations, we set $\gamma_\beta = 1$ and $\gamma_u = 2$. The model is estimated after standardizing the input variables.
[27] The production function is exactly the same as Horrace and Schmidt (2000), and the LASSO selects the full model.

by LASSO (yellow) and those included in the best subset by the R&S (red). The result is showing that all of the farms in the best subset by R&S are estimated as zero inefficiency in the LASSO. 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 whole model, which is similar to an F-test procedure. The only difference is that the LASSO is based on the BIC criterion, whereas R&S is based on a pre-specified error rate and the multivariate confidence intervals it implies. The multivariate confidence intervals are based on all $N(N-1)$ differences, $\tilde{\alpha}_i - \tilde{\alpha}_{i \neq j}$. If we use a different BIC criterion or a different error rate, we would get a different set of results. [28]

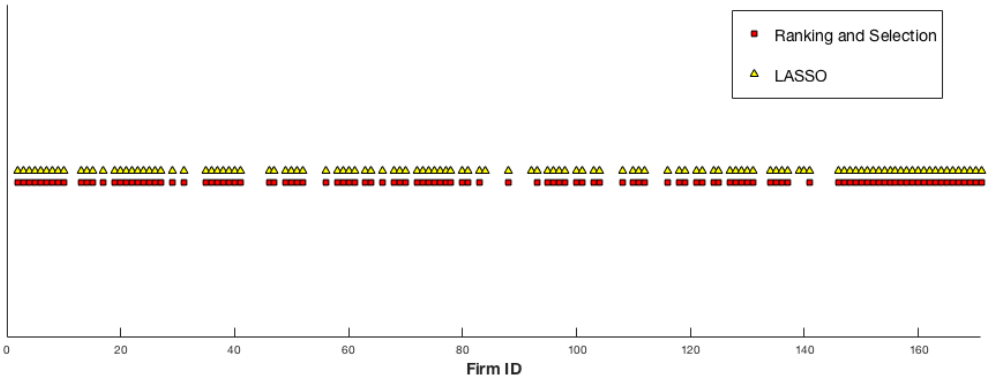


Figure 5: Firm IDs estimated as fully efficient by LASSO or included in the best subset by R&S

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 LSDV in many aspects. The empirical application shows that our methodology and the R&S procedure produce similar results. Consequently, there may be scope for using the LASSO whenever R&S is prescribed. For example, Horrace (2005b) uses R&S to determine a subset of industries with the largest wage gap. Using Oaxaca-Blinder decomposition, the LASSO could be applied to select these industries.

For future research, it may be interesting to develop a zero inefficiency SF model that allows the individual inefficiencies to vary over time. The standard fixed effect model can be seen as a special case of the factor

[28] In this application, the absolute values of the individual inefficiencies from the preliminary LSDV estimation were small whereas the model fit (measured by the mean squared error of the model) was bad, which leads the LASSO to keep a small number of large inefficiency estimates in the model. For the same reason, R&S resulted in wide confidence intervals, wide enough for a large portion of farms to be categorized as the best.

error structure panel data model when the factors are constant over time. If we assume the variation in the inefficiencies is due to time-varying factors, the factor error structure model could be a good baseline model for the analysis.[29] One way to apply the zero inefficiency concept to the factor model would be to estimate some individual factor loadings as zeros using the shrinkage technique as in Cheng et al (2016). However, this may lead to an unsatisfying result, as some firms will have zero inefficiencies all the time. We may consider of a model that allows for factor loadings to be zero for some periods but nonzero for others. However, as shown in Cheng et al (2016), this would make the model and estimation procedures prohibitively complex. Perhaps, an alternative strategy could be developed.

[29] For studies of the factor error structure panel model, see Pesaran (2006), Bai (2009), and Ahn et al (2013).

A Proofs of the lemmas and the theorem

We denote

- LSDV estimates for individual intercepts : $\hat{\alpha}_i(0)$ for $i = 1, ..., N$
- LASSO estimates for individual intercepts : $\hat{\alpha}_i(\Pi)(= \hat{\alpha}(\Pi) - \hat{u}(\Pi)_i)$ for $i = 1, ..., N$

Lemma 4.1 Under the assumption 4.1, we have $E((\hat{\alpha}(0) - \alpha_0)^2) = O(\frac{(\log N)^2}{T})$, $E(||\hat{\beta}(0) - \beta_0||_2^2) = O(\frac{1}{NT})$, and $E((\hat{u}_i(0) - u_{0,i})^2) = O(\frac{(\log N)^2}{T})$ for $\forall i$ as $(N, T) \to \infty$

Proof $E(||\hat{\beta}(0) - \beta_0||_2^2) = O(\frac{1}{NT})$ can be easily proved as it is a standard fixed effect panel data model problem in the literature. We only prove $E((\hat{\alpha}(0) - \alpha_0)^2) = O(\frac{(\log N)^2}{T})$ and $E((\hat{u}_i(0) - u_{0,i})^2) = O(\frac{(\log N)^2}{T})$ here.

It is easy to show $(\hat{\alpha}_i(0) - \alpha_{0,i})^2 = O(\frac{1}{T})$ for $\forall i$ from $\hat{\alpha}(0)_i = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}^{'}\hat{\beta}(0))$ which implies

$$\hat{\alpha}_i(0) - \alpha_{0,i} = \frac{1}{T}\sum_{t=1}^{T} x_{it}^{'}(\beta_0 - \hat{\beta}(0)) + \frac{1}{T}\sum_{t=1}^{T} v_{it} \Rightarrow E((\hat{\alpha}_i(0) - \alpha_{0,i})^2) = O(\frac{1}{T}) \tag{A.1}$$

due to $E(||\hat{\beta}(0) - \beta_0||_2^2) = O(\frac{1}{NT})$. Then, using the results from Theorem 3.2 (ii) of Park, Schmidt and Simar (1998), we can show

$$\hat{\alpha}(0) - \alpha_0 = O_p\left(\frac{\log N}{\sqrt{T}}\right) \Rightarrow E((\hat{\alpha}(0) - \alpha_0)^2) = O\left(\frac{(\log N)^2}{T}\right) \tag{A.2}$$

which is due to $E_{max_i}|\frac{1}{\sqrt{T}}\sum_i v_{it}| = O(\log N)$. Next, we derive the convergence rate of $\hat{u}_i(0)$. By definition, $\hat{u}_i(0) = \hat{\alpha}(0) - \hat{\alpha}_i(0)$. Then, from above results, we have

$$\begin{aligned} &\hat{u}_i(0) = \alpha_0 + O_p\left(\frac{\log N}{\sqrt{T}}\right) - \left(\alpha_{0,i} + O_p(\frac{1}{\sqrt{T}})\right) \Rightarrow \sup|\hat{u}_i(0) - (\alpha_0 - \alpha_{0,i})| = O_p\left(\frac{\log N}{\sqrt{T}}\right) \\ &\Rightarrow E((\hat{u}_i(0) - u_{0,i})^2) = O\left(\frac{(\log N)^2}{T}\right) \end{aligned} \tag{A.3}$$

The lemma is proved. ■

Lemma 4.2 Under the assumption 4.1 and 4.2, we have (i) $E((\hat{\alpha}(\Pi) - \alpha_0)^2) = O(\frac{1}{\delta_0 NT})$, (ii) $E(||\hat{\beta}_A(\Lambda) - \beta_{0,A}||_2^2) = O(\frac{1}{NT})$, (iii) $E((\hat{u}_{A,i}(\Pi) - u_{0,A,i})^2) = O(\frac{1}{T})$ for $\forall i$ $(N, T) \to \infty$

Proof (ii) has been proved in the literature. It can be proved similarly using the first step equation of (3.1). To prove (i), we use the coordinate decent algorithm in 3.1. The algorithm implies that $\hat{\alpha}(\Pi)$ is

estimated as a common intercept of the firms categorized as fully efficient by LASSO technique. Therefore, if we know the true model, $\hat{\alpha}(\Pi)$ is given by $\hat{\alpha}(\Pi) = \frac{1}{\delta_0 NT} l'_{NT,\delta}(Y - X_{1,A}\hat{\beta}_A(\Lambda))$ where $l_{NT,\delta}$ a vector with ones at the observations with zero inefficiency and zero at the others. From the model assumption and $E(||\hat{\beta}_A(\Lambda) - \beta_{0,A}||_2^2) = O_p(\frac{1}{NT})$, we have

$$(\hat{\alpha}(\Pi) - \alpha_0) = \frac{1}{\delta_0 NT} l'_{NT,\delta} X_{1,A}(\beta_{0,A} - \hat{\beta}_A(\Lambda)) + \frac{1}{\delta_0 NT} l'_{NT,\delta} v \Rightarrow E(\hat{\alpha}(\Pi) - \alpha_0)^2 = O\left(\frac{1}{\delta_0 NT}\right) \quad \text{(A.4)}$$

Next, we prove (iii). The algorithm gives us

$$\hat{u}_{A,i}(\Pi) = \hat{\alpha}(\Pi) - \hat{\alpha}_i(0) - \frac{\Pi\hat{\pi}_i}{2T} = \alpha_0 + O_p\left(\frac{1}{\sqrt{\delta_0 NT}}\right) - \left(\alpha_{0,i} + O_p(\frac{1}{\sqrt{T}})\right) - \frac{\Pi\hat{\pi}_i}{2T} \quad \text{(A.5)}$$

which implies

$$\sup|\hat{u}_{A,i}(\Pi) - u_{0,i}| = O_p(\frac{1}{\sqrt{T}}) - \frac{\Pi\hat{\pi}_i}{2T} \leq O_p(\frac{1}{\sqrt{T}}) - O_p(\frac{1}{2\sqrt{T}\log N})\frac{\Pi}{\sqrt{T^*}}\eta^{-\gamma_u}\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma_u} \quad \text{(A.6)}$$

where $\eta = \min_i(|u_{0,i}|)$ and $\hat{\eta} = \min_i(|\hat{u}_i(\Pi)|)$ and $T^* = \frac{T}{(\log N)^2}$. We show that

$$\begin{aligned} E\left[\left(\frac{\hat{\eta}}{\eta}\right)^2\right] &= E\left[\left(\frac{\hat{\eta} - \eta + \eta}{\eta}\right)^2\right] \leq \frac{2}{\eta^2}E\left[(\hat{\eta} - \eta)^2\right] + 2 \\ &\leq \frac{2 \cdot E(|\hat{u}_a(0) - u_{0,a}|^2) + 2E(|\hat{u}_b(0) - u_{0,b}|^2)}{\eta^2} + 2 = \frac{2}{\eta^2}O\left(\frac{(\log N)^2}{T}\right) + 2 = O(1) \end{aligned} \quad \text{(A.7)}$$

where a and b is defined from $\eta = |u_{0,a}|$ and $\hat{\eta} = |\hat{u}_b(\Pi)|$. The second inequality is straightforward if $a = b$. Even if not, when $\hat{\eta} > \eta$, $E\left[(\hat{\eta} - \eta)^2\right] < E(|\hat{u}_a(0) - u_{0,a}|^2)$ and when $\hat{\eta} < \eta$, $E\left[(\hat{\eta} - \eta)^2\right] < E(|\hat{u}_b(0) - u_{0,b}|^2)$ so the inequality holds. The next equality is due to the fact that the convergence rate for the individual LSDV estimator is $O_p\left(\frac{(\log N)^2}{T}\right)$. The last equality holds because

$$O\left(\frac{(\log N)^2}{T}\frac{1}{\eta^2}\right) = O\left((\frac{\Pi}{\sqrt{T^*}}\eta^{-\gamma_u})^{2/\gamma_u}(\log N \cdot \frac{\Pi}{\sqrt{T}} \cdot T^{*\gamma_u/2})^{-2/\gamma_u}\right) = o(1) \quad \text{(A.8)}$$

due to the assumption 4.2. Therefore, (A.7) gives us

$$\sup|\hat{u}_{A,i}(\Pi) - u_{0,i}| \leq O_p(\frac{1}{\sqrt{T}}) - O_p(\frac{1}{\sqrt{T}\log N})o_p(1) = O_p(\frac{1}{\sqrt{T}}) \quad \text{(A.9)}$$

which implies $E((\hat{u}_{A,i}(\Pi) - u_{0,A,i})^2) = O(\frac{1}{T})$ for $\forall i$ ■

Lemma 4.2.B Under the assumption 4.1, 4.2 and 4.3, the adaptive LASSO estimator for β_0, $\hat{\beta}(\Lambda)$, has the

oracle property; that is, the estimator satisfy:

1. Consistency in selection: $Pr(\{j : \hat{\beta}(\Lambda)_j \neq 0\} = A) \to 1$
2. Aysmptotic normality: $\sqrt{NT}(\hat{\beta}(\Lambda) - \beta_0) \to_d N(0, \sigma^2 \Sigma_{1,A}^{-1})$

Proof See Zou (2006) for proof. Only difference is the regressors in our problem involves a within transformation.

Lemma 4.3 Under the assumption 3.1 and 3.2, $[\hat{a}(\Pi), (\hat{\beta}_A(\Lambda)', 0'), (\hat{U}_A(\Pi)', 0')]'$ is the solution to the minimization problem of (2.3) w.p.a 1 $(N, T) \to \infty$

Proof By the lemma 4.2.B, the selection consistency of $\hat{\beta}(\Lambda)$ is already proved. To show $(\hat{U}_A(\Pi)', 0')$ is the solution to the minimization problem of (2.3) w.p.a 1, we need to show it satisfies [30]

$$P\left\{\text{For} \quad j \in A^c, |2x_j'(Y - X_A \cdot \hat{\theta}_A(\Lambda, \Pi))| \leq \Pi \cdot \hat{\pi}_j\right\} \to 1$$

or equivalently,

$$\Psi_j \equiv P\left\{\text{For} \quad j \in A^c, |2x_j'(Y - X_A \cdot \hat{\theta}_A(\Lambda, \Pi))| > \Pi \cdot \hat{\pi}_j\right\} \to 0 \tag{A.10}$$

With some set theorems, we have

$$\begin{aligned}
\Psi_j &\equiv P\left\{|2x_j'(Y - X_A \cdot \hat{\theta}_A(\Lambda, \Pi))| > \Pi \cdot \hat{\pi}_j\right\} \\
&= P\left\{|2x_j'(Y - X_A \cdot \hat{\theta}_A(\Lambda, \Pi))| > \Pi \cdot \hat{\pi}_j \quad \text{and} \quad \hat{\eta} > \eta/2\right\} \\
&\qquad + P\left\{|2x_j'(Y - X_A \cdot \hat{\theta}_A(\Lambda, \Pi))| > \Pi \cdot \hat{\pi}_j \text{and} \quad \hat{\eta} \leq \eta/2\right\} \\
&\leq P\left\{|2x_j'(Y - X_A \cdot \hat{\theta}_A(\Lambda, \Pi))| > \Pi \cdot \hat{\pi}_j, \hat{\eta} > \eta/2\right\} + P\{\hat{\eta} \leq \eta/2\}
\end{aligned} \tag{A.11}$$

As in (A.7), a and b is defined from $\eta = |u_{0,a}|$ and $\hat{\eta} = |\hat{u}_b(\Pi)|$, then, we can show

$$\begin{aligned}
P\{\hat{\eta} \leq \eta/2\} &\leq P\{|\hat{u}_a(0) - u_{0,a}| \geq \eta/2\} + P\{|\hat{u}_b(0) - u_{0,b}| \geq \eta/2\} \\
&\leq \frac{E(|\hat{u}_a(0) - u_{0,a}|^2) + E(|\hat{u}_b(0) - u_{0,b}|^2)}{\eta^2/4} = O\left(\frac{(\log N)^2}{T}\right) \cdot \frac{1}{\eta^2}
\end{aligned} \tag{A.12}$$

w.p.a 1. The first inequality is due to that if $a = b = e^*$, $\hat{\eta} \leq \frac{\eta}{2}$ implies $|\hat{\eta} - \eta| = |\hat{u}_{e^*}(0) - u_{0,e^*}| \geq \frac{\eta}{2}$. Similiarly, if $a \neq b$, $\hat{\eta} \leq \frac{\eta}{2}$ implies $|\hat{\eta} - u_{0,b}| \geq |\hat{\eta} - \eta| \geq \frac{\eta}{2}$ due to $|u_{0,b}| > \eta$ by definition of η. So the inequality

[30] Zou and Zhang (2009) considers the selection consistency for all the coefficients at the same time, but in our problem, even though there is an infinite number of inefficiency terms in the limit, they are independent due to the independency between the dummy variables, therefore it is suffice to consider of the selection consistency individually.

holds. Let $M = \left(\frac{\Pi}{T/\log N}\right)^{1/(1+\gamma_u)}$, then, we can show

$$
\begin{aligned}
&P\left\{|2x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))| > \Pi \cdot \hat{\pi_j}, \hat{\eta} > \eta/2\right\} \\
&\leq P\left\{|2x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))| > \Pi \cdot \hat{\pi_j}, \hat{\eta} > \eta/2, |\hat{u}_j(0)| \leq M\right\} + P\left\{|\hat{u}_j(0)| > M\right\} \\
&\leq P\left\{|2x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))| > \Pi \cdot M^{-\gamma_u}, \hat{\eta} > \eta/2\right\} + P\left\{|\hat{u}_j(0)| > M\right\} \\
&\leq 4\frac{M^{2\gamma_u}}{\Pi^2} \cdot E\left[|x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))|^2 \cdot I_{\hat{\eta}>\eta/2}\right] + \frac{1}{M^2} \cdot E\left[|\hat{u}_j(0)|^2\right] \\
&\leq 4\frac{M^{2\gamma_u}}{\Pi^2} \cdot E\left[|x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))|^2 \cdot I_{\hat{\eta}>\eta/2}\right] + \frac{1}{M^2} \cdot E(|\hat{u}_j(0) - u_{0,j}|^2) \quad \text{due to } u_{0,j} \in A^c \\
&= 4\frac{M^{2\gamma_u}}{\Pi^2} \cdot \underbrace{E\left[|x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))|^2 \cdot I_{\hat{\eta}>\eta/2}\right]}_{\text{(A)}} + \frac{1}{M^2} \cdot O\left(\frac{(\log N)^2}{T}\right) \quad \text{w.p.a 1 by the lemma 4.1}
\end{aligned}
\tag{A.13}
$$

Next, we derive a upper bound of (A). By the model assumption, we have

$$
\begin{aligned}
&E\left[|x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))|^2\right] = E\left[|x_j^{'}(X_A \cdot \theta_0 + v - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))|^2\right] \\
&\leq 2 \cdot E\left[|x_j^{'} \cdot X_A(\theta_0 - \hat{\theta}_A(\Lambda,\Pi))|^2\right] + 2 \cdot E\left[|x_j^{'}v|^2\right]
\end{aligned}
\tag{A.14}
$$

Due to $j \in A^c$, there's no inefficiency estimate and parameter in this firm, therefore, we have

$$
\begin{aligned}
&\leq 2 \cdot E(|x_j^{'}X_{1,A}(\beta_0 - \hat{\beta}_A(\Lambda)) + x_j^{'}l_{NT}(a_0 - \hat{a}(\Pi))|^2) + 2 \cdot T \cdot \sigma^2 \\
&\leq 4T \cdot E(||X_{1,A}(\beta_0 - \hat{\beta}_A(\Lambda))||_2^2) + 4 \cdot T^2 \cdot E(|a_0 - \hat{a}(\Pi))|^2) + 2 \cdot T \cdot \sigma^2 \\
&\leq 4 \cdot BNT^2 \cdot E(||\beta_0 - \hat{\beta}_A(\Lambda)||_2^2) + 4 \cdot T^2 \cdot E(||a_0 - \hat{a}(\Pi)||_2^2) + 2 \cdot T \cdot \sigma^2 \\
&\leq 4B \cdot NT^2 \cdot \left\{E(||a_0 - \hat{a}(\Pi)||_2^2) + E(||\beta_0 - \hat{\beta}_A(\Lambda)||_2^2)\right\} + 2 \cdot T \cdot \sigma^2
\end{aligned}
\tag{A.15}
$$

Then, (A.31) and (A.32) give us the inequality

$$
\begin{aligned}
&E\left[\sum_{j \in A^c} |x_j^{'}(Y - X_A \cdot \hat{\theta}_A(\Lambda,\Pi))|^2 \cdot I_{\hat{\eta}>\eta/2}\right] \\
&\leq 4BNT^2 \cdot \left\{E(||a_0 - \hat{a}(\Pi)||_2^2 \cdot I_{\hat{\eta}>\eta/2}) + E(||\beta_0 - \hat{\beta}_A(\Lambda)||_2^2 \cdot I_{\hat{\eta}>\eta/2})\right\} + 2T \cdot \sigma^2
\end{aligned}
\tag{A.16}
$$

Combining the above results and by the lemma 3.2, we have the upper bound of Ψ_j as

$$
\begin{aligned}
\Psi_j &\leq 4\frac{M^{2\gamma_u}}{\Pi^2} \cdot (4BNT^2 \cdot O(\frac{1}{NT}) + 2T\sigma^2) + \frac{1}{M^2} \cdot O\left(\frac{(\log N)^2}{T}\right) + O\left(\frac{(\log N)^2}{T}\right) \cdot \frac{1}{\eta^2} \\
&\equiv \Psi_{j,1} + \Psi_{j,2} + \Psi_{j,3}
\end{aligned}
\tag{A.17}
$$

w.p.a 1. Under the assumption 3.1 and 3.2, and due to (A.20) we can show

$$\begin{aligned}
\Psi_{j,1} &= O\left(\frac{M^{2\gamma_u}}{\Pi^2}\cdot T\right) = O\left(\left(\frac{\Pi}{\sqrt{T}}\cdot T^{*\gamma_u/2}\right)^{-2/(1+\gamma_u)}\right) \to 0 \\
\Psi_{j,2} &= O\left(\frac{1}{M^2}\cdot\frac{1}{T^*}\right) = O\left(\left(\frac{\Pi}{\sqrt{T}}\cdot T^{*\gamma_u/2}\right)^{-2/(1+\gamma_u)}\right) \to 0 \\
\Psi_{j,3} &= O(\frac{1}{T^*}\frac{1}{\eta^2}) = O\left(\left(\frac{\Pi}{\sqrt{T^*}}\eta^{-\gamma_u}\right)^{2/\gamma_u}\left(\log N\frac{\Pi}{\sqrt{T}}\cdot T^{*\gamma_u/2}\right)^{-2/\gamma_u}\right) \to 0
\end{aligned} \tag{A.18}$$

Thus, the proof is complete ■

Theorem (Oracle Property) Under the assumption 4.1, 4.2 and 4.3, the LASSO estimator for θ_0, $\hat{\theta}(\Lambda,\Pi)$, has the oracle property; that is, the estimator satisfy:

1. Consistency in selection: $Pr(\{j:\hat{\theta}(\Lambda,\Pi)_j \neq 0\} = A) \to 1$ as $(N,T)\to\infty$
2. Asymptotic normality
   1) $\sqrt{NT}(\hat{\beta}_A(\Lambda)-\beta_{0,A}) \to_d N(0,\sigma^2_{v,0}\Sigma^{-1}_{A,1})$
   2) $\sqrt{\delta_0 NT}(\hat{\alpha}(\Pi)-\alpha_0) \to_d N(0,\sigma^2_{v,0}\Sigma_\alpha)$
   3) $\sqrt{T}(\hat{u}_{i,A}(\Pi)-u_{i,0}) \to_d N(0,\sigma^2_{v,0})$

w.p.a 1 as $(N,T)\to\infty$ where $\Sigma_\alpha = 1+\delta_0\frac{l'_\delta X_{1,A}}{\delta_0 NT}(\frac{X'_{1,A}QX_{1,A}}{NT})^{-1}\frac{X'_{1,A}l_\delta}{\delta_0 NT}$.

Proof Using Lemma 4.2.B and 4.3, we only need to show $P\{\min_{i\in A}\hat{u}_i(\Pi)>0\}\to 1$. We can show that

$$\hat{u}_a(\Pi) > \min_{i\in A}u_{0,i} - |\hat{u}_a(\Pi)-u_{0,a}| \tag{A.19}$$

where a is the index for $\hat{u}_a(\Pi)=\min_{i\in A}\hat{u}_i(\Pi)$. The inequality is straightforward if $\hat{u}(\Pi)_a > \min_{i\in A}u_{0,i}$. When $\hat{u}_a(\Pi)<\min_{i\in A}u_{0,i}$, it implies $\hat{u}_a(\Pi)<\min_{i\in A}u_{0,i}<u_{0,a}$, which in turn implies $\hat{u}_a(\Pi)-\min_{i\in A}u_{0,i} > -|\hat{u}_a(\Pi)-u_{0,a}|$. Then, due to the Lemma 4.2, we have

$$\min_{j\in A}\hat{u}(\Pi)_j > \eta - O_p(\sqrt{\frac{1}{T}}) \tag{A.20}$$

The proof is complete because we have already shown that $\sqrt{\frac{(\log N)^2}{T}}$ converges to zero faster than η by (A.8).

Next, we prove the asymptotic normality. From the selection consistency, we only have nonzero coefficients now. Then, each estimators' asymptotic distribution will be given as follows: First, we already showed that

$\sqrt{NT}(\hat{\beta}_A(\Lambda)-\beta_{0,A}) \to_d N(0, \sigma^2_{v,0}\Sigma^{-1}_{A,1})$ from Lemma 4.2.B. Second, as we've seen in the Lemma 4.2, we have $(\hat{\alpha}(\Pi)-\alpha_0) = \frac{1}{\delta_0 NT} l_\delta^{'}(Y - X_{1,A}\hat{\beta}_A(\Lambda))$. As $(N,T)\to\infty$, we will have $\hat{\beta}_A(\Lambda) = (X_{1,A}^{'}QX_{1,A})^{-1}X_{1,A}^{'}QY$ w.p.a 1. Then, we can show

$$\begin{aligned}\sqrt{\delta_0 NT}(\hat{\alpha}(\Pi)-\alpha_0) &= \frac{1}{\sqrt{\delta_0 NT}} l_\delta^{'}(v - X_{1,A}(X_{1,A}^{'}QX_{1,A})^{-1}X_{1,A}^{'}Qv) \\ &= \frac{1}{\sqrt{\delta_0 NT}} l_\delta^{'}v - \sqrt{\delta_0}\frac{l_\delta^{'}X_{1,A}}{\delta_0 NT}(\frac{X_{1,A}^{'}QX_{1,A}}{NT})^{-1}\frac{1}{\sqrt{NT}}X_{1,A}^{'}Qv\end{aligned} \tag{A.21}$$

where we omit the inefficiency terms as they will be removed eventually. By CLT, we can show

$$\sqrt{\delta_0 NT}(\hat{\alpha}(\Pi)-\alpha_0) \to_d N(0, \sigma^2_{v,0}\Sigma_\alpha) \tag{A.22}$$

where $\left(\frac{1}{\sqrt{\delta_0 NT}} l_\delta^{'} - \sqrt{\delta_0}\frac{l_\delta^{'}X_{1,A}}{\delta_0 NT}(\frac{X_{1,A}^{'}QX_{1,A}}{NT})^{-1}\frac{1}{\sqrt{NT}}X_{1,A}^{'}Q\right)\left(\frac{1}{\sqrt{\delta_0 NT}} l_\delta^{'} - \sqrt{\delta_0}\frac{l_\delta^{'}X_{1,A}}{\delta_0 NT}(\frac{X_{1,A}^{'}QX_{1,A}}{NT})^{-1}\frac{1}{\sqrt{NT}}X_{1,A}^{'}Q\right)^{'} \to_p \Sigma_\alpha$. This reduces to $1 + \delta_0\frac{l_\delta^{'}X_{1,A}}{\delta_0 NT}(\frac{X_{1,A}^{'}QX_{1,A}}{NT})^{-1}\frac{X_{1,A}^{'}l_\delta}{\delta_0 NT} \to_p \Sigma_\alpha$ due to $Ql_\delta = 0$

Lastly, for each nonzero inefficiencies, we have F.O.C as

$$\sum^{i}(y_{it,A} - \hat{\alpha}(\Pi) - x_{it,1,A}\cdot\hat{\beta}_A(\Lambda) + \hat{u}_{i,A}(\Pi)) + \frac{\Pi}{2}\cdot\hat{\pi}_i = 0 \quad \text{for } \hat{u}_{i,A}(\Pi) > 0 \tag{A.23}$$

w.p.a 1. This leads to

$$\sqrt{T}(\hat{u}_{i,A}(\Pi) - u_{i,0}) = \underbrace{\sqrt{T}(\hat{\alpha}(\Pi)-\alpha_0) + \frac{1}{\sqrt{T}}\sum_i x_{it,1,A}(\hat{\beta}_A(\Lambda)-\beta_{A,0})}_{(1)} - \frac{1}{\sqrt{T}}\sum^{i} v_{it} - \underbrace{\frac{\Pi}{2\sqrt{T}}\cdot\hat{\pi}_i}_{(2)} \tag{A.24}$$

As $(1) = o_p(1)$ and $(2) = o_p(1)$ (due to A.6), by CLT, we have

$$\sqrt{T}(\hat{u}_{i,A}(\Pi) - u_{i,0}) \to_d N(0, \sigma^2_{v,0}) \tag{A.25}$$

w.p.a 1 as $(N,T)\to\infty$. This completes the proof. ■

B The equivalence between the one step estimation of (2.3) and the two step estimation of (2.5) and (2.6)

In order to show the equivalence between the one step estimation of (2.3) and the two step estimation of (2.5) and (2.6), we first derive the F.O.C for $\hat{U}(\Pi)$ in (2.3), which gives us the close form solution for $\hat{U}(\Pi)$ such that

$$\hat{U}(\Pi) = \left[(X_2^{'} X_2)^{-1} \left(-X_2^{'} (Y - \alpha \cdot l_{NT} - X_1 \beta) - \Pi \cdot i_N (\frac{\hat{\pi_i}}{2}) \right) \right]_+ \tag{B.1}$$

where $(x)_+ = x$ if $x \geq 0$ and $= 0$ otherwise and $i_N(\hat{\pi_i})$ is $N \times 1$ vector with $\hat{\pi_i}$ for i^{th} element. Plugging this result into (2.3) give us the equation

$$\begin{aligned} &\left\{ Q \cdot (Y - \alpha \cdot l_{NT} - X_1 \beta) - X_2 \Pi \cdot i_N \left(\frac{\hat{\pi_i}}{2T} + \epsilon_i \right) \right\}^{'} \{...\} + \Lambda \sum_{j=1}^{p} \hat{\lambda_j} |\beta_j| + \Pi \sum_{k=1}^{N} \hat{\pi_k} |u_k| \\ &= \left\{ Q \cdot (Y - X_1 \beta) - X_2 \Pi \cdot i_N \left(\frac{\hat{\pi_i}}{2T} + \epsilon_i \right) \right\}^{'} \{...\} + \Lambda \sum_{j=1}^{p} \hat{\lambda_j} |\beta_j| + \Pi \sum_{k=1}^{N} \hat{\pi_k} |u_k| \end{aligned} \tag{B.2}$$

where ϵ_i is for removing the positive constraints on the inefficiency estimates so β has no corresponding constraint in (B.3). The equality is due to $Q \cdot l_{NT} = 0$. Next we derive the F.O.C for β in (B.2), which is given by

$$\begin{aligned} &- 2 \cdot X_1^{'} Q \left\{ Q \cdot (Y - X_1 \beta) - X_2 \Pi \cdot i_N \left(\frac{\hat{\pi_i}}{T} + \epsilon_i \right) \right\} + \Lambda \cdot i_p(\hat{\lambda_i}) \\ &= -2 \cdot X_1^{'} Q \left(Y - X_1 \beta \right) + \Lambda \cdot i_p(\hat{\lambda_i}) \end{aligned} \tag{B.3}$$

This equality is due to $Q \cdot Q = Q$ and $Q \cdot X_2 = 0$. This is the same with the F.O.C for β in (2.5) so we show the equivalence between the two. ■

C The proximity of the new algorithm and the standard constrained optimization algorithm

The basic set up here is the same with that in the simulation exercise. We compare the series of estimation results on $U(\hat{\Pi})$ (RMSE, Pr_{U_A}, $\hat{\alpha}$, Sum of squared error) from the new algorithm (NA) and one of the standard constrained algorithm (SA) (SQP in Matlab) in the 20 replications of two models: (1) (N,T)=(20,10), $\sigma_u = 1$, and (2) (N,T)=(10,20) $\sigma_u = 2$.

Figure 6 is from model (1) and Figure 7 is from model (2). The black dashed line is from NA and the red dotted line is from SA. In model (1), there are one or two replications showing the two algorithms produced slightly different results, but the differences are small enough to conclude that the two algorithms generally produce very similar results. In model (2), the two produced almost identical results in every replication. The new algorithm is incomparably faster than SA in estimation. For one replication, the new algorithm took on average 1.5 second whereas the standard algorithm took 345 seconds. This difference will be pronounced as the sample size increases.

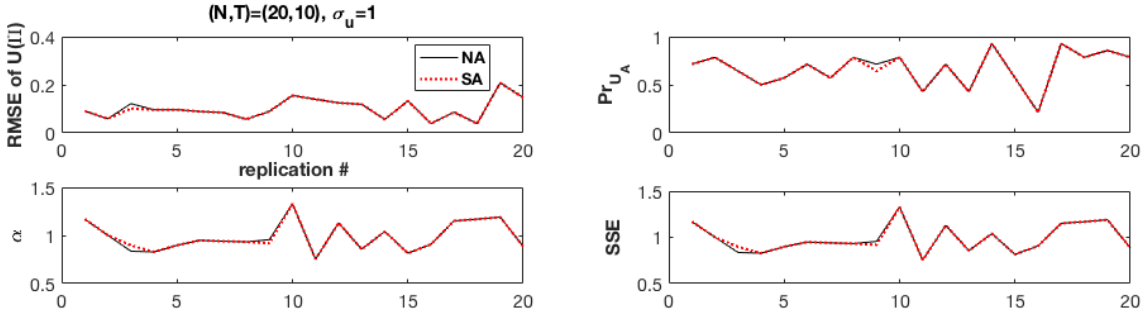


Figure 6: Model (1)

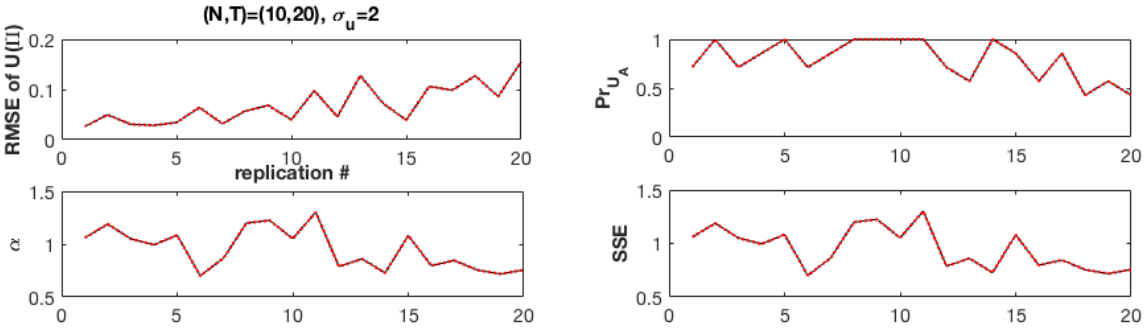
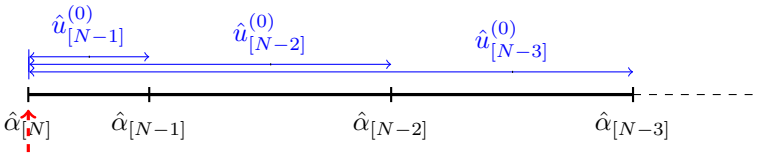


Figure 7: Model (2)

D New algorithm

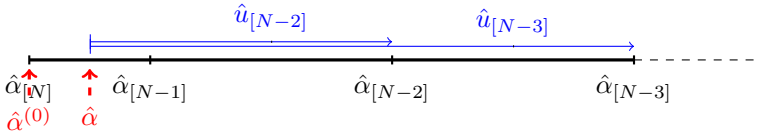
- Using $\beta(\hat{\Lambda})$ from the first step, compute $\hat{\alpha}_i = \frac{1}{T}\sum_i y_{it} - x_{it}^{'}\beta(\hat{\Lambda})$ and $\hat{u}_i = \text{Max}_{j=1}^{N}\hat{\alpha}_j - \hat{\alpha}_i$ for all i. Let $\hat{\alpha}_{[1]} \leq \hat{\alpha}_{[2]} \leq ... \leq \hat{\alpha}_{[N]}$ be the rankings of the $\hat{\alpha}_i$, so $\hat{\alpha}_{[N]} = \text{Max}_{j=1}^{N}\hat{\alpha}_j$. Similarly, let $\hat{u}_{[N]} \leq \hat{u}_{[N-1]} \leq ... \leq \hat{u}_{[1]}$ be the rankings of the $\hat{u}_i$, so $\hat{u}_{[N]} = \text{Min}_{j=1}^{N}\hat{u}_j$. We set the initial value for α as $\hat{\alpha}_{[N]}$. Denote the current values for $\hat{u}_{[i]}$ and $\hat{\alpha}$ as $\hat{u}_{[i]}^{(0)}$ and $\hat{\alpha}^{(0)}$, then we have below initial settings. Note that as $\hat{\alpha}^{(0)} = \hat{\alpha}_{[N]}$, we have one fully efficient firm, $\hat{u}_{[N]}^{(0)} = 0$, now.



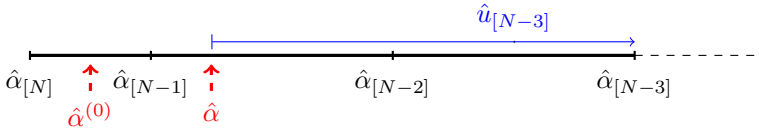
$\hat{\alpha}^{(0)} = \hat{\alpha}_{[N]}, \hat{u}_{[N]}^{(0)} = 0$

- For a given Π, check the KKT condition for the second best firm($=[N-1]$), that is, check the sign of $\Delta_{[N-1]} = \hat{u}_{[N-1]}^{(0)} - \Pi\frac{\hat{\pi}_{[N-1]}}{2T}$

  1. $\Delta_{[N-1]} \leq 0$, update $\hat{u}_{[N-1]}^{(0)}$ as $\hat{u}_{[N-1]} = 0$, and update $\hat{\alpha}^{(0)}$ as $\hat{\alpha} = \frac{1}{2}(\hat{\alpha}_{[N]} + \hat{\alpha}_{[N-1]})$. As we have a new frontier ($=\hat{\alpha}$), we update the rest of the inefficiencies (from $[N-2]$ to $[1]$) as below.



     (a) Check the KKT condition for the third best firm($=[N-2]$), and if $\Delta_{[N-2]} \leq 0$, update $\hat{u}_{[N-2]}^{(0)}$ as $\hat{u}_{[N-2]} = 0$, and update $\hat{\alpha}^{(0)}$ as $\hat{\alpha} = \frac{1}{3}(\hat{\alpha}_{[N]} + \hat{\alpha}_{[N-1]} + \hat{\alpha}_{[N-2]})$, and update the rest of the inefficiencies (from $[N-3]$ to $[1]$) as below.



     (b) if $\Delta_{[N-2]} > 0$, do some minor updating for $\hat{\alpha}$ and $\hat{u}_{[i]}$ for $i = N-2, N-3, ..., 1$ as 2. (b) of the algorithm in 4.2 and report the results.

E Simulation results for β

Table 4: Simulation results for β

	Estimation accuracy: RMSE			Selection accuracy	
(N, T, σ_u)	LASSO	Oracle	LSDV	$\lvert\hat{\beta_A}\rvert$	$\lvert\hat{\beta_{A^c}}\rvert$
(100,10,1)	0.023	0.013	0.040	3	4.92
	(0.011)	(0.002)	(0.011)	(0)	(0.265)
(100,50,1)	0.009	0.005	0.017	3	4.97
	(0.004)	(0.001)	(0.004)	(0)	(0.146)
(200,50,4)	0.006	0.004	0.012	3	4.98
	(0.003)	(0.000)	(0.003)	(0)	(0.117)

The numbers in main entries are the average RMSE over 1,000 replications and $\lvert\hat{\beta_A}\rvert$ represents the average number of nonzero estimate for β_A (the true = 3) over the replications, and $\lvert\hat{\beta_{A^c}}\rvert$ represents the average number of zero estimate for β_{A^c} (the true = 5). The corresponding variances of each measures are in next row in parentheses. The oracle is the RMSE calculated from when we know the true model and apply LSDV estimation to the model.

References

[1] Aigner, D., Lovell, C. A. K., Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. Journal of Econometrics, 6(1), 21-37.

[2] Ahn, S. G., Lee, Y. H. and Schmidt, P. (2013): Panel data models with multiple time-varying effects. J of Econometrics, 174, 1-14.

[3] Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77, 1229-1279.

[4] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain, Econometrica, 80, 2369-2431.

[5] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of LASSO and Dantzig selector. Ann. Statist., 37, 1705-1732.

[6] Bonhomme, S., and E. Manresa (2015): Grouped Patterns of Heterogeneity in Panel Data, Econometrica, 83, 1147-1184.

[7] Caner, M. (2009). LASSO type GMM estimator, Econometric Theory, 25, 270-290.

[8] Caner, M., Han, X., and Lee, Y. (2016) Adaptive Elastic Net GMM Estimation with Many Invalid Moment Conditions: Simultaneous Model and Moment Selection. J Bus Econ Stat 32 (1), 30-47.

[9] Chan, N. H., C. Y. Yau, and R.-M. Zhang (2014): Group LASSO for Structural Break Time Series, Journal of the American Statistical Association, 109, 590-599.

[10] Cheng,.X. and Liao, Z. (2015) Select the Valid and Relevant Moments: An Information-Based LASSO for GMM with Many Moments. J of Econometrics, 2015, 186(2), 443-464.

[11] Cheng,.X., Liao, Z., and Schorfheide, F. (2016) Shrinkage Estimation of High-Dimensional Factor Models with Structural Instabilities. The Review of Econ Studies, 83, 1511-1543

[12] Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression, Annals of Statistics, 32, 407-499.

[13] 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.

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

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

[16] Flores-Lagunes, A., W. C. Horrace, and K. E. Schnier. (2007). Identifying technically efficient fishing vessels: A nonempty, minimal subset approach. Journal of Applied Econometrics 22:729-45.

[17] Friedman, J., Hastie, T., and Tibshirani, R. (2008), Regularization Paths for Generalized Linear Models via Coordinate Descent, Technical Report, Available at http://www-stat.stanford.edu/jhf/ftp/glmnet.pdf.

[18] Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics 126 (2), 269-303.

[19] Horrace, W. C. (2005a). On ranking and selection from independent truncated normal distributions. Journal of Econometrics, 126 (2), 335-354,

[20] Horrace, W. C. (2005b). On the ranking uncertainty of labor market wage gaps. Journal of Population Economics 18:181?7.

[21] Horrace, W.C. and Oaxaca, R.L., (2001). Inter-industry wage differentials and the gender wage gap: An identification problem. ILR Review, 54(3), pp.611-618.

[22] Horrace, W. C. and Schmidt, P. (1996). Confidence Statements for Efficiency Estimates from Stochastic Frontier Models. J Productivity Analysis 7: 257-282

[23] Horrace, W. C., and Schmidt, P. (2000). Multiple comparisons with the best, with economic applications. Journal of Applied Econometrics 15:1-26.

[24] Hu, Q., Zeng, P., Lin, L. (2015). The dual and degrees of freedom of linearly constrained generalized LASSO. Comput Stat Data Anal 2015;86:13-26.

[25] Hui, F.K., 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), pp.262-269.

[26] Jondrow, J., Lovell, C. A. K., Materov. I. S., Schmidt. P. (1982), On the estimation of technical inefficiency in the stochastic frontier production function model, Journal of Econometrics, 19(2), 233-238.

[27] Kumbhakar S. C., Parmeter C. F., Tsionas E. G. (2013) A zero inefficiency stochastic Frontier model. J Econ 172:66-76

[28] Lee, S., Seo, M. H. and Shin, Y. (2016), The LASSO for High-Dimensional Regression with a Possible Change-Point, Journal of the Royal Statistical Society: Series B, 78, Part 1, pp. 193-210.

[29] Lu, X. and Su, L., (2016). Shrinkage estimation of dynamic panel data models with interactive fixed effects. Journal of Econometrics, 190(1), pp.148-175.

[30] Park, B.U., Sickles, R.C., Simar, L. (1998) Stochastic panel frontiers: A semiparametric approach, Journal of Econometrics, Volume 84, Issue 2, 273-301.

[31] Pesaran, M. H. (2006): Estimation and Inference in Large Heterogeneous PanelsWith a Multifactor Error Structure, Econometrica, 74, 967-1012.

[32] Qian, J., And L. Su (2015): Shrinkage Estimation of Regression Models With Multiple Structural Changes, Econometric Theory, first published online 23 June 2015, 1?58, DOI:10.1017/S0266466615000237.

[33] Rho, S. and Schmidt, P. (2015). Are All Firms Inefficient? Journal of Productivity Analysis 43:327?349.

[34] Schmidt, P. and Sickles, R.C., (1984). Production frontiers and panel data. J Bus Econ Stat, 2(4), pp.367-374.

[35] Simar, L. and Wilson, P. (2010), Inferences from Cross-Sectional, Stochastic Frontier Models, Econometric Reviews, 29, (1), 62-98

[36] Su, L., Shi, Z. and Phillips, P.C., (2016). Identifying latent structures in panel data. Econometrica, 84(6), pp.2215-2264.

[37] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[38] Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005): Sparsity and Smoothness via the Fused LASSO, Journal of the Royal Statistical Society, Series B, 67, 91-108.

[39] Wang, H., Li, G., and Tsai, C. L. (2007a). Regression coefficient and autoregressive order shrinkage and selection via the LASSO. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(1), 63-78.

[40] Wang, H., Li, R. and Tsai, C.L., (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94(3), p.553.

[41] Wheat, P., Greene, W. and Smith, (2014). Journal of Productivity Analysis 42: 55.

[42] Zhu, J., Huang, H.C. and Reyes, P.E., (2010). On selection of spatial linear models for lattice data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(3), pp.389-402.

[43] Zou, H., (2006). The adaptive LASSO and its oracle properties. Journal of the American statistical association, 101(476), pp.1418-1429.

[44] Zou, H. and Zhang, H.H., (2009). On the adaptive elastic-net with a diverging number of parameters. Annals of statistics, 37(4), p.1733.