Adaptive Elastic Net GMM Estimator with Many Invalid Moment Conditions: A Simultaneous Model and Moment Selection

Mehmet Caner* Xu Han† Yoonseok Lee‡

September 1, 2013

Abstract

This paper develops an adaptive elastic-net GMM estimator with many possibly invalid moment conditions. We allow for the number of structural parameters (p_0) as well as the number of moment conditions increasing with the sample size (n). The new estimator conducts simultaneous model and moment selection. We estimate the structural parameters along with parameters associated with the invalid moments. The basic idea is to conduct the standard GMM combined with two penalty terms: the quadratic regularization and the adaptively weighted LASSO shrinkage. The new estimator uses information only from the valid moment conditions to estimate the structural parameters and achieve the semiparametric efficiency bound. The estimator is thus very useful in practice since it conducts the consistent moment selection and efficient estimation of the structural parameters simultaneously. We also establish the order of magnitude for the smallest local to zero coefficient to be selected as nonzero. We apply the new estimation procedure to dynamic panel data models, where both time and cross section dimensions are large. The new estimator is robust to possible serial correlations in the error terms of dynamic panel models.

Keywords and phrases: Adaptive Elastic-Net, GMM, many parameters, many invalid moments, semiparametric efficiency, dynamic panel.

JEL classification: C13, C23, C26.

*North Carolina State University, Department of Economics, 4168 Nelson Hall, Raleigh, NC 27695. Email: mcaner@ncsu.edu

†City University of Hong Kong, Department of Economics and Finance. Email: xuhan25@cityun.edu.hk

‡University of Michigan, Department of Economics, 611 Tappan Street, Ann Arbor, MI 48109-1220, USA. Email: yoolee@umich.edu

1 Introduction

The structural parameter estimation in systems with endogenous regressors is a very common issue in applied econometrics. To deal with the endogeneity, economists have to choose the valid moments as well as the structural parameters in the model. The moment selection in systems with a fixed number of moments is usually achieved by the J test. For the model selection, applied researchers usually justify the model via some economic theory or intuition. However, mistakes in moment selection can be carried over to model selection and lead to inconsistent estimates. Additionally, *ad hoc* model selection may result in missing regressors which generate an endogeneity problem in the estimation stage. These issues become more serious in high dimension models. With many endogenous regressors and many moments, we have a higher chance of misspecification, so more attentions must be paid to the moment validity and model selection.

This paper tries to bridge the gap between the model and moment selection. We propose an adaptive elastic net GMM for linear models with many structural parameters and many possibly invalid moment conditions. The new estimator conducts selection and estimation simultaneously. We prove that our estimator can select the correct model and valid moments with probability converging to one. In addition, we show that the estimates for the structural parameters reach the semiparametric efficiency bound. This is due to the fact our method selects all valid moments through penalization and use them to estimate the structural parameters. The invalid instruments only serve to estimate the parameters associated with invalid moments and do not affect the asymptotic variance of the estimates for the structural parameters. This is new in the literature and valuable in practice. The method can be applied to dynamic panel models where the error terms have potential serial correlation. Simulations confirm our theoretical results and show that our estimator performs well in finite samples.

In addition, this paper shows that the LARS algorithm proposed by Efron *et al.* (2004) can be extended into a linear GMM framework. This gives our estimator a great computational advantage over downward or upward testing procedures, especially in a high dimensional setup. Andrews (1999) develops information criteria for moment selection based on the J test, and Andrews and Lu (2001) extend these criteria to allow for parameter selection in the structural equation. While these methods are able to consistently select the correct model and valid moments, the computation cost grows at a geometric rate as the number of parameters and moments diverges.

In the shrinkage estimation literature, a few papers focus on high dimension model or moment

selection. In a seminal paper, Belloni, Chernozhukov, Chen, and Hansen (2012) introduce a heteroskedasticity consistent LASSO estimator and obtain the finite sample performance bound in a large heteroskedastic data context. They deal with optimal instrument selection given that all instruments are valid. Gautier and Tsybakov (2011) provide the finite sample performance bound for the Danzig selector when there are a large number of invalid instruments. Cheng and Liao (2012) provide asymptotic results for the adaptive LASSO estimator when there are many invalid and irrelevant instruments. Caner and Zhang (2013) propose an adaptive elastic net GMM estimator for model selection assuming that all moments are valid. Our paper is different from the papers above in the sense that we conduct model and moment selection simultaneously. By using the adaptive elastic net, we are able to control the problem of multicollinearity in the high-dimensional models. Compared to Caner and Zhang (2013), we also allow for many invalid instruments. This is a nontrivial extension since many invalid moments can affect the analysis of the variance covariance matrix and require a different proof technique.

Recently, Qian and Su (2013) use shrinkage estimators to determine the number of structural changes in multiple linear regression models. Also, Lu and Su (2013) use adaptive LASSO to determine the number of factors and select the proper regressors in linear dynamic panel data models with interactive fixed effects. These will make important contributions to the literature since structural change models and factor model structures are relevant empirically.

Section 2 provides the model and assumptions. Section 3 introduces our estimator and demonstrates how it can be applied to dynamic panel data. Section 4 shows how to choose tuning parameters and proves that Least Angle Regression (LARS) of Efron *et al.* (2004) is applicable to our adaptive elastic net GMM estimator. Section 5 provides simulations. We conclude in section 6. Proofs are contained in the appendix. Let $||A|| = [tr(A'A)]^{1/2}$ for any matrix A.

2 Model

We consider a structural equation given by

$$Y = X\beta_0 + u, \tag{1}$$

where X is the $n \times p$ matrix of endogenous variables. β_0 is the $p \times 1$ true structural parameter vector, where some of the components are zero and some are non-zero. We assume an $n \times q$ instrument matrix Z yielding q moment conditions. However, out of q moment restrictions, we assume that at

most s of them could be invalid and that all the valid instruments are strongly correlated with the endogenous regressors.

We allow that p, q, and s increase with the sample size n satisfying $s/q \to \varphi \in [0,1)$ and $s < n$. We further impose that $p + s \leq q$ for identification purposes. More precisely, we rewrite the q moment conditions as, for each $i = 1, 2 \cdots, n$

$$E[Z_i u_i] - F_{q,s}\tau_0 = 0, \tag{2}$$

where $F_{q,s}$ is a $q \times s$ matrix given by

$$F_{q,s} = \begin{bmatrix} 0_{q-s,s} \\ I_s \end{bmatrix},$$

with $0_{q-s,s}$ being a matrix of zeroes with dimension $(q-s) \times s$, and $\tau_0 \in \mathbb{R}^s$ for some $0 \leq s \leq q-p$. The particular case of $s = 0$ shows that the researcher believes that all moment conditions are valid. This results in a linear GMM estimation of structural parameters. Some elements of τ_0 could be zero, so out of q moment restrictions, we assume that at most s of them could be invalid. Set $Y_z = Z'Y$ is a $q \times 1$ vector, $X_{zF} = [Z'X, nF_{q,s}]$ is a $q \times (p+s)$ matrix and $\theta_0 = (\beta_0', \tau_0')'$ is a $(p+s) \times 1$ vector.

In this setup, the adaptive elastic-net GMM estimator is given as

$$\hat{\theta} = \left(1 + \frac{\lambda_2}{n^2}\right) \arg\min_{\theta} \left\{ (Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_1^* \sum_{j=1}^{p+s} \hat{w}_j |\theta_j| + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\}, \tag{3}$$

where $\hat{W}$ is some symmetric positive definite weight matrix, and λ_1^* and λ_2 are some positive tuning parameters. We have $\hat{w}_j = |\hat{\theta}_{j,enet}|^{-\gamma}$ with $\gamma > 1$ as the data dependent weight, where $\hat{\theta}_{j,enet}$ denotes the elastic-net estimator.[1] In practice, we run elastic-net and obtain data dependent weights $\hat{w}_j$ in the first stage; and we run the adaptive elastic-net using $\hat{w}_j$ in second stage. See Zou and Zhang (2009) for further details in the context of the least squares adaptive elastic-net estimator. An important point is that we use a finite sample correction of $1 + \lambda_2/n^2$, rather than the one $1 + \lambda_2/n$ used in Zou and Zhang (2009) and Caner and Zhang (2013). This is discussed in details in section 4 below.

We work with triangular arrays ξ_{in}, $\{i = 1, \cdots, n\}, \{n = 1, 2, 3, \cdots\}$ defined on the probability space $(\Omega, \mathcal{B}, P_n)$ where $P = P_n$ can change with n. At each $\xi_{in} = (X_{in}', Z_{in}', u_{in})'$, X_{in} is a $p \times 1$ vector, Z_{in} is a $q \times 1$ vector. Each of these vectors are independent across i, but they are not

[1] Note that the elastic-net objective function is given as (3) with $\hat{w}_j = 1$ for all j.

necessarily identically distributed. All parameters that characterize the distribution of ξ_{in} are implicitly indexed by P_n, and hence by n.

Leeb and Potscher (2005) make a very important point in the analysis on the case of local to zero parameters. They show that one cannot select the true model with probability approaching one uniformly. Their research has deep implications for post selection estimators, which have bi-modal empirical distribution functions due to the this uniformity problem. This is in the least squares framework when the interest centers on one set of coefficients and the other set are local to zero. We also allow local to zero parameters, and establish a lower bound for nonzero parameters to be selected as nonzero.

The conditions for theorems are presented below. Define for each $i = 1, 2 \cdots, n$, $e_i = Z_i u_i - F_{q,s}\tau_0$, $e = Z'u - nF_{q,s}\tau_0$. The first assumption is useful to prove Theorem 1.

Assumption 1. *(i)* $\|\hat{W} - W\| \xrightarrow{p} 0$*, where* W *is a* $q \times q$*, symmetric, positive definite and finite matrix. (ii)* $\{X_i, Z_i, u_i\}_{i=1}^n$ *are independent across* i*. Also, we have* $\|n^{-1}\sum_{i=1}^n e_i e_i' - V\| \xrightarrow{p} 0$*, where* V *is a* $q \times q$ *symmetric, positive definite and finite matrix. (iii)* $\|Z'X/n - \Sigma_{xz}\| \xrightarrow{p} 0$*, where* Σ_{xz} *is a* $q \times p$ *matrix of full column rank* p*. (iv) Eigmax* $\left(n^{-1}\hat{W}X_{zF}X_{zF}'\hat{W}n^{-1}\right) \leq B < \infty$

Assumption $1(i)$ is used in many weak moments literature. Specifically, a more restrictive version is used in Assumption $3(iii)$ in Newey and Windmeijer (2009). This type of assumption restricts how q grows with sample size n. Assumption $1(ii)$ is used for the estimation of the variance matrix. This is an infeasible estimator, but takes into account the effect of moment invalidity. Note that Assumption $1(iii)$ implies that

$$\|X_{zF}n^{-1} - \Sigma_{xzF}\| \xrightarrow{p} 0, \tag{4}$$

where $\Sigma_{xzF} = [\Sigma_{xz}, F_{q,s}]$ is a $q \times (p+s)$ matrix of full column rank $p+s$. In addition, W is nonsingular, symmetric and positive definite and Σ_{xzF} is of full rank. So we can show that,

$$0 < \text{Eigmin}(\Sigma_{xzF}'W\Sigma_{xzF}), \;\; \text{Eigmax}(\Sigma_{xzF}'W\Sigma_{xzF}) < \infty. \tag{5}$$

From Assumption 1 and results (4) and (5), we can show that there exists some positive absolute constants b and B which do not depend on n such that

$$0 < b \leq \text{Eigmin}\left(n^{-1}X_{zF}'\hat{W}X_{zF}n^{-1}\right),$$
$$\text{Eigmax}\left(n^{-1}X_{zF}'\hat{W}X_{zF}n^{-1}\right) \leq B < \infty \tag{6}$$

with probability approaching one (w.p.a.1, hereafter), by Lemma A0 of Newey and Windmeijer (2009). Assumption 1(iv) is needed to control the second moment of the estimators when there are many invalid instruments.

We impose further conditions. We let $\mathcal{A} = \{j : \theta_{j0} \neq 0, j = 1, 2, \cdots p+s\}$, which collects the indexes of nonzero coefficients in θ_0. Set $\eta = \min_{j\in\mathcal{A}} |\theta_{j0}|$, so η represents the minimum of nonzero (i.e. also allowing for local to zero coefficients) coefficients. Also set $p+s = O(n^\nu)$, where $0 \leq \nu \leq \alpha < 1$. Note that λ_1, λ_1^*, and λ_2 diverge to infinity when $n \to \infty$.

Assumption 2. *(i)* $\lambda_2(p+s)^{1/2}/n^{3/2} \to 0$ *and* $\lambda_1^2/n^3 \to 0$ *as* $n \to \infty$. *(ii) There exist absolute constants* α, γ *and* κ *satisfying* $0 \leq \alpha < 1$ *and* $3 + \alpha < \kappa < 2 + \gamma(1-\alpha) - \nu$. *(iii)* q *grows with sample size but* $q = O(n^\alpha)$ *and* $(p+s) \leq q$. *(iv)* $\frac{(\lambda_1^*)^2}{n^2}\frac{p+s}{\eta^{2\gamma}} \to 0$ *and* $\frac{\lambda_1^{*2}}{n^{\kappa-\gamma(1-\alpha)}} \to \infty$ *as* $n \to \infty$.

Assumption 3. $n^{-1}\max_{1\leq i\leq n} \|Z_i u_i - F_{q,s_0}\tau_{\mathcal{A}}\|^2 = o_p(1)$, *where* s_0 *is the true number of invalid instruments with* $0 \leq s_0 \leq s$, $F_{q,s_0} = [0'_{q-s_0,s_0}, I_{s_0}]'$, *and* $\tau_{\mathcal{A}}$ *is an* $s_0 \times 1$ *nonzero vector that represents invalid moment conditions.*

Assumption 2 establishes the rates for tuning parameters as well as the number of orthogonality restrictions, which are useful to prove Theorem 2. α controls the rate of the number of moment conditions. Also, it is noteworthy that we need $3+\alpha < \kappa < 2+\gamma(1-\alpha)-\nu$, which is used in the proof of Theorem 2. κ is mainly a parameter of technical nature and needed for selection consistency. To see that Assumption 2(ii) is not restrictive, consider the following system with $q = p+s$. The total number of moments is equal to the sum of the number of structural parameters and the number of potential invalid moments. Hence, $\alpha = \nu$ in our example. If $\alpha = \nu = 1/2$ with $\gamma = 5$ we can have $\kappa = 3.75$. Obviously, Assumption 2(ii) is satisfied. This is an example that the number of total moments grows with square root of the sample size as we all as all structural and invalidity parameters. The model is just identified in this example. Also, by putting $\gamma = 5$ we penalize the small coefficients a lot in the first stage of elastic net. This may be due to suspecting a lot of zeros a priori in the problem. Assumption 2(iv) has an important implication that we are able to come up with lower bounds on local to zero coefficients to be selected. In other words, in theorems below with Assumption 2(iv), we show that anything above the lower bound can be selected as nonzero. We also show that the lower bound depends on the number of parameters, the number of invalid moments, and the number of valid moments. If either the number of moments or parameters increases, then the lower bound becomes larger, meaning that only larger local to zero coefficients can be selected. This is an extension of Leeb and Pőtscher's (2005) result to

many invalid moments/parameters case. To obtain this bound, we first set $p+s = O(n^{\nu})$, where $0 < \nu \leq \alpha$, and assume $\eta = O(n^{-1/m})$ with $m > 0$. Assumption 2(iv) implies a lower bound for m which will be shown next. Assumption 2(iv) conditions are

$$\frac{(\lambda_1^*)^2}{n^{\kappa}} n^{\gamma(1-\alpha)} \to \infty,$$

and

$$\frac{(\lambda_1^*)^2}{n^2}\frac{p+s}{\eta^{2\gamma}} = \frac{(\lambda_1^*)^2}{n^2} n^{\nu+\frac{2\gamma}{m}} \to 0.$$

Hence, the only way so that both conditions hold is

$$\frac{n^{\gamma(1-\alpha)-\kappa}}{n^{\nu+\frac{2\gamma}{m}-2}} \to \infty,$$

which is possible if $\gamma(1-\alpha) - \kappa + 2 > \nu + 2\gamma/m$ or

$$m > \frac{2\gamma}{\gamma(1-\alpha) - \nu - \kappa + 2} = m^*. \tag{7}$$

(7) shows a lower bound for m, which will become a lower bound for η to be selected as a nonzero coefficient in Theorem 3. Clearly for a larger α or ν, m^* becomes larger and the lower bound for nonzero coefficients to be selected becomes larger as well. As a minor note, in order to have $m > 0$, we need $\gamma(1-\alpha) - \nu - \kappa + 2 > 0$, but this means that $(\gamma - \nu - \kappa + 2)/\gamma > \alpha$, which is already satisfied by Assumption 2(ii). Note that with the above example for Assumption 2(ii) with $\kappa = 3.75, \alpha = \nu = 1/2, \gamma = 5$ we get $m^* = 40$, so the coefficient is local to zero but much larger than the $\sqrt{n}$ rate.

Assumption 3 is useful to obtain the Lyapunov condition in Theorem 4, which is similar to the assumption used in least squares case of Zou and Zhang (2009).

3 Adaptive Elastic-Net GMM Estimation

We first obtain one of the main results of the paper by deriving the upper bounds of the mean square error of the estimates. We obtain the bound for the adaptive elastic net and the bound for the elastic net estimator where $\hat{w}_j = 1$, $\forall j = 1, 2 \cdots, p$. Given the data (y_i, X_i, Z_i), let $\hat{\omega} = (\hat{\omega}_1, \cdots, \hat{\omega}_{p+s})$ be a vector whose components are all nonnegative and can depend on data. Set $\theta = (\beta', \tau')'$. Then define

$$\hat{\theta}_W = argmin_{\theta}\{(Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_2\|\theta\|^2 + \lambda_1 \sum_{j=1}^{p+s} \hat{w}_j|\theta_j|\}$$

where λ_1 and λ_2 are nonnegative tuning parameters.

If we substitute $\hat{\omega}_j = 1$, $j = 1, \cdots, p+s$, in $\hat{\theta}_W$ above, then we obtain the elastic net estimator and denote it as $\hat{\theta}_{enet}$.

Theorem 1. *Under the model (1), (2) and Assumption 1, we have w.p.a.1*

$$\begin{aligned} \textit{(i)}\ E\|\hat{\theta}_W - \theta_0\|^2 &\leq 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2 E(\sum_{j=1}^{p+s}\hat{w}_j^2)}{(bn^2+\lambda_2)^2} \textit{ and} \\ \textit{(ii)}\ E\|\hat{\theta}_{enet} - \theta_0\|^2 &\leq 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2}, \end{aligned}$$

B and b are some positive absolute constants given in (6).

This result clearly shows the upper bound on the mean square error of our estimators and is used to obtain Theorems 2 and 3. [2]

Next we obtain the selection consistency. This result is important since it shows that the adaptive elastic-net procedure automatically selects the valid moment conditions as well as the relevant regressors in the structural equation. We further define an estimator given by

$$\tilde{\theta}_{\mathcal{A}} = \arg\min_{\theta} \left\{ (Y_z - X_{zF\mathcal{A}}\theta)'\hat{W}(Y_z - X_{zF\mathcal{A}}\theta) + \lambda_1^* \sum_{j\in\mathcal{A}} \hat{w}_j|\theta_j| + \lambda_2 \sum_{j\in\mathcal{A}} \theta_j^2 \right\}, \tag{8}$$

where $X_{zF\mathcal{A}}$ consists of the sub-columns of X_{zF} that correspond to nonzero elements in $\theta_0 = (\beta_0', \tau_0')'$. The following result is useful to derive the selection consistency.

Theorem 2. *Under Assumptions 1-2, w.p.a.1, $((1 + (\lambda_2/n^2))\tilde{\theta}_{\mathcal{A}}, 0)$ is the solution to the minimization problem of adaptive elastic-net in (3).*

The next theorem obtains the selection consistency of the adaptive elastic-net estimator. This extends Zou and Zhang (2009) from finding the relevant regressors in least squares to linear GMM. We also find the invalid moments compared to their case.

Theorem 3. *Under Assumptions 1-2, the adaptive elastic-net estimator $\hat{\theta}$ in (3) satisfies the selection consistency property: $P(\{j : \hat{\theta}_j \neq 0\} = \mathcal{A}) \to 1$.*

The main difference between Theorems 2 and 3 is that we can get local to zero coefficients as nonzero above a certain threshold in Theorem 3. The minimum coefficient that can be selected correctly should be of the order $n^{-1/m}$, where $m > m^*$ in (7). This shows that in an environment with many moments/parameters, it will be difficult to do perfect model selection if the coefficients

[2] Another distinction is that the bound results by Zou and Zhang (2009) are exact since they take the regressors to be deterministic. In contrast, our result is obtained w.p.a.1 since we consider the stochastic regressors.

are small. To give an example, take $\nu = 1/5, \alpha = 2/5, \gamma = 3, \kappa = 3.5$, then $m^* = 60$ which means the order of the smallest coefficient to be selected should be larger than $n^{-1/60}$. This theorem extends Leeb and Pőtscher's (2005) criticism to the many parameters context. In the case with a fixed number of parameters, they found that the order of the minimum coefficient to be selected should be larger than $n^{-1/2}$.

In addition, we provide the limit distribution of the adaptive elastic-net estimator of the nonzero parameters $\theta_{\mathcal{A}} = (\beta'_{\mathcal{A}}, \tau'_{\mathcal{A}})'$. Without losing any generality, we denote the true number of nonzero structural parameters as p_0 with $1 \leq p_0 \leq p$ and the true number of invalid instruments as s_0 with $1 \leq s_0 \leq s$, so that $\beta_{\mathcal{A}}$ is $p_0 \times 1$ and $\tau_{\mathcal{A}}$ is $s_0 \times 1$. We further define a $(p_0 + s_0) \times (p_0 + s_0)$ matrix

$$\Sigma_{\mathcal{A}} = \Sigma'_{xzF\mathcal{A}} V^{-1} \Sigma_{xzF\mathcal{A}},$$

where $\Sigma_{xzF\mathcal{A}} = [\Sigma_{xz\mathcal{A}}, F_{q,s_0}]$ is a full column rank $q \times (p_0 + s_0)$ matrix and $\Sigma_{xz\mathcal{A}}$ is a full column rank $q \times p_0$ matrix. $F_{q,s_0} = [0'_{q-s_0,s_0}, I_{s_0}]'$ is a $q \times s_0$ matrix that is defined similarly to $F_{q,s}$ above. Note that $\Sigma_{xz\mathcal{A}}$ is defined from $\|Z'X_{\mathcal{A}}/n - \Sigma_{xz\mathcal{A}}\| \xrightarrow{p} 0$, which holds from Assumption 1-(iii), where $X_{\mathcal{A}}$ is an $n \times p_0$ matrix that consists of the (endogenous) regressors corresponding to the nonzero structural parameters. Then using a similar argument as (4), we have

$$\|X_{zF\mathcal{A}} n^{-1} - \Sigma_{xzF\mathcal{A}}\| \xrightarrow{p} 0. \tag{9}$$

Now we introduce one of the main theorems.

Theorem 4. *We let $\hat{\theta}_{\mathcal{A}}$ be the adaptive elastic-net GMM estimator in (3) that corresponds to $\theta_{\mathcal{A}}$. Under Assumptions 1-3, the limit distribution of $\hat{\theta}_{\mathcal{A}}$ is given by*

$$\zeta' \frac{\left(I_{p_0+s_0} + \lambda_2 \hat{\Sigma}_{\mathcal{A}}^{-1}\right)}{1 + (\lambda_2/n^2)} \hat{\Sigma}_{\mathcal{A}}^{1/2} n^{-1/2} (\hat{\theta}_{\mathcal{A}} - \theta_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty,$$

where $\hat{\Sigma}_{\mathcal{A}} = X'_{zF\mathcal{A}} \hat{V}^{-1} X_{zF\mathcal{A}}$, $\hat{V}$ is some consistent estimator of V, and ζ is an arbitrary $(p_0+s_0)\times 1$ vector with $\|\zeta\| = 1$.

Remarks: 1. Note that from (9) and Assumption 1, it can be verified that the minimum eigenvalue of $\hat{\Sigma}_{\mathcal{A}}$ is $O_p(n^2)$ and the maximum eigenvalue of $\hat{\Sigma}_{\mathcal{A}}^{1/2}$ is $O_p(n)$. By Assumption 2, we have $\|\lambda_2 \hat{\Sigma}_{\mathcal{A}}^{-1}\| \xrightarrow{p} 0$. Therefore, we obtain

$$\left\| \frac{I_{p_0+s_0} + \lambda_2 \hat{\Sigma}_{\mathcal{A}}^{-1}}{1 + (\lambda_2/n^2)} - I_{p_0+s_0} \right\| = o_p(1)$$

as $\lambda_2/n^2 \to 0$ with $\|\zeta\| = 1$. ζ is a $(p_0 + s_0)$ vector, which indicates the rate of convergence of $\hat{\theta}_{\mathcal{A}}$ equal to $\sqrt{n/(p_0 + s_0)}$. The rate of convergence of the structural parameters and the invalid moment parameters is slower than $\sqrt{n}$, which is affected by the number of invalid moments.

2. Caner and Zhang (2013) also obtain the asymptotics of adaptive elastic net estimators in a GMM framework. However, their exercise is relatively limited in the sense that they only analyze structural parameters and assume that all the moments are valid.

3. An interesting question is the analysis of many weak moments. In GMM case, we know from the work of Newey and Windmeijer (2009) that this is an inconsistent estimator. Only GEL estimators will be consistent. For LASSO type estimators, the same problem is pointed out by Caner (2009). Caner (2009) shows that with a fixed number of instruments, only nearly-weak asymptotics can give consistent estimates. We think that the case of many weak moments will be very interesting but has to be handled in GEL or CUE framework, so its analysis is beyond the scope of this paper.

Another interesting question is whether we can achieve the semiparametric efficiency bound from the adaptive elastic-net procedure. Note that it is generally the case if we use the entire set of valid (and strong) instruments. The following result shows that the adaptive elastic-net GMM estimator of the nonzero structural parameter β indeed achieves the semiparametric efficiency bound. Therefore, even with many invalid moments, it is still possible to construct an estimator that reaches the semiparametric efficiency bound. We let $Z = (Z_1, Z_2)$, where Z_1 represents the $n \times (q - s_0)$ valid instruments, and Z_2 represents $n \times s_0$ invalid instruments. More precisely, $\|n^{-1} \sum_{i=1}^n Z_{1i} u_i\| \xrightarrow{p} 0$ and $\|n^{-1} \sum_{i=1}^n Z_{2i} u_i - \tau_{\mathcal{A}}\| \xrightarrow{p} 0$, where $\tau_{\mathcal{A}}$ is an $s_0 \times 1$ vector whose elements are all nonzero.

Theorem 5. *Under Assumptions 1-3 the limit variance of the true nonzero structural parameter estimator $\hat{\beta}_{\mathcal{A}}$ is*

$$(\Sigma'_{xz1\mathcal{A}} V_{11}^{-1} \Sigma_{xz1\mathcal{A}})^{-1},$$

where $\|Z_1' X_{\mathcal{A}}/n - \Sigma_{xz1\mathcal{A}}\| \xrightarrow{p} 0$ and $\|n^{-1} \sum_{i=1}^n Z_{1i} Z_{1i}' u_i^2 - V_{11}\| \xrightarrow{p} 0$.

This result implies that even though we have some invalid instruments and there may be many of them, we can still estimate β as if we were using only the valid instruments. It can be done by one-step estimation (i.e., the adaptive elastic-net GMM) instead of using some two-step estimation depending on pre-testing for the instruments validity. This is the oracle result.

3.1 An Application to Dynamic Panel Data Estimation

As an application, we consider the following dynamic panel regression model given by

$$y_{i,t} = \rho y_{i,t-1} + x_{i,t}'\beta + \mu_i + u_{i,t} \tag{10}$$

for $i = 1, \cdots, N$ and $t = 1, \cdots, T$, where $|\rho| < 1$, $y_{i,t}$ is a scalar, $x_{i,t}$ is a $K \times 1$ vector of exogenous regressors and μ_i is the unobserved individual effects that can be correlated with $y_{i,t-1}$ or $x_{i,t}$. Under the condition that

$$E[u_{i,t}|\mu_i, y_i^{t-1}, x_i^T] = 0, \tag{11}$$

where $y_i^{t-1} = (y_{i,1}, \cdots, y_{i,t-1})'$ and $x_i^T = (x_{i,1}', \cdots, x_{i,T}')'$, we have the moment conditions given by

$$E[\Delta x_{i,t}\Delta u_{i,t}] = E[\Delta x_{i,t}(\Delta y_{i,t} - \rho\Delta y_{i,t-1} - \Delta x_{i,t}'\beta)] = 0 \tag{12}$$

$$E[y_i^{t-2}\Delta u_{i,t}] = E[y_i^{t-2}(\Delta y_{i,t} - \rho\Delta y_{i,t-1} - \Delta x_{i,t}'\beta)] = 0 \tag{13}$$

for $t \geq 2$ as Arellano and Bond (1991). But note that the second set of $(T-2)(T-1)/2$ number of moment conditions (13) heavily depend on the condition that $E[u_{i,t}u_{i,t-s}] = 0$ for all $s \geq 1$, which is indeed implied by (11), whereas the first set of $(T-2)K$ number of moment conditions (12) is robust to the possible serial correlation in $u_{i,t}$. Therefore, if the error term $u_{i,t}$ in (10) has serial correlation, then some moment conditions in (13) become invalid.[3]

In this case, we have $q \equiv (T-2)(T-1)/2 + (T-2)K$ number of total moment conditions, whereas we have $p \equiv K+1$ number of parameters of interest. Among the q number of moment conditions (or instruments) we have $s \equiv (T-2)(T-1)/2$ number of moment conditions that are potentially invalid under the possible serial correlation in $u_{i,t}$, which indeed happens frequently in practice. We allow for $N, T, K \to \infty$ and thus $q, s, p \to \infty$ in this case. For identification purposes, we assume $p + s \leq q \iff K + 1 \leq (T-2)K$ for all T and K, which is satisfied with $T \geq 4$ and $K \geq 1$.

Note that one of the condition on q is $q = O(n^\alpha)$ for some $0 \leq \alpha < 1$, and thus $q/n \to 0$ as the sample size grows. In this dynamic panel case, we have $q/n = [(T-2)(\frac{T-1}{2} + K)]/NT =$

[3]Under an additional condition of the mean stationarity (i.e., $E[y_{i,t}] = \mu$ for all i and t), we further have $E[\Delta y_{i,t-1}(y_{i,t} - \rho y_{i,t-1} - x_{i,t}'\beta)] = 0$ for $t \geq 2$ as Blundell and Bond (1998) and Bun and Kleibergen (2013). When ρ is close to one, the moment condition (13) is prone to have weak identification (i.e., weak instrument problem) whereas this new moment condition is robust to such a persistence. We could find more moment conditions (e.g., $E[\Delta x_{i,s}\Delta v_{i,t}] = 0$ for $s = 2, \cdots, T$ under strict exogeneity of $x_{i,t}$ or second moment restrictions with homoskedasticity assumption), but we only consider the most conventional moment conditions given as (13).

$O(\max\{K,T\}/N)$ and thus we need $\max\{K,T\}/N \to 0$ as $N,T,K \to \infty$. However, in general the (system) GMM approach using first-differenced panel is normally used in large cross section case (i.e., $N \gg T$). Unless K is extremely large, this condition is usually satisfied in practice.

More precisely, we consider the vector form of the first-differenced equation of (10) as $\Delta y_{i,t} = \rho \Delta y_{i,t-1} + \Delta x_{i,t}'\beta + \Delta u_{i,t}$ or in a matrix form

$$\Delta y_i = X_i \delta + \Delta u_i,$$

where $\Delta y_i = (\Delta y_{i,3}, \cdots, \Delta y_{i,T})'$, $X_i = [\Delta y_{i(-1)}, \Delta x_i]$ with $\Delta y_{i(-1)} = (\Delta y_{i,2}, \cdots, \Delta y_{i,T-1})'$ and $\Delta x_i = (\Delta x_{i,3}, \cdots, \Delta x_{i,T})'$, $\delta = (\rho, \beta')'$ and $\Delta u_i = (\Delta u_{i,3}, \cdots, \Delta u_{i,T})'$. We denote the $((t-2)+K) \times 1$ instrumental variable as

$$z_{i,t} = \begin{pmatrix} y_i^{t-2} \\ \Delta x_{i,t} \end{pmatrix}.$$

Note that with possible serial correlation in $u_{i,t}$, we have the following set of moment conditions in this case:

$$E\left[z_{i,t}(\Delta y_{i,t} - \rho \Delta y_{i,t-1} - \Delta x_{i,t}'\beta) - \begin{pmatrix} \tau_{t-2} \\ 0_K \end{pmatrix} \right] = 0$$

for all $i = 1, \cdots, N$ and for each $t = 3, \cdots, T$, where τ_{t-2} is some $(t-2) \times 1$ vector. We have $Z_i = [Z_{1i}, Z_{2i}]$, where

$$Z_{1i} = \begin{pmatrix} \Delta x_{i,3}' & 0 & \cdots & 0 \\ 0 & \Delta x_{i,4}' & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & \Delta x_{i,T}' \end{pmatrix}_{(T-2)\times(T-2)K} \quad \text{and}$$

$$Z_{2i} = \begin{pmatrix} y_{i,1} & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & y_{i,1} & y_{i,2} & & 0 & & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & \cdots & y_{i,1} & \cdots & y_{i,T-2} \end{pmatrix}_{(T-2)\times(T-2)(T-1)/2}.$$

Then the adaptive elastic-net GMM estimator is given by

$$\begin{aligned} \widehat{\theta} = & \left(1 + \frac{\lambda_2}{(NT)^2}\right) \arg\min_{\theta=(\delta,\tau)} \left\{ \sum_{i=1}^{N} (Z_i'(\Delta y_i - X_i\delta) - F_{q,s}\tau)' \hat{W} (Z_i'(\Delta y_i - X_i\delta) - F_{q,s}\tau) \right. \\ & \left. + \lambda_1^* \sum_{j=1}^{p+s} \hat{w}_j |\theta_j| + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\} \end{aligned} \tag{14}$$

from (3), where $q = (T-2)(T-1)/2 + (T-2)K$, $s = (T-2)(T-1)/2$, $p = K+1$ and $\tau = (\tau_1', \cdots, \tau_{T-2}')'$. Assuming that the data $\{y_i, x_i\}_{i=1}^N$ are i.i.d. across i, the theoretical results in the previous section extend to this example. Note that we choose $\hat{w}_j$ such that $\hat{w}_j = |\hat{\theta}_{j,enet}|^{-\gamma}$ for some $\gamma > 1$, where $\hat{\theta}_{j,enet}$ is the elastic-net estimator that minimizes (14) with $\hat{w}_j = 1$ for all j. The optimal weight matrix $\hat{W}$ can be obtained via the standard two-step GMM estimation.

4 Algorithm for Optimization and Tuning Parameter Selection

We first start this section by showing that the LARS algorithm of Efron *et al.* (2004) can be applied in the linear adaptive elastic net estimator for least squares, and then we show that it can be applied to adaptive elastic net for GMM. We extend Lemma 1 of Zou and Hastie (2005) from the elastic net to adaptive elastic net by using Algorithm 1 in section 3.5 of Zou (2006). It shows that the adaptive elastic net in linear models can be optimized as LASSO. Consider the linear regression model:

$$y = x\phi + \varepsilon$$

where y is $n \times 1$, $x = [x_1, x_2, ..., x_r]$ is $n \times r$ and ϕ is $r \times 1$. The naive elastic net estimator in a linear regression is

$$\hat{\phi}_{nenet} = argmin_\phi \{\|y - \sum_{j=1}^r x_j'\phi_j\|^2 + \lambda_1 \sum_{j=1}^r \hat{w}_j|\phi_j| + \lambda_2 \sum_{j=1}^r \phi_j^2\}. \tag{15}$$

The naive elastic net is the adaptive elastic net without the extra scaling of $(1+\lambda_2/n)$. Now, form the following LASSO problem

$$\hat{\phi}^* = argmin_\phi \{\|y^* - \sum_{j=1}^r x_j^{*\prime}\phi_j^*\|^2 + \lambda_1^* \sum_{j=1}^r |\phi_j^*|\}, \tag{16}$$

where for $j = 1, \cdots, r$, $\phi_j^* = \hat{w}_j\phi_j$, and

$$x_j^* = \hat{w}_j^{-1} \begin{pmatrix} x_j \\ \sqrt{\lambda_2} e_j \end{pmatrix}.$$

Note that x_j is an $(n+r) \times 1$ vector, and $e_j = (0, \cdots 1, \cdots 0)'$, where the jth element is 1, and all other elements are 0. Also, we set the $(n+r)-$dimensional vector y^* as

$$y^* = \begin{pmatrix} y \\ 0_{r\times 1} \end{pmatrix}.$$

Lemma 1. *Given the above data (y, x_j) and transformed data (y^*, x_j^*) for $j = 1, \cdots, r$, the relation between the naive elastic net estimator $(\hat{\phi}_{nenet})$ and the LASSO estimator $(\hat{\phi}^*)$ is*

$$\hat{\phi}_{nenet,j} = \hat{w}_j^{-1} \hat{\phi}_j^*.$$

Now, we return to the discussion on our adaptive elastic net estimator in the GMM framework. The naive elastic net estimator $\hat{\theta}_{nenet}$ can be computed by substituting $y = \hat{W}^{1/2} Y_z$, $x = \hat{W}^{1/2} X_{zF}$ and $\hat{w}_j = |\hat{\theta}_{j,enet}|^{-\gamma}$ into (15). $\hat{W}$ can be computed via the conventional efficient two-step GMM. Given the adaptive weight $\hat{w}_j$, y and x are transformed into y^* and x^*, and the transformed coefficient $\hat{\theta}_j^*$ can be computed by the LARS algorithm as in (16). The naive elastic net estimator is thus defined as $\hat{\theta}_{nenet,j} = \hat{w}_j^{-1} \hat{\theta}_j^*$.

Then our adaptive elastic net estimator is computed as

$$\hat{\theta} = \left(1 + \frac{\lambda_2}{n^2}\right) \hat{\theta}_{nenet}.$$

The naive elastic net estimator $\hat{\theta}_{nenet}$ is rescaled by $1+\lambda_2/n^2$ instead of $1+\lambda_2/n$. To see the reason, note that the maximum eigenvalue of $x'x$ is $O_p(n)$ in Zou and Zhang's (2009) linear models, whereas the maximum eigenvalue of $X_{zF}' \hat{W} X_{zF}$ is $O_p(n^2)$ by (6) in our GMM setup. If only L_2 penalty were used, then we would get the ridge estimate for θ:

$$\left(\frac{X_{zF}' \hat{W} X_{zF}}{n^2} + \frac{\lambda_2}{n^2} I_{p+s}\right)^{-1} \frac{X_{zF}' \hat{W} Y_z}{n^2}.$$

It is clear that the normalization involves n^2 rather than n, so we suggest using $1 + \lambda_2/n^2$ to the scale the naive elastic net estimator in the GMM context.

Let $\hat{\theta}$ be partitioned as $\hat{\theta} \equiv [\hat{\beta}', \hat{\tau}']'$, where $\hat{\beta}$ and $\hat{\tau}$ are the estimates for β_0 and τ_0, respectively. Let $\hat{p}_0 \equiv \|\hat{\beta}\|_0$ and $\hat{s}_0 \equiv \|\hat{\tau}\|_0$, where $\|.\|_0$ denotes the number of nonzero elements of a vector. We select the tuning parameters by minimizing the following criterion,

$$IC(\lambda_1^*, \lambda_2) = J(\hat{\theta}) + (\hat{p}_0 + \hat{s}_0) \ln(n) \max\{\ln[\ln(p+s)], 1\}, \tag{17}$$

where $J(\hat{\theta}) = n^{-1}(Y_z - X_{zF}\hat{\theta})' \hat{W} (Y_z - X_{zF}\hat{\theta})$, and we use abbreviated notations, $\hat{\theta}$, $\hat{p}_0$ and $\hat{s}_0$, to denote $\hat{\theta}(\lambda_1^*, \lambda_2)$, $\hat{p}_0(\lambda_1^*, \lambda_2)$ and $\hat{s}_0(\lambda_1^*, \lambda_2)$. Wang *et al.* (2009) show that BIC can be applied to select the tuning parameter that produces correct model selection w.p.a.1 for shrinkage estimation of linear models. The criterion (17) is an analog of BIC proposed by Wang *et al.* (2009) under our adaptive elastic net GMM setup. Recall that we set $y = \hat{W}^{1/2} Y_z$ and $x = \hat{W}^{1/2} X_{zF}$ to transform (3) into a linear regression so that the LARS algorithm can be applied. Thus, $J(\hat{\theta}) =$

$n^{-1}(y - x\hat{\theta})'(y - x\hat{\theta})$ corresponds to the mean of squared errors in BIC proposed by Wang *et al.* (2009). $J(\hat{\theta})$ in (17) prevents under-fitting. Note that there will be endogeneity and $J(\hat{\theta})$ will diverge if any nonzero element of β_0 or τ_0 is estimated as zero. The second term in (17) prevents over-fitting. The term $\ln[\ln(p+s)]$ follows the suggestion by Wang *et al.* (2009) for a diverging number of parameters.

5 Monte Carlo Simulation

In this section, we study the finite sample performance of our estimator. Let ι_j denote a $j \times 1$ vector of ones. We consider the following data generating processes (DGPs) for $i = 1, ..., n$.

$$\begin{aligned} Y_i &= X_i'\beta_0 + u_i \\ X_i &= Z_{1i}'\pi + v_i \\ u_i &= \sqrt{\rho_{uv}}\varepsilon_{1i} + \sqrt{1-\rho_{uv}}\varepsilon_{2i} \\ v_i &= \sqrt{\rho_{uv}}\varepsilon_{1i} \cdot \iota_p + \sqrt{1-\rho_{uv}}\varepsilon_{3i} \end{aligned}$$

where Z_{1i} is a $(q - s_0) \times 1$ vector of valid instruments, $Z_{1i} \overset{i.i.d.}{\sim} N(0, \Omega_z)$, Z_{2i} is an $s_0 \times 1$ vector of invalid instruments, $Z_{2i} = \varepsilon_{4i} + \tau_{\mathcal{A}} u_i \cdot \iota_{s_0}$ and

$$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \overset{i.i.d.}{\sim} N\left(0, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & I_p & 0 \\ 0 & 0 & 0 & I_{s_0} \end{bmatrix}\right).$$

Let $Z_i \equiv [Z_{1i}', Z_{2i}']'$. We set $\rho_{uv} = 0.5$, $2p = q - s_0$, $\beta_0 = b(\iota_{p_0}', 0_{1\times(p-p_0)})'$ and

$$\pi = \frac{1}{\sqrt{2}} \begin{bmatrix} I_p \\ I_p \end{bmatrix}.$$

The (i,j)th element of Ω_z is set equal to $\rho_z^{|i-j|}$. We set $\rho_z = 0.5$ and 0.8 to vary the dependence among the valid instruments. $\tau_{\mathcal{A}}$ controls the severity of the invalid moment conditions. Note that $\Sigma_{xzF} = [\Sigma_{xz}, F_{q,s}]$ has full column rank even if Z_{2i} is generated to be uncorrelated with X_i. Cheng and Liao (2012) and Liao (2013) generate invalid instruments using a similar DGP. We set $\tau_{\mathcal{A}}$ equal to 0.5 and 0.3. The parameter b is the value of nonzero structural parameters and set equal to 0.25, 0.5 and 1. The sample size n is set equal to 250 and 1000, and $p = 20$, $p_0 = 3$, $s = 10$, $s_0 = 3$ and $q = 43$. The number of replications is 2000.

We summarize the simulation results in Tables 1, 2, and 3. AENet is the estimator proposed in (3) and is solved by the algorithm provided in Section 4. ALASSO-LARS is the same as AENet except that λ_2 is restricted to be zero, so ALASSO-LARS is the adaptive LASSO GMM estimator solved by LARS. ALASSO-CL is the adaptive LASSO estimator proposed by Cheng and Liao (2012). The main difference between ALASSO-CL and our estimator is that ALASSO-CL does not select variables in the structural equation (1) but only selects moments in (2). Also, ALASSO-CL is solved by the algorithm proposed by Schmidt (2010), and the tuning parameter is selected by cross validation instead of (17). Let $\beta_{\mathcal{A}^c}$ and $\tau_{\mathcal{A}^c}$ denote the zero elements in β_0 and τ_0, respectively, so that $\beta_0 = (\beta'_{\mathcal{A}}, \beta'_{\mathcal{A}^c})'$ and $\tau_0 = (\tau'_{\mathcal{A}}, \tau'_{\mathcal{A}^c})'$.

Table 1 reports the root of mean squared errors (RMSE) of three estimators. RMSEs of estimators of $\tau_{\mathcal{A}^c}$, $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$, denoted by $rmse_1$, $rmse_2$, $rmse_3$ and $rmse_4$, respectively. First, the RMSEs of $\tau_{\mathcal{A}^c}$ are similar when $\rho_z = 0.5$. When $\rho_z = 0.8$, however, both AENet and ALASSO-LARS tend to estimate $\tau_{\mathcal{A}^c}$ more accurately than ALASSO-CL. For example, when $(\rho_z,\ \tau_{\mathcal{A}},\ b) = (0.8,\ 0.5,\ 0.25)$ and $n = 250$, the RMSEs of $\tau_{\mathcal{A}^c}$ for AENet, ALASSO-LARS, and ALASSO-CL are 0.006, 0.006, and 0.015, respectively. Second, for invalid moment conditions $\tau_{\mathcal{A}}$, AENet and ALASSO-LARS have smaller RMSEs than ALASSO-CL. For instance, when $(\rho_z,\ \tau_{\mathcal{A}},\ b) = (0.5,\ 0.5,\ 0.25)$ and $n = 250$, the RMSEs of $\tau_{\mathcal{A}}$ for AENet, ALASSO-LARS, and ALASSO-CL are 0.187, 0.181, and 0.388, respectively. Third, $rmse_3$ shows that AENet and ALASSO-LARS produce smaller RMSEs than ALASSO-CL. This is expected because ALASSO-CL does not shrink estimates to zero for $\beta_{\mathcal{A}^c}$. Lastly, for nonzero structural parameters $\beta_{\mathcal{A}}$, none of these estimator outperforms others uniformly. AENet performs better than ALASSO-LARS when instruments are highly correlated ($\rho_z = 0.8$). Hence, the introduction of ridge penalty can reduce the severity of multicollinearity. The RMSE of ALASSO-CL is smaller than the other two estimators when $\beta_{\mathcal{A}} = 0.25 \cdot \iota_{p_0 \times 1}$.

Table 2 reports the accuracy of moment selection by different estimators. Pr_1 is the percentage of replications that yield zero estimates for $\tau_{\mathcal{A}^c}$. Pr_2 is the percentage of replications that yield nonzero estimates for $\tau_{\mathcal{A}}$. First, for the unsure-but-valid moments, ALASSO-CL is slightly better than AENet when $n = 250$ and $\rho_z = 0.5$. When $n = 250$ and $\rho_z = 0.8$, however, our estimators outperform ALASSO-CL. For example, AENet estimates $\tau_{\mathcal{A}^c}$ as zero for 98.1% of replications, whereas ALASSO-CL estimates $\tau_{\mathcal{A}^c}$ as zero for 92% of replications when $(\rho_z, \tau_{\mathcal{A}}, b) = (0.8, 0.3, 0.25)$. Second, for invalid moments, our estimators always outperform ALASSO-CL except for the cases where $\rho_z = 0.8$, $\tau_{\mathcal{A}} = 0.3$, and $n = 250$. It is expected that detecting invalid moments is difficult

when $\tau_{\mathcal{A}}$ is small. When $(\rho_z, \tau_{\mathcal{A}}, b) = (0.5, 0.3, 0.25)$ and $n = 250$, for example, ALASSO-CL only detects 35% of the invalid moments, but AENet detects 86.1%. As n increases to 1000, our estimators always capture all the invalid moments.

Table 3 reports the accuracy of structural parameter selection by different estimators. Pr_3 is the percentage of replications that yield zero estimates for $\beta_{\mathcal{A}^c}$. Pr_4 is the percentage of replications that yield nonzero estimates for $\beta_{\mathcal{A}}$. Note that ALASSO-CL cannot select irrelevant variables in (1) since it only focuses on selecting moments, i.e., all β's will be estimated as nonzero by ALASSO-CL. In addition, AENet is slightly better than ALASSO-LARS in terms of the percentage that yields zero estimate for $\beta_{\mathcal{A}^c}$. Moreover, for the selection of nonzero structural parameters, AENet tends to outperform ALASSO-LARS, especially for $n = 250$. It is also expected from Theorem 3 that selecting relevant regressors becomes more difficult when $\beta_{\mathcal{A}}$ is small. This can be seen in the case where AENet only detects around 70% of the relevant regressors with $n = 250$, $\rho_z = 0.8$, and $b = 0.25$. When $n = 1000$, $\rho_z = 0.8$, and $b = 0.25$, however, our estimators can detect over 90% of the relevant regressors with small coefficients.

6 Conclusion

This paper develops an adaptive elastic-net estimator with many possibly invalid moment conditions. The number of structural parameters as well as the number of moment conditions are allowed to increase with the sample size. The moment selection and model selection are conducted simultaneously. The moment conditions are constructed in a way to take into account the possibly invalid instruments. We use the penalized GMM to estimate both structural parameters along with the parameters associated with the invalid moments. The penalty contains two terms: the quadratic regularization and the adaptively weighted LASSO penalty. We show that our estimator uses information only from the valid moment conditions to achieve the semiparametric efficiency bound. The estimator is thus very useful in practice since it conducts the consistent moment selection and efficient estimation of the structural parameters simultaneously. We also establish the order of magnitude for the smallest local to zero coefficient to be selected as nonzero. An algorithm is proposed based on LARS for the implementation of our estimator. Simulation results show that our estimator has good finite-sample performance.

Appendix: Mathematical Proofs

Proof of Theorem 1 We define a ridge type estimator

$$\hat{\theta}_R = \arg\min_{\theta} \left\{ (Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\}. \tag{A.1}$$

We will benefit from the following inequality:

$$E\|\hat{\theta}_W - \theta_0\|^2 \leq 2E\|\hat{\theta}_R - \theta_0\|^2 + 2E\|\hat{\theta}_W - \hat{\theta}_R\|^2. \tag{A.2}$$

We try to bound the term $E\|\hat{\theta}_W - \hat{\theta}_R\|^2$. For that reason note that

$$\begin{aligned} & (Y_z - X_{zF}\hat{\theta}_W)'\hat{W}(Y_z - X_{zF}\hat{\theta}_W) + \lambda_1 \sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,W}| + \lambda_2 \sum_{j=1}^{p+s} \hat{\theta}_{j,W}^2 \\ \leq \quad & (Y_z - X_{zF}\hat{\theta}_R)'\hat{W}(Y_z - X_{zF}\hat{\theta}_R) + \lambda_1 \sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,R}| + \lambda_2 \sum_{j=1}^{p+s} \hat{\theta}_{j,R}^2, \end{aligned} \tag{A.3}$$

which is derived from the definition of $\hat{\theta}_W$ in the statement of Theorem 1. But we can rewrite (A.3) as

$$\begin{aligned} \lambda_1 \sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,R}| - \lambda_1 \sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,W}| \quad \geq \quad & [(Y_z - X_{zF}\hat{\theta}_W)'\hat{W}(Y_z - X_{zF}\hat{\theta}_W) + \lambda_2 \sum_{j=1}^{p+s} \hat{\theta}_{j,W}^2] \\ & -[(Y_z - X_{zF}\hat{\theta}_R)'\hat{W}(Y_z - X_{zF}\hat{\theta}_R) + \lambda_2 \sum_{j=1}^{p+s} \hat{\theta}_{j,R}^2], \end{aligned} \tag{A.4}$$

where it holds that

$$\sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,R}| - \sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,W}| \leq \sum_{j=1}^{p+s} \hat{w}_j|\hat{\theta}_{j,R} - \hat{\theta}_{j,W}| \leq \left[\sum_{j=1}^{p+s} \hat{w}_j^2\right]^{1/2} \|\hat{\theta}_R - \hat{\theta}_W\|. \tag{A.5}$$

Now we try to get a lower bound for the right hand side of (A.4). So we find the ridge solution from (A.1) as

$$\hat{\theta}_R = [(X'_{zF}\hat{W}X_{zF}) + \lambda_2 I_{p+s}]^{-1}[X'_{zF}\hat{W}Y_z] \tag{A.6}$$

yielding

$$\begin{aligned} & (Y_Z - X'_{zF}\hat{\theta}_R)'\hat{W}(Y_Z - X'_{zF}\hat{\theta}_R) + \lambda_2\|\hat{\theta}_R\|^2 \\ = \quad & Y'_z\hat{W}Y_z - 2\hat{\theta}'_R X'_{zF}\hat{W}Y_z + \hat{\theta}'_R[X'_{zF}\hat{W}X_{zF} + \lambda_2 I_{p+s}]\hat{\theta}_R \\ = \quad & Y'_z\hat{W}Y_z - \hat{\theta}'_R[X'_{zF}\hat{W}X_{zF} + \lambda_2 I_{p+s}]\hat{\theta}_R \end{aligned} \tag{A.7}$$

since from (A.6)

$$\hat{\theta}'_R(X'_{zF}\hat{W}Y_z) = (Y'_z\hat{W}X_{zF})[(X'_{zF}\hat{W}X_{zF}) + \lambda_2 I_{p+s}]^{-1}(X'_{zF}\hat{W}Y_z)$$

and

$$(Y_z'\hat{W}X_{zF})[X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}]^{-1}(X_{zF}'\hat{W}Y_z)=\hat{\theta}_R'[(X_{zF}'\hat{W}X_{zF})+\lambda_2 I_{p+s}]\hat{\theta}_R.$$

Similarly, we also have

$$\begin{aligned}
&(Y_z-X_{zF}\hat{\theta}_W)'\hat{W}(Y_z-X_{zF}\hat{\theta}_W)+\lambda_2\|\hat{\theta}_W\|^2\\
=\;& Y_z'\hat{W}Y_z-2\hat{\theta}_W'X_{zF}'\hat{W}Y_z+\hat{\theta}_W'(X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s})\hat{\theta}_W\\
=\;& Y_z'\hat{W}Y_z-2\hat{\theta}_W'(X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s})\hat{\theta}_R+\hat{\theta}_W'(X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s})\hat{\theta}_W \qquad \text{(A.8)}
\end{aligned}$$

since from (A.6)

$$\hat{\theta}_W'X_{zF}'\hat{W}Y_z=\hat{\theta}_W'[X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}]\hat{\theta}_R.$$

Subtracting (A.7) from (A.8), we thus have

$$\begin{aligned}
&\left\{(Y_z-X_{zF}\hat{\theta}_W)'\hat{W}(Y_z-X_{zF}\hat{\theta}_W)+\lambda_2\|\hat{\theta}_W\|^2\right\}\\
&-\left\{(Y_Z-X_{zF}\hat{\theta}_R)'\hat{W}(Y_Z-X_{zF}\hat{\theta}_R)+\lambda_2\|\hat{\theta}_R\|^2\right\}\\
=\;& (\hat{\theta}_W-\hat{\theta}_R)'[X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}](\hat{\theta}_W-\hat{\theta}_R), \qquad \text{(A.9)}
\end{aligned}$$

where with some symmetric $\hat{W}$

$$(\hat{\theta}_W-\hat{\theta}_R)'[X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}](\hat{\theta}_W-\hat{\theta}_R)\geq[\text{Eigmin}(X_{zF}'\hat{W}X_{zF})+\lambda_2]\|\hat{\theta}_W-\hat{\theta}_R\|^2 \qquad \text{(A.10)}$$

by exercise 7.25, p.167 of Abadir and Magnus (2005). Therefore, using (A.9), (A.10), (A.5) and (A.4), we have

$$\|\hat{\theta}_W-\hat{\theta}_R\|\leq\frac{\lambda_1\left[\sum_{j=1}^{p+s}\hat{w}_j^2\right]^{1/2}}{\text{Eigmin}(X_{zF}'\hat{W}X_{zF})+\lambda_2}. \qquad \text{(A.11)}$$

Second, for the bound of $||\hat{\theta}_R-\theta_0||$, we note that from (1)

$$Y_z=Z'Y=Z'X\beta_0+Z'u=Z'X\beta_0+nF_{q,s}\tau_0+(Z'u-nF_{q,s}\tau_0)=X_{zF}\theta_0+e, \qquad \text{(A.12)}$$

where we let $X_{zF}=[Z'X,nF_{q,s}]$, $\theta_0=(\beta_0',\tau_0')'$ and $e=Z'u-nF_{q,s}\tau_0$. Using (A.6), we have

$$\begin{aligned}
\hat{\theta}_R &= [X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}]^{-1}[X_{zF}'\hat{W}Y_z]\\
&= [X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}]^{-1}[X_{zF}'\hat{W}X_{zF}\theta_0+e+\lambda_2\theta_0-\lambda_2\theta_0],
\end{aligned}$$

and

$$\hat{\theta}_R-\theta_0=[X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}]^{-1}[X_{zF}'\hat{W}e]-\lambda_2[X_{zF}'\hat{W}X_{zF}+\lambda_2 I_{p+s}]^{-1}\theta_0. \qquad \text{(A.13)}$$

Then we can write that

$$\|\hat{\theta}_R-\theta_0\|^2\leq[\text{Eigmin}(X_{zF}'\hat{W}X_{zF})+\lambda_2]^{-2}[\lambda_2^2\|\theta_0\|^2+\|X_{zF}'\hat{W}e\|^2].$$

But by Assumption 1 and (6), we can rewrite it as (w.p.a.1)

$$\|\hat{\theta}_R-\theta_0\|^2\leq[bn^2+\lambda_2]^{-2}[\lambda_2^2\|\theta_0\|^2+\|X_{zF}'\hat{W}e\|^2],$$

where from Assumptions 1, (6),

$$\begin{aligned} \|X'_{zF}\hat{W}e\|^2 &= |e'\hat{W}X_{zF}X'_{zF}\hat{W}e| \\ &\leq Eigmax(\hat{W}X_{zF}X'_{zF}\hat{W})\|e\|^2 \\ &\leq n^2B\|e\|^2 \end{aligned} \tag{A.14}$$

wpa1. Next, given L is a finite constant

$$E\|X'_{zF}\hat{W}e\|^2 \leq n^2BE\|e\|^2 \leq qn^3BL. \tag{A.15}$$

We want to prove (A.15). This means showing

$$E\|e\|^2 \leq nqL. \tag{A.16}$$

Before proving (A.16) we introduce some notation. For $i = 1, \cdots, n$, let $e_i = Z_iu_i - F_{q,s}\tau_0$, $\forall j = 1, \cdots q$. See that e_i is a $q \times 1$ vector, and we can see that each cell in e_i is $e_{ij} = Z_{ij}u_i - (F_{q,s}\tau_0)_j$. For $k = 1, \cdots, n$ and $i \neq k$, let $e_k = Z_ku_k - F_{q,s}\tau_0$. See that e_k is a $q \times 1$ vector, and we can see that each cell in e_k is $e_{kj} = Z_{kj}u_k - (F_{q,s}\tau_0)_j$. Note that $(F_{q,s}\tau_0)_j$ represents the j th element in $q \times 1$ vector of $F_{q,s}\tau_0$. Given the independence of Z_i, u_i across i, if the moments are nonzero for $EZ_{ij}u_i = (F_{q,s}\tau_0)_j$, and $EZ_{kj}u_k = (F_{q,s}\tau_0)_j$, then it is easy to see that

$$Ee_{ij}e_{kj} = 0. \tag{A.17}$$

This last equation will be also true if the moments EZ_iu_i EZ_ku_k are zero or they have different nonzero moments. To see (A.16)

$$E\|e\|^2 = nE|(e'e)/n|.$$

Next

$$\begin{aligned} E[e'e/n] &= \frac{1}{n}E[(\sum_{i=1}^{n} e_i)'(\sum_{i=1}^{n} e_i)] \\ &= E\sum_{j=1}^{q}(\frac{1}{n^{1/2}}\sum_{i=1}^{n} e_{ij})^2 \\ &= \sum_{j=1}^{q}[\frac{1}{n}E(\sum_{i=1}^{n}\sum_{k=1}^{n} e_{ij}e_{kj})] \\ &= \sum_{j=1}^{q}[\frac{1}{n}E(\sum_{i=1}^{n} e_{ij}^2)] \leq qL, \end{aligned} \tag{A.18}$$

where the last equality is obtained through (A.17) and the inequality through Assumption 3. So (A.16) is proved.

Therefore, by (A.15) and seeing that we can write $B = BL$, it holds that

$$E\|\hat{\theta}_R - \theta_0\|^2 \leq 2\left[\frac{\lambda_2^2\|\theta_0\|^2 + qn^3B}{(bn^2 + \lambda_2)^2}\right]. \tag{A.19}$$

Finally, by taking expectations in (A.11) with Assumption 1, and combining it with (A.19) into (A.2), we have

$$E\|\hat{\theta}_W - \theta_0\|^2 \leq 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2 E(\sum_{j=1}^{p+s}\hat{w}_j^2)}{(bn^2+\lambda_2)^2}$$

w.p.a.1. The bounds for $E\|\hat{\theta}_W - \theta_0\|^2$ can be obtained by letting $\hat{w}_j = 1$ for all j. See that b, B are absolute positive constants, and they do not depend on n. **Q.E.D.**

Proof of Theorem 2 We have to show that $((1+\lambda_2/n)\tilde{\theta}_{\mathcal{A}}, 0)$ satisfies the Karush-Kuhn-Tucker conditions of the optimization of adaptive elastic-net equation (3) w.p.a.1. More precisely, we need to show

$$\Psi_n \equiv P\left\{|-2X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| \leq \lambda_1^* \hat{w}^j \text{ for all } j \in \mathcal{A}^c\right\} \to 1, \tag{A.20}$$

where $X_{zF\mathcal{A}}$ consists of columns of X_{zF} that correspond to nonzero elements in θ_0 and $X_{zF,j}$ is the jth column of X_{zF}. Then the next steps follow exactly as in equations (6.7) and (6.8) of Zou and Zhang (2009). We let $\eta = \min_{j\in\mathcal{A}}|\theta_{j0}|$ and $\hat{\eta} = \min_{j\in\mathcal{A}}|\hat{\theta}_{enet,j}|$. Since Ψ_n is equivalent to

$$P\left\{|-2X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^* \hat{w}^j, \quad \exists j \in \mathcal{A}^c\right\} \to 0.$$

So (A.20) satisfies

$$\Psi_n \leq \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^* \hat{w}^j \text{ and } \hat{\eta} > \eta/2\right\} + P(\hat{\eta} \leq \eta/2).$$

But from Theorem 1, w.p.a.1,

$$\begin{aligned} P(\hat{\eta} \leq \eta/2) &\leq P(\|\hat{\theta}_{enet} - \theta_0\| > \eta/2) \leq E\|\hat{\theta}_{enet} - \theta_0\|^2/(\eta^2/4) \\ &\leq 16\frac{\lambda_2^2\|\theta_0\|_2^2 + Bqn^3 + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2\eta^2}. \end{aligned} \tag{A.21}$$

In addition, letting $M = \left(\lambda_1^{*2}/n^{\kappa}\right)^{1/2\gamma}$,

$$
\begin{aligned}
& \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^*\hat{w}^j \text{ and } \hat{\eta} > \eta/2\right\} \\
\leq & \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^*\hat{w}^j, \hat{\eta} > \eta/2 \text{ and } |\hat{\theta}_{enet,j}| \leq M\right\} \\
& + \sum_{j\in\mathcal{A}^c} P\left(|\hat{\theta}_{enet,j}| > M\right) \\
\leq & \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^* M^{-\gamma} \text{ and } \hat{\eta} > \eta/2\right\} + \sum_{j\in\mathcal{A}^c} P\left(|\hat{\theta}_{enet,j}| > M\right) \\
\leq & \frac{4M^{2\gamma}}{\lambda_1^{*2}} E\left[\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] + \frac{1}{M^2} E\left[\sum_{j\in\mathcal{A}^c} |\hat{\theta}_{enet,j}|^2\right] \\
\leq & \frac{4M^{2\gamma}}{\lambda_1^{*2}} E\left[\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] + \frac{E\|\hat{\theta}_{enet} - \theta_0\|^2}{M^2} \\
\leq & \frac{4M^{2\gamma}}{\lambda_1^{*2}} E\left[\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] \qquad \text{(A.22)} \\
& + 4\frac{\lambda_2^2\|\theta_0\|_2^2 + Bn^3 q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2 M^2}
\end{aligned}
$$

w.p.a.1, where the last inequality follows from Theorem 1. Note that equations (A.21) and (A.22) are linear GMM counterparts of (6.7) and (6.8) in Zou and Zhang (2009). However, M definition in Zou and Zhang (2009) least squares proof does not extend here. So deriving (A.22) and finding a new M for linear GMM that will make the new proof workable is not trivial.

The last inequality (A.22) can be further bounded as follows. Given the fact that $\theta_{\mathcal{A}}$ represents all the nonzero elements in the true model θ_0 with (A.12), we can see that

$$
\begin{aligned}
\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 &= \sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(X_{zF\mathcal{A}}\theta_{\mathcal{A}} - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}}) + X'_{zF,j}\hat{W}e|^2 \\
&\leq 2\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(X_{zF\mathcal{A}}\theta_{\mathcal{A}} - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 + 2\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}e|^2.
\end{aligned}
$$

$\theta_{\mathcal{A}}$ represent the true model parameters that corresponds to active set $\mathcal{A}$. However, with $\hat{W}$ being symmetric and positive definite, we have

$$
\begin{aligned}
\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(X_{zF\mathcal{A}}\theta_{\mathcal{A}} - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 &\leq Bn^2\|\hat{W}^{1/2}X_{zF\mathcal{A}}(\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}})\|^2 \\
&\leq Bn^2 \times Bn^2\|\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}}\|^2, \qquad \text{(A.23)}
\end{aligned}
$$

from Assumption 1 and (6) w.p.a.1. It thus follows that by (A.15)(A.23)

$$
E\left[\sum_{j\in\mathcal{A}^c} |X'_{zF,j}\hat{W}(Y_z - X_{zF\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] \leq 2B^2n^4 E(\|\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}}\|_2^2 1_{\{\hat{\eta}>\eta/2\}}) + 2Bn^3 q. \qquad \text{(A.24)}
$$

Furthermore, by defining

$$\tilde{\theta}_{\mathcal{A}R} = \arg\max_{\theta} \left\{ (Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_2 \sum_{j\in\mathcal{A}} \theta_j^2 \right\},$$

we have by the analysis in (A.11), since $\hat{w}_j \leq \hat{\eta}^{-\gamma}$

$$\|\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}R}\| \leq \frac{\lambda_1^* \hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2} \tag{A.25}$$

w.p.a.1 and thus

$$E(\|\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}}\|^2 1_{\{\hat{\eta}>\eta/2\}}) \leq 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^{*2}(\eta/2)^{-2/\gamma}(p+s)}{(bn^2+\lambda_2)^2} \tag{A.26}$$

by the last equation in the proof of Theorem 1 above. Therefore, by combining (A.21), (A.22), (A.24) and (A.26), we have (w.p.a.1)

$$\begin{aligned}
\Psi_n \quad \leq \quad & \frac{4M^{2\gamma}}{\lambda_1^{*2}} \left\{ 2B^2n^4 \times 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^{*2}(\eta/2)^{-2/\gamma}(p+s)}{(bn^2+\lambda_2)^2} + 2Bn^3q \right\} && \text{(A.27)} \\
& +4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2M^2} && \text{(A.28)} \\
& +16\frac{\lambda_2^2\|\theta_0\|^2 + Bqn^3 + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2\eta^2}. && \text{(A.29)}
\end{aligned}$$

Now we prove that each of (A.27), (A.28) and (A.29) all converges to zero to complete the proof. First, (A.27) is

$$O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}\lambda_2^2(p+s)\right) + O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}n^3q\right) + O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}\frac{(\lambda_1^*)^2(p+s)}{\eta^{2\gamma}}\right) + O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}n^3q\right),$$

where the second and the last terms are $o_p(1)$ since

$$\frac{M^{2\gamma}}{\lambda_1^{*2}}n^3q = \frac{\lambda_1^{*2}}{n^\kappa}\frac{1}{\lambda_1^{*2}}n^3q = \frac{n^{3+\alpha}}{n^\kappa} \to 0$$

from $q = O(n^\alpha)$ and Assumption 2-(ii). In addition, the first term is all dominated by the last or second terms: for the first term, it is because $\lambda_2^2/n^3 \to 0$ by Assumption 2(i). Next see that

$$\frac{M^{2\gamma}}{(\lambda_1^*)^2}(\lambda_1^*)^2\frac{(p+s)}{\eta^{2\gamma}} = \frac{(\lambda_1^*)^2}{n^\kappa}\frac{(p+s)}{\eta^{2\gamma}} \to 0, \tag{A.30}$$

by Assumption 2(iv) and κ definition in Assumption 2(ii), and $M = ((\lambda_1^*)^2/n^\kappa)^{1/2\gamma}$. Therefore, (A.27) is $o_p(1)$.
Second, (A.28) is

$$O_p\left(\frac{\lambda_2^2}{n^3}\frac{(p+s)}{n}\frac{1}{M^2}\right) + O_p\left(\frac{n^3}{n^4}\frac{q}{M^2}\right) + O_p\left(\frac{\lambda_1^2}{n^4}\frac{(p+s)}{M^2}\right).$$

But note that the second term dominates the other two terms since $\lambda_1^2/n^3 \to 0$ and $\lambda_2^2/n^3 \to 0$ by Assumption 2-(i). Moreover, the second term is $o_p(1)$ since

$$\frac{q}{nM^2} = \frac{q}{n} \times \frac{1}{M^2} \leq \frac{n^{\alpha-1}}{M^2} = \frac{n^{\alpha-1+\kappa/\gamma}}{(\lambda_1^*)^{(2/\gamma)}} \to 0$$

$q = O(n^\alpha)$, Assumption 2-(iv) and the definition of M.
Finally, (A.29) is

$$O_p\left(\frac{\lambda_2^2(p+s)}{n^4\eta^2}\right) + O_p\left(\frac{qn^3}{n^4\eta^2}\right) + O_p\left(\frac{\lambda_1^2(p+s)}{n^4\eta^2}\right) = o_p(1) \tag{A.31}$$

We prove (A.31). Since $(p+s) \leq q$, $\lambda_2^2/n^3 \to 0$, $\lambda_1^2/n^3 \to 0$ by Assumption 2, the second term dominates the others in (A.31). Then we consider the second term above

$$\frac{qn^3}{n^4\eta^2} == \frac{q}{n}\frac{1}{\eta^2} = \frac{n^\alpha}{n\eta^2} = \frac{n^{\alpha-1}}{\eta^2} \to 0, \tag{A.32}$$

with $\eta = O(n^{-1/m})$, this means that $n^{\alpha-1}n^{2/m} \to 0$, but this indicates a lower bound on m to be true

$$m > \frac{2}{1-\alpha},$$

but this lower bound is implied by the lower bound that comes from Assumption 2(iv)(equation (7)), since $2\gamma/[\gamma(1-\alpha) - \nu - \kappa + 2] > 2/(1-\alpha)$ with $0 < \nu \leq \alpha < 1$ with $\gamma > 1$, $\kappa > 3 + \alpha$ by Assumption 2(ii). So Assumption 2(iv) provides (A.32). **Q.E.D.**

Proof of Theorem 3 Using Theorem 2, in order to prove the selection consistency, we only need to show that the minimal element of the estimator of nonzero coefficients is larger than zero w.p.a.1: $P\left\{\min_{j\in\mathcal{A}} |\tilde{\theta}_j| > 0\right\} \to 1$. Note that by (A.25)

$$\min_{j\in\mathcal{A}} |\tilde{\theta}_j| > \min_{j\in\mathcal{A}} |\tilde{\theta}_{\mathcal{A}R,j}| - \frac{\lambda_1^* \hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2 + \lambda_2}, \tag{A.33}$$

and also

$$\min_{j\in\mathcal{A}} |\tilde{\theta}_{\mathcal{A}R,j}| > \min_{j\in\mathcal{A}} |\theta_{\mathcal{A}j}| - \|\tilde{\theta}_{\mathcal{A}R} - \theta_{\mathcal{A}}\|. \tag{A.34}$$

But from (A.19), it holds that

$$\begin{aligned} E(\|\tilde{\theta}_{\mathcal{A}R} - \theta_{\mathcal{A}}\|^2) &\leq 2\left[\frac{\lambda_2^2\|\theta_0\|_2^2 + qn^3B}{(bn^2+\lambda_2)^2}\right] \\ &= O\left(\frac{\lambda_2^2(p+s)}{n^4}\right) + O\left(\frac{qn^3}{n^4}\right) \\ &= O\left(\frac{q}{n}\right) \end{aligned} \tag{A.35}$$

w.p.a.1 since $\lambda_2^2/n^3 \to 0$ and $p + s \leq q$. Next,

$$\frac{\lambda_1^* \hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2} = O\left(\frac{\lambda_1^*\sqrt{p+s}}{n^2\eta^\gamma}\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma}\right) \tag{A.36}$$

where

$$\frac{\lambda_1^*}{n^2}\frac{\sqrt{p+s}}{\eta^\gamma} = \frac{1}{n}\left(\frac{\lambda_1^*}{n}\frac{\sqrt{p+s}}{\eta^\gamma}\right) = o(\frac{1}{n}), \tag{A.37}$$

by Assumption 2(iv). Next we consider

$$\begin{aligned} E\left[\left(\frac{\hat{\eta}}{\eta}\right)^2\right] &\leq 2 + \frac{2}{\eta^2}E[(\hat{\eta}-\eta)^2] \\ &\leq 2 + \frac{2}{\eta^2}E\|\hat{\theta}_e - \theta_0\|^2 \\ &\leq 2 + \frac{2}{\eta^2}\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2} \to 2 \end{aligned} \tag{A.38}$$

by (A.31). Note that

$$\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma} = \left[\left(\frac{\hat{\eta}}{\eta}\right)^2\right]^{-\gamma/2}. \tag{A.39}$$

Then by (A.38),

$$E\left(\frac{\hat{\eta}}{\eta}\right)^2 = O(1),$$

so by Markov's inequality

$$\left(\frac{\hat{\eta}}{\eta}\right)^2 = O_p(1).$$

Then by (A.39) and the last equation above we have

$$\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma} = O_p(1), \tag{A.40}$$

since if a generic random variable $\Gamma = O_p(1)$ we have $\Gamma^{-\gamma/2} = O_p(1)$. Plugging (A.35)-(A.38) in (A.33) and (A.34)

$$min_{j\in\mathcal{A}}|\tilde{\theta}_j| > min_{j\in\mathcal{A}}|\theta_{\mathcal{A}j}| - (\sqrt{q/n})O(1) - (1/n)o_p(1),$$

since $\sqrt{q/n}$ converges to zero faster than η by (A.32) we have the desired result .

Proof of Theorem 4 We define

$$\Phi_n = \zeta'\frac{(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})}{1+\lambda_2/n}\hat{\Sigma}_{\mathcal{A}}^{1/2}n^{-1/2}(\hat{\theta}_{\mathcal{A}} - \theta_{\mathcal{A}}).$$

Using $\tilde{\theta}_{\mathcal{A}}$ in (8), and noting its scaled difference from the definition of $\hat{\theta}_{\mathcal{A}}$ we write

$$\begin{aligned} & \zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2}n^{-1/2}\left(\tilde{\theta}_{\mathcal{A}} - \frac{\theta_{\mathcal{A}}}{1+\lambda_2/n}\right) \\ = \; & \zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2}n^{-1/2}\left(\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}R} + \tilde{\theta}_{\mathcal{A}R} - \frac{\theta_{\mathcal{A}}}{1+\lambda_2/n}\right) \\ = \; & \zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2}n^{-1/2}(\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}R}) \\ & +\zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2}n^{-1/2}(\tilde{\theta}_{\mathcal{A}R} - \theta_{\mathcal{A}}) \\ & +\zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2}n^{-1/2}\left(\theta_{\mathcal{A}} - \frac{\theta_{\mathcal{A}}}{1+\lambda_2/n}\right), \end{aligned} \tag{A.41}$$

where

$$\tilde{\theta}_{\mathcal{A}R} = \arg\min_{\theta} \left\{ (Y_z - X_{zF\mathcal{A}}\theta)'\hat{W}(Y_z - X_{zF\mathcal{A}}\theta) + \lambda_2 \sum_{j\in\mathcal{A}} \theta_j^2 \right\}.$$

Define $e_{\mathcal{A}} = Z'u - F_{q,s_0}\tau_{\mathcal{A}}$, and an $s_0 \times 1$ vector $\tau_{\mathcal{A}}$ represents the nonzero s_0 elements in τ. But note that $\tilde{\theta}_{\mathcal{A}R} - \theta_{\mathcal{A}} = (\hat{\Sigma}_{\mathcal{A}} + \lambda_2 I_{p_0+s_0})^{-1}(X'_{zF\mathcal{A}}\hat{W}e_{\mathcal{A}}) - \lambda_2(\hat{\Sigma}_{\mathcal{A}} + \lambda_2 I_{p_0+s_0})^{-1}\theta_{\mathcal{A}}$ from (A.13) and thus the second term in (A.41) satisfies

$$\begin{aligned}
& (I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2} n^{-1/2}(\tilde{\theta}_{\mathcal{A}R} - \theta_{\mathcal{A}}) \\
= \; & \hat{\Sigma}_{\mathcal{A}}^{-1/2}(\hat{\Sigma}_{\mathcal{A}}^{1/2} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2})\hat{\Sigma}_{\mathcal{A}}^{1/2} n^{-1/2}(\tilde{\theta}_{\mathcal{A}R} - \theta_{\mathcal{A}}) \\
= \; & \hat{\Sigma}_{\mathcal{A}}^{-1/2}(\hat{\Sigma}_{\mathcal{A}}^{1/2} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2}) \\
& \times \left\{ (\hat{\Sigma}_{\mathcal{A}}^{1/2} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2})^{-1} n^{-1/2}(X'_{zF\mathcal{A}}\hat{W}e_{\mathcal{A}}) - \lambda_2(\hat{\Sigma}_{\mathcal{A}}^{1/2} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2})^{-1} n^{-1/2}\theta_{\mathcal{A}} \right\} \\
= \; & \hat{\Sigma}_{\mathcal{A}}^{-1/2} X'_{zF\mathcal{A}}\hat{W} n^{-1/2} e_{\mathcal{A}} - \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2} n^{-1/2}\theta_{\mathcal{A}},
\end{aligned}$$

Moreover, the third term in (A.41) can be simply written as

$$\zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}^{-1})\hat{\Sigma}^{1/2} n^{-1/2}\left(\theta_{\mathcal{A}} - \frac{\theta_{\mathcal{A}}}{1+\lambda_2/n}\right) = \zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}^{-1})\hat{\Sigma}^{1/2} n^{-1/2}\left(\frac{\lambda_2\theta_{\mathcal{A}}}{\lambda_2+n}\right).$$

Therefore, using these expressions as well as Theorem 3, we can write

$$\Phi_n = \Phi_{1,n} + \Phi_{2,n} + \Phi_{3,n}$$

w.p.a.1, where

$$\begin{aligned}
\Phi_{1,n} &= \zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2} n^{-1/2}\frac{\lambda_2\theta_{\mathcal{A}}}{n+\lambda_2} - \zeta'\lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2} n^{-1/2}\theta_{\mathcal{A}} \\
\Phi_{2,n} &= \zeta'(I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2} n^{-1/2}(\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}R}) \\
\Phi_{3,n} &= \zeta'\hat{\Sigma}_{\mathcal{A}}^{-1/2} X'_{zF\mathcal{A}}\hat{W} n^{-1/2} e_{\mathcal{A}}.
\end{aligned}$$

We will look at each term to obtained the desired result. First note that w.p.a.1

$$\begin{aligned}
\Phi_{1,n}^2 &\leq \frac{2}{n}\left\| (I_{p_0+s_0} + \lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1})\hat{\Sigma}_{\mathcal{A}}^{1/2}\frac{\lambda_2\theta_{\mathcal{A}}}{n^2+\lambda_2}\right\|^2 + \frac{2}{n}\|\lambda_2\hat{\Sigma}_{\mathcal{A}}^{-1/2}\theta_{\mathcal{A}}\|^2 \\
&\leq \frac{2}{n}\frac{\lambda_2^2}{(n^2+\lambda_2)^2}\|\hat{\Sigma}_{\mathcal{A}}^{1/2}\theta_{\mathcal{A}}\|^2\left(1+\frac{\lambda_2}{bn^2}\right)^2 + \frac{2}{n}\lambda_2^2\|\theta_{\mathcal{A}}\|^2\frac{1}{bn^2} \\
&\leq \frac{2\lambda_2^2}{n(n^2+\lambda_2)^2} Bn^2\left(1+\frac{\lambda_2}{bn^2}\right)^2\|\theta_{\mathcal{A}}\|^2 + \frac{2\lambda_2^2\|\theta_{\mathcal{A}}\|^2}{bn^3} \to 0
\end{aligned}$$

from (6) and Assumption 2-(i), where $\lambda_2^2(p+s)/n^3 \to 0$, $\|\theta_{\mathcal{A}}\|^2 \leq (p+s)$ and $(p+s)/n \to 0$. Second, in the same way, we have

$$\begin{aligned}
\Phi_{2,n}^2 &\leq \frac{1}{n}\left(1+\frac{\lambda_2}{bn^2}\right)^2 \|\hat{\Sigma}_{\mathcal{A}}^{1/2}(\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}R})\|^2 \\
&\leq \frac{1}{n}\left(1+\frac{\lambda_2}{bn^2}\right)^2 Bn^2\|\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}R}\|^2 \leq \frac{1}{n}\left(1+\frac{\lambda_2}{bn^2}\right)^2 Bn^2\left(\frac{\lambda_1^*\hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2}\right)^2 \\
&= Bn\left(\frac{\lambda_1^*\hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2}\right)^2 + o(1) \\
&= B\left(\frac{\lambda_1^*\hat{\eta}^{-\gamma}\sqrt{p+s}\sqrt{n}}{bn^2+\lambda_2}\right)^2 + o(1) \\
&= O\left(n\left[\frac{\lambda_1^*\sqrt{p+s}}{n^2\eta^\gamma}\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma}\right]^2\right) = O\left(\frac{1}{n}\left[\frac{\lambda_1^*\sqrt{p+s}}{n\eta^\gamma}\left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma}\right]^2\right) \\
&= o_p(1)
\end{aligned}$$

where we use $(1+\lambda_2/bn^2) \to 1$, (A.25)(A.36)-(A.37) and (A.40). So we have $\Phi_{2,n}^2 = o_p(1)$. Finally, we prove that $\Phi_{3,n} \xrightarrow{d} \mathcal{N}(0,1)$. We denote the ith element of $\Phi_{3,n}$ as

$$\hat{r}_i = \zeta'\hat{\Sigma}_{\mathcal{A}}^{-1/2}X_{zF\mathcal{A}}'\hat{W}n^{-1/2}e_i,$$

where $e_i = Z_iu_i - F_{q,s_0}\tau_{\mathcal{A}} = Z_iu_i - F_{q,s}\tau_0$. We also let $r_i = \zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma_{xzF\mathcal{A}}'V^{-1}n^{-1/2}e_i$, where we use $W = V^{-1}$ as the optimal weight. Then by (9), Assumption 1-(i) and the definition of $\hat{\Sigma}_{\mathcal{A}}$, we have

$$\|\hat{\Sigma}_{\mathcal{A}}^{-1/2}X_{zF\mathcal{A}}'\hat{W} - \Sigma_{\mathcal{A}}^{-1/2}\Sigma_{xzF\mathcal{A}}'V^{-1}\| = \|\hat{\Sigma}_{\mathcal{A}}^{-1/2}n(n^{-1}X_{zF\mathcal{A}})'\hat{W} - \Sigma_{\mathcal{A}}^{-1/2}\Sigma_{xzF\mathcal{A}}'V^{-1}\| \xrightarrow{p} 0$$

and $\sum_{i=1}^n(\hat{r}_i - r_i) \xrightarrow{p} 0$. We now verify the Lyapunov condition to obtain the CLT. Since $\Sigma_{\mathcal{A}} = \Sigma_{xzF\mathcal{A}}'V^{-1}\Sigma_{xzF\mathcal{A}}$ and by Assumption 1 $\|n^{-1}\sum_{i=1}^n e_ie_i' - V\| \xrightarrow{p} 0$, we have

$$\begin{aligned}
\lim_{n\to\infty}\sum_{i=1}^n E[r_i^2] &= \zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma_{xzF\mathcal{A}}'V^{-1}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}\zeta \\
&= \zeta'(\Sigma_{xzF\mathcal{A}}'V^{-1}\Sigma_{xzF\mathcal{A}})^{-1/2}(\Sigma_{xzF\mathcal{A}}'V^{-1}\Sigma_{xzF\mathcal{A}})(\Sigma_{xzF\mathcal{A}}'V^{-1}\Sigma_{xzF\mathcal{A}})^{-1/2}\zeta = 1
\end{aligned}$$

using $W = V^{-1}$ as the optimal weight, and $\Sigma_{\mathcal{A}}$ definition. Next, for $\delta > 0$ we need to show that

$$\lim_{n\to\infty}\sum_{i=1}^n E|r_i|^{2+\delta} = 0.$$

But since we show that $\lim_{n\to\infty}\sum_{i=1}^n E|r_i|^2 = 1$ above,

$$\lim_{n\to\infty}\sum_{i=1}^n E|r_i|^{2+\delta} \leq \lim_{n\to\infty}\sum_{i=1}^n E|r_i|^2 \max_{1\leq i\leq n}|r_i|^\delta \leq \left(\max_{1\leq i\leq n}|r_i^2|\right)^{\delta/2}.$$

Note that

$$|r_i^2| \leq n^{-1}\|e_i\|^2\|V^{-1}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}\zeta\|^2 \tag{A.42}$$

by Cauchy-Schwartz inequality. For (A.42), we have

$$\begin{aligned}\|V^{-1}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}\zeta\|^2 &= \zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}\zeta \\ &\leq \text{Eigmax}(\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2})\|\zeta\|^2 \\ &= \text{Eigmax}(\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}) < \infty, \qquad \text{(A.43)}\end{aligned}$$

where $\|\zeta\|^2 = 1$ and the first inequality is obtained by $\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}$ being symmetric and using the bounds of Rayleigh quotient (e.g., Exercise 7.53a of Abadir and Magnus, 2005). Since $\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}$ is positive definite, so is $(\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2})^{-1}$, which gives that the minimal eigenvalue of $(\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2})^{-1}$ is greater than zero so the maximal eigenvalue of $\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{xzF\mathcal{A}}V^{-2}\Sigma_{xzF\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}$ is finite. Therefore, given (A.42), (A.43) and using Assumption 3, we have $(\max_i |r_i^2|)^{\delta/2} = o_p(1)$ so that $\lim_{n\to\infty}\sum_{i=1}^n E|r_i|^{2+\delta} = 0$, which proves the conditions for CLT, and hence we have the desired result. **Q.E.D.**

Proof of Theorem 5 Without losing any generality, we divide the instruments into two sets $Z_i = [Z_{1i}, Z_{2i}]$ satisfying

$$\sum_{i=1}^n E[Z_{1i}u_i] = 0_{q-s_0} \quad \text{and} \quad \sum_{i=1}^n E[Z_{2i}u_i] = \tau_{\mathcal{A}},$$

where Z_{1i} are $(q - s_0)$ number of valid instruments whereas $\tau_{\mathcal{A}}$ is an $s_0 \times 1$ vector, whose elements are all nonzero, so that Z_{2i} consists of s_0 number of invalid instruments. Accordingly we also decompose the $q \times (p_0 + s_0)$ matrix $\Sigma_{xzF\mathcal{A}}$ as

$$\Sigma_{xzF\mathcal{A}} = [\Sigma_{xz\mathcal{A}}, F_{q,s_0}] = \begin{bmatrix} \Sigma_{xz1\mathcal{A}} & 0_{q-s_0,s_0} \\ \Sigma_{xz2\mathcal{A}} & I_{s_0} \end{bmatrix},$$

where $\|Z'X_{\mathcal{A}}/n - \Sigma_{xz\mathcal{A}}\| \xrightarrow{p} 0$, $\|Z_1'X_{\mathcal{A}}/n - \Sigma_{xz1\mathcal{A}}\| \xrightarrow{p} 0$ and $\|Z_2'X_{\mathcal{A}}/n - \Sigma_{xz2\mathcal{A}}\| \xrightarrow{p} 0$. Note that $\Sigma_{xz1\mathcal{A}}$ is of dimension $(q - s_0) \times p_0$ and $\Sigma_{xz2\mathcal{A}}$ is of dimension $s_0 \times p_0$. Similarly, we let

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{bmatrix} \begin{matrix} (q-s_0) \\ s_0 \end{matrix}, \qquad \begin{matrix} \\ (q-s_0) \quad s_0 \end{matrix}$$

and note that we show the number of rows and columns of partitioned matrices on the side. For notational convenience, we also define

$$V^{-1} = \begin{bmatrix} V^{11} & V^{12} \\ (V^{12})' & V^{22} \end{bmatrix} \begin{matrix} (q-s_0) \\ s_0 \end{matrix}, \qquad \begin{matrix} \\ (q-s_0) \quad s_0 \end{matrix}$$

where explicit expressions of each term become clear at the end of this proof.

Given $\Sigma_{xzF\mathcal{A}}$ and V^{-1} decompositions above, we can write

$$\Sigma_{\mathcal{A}} = \Sigma'_{xzF\mathcal{A}}V^{-1}\Sigma_{xzF\mathcal{A}} = \begin{bmatrix} \Sigma_{\mathcal{A}11} & \Sigma_{\mathcal{A}12} \\ \Sigma'_{\mathcal{A}12} & \Sigma_{\mathcal{A}22} \end{bmatrix} \begin{matrix} p_0 \\ s_0 \end{matrix}, \qquad \begin{matrix} \\ p_0 \quad s_0 \end{matrix}$$

where

$$\begin{aligned}
\Sigma_{\mathcal{A}11} &= \Sigma'_{xz1\mathcal{A}}V^{11}\Sigma_{xz1\mathcal{A}} + \Sigma'_{xz1\mathcal{A}}V^{12}\Sigma_{xz2\mathcal{A}} + \Sigma'_{xz2\mathcal{A}}(V^{12})'\Sigma_{xz1\mathcal{A}} + \Sigma'_{xz2\mathcal{A}}V^{22}\Sigma_{xz2\mathcal{A}} \quad \text{(A.44)} \\
\Sigma_{\mathcal{A}12} &= \Sigma'_{xz1\mathcal{A}}V^{12} + \Sigma'_{xz2\mathcal{A}}V^{22} \\
\Sigma_{\mathcal{A}22} &= V^{22}.
\end{aligned}$$

We let $\Sigma^{11}_{\mathcal{A}}$ be the north-west (upper left $p_0 \times p_0$) block of $\Sigma^{-1}_{\mathcal{A}}$. Then using the formula for partitioned inverses (e.g., Exercises 5.16a and 5.17 of Abadir and Magnus, 2005), we have

$$\begin{aligned}
\Sigma^{11}_{\mathcal{A}} &= \left[\Sigma_{\mathcal{A}11} - \Sigma_{\mathcal{A}12}\Sigma^{-1}_{\mathcal{A}22}\Sigma'_{\mathcal{A}12}\right]^{-1} \\
&= [\Sigma'_{xz1\mathcal{A}}V^{11}\Sigma_{xz1\mathcal{A}} - \Sigma'_{xz1\mathcal{A}}V^{12}(V^{22})^{-1}(V^{12})'\Sigma_{xz1\mathcal{A}}]^{-1} \\
&= [\Sigma'_{xz1\mathcal{A}}\left\{V^{11} - V^{12}(V^{22})^{-1}(V^{12})'\right\}\Sigma_{xz1\mathcal{A}}]^{-1} \\
&= [\Sigma'_{xz1\mathcal{A}}V_{11}^{-1}\Sigma_{xz1\mathcal{A}}]^{-1}, \qquad \text{(A.45)}
\end{aligned}$$

where the last equality is from the fact that (e.g., Exercise 5.16a of Abadir and Magnus, 2005) $V^{11} = V_{11}^{-1} + V_{11}^{-1}V_{12}V^{22}V'_{12}V_{11}^{-1}$ and $V^{12} = -V_{11}^{-1}V_{12}V^{22}$. The result follows since $\Sigma^{11}_{\mathcal{A}}$ corresponds to the asymptotic variance of $\hat{\beta}_{\mathcal{A}}$ from Theorem 4. **Q.E.D.**

References

Abadir and Magnus (2005). *Matrix Algebra*, Cambridge University Press.

Andrews, D. (1999). Consistent Moment Selection Procedures for Generalized Method of Moments Estimation, *Econometrica*, 67, 543-564.

Andrews, D. and B. Lu (2001). Consistent Model and Moment Selection Criteria for GMM Estimation with Applications to Dynamic Panel Models, *Journal of Econometrics*, 101, 123-164.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application of employment equations, *Review of Economics Studies*, 58, 277-297.

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

Blundell, R. and S. Bond. (1998). Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics*, 87(1), 115-143.

Bun M. and F. Kleibergen (2013). Identification and inference in moments based analysis of linear dynamic panel data models. University of Amsterdam-Econometrics Discussion Paper 2013/07.

Caner, M. (2009). Lasso type GMM estimator. *Econometric Theory*, 25,270-290.

Caner, M., and H.H. Zhang (2013). Adaptive Elastic Net GMM Estimator. Forthcoming *Journal of Business and Economics Statistics*.

Cheng, X. and Z. Liao (2012). Select the valid and relevant moments: A one step procedure for GMM with many moments. Working Paper. Department of Economics, University of Pennsylvania and UCLA.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani (2004). Least Angle Regression. *Annals of Statistics*, 32, 407-499.

Gautier E. and A. Tsybakov (2011). High dimensional instrumental variable regression and confidence sets. arXIV 1105.2454.

Leeb, H., and B. Potscher (2005). Model selection and inference:facts and fiction. *Econometric Theory*, 21, 21-59.

Liao, Z. (2013). Adaptive GMM Shrinkage Estimation with Consistent Moment Selection, *Econometric Theory*, forthcoming.

Lu, X. and L. Su (2013). Shrinkage estimation of dynamic panels with interactive fixed effects. Working paper, Department of Economics, Singapore Management University.

Newey, W. K. and Windmeijer, F. (2009). GMM with many weak moment conditions, *Econometrica*, 77, 687–719.

Qian, J and L. Su (2013). Shrinkage estimation of regression models with multiple structural change. Working paper, Department of Economics, Singapore Management University.

Schmidt, M. (2010). Graphical model structure learning with L-1 regularization. Thesis. University of British Columbia.

Wang, H., R. Li, C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B*, 71, 671-683.

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

Zou, H., and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67-part 2, 301-320.

Zou, H. and H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters, *Annals of Statistics*, 37, 1733-1751.

Table 1: RMSE of estimators of $\tau_{\mathcal{A}^c}$, $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$

$n = 250$, $p = 20$, $p_0 = 3$, $s = 10$, $s_0 = 3$ and $q = 43$												
	AENet				ALASSO-LARS				ALASSO-CL			
$\rho_z, \tau_{\mathcal{A}}, b$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$
.5, .5, .25	0.012	0.187	0.030	0.127	0.013	0.181	0.029	0.125	0.010	0.388	0.088	0.087
.5, .5, .5	0.011	0.194	0.026	0.090	0.011	0.193	0.025	0.088	0.010	0.388	0.088	0.087
.5, .5, 1	0.011	0.195	0.026	0.086	0.011	0.196	0.025	0.084	0.010	0.388	0.088	0.087
.5, .3, .25	0.014	0.167	0.033	0.127	0.014	0.165	0.032	0.122	0.011	0.272	0.087	0.086
.5, .3, .5	0.013	0.176	0.029	0.092	0.013	0.176	0.028	0.089	0.011	0.272	0.087	0.086
.5, .3, 1	0.012	0.178	0.028	0.088	0.012	0.178	0.028	0.086	0.011	0.272	0.087	0.086
.8, .5, .25	0.006	0.201	0.038	0.205	0.006	0.195	0.037	0.208	0.015	0.263	0.136	0.129
.8, .5, .5	0.004	0.208	0.030	0.166	0.005	0.205	0.030	0.174	0.014	0.265	0.136	0.129
.8, .5, 1	0.004	0.211	0.028	0.118	0.004	0.211	0.028	0.119	0.015	0.263	0.136	0.129
.8, .3, .25	0.006	0.198	0.040	0.206	0.007	0.195	0.040	0.208	0.016	0.206	0.138	0.131
.8, .3, .5	0.005	0.206	0.034	0.170	0.005	0.204	0.034	0.179	0.016	0.206	0.138	0.131
.8, .3, 1	0.005	0.213	0.029	0.125	0.005	0.214	0.029	0.127	0.016	0.206	0.138	0.131
$n = 1000$, $p = 20$, $p_0 = 3$, $s = 10$, $s_0 = 3$ and $q = 43$												
	AENet				ALASSO-LARS				ALASSO-CL			
$\rho_z, \tau_{\mathcal{A}}, b$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$
.5, .5, .25	0.003	0.060	0.006	0.045	0.003	0.057	0.006	0.045	0.004	0.151	0.040	0.039
.5, .5, .5	0.003	0.060	0.005	0.040	0.003	0.059	0.005	0.040	0.004	0.151	0.040	0.039
.5, .5, 1	0.002	0.061	0.005	0.040	0.002	0.061	0.005	0.040	0.005	0.152	0.040	0.040
.5, .3, .25	0.003	0.049	0.006	0.045	0.003	0.049	0.006	0.045	0.004	0.187	0.042	0.042
.5, .3, .5	0.003	0.050	0.006	0.040	0.003	0.050	0.005	0.040	0.004	0.187	0.042	0.042
.5, .3, 1	0.003	0.050	0.006	0.040	0.003	0.050	0.005	0.040	0.004	0.187	0.042	0.042
.8, .5, .25	0.001	0.062	0.008	0.108	0.001	0.058	0.008	0.112	0.005	0.128	0.068	0.063
.8, .5, .5	0.001	0.062	0.005	0.063	0.001	0.062	0.005	0.064	0.005	0.128	0.068	0.063
.8, .5, 1	0.001	0.063	0.005	0.058	0.001	0.063	0.005	0.058	0.005	0.128	0.068	0.063
.8, .3, .25	0.001	0.055	0.009	0.103	0.001	0.054	0.009	0.108	0.005	0.173	0.073	0.068
.8, .3, .5	0.001	0.056	0.007	0.062	0.001	0.056	0.007	0.063	0.005	0.173	0.073	0.068
.8, .3, 1	0.001	0.057	0.006	0.058	0.001	0.057	0.006	0.058	0.004	0.174	0.073	0.070

Note: AENet is the estimator defined in (3) and solved by the LARS algorithm. ALASSO-LARS is the same as AENet except that λ_2 is restricted to be zero. ALASSO-CL is the estimator proposed by Cheng and Liao (2012). ρ_z controls the correlation among valid instruments. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. b is the value of nonzero structural parameters. $rmse_1$, $rmse_2$, $rmse_3$ and $rmse_4$ denote the RMSE of $\tau_{\mathcal{A}^c}$, $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$, respectively.

Table 2: Moment Selection Accuracy

$n = 250$, $p = 20$, $p_0 = 3$, $s = 10$, $s_0 = 3$ and $q = 43$						
	AENet		ALASSO-LARS		ALASSO-CL	
$\rho_z, \tau_{\mathcal{A}}, b$	Pr_1	Pr_2	Pr_1	Pr_2	Pr_1	Pr_2
.5, .5, .25	0.974	0.999	0.968	1.000	0.989	0.823
.5, .5, .5	0.980	0.999	0.979	0.999	0.989	0.823
.5, .5, 1	0.979	0.999	0.979	0.999	0.989	0.823
.5, .3, .25	0.966	0.861	0.958	0.881	0.983	0.350
.5, .3, .5	0.970	0.828	0.969	0.832	0.983	0.350
.5, .3, 1	0.970	0.819	0.969	0.819	0.983	0.350
.8, .5, .25	0.985	0.991	0.982	0.994	0.944	0.954
.8, .5, .5	0.991	0.989	0.990	0.991	0.949	0.954
.8, .5, 1	0.990	0.988	0.991	0.989	0.944	0.954
.8, .3, .25	0.981	0.711	0.977	0.735	0.920	0.716
.8, .3, .5	0.986	0.668	0.985	0.684	0.920	0.717
.8, .3, 1	0.988	0.626	0.988	0.622	0.920	0.717

$n = 1000$, $p = 20$, $p_0 = 3$, $s = 10$, $s_0 = 3$ and $q = 43$.						
	AENet		ALASSO-LARS		ALASSO-CL	
$\rho_z, \tau_{\mathcal{A}}, b$	Pr_1	Pr_2	Pr_1	Pr_2	Pr_1	Pr_2
.5, .5, .25	0.998	1.000	0.997	1.000	0.991	1.000
.5, .5, .5	0.998	1.000	0.998	1.000	0.991	1.000
.5, .5, 1	0.998	1.000	0.998	1.000	0.991	1.000
.5, .3, .25	0.997	1.000	0.996	1.000	0.991	0.982
.5, .3, .5	0.997	1.000	0.997	1.000	0.991	0.982
.5, .3, 1	0.997	1.000	0.997	1.000	0.991	0.982
.8, .5, .25	0.998	1.000	0.998	1.000	0.990	1.000
.8, .5, .5	0.999	1.000	0.999	1.000	0.990	1.000
.8, .5, 1	0.999	1.000	0.999	1.000	0.990	1.000
.8, .3, .25	0.998	1.000	0.997	1.000	0.990	0.984
.8, .3, .5	0.999	1.000	0.999	1.000	0.990	0.984
.8, .3, 1	0.999	1.000	0.999	1.000	0.991	0.982

Note: AENet is the estimator defined in (3) and solved by the LARS algorithm. ALASSO-LARS is the same as AENet except that λ_2 is restricted to be zero. ALASSO-CL is the estimator proposed by Cheng and Liao (2012). ρ_z controls the correlation among valid instruments. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. b is the value of nonzero structural parameters. Pr_1 is the percentage of replications that yield zero estimates for $\tau_{\mathcal{A}^c}$. Pr_2 is the percentage of replications that yield nonzero estimates for $\tau_{\mathcal{A}}$.

Table 3: Model Selection Accuracy

	$n = 250$				$n = 1000$			
	AENet		ALASSO-LARS		AENet		ALASSO-LARS	
$\rho_z, \tau_{\mathcal{A}}, b$	Pr_3	Pr_4	Pr_3	Pr_4	Pr_3	Pr_4	Pr_3	Pr_4
.5, .5, .25	0.934	0.908	0.930	0.904	0.993	1.000	0.994	1.000
.5, .5, .5	0.947	1.000	0.946	1.000	0.995	1.000	0.996	1.000
.5, .5, 1	0.947	1.000	0.947	1.000	0.995	1.000	0.995	1.000
.5, .3, .25	0.918	0.920	0.913	0.919	0.992	1.000	0.992	1.000
.5, .3, .5	0.930	1.000	0.928	1.000	0.993	1.000	0.994	1.000
.5, .3, 1	0.930	1.000	0.930	1.000	0.993	1.000	0.994	1.000
.8, .5, .25	0.921	0.680	0.919	0.669	0.988	0.922	0.987	0.917
.8, .5, .5	0.942	0.975	0.941	0.971	0.995	1.000	0.995	1.000
.8, .5, 1	0.947	1.000	0.947	1.000	0.996	1.000	0.996	1.000
.8, .3, .25	0.915	0.690	0.908	0.680	0.984	0.934	0.983	0.928
.8, .3, .5	0.933	0.975	0.929	0.971	0.990	1.000	0.990	1.000
.8, .3, 1	0.944	1.000	0.943	1.000	0.990	1.000	0.990	1.000

Note: AENet is the estimator defined in (3) and solved by the LARS algorithm. ALASSO-LARS is the same as AENet except that λ_2 is restricted to be zero. ρ_z controls the correlation among valid instruments. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. b is the value of nonzero structural parameters. Pr_3 is the percentage of replications that yield zero estimates for $\beta_{\mathcal{A}^c}$. Pr_4 is the percentage of replications that yield nonzero estimates for $\beta_{\mathcal{A}}$.