Spring 1-2015

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

Mehmet Caner
North Carolina State University at Raleigh, mcaner@ncsu.edu

Xu Han
City University of Hong Kong, xuhan25@cityu.edu.hk

Yoonseok Lee
Syracuse University, ylee41@maxwell.syr.edu

THE MAXWELL SCHOOL

CENTER FOR POLICY RESEARCH

WORKING PAPER SERIES

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

Mehmet Caner, Xu Han, and Yoonseok Lee

ISSN: 1525-3066

426 Eggers Hall
Syracuse University
Syracuse, NY 13244-1020
(315) 443-3114 / email: ctrpol@syr.edu

Paper No. 177
January 2015

http://www.maxwell.syr.edu/CPR_Working_Papers.aspx

CENTER FOR POLICY RESEARCH –Spring 2015

Leonard M. Lopoo, Director
Associate Professor of Public Administration and International Affairs (PAIA)

Associate Directors

Margaret Austin
Associate Director
Budget and Administration

John Yinger
Trustee Professor of Economics and PAIA
Associate Director, Metropolitan Studies Program

SENIOR RESEARCH ASSOCIATES

Badi H. Baltagi.. Economics
Robert Bifulco...PAIA
Thomas Dennison ....PAIA
Alfonso Flores-Lagunes .... Economics
Sarah Hamersma ....PAIA
William C. Horrace .... Economics
Yilin Hou ....PAIA
Duke Kao... Economics
Sharon Kioko .... PAIA
Jeffrey Kubik.. Economics
Yoonseok Lee .... Economics
Amy Lutz.. Sociology
Yingyi Ma... Sociology
Jerry Miner ....Economics
Cynthia Morrow .... PAIA
Jan Ondrich..Economics
John Palmer.. PAIA
David Popp .... PAIA
Stuart Rosenthal ....Economics
Ross Rubenstein... PAIA
Rebecca Schewe ....Sociology
Amy Ellen Schwartz .... PAIA/Economics
Perry Singleton………………………….....Economics
Abbey Steele... PAIA
Michael Wasylenko... ………………………Economics
Peter Wilcoxen...…PAIA

GRADUATE ASSOCIATES

Emily Cardon... PAIA
Hannah Dalager .... PAIA
Maidel De La Cruz... PAIA
Carlos Diaz ....Economics
Vantiel Elizabeth Duncan .... PAIA
Alex Falevich ....Economics
Lincoln Groves .... PAIA
Ruby Jennings... PAIA
Yusun Kim .... PAIA
Bridget Lenkiewicz .... PAIA
Michelle Lofton .... PAIA
Roberto Martinez.. PAIA
Qing Miao ....PAIA
Nuno Abreu Faro E Mota............................ Economics
Judson Murchie ....PAIA
Sun Jung Oh .... Social Science
Brian Ohl...PAIA
Laura Rodriquez-Ortiz ....PAIA
Timothy Smilnak ....PAIA
Kelly Stevens ....PAIA
Rebecca Wang .... Sociology
Pengju Zhang .... Economics
Xirui Zhang .... Economics

STAFF

Kelly Bogart....……………..….Administrative Specialist
Karen Cimilluca.....…………………..….Office Coordinator
Kathleen Nasto........................Administrative Assistant
Candi Patterson.……………..……Computer Consultant
Mary Santy.…………………….....Administrative Assistant
Katrina Wingle......………….…Administrative Assistant

Abstract

This paper develops the adaptive elastic net GMM estimator in large dimensional models with many possibly invalid moment conditions, where both the number of structural parameters and the number of moment conditions may increase with the sample size. The basic idea is to conduct the standard GMM estimation combined with two penalty terms: the quadratic regularization and the adaptively weighted lasso shrinkage. The new estimation procedure consistently selects both the nonzero structural parameters and the valid moment conditions. At the same time, it uses information only from the valid moment conditions to estimate the selected structural parameters and thus achieves the standard GMM efficiency bound as if we know the valid moment conditions ex ante. It is shown that the quadratic regularization is important to obtain the efficient estimator. We also study the tuning parameter choice, with which we show that selection consistency still holds without assuming Gaussianity. We apply the new estimation procedure to dynamic panel data models, where both the time and cross section dimensions are large. The new estimator is robust to possible serial correlations in the regression error terms.

Keywords and phrases: Adaptive Elastic Net, GMM, many invalid moments, large dimensional models, efficiency bound, tuning parameter choice, dynamic panel.

JEL No. C13, C23, D26

Keywords: Adaptive Elastic Net, GMM, many invalid moments, large dimensional models, efficiency bound, turning parameter choice, dynamic panel

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

Xu Han-City University of Hong Kong, Department of Economics and Finance, Hong Kong.
Email: xuhan25@cityu.edu.hk

Yoonseok Lee-Syracuse University, Department of Economics and Center for Policy Research, 426 Eggers Hall, Syracuse, NY 13244
Email: ylee41@maxwell.syr.edu

1 Introduction

Structural parameter estimation with endogenous regressors is a very common issue in applied econometrics. For proper inferences, however, researchers need to choose the valid instruments or moment conditions as well as the correct structural model before estimation. When the number of moment conditions is small and fixed, the moments/models are normally justified using some based on some economic theory or intuition; pre-testing procedures based on over-identifying restrictions tests are also commonly used, though any ad hoc moment/model selection could affect the post-selection inferences (e.g., Leeb and Pötscher, 2005). Such issues gain more importance in high dimensional models since we have a higher chance of misspecification with many endogenous regressors and many instruments/moment conditions. Unfortunately, the standard statistical tool may not be used immediately with a large number of instruments. For example, the standard over-identifying restrictions test or some moment/model selection procedure (e.g., Andrews, 1999; Andrews and Lu, 2001) are computationally challenging and hard to be generalized to the case of large dimensional models; they may not even follow the standard asymptotics (e.g., Newey and Windmeijer, 2009; Lee and Okui, 2012). Therefore, with many endogenous regressors and many moment conditions, the validity of moments and selection of the model come to the fore. In this case, shrinkage methods can be useful.

Since the valid moment conditions depend on the validity of the instruments as well as the correct model specification, in fact, they should be considered together. However, the existing literature on the shrinkage GMM/two-step method mostly does not question on the validity of the instruments; it is normally assumed that all the available instruments are valid (i.e., orthogonal to the structural error). For example, a seminal paper by Belloni, Chernozhukov, Chen, and Hansen (2012) introduces a heteroskedasticity consistent lasso estimator and provides finite sample performance bounds, but it focuses on the optimal instrument selection given that all the instruments are valid. Caner and Zhang (2014) consider the adaptive elastic net GMM estimation with many structural parameters and instruments, but they assume all the instruments are valid.

This paper develops the adaptive elastic net GMM estimator in large dimensional models with many possibly invalid moment conditions, where both the number of structural parameters and the number of moment conditions may increase with the sample size. The basic idea is to conduct the standard GMM estimation combined with two penalty terms: the quadratic regularization ℓ_2-penalty and the adaptively weighted lasso shrinkage ℓ_1-penalty. So the key contribution of this paper is to handle both the valid moment condition (or instruments) selection problem and the correct model selection problem simultaneously. The new estimation procedure is shown to consistently select both the nonzero structural parameters and the valid moment conditions.

Furthermore, the new estimator uses information only from the valid moment conditions to estimate the selected structural parameters and thus achieves the standard GMM efficiency bound as if we know the valid moment conditions *ex ante*. To achieve the efficiency bound, it is shown that including the ℓ_2-penalty of the quadratic regularization is important in this particular problem. It is because this ridge penalty controls for the possible (near) multicollinearity problem among the instruments in the first stage regression, so that it allows for the estimation procedure to select all the valid instruments even when they are highly correlated with each other. Apparently, including more valid instruments will improve the efficiency of the GMM estimator.

We also discuss about the tuning parameter choice by developing a BIC-type criterion, based on which we can still achieve the model/moment selection consistency. Unlike the results in the statistics literature (e.g., Wang, Li, and Leng, 2009), our selection consistency result is obtained without assuming Gaussianity. In addition, this paper shows that the Least Angle Regression (LARS) algorithm by Efron, Hastie, Johnstone and Tibshirani (2004) can be extended to our large-dimensional GMM framework. This algorithm gives a great computational advantage over downward or upward testing procedures, especially in this large-dimensional setup. As an illustration, we apply the new estimation procedure to dynamic panel regression models with fixed effect, where both the time and cross section dimensions are large. The new estimator can be useful since it is robust to possible serial correlations in the error terms of dynamic panel regressions.

There are some studies, that are closely related with the current paper, on the shrinkage method with increasing number of moment conditions. Gautier and Tsybakov (2011) provide finite sample performance bounds for Danzig selector when there are large number of invalid instruments. Fan and Liao (2012) analyze the ultra high dimensional case when the number of moments are larger than the sample size. Cheng and Liao (2013) provide asymptotic results in adaptive lasso when there are many invalid moments. However, the current paper is different from the aforementioned ones in the following sense. First, we develop the adaptive elastic net GMM estimation procedure, which selects both the correct model and the valid moment conditions at the same time, when both dimensions are large. Second, unlike the lasso method, by including the ℓ_2 penalty as well as the ℓ_1 penalty, we are able to control for the multicollinearity problem among the instruments so that we can achieve the efficient GMM estimator. Also note that, though the current paper appears similar to Caner and Zhang (2014), their technical aspects are fundamentally different because we allow for invalid moment conditions.

The remainder of the paper is organized as follows. Section 2 introduces the basic setup and the adaptive elastic net GMM estimator. Section 3 provides some technical assumptions and develops the oracle property of the new estimator. As an illustration, the new estimation procedure is applied to dynamic panel data regressions. Section 4 discusses some computational issues including tuning parameter choice based on BIC-type criterion and computational algorithm using the LARS. Section 5 provides simulation results and Section 6 concludes. Proofs are given in the Appendix.

2 Adaptive Elastic Net GMM

2.1 The setup

We consider a structural equation given by

$$Y_i = X_i' \beta_0 + u_i \tag{1}$$

for $i = 1, 2, \cdots, n$, where X_i is the $p \times 1$ vector of endogenous regressors and β_0 is the $p \times 1$ true structural parameter vector. We assume the $q \times 1$ vector of instrumental variables Z_i. For simplicity, we assume all variables are demeaned. We allow that both the number of endogenous regressors p and the number of instruments q increase with the sample size n. However, we assume that some components of β_0 are zero so that the true model has a sparse representation. We denote p_0 as the number of nonzero components of β_0. Similarly, we assume that the set of q-number of instrumental variables Z_i is a mixture of valid (i.e., they are uncorrelated with u_i) and invalid (i.e., they are correlated with u_i) instruments. We denote s_0 as the number of invalid instruments in Z_i, that is s_0 corresponds to the number of nonzero components of the $q \times 1$ moment condition vector $E[Z_i u_i]$.

Though we have limited information about the validity of the moment conditions, we presume a minimal set of valid moment conditions in Z_i. We assume that at least $(q - s)$-number of moment conditions are valid, where $p \leq (q - s) \leq (q - s_0)$ so that β_0 in (1) is well identified. In this case, s is the maximal number of invalid moment conditions such that $s_0 \leq s$, which is restricted as $(p + s) \leq q < n$. Note that we assume $p < n$ and $q < n$ but $p + q$ can be larger than n.[1] More

[1] We do not consider the case of $q > n$ in this paper, which is completely a different problem. For the case of convex loss with exogenous covariates, recently Caner and Kock (2014) handle oracle inequalities and estimation errors. Note that the techniques are entirely different from this paper because of the singularity of Gram matrix.

precisely, we rewrite the q moment conditions as

$$E[Z_i u_i] - F\tau_0 = E[Z_i(Y_i - X_i'\beta_0) - F\tau_0] = 0 \tag{2}$$

for each $i = 1, 2 \cdots, n$, where F is the $q \times s$ matrix given by

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

with $0_{q-s,s}$ being the $(q-s) \times s$ matrix of zeros and I_s is the identity matrix with rank s. The $s \times 1$ vector τ_0 includes all the s_0-number of nonzero components of the moment condition $E[Z_i u_i]$. For example, when all the moment conditions are valid and the researcher also believes so, it is simply the case of $s = 0$, yielding the standard GMM case with $E[Z_i u_i] = 0$.

Note that some of the elements of τ_0 can be zero since $s_0 \leq s$. Therefore, the subset of moment conditions $E[Z_i u_i]$ that correspond to the first-block of F (i.e., $0_{q-s,s}$) are the minimal set of valid moment conditions, which are required to know for identification, whereas the moment conditions that correspond to the second-block of F (i.e., I_s) are potentially invalid moment conditions. The main purpose of this paper is to develop a simultaneous procedure of model selection (i.e., choosing the regressors that correspond to nonzero β_0) and valid moment condition selection (i.e., choosing the instrumental variables in Z_i that are uncorrelated with u_i) among this set of potentially invalid moment conditions, as well as efficient estimation of the nonzero components of the structural parameters β_0.

2.2 The adaptive elastic net GMM estimator

For estimation, we generalize the adaptive elastic net estimation of Zou and Zhang (2009) to the GMM setup. The basic idea is to conduct the standard GMM combined with two penalty terms (i.e., the quadratic regularization and the adaptively weighted lasso shrinkage) so that both nonzero β_0 and valid moment conditions in (2) are correctly chosen as well as their consistent estimators are obtained simultaneously. We let $X = (X_1, \cdots, X_n)'$ and similarly for Y, Z and u. We define a $q \times 1$ vector $Y_z = Z'Y = \sum_{i=1}^n Z_i Y_i$, a $q \times (p+s)$ matrix $X_{zF} = [Z'X, nF] = \sum_{i=1}^n [Z_i X_i', F]$, and a $(p+s) \times 1$ vector of parameters $\theta_0 = (\beta_0', \tau_0')' \in \mathbb{R}^{p+s}$. In this setup, the *adaptive elastic net GMM estimator* for θ_0 is defined as

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

from the moment condition (2), where $\hat{W}$ is some $q \times q$ symmetric and positive definite weight matrix, and λ_1^* and λ_2 are some positive tuning parameters. $\{\hat{\pi}_j\}_{j=1}^{p+s}$ are some data-dependent weights and they are usually obtained as $\hat{\pi}_j = |\hat{\theta}_{j,enet}|^{-\gamma}$ with some $\gamma > 1$, where $\hat{\theta}_{enet} = (\hat{\theta}_{1,enet}, \cdots, \hat{\theta}_{p+s,enet})'$ denotes the (*naive*) elastic net estimator. $\hat{\theta}_{enet}$ is obtained by minimizing (3) with $\hat{\pi}_j = 1$ for all j and without the scaling factor $1 + \lambda_2/n^2$:

$$\hat{\theta}_{enet} = \arg\min_{\theta} \left\{ (Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_1 \sum_{j=1}^{p+s} |\theta_j| + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\}, \tag{4}$$

in which λ_1 can be difference from λ_1^* in (3). So in practice, we run elastic net to obtain data dependent weights $\hat{\pi}_j$ in the first step and run the adaptive elastic net using $\hat{\pi}_j$ in the second step. See Zou and Zhang (2009) for further details in the context of the least squares adaptive elastic net estimator.

Note that the scaling factor $1 + \lambda_2/n^2$ in (3) will undo the shrinkage from the ridge penalty and thus reduce extra bias caused by double shrinkage. Unlike the least squares case, however, we use a finite sample correction of $1 + \lambda_2/n^2$ instead of $1 + \lambda_2/n$ that is used by Zou and Zhang (2009) and Caner and Zhang (2014). The reason for this different scaling factor becomes clear later, but intuitively it is because the GMM objective function is of the quadratic form of the sample average, whereas the least squares objective function is simply the sample average.

The objective function in (3) includes two penalty terms. The first ℓ_1 penalty term corresponds to the adaptively weighted lasso shrinkage of Zou (2006) for both β_0 and τ_0, which results in consistent model (for nonzero components of β_0) and moment (for nonzero components of τ_0) selections. On the other hand, the second ℓ_2 penalty term, which corresponds to the quadratic regularization, is included mainly for the moment selection problem.[2] Basically it will resolve possible collinearity problem particularly among the instrumental variables Z_i. Note that introducing the ℓ_2 penalty allows the procedure to select all the valid instruments even when they are highly correlated with each other. Using more number of instruments will improve the predictability from the first stage regression, which results in more efficient estimator of β_0. Moreover, in this setup, including the ℓ_2 penalty will result in a less biased estimator of β_0.

For example, suppose that there are two invalid instruments that are highly correlated. Without this ℓ_2 penalty, the ℓ_1 penalty will choose only one of them as an invalid one, which will result in that the other remains in the pool of valid instruments. Apparently, this result will yield a biased

[2]Note that if the main purpose is to predict Y_i, then adding the ℓ_2-penalty also improves the predictability of (1) as emphasized in Zou and Hastie (2005).

estimator of β_0. On the other hand, when both of the highly correlated instruments are valid and if the ℓ_1 penalty only choose one of them, then it will result in less efficient GMM estimator. Note that when we include the ℓ_2 penalty, those two instruments are more likely to be selected (or not) together. This idea is confirmed in the simulation studies below, in which the RMSE estimate of the structural parameter estimators is smaller with the inclusion of the ℓ_2 penalty when the correlation between valid instruments is high.

3 Statistical Theory

3.1 Assumptions

We first provide technical conditions for the main theorems. We suppose triangular arrays $\Upsilon_{in} = (X'_{in}, Z'_{in}, u_{in})' \in \mathbb{R}^{p+q+1}$ for $i = 1, 2, \cdots, n$ and $n = 1, 2, \cdots$ defined on the probability space $(\Omega, \mathcal{B}, P_n)$, where the probability measure P_n can change with n. For the sake of simplicity, we assume $\{\Upsilon_{in}\}$ are independent and identically distributed across i for each n, though they do not need to be identically distributed. All parameters that characterize the distribution of Υ_{in} are implicitly indexed by P_n and hence by n. We suppress the subscript n to simplify the notation though.

We first let

$$e_i = Z_i u_i - E[Z_i u_i] = Z_i u_i - F\tau_0 \tag{5}$$

for each i, based on which the moment condition (2) can be simply rewritten as $E[e_i] = 0$. Throughout the paper, we let $\xrightarrow{p}$ denote the convergence in probability, and $||D|| = [tr(D'D)]^{1/2}$ for any matrix D.

Assumption 1. *(i) $\|\hat{W} - W\| \xrightarrow{p} 0$ as $n \to \infty$, where W is a $q \times q$ symmetric, bounded and positive definite matrix. (ii)$\{X_i, Z_i, u_i\}_{i=1}^n$ are independent and identically distributed over i. We also have $\|n^{-1}\sum_{i=1}^n e_i e_i' - V\| \xrightarrow{p} 0$ as $n \to \infty$, where V is a $q \times q$ symmetric, bounded and and positive definite matrix. (iii) $\|n^{-1}Z'X - \Sigma_{zx}\| \xrightarrow{p} 0$ as $n \to \infty$, where Σ_{zx} is a $q \times p$ bounded matrix of full column rank p.*

As noted in Newey and Windmeijer (2009), Assumption 1 restricts the rate at which q can grow with the sample size n, which is frequently used in the many (weak) moment literature. Assumption 1-(ii) defines the variance matrix of e_i or equivalently that of u_i, which takes into account the effect of moment invalidity. Assumption 1-(iii) assumes that all the instruments Z_i are strongly correlated with the endogenous regressors X_i so that Σ_{zx} has full column rank.[3] It also implies that

$$\|n^{-1}X_{zF} - \Sigma_{zxF}\| \xrightarrow{p} 0, \tag{6}$$

where $\Sigma_{zxF} = [\Sigma_{zx}, F]$ is a $q \times (p+s)$ matrix of full column rank $(p+s)$. It follows that, for each given q,

$$0 < \text{Eigmin}(\Sigma'_{zxF} W \Sigma_{zxF}) \quad \text{and} \quad \text{Eigmax}(\Sigma'_{zxF} W \Sigma_{zxF}) < \infty, \tag{7}$$

where Eigmax$(\cdot)$ and Eigmin$(\cdot)$ denote the maximal and the minimal eigenvalues of a matrix, respectively, since W is a $q \times q$ symmetric, bounded and positive definite matrix. From Assumption 1-(i), therefore, (6) and (7) implies that there exist positive constants b and B, which do not depend on n, such that

$$\text{Eigmax}(n^{-2} X'_{zF} \hat{W} X_{zF}) \leq B < \infty \quad \text{and} \quad \text{Eigmin}(n^{-2} X'_{zF} \hat{W} X_{zF}) \geq b > 0 \tag{8}$$

with probability approaching one (w.p.a.1, hereafter) from Newey and Windmeijer (2009, Lemma A0). Similarly, from Assumption 1-(iii), we also have

$$\text{Eigmax}(n^{-2} \hat{W} X_{zF} X'_{zF} \hat{W}) \leq B < \infty \tag{9}$$

w.p.a.1, which will control for the second moment of the estimators when there are many invalid instruments.

We let $\mathcal{A} = \{j : \theta_{j,0} \neq 0, j = 1, 2, \cdots, p+s\}$, which collects the index of nonzero coefficients in θ_0. The minimum absolute value of the nonzero coefficients is denoted as

$$\eta = \min_{j \in \mathcal{A}} |\theta_{j,0}|,$$

which may depend on n and possibly local to zero. We impose conditions on the tuning parameters as follows. Note that the tuning parameters λ_1, λ_1^*, and λ_2 all diverge to infinity as $n \to \infty$.

Assumption 2. *(i)* $\lim_{q,n\to\infty} q/n^{\alpha} < \infty$ *and* $\lim_{p,s,n\to\infty}(p+s)/n^{\nu} < 1$ *for some* $0 \leq \nu \leq \alpha < 1$, *where* $(p+s) \leq q$ *for any* n. *(ii) There exist positive constants* γ *and* κ *satisfying* $\alpha < (\kappa - 3) < \gamma(1-\alpha) - \nu$. *(iii)* $\lambda_2^2 \|\theta_0\|^2 / n^3 \to 0$ *and* $\lambda_1^2/n^3 \to 0$ *as* $n \to \infty$. *(iv)* $\lambda_1^{*2}(p+s)/n^3\eta^{2\gamma} \to 0$ *but* $\lambda_1^{*2}/n^{\kappa - \gamma(1-\alpha)} \to \infty$ *as* $n \to \infty$.

[3] Under the case with many weak moment conditions, GMM estimation normally yield inconsistent estimators (e.g., Newey and Windmeijer, 2009). For Lasso type estimators, same problem is pointed in Caner (2009) by showing that even with fixed number of instruments only nearly-weak asymptotics can give consistent estimates. We think that many weak moment case will be interesting but it needs to be handled in the GEL or CUE framework, which is outside the scope of this paper.

Assumption 2 establishes the rates for tuning parameters as a function of the total number of moment conditions and the number of parameters. Note that the total number of moment conditions q can come arbitrarily close to the full sample size n when α is close to one.[4] Recall that γ is chosen for defining the weights in the lasso penalty (i.e., $\hat{\pi}_j = |\hat{\theta}_{j,enet}|^{-\gamma}$), and thus γ is closely related with the degree of penalty on the small coefficients. Assumption 2-(ii) requires that γ is to be chosen such that $\gamma > (\alpha+\nu)/(1-\alpha)$, where the same conditions can be found in Zou and Zhang (2009) when $\alpha = \nu$. Once γ is determined, the tuning parameters λ_1, λ_1^*, and λ_2 are to be chosen according to Assumptions 2-(iii) and (iv). Note that these conditions allows for larger values of the tuning parameters than those of Zou and Zhang (2009) or Caner and Zhang (2014), though we can still choose smaller values similar to theirs. The constant κ is introduced for the technical reason proving selection consistency.

It is important to note that Assumption 2-(iv) allows the nonzero coefficients to be local-to-zero but it restricts the rate at which the nonzero coefficients should vanish so that they can be distinguished from the true zero coefficients. In fact, from this condition, we are able to come up with the lower bound of the local-to-zero rate of η: if a nonzero coefficient is local-to-zero but it vanishes faster than this rate, it cannot be selected as nonzero in our adaptive elastic net GMM procedure. Apparently this condition also imposes restrictions on the tuning parameter λ_1^* for the ℓ_1 penalty in (3) so that we can achieve the selection consistency in Theorem 1 below. In comparison, Assumption 2-(iii) on λ_1 for the ℓ_1 penalties in (4) and (11) is required to obtain an asymptotically negligible upper bound of the estimation error of the elastic net estimator.

The lower bound of the local-to-zero rate of η depends on the number of structural parameters p, the number of moment conditions q, and the maximal number of (potentially) invalid moment conditions s. If either the number of moments or parameters increases, then the required lower bound of the threshold defining the local-to-zero parameter gets larger. For example, we suppose $\eta = n^{-1/\xi}$ for some $\xi > 0$. Then Assumption 2-(ii) and (iv) imply that

$$\lambda_1^{*2}(p+s)/n^3\eta^{2\gamma} = c\lambda_1^{*2}n^{\nu+(2\gamma/\xi)-3} \to 0$$

for some positive constant $c < 1$, and thus

$$n^{-\kappa+\gamma(1-\alpha)}/n^{\nu+(2\gamma/\xi)-3} \to \infty$$

[4] If $\max_{1\leq i\leq n}\max_{1\leq j\leq q} E(Z_{j,i}u_i)^4 < \infty$ as Newey and Windmeijer (2009, p.706), where $Z_{j,i}$ is the jth instrument of Z_i, we need $q^2/n \to 0$ to satisfy Assumption 1-(i). Since $q = O(n^\alpha)$ and $(p+s) = O(n^\nu)$, however, it restricts that $0 \leq \nu \leq \alpha < 1/2$.

as $\lambda_1^{*2}/n^{\kappa-\gamma(1-\alpha)} \to \infty$, which requires $-\kappa+\gamma(1-\alpha) > \nu + (2\gamma/\xi) - 3$ or

$$\xi > \frac{2\gamma}{\gamma(1-\alpha)-\nu-\kappa+3} \equiv \xi^*. \tag{10}$$

Note that $\xi^* > 0$ since $\gamma > 0$ and $\gamma(1-\alpha)-\nu-\kappa+3 > 0$ from Assumption 2-(ii). Since $q = O(n^\alpha)$ and $(p+s) = O(n^\nu)$ in Assumption 2-(i), it thus follows that the lower bound of the local-to-zero rate of η increases as the number of moment conditions q (i.e., the size of α) or the number of parameters $p+s$ (i.e., the size of ν) gets larger. As an illustration, when the system is just identified with $\alpha = \nu = 1/2$, we have $\gamma > 2$ and $7 < 2\kappa < 5+\gamma$ from Assumption 2-(ii), which can be satisfied with $\gamma = 4$ and $\kappa = 4$. The true local-to-zero but nonzero coefficients thus cannot vanish faster than $n^{-1/\xi^*} = n^{-1/16}$ in order to be selected as nonzero. For example, even when $n = 10^5$, $\eta = \min_{j\in\mathcal{A}} |\theta_{j,0}|$ is about 0.5 in this case. This illustration shows that in an environment with many moments and/or parameters, it will be difficult to do perfect model selection unless the coefficients are large enough.

We let $\theta_{\mathcal{A}} = (\beta'_{\mathcal{A}}, \tau'_{\mathcal{A}})'$, which collects the nonzero parameters in θ_0. Then $\tau_{\mathcal{A}}$ is an $s_0 \times 1$ nonzero subvector of τ_0 that represents invalid moment conditions. We also let $F_{\mathcal{A}} = [0'_{q-s_0,s_0}, I_{s_0}]'$. The last condition is useful to obtain the Lyapunov condition in Theorem 2 below, which is similar to what Zou and Zhang (2009) assumes in the context of simple least squares.

Assumption 3. $\max_{1\le i\le n} \|n^{-1/2}(Z_i u_i - F_{\mathcal{A}}\tau_{\mathcal{A}})\|^2 \xrightarrow{p} 0$ *as* $n \to \infty$.

3.2 The oracle property

For analytical convenience, we define

$$\hat{\theta}_\pi = \arg\min_\theta \left\{ (Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_1 \sum_{j=1}^{p+s} \hat{\pi}_j |\theta_j| + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\} \tag{11}$$

for nonnegative tuning parameters λ_1 and λ_2, where λ_1 is introduced in (4) and $\hat{\pi}_j = |\hat{\theta}_{j,enet}|^{-\gamma}$ in (3). Apparently, this estimator becomes the elastic net GMM estimator in (4) when $\hat{\pi}_j = 1$ for all j; or the adaptive elastic net GMM estimator in (3) with the scaling factor $1+\lambda_2/n^2$ and the tuning parameter λ_1^* instead of λ_1. We first obtain the risk bound of this interim estimator, which is useful to obtain the selection consistency in Theorem 1 and the asymptotic normality in Theorem 2 below.[5]

[5] This result seems quite similar to that of Zou and Zhang (2009). But this result is obtained w.p.a.1 since we consider stochastid regressors in the GMM setup; whereas Zou and Zhang's result is exact since they consider deterministic regressors in the least squares setup.

Lemma 1. *Under the model (1), (2) and Assumption 1, we have*

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

w.p.a.1, where B and b are some positive constants given in (8) and (9).

Since the risk bounds in Lemma 1 go to zero as $n \to \infty$ under Assumptions 1 and 2, both $\hat{\theta}_\pi$ and $\hat{\theta}_{enet}$ are consistent estimators of θ_0. Therefore, we can use $\hat{\theta}_{enet}$ to construct the adaptive weight $\hat{\pi}_j$ in (3). Notice that, as in Zou and Zhang (2009) and Caner and Zhang (2014), the upper bounds are formulated using $\|\theta_0\|^2$ instead of the sparsity of the vector θ_0 (i.e., $p+s$). In comparison, the sparsity index is commonly used in the high dimensional models literature (e.g., Belloni, Chen, Chernozhukov, and Hansen, 2012), which is because of singularity of the design matrix.

Note that the mean squared error expressions depend inversely on b, which is a lower bound of the minimum eigenvalue of $n^{-2}X_{zF}'\hat{W}X_{zF}$ defined in (8). When the regressors are highly correlated with each other, b can be close to zero. In this case, without the tuning parameter λ_2 in the denominator (e.g., lasso or adaptive lasso), the error bounds in Lemma 1 can be quite large. This implies that the error bound of the adaptive elastic net estimator will be smaller than that of the adaptive lasso estimator when b is close to zero because of highly correlated regressors.

We now obtain one of the main results: the selection consistency. This result shows that the adaptive elastic net GMM procedure automatically selects the valid moment conditions as well as the relevant regressors in the structural equation. This extends Zou and Zhang (2009) in two ways: it finds the relevant regressors in the linear GMM setup instead of the linear least squares case; it also tells if each moment condition is valid or not.

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

The selection consistency in Theorem 1, which mainly comes from the ℓ_1 penalty in (3),[6] means that the true nonzero coefficients are to be selected as nonzero. For some local-to-zero coefficients, they are to be concluded as nonzero coefficients provided that they vanish slower than the lower bound of the local-to-zero rate in Assumption 2. As we discussed after Assumption 2, however, it

[6] In principle, we expect that the same goals can be achieved using different types of penalties that possess the oracle property such as Bridge (e.g., Huang, Horowitz and Ma, 2008) and SCAD (e.g., Fan and Li, 2001).

will be difficult to do perfect model selection unless the nonzero coefficients are large enough in an environment with many moments and/or parameters. For example, if we suppose $\eta = O(n^{-1/\xi})$ for some $\xi > 0$, the lower bound of the local-to-zero rate is given by $O(n^{-1/\xi^*})$, where ξ^* given in (10) is normally much larger than 2. In this sense, Theorem 1 extends Leeb and Pötscher (2005)'s criticism to the context of many parameters. Recall that, in the case with fixed number of parameters, they found that the minimum order of which a local-to-zero coefficient is to be distinguished from zero is $n^{-1/2}$, that is smaller than n^{-1/ξ^*} for $\xi^* > 2$.

As the second main result, we derive the limiting distribution of the adaptive elastic net GMM estimator of the nonzero coefficients $\theta_{\mathcal{A}} = (\beta'_{\mathcal{A}}, \tau'_{\mathcal{A}})'$ in the following theorem. Caner and Zhang (2014) also obtain the limiting distribution of the GMM estimator via adaptive elastic net, but their focus is in choosing the nonzero structural parameters of β_0 when all the moment conditions are assumed to be valid. 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 given by

$$\Sigma_{\mathcal{A}} = \Sigma'_{zxF:\mathcal{A}} V^{-1} \Sigma_{zxF:\mathcal{A}}, \tag{12}$$

where $\Sigma_{zxF:\mathcal{A}} = [\Sigma_{zx:\mathcal{A}}, F_{\mathcal{A}}]$ is a full column rank $q \times (p_0 + s_0)$ matrix and $\Sigma_{zx:\mathcal{A}}$ is a full column rank $q \times p_0$ matrix. Recall that $F_{\mathcal{A}} = [0'_{q-s_0,s_0}, I_{s_0}]'$. From the conditions on Σ_{zx} and V in Assumption 1, $\Sigma_{\mathcal{A}}$ is symmetric, bounded and positive definite.

Note that $\Sigma_{zx:\mathcal{A}}$ is defined from $\|n^{-1} Z' X_{\mathcal{A}} - \Sigma_{zx:\mathcal{A}}\| \xrightarrow{p} 0$, 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 (6), we have

$$\|n^{-1} X_{zF:\mathcal{A}} - \Sigma_{zxF:\mathcal{A}}\| \xrightarrow{p} 0, \tag{13}$$

where $X_{zF:\mathcal{A}} = [Z' X_{\mathcal{A}}, n F_{\mathcal{A}}]$ is a $q \times (p_0 + s_0)$ matrix.

Theorem 2. *We let $\hat{\theta}_{\mathcal{A}}$ be the adaptive elastic net GMM estimator in (3) that corresponds to $\theta_{\mathcal{A}}$. We also let $\hat{W} = \hat{V}^{-1}$, where $\hat{V}$ is some consistent estimator of V. Under Assumptions 1-3, the limiting distribution of $\hat{\theta}_{\mathcal{A}}$ is given by*

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

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

From Assumption 1-(ii), $\hat{V}$ can be consistently estimated by $\hat{V} = n^{-1}\sum_{i=1}^{n}\hat{e}_i\hat{e}_i'$ and so is $\hat{W}$, where $\hat{e}_i = Z_i(Y_i - X_i'\tilde{\beta}) - F\tilde{\tau}$ and $(\tilde{\beta}', \tilde{\tau}')'$ is the $(p+s)\times 1$ vector of initial adaptive elastic net GMM estimators in (3) with the weight $\hat{W} = I_q$. Since $\hat{e}_i = e_i - Z_iX_i'(\tilde{\beta} - \beta_0) - F(\tilde{\tau} - \tau_0)$ and $(\tilde{\beta}', \tilde{\tau}')'$ is consistent to $(\beta_0', \tau_0')'$, it can be verified that $\|\hat{V} - V\| \leq \|n^{-1}\sum_{i=1}^{n}\hat{e}_i\hat{e}_i' - n^{-1}\sum_{i=1}^{n}e_ie_i'\| + \|n^{-1}\sum_{i=1}^{n}e_ie_i' - V\| \xrightarrow{p} 0$ as $n \to \infty$.

From (12) and (13), we have $\|\hat{H}_{\mathcal{A}}\| \leq \|\hat{H}_{\mathcal{A}}^{1/2}\|^2 = n^2 tr(n^{-2}X_{zF:\mathcal{A}}'\hat{V}^{-1}X_{zF:\mathcal{A}}) \leq n^2(p_0 + s_0)\text{Eigmax}(n^{-2}X_{zF:\mathcal{A}}'\hat{V}^{-1}X_{zF:\mathcal{A}}) = O_p(n^2(p_0+s_0))$ since $\|n^{-2}X_{zF:\mathcal{A}}'\hat{V}^{-1}X_{zF:\mathcal{A}} - \Sigma_{\mathcal{A}}\| \xrightarrow{p} 0$ from Assumption 1. Therefore, for some positive constant $c < \infty$, we can obtain

$$\begin{aligned} & \left\| \frac{I_{p_0+s_0} + \lambda_2\hat{H}_{\mathcal{A}}^{-1}}{1+(\lambda_2/n^2)} - I_{p_0+s_0} \right\| \qquad (14) \\ \leq & \left\| \frac{I_{p_0+s_0} + c(\lambda_2/n^2)I_{p_0+s_0}}{1+(\lambda_2/n^2)} - I_{p_0+s_0} \right\| + \left\| \frac{(\lambda_2/n^2)\left\{(n^{-2}\hat{H}_{\mathcal{A}})^{-1} - cI_{p_0+s_0}\right\}}{1+(\lambda_2/n^2)} \right\| = o_p(1) \end{aligned}$$

as $n \to \infty$, for $\lambda_2/n^2 \to 0$ from Assumptions 2-(iii). This implies that, though $\hat{\theta}_{\mathcal{A}}$ is a consistent estimator of $\theta_{\mathcal{A}}$, the rate of convergence of $\hat{\theta}_{\mathcal{A}}$ is $\sqrt{n/(p_0+s_0)}$, which is affected by the true number of invalid moment conditions s_0. Therefore, existence of invalid moment conditions makes the rate of convergence of $\hat{\theta}_{\mathcal{A}}$ slower. When $p_0 + s_0$ is fixed, however, it retrieves the optimal rate of $\sqrt{n}$.

Finally, an interesting question is whether we can achieve the efficiency bound of the structural parameter estimators of β from the adaptive elastic net GMM procedure as if we had known all valid instruments. Note that it is generally the case if we use the entire valid (and strong) instruments. Since $\|n^{-2}\hat{H}_{\mathcal{A}} - \Sigma_{\mathcal{A}}\| \xrightarrow{p} 0$, however, Theorem 2 and the result (14) show that $\Sigma_{\mathcal{A}}^{-1} = (\Sigma_{zxF:\mathcal{A}}'V^{-1}\Sigma_{zxF:\mathcal{A}})^{-1}$ corresponds to the asymptotic variance of $\hat{\theta}_{\mathcal{A}}$, which is obtained by letting $W = V^{-1}$ as in the conventional efficient GMM theory. The following theorem shows that the asymptotic variance of the true nonzero structural parameter estimator $\hat{\beta}_{\mathcal{A}}$, which is given by the $p_0 \times p_0$ north-west block of $\Sigma_{\mathcal{A}}^{-1}$, is the same as the efficiency bound obtained when we only use all the valid moment conditions. We decompose $Z = [Z^1, Z^2]$, where Z^1 represents the $n \times (q - s_0)$ valid instruments and Z^2 represents $n \times s_0$ invalid instruments. More precisely, they satisfy $\|n^{-1}\sum_{i=1}^{n}Z_i^1u_i\| \xrightarrow{p} 0$ and $\|n^{-1}\sum_{i=1}^{n}Z_i^2u_i - \tau_{\mathcal{A}}\| \xrightarrow{p} 0$, where $Z_i^{1\prime}$ and $Z_i^{2\prime}$ are the ith row of Z^1 and Z^2, respectively, and $\tau_{\mathcal{A}}$ is the $s_0 \times 1$ vector collecting all the nonzero elements of τ_0.

Theorem 3. *We let $\Sigma_{\mathcal{A}}^{11}$ be the $p_0 \times p_0$ north-west block of $\Sigma_{\mathcal{A}}^{-1}$, which corresponds to the asymptotic variance of the true nonzero structural parameter estimator $\hat{\beta}_{\mathcal{A}}$. Under Assumptions 1-3, it holds that $\Sigma_{\mathcal{A}}^{11} = (\Sigma_{z_1x:\mathcal{A}}'V_{11}^{-1}\Sigma_{z_1x:\mathcal{A}})^{-1}$, where $\|n^{-1}Z^{1\prime}X_{\mathcal{A}} - \Sigma_{z_1x:\mathcal{A}}\| \xrightarrow{p} 0$ and $\|n^{-1}\sum_{i=1}^{n}Z_i^1Z_i^{1\prime}u_i^2 - V_{11}\| \xrightarrow{p} 0$ as $n \to \infty$.*

Note that the efficiency bound of the true nonzero structural parameter estimator is given by $(\Sigma'_{z_1x:\mathcal{A}}V_{11}^{-1}\Sigma_{z_1x:\mathcal{A}})^{-1}$, which can be obtained by only using all the valid moment conditions (i.e., excluding all the invalid moment conditions). Therefore, this result implies that even when we have potentially many invalid instruments, we can still estimate the nonzero elements in β_0 as if we were using only the valid instruments, which gives the oracle result. The adaptive elastic net GMM estimation does not require any pre-testing for the instruments validity.

It is well known that adding more valid instruments will improve the efficiency in the standard GMM setup and Theorem 3 manifests that including all the valid instruments is important to achieve the efficiency bound. As discussed in Section 2.2, therefore, adding the ℓ_2 penalty here is quite important to achieve the efficiency. As shown in Zou and Hastie (2005), the (adaptive) lasso, which only has the ℓ_1 penalty, is more likely to select only one variable among the relevant variables if they are highly correlated. In our setup, omitting ℓ_2 penalty thus could result in selection inconsistency, which means that the procedure fails to select all the valid moment conditions when they are highly correlated with each other. Apparently, omitting valid instruments results in efficiency loss.

3.3 An example: Dynamic panel regression

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

$$y_{i,t} = \rho_0 y_{i,t-1} + x'_{i,t}\delta_0 + \mu_i + u_{i,t} \tag{15}$$

for $i = 1, \cdots, N$ and $t = 1, \cdots, T$, where $|\rho_0| < 1$, $y_{i,t}$ is a scalar, $x_{i,t}$ is a $K \times 1$ vector of strictly 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

$$E[u_{i,t}|\mu_i, y_i^{t-1}, x_i^T] = 0, \tag{16}$$

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 can estimate ρ_0 and δ_0 using the moment conditions given by

$$\begin{aligned} E[\Delta x_{i,t}\Delta u_{i,t}] &= 0 \qquad (17) \\ E[y_i^{t-2}\Delta u_{i,t}] &= 0 \qquad (18) \end{aligned}$$

for $t \geq 2$ as Arellano and Bond (1991), where $\Delta x_{i,t} = x_{i,t} - x_{i,t-1}$ and similarly for $\Delta y_{i,t}$ and $\Delta u_{i,t}$. However, the moment condition (18) is vulnerable since it heavily depend on the condition that $u_{i,t}$ is serially uncorrelated, whereas the moment condition (17) is robust to the possible serial correlations in $u_{i,t}$.[7] So, with potentially serially correlated $u_{i,t}$, we have $q \equiv (T-2)K + (T-2)(T-1)/2$ number of total moment conditions, among which (at most) $s \equiv (T-2)(T-1)/2$ number of moment conditions are potentially invalid under the possible serial correlation in $u_{i,t}$. For identification purposes, however, we need $p \leq (q-s)$, which corresponds to $K+1 \leq (T-2)K$ for all T and K and is thus satisfied with $T \geq 4$. With $N, T, K \to \infty$, we allow for $q, s, p \to \infty$ in this case. However, the ratio of the number of moment conditions q to the sample size $n = NT$ is $q/NT = O(K/N + T/N)$. So in order to satisfy Assumption 2-(i), we need $\max\{K, T\}/N \to 0$ as $N, T, K \to \infty$. But this condition naturally holds for the conventional panel data with large cross-sectional observations.

More precisely, we let $\Delta y_i = (\Delta y_{i,3}, \cdots, \Delta y_{i,T})'$, $\Delta y_{i(-1)} = (\Delta y_{i,2}, \cdots, \Delta y_{i,T-1})'$, $\Delta x_i = (\Delta x_{i,3}, \cdots, \Delta x_{i,T})'$, $\Delta u_i = (\Delta u_{i,3}, \cdots, \Delta u_{i,T})'$, and the $q \times (T-2)$ instrument matrix $Z_i = [Z_{xi}, Z_{yi}]'$, where

$$Z_{xi} = \begin{pmatrix} \Delta x_{i,3} & 0 & \cdots & 0 \\ 0 & \Delta x_{i,4} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Delta x_{i,T} \end{pmatrix}_{(T-2)\times(T-2)K} \quad \text{and}$$

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

Then, with possible serial correlations in $u_{i,t}$, we have the $q \times 1$ moment condition

$$E\left[Z_i \Delta u_i - F\tau_0\right] = 0$$

for each i as (2), where τ_0 is a $s \times 1$ vector of unknown parameters. The adaptive elastic net GMM

[7] Under an additional condition of the mean stationarity (i.e., $E[y_{i,t}] = \mu$ for all i and t), we further have $E[\Delta y_{i,t-1}(y_{i,t} - \rho y_{i,t-1} - x_{i,t}'\delta)] = 0$ for $t \geq 2$ as Blundell and Bond (1998) and Bun and Kleibergen (2013). When ρ is close to one, the moment condition (18) is prone to have weak identification (i.e., weak instrument problem) whereas this new moment condition is robust to such a persistency. We could find valid more moment conditions (e.g., $E[y_i^{t-h} \Delta u_{i,t}] = 0$ for some $h > 2$ under m-dependence type restriction on $u_{i,t}$; $E[\Delta x_{i,s} \Delta u_{i,t}] = 0$ 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 (18).

estimator of $\theta_0 = (\rho_0, \delta_0', \tau_0')'$ is then given by

$$\hat{\theta} = \left(1 + \frac{\lambda_2}{(NT)^2}\right) \arg \min_{\theta=(\rho,\delta',\tau')'} \left\{ \sum_{i=1}^{N} (Z_i(\Delta y_i - \rho \Delta y_{i(-1)} - \Delta x_i \delta) - F\tau)' \hat{W} \right.$$
$$\left. \times (Z_i(\Delta y_i - \rho \Delta y_{i(-1)} - \Delta x_i \beta) - F\tau) + \lambda_1^* \sum_{j=1}^{p+s} \hat{\pi}_j |\theta_j| + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\}$$

for some positive definite $q \times q$ weight matrix $\hat{W}$, where $p + s = (K+1) + (T-2)(T-1)/2$. For the choice of the optimal weight matrix, we first obtain the elastic net estimator $\tilde{\theta} = (\tilde{\rho}, \tilde{\delta}', \tilde{\tau}')'$ with $\hat{W} = I_q$ and let $\hat{e}_i = Z_i(\Delta y_i - \tilde{\rho} \Delta y_{i(-1)} - \Delta x_i \tilde{\delta}) - F\tilde{\tau}$. Then the optimal adaptive elastic net GMM estimator can be obtained using $\hat{\pi}_j = |\tilde{\theta}_j|^{-\gamma}$ and $\hat{W} = (\sum_{i=1}^{N} \hat{e}_i \hat{e}_i')^{-1}$ as discussed after Theorem 2.

4 Computation

4.1 Tuning Parameter Selection

One important issue of penalized estimation is the choice of the tuning parameters. The conventional approach is to choose the tuning parameters based on some information criterion. For example, Wang, Li, and Leng (2009) show that, for shrinkage estimation of linear models, BIC can be used to select the tuning parameter that produces correct model selection w.p.a.1. Similarly, we also consider the following BIC-type information criterion:

$$IC_\lambda = J(\hat{\theta}(\lambda)) + |\mathcal{S}_\lambda| \ln(n) \max\{\ln[\ln(p+s)], 1\}, \tag{19}$$

where $\lambda = (\lambda_1^*, \lambda_2)'$ is the vector of tuning parameters and $\hat{\theta}(\lambda)$ is the adaptive elastic net GMM estimator defined in (3) indexed by λ. $J(\hat{\theta}(\lambda)) = n^{-1}(Y_z - X_{zF}\hat{\theta}(\lambda))'\hat{W}(Y_z - X_{zF}\hat{\theta}(\lambda))$ is the J-statistic of GMM and $\mathcal{S}_\lambda = \{j : \hat{\theta}_j(\lambda) \neq 0\}$ denotes the collection of nonzero estimates given λ. This criterion is also similar to the BIC-type criteria of Andrews (1999) and Andrews and Lu (2001). But note that the J-statistic $J(\hat{\theta}(\lambda))$ in (19) is based on the recentered moment conditions (i.e., $E[Z_i(Y_i - X_i'\beta_0) - F\tau_0] = 0$) that accommodate the invalid moment conditions, whereas Andrews (1999) and Andrews and Lu (2001) do not consider such recentering. We choose the tuning parameters by minimizing IC_λ:

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda} IC_\lambda \tag{20}$$

for some finite choice set Λ. Note that the term $\ln[\ln(p+s)]$ is to handle the case of a diverging number of parameters as in the current setup.

In this case, the key question is, even with the tuning parameters $\hat{\lambda}$ selected from (20), whether or not the adaptive elastic net GMM estimation still yields selection consistency under the case of a diverging number of parameters: $P(\mathcal{S}_{\hat{\lambda}} = \mathcal{A}) \to 1$ as $n \to \infty$. In order to verify this result, we need the following condition. We let a $q \times 1$ vector $\tilde{e}_i = V^{-1/2} e_i$ and $\tilde{e}_{ik}$ be the kth element of $\tilde{e}_i$, where e_i is given in (5).

Assumption 4. *For any $\varsigma > 0$, there exists a positive constant $c_0 < \infty$ not depending on n such that*

$$P\left(\max_{1\le j\le m}\left|\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{c_{jk}}{\sqrt{n}}\tilde{e}_{ik}\right| > \varsigma\right) \le \frac{c_0(\log m)^{1/2}}{\varsigma},$$

where c_{jk} are some constants satisfying $\max_{1\le j\le m}\sum_{k=1}^{q} c_{jk}^2 \le \bar{c} < \infty$, and m is some integer with $m \le q$ but $m \to \infty$ as $n \to \infty$.

This assumption regulates the tail behavior of $\tilde{e}_i$, but it does not restrict the distribution family or the dependence structure in $\tilde{e}_{ik}$ over k. Note that we can obtain the same result as Lemma 1-(ii)-(a) of Huang, Ma, and Zhang (2007) if we let $\tilde{e}_{ik}$ be uncorrelated over i and k with $P(|\tilde{e}_{ik}| > \psi_0) \le \psi_1 \exp(-\psi_2 \psi_0^d)$ for some positive constants ψ_0, ψ_1, ψ_2, and $1 \le d \le 2$, which can be easily satisfied when $\tilde{e}_i$ is independent over i and uncorrelated multivariate normal for each i. (In this sub-Gaussian case, $d = 2$.)

The following theorem obtains the selection consistency of the adaptive elastic net GMM estimation even when the tuning parameters are selected by (20). We do not suppose that $\tilde{e}_{ik}$ is Gaussian here. So this result extends that of Wang, Li, and Leng (2009), which assumes the Gaussianity of the linear regression error terms. Recall that $\eta = \min_{j\in\mathcal{A}} |\theta_{j,0}|$.

Theorem 4. *If $n\eta^2/((p+s)\ln(n)\ln[\ln(p+s)]) \to \infty$, then under Assumptions 1, 2 and 4, we have $P(\mathcal{S}_{\hat{\lambda}} = \mathcal{A}) \to 1$ as $n \to \infty$, where $\hat{\lambda}$ is given in (20).*

4.2 Optimization algorithm

Though we can run the adaptive elastic net GMM as (3) directly, we can consider more efficient algorithm for computation. More precisely, we discuss how to apply the LARS algorithm of Efron, Hastie, Johnstone and Tibshirani (2004) to the case of the adaptive elastic net GMM estimation. Basically, we first reformulate the adaptive elastic net GMM problem into a lasso problem by extending Lemma 1 of Zou and Hastie (2005) using the Algorithm 1 of Zou (2006, Section 3.5). Then we apply the LARS algorithm on this transformed problem, in which we also consider tuning parameter selection as described above. Note that the tuning parameter choice preserves selection consistency of the parameter estimates.

To describe the algorithm, we define the *naive* adaptive elastic net GMM estimator as

$$\hat{\theta}_{naive} = \arg \min_{\theta=(\theta_1,\cdots,\theta_{p+s})} \left\{ \left\| \hat{W}^{1/2} Y_z - \sum_{j=1}^{p+s} \hat{W}^{1/2} X_{j,zF} \theta_j \right\|^2 + \lambda_1^* \sum_{j=1}^{p+s} \hat{\pi}_j |\theta_j| + \lambda_2 \sum_{j=1}^{p+s} \theta_j^2 \right\},$$

whose objective function is simply an unscaled version of (3). $\hat{W}$ is obtained as the conventional efficient GMM procedure as described in the previous section. We denote $X_{j,zF}$ as the jth column of X_{zF}. Given the adaptive weights $\hat{\pi}_j$, we further define $(q + (p + s)) \times 1$ vectors given by

$$\hat{X}_{j,zF}^{\hat{W}} = \hat{\pi}_j^{-1} \begin{pmatrix} \hat{W}^{1/2} X_{j,zF} \\ d_j \sqrt{\lambda_2} \end{pmatrix} \quad \text{and} \quad \hat{Y}_z^{\hat{W}} = \begin{pmatrix} \hat{W}^{1/2} Y_z \\ 0_{p+s,1} \end{pmatrix},$$

where d_j is the jth column of I_{p+s}. Then we formulate a lasso problem as

$$\hat{\varphi}_{lasso} = \arg \min_{\varphi=(\varphi_1,\cdots,\varphi_{p+s})} \left\{ \left\| \hat{Y}_z^{\hat{W}} - \sum_{j=1}^{p+s} \hat{X}_{j,zF}^{\hat{W}} \varphi_j \right\|^2 + \lambda_1^* \sum_{j=1}^{p+s} |\varphi_j| \right\},$$

and we can apply the LARS algorithm on this lasso problem. See Efron, Hastie, Johnstone and Tibshirani (2004) for further details. Note that for each given $\lambda = (\lambda_1^*, \lambda_2)$, we calculate $\hat{\varphi}_{lasso}$ as well as IC_λ in (19). The choice of λ is determined at the smallest level of IC_λ as we discussed in the previous subsection and the adaptive elastic net GMM estimator can be obtained as

$$\hat{\theta}_j = \left(1 + \frac{\lambda_2}{n^2}\right) \hat{\theta}_{j,naive} = \left(1 + \frac{\lambda_2}{n^2}\right) \hat{\pi}_j^{-1} \hat{\varphi}_{j,lasso}$$

for all $j = 1, \cdots, p + s$ at given λ.

5 Monte Carlo Simulation

In this section, we study the finite sample performance of our estimator. We let ι_a denote a $a \times 1$ vector of ones and define the set of valid instruments as

$$Z_{1i} = \begin{bmatrix} Z_{1i}^0 \\ Z_{1i}^1 \end{bmatrix} \sim i.i.d.\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega_0 & 0 \\ 0 & \Omega_1 \end{bmatrix} \right)$$

for $i = 1, 2, \cdots, n$, where $\dim(Z_{1i}^0) = 2(q - s_0)/3$, $\dim(Z_{1i}^1) = (q - s_0)/3$, Ω_0 is $\dim(Z_{1i}^0)\times\dim(Z_{1i}^0)$, and Ω_1 is $\dim(Z_{1i}^1)\times\dim(Z_{1i}^1)$. The (i, j)-th element of Ω_0 is set equal to $0.5^{|i-j|}$ and the (i, j)-th element of Ω_1 is set equal to $\rho_z^{|i-j|}$. We consider $\rho_z \in \{0.50,\ 0.95,\ 0.99\}$ to allow for the case where some of the instruments are highly correlated (when $\rho_z = 0.95$ in particular). We also define

$$X_i = \begin{bmatrix} X_{0i} \\ X_{1i} \end{bmatrix} = \begin{bmatrix} Z_{1i}^{0\prime}\pi_0 + v_{0i} \\ Z_{1i}^{1\prime}\pi_1 + v_{1i} \end{bmatrix},$$

where $\pi_0 = 2^{-1/2}(\iota_2 \otimes I_{\dim(Z_{1i}^0)})$ and $\pi_1 = (2 + 2\rho_z^{\dim(Z_{1i}^1)})^{-1/2}(\iota_2 \otimes I_{\dim(Z_{1i}^1)})$. In this setup, all the instruments Z_{1i} are valid, and $p = \dim(X_i) = (q - s_0)/2$. Note that we choose (q, s_0) such that $q - s_0$ is a multiple of 6 to ensure all the dimensions are integers. For the random components, we let

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

and define

$$\begin{aligned} u_i &= \sqrt{\rho_{uv}}\varepsilon_{1i} + \sqrt{1 - \rho_{uv}}\varepsilon_{2i} \\ v_i &= [v_{0i}', v_{1i}']' = \sqrt{\rho_{uv}}\varepsilon_{1i} \cdot \iota_p + \sqrt{1 - \rho_{uv}}\varepsilon_{3i} \end{aligned}$$

with $\rho_{uv} \in \{0.5,\ 0.95\}$. Similarly as Cheng and Liao (2013) and Liao (2013), the $s_0 \times 1$ vector of invalid instruments are generated from

$$Z_{2i} = \varepsilon_{4i} + \tau_{\mathcal{A}} u_i \cdot \iota_{s_0},$$

where we set $\tau_{\mathcal{A}} \in \{0.3,\ 0.6,\ 0.9\}$ that controls the severity of the invalid moment conditions. Note that $\Sigma_{zxF} = [\Sigma_{zx}, F]$ has full column rank even if Z_{2i} is uncorrelated with X_i. Y_i is defined as $Y_i = X_i'\beta_0 + u_i$. For the simulation, we consider the sample size $n = 250, 1000$ and $(p, p_0, s, s_0, q) = (18, 3, 15, 6, 42)$, which implies that $\dim(Z_{1i}^0) = 24$, $\dim(Z_{1i}^1) = 12$, $\dim(Z_{2i}) = 6$, $\dim(X_{0i}) = 12$, and $\dim(X_{1i}) = 6$. We also set

$$\beta_0 = [C, C, 0_{1\times 10}, C, 0_{1\times 5}]', \tag{21}$$

where the nonzero structural parameter C is set as $C \in \{0.25,\ 0.5,\ 0.75\}$. In order to guarantee the identification of β_0, we suppose that we know the first 18 elements of Z_{1i}^0 and the first 9 elements of Z_{1i}^1 are the $(q-s)$-number of valid instruments; we call them "surely-valid" instruments whereas we call the rest of the instruments in Z_{1i} as "unsurely-valid" instruments. We conduct 2000 replications.

We summarize the simulation results in Tables 1 to 4. AENet is the estimator proposed in (3) and is solved by the algorithm in Section 4. We set $\gamma = 2$ and use the conventional two-step GMM estimator to obtain the adaptive weight $\hat{\pi}_j$ and the optimal weighting matrix $\hat{W}$. The optimal tuning parameters for λ_1^* and λ_2 are selected by (19) from the grid search over $\{0.01n, 0.025n, 0.05n, 0.075n, (0.1 : 0.05 : 1)n\}$ and $\{0.01n, 0.05n, (0.1 : 0.1 : 2)n, 2.5n, 3n, 4n, 5n\}$, respectively, where the grids $(0.1 : 0.05 : 1)n$ means it starts with $0.1n$ and increases by $0.05n$ until $1n$. 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 the LARS algorithm. ALasso-CL is the adaptive lasso estimator proposed by Cheng and Liao (2013). 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 valid instruments in (2). ALasso-CL is solved by the algorithm proposed by Schmidt (2010) and the tuning parameter is selected following Cheng and Liao's (2013) suggestion. We 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 $\tau_0 = (\tau'_{\mathcal{A}}, \tau'_{\mathcal{A}^c})'$, where $\beta_{\mathcal{A}}$ and $\tau_{\mathcal{A}}$ collect the nonzero elements in β_0 and τ_0, respectively.

Table 1 reports the root mean squared errors (RMSE) of the three estimators, AENet, ALasso-LARS, and ALasso-CL, when $\rho_{uv} = 0.5$ and all the valid instruments are relevant. For each case, the RMSEs of estimators of $\tau_{\mathcal{A}^c}$, $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$ are denoted by $rmse_1$, $rmse_2$, $rmse_3$ and $rmse_4$, respectively. First, for each n, the RMSEs of $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$ increase when ρ_z increases from 0.5 to 0.95. This is intuitive because the instruments in Z_{1i}^1 are highly correlated when $\rho_z = 0.95$ and thus provide less information about X_{1i} than the case of $\rho_z = 0.5$. Second, the RMSEs of the nonzero structural parameter $\beta_{\mathcal{A}}$ are very similar for all three estimators when $\rho_z = 0.5$. However, both AENet and ALasso-LARS have smaller RMSEs for $\beta_{\mathcal{A}}$ than ALasso-CL when $\rho_z = 0.95$. Since ALasso-CL does not conduct variable selection, it is not surprising that its estimator for $\beta_{\mathcal{A}}$ is less efficient especially when $\rho_z = 0.95$. To make the comparison fair, we also look at the moment selection performance in Table 2 and consider the case with redundant instruments in Table 4.

Table 2 reports the accuracy of moment/instrument selection by different estimation procedures under the same condition as Table 1. Pr_1 stands for the percentage of replications that yield zero estimates for $\tau_{\mathcal{A}^c}$; Pr_2 for the percentage yielding nonzero estimates for $\tau_{\mathcal{A}}$. First, for the unsurely-valid instruments, all selection procedures perform well in the sense that Pr_1 is close to one. Second, for invalid instruments, it is not surprising that Pr_2 increases with $\tau_{\mathcal{A}}$ because detecting invalid instruments gets easier as $\tau_{\mathcal{A}}$ gets larger. Third, our estimators outperform ALasso-CL in instrument selection when $n = 250$ or $\rho_z = 0.95$. When $(\tau_{\mathcal{A}}, C, \rho_z) = (0.3, 0.25, 0.95)$ and $n = 250$, for example, AENet detects 75.8% of the invalid instruments, but ALasso-CL only detects 1.7%. As n increases to 1000, our estimators can pick up the invalid instruments all the time while ALasso-CL performs well only for $\rho_z = 0.5$. Recall that the differences between ALasso-LARS and ALasso-CL are the algorithm and tuning parameter selection method. It seems that ALasso-CL tends to select relatively large tuning parameters for τ, which results in over-selection of valid moment conditions.

In order to manifest the difference between AENet and ALasso-LARS further, Table 3 summarizes the simulation results when the instruments Z_i are highly correlated with each other (and so are the regressors X_i). In particular, we set $\rho_z = 0.99$ and $\rho_{uv} = 0.95$, under which we can also calculate the population correlation between any two variables in X_{1i} lies between 0.963 and 0.972 with our data generating process. In general, AENet yields smaller RMSEs for $\beta_{\mathcal{A}^c}$ and $\beta_{\mathcal{A}}$ than ALasso-LARS. The difference between AENet and ALasso-LARS vanishes as n increases from 250 to 1000. These results show that adding a ridge penalty is helpful to reduce the RMSEs of the estimated structural parameters when some instruments are highly correlated or the structural regression suffers near-multicollinearity.

As for the last experiment, we investigate the performance of our estimators when some instruments are redundant. An additional instrument is considered redundant for a given set of instruments if it does not improve the efficiency of the GMM estimator based on the given set of instruments. Following Cheng and Liao's (2013) simulation design, we use irrelevant instruments to play the role of redundant ones. From the simulation design above, we replace 6 of the unsurely-valid instruments in Z_{1i} with 6 independent standard normal random variables. The simulation results are reported in Table 4, in which Pr_5 denotes the percentage of replications that yield nonzero estimates for τ for the redundant instruments. Since ALasso-CL is designed to detect the redundant instruments, its Pr_5 increases with n. Though not reported in the table to save space, extra simulations show that Pr_5 of ALasso-CL is equal to 60.8% for $n = 2500$. Also, our estimators cannot detect redundant instruments by design. Pr_5 of AENet and ALasso-LARS is close to zero regardless of the sample size. However, our estimators have very similar RMSEs to ALasso-CL for nonzero β_0 and smaller RMSEs for zero β_0. Hence, our estimation procedure is robust to presence of a moderate amount of redundant instruments.

6 Conclusion

This paper develops an adaptive elastic net GMM estimator with possibly many 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 written 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 efficient GMM variance as if we had known all the valid instruments. 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 have good finite-sample performance.

Appendix

Proof of Lemma 1 We first define a ridge estimator given by

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

We obtain the desired results using

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

where $\hat{\theta}_\pi$ is defined in (11) and by deriving the bounds for $E\|\hat{\theta}_\pi - \hat{\theta}_R\|^2$ and $E\|\hat{\theta}_R - \theta_0\|^2$.

For the first term $E\|\hat{\theta}_\pi - \hat{\theta}_R\|^2$ in (A.2), we note that

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

which is from the definition of $\hat{\theta}_\pi$ and $\hat{\theta}_R$. But we can rearrange (A.3) as

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

where

$$\sum_{j=1}^{p+s} \hat{\pi}_j \left\{ |\hat{\theta}_{j,R}| - |\hat{\theta}_{j,\pi}| \right\} \leq \sum_{j=1}^{p+s} \hat{\pi}_j |\hat{\theta}_{j,R} - \hat{\theta}_{j,\pi}| \leq \left(\sum_{j=1}^{p+s} \hat{\pi}_j^2 \right)^{1/2} \|\hat{\theta}_\pi - \hat{\theta}_R\|. \tag{A.5}$$

Moreover, 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

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

since from (A.6)

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

Similarly, we also have

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

from (A.6). By subtracting (A.7) from (A.8), the right-hand-side of (A.4) satisfies

$$
\begin{aligned}
& \left\{(Y_z - X_{zF}\hat{\theta}_\pi)'\hat{W}(Y_z - X_{zF}\hat{\theta}_\pi) + \lambda_2\|\hat{\theta}_\pi\|^2\right\} \\
& -\left\{(Y_z - X_{zF}\hat{\theta}_R)'\hat{W}(Y_z - X_{zF}\hat{\theta}_R) + \lambda_2\|\hat{\theta}_R\|^2\right\} \\
= \;& (\hat{\theta}_\pi - \hat{\theta}_R)'[X_{zF}'\hat{W}X_{zF} + \lambda_2 I_{p+s}](\hat{\theta}_\pi - \hat{\theta}_R) \\
\geq \;& [\text{Eigmin}(X_{zF}'\hat{W}X_{zF}) + \lambda_2]\|\hat{\theta}_\pi - \hat{\theta}_R\|^2 > 0
\end{aligned} \tag{A.9}
$$

as $\hat{W}$ is positive definite. Therefore, using (A.4), (A.5), and (A.9), we have

$$
\|\hat{\theta}_\pi - \hat{\theta}_R\| \leq \frac{\lambda_1(\sum_{j=1}^{p+s}\hat{\pi}_j^2)^{1/2}}{\text{Eigmin}(X_{zF}'\hat{W}X_{zF}) + \lambda_2} \leq \frac{\lambda_1(\sum_{j=1}^{p+s}\hat{\pi}_j^2)^{1/2}}{bn^2 + \lambda_2} \tag{A.10}
$$

w.p.a.1 or

$$
E\|\hat{\theta}_\pi - \hat{\theta}_R\|^2 \leq \frac{\lambda_1^2 E(\sum_{j=1}^{p+s}\hat{\pi}_j^2)}{(bn^2 + \lambda_2)^2} \tag{A.11}
$$

w.p.a.1 as $\text{Eigmin}(X_{zF}'\hat{W}X_{zF}) + \lambda_2 > bn^2 + \lambda_2 > 0$ from (8).

For the second term $E||\hat{\theta}_R - \theta_0||^2$ in (A.2), we note that from (1)

$$
Y_z = Z'X\beta_0 + nF\tau_0 + (Z'u - nF\tau_0) = X_{zF}\theta_0 + e,
$$

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

$$
\begin{aligned}
\hat{\theta}_R &= [(X_{zF}'\hat{W}X_{zF}) + \lambda_2 I_{p+s}]^{-1}[X_{zF}'\hat{W}(X_{zF}\theta_0 + e) + \lambda_2\theta_0 - \lambda_2\theta_0] \\
&= \theta_0 + [X_{zF}'\hat{W}X_{zF} + \lambda_2 I_{p+s}]^{-1}[X_{zF}'\hat{W}e] - \lambda_2[X_{zF}'\hat{W}X_{zF} + \lambda_2 I_{p+s}]^{-1}\theta_0
\end{aligned} \tag{A.12}
$$

yielding

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

w.p.a.1 similarly as above. However,

$$
E\|X_{zF}'\hat{W}e\|^2 \leq \text{Eigmax}(\hat{W}X_{zF}X_{zF}'\hat{W})E\|e\|^2,
$$

where $\text{Eigmax}(\hat{W}X_{zF}X_{zF}'\hat{W}) \leq n^2 B$ w.p.a.1 from (9) and

$$
E\|e\|^2 = E\left[\sum_{i=1}^n\sum_{j=1}^n e_i'e_j\right] \leq n \times \max_{1\leq i\leq n} E\left[e_i'e_i\right] = nqc
$$

for some positive constant $c < \infty$ since $e = \sum_{i=1}^n e_i = \sum_{i=1}^n (Z_i u_i - F\tau_0)$ from (5) and the $q \times 1$ vector e_i is independent over i with the bounded second moment from Assumption 1-(ii). It follows that

$$
E\|\hat{\theta}_R - \theta_0\|^2 \leq \frac{2\{\lambda_2^2\|\theta_0\|^2 + E\|X_{zF}'\hat{W}e\|^2\}}{(bn^2 + \lambda_2)^2} \leq \frac{2\{\lambda_2^2\|\theta_0\|^2 + Bn^3 q\}}{(bn^2 + \lambda_2)^2} \tag{A.13}
$$

w.p.a.1, in which B is redefined as cB without loss of generality.

Therefore, by combining the results in (A.2), (A.11), and (A.13), we obtain the desired result as

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

w.p.a.1. The result for $\hat{\theta}_{enet}$ readily follows by letting $\hat{\pi}_j=1$ for all j. **Q.E.D.**

We now provide a useful lemma before we proof Theorem 1. Note that the difference between this lemma and Theorem 1 is that we can tell the local-to-zero coefficients as nonzero from Theorem 1, provided they vanish at the rate slower than certain threshold.

Lemma A.1. *We define*

$$\tilde{\theta}_{\mathcal{A}} = \arg\min_{\theta}\left\{(Y_z - X_{zF:\mathcal{A}}\theta)'\hat{W}(Y_z - X_{zF:\mathcal{A}}\theta)+\lambda_1^*\sum_{j\in\mathcal{A}}\hat{\pi}_j|\theta_j|+\lambda_2\sum_{j\in\mathcal{A}}\theta_j^2\right\}, \tag{A.14}$$

where $X_{zF,\mathcal{A}}$ consists of the sub-columns of X_{zF} that correspond to nonzero elements in θ_0. Under Assumptions 1 and 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).

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

$$P\left\{\text{For all } j\in\mathcal{A}^c,\ |-2X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|\leq\lambda_1^*\hat{\pi}_j\right\}\to 1$$

or equivalently

$$\Psi_n\equiv P\left\{\text{There exist } j\in\mathcal{A}^c \text{ such that } |-2X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|>\lambda_1^*\hat{\pi}_j\right\}\to 0 \tag{A.15}$$

as $n\to\infty$, where $X_{j,zF}$ is the jth column of X_{zF}. We let $\eta=\min_{j\in\mathcal{A}}|\theta_{j,0}|$ and $\hat{\eta}=\min_{j\in\mathcal{A}}|\hat{\theta}_{j,enet}|$. Similarly as Zou and Zhang (2009), we have

$$\Psi_n\leq\sum_{j\in\mathcal{A}^c}P\left\{|-2X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|>\lambda_1^*\hat{\pi}_j \text{ and } \hat{\eta}>\eta/2\right\}+P\left\{\hat{\eta}\leq\eta/2\right\}, \tag{A.16}$$

where the second term is bounded by

$$\begin{aligned} P\{\hat{\eta}\leq\eta/2\} &\leq P\left\{\|\hat{\theta}_{enet}-\theta_0\|>\eta/2\right\} \\ &\leq \frac{E\|\hat{\theta}_{enet}-\theta_0\|^2}{\eta^2/4}\leq 16\frac{\lambda_2^2\|\theta_0\|_2^2+Bqn^3+\lambda_1^2(p+s)}{(bn^2+\lambda_2)^2\eta^2} \end{aligned} \tag{A.17}$$

w.p.a.1 from Lemma 1. For the first term in (A.16), letting $M = \left(\lambda_1^{*2}/n^{\kappa}\right)^{1/2\gamma}$, we have

$$\begin{aligned}
& \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^*\hat{\pi}_j \text{ and } \hat{\eta} > \eta/2\right\} \\
\leq \; & \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^*\hat{\pi}_j,\, \hat{\eta} > \eta/2,\, |\hat{\theta}_{j,enet}| \leq M\right\} + \sum_{j\in\mathcal{A}^c} P\left\{|\hat{\theta}_{j,enet}| > M\right\} \\
\leq \; & \sum_{j\in\mathcal{A}^c} P\left\{|-2X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})| > \lambda_1^* M^{-\gamma} \text{ and } \hat{\eta} > \eta/2\right\} + \sum_{j\in\mathcal{A}^c} P\left\{|\hat{\theta}_{j,enet}| > M\right\} \\
\leq \; & \frac{4M^{2\gamma}}{\lambda_1^{*2}} E\left[\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] + \frac{1}{M^2} E\left[\sum_{j\in\mathcal{A}^c} |\hat{\theta}_{j,enet}|^2\right] \\
\leq \; & \frac{4M^{2\gamma}}{\lambda_1^{*2}} E\left[\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] + \frac{E\|\hat{\theta}_{enet} - \theta_0\|^2}{M^2} \\
\leq \; & \frac{4M^{2\gamma}}{\lambda_1^{*2}} E\left[\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] + 4\frac{\lambda_2^2\|\theta_0\|_2^2 + Bn^3 q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2 M^2} \qquad \text{(A.18)}
\end{aligned}$$

w.p.a.1 from Lemma 1, where $1_{\{\cdot\}}$ is the binary indicator.[8] The last expression in (A.18) can be further bounded as

$$E\left[\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 1_{\{\hat{\eta}>\eta/2\}}\right] \leq 2B^2 n^4 E(\|\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}}\|_2^2 1_{\{\hat{\eta}>\eta/2\}}) + 2Bn^3 q \qquad \text{(A.19)}$$

w.p.a.1 since

$$\begin{aligned}
\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(Y_z - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 &= \sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(X_{zF:\mathcal{A}}\theta_{\mathcal{A}} - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}}) + X'_{j,zF}\hat{W}e|^2 \\
&\leq 2\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}(X_{zF:\mathcal{A}}\theta_{\mathcal{A}} - X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})|^2 + 2\sum_{j\in\mathcal{A}^c} |X'_{j,zF}\hat{W}e|^2 \\
&\leq 2Bn^2\|\hat{W}^{1/2}X_{zF:\mathcal{A}}(\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}})\|^2 + 2Bn^2\|e\|^2 \\
&\leq 2Bn^2 \times Bn^2\|\theta_{\mathcal{A}} - \tilde{\theta}_{\mathcal{A}}\|^2 + 2Bn^3 q
\end{aligned}$$

w.p.a.1 from (8) and (9). Note that $\text{Eigmax}(n^{-2}X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}) \leq \text{Eigmax}(n^{-2}X'_{zF}\hat{W}X_{zF}) \leq B$. Furthermore, by defining

$$\tilde{\theta}_{R\mathcal{A}} = \arg\max_{\theta}\left\{(Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta) + \lambda_2\sum_{j\in\mathcal{A}}\theta_j^2\right\}, \qquad \text{(A.20)}$$

we have

$$\|\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{R\mathcal{A}}\| \leq \frac{\lambda_1^*\hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2} \qquad \text{(A.21)}$$

[8] Though the proof steps are similar to Zou and Zhang (2009, p.1746), note the difference in M; finding a new M for the GMM setup is not trivial.

w.p.a.1 similarly as (A.10) since $\max_{j\in\mathcal{A}}\hat{\pi}_j^2 = (\min_{j\in\mathcal{A}}|\hat{\theta}_{j,enet}|)^{-2\gamma} \leq \hat{\eta}^{-2\gamma}$, $p_0+s_0 \leq p+s$, and $\mathrm{Eigmin}(n^{-2}X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}) \geq \mathrm{Eigmin}(n^{-2}X'_{zF}\hat{W}X_{zF}) \geq b$. Similarly as the proof of Lemma 1 above, we thus have

$$E(\|\theta_{\mathcal{A}}-\tilde{\theta}_{\mathcal{A}}\|^2 1_{\{\hat{\eta}>\eta/2\}}) \leq 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^{*2}(\eta/2)^{-2\gamma}(p+s)}{(bn^2+\lambda_2)^2} \tag{A.22}$$

w.p.a.1. Therefore, by combining the results in (A.17), (A.18), (A.19) and (A.22), we can find the upper bound of Ψ_n in (A.16) as

$$\begin{aligned}\Psi_n &\leq \frac{4M^{2\gamma}}{\lambda_1^{*2}}\left\{2B^2n^4 \times 4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^{*2}(\eta/2)^{-2\gamma}(p+s)}{(bn^2+\lambda_2)^2} + 2Bn^3q\right\} \\ &\quad +4\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2M^2} \\ &\quad +16\frac{\lambda_2^2\|\theta_0\|^2 + Bn^3q + \lambda_1^2(p+s)}{(bn^2+\lambda_2)^2\eta^2} \\ &\equiv \bar{\Psi}_{1,n} + \bar{\Psi}_{2,n} + \bar{\Psi}_{3,n}\end{aligned}$$

w.p.a.1, in which each term of the bound goes to zero as $n\to\infty$ yielding the desired result in (A.15).

More precisely, note that

$$\bar{\Psi}_{1,n} = O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}\lambda_2^2\|\theta_0\|^2\right) + O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}n^3q\right) + O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}\frac{(\lambda_1^*)^2(p+s)}{\eta^{2\gamma}}\right) + O_p\left(\frac{M^{2\gamma}}{\lambda_1^{*2}}n^3q\right) = o_p(1),$$

where the second (and the last) term satisfies

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

by Assumption 2-(ii) as $M = \left(\lambda_1^{*2}/n^\kappa\right)^{1/2\gamma}$ and $q = O(n^\alpha)$ from Assumption 2-(i); the first term is dominated by the second term since $\lambda_2^2\|\theta_0\|^2 \leq \lambda_2^2(p+s) \leq (\lambda_2^2/n^3)(n^3q)$ and λ_2^2/n^3 is bounded by Assumption 2-(iii); and the third term satisfies

$$\frac{M^{2\gamma}}{(\lambda_1^*)^2}(\lambda_1^*)^2\frac{(p+s)}{\eta^{2\gamma}} = \frac{\lambda_1^{*2}}{n^\kappa}\frac{(p+s)}{\eta^{2\gamma}} \to 0 \tag{A.23}$$

by Assumption 2-(iv) for $\kappa > 3$. Similarly,

$$\bar{\Psi}_{2,n} = O_p\left(\frac{\lambda_2^2}{n^3}\frac{\|\theta_0\|^2}{n}\frac{1}{M^2}\right) + O_p\left(\frac{q}{n}\frac{1}{M^2}\right) + O_p\left(\frac{\lambda_1^2}{n^3}\frac{(p+s)}{n}\frac{1}{M^2}\right) = o_p(1),$$

where the second term is dominating as $\lambda_1^2/n^3 \to 0$ and $\lambda_2^2/n^3 \to 0$ by Assumption 2-(i) and $\|\theta_0\|^2/n \leq (p+s)/n \leq (q/n)$; and the second term satisfies

$$O_p\left(\frac{q}{nM^2}\right) = O_p\left(\frac{n^{\alpha-1}}{M^2}\right) = O_p\left(\left[\frac{n^{\kappa-\gamma(1-\alpha)}}{(\lambda_1^*)^2}\right]^{1/\gamma}\right) \to 0$$

by Assumption 2-(iv) as $\gamma > 0$. Finally, $\bar{\Psi}_{3,n} = o_p(1)$ can be shown similarly as $\bar{\Psi}_{2,n}$ since

$$\bar{\Psi}_{3,n} = O_p\left(\frac{\lambda_2^2}{n^3}\frac{\|\theta_0\|^2}{n}\frac{1}{\eta^2}\right) + O_p\left(\frac{q}{n}\frac{1}{\eta^2}\right) + O_p\left(\frac{\lambda_1^2}{n^3}\frac{(p+s)}{n}\frac{1}{\eta^2}\right)$$

and

$$O_p\left(\frac{q}{n\eta^2}\right) = O_p\left(\frac{1}{n^{1-\alpha}\eta^2}\right) = O_p\left(\left[\frac{n^{\kappa-\gamma(1-\alpha)}}{(\lambda_1^*)^2} \times \frac{\lambda_1^{*2}(p+s)}{n^3\eta^{2\gamma}} \times \frac{1}{(p+s)n^{\kappa-3}}\right]^{1/\gamma}\right) \to 0 \quad \text{(A.24)}$$

by Assumption 2-(iv) as $\gamma > 0$ and $\kappa > 3$. **Q.E.D.**

Proof of Theorem 1 Using Lemma A.1, we only need to show that the minimal element of the estimator of nonzero coefficients is larger than zero w.p.a.1: $P\{\min_{j\in\mathcal{A}} |\tilde{\theta}_{j,\mathcal{A}}| > 0\} \to 1$, where $\tilde{\theta}_{j,\mathcal{A}}$ is the jth element of $\tilde{\theta}_{\mathcal{A}}$ in (A.14). Note that by (A.21)

$$\min_{j\in\mathcal{A}} |\tilde{\theta}_{j,\mathcal{A}}| > \min_{j\in\mathcal{A}} |\tilde{\theta}_{j,R\mathcal{A}}| - \frac{\lambda_1^*\hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2}, \quad \text{(A.25)}$$

where $\tilde{\theta}_{j,R\mathcal{A}}$ is the jth element of $\tilde{\theta}_{R\mathcal{A}}$ in (A.20) and

$$\min_{j\in\mathcal{A}} |\tilde{\theta}_{j,R\mathcal{A}}| > \min_{j\in\mathcal{A}} |\theta_{j,\mathcal{A}}| - \|\tilde{\theta}_{R\mathcal{A}} - \theta_{\mathcal{A}}\|. \quad \text{(A.26)}$$

But from (A.13), it holds that

$$\begin{aligned} E(\|\tilde{\theta}_{R\mathcal{A}} - \theta_{\mathcal{A}}\|^2) &\leq 2\left[\frac{\lambda_2^2\|\theta_0\|_2^2 + qn^3B}{(bn^2+\lambda_2)^2}\right] \\ &= O\left(\frac{\lambda_2^2(p+s)}{n^4}\right) + O\left(\frac{qn^3}{n^4}\right) = O\left(\frac{q}{n}\right) \end{aligned} \quad \text{(A.27)}$$

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

$$\frac{\lambda_1^*\hat{\eta}^{-\gamma}\sqrt{p+s}}{bn^2+\lambda_2} = O\left(\frac{\lambda_1^*\sqrt{p+s}}{n^2\eta^\gamma} \times \left(\frac{\hat{\eta}}{\eta}\right)^{-\gamma}\right) = o\left(\frac{1}{\sqrt{n}}\right) O_p(1), \quad \text{(A.28)}$$

where

$$\frac{\lambda_1^*}{n^2}\frac{\sqrt{p+s}}{\eta^\gamma} = \frac{1}{\sqrt{n}} \times \frac{\lambda_1^*\sqrt{p+s}}{n^{3/2}\eta^\gamma} = o\left(\frac{1}{\sqrt{n}}\right) \quad \text{(A.29)}$$

by Assumption 2-(iv) and

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

as we showed $\bar{\Psi}_{3,n} = o_p(1)$ in the proof of Lemma A.1 above that gives

$$(\hat{\eta}/\eta)^{-\gamma} = [(\hat{\eta}/\eta)^2]^{-\gamma/2} = O_p(1). \quad \text{(A.30)}$$

Therefore, we have

$$\min_{j\in\mathcal{A}} |\tilde{\theta}_{j,R\mathcal{A}}| > \min_{j\in\mathcal{A}} |\theta_{j,\mathcal{A}}| - \sqrt{\frac{q}{n}} O_p(1) - o\left(\frac{1}{\sqrt{n}}\right) O_p(1)$$

and we have the desired result since $\sqrt{1/n} \leq \sqrt{q/n}$ converges to zero faster than η by (A.24). **Q.E.D.**

Proof of Theorem 2 Using $\tilde{\theta}_{\mathcal{A}}$ in (A.14), we note that

$$\begin{aligned}
& \zeta'(I_{p_0+s_0} + \lambda_2 \hat{H}_{\mathcal{A}}^{-1}) \hat{H}_{\mathcal{A}}^{1/2} n^{-1/2} \left(\tilde{\theta}_{\mathcal{A}} - \frac{\theta_{\mathcal{A}}}{1+\lambda_2/n^2} \right) \\
= \; & \zeta'(I_{p_0+s_0} + \lambda_2 \hat{H}_{\mathcal{A}}^{-1}) \hat{H}_{\mathcal{A}}^{1/2} n^{-1/2} (\tilde{\theta}_{\mathcal{A}} - \tilde{\theta}_{R\mathcal{A}}) \\
& + \zeta'(I_{p_0+s_0} + \lambda_2 \hat{H}_{\mathcal{A}}^{-1}) \hat{H}_{\mathcal{A}}^{1/2} n^{-1/2} (\tilde{\theta}_{R\mathcal{A}} - \theta_{\mathcal{A}}) \\
& + \zeta'(I_{p_0+s_0} + \lambda_2 \hat{H}_{\mathcal{A}}^{-1}) \hat{H}_{\mathcal{A}}^{1/2} n^{-1/2} \left(\frac{\lambda_2 \theta_{\mathcal{A}}}{n^2 + \lambda_2} \right),
\end{aligned} \tag{A.31}$$

where $\tilde{\theta}_{R\mathcal{A}}$ is given in (A.20). Since $\tilde{\theta}_{R\mathcal{A}} - \theta_{\mathcal{A}} = (\hat{H}_{\mathcal{A}} + \lambda_2 I_{p_0+s_0})^{-1}(X'_{zF:\mathcal{A}} \hat{W} e_{\mathcal{A}}) - \lambda_2(\hat{H}_{\mathcal{A}} + \lambda_2 I_{p_0+s_0})^{-1} \theta_{\mathcal{A}}$ from (A.12), the second term in (A.31) satisfies

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

where $e = Z'u - nF\tau_0 = Z'u - nF_{\mathcal{A}}\tau_{\mathcal{A}}$. Therefore, by Theorem 1, we can write

$$\begin{aligned}
\Phi_n \; & \equiv \; \zeta' \frac{(I_{p_0+s_0} + \lambda_2 \hat{H}_{\mathcal{A}}^{-1})}{1+\lambda_2/n} \hat{H}_{\mathcal{A}}^{1/2} n^{-1/2} (\hat{\theta}_{\mathcal{A}} - \theta_{\mathcal{A}}) \\
& = \; \Phi_{1,n} + \Phi_{2,n} + \Phi_{3,n}
\end{aligned}$$

w.p.a.1, where

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

We will show that $\Phi_{1,n} = o_p(1)$, $\Phi_{2,n} = o_p(1)$, and $\Phi_{3,n} \xrightarrow{d} \mathcal{N}(0,1)$ to obtain the desired result. First note that w.p.a.1 we have

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

from (A.21), (A.28), (A.29) and (A.30). In a similar way, w.p.a.1, we have

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

from (8) and Assumption 2-(iii) for $\|\theta_{\mathcal{A}}\|^2 \leq \|\theta_0\|^2$. Finally, we prove that $\Phi_{3,n} \xrightarrow{d} \mathcal{N}(0,1)$. Recall that $e_i = Z_i u_i - F\tau_0 = Z_i u_i - F_{\mathcal{A}}\tau_{\mathcal{A}}$ so that $e = \sum_{i=1}^n e_i$, and we can rewrite $\Phi_{3,n} = \sum_{i=1}^n \hat{r}_i$ with $\hat{r}_i = \zeta' \hat{H}_{\mathcal{A}}^{-1/2} X'_{zF:\mathcal{A}} \hat{V}^{-1} n^{-1/2} e_i$ by letting $\hat{W} = \hat{V}^{-1}$ as the optimal weight. We also define $r_i = \zeta' \Sigma_{\mathcal{A}}^{-1/2} \Sigma'_{zxF:\mathcal{A}} V^{-1} n^{-1/2} e_i$. Then, since $\|n^{-2}\hat{H}_{\mathcal{A}} - \Sigma_{\mathcal{A}}\| \xrightarrow{p} 0$, $\|n^{-1}X_{zF:\mathcal{A}} - \Sigma_{zxF:\mathcal{A}}\| \xrightarrow{p} 0$, and $\|\hat{V} - V\| \xrightarrow{p} 0$ with $\text{Eigmax}(V) < \infty$ from Assumption 1-(ii), we have $\Phi_{3,n} = \sum_{i=1}^n r_i + o_p(1)$. Also note that $\sum_{i=1}^n r_i^2 = 1 + o_p(1)$ since

$$
\begin{aligned}
\left|\sum_{i=1}^n r_i^2 - 1\right| &= \Bigg|\zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1}\left(\frac{1}{n}\sum_{i=1}^n e_i e_i'\right)V^{-1}\Sigma_{zxF:\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}\zeta \\
&\quad -\zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1}VV^{-1}\Sigma_{zxF:\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2}\zeta\Bigg| \xrightarrow{p} 0
\end{aligned}
$$

for $\zeta'\zeta = 1$, $\Sigma_{\mathcal{A}} = \Sigma'_{zxF:\mathcal{A}}V^{-1}\Sigma_{zxF:\mathcal{A}}$, and $\|n^{-1}\sum_{i=1}^n e_i e_i' - V\| \xrightarrow{p} 0$. Therefore, for some $\epsilon > 0$, w.p.a.1 we obtain the Lyapunov condition as

$$
\sum_{i=1}^n E|r_i|^{2+\epsilon} \leq E\left(\max_{1\leq i\leq n}|r_i|^\epsilon \sum_{i=1}^n r_i^2\right) \leq \left(E\max_{1\leq i\leq n} r_i^2\right)^{\epsilon/2} \to 0
$$

that gives the desired CLT, where the second inequality is from Jensen's. Note that using Cauchy-Schwartz

$$
E\max_{1\leq i\leq n} r_i^2 \leq \left\|\zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1}\right\|^2 \times \frac{1}{n}E\max_{1\leq i\leq n}\|e_i\|^2,
$$

in which $||\zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1}||^2$ is bounded since

$$
\begin{aligned}
\left\|\zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1}\right\|^2 &\leq \text{Eigmax}(V^{-1})\left\|\zeta'\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1/2}\right\|^2 \\
&\leq \text{Eigmax}(V^{-1})\text{Eigmax}(\Sigma_{\mathcal{A}}^{-1/2}\Sigma'_{zxF:\mathcal{A}}V^{-1}\Sigma_{zxF:\mathcal{A}}\Sigma_{\mathcal{A}}^{-1/2})\left\|\zeta\right\|^2 \\
&= [\text{Eigmin}(V)]^{-1}\text{Eigmax}(I_{p_0+s_0})\left\|\zeta\right\|^2 < \infty
\end{aligned}
$$

from Assumption 1-(ii) but $n^{-1}E\max_{1\leq i\leq n}\|e_i\|^2 \to 0$ from Assumption 3. **Q.E.D.**

Proof of Theorem 3 Conformable to the decomposition of $Z_i = [Z_i^{1\prime}, Z_i^{2\prime}]'$, we decompose the $q\times(p_0+s_0)$ matrix $\Sigma_{zxF:\mathcal{A}}$ as

$$
\Sigma_{zxF:\mathcal{A}} = [\Sigma_{zx:\mathcal{A}}, F_{\mathcal{A}}] = \begin{array}{cc} \left[\begin{array}{cc} \Sigma_{z_1x:\mathcal{A}} & 0_{q-s_0,s_0} \\ \Sigma_{z_2x:\mathcal{A}} & I_{s_0} \end{array}\right] & \begin{array}{c} {\scriptstyle (q-s_0)} \\ {\scriptstyle s_0} \end{array} \\ \begin{array}{cc} {\scriptstyle p_0} & \quad\quad {\scriptstyle s_0} \end{array} & \end{array},
$$

where $\|n^{-1}Z^{1\prime}X_{\mathcal{A}} - \Sigma_{z_1x:\mathcal{A}}\| \xrightarrow{p} 0$ and $\|n^{-1}Z^{2\prime}X_{\mathcal{A}} - \Sigma_{z_2x:\mathcal{A}}\| \xrightarrow{p} 0$. Note that $\Sigma_{z_1x:\mathcal{A}}$ is $(q-s_0)\times p_0$ and $\Sigma_{z_2x:\mathcal{A}}$ is $s_0\times p_0$. Similarly, we let

$$
V = \begin{array}{cc} \left[\begin{array}{cc} V_{11} & V_{12} \\ V'_{12} & V_{22} \end{array}\right] & \begin{array}{c} {\scriptstyle (q-s_0)} \\ {\scriptstyle s_0} \end{array} \\ \begin{array}{cc} {\scriptstyle (q-s_0)} & {\scriptstyle s_0} \end{array} & \end{array}
$$

and

$$
V^{-1} = \begin{array}{cc} \left[\begin{array}{cc} V^{11} & V^{12} \\ \left(V^{12}\right)' & V^{22} \end{array}\right] & \begin{array}{c} {\scriptstyle (q-s_0)} \\ {\scriptstyle s_0} \end{array} \\ \begin{array}{cc} {\scriptstyle (q-s_0)} & {\scriptstyle s_0} \end{array} & \end{array},
$$

where explicit expressions of each term become clear at the end of this proof. Given $\Sigma_{zxF:\mathcal{A}}$ and V^{-1} decompositions above, we can write

$$
\Sigma_{\mathcal{A}} = \Sigma'_{zxF:\mathcal{A}}V^{-1}\Sigma_{zxF:\mathcal{A}} = \begin{array}{cc} \left[\begin{array}{cc} \Sigma_{\mathcal{A}11} & \Sigma_{\mathcal{A}12} \\ \Sigma'_{\mathcal{A}12} & \Sigma_{\mathcal{A}22} \end{array}\right] & \begin{array}{c} {\scriptstyle p_0} \\ {\scriptstyle s_0} \end{array} \\ \begin{array}{cc} {\scriptstyle p_0} & \quad {\scriptstyle s_0} \end{array} & \end{array},
$$

where

$$
\begin{aligned}
\Sigma_{\mathcal{A}11} &= \Sigma'_{z_1x:\mathcal{A}}V^{11}\Sigma_{z_1x:\mathcal{A}} + \Sigma'_{z_1x:\mathcal{A}}V^{12}\Sigma_{z_2x:\mathcal{A}} \\
&\quad +\Sigma'_{z_2x:\mathcal{A}}(V^{12})'\Sigma_{z_1x:\mathcal{A}} + \Sigma'_{z_2x:\mathcal{A}}V^{22}\Sigma_{z_2x:\mathcal{A}}, \\
\Sigma_{\mathcal{A}12} &= \Sigma'_{z_1x:\mathcal{A}}V^{12} + \Sigma'_{z_2x:\mathcal{A}}V^{22}, \\
\Sigma_{\mathcal{A}22} &= V^{22}.
\end{aligned}
\tag{A.32}
$$

We now let $\Sigma_{\mathcal{A}}^{11}$ be the $p_0\times p_0$ north-west block of $\Sigma_{\mathcal{A}}^{-1}$. From the partitioned inverse matrix formula, we thus have

$$
\begin{aligned}
\Sigma_{\mathcal{A}}^{11} &= \left[\Sigma_{\mathcal{A}11} - \Sigma_{\mathcal{A}12}\Sigma_{\mathcal{A}22}^{-1}\Sigma'_{\mathcal{A}12}\right]^{-1} \\
&= [\Sigma'_{z_1x:\mathcal{A}}V^{11}\Sigma_{z_1x:\mathcal{A}} - \Sigma'_{z_1x:\mathcal{A}}V^{12}(V^{22})^{-1}\left(V^{12}\right)'\Sigma_{z_1x:\mathcal{A}}]^{-1} \\
&= [\Sigma'_{z_1x:\mathcal{A}}\left\{V^{11} - V^{12}(V^{22})^{-1}(V^{12})'\right\}\Sigma_{z_1x:\mathcal{A}}]^{-1} \\
&= [\Sigma'_{z_1x:\mathcal{A}}V_{11}^{-1}\Sigma_{z_1x:\mathcal{A}}]^{-1}
\end{aligned}
$$

from (A.32), where the last equality is from the fact that $V^{11} = V_{11}^{-1} + V_{11}^{-1}V_{12}V^{22}V_{12}'V_{11}^{-1}$ and $V^{12} = -V_{11}^{-1}V_{12}V^{22}$. **Q.E.D.**

Recall that $\mathcal{A} = \{j : \theta_{j0} \neq 0, j = 1, 2, \cdots p+s\}$ denotes the true model (i.e., collection of nonzero coefficients). We let $\mathcal{S} = \{j : \theta_j \neq 0,\ j = 1, 2, \cdots, p+s\}$ denote a generic candidate model and $|\mathcal{S}| = \text{card}(\mathcal{S})$, where θ_j denotes a generic estimate for θ_{j0}. We also let $S_F = \{1, \cdots, p+s\}$ denote the full model and $\mathcal{A}^c = S_F \backslash \mathcal{A}$. We define

$$\ddot{\theta}_{\mathcal{S}} \equiv \arg \min_{\{\theta \in \mathbb{R}^{p+s} : \theta_j = 0\ \forall j \notin \mathcal{S}\}} (Y_z - X_{zF}\theta)'\hat{W}(Y_z - X_{zF}\theta), \tag{A.33}$$

which is the unpenalized GMM estimator under the restriction that $\theta_j = 0$ and all $j \notin \mathcal{S}$. We also define

$$IC_{\mathcal{S}} = J(\ddot{\theta}_{\mathcal{S}}) + |\mathcal{S}| \ln(n) \max\{\ln[\ln(p+s)], 1\}.$$

The following two lemmas are useful to prove Theorem 4.

Lemma A.2. *Under Assumptions 1 and 2, we have* $P\{\min_{\mathcal{S} \not\supseteq \mathcal{A}}(IC_{\mathcal{S}}) > IC_{\mathcal{S}_F}\} \to 1$ *as* $n \to \infty$, *provided that* $n\eta^2/(p+s)\ln(n)\ln[\ln(p+s)] \to \infty$.

Proof of Lemma A.2 We let $\ddot{\theta}_{GMM}$ denote the unpenalized GMM estimator for θ, i.e., $\ddot{\theta}_{GMM} = \ddot{\theta}_{\mathcal{S}_F}$ from (A.33). Since $\ddot{\theta}_{GMM}$ is a special case of the ridge estimator $\hat{\theta}_R$ with $\lambda_2 = 0$ in (A.1), by the result in (A.13) we have

$$\begin{aligned} E\|\ddot{\theta}_{GMM} - \theta_0\|^2 &= E\|(X_{zF}'\hat{W}X_{zF})^{-1}X_{zF}'\hat{W}e\|^2 \\ &\leq [\text{Eigmin}(n^{-2}X_{zF}'\hat{W}X_{zF})]^{-2}E\|n^{-2}X_{zF}'\hat{W}e\|^2 \leq \frac{2qB}{b^2 n} \end{aligned}$$

and hence $\|\ddot{\theta}_{GMM} - \theta_0\|^2 = O_p(q/n)$. We can obtain

$$\min_{\mathcal{S} \not\supseteq \mathcal{A}} \|\ddot{\theta}_{\mathcal{S}} - \ddot{\theta}_{GMM}\|^2 \geq \frac{1}{2}\min_{\mathcal{S} \not\supseteq \mathcal{A}} \|\ddot{\theta}_{\mathcal{S}} - \theta_0\|^2 - \|\ddot{\theta}_{GMM} - \theta_0\|^2 \geq \frac{1}{2}\min_{j \in \mathcal{A}}(\theta_{0j}^2) - O_p(q/n) = \frac{1}{2}\eta^2 - O_p(q/n), \tag{A.34}$$

in which $\eta^2 n/q \to \infty$ from (A.24) so that the right hand side of (A.34) is positive w.p.a.1. Note that $X_{zF}'\hat{W}Y_z = (X_{zF}'\hat{W}X_{zF})\ddot{\theta}_{GMM}$, so we have

$$\begin{aligned} J(\ddot{\theta}_{\mathcal{S}}) - J(\ddot{\theta}_{GMM}) &= n^{-1}(Y_z - X_{zF}\ddot{\theta}_{\mathcal{S}})'\hat{W}(Y_z - X_{zF}\ddot{\theta}_{\mathcal{S}}) - n^{-1}(Y_z - X_{zF}\ddot{\theta}_{GMM})'\hat{W}(Y_z - X_{zF}\ddot{\theta}_{GMM}) \\ &= n^{-1}(\ddot{\theta}_{\mathcal{S}} - \ddot{\theta}_{GMM})'(X_{zF}'\hat{W}X_{zF})(\ddot{\theta}_{\mathcal{S}} - \ddot{\theta}_{GMM}) \\ &\geq n \cdot \text{Eigmin}(n^{-2}X_{zF}'\hat{W}X_{zF})\|\ddot{\theta}_{\mathcal{S}} - \ddot{\theta}_{GMM}\|^2 = nb\left[\frac{\eta^2}{2} - O_p\left(\frac{q}{n}\right)\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \min_{\mathcal{S} \not\supseteq \mathcal{A}}(IC_{\mathcal{S}} - IC_{\mathcal{S}_F}) &\geq \min_{\mathcal{S} \not\supseteq \mathcal{A}}\left[J(\ddot{\theta}_{\mathcal{S}}) - J(\ddot{\theta}_{GMM})\right] - (p+s)\ln(n)\max\{\ln[\ln(p+s)], 1\} \\ &\geq nb[\eta^2/2 - O_p(q/n)] - (p+s)\ln(n)\max\{\ln[\ln(p+s)], 1\}, \end{aligned}$$

which is positive w.p.a.1 from the condition that $n\eta^2/(p+s)\ln(n)\ln[\ln(p+s)] \to \infty$. **Q.E.D.**

Lemma A.3. *Under Assumptions 1, 2, and 4,* $P\{\min_{\mathcal{S} \supset \mathcal{A}, \mathcal{S} \neq \mathcal{A}}(IC_{\mathcal{S}}) > IC_{\mathcal{A}}\} \to 1$ *as* $n \to \infty$.

Proof of Lemma A.3 In this case, we have $\mathcal{A} \subset \mathcal{S}$ but $\mathcal{A} \neq \mathcal{S}$. Hence, $\mathcal{A}^c = \mathcal{S} \backslash \mathcal{A} \neq \varnothing$. Recall that $X_{zF:\mathcal{A}}$ consists of the columns of X_{zF} that correspond to nonzero elements in θ_0. We let $X_{zF:\mathcal{S}}$ be the columns of X_{zF} that are selected by model $\mathcal{S}$. We also let $X_{zF:\mathcal{A}^c}$ denote the columns of $X_{zF:\mathcal{S}}$ that correspond to zero elements in θ_0. We define

$$\begin{aligned} \check{\theta}_{\mathcal{A}} &= \arg \min_{\theta \in \mathbb{R}^{p_0+s_0}} (Y_z - X_{zF:\mathcal{A}}\theta)'\hat{W}(Y_z - X_{zF:\mathcal{A}}\theta), \\ \check{\theta}_{\mathcal{S}} &= \arg \min_{\theta \in \mathbb{R}^{|\mathcal{S}|}} (Y_z - X_{zF:\mathcal{S}}\theta)'\hat{W}(Y_z - X_{zF:\mathcal{S}}\theta), \\ \check{\theta}_{\mathcal{A}^c} &= \arg \min_{\theta \in \mathbb{R}^{|\mathcal{S}|-|\mathcal{A}|}} (Y_z - X_{zF:\mathcal{A}^c}\theta)'\hat{W}(Y_z - X_{zF:\mathcal{A}^c}\theta). \end{aligned} \tag{A.35}$$

In this case, we have

$$J(\check{\theta}_{\mathcal{A}}) = n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF:\mathcal{A}}\check{\theta}_{\mathcal{A}}\|^2.$$

For the model $\mathcal{S}$, by the Frisch-Waugh Theorem, we have

$$\begin{aligned} J(\check{\theta}_{\mathcal{S}}) &= \min_{\theta \in R^{|\mathcal{S}|}} n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF:\mathcal{S}}\theta\|^2 \\ &= \min_{\theta \in R^{|\mathcal{S}|-|\mathcal{A}|}} \|\tilde{\mathbb{Y}}_{\mathcal{A}^c} - \tilde{\mathbb{X}}_{\mathcal{A}^c}\theta\|^2, \end{aligned} \tag{A.36}$$

where $\tilde{\mathbb{Y}}_{\mathcal{A}^c} = M_{\mathcal{A}}\hat{W}^{1/2}Y_z$ and $\tilde{\mathbb{X}}_{\mathcal{A}^c} = M_{\mathcal{A}}\hat{W}^{1/2}X_{zF:\mathcal{A}^c}$. Note that we define $M_{\mathcal{A}} = I_q - P_{\mathcal{A}}$ and $P_{\mathcal{A}} = \hat{W}^{1/2}X_{zF:\mathcal{A}}(X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}})^{-1}X'_{zF:\mathcal{A}}\hat{W}^{1/2}$. Therefore,

$$\begin{aligned} J(\check{\theta}_{\mathcal{A}}) - J(\check{\theta}_{\mathcal{S}}) &= n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF\mathcal{A}}\check{\theta}_{\mathcal{A}}\|^2 - n^{-1}\|\tilde{\mathbb{Y}}_{\mathcal{A}^c} - \tilde{\mathbb{X}}_{\mathcal{A}^c}\check{\theta}_{\mathcal{A}^c}\|^2 \\ &= n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF\mathcal{A}}\check{\theta}_{\mathcal{A}}\|^2 - n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF\mathcal{A}}\check{\theta}_{\mathcal{A}} - \tilde{\mathbb{X}}_{\mathcal{A}^c}\check{\theta}_{\mathcal{A}^c}\|^2 \\ &= n^{-1}\check{\theta}'_{\mathcal{A}^c}(\tilde{\mathbb{X}}'_{\mathcal{A}^c}\tilde{\mathbb{X}}_{\mathcal{A}^c})\check{\theta}_{\mathcal{A}^c} \\ &= n^{-1}\tilde{\mathbb{Y}}'_{\mathcal{A}^c}\tilde{\mathbb{X}}_{\mathcal{A}^c}(\tilde{\mathbb{X}}'_{\mathcal{A}^c}\tilde{\mathbb{X}}_{\mathcal{A}^c})^{-1}\tilde{\mathbb{X}}'_{\mathcal{A}^c}\tilde{\mathbb{Y}}_{\mathcal{A}^c} \\ &= n^{-1}e'\hat{W}^{1/2}\tilde{\mathbb{X}}_{\mathcal{A}^c}(\tilde{\mathbb{X}}'_{\mathcal{A}^c}\tilde{\mathbb{X}}_{\mathcal{A}^c})^{-1}\tilde{\mathbb{X}}'_{\mathcal{A}^c}\hat{W}^{1/2}e, \end{aligned} \tag{A.37}$$

where the last equality uses the fact that $\tilde{\mathbb{Y}}_{\mathcal{A}^c} = M_{\mathcal{A}}\hat{W}^{1/2}Y_z = M_{\mathcal{A}}\hat{W}^{1/2}e$ and since $e = Y_z - X_{zF:\mathcal{A}}\theta_{\mathcal{A}}$. We now let $P_{\tilde{\mathbb{X}}_{\mathcal{A}^c}} \equiv \tilde{\mathbb{X}}_{\mathcal{A}^c}(\tilde{\mathbb{X}}'_{\mathcal{A}^c}\tilde{\mathbb{X}}_{\mathcal{A}^c})^{-1}\tilde{\mathbb{X}}'_{\mathcal{A}^c}$, which is positive semi-definite with rank $|\mathcal{S}| - p_0 - s_0$. Recall that the optimal weighting matrix $\hat{W}$ is chosen such that $||\hat{W} - V^{-1}|| \xrightarrow{p} 0$ and it holds that $||n^{-1}X_{zF} - \Sigma_{zxF}|| \xrightarrow{p} 0$, so there exists a nonrandom matrix $P^0_{\tilde{\mathbb{X}}_{\mathcal{A}^c}}$ such that $||P_{\tilde{\mathbb{X}}_{\mathcal{A}^c}} - P^0_{\tilde{\mathbb{X}}_{\mathcal{A}^c}}|| \xrightarrow{p} 0$, where $P^0_{\tilde{\mathbb{X}}_{\mathcal{A}^c}}$ is symmetric and idempotent. Therefore, by the Spectral decomposition, we can decompose $P^0_{\tilde{\mathbb{X}}_{\mathcal{A}^c}} = LL'$, where L is an $q \times (|\mathcal{S}| - p_0 - s_0)$ orthonormal matrix consisting of the eigenvectors corresponding to the eigenvalues of 1. If we let L_j denote the jth column of L, then (A.37) can be rewritten as

$$\begin{aligned} J(\check{\theta}_{\mathcal{A}}) - J(\check{\theta}_{\mathcal{S}}) &= n^{-1}e'V^{-1/2}LL'V^{-1/2}e[1 + o_p(1)], \\ &= \sum_{j=1}^{|\mathcal{S}|-p_0-s_0} \left| n^{-1/2}L'_jV^{-1/2}e \right|^2 [1 + o_p(1)] \\ &\leq \left(\max_{j=1,\cdots,|\mathcal{S}|-p_0-s_0} \left| n^{-1/2}\sum_{i=1}^{n} L'_jV^{-1/2}e_i \right|^2 \right) (|\mathcal{S}| - p_0 - s_0) \times [1 + o_p(1)]. \end{aligned}$$

However, $\sum_{k=1}^{q} L_{jk}^2 = 1$ for all j and $n^{-1/2}\sum_{i=1}^{n} L_j' V^{-1/2} e_i = \sum_{i=1}^{n}\sum_{k=1}^{q} n^{-1/2} L_{jk}\tilde{e}_{ik}$, where L_{jk} is the kth element of L_j and $\tilde{e}_{ik}$ is the kth element of $V^{-1/2}e_i$. Moreover, by the Assumption 4, we have

$$P\left(\max_{j=1,\cdots,|\mathcal{S}|-p_0-s_0}\left|\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{L_{jk}}{\sqrt{n}}\tilde{e}_{ik}\right| > \log(|\mathcal{S}|)\right) \leq \frac{c_0\{\log[|\mathcal{S}|-p_0-s_0)]\}^{1/2}}{\log(|\mathcal{S}|)} \leq \frac{c_0}{[\log(|\mathcal{S}|)]^{-1/2}}$$

for some positive constant $c_0 < \infty$. It thus follows that

$$J(\check{\theta}_{\mathcal{A}}) - J(\check{\theta}_{\mathcal{S}}) \leq \log(|\mathcal{S}|)(|\mathcal{S}|-p_0-s_0)[1+o_p(1)]$$

w.p.a.1. Therefore,

$$\begin{aligned} IC_{\mathcal{S}} - IC_{\mathcal{A}} &= J(\check{\theta}_{\mathcal{S}}) - J(\check{\theta}_{\mathcal{A}}) + (|\mathcal{S}|-p_0-s_0)\ln(n)\max\{\ln[\ln(p+s)],1\} \\ &\geq (|\mathcal{S}|-p_0-s_0)\left(\ln(n)\max\{\ln[\ln(p+s)],1\} - \log(|\mathcal{S}|)[1+o_p(1)]\right) \end{aligned}$$

yielding

$$\begin{aligned} \min_{\mathcal{S}\supset\mathcal{A},\mathcal{S}\neq\mathcal{A}}\left\{\frac{IC_{\mathcal{S}} - IC_{\mathcal{A}}}{(|\mathcal{S}|-p_0-s_0)}\right\} &\geq \ln(n)\max\{\ln[\ln(p+s)],1\} - \log(p+s)[1+o_p(1)] \\ &\geq \ln(n)\left(\max\{\ln[\ln(p+s)],1\} - \nu[1+o_p(1)]\right), \qquad \text{(A.38)} \end{aligned}$$

where $\lim_{n,p,s\to\infty}[(p+s)/n^{\nu}] < 1$ for some $0 \leq \nu \leq \alpha < 1$ from Assumption 2-(i). Since $1-\nu[1+o_p(1)] > 0$ w.p.a.1 and $\ln[\ln(p+s)] \to \infty$ as $p+s\to\infty$, the right hand side of the above inequality is positive w.p.a.1. **Q.E.D.**

Proof of Theorem 4 We note that $\check{\theta}_{\mathcal{A}} = (X_{zF:\mathcal{A}}'\hat{W}X_{zF:\mathcal{A}})^{-1}X_{zF:\mathcal{A}}'\hat{W}Y_z$ from (A.35), so we have

$$\begin{aligned} \tilde{\theta}_{\mathcal{A}} &= \left(X_{zF:\mathcal{A}}'\hat{W}X_{zF:\mathcal{A}} + \lambda_2 I_{p_0+s_0}\right)^{-1}\left[X_{zF:\mathcal{A}}'\hat{W}Y_z - \frac{1}{2}\lambda_1^*\text{diag}\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}\hat{\pi}_{\mathcal{A}}\right], \\ &= \left(X_{zF:\mathcal{A}}'\hat{W}X_{zF:\mathcal{A}} + \lambda_2 I_{p_0+s_0}\right)^{-1}[X_{zF:\mathcal{A}}'\hat{W}X_{zF:\mathcal{A}}\check{\theta}_{\mathcal{A}} - \frac{1}{2}\lambda_1^*\text{diag}\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}\hat{\pi}_{\mathcal{A}}] \end{aligned}$$

from (A.14), where $\hat{\pi}_{\mathcal{A}}$ is the column vector of the adaptive weights corresponding to nonzero coefficients of $\hat{\theta}$. Recall that $\hat{\theta}_{\mathcal{A}}$ is the adaptive elastic net GMM estimator $\hat{\theta}$ in (3) that corresponds to $\theta_{\mathcal{A}}$. diag$\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}$ is the diagonal matrix whose diagonal elements are sgn$(\hat{\theta}_{\mathcal{A}})$. We further let λ_n be the sequence of λ such that $\hat{\theta}$ defined in (3) has the properties shown in Theorem 1. Then $P(\mathcal{S}_{\lambda_n} = \mathcal{A}) \to 1$ by Theorem 1, where $\mathcal{S}_{\lambda_n} = \{j : \hat{\theta}_{\lambda_n,j} \neq 0\}$. It follows that w.p.a.1

$$\begin{aligned} J(\tilde{\theta}_{\mathcal{A}}) - J(\check{\theta}_{\mathcal{S}_{\lambda_n}}) &= n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}}\|^2 - n^{-1}\|\hat{W}^{1/2}Y_z - \hat{W}^{1/2}X_{zF:\mathcal{A}}\check{\theta}_{\mathcal{A}}\|^2 \\ &= n^{-1}\tilde{\theta}_{\mathcal{A}}'X_{zF:\mathcal{A}}'\hat{W}X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}} - 2n^{-1}Y_z'\hat{W}X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}} + n^{-1}\check{\theta}_{\mathcal{A}}'X_{zF:\mathcal{A}}'\hat{W}X_{zF:\mathcal{A}}\check{\theta}_{\mathcal{A}} \\ &= n(\tilde{\theta}_{\mathcal{A}} - \check{\theta}_{\mathcal{A}})'(n^{-2}X_{zF}'\hat{W}X_{zF})(\tilde{\theta}_{\mathcal{A}} - \check{\theta}_{\mathcal{A}}) \\ &\leq n\text{Eigmax}(n^{-2}X_{zF}'\hat{W}X_{zF})\|\tilde{\theta}_{\mathcal{A}} - \check{\theta}_{\mathcal{A}}\|^2. \end{aligned}$$

Note that, however, w.p.a.1

$$
\begin{aligned}
\|\tilde{\theta}_{\mathcal{A}}-\check{\theta}_{\mathcal{A}}\|^2 &= \left\|\left(X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}+\lambda_2 I_{p_0+s_0}\right)^{-1}[X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}\check{\theta}_{\mathcal{A}}-\frac{1}{2}\lambda_1^*\text{diag}\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}\hat{\pi}_{\mathcal{A}}]-\check{\theta}_{\mathcal{A}}\right\|^2 \\
&= \left\|\left(\frac{X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}}{n^2}+\frac{\lambda_2}{n^2}I_{p_0+s_0}\right)^{-1}\left[\frac{\lambda_2\check{\theta}_{\mathcal{A}}}{n^2}+\frac{\lambda_1^*}{2n^2}\text{diag}\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}\hat{\pi}_{\mathcal{A}}\right]\right\|^2 \\
&\leq \left[\text{Eigmin}\left(\frac{X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}}{n^2}+\frac{\lambda_2}{n^2}I_{p_0+s_0}\right)\right]^{-2}\left\|\frac{\lambda_2\check{\theta}_{\mathcal{A}}}{n^2}+\frac{\lambda_1^*}{2n^2}\text{diag}\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}\hat{\pi}_{\mathcal{A}}\right\|^2 \\
&\leq 2b^{-2}\left(\|n^{-2}\lambda_2\check{\theta}_{\mathcal{A}}\|^2+\|n^{-2}2\lambda_1^*\text{diag}\{\text{sgn}(\hat{\theta}_{\mathcal{A}})\}\hat{\pi}_{\mathcal{A}}\|^2\right) \\
&= O_p\left(\frac{\lambda_2^2(p+s)}{n^4}\right)+O_p\left(\frac{\lambda_1^{*^2}(p+s)}{n^4\eta^{2\gamma}}\right)=o_p(n^{-1}) \qquad \text{(A.39)}
\end{aligned}
$$

since $\text{Eigmin}(X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}+\lambda_2 I_{p_0+s_0})\geq\text{Eigmin}(X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}})\geq\text{Eigmin}(X'_{zF}\hat{W}X_{zF})$. Hence, $J(\tilde{\theta}_{\mathcal{A}})-J(\check{\theta}_{\mathcal{S}_{\lambda_n}})=o_p(1)$ w.p.a.1 from Assumption 2-(iii) and (iv). Also note that $P[\hat{\theta}_{\mathcal{A}}=(1+\lambda_2/n^2)\tilde{\theta}_{\mathcal{A}}]\to 1$ by Lemma A.1, so w.p.a.1 we have

$$
\begin{aligned}
&J(\hat{\theta}_{\mathcal{A}})-J(\tilde{\theta}_{\mathcal{A}}) \\
&= n^{-1}\|\hat{W}^{1/2}Y_z-\hat{W}^{1/2}X_{zF:\mathcal{A}}\hat{\theta}_{\mathcal{A}}\|^2-n^{-1}\|\hat{W}^{1/2}Y_z-\hat{W}^{1/2}X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}}\|^2 \\
&= n^{-3}\lambda_2^2\tilde{\theta}_{\mathcal{A}}\left(n^{-2}X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}\right)\tilde{\theta}_{\mathcal{A}}-2n^{-3}\lambda_2(Y_z-X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}})'\hat{W}X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}} \\
&\leq n^{-3}\lambda_2^2\text{Eigmax}(n^{-2}X'_{zF}\hat{W}X_{zF})\|\tilde{\theta}_{\mathcal{A}}\|^2-2n^{-3}\lambda_2[Y_z-X_{zF:\mathcal{A}}\check{\theta}_{\mathcal{A}}+X_{zF:\mathcal{A}}(\check{\theta}_{\mathcal{A}}-\tilde{\theta}_{\mathcal{A}})]'\hat{W}X_{zF:\mathcal{A}}\tilde{\theta}_{\mathcal{A}} \\
&\leq n^{-3}\lambda_2^2B\|\tilde{\theta}_{\mathcal{A}}\|^2-2n^{-1}\lambda_2(\check{\theta}_{\mathcal{A}}-\tilde{\theta}_{\mathcal{A}})'\left(n^{-2}X'_{zF:\mathcal{A}}\hat{W}X_{zF:\mathcal{A}}\right)\tilde{\theta}_{\mathcal{A}} \\
&\leq n^{-3}\lambda_2^2B\|\tilde{\theta}_{\mathcal{A}}\|^2+2n^{-1}\lambda_2B\|\check{\theta}_{\mathcal{A}}-\tilde{\theta}_{\mathcal{A}}\|\cdot\|\tilde{\theta}_{\mathcal{A}}\| \\
&= O_p\left(\frac{\lambda_2^2(p+s)}{n^3}\right)+O_p\left(\frac{\lambda_2(p+s)^{1/2}}{n^{3/2}}\right)\cdot\left[O_p\left(\frac{\lambda_2^2(p+s)}{n^3}\right)+O_p\left(\frac{\lambda_1^{*^2}(p+s)}{n^3\eta^{2\gamma}}\right)\right]^{1/2}=o_p(1),
\end{aligned}
$$

where we use (A.39) and Assumptions 2-(iii) and (iv) in the last equality. Hence, $J(\hat{\theta}_{\mathcal{A}})-J(\check{\theta}_{\mathcal{S}_{\lambda_n}})=o_p(1)$ w.p.a.1, which implies

$$IC_{\lambda_n}-IC_{\mathcal{S}_{\lambda_n}}=o_p(1)$$

w.p.a.1. We now define $\Omega_-=\{\lambda:\mathcal{S}_\lambda\not\supseteq\mathcal{A}\}$, $\Omega_+=\{\lambda:\mathcal{A}\subseteq\mathcal{S},\mathcal{A}\neq\mathcal{S}\}$ and $\Omega_0=\{\lambda:\mathcal{S}_\lambda=\mathcal{A}\}$. Note that $P(\mathcal{S}_{\lambda_n}=\mathcal{A})\to 1$, so w.p.a.1 we have

$$
\begin{aligned}
\inf_{\lambda\in\Omega_-}(IC_\lambda)-IC_{\lambda_n} &\geq \inf_{\lambda\in\Omega_-}(IC_\lambda)-IC_{\mathcal{A}}+o_p(1) \\
&\geq \min_{\mathcal{S}_\lambda\not\supseteq\mathcal{A}}(IC_{\mathcal{S}})-IC_{\mathcal{A}}+o_p(1) \\
&= \min_{\mathcal{S}_\lambda\not\supseteq\mathcal{A}}(IC_{\mathcal{S}})-IC_{\mathcal{S}_F}+o_p(1)+IC_{\mathcal{S}_F}-IC_{\mathcal{A}}.
\end{aligned}
$$

From Lemma A.2, we know that $\min_{\mathcal{S}_\lambda \not\supseteq \mathcal{A}}(IC_{\mathcal{S}}) - IC_{\mathcal{S}_F} + o_p(1) > 0$ w.p.a.1. In addition, Lemma A.3 shows that $P(IC_{\mathcal{S}_F} - IC_{\mathcal{A}} > 0) \to 1$. We thus have $P\{\inf_{\lambda\in\Omega_-}(IC_\lambda) - IC_{\lambda_n} > 0\} \to 1$ as $n \to \infty$. Similarly, we have

$$\inf_{\lambda\in\Omega_+}(IC_\lambda) - IC_{\lambda_n} \geq \inf_{\lambda\in\Omega_+}(IC_\lambda) - IC_{\mathcal{A}} + o_p(1) \geq \min_{\mathcal{A}\subseteq\mathcal{S},\mathcal{A}\neq\mathcal{S}}(IC_{\mathcal{S}}) - IC_{\mathcal{A}} + o_p(1).$$

Recalling that

$$\min_{\mathcal{A}\subseteq\mathcal{S},\mathcal{A}\neq\mathcal{S}}\left\{\frac{IC_{\mathcal{S}} - IC_{\mathcal{A}}}{(|\mathcal{S}| - p_0 - s_0)}\right\} \geq \ln(n)\left(\max\{\ln[\ln(p+s)], 1\} - \nu[1 + o_p(1)]\right) \to \infty$$

in (A.38), we can conclude that $P\{\inf_{\lambda\in\Omega_+}(IC_\lambda) - IC_{\lambda_n} > 0\} \to 1$, which gives the desired result. **Q.E.D.**

References

Andrews, D.W.K. (1999). Consistent moment selection procedures for generalized method of moments estimation, *Econometrica*, 67, 543-564.

Andrews, D.W.K. and B. Lu (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data 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, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain, *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 A.B. Kock (2014). An oracle inequality for convex loss with elastic net, Working paper, Department of Economics, North Carolina State University.

Caner, M. and H.H. Zhang (2014). Adaptive Elastic Net GMM Estimator, *Journal of Business and Economics Statistics*, 32, 30-47.

Cheng, X. and Z. Liao (2013). 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, and R. Tibshirani (2004). Least Angle Regression, *Annals of Statistics*, 32, 407-499.

Fan, J. and Y. Liao (2012). Endogeneity in Ultrahigh Dimension, Working Paper, Department of Economics, Princeton University.

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.

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

Huang, J., J. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *The Annals of Statistics*, 36, 587–613.

Huang, J., S. Ma, and C.–H. Zhang (2007). Adaptive LASSO for sparse high-dimensional regression models, *Technical Report 374*, Department of Statistics and Actuarial Science, University of Iowa.

Lee, Y. and R. Okui (2012). Hahn-Hausman Test as a Specification Test, 2012, *Journal of Econometrics*, 167, 133-139.

Leeb, H. and B. M. Pötscher (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.

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

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

Wang, H., R. Li, and 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 $\beta_{\mathcal{A}}$

$n = 250$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$ and $\rho_{uv} = 0.5$												
	AENet				ALasso-LARS				ALasso-CL			
$\tau_{\mathcal{A}}$, C, ρ_z	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$
.3, .25, .5	0.014	0.160	0.037	0.112	0.015	0.159	0.037	0.111	0.005	0.282	0.096	0.094
.3, .50, .5	0.013	0.166	0.033	0.089	0.013	0.166	0.033	0.087	0.005	0.282	0.096	0.094
.3, .75, .5	0.013	0.166	0.033	0.088	0.013	0.167	0.032	0.086	0.005	0.282	0.096	0.094
.6, .25, .5	0.013	0.203	0.034	0.111	0.013	0.197	0.033	0.112	0.005	0.557	0.105	0.104
.6, .50, .5	0.012	0.213	0.029	0.083	0.012	0.210	0.029	0.083	0.005	0.557	0.105	0.104
.6, .75, .5	0.012	0.215	0.029	0.082	0.012	0.212	0.028	0.081	0.005	0.557	0.105	0.104
.9, .25, .5	0.013	0.278	0.033	0.111	0.014	0.261	0.032	0.111	0.005	0.820	0.107	0.106
.9, .50, .5	0.012	0.289	0.028	0.082	0.012	0.280	0.028	0.082	0.005	0.820	0.107	0.106
.9, .75, .5	0.012	0.291	0.028	0.081	0.012	0.283	0.027	0.080	0.005	0.820	0.107	0.106
.3, .25, .95	0.009	0.202	0.079	0.181	0.009	0.202	0.080	0.183	0.010	0.298	0.225	0.191
.3, .50, .95	0.008	0.209	0.072	0.164	0.008	0.210	0.073	0.167	0.010	0.298	0.225	0.190
.3, .75, .95	0.007	0.212	0.066	0.145	0.007	0.213	0.066	0.147	0.010	0.298	0.225	0.190
.6, .25, .95	0.008	0.234	0.071	0.176	0.008	0.229	0.072	0.177	0.010	0.594	0.232	0.203
.6, .50, .95	0.007	0.249	0.065	0.156	0.007	0.246	0.066	0.157	0.010	0.594	0.232	0.203
.6, .75, .95	0.007	0.253	0.058	0.133	0.007	0.250	0.058	0.134	0.010	0.594	0.232	0.203
.9, .25, .95	0.008	0.297	0.068	0.176	0.008	0.284	0.067	0.176	0.010	0.887	0.235	0.207
.9, .50, .95	0.007	0.320	0.062	0.154	0.007	0.309	0.062	0.155	0.010	0.887	0.235	0.207
.9, .75, .95	0.006	0.327	0.054	0.129	0.007	0.316	0.055	0.130	0.010	0.887	0.235	0.207

$n = 1000$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$ and $\rho_{uv} = 0.5$												
	AENet				ALasso-LARS				ALasso-CL			
$\tau_{\mathcal{A}}$, C, ρ_z	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$
.3, .25, .5	0.004	0.049	0.008	0.039	0.004	0.049	0.007	0.039	0.004	0.054	0.040	0.038
.3, .50, .5	0.003	0.050	0.007	0.037	0.003	0.050	0.007	0.037	0.004	0.054	0.040	0.038
.3, .75, .5	0.003	0.050	0.007	0.037	0.003	0.050	0.007	0.037	0.004	0.054	0.040	0.038
.6, .25, .5	0.004	0.065	0.007	0.039	0.004	0.063	0.007	0.039	0.004	0.097	0.040	0.038
.6, .50, .5	0.003	0.067	0.007	0.037	0.003	0.066	0.007	0.036	0.004	0.097	0.040	0.038
.6, .75, .5	0.003	0.067	0.007	0.037	0.003	0.066	0.007	0.036	0.004	0.097	0.040	0.038
.9, .25, .5	0.003	0.087	0.007	0.039	0.004	0.084	0.007	0.039	0.004	0.140	0.040	0.038
.9, .50, .5	0.003	0.090	0.007	0.036	0.003	0.088	0.007	0.036	0.004	0.140	0.040	0.038
.9, .75, .5	0.003	0.091	0.007	0.036	0.003	0.089	0.006	0.036	0.004	0.140	0.040	0.038
.3, .25, .95	0.002	0.053	0.029	0.088	0.002	0.053	0.029	0.088	0.007	0.291	0.157	0.121
.3, .50, .95	0.002	0.056	0.022	0.067	0.002	0.056	0.022	0.067	0.007	0.291	0.157	0.121
.3, .75, .95	0.002	0.056	0.020	0.057	0.002	0.056	0.019	0.057	0.007	0.291	0.157	0.121
.6, .25, .95	0.002	0.066	0.028	0.088	0.002	0.064	0.028	0.088	0.007	0.574	0.172	0.137
.6, .50, .95	0.002	0.071	0.021	0.066	0.002	0.069	0.021	0.066	0.007	0.574	0.172	0.137
.6, .75, .95	0.002	0.072	0.018	0.055	0.002	0.070	0.018	0.055	0.007	0.574	0.172	0.137
.9, .25, .95	0.002	0.089	0.027	0.088	0.002	0.083	0.027	0.089	0.007	0.854	0.176	0.141
.9, .50, .95	0.002	0.094	0.021	0.066	0.002	0.090	0.021	0.066	0.007	0.854	0.176	0.141
.9, .75, .95	0.002	0.096	0.018	0.054	0.002	0.092	0.018	0.055	0.007	0.854	0.176	0.141

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 (2013). ρ_z controls the correlation of Z_{1i}^1. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. C is the value of nonzero structural parameters. $rmse_1$, $rmse_2$, $rmse_3$ and $rmse_4$ denote the RMSEs of $\tau_{\mathcal{A}^c}$, $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$, respectively.

Table 2: Moment Selection Accuracy

$n = 250$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$ and $\rho_{uv} = 0.5$						
	AENet		ALasso-LARS		ALasso-CL	
$\tau_{\mathcal{A}}$, C, ρ_z	Pr_1	Pr_2	Pr_1	Pr_2	Pr_1	Pr_2
.3, .25, .5	0.957	0.920	0.956	0.922	0.997	0.256
.3, .50, .5	0.963	0.912	0.962	0.913	0.997	0.256
.3, .75, .5	0.964	0.912	0.963	0.912	0.997	0.256
.6, .25, .5	0.965	1.000	0.962	1.000	0.997	0.368
.6, .50, .5	0.970	1.000	0.968	1.000	0.997	0.368
.6, .75, .5	0.970	1.000	0.969	1.000	0.997	0.368
.9, .25, .5	0.967	1.000	0.961	1.000	0.997	0.424
.9, .50, .5	0.972	1.000	0.967	1.000	0.997	0.424
.9, .75, .5	0.972	1.000	0.968	1.000	0.997	0.424
.3, .25, .95	0.977	0.758	0.977	0.762	0.988	0.017
.3, .50, .95	0.980	0.732	0.980	0.734	0.988	0.017
.3, .75, .95	0.982	0.725	0.981	0.726	0.988	0.017
.6, .25, .95	0.981	0.995	0.980	0.995	0.987	0.024
.6, .50, .95	0.983	0.993	0.982	0.994	0.987	0.024
.6, .75, .95	0.985	0.993	0.984	0.993	0.987	0.024
.9, .25, .95	0.982	0.999	0.980	0.999	0.987	0.035
.9, .50, .95	0.984	0.999	0.982	0.999	0.987	0.035
.9, .75, .95	0.986	0.999	0.984	0.999	0.987	0.035

$n = 1000$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$ and $\rho_{uv} = 0.5$						
	AENet		ALasso-LARS		ALasso-CL	
$\tau_{\mathcal{A}}$, C, ρ_z	Pr_1	Pr_2	Pr_1	Pr_2	Pr_1	Pr_2
.3, .25, .5	0.993	1.000	0.992	1.000	0.995	1.000
.3, .50, .5	0.993	1.000	0.993	1.000	0.995	1.000
.3, .75, .5	0.993	1.000	0.993	1.000	0.995	1.000
.6, .25, .5	0.993	1.000	0.992	1.000	0.995	1.000
.6, .50, .5	0.993	1.000	0.993	1.000	0.995	1.000
.6, .75, .5	0.994	1.000	0.993	1.000	0.995	1.000
.9, .25, .5	0.993	1.000	0.992	1.000	0.995	1.000
.9, .50, .5	0.994	1.000	0.993	1.000	0.995	1.000
.9, .75, .5	0.994	1.000	0.993	1.000	0.995	1.000
.3, .25, .95	0.997	1.000	0.997	1.000	0.977	0.066
.3, .50, .95	0.997	1.000	0.997	1.000	0.977	0.066
.3, .75, .95	0.997	1.000	0.997	1.000	0.977	0.066
.6, .25, .95	0.997	1.000	0.996	1.000	0.977	0.109
.6, .50, .95	0.998	1.000	0.997	1.000	0.977	0.109
.6, .75, .95	0.998	1.000	0.997	1.000	0.977	0.109
.9, .25, .95	0.997	1.000	0.996	1.000	0.977	0.139
.9, .50, .95	0.998	1.000	0.997	1.000	0.977	0.139
.9, .75, .95	0.998	1.000	0.997	1.000	0.977	0.139

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 (2013). ρ_z controls the correlation of Z_{1i}^1. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. C 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: RMSEs in the Case of Highly Correlated Endogenous Variables

$n = 250$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$, $\rho_{uv} = 0.95$, and $\rho_z = 0.99$								
	AENet				ALasso-LARS			
$\tau_{\mathcal{A}}$, C	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$
.3, .25	0.004	0.277	0.126	0.227	0.004	0.276	0.139	0.236
.3, .50	0.003	0.282	0.117	0.241	0.003	0.283	0.128	0.247
.3, .75	0.005	0.434	0.125	0.223	0.005	0.431	0.139	0.231
.6, .25	0.005	0.434	0.125	0.223	0.005	0.431	0.139	0.231
.6, .50	0.004	0.460	0.118	0.237	0.004	0.459	0.130	0.243
.6, .75	0.004	0.463	0.120	0.256	0.004	0.464	0.130	0.262
.9, .25	0.005	0.560	0.120	0.217	0.005	0.546	0.133	0.223
.9, .50	0.004	0.601	0.114	0.229	0.004	0.594	0.125	0.234
.9, .75	0.004	0.611	0.117	0.249	0.004	0.604	0.127	0.253

$n = 1000$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$, $\rho_{uv} = 0.95$, and $\rho_z = 0.99$								
	AENet				ALasso-LARS			
$\tau_{\mathcal{A}}$, C	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$	$rmse_1$	$rmse_2$	$rmse_3$	$rmse_4$
.3, .25	0.001	0.083	0.059	0.122	0.001	0.083	0.065	0.125
.3, .50	0.001	0.088	0.063	0.141	0.001	0.090	0.069	0.144
.3, .75	0.001	0.089	0.060	0.139	0.001	0.091	0.065	0.142
.6, .25	0.001	0.112	0.056	0.121	0.001	0.107	0.063	0.123
.6, .50	0.001	0.123	0.061	0.138	0.001	0.119	0.066	0.140
.6, .75	0.001	0.125	0.057	0.134	0.001	0.121	0.063	0.137
.9, .25	0.001	0.155	0.056	0.121	0.001	0.141	0.061	0.123
.9, .50	0.001	0.170	0.060	0.138	0.001	0.157	0.065	0.139
.9, .75	0.001	0.174	0.057	0.134	0.001	0.161	0.062	0.136

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 of Z_{1i}^1. ρ_{uv} is the correlation between u and v. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. C is the value of nonzero structural parameters. $rmse_1$, $rmse_2$, $rmse_3$ and $rmse_4$ denote the RMSEs of $\tau_{\mathcal{A}^c}$, $\tau_{\mathcal{A}}$, $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$, respectively.

Table 4: Redundant Instruments

$n = 250$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$ and $\rho_{uv} = 0.5$									
	AENet			ALasso-LARS			ALasso-CL		
$\tau_{\mathcal{A}},\ C,\ \rho_z$	$rmse_3$	$rmse_4$	Pr_5	$rmse_3$	$rmse_4$	Pr_5	$rmse_3$	$rmse_4$	Pr_5
.3, .25, .5	0.044	0.113	0.022	0.043	0.113	0.022	0.114	0.092	0.020
.3, .50, .5	0.039	0.088	0.019	0.039	0.087	0.019	0.114	0.092	0.020
.3, .75, .5	0.039	0.088	0.019	0.038	0.086	0.019	0.114	0.092	0.020
.6, .25, .5	0.039	0.111	0.016	0.038	0.112	0.016	0.126	0.100	0.019
.6, .50, .5	0.034	0.083	0.015	0.034	0.083	0.015	0.125	0.100	0.019
.6, .75, .5	0.034	0.082	0.015	0.033	0.082	0.015	0.125	0.100	0.019
.9, .25, .5	0.039	0.112	0.017	0.038	0.112	0.019	0.127	0.101	0.018
.9, .50, .5	0.033	0.083	0.014	0.032	0.082	0.016	0.127	0.101	0.018
.9, .75, .5	0.032	0.081	0.013	0.032	0.080	0.015	0.127	0.101	0.018

$n = 1000$, $p = 18$, $p_0 = 3$, $s = 15$, $s_0 = 6$, $q = 42$ and $\rho_{uv} = 0.5$									
	AENet			ALasso-LARS			ALasso-CL		
$\tau_{\mathcal{A}},\ C,\ \rho_z$	$rmse_3$	$rmse_4$	Pr_5	$rmse_3$	$rmse_4$	Pr_5	$rmse_3$	$rmse_4$	Pr_5
.3, .25, .5	0.010	0.040	0.003	0.009	0.040	0.004	0.048	0.039	0.286
.3, .50, .5	0.009	0.038	0.003	0.009	0.038	0.003	0.048	0.039	0.286
.3, .75, .5	0.009	0.038	0.003	0.009	0.038	0.003	0.048	0.039	0.286
.6, .25, .5	0.009	0.040	0.003	0.009	0.040	0.003	0.048	0.039	0.284
.6, .50, .5	0.009	0.037	0.003	0.008	0.037	0.003	0.048	0.039	0.284
.6, .75, .5	0.009	0.037	0.003	0.008	0.037	0.003	0.048	0.039	0.284
.9, .25, .5	0.009	0.040	0.003	0.009	0.040	0.003	0.048	0.039	0.284
.9, .50, .5	0.008	0.037	0.003	0.008	0.037	0.003	0.048	0.039	0.284
.9, .50, .5	0.008	0.037	0.003	0.008	0.037	0.003	0.048	0.039	0.284

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 (2013). ρ_z controls the correlation of Z_{1i}^1. $\tau_{\mathcal{A}}$ is the expectation of the invalid moment conditions. C is the value of nonzero structural parameters. $rmse_3$ and $rmse_4$ denote the RMSEs of $\beta_{\mathcal{A}^c}$, and $\beta_{\mathcal{A}}$, respectively. Pr_5 is the percentage of replications that yield nonzero estimates for τ for redundant instruments.