Published in final edited form as:

J Bus Econ Stat. 2014 January 30; 32(1): 30–47. doi:10.1080/07350015.2013.836104.

Adaptive Elastic Net for Generalized Methods of Moments

Mehmet Caner* and

Department of Economics, 4168 Nelson Hall, North Carolina State University, Raleigh, NC 27518.

Hao Helen Zhang

Department of Mathematics, University of Arizona, Tucson, AZ 85718.

hzhang@math.arizona.edu; and Department of Statistics, North Carolina State University, Raleigh, NC 27695. hzhang@stat.ncsu.edu.

Abstract

Model selection and estimation are crucial parts of econometrics. This paper introduces a new technique that can simultaneously estimate and select the model in generalized method of moments (GMM) context. The GMM is particularly powerful for analyzing complex data sets such as longitudinal and panel data, and it has wide applications in econometrics. This paper extends the least squares based adaptive elastic net estimator of Zou and Zhang (2009) to nonlinear equation systems with endogenous variables. The extension is not trivial and involves a new proof technique due to estimators lack of closed form solutions. Compared to Bridge-GMM of Caner (2009), we allow for the number of parameters to diverge to infinity as well as collinearity among a large number of variables, also the redundant parameters set to zero via a data dependent technique. This method has the oracle property, meaning that we can estimate nonzero parameters with their standard limit and the redundant parameters are dropped from the equations simultaneously. Numerical examples are used to illustrate the performance of the new method.

Keywords

Penalized estimators; Oracle property; Method of moments; GMM

1 Introduction

One of the most commonly used estimation techniques is Generalized Method of Moments (GMM) estimation. The GMM provides a unified framework for parameter estimation by encompassing many common estimation methods such as ordinary least squares (OLS), maximum likelihood estimator (MLE), and instrumental variables. We can estimate the parameters by two-step efficient Generalized Method of Moments (GMM) of Hansen (1982). The GMM is an important tool in econometrics, finance, accounting, and strategic planning literature as well. In this paper we are concerned about model selection in GMM when the number of parameters diverges. These situations can arise in labor economics, international finance (see Alfaro, Kalemli-Ozcan, Volosovych (2008)), and so on. In linear models when the some of the regressors are correlated with errors and there are a large number of covariates, the model selection tools are essential, since they can improve finite sample performance of the estimators.

*mcaner@ncsu.edu. .

Model selection techniques are very useful and widely used in statistics. For example Tibshirani (1996) propose the lasso method, Knight and Fu (2000) derive the asymptotic properties of the lasso, and Fan and Li (2001) propose the SCAD estimator. In econometrics, Knight (2008), and Caner (2009) offer Bridge-least squares and Bridge-GMM estimators respectively. But these procedures all consider finite dimensions and do not take into account the collinearity among variables. Recently model selection with a large number of parameters has been analyzed in least squares by Huang, Horowitz, and Ma (2008), and Zou and Zhang (2009), where the first article analyzes the Bridge estimator, and the second paper is concerned with the adaptive elastic net estimator.

Adaptive elastic net estimator has the oracle property when the number of parameters diverges with the sample size. Furthermore, this method can handle the collinearity arising from a large number of regressors when the system is linear with endogenous regressors. When some of the parameters are redundant (i.e. when the true model has a sparse representation), this estimator can estimate the zero parameters as exactly zero.

In this paper we extend the least squares based adaptive elastic net of Zou and Zhang (2009) to GMM. The following issues are pertinent to model selection in GMM: (i) handling a large number of control variables in the structural equation in a simultaneous equation system, or a large number of parameters in a nonlinear system with endogenous and control variables; (ii) taking into account correlation among variables; (iii) achieving selection consistency and estimation efficiency simultaneously. All of these are successfully addressed in this work. In the least squares case of Zou and Zhang (2009), they do not need an explicit consistency proof since the least squares estimator has a simple and closed form solution. However, in this paper, since the GMM estimator does not have a closed form solution, an explicit consistency proof is needed before deriving the finite sample risk bounds. This is one major contribution of this paper. Furthermore, in order to get a consistency proof, we have substantially extended the technique used in the consistency proof of Bridge least squares estimator of Huang, Horowitz and Ma (2008) to the GMM with adaptive elastic net penalty. To derive the finite sample risk bounds, we use the mean value theorem and benefit from consistency proof, unlike the least squares case of Zou and Zhang (2009). The nonlinear nature of the functions introduces additional difficulties. The GMM case involves partial derivatives of the sample moments which depend on parameter estimates. This is unlike the least squares case where the same quantity does not depend on parameter estimates. This results in the need for consistency proof as mentioned above. Also we extend Zou and Zhang (2009) to conditionally heteroscedastic data, and this results in tuning parameter for l_1 norm to be larger than the one in least squares case. We also pinpoint ways to handle stationary time series cases. The estimator also has the oracle property, and the nonzero coefficients are estimated converging to a normal distribution. This is their standard limit and furthermore, the zero parameters are estimated as zero. Note that the oracle property is a pointwise criterion.

Earlier works on diverging parameters include Huber (1988) and Portnoy (1984), He and Shao (2000). In recent years, there are a few works on penalized methods for standard linear regression with diverging parameters. Fan and Peng (2004) study the nonconcave penalized likelihood with a growing number of nuisance parameters, Lam and Fan (2007) analyze the profile likelihood ratio inference with a growing number of parameters, and Huang et al. (2008) study asymptotic properties of bridge estimators in sparse linear regression models. As far as we know, this is the first paper to estimate and select the model in GMM with a diverging number of parameters. In econometrics, sieve estimation will be a natural application of shrinkage estimators. There are several articles that use sieves, Ai and Chen (2003), Newey and Powell (2003), Chen (2007), Chen and Ludvigson (2009). In these articles, we see that sieve dimension is determined by trying several possibilities or left for

future work. Adaptive elastic net GMM can simultaneously determine the sieve dimension and estimate the structural parameters. We also see in unpenalized GMM with many parameters case, there is a paper by Han and Phillips (2006). Liao (2011) considers adaptive lasso with fixed number of invalid moment conditions.

Section 2 presents the model and the new estimator. Then in Section 3 we derive the asymptotic results for the proposed estimator. Section 4 conducts simulations. Section 5 provides an asset pricing example used in Chen and Ludvigson (2009). Section 6 concludes. Appendix includes all the proofs.

2 Model

Let β be a p -dimensional parameter vector, where $\beta \in B_p$ which is a compact subset in R^p . The true value of β is β_0 . We allow p to grow with the sample size n , so when $n \rightarrow \infty$, we have $p \rightarrow \infty$, but $p/n \rightarrow 0$ as $n \rightarrow \infty$. We do not provide a subscript of n for parameter space not to burden ourselves with the notation. The population orthogonality conditions are

$$E[g(X_i, \beta_0)] = 0,$$

where the data are $\{X_i : i = 1, 2, \dots, n\}$, $g(\cdot)$ is a known function, and the number of orthogonality restrictions is q , $q \leq p$. So we also allow q to grow with the sample size n , but $q/n \rightarrow 0$ as $n \rightarrow \infty$. From now on, we denote $g(X_i, \beta)$ as $g_i(\beta)$ for simplicity. Also assume that $g_i(\beta)$ are independent, and we do not use $g_{ni}(\beta)$ just to simplify the notation.

2.1 The Estimators

We first define the estimators that we use. The estimators that we are interested in aim to answer the following questions. If we have a large number of control variables, some of which may be irrelevant (we may have also a large number of endogenous variables and control variables) in the structural equation in a simultaneous equation system or a large number of parameters in a nonlinear system with endogenous and control variables, can we select the relevant ones as well as estimate the selected system simultaneously? If we have a large number of variables among which there may be possibly some correlation among the variables, can this method handle that? Is it also possible for the estimator to achieve the oracle property? The answers to all three questions are affirmative. First of all, the adaptive elastic net estimator simultaneously selects and estimates the model when there are a large number of parameters/regressors. It can also take into account the possible correlation among the variables. By achieving the oracle property, the nonzero parameters are estimated with their standard limits, and the zero ones are estimated as exactly zero. This method is computationally easy and uses data dependent methods to set small coefficient estimates to zero. A subcase of the adaptive elastic net estimator is the adaptive lasso estimator which can handle the first and third questions but does not handle correlation among a large number of variables.

First we introduce the notation: we use the following norms for the vector $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, $\|\beta\|_2^2 = \sum_{j=1}^p |\beta_j|^2$, also $\|\beta\|_{2+l}^{2+l} = \sum_{j=1}^p |\beta_j|^{2+l}$, where $l > 0$ is a positive number. For a matrix A , the norm is $\|A\|_2^2 = \text{tr}(A' A)$. We start by introducing the adaptive elastic net estimator, given the positive and diverging tuning parameters λ_1 , λ_1^* , λ_2 (how to choose them in finite samples and its asymptotic properties will be discussed below in Assumptions and then in Simulation Section)

$$\hat{\beta}_{aenet} = (1 + \lambda_2/n) \left\{ \arg \min_{\beta \in B_p} \left[\left(\sum_{i=1}^n g_i(\beta) \right)' W_n \left(\sum_{i=1}^n g_i(\beta) \right) + \lambda_2 \|\beta\|_2^2 + \lambda_1^* \sum_{j=1}^p \hat{w}_j |\beta_j| \right] \right\}, \quad (1)$$

where $\hat{w}_j = \frac{1}{|\hat{\beta}_{enet}|^\gamma}$, $\hat{\beta}_{enet}$ is a consistent estimator immediately explained below, γ is a positive constant, and $p = n^\alpha$, $0 < \alpha < 1$. The assumption on γ will be explained in detail in Assumption 3(iii). W_n is a $q \times q$ weight matrix that will be defined in Assumptions below.

The elastic net estimator, which is used in the weights of the penalty above, is

$$\hat{\beta}_{enet} = (1 + \lambda_2/n) \left\{ \arg \min_{\beta \in B_p} S_n(\beta) \right\},$$

where

$$S_n(\beta) = \left[\sum_{i=1}^n g_i(\beta) \right]' W_n \left[\sum_{i=1}^n g_i(\beta) \right] + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1, \quad (2)$$

λ_1, λ_2 are positive and diverging sequences that will be defined in Assumption 5.

We now discuss the penalty functions in both estimators and explain why we need $\hat{\beta}_{enet}$. The elastic net estimator has both l_1 and l_2 penalties. The l_1 penalty is used to perform automatic variable selection, and the l_2 penalty is used to improve the prediction and handles the collinearity that may arise with a large number of variables. However, the standard elastic net estimator does not provide the oracle property. It turns out that, by introducing an adaptive weight in the elastic net, we can obtain the oracle property. The adaptive weights play crucial roles, since they provide data dependent penalization.

An important point to remember is when we set $\lambda_2 = 0$ in the adaptive elastic net estimator (1), we obtain the adaptive lasso estimator. This is simple and we can also get the oracle property. However, with a large number of parameters/variables which may be highly collinear, an additional ridge-type penalty as in adaptive elastic net offers estimation stability and better selection. Before the assumptions we introduce the following notations. Let the collection of nonzero parameters be the set $A = \{j : |\beta_{j0}| > 0\}$ and denote the absolute value of the minimum of the nonzero coefficients as $\eta = \min_{j \in A} |\beta_{j0}|$. Also the cardinality of A is p_A (the number of nonzero coefficients). We now provide the main assumptions.

Assumptions—

1. Define the following $q \times p$ matrix $\hat{G}_n(\beta) = \frac{\sum_{i=1}^n \partial g_i(\beta)}{\partial \beta'}$. Assume the following uniform law of large numbers

$$\sup_{\beta \in B_p} \|\hat{G}_n(\beta)/n - G(\beta)\|_2^2 \xrightarrow{P} 0,$$

where $G(\beta)$ is continuous in β and has a full column rank p . Also $\beta \in B_p \subset \mathbb{R}^p$, B_p is compact and all the absolute value of individual components of the vector β , are

uniformly bounded by a constant a , $|\beta_{j0}|, 0 < a < \infty, j = 1, \dots, p$. Note that

specifically $\sup_{\beta \in B_p} \left\| \frac{1}{2} \sum_{i=1}^n E \frac{\partial g_i(\beta)}{\partial \beta'} - G(\beta) \right\|_2^2 \rightarrow 0$, defines $G(\beta)$.

2. W_n is a symmetric, positive definite matrix, and $\|W_n - W\|_2^2 \xrightarrow{P} 0$, where W is finite and positive definite matrix as well.
3. (i). $\left\| \left[n^{-1} \sum_{i=1}^n E g_i(\beta_0) g_i(\beta_0)' \right]^{-1} - \Omega^{-1} \right\|_2^2 \rightarrow 0$.
 (ii). Assume $p = n^\alpha, 0 < \alpha < 1, p/n \rightarrow 0, q/n \rightarrow 0$ as $n \rightarrow \infty$, and $p, q \rightarrow \infty, q \ll p$, so $q = n^\nu, 0 < \nu < 1, \alpha > \nu$.
 (iii). The coefficient on the weights: γ satisfies the following bound $\gamma > \frac{2+\alpha}{1-\alpha}$.
4. Assume that

$$\max_i \frac{E \|g_i(\beta_{A,0})\|_{2+l}^{2+l}}{n^{l/2}} \rightarrow 0,$$

for $l > 0$, and where $\beta_{A,0}$ represents the true values of the nonzero parameters and it is of p_A dimension. The dimension also increases with the sample size; $p_A \rightarrow \infty$, as $n \rightarrow \infty$, and $0 < p_A \ll p$.

5. (i). $\lambda_1/n \rightarrow 0, \lambda_2/n \rightarrow 0, \lambda_1^*/n \rightarrow 0$
 (ii).

$$\frac{\lambda_1^*}{n^{3+\alpha}} n^{\gamma(1-\alpha)} \rightarrow \infty, \quad (3)$$

(iii).

$$n^{1-\nu} \eta^2 \rightarrow \infty. \quad (4)$$

(iv).

$$n^{1-\alpha} \eta^\gamma \rightarrow \infty. \quad (5)$$

(v). Set $\eta = O(n^{-m}), 0 < m < \alpha/2$.

Now we provide some discussions on the above assumptions. Most of them are standard and used in other papers which establish asymptotic results for penalized estimation in the context of diverging parameters. The rest of them are typically used for the GMM to deal with nonlinear equation systems with endogenous variables. Since $p \rightarrow \infty$, Assumption 1 can be thought of uniform convergence over sieve spaces B_p . For iid subcase, primitive conditions are available in Condition 3.5M in Chen (2007).

Assumptions 1-2 are standard in the literature of GMM (Newey and Windmeijer, 2009, Chen 2007). They are similar to Assumptions 7 and 3 in Newey and Windmeijer (2009a). It is also important to see that Assumption 2 is needed for the two-step nature of the GMM problem. In the first step we can use any consistent estimator (i.e. elastic net) and substitute

in $W_n = n^{-1} \sum_{i=1}^n g_i(\hat{\beta}_{enet}) g_i(\hat{\beta}_{enet})'$ in the adaptive elastic net estimation, where $\hat{\beta}_{enet}$ is the elastic net estimator. Also note that with different estimators, we can define the limit weight W differently. Depending on the estimators W_n , W can change. Assumption 3 provides a definition of variance covariance matrix, and then establishes that the number of diverging parameters cannot exceed the sample size. This is also used in Zou and Zhang (2009). For the penalty exponent in the weight, our condition is more stringent than in the least squares case of Zou and Zhang (2009). This is needed for model selection for local to zero coefficients in GMM format. Assumption 4 is a primitive condition for the triangular array central limit theorem. This is also restraining the number of orthogonality conditions q .

The main issue are the tuning parameter assumptions that reduce bias. We first compare with Bridge estimator of Knight and Fu (2000), there in Theorem 3, they need $\lambda/n^{a/2} \rightarrow \lambda_0$, where $0 < a < 1$, and λ represented the only tuning parameter. In our Assumption 5(i), we need $\lambda_1/n \rightarrow 0$, $\lambda_1^*/n \rightarrow 0$, $\lambda_2/n \rightarrow 0$, so ours can be larger than the Knight and Fu (2000) estimator, we can penalize more in our case. This is due to Bridge type of penalty, which requires less penalization in order to reduce bias and get the oracle property. Theorem 2 of Zou (2006) for the adaptive lasso in least squares assumes $\lambda/n^{1/2} \rightarrow 0$. We think the reason that the GMM estimator requires large penalization is due to its complex model nature, since there are more elements that contribute to bias here. Theorem 1 of Gao and Huang (2010) display the same tuning analysis as Zou (2006).

The rates on λ_1 , λ_2 are standard, but the rate of λ_1^* depends on α and γ . The conditions on λ_1 , λ_1^* , λ_2 are needed for consistency and the bounds on the moments of the estimators. We also allow for local to zero (nonzero) coefficients, but Assumptions 3 and 5, (4)(5) restrict their magnitude. This is tied to the possible number of nonzero coefficients and seeing that α

ν . If there are too many nonzero coefficients (α near 1), then for model selection purpose, the coefficients should slowly approach zero. If there are few nonzero coefficients, to give an example (α near 0), then the order of η should be slightly larger than $n^{-1/2}$. This also confirms and extends Leeb and P tscher (2005) finding that local to zero coefficients should be larger than $n^{-1/2}$ in order to be differentiated from zero coefficients. This is shown in Proposition A.1(2) in their paper. Our results extend that result to the diverging parameter case. Assumption 5(iii) is needed to get consistency for local to zero coefficients. Assumptions 5(iv)(v) are needed for model selection consistency of local to zero coefficients.

To be specific about implications of Assumptions 5(iii)(iv)(v) on the order of η , since $\eta = O(n^{-m})$, then Assumption 5(iii) implies that

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

and Assumption 5(iv) implies that

$$\frac{1-\alpha}{\gamma} > m.$$

Combining these two inequalities with Assumption 5(v), we obtain

$$m < \min \left(\frac{1-\alpha}{\gamma}, \frac{1-\nu}{2}, \frac{\alpha}{2} \right).$$

Now we can see that with a large number of moments, or a large number of parameters, m may get small, so the magnitude of η should be large. To give an example take $\gamma = 5$, $\alpha = 1/3$, $\nu = 2/3$, which gives us an upper bound of $m < 2/15$. So in that scenario, $\eta = O(n^{-2/15})$, to get selected as non zero. It is clear that this is much larger than $n^{-1/2}$ which Leeb and Pötscher (2005) found.

Note also that with $\gamma > \frac{2+\alpha}{1-\alpha}$ (i.e. Assumption 3(iii)), we can assure that the conditions on λ_1^* in Assumptions 5(i) and (ii) are compatible with each other.

Using Assumptions 1, 2, 3(i), we can see that

$$0 < b \leq \text{Eig}_{\min} \left[\left(\hat{G}(\beta)' / n \right) W_n \left(\hat{G}(\beta) / n \right) \right], \quad (6)$$

and

$$\text{Eig}_{\max} \left[\left(\hat{G}(\beta)' / n \right) W_n \left(\hat{G}(\beta) / n \right) \right] \leq B < \infty, \quad (7)$$

with probability approaching one, where $\beta \in [\beta_0, \hat{\beta}_w]$, $B > 0$. These are obtained by Exercise 8.26b of Abadir and Magnus (2005), and Lemma A0 of Newey and Windmeijer (2009b) result (eigenvalue inequality for increasing number of dimension case). $\hat{\beta}_w$ is related to adaptive elastic net estimator and immediately defined below. Here $\text{Eig}_{\min}(M)$ and $\text{Eig}_{\max}(M)$ respectively represent the minimal and maximal eigenvalue of a generic matrix M .

3 Asymptotics

We define an estimator which is related to the adaptive elastic net estimator in (1) and also used in the risk bound calculations.

$$\hat{\beta}_w = \arg \min_{\beta \in B_p} \left[\left(\sum_{i=1}^n g_i(\beta) \right)' W_n \left(\sum_{i=1}^n g_i(\beta) \right) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| \right], \quad (8)$$

The following theorem provides consistency for both the elastic net estimator and $\hat{\beta}_w$.

Theorem 1

Under Assumptions 1-3, 5

(i).

$$\|\hat{\beta}_{enet} - \beta_0\|_2^2 \xrightarrow{P} 0.$$

(ii).

$$\|\hat{\beta}_w - \beta_0\|_2^2 \xrightarrow{P} 0.$$

Remark—It is clear from Theorem 1(ii) that adaptive elastic net estimator in (1) is also consistent. We should note that in Zou and Zhang (2009) where the least squares adaptive elastic net estimator is studied, there is no explicit consistency proof. This is due to using a simple linear model. However, for the GMM adaptive elastic net estimator we have the partial derivative of $g(\cdot)$ which depends on estimators unlike in the linear model case. Specifics are in equations (45)-(50). Therefore we need a new and different consistency proof compared with the least squares case. We need to introduce an estimator that is closely tied to the elastic net estimator above.

$$\hat{\beta}(\lambda_2, \lambda_1) = \arg \min_{\beta \in B_p} S_n(\beta), \quad (9)$$

where $S_n(\beta)$ is defined in (2). This is also the estimator we get when we set for all j , $\hat{w}_j=1$ in $\hat{\beta}_w$. Next, we provide bounds for our estimators. These are used then in the proofs for oracle property, and the limits of the estimators.

Theorem 2

Under Assumptions 1-3, 5

$$E\left(\|\hat{\beta}_w - \beta_0\|_2^2\right) \leq 4 \frac{\lambda_2^2 \|\beta_0\|_2^2 + n^3 p B + \lambda_1^2 E \sum_{j=1}^p \hat{w}_j^2 + o(n^2)}{[n^2 b + \lambda_2 + o(n^2)]^2},$$

and

$$E\left(\|\hat{\beta}(\lambda_2, \lambda_1) - \beta_0\|_2^2\right) \leq 4 \frac{\lambda_2^2 \|\beta_0\|_2^2 + n^3 p B + \lambda_1^2 p + o(n^2)}{[n^2 b + \lambda_2 + o(n^2)]^2}.$$

Remark—Note that the first bound is related to the estimator in (8). The second bound is related to the estimator in (9). $\hat{\beta}_w$ is related to the adaptive elastic net estimator in (1), and $\hat{\beta}(\lambda_1, \lambda_2)$ is related to the estimator in (2). Even though $\|\beta_0\|_2^2 = O(p)$, and $p \rightarrow \infty$, the bound depends on $\lambda_2^2 \|\beta_0\|_2^2 / n^2 \rightarrow 0$ in large samples by Assumptions 3(ii) and 5. Also $\lambda_1^2 E \sum_{j=1}^p \hat{w}_j^2$, is dominated by n^4 in the denominator in large samples as seen in the proof of Theorem 3(i) below. It is clear from the last result that the elastic net estimator is converging at the rate of $\sqrt{n/p}$.

Theorem 2 extends the least squares case of Theorem 3(i) in Zou and Zhang (2009) to the nonlinear GMM case. The risk bounds are different from their case due to the nonlinear nature of our problem. The partial derivative of the sample moment depends on parameter estimates in our case which complicates the proofs.

Write $\beta_0 = (\beta'_{A,0}, 0'_{p-p_A})'$ where $\beta_{A,0}$ represents the vector of nonzero parameters (true values) Its dimension grows with the sample size, and 0_{p-p_A} vector of $p - p_A$ elements

represents the zero (redundant) parameters. Let β_A represent the nonzero parameters and of dimension $p_A \times 1$.

Then define

$$\tilde{\beta} = \arg \min_{\beta} \left\{ \left[\sum_{i=1}^n g_i(\beta_A) \right]' W_n \left[\sum_{i=1}^n g_i(\beta_A) \right] + \lambda_2 \sum_{j \in A} \beta_j^2 + \lambda_1^* \sum_{j \in A} \hat{w}_j |\beta_j| \right\},$$

where $A = \{j : \beta_{j0} \neq 0, j = 1, 2, \dots, p\}$. Our next goal is to show that, with probability one, $\left[(1 + \lambda_2/n) \tilde{\beta}, 0_{p-p_A} \right]$ converges to the solution of the adaptive elastic net estimator in (1).

Theorem 3

Given Assumptions 1-3, and 5,

- (i). with probability tending to one, $\left(\left(1 + \frac{\lambda_2}{n} \right) \tilde{\beta}, 0 \right)$ is the solution to (1).
- (ii). (Consistency in Selection) we also have

$$P \left(\left\{ j : \hat{\beta}_{aenet,j} \neq 0 \right\} = A \right) \rightarrow 1.$$

Remarks—1. Theorem 3(i) shows that ideal estimator $\tilde{\beta}$ becomes the same as the adaptive elastic net estimator in large samples. So the GMM elastic adaptive net estimator has the same solution as $\left((1 + \lambda_2/n) \tilde{\beta}, 0_{p-p_A} \right)$. Theorem 3(ii) shows that the nonzero adaptive elastic net estimates display the oracle property with Theorem 4 below. This is a sharper result than the one in Theorem 3(i). This is an important extension of the least squares case of Theorems 3.2, 3.3 of Zou and Zhang (2009) to the GMM estimation.

2. We allow for local to zero parameters and also provide an assumption when they may be considered as nonzero. This is Assumption 5 (iii)-(iv), $n^{1-\nu} \eta^2 \rightarrow \infty$, $n^{1-\alpha} \eta^\gamma \rightarrow \infty$ where $q = n^\nu$, $p = n^\alpha$, $0 < \alpha < \nu < 1$. The implications of the assumptions on the magnitude of the smallest nonzero coefficient is discussed after Assumptions. The proof of Theorem 3(ii) clearly shows that, as long as Assumption 5 is satisfied, the model selection for local to zero coefficients is possible. However, the local to zero coefficients cannot be arbitrarily close to zero in order to be selected. This is well established in Leeb and Pötscher (2005). Leeb and Pötscher (2005) show, in their Proposition A1(2), as long as the order of local to zero coefficients is larger than $n^{-1/2}$ in magnitude they can be selected. So this is like a lower bound for nonzeros to be selected as zeros. Our Assumption 5 is the extension of their result to the GMM estimator with a diverging number of parameters. In the diverging parameter case, there is a tradeoff between the number of local to zero coefficients and the requirement on their order of the magnitude.

Now we provide the limit law for the estimates of the nonzero parameter values (true values). Denote the adaptive elastic net estimators that correspond to nonzero true parameter values as the vector $\hat{\beta}_{aenet,A}$, which is of dimension $p_A \times 1$. Define a consistent variance

estimator for nonzero parameters that can be derived from elastic net estimators as: $\hat{\Omega}_*$. We

also define Ω_*^{-1} as $\left\| \left[n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})' \right]^{-1} - \Omega_*^{-1} \right\|_2^2 \rightarrow 0$.

Theorem 4

Under Assumptions 1-5, given $W_n = \hat{\Omega}_*^{-1}$, set $W = \Omega_*^{-1}$

$$\delta' K_n \left[\hat{G}(\hat{\beta}_{aenet,A})' \hat{\Omega}_*^{-1} \hat{G}(\hat{\beta}_{aenet,A}) \right]^{1/2} n^{-1/2} (\hat{\beta}_{aenet,A} - \beta_{A,0}) \xrightarrow{d} N(0, 1),$$

where $K_n = \left[\frac{I + \lambda_2 \left(\hat{G}(\hat{\beta}_{aenet,A})' \hat{\Omega}_*^{-1} \hat{G}(\hat{\beta}_{aenet,A}) \right)^{-1}}{1 + \lambda_2/n} \right]$ is a square matrix of dimension p_A and δ is a vector of Euclidean norm 1.

Remarks—1. First we see that

$$\|K_n - I_{p_A}\|_2^2 \xrightarrow{P} 0,$$

due to Assumptions 1, 2, and $\lambda_2 = o(n)$.

2. This theorem clearly extends Zou and Zhang (2009) from the least squares case to the GMM estimation. This result generalizes theirs to nonlinear functions of endogenous variables which are heavily used in econometrics and finance. The extension is not straightforward, since the new limit result depends on an explicit separate consistency proof unlike the least squares case of Zou and Zhang (2009). This is mainly because the partial derivative of the sample moment function depends on the parameter estimates, which is not shared by the least squares estimator. The limit that we derive also corresponds to the standard GMM limit in Hansen (1982), where the same result was derived for a fixed number of parameters with a well specified model. In this way, Theorem 4 also generalizes the result of Hansen (1982) to the direction of a large number of parameters with model selection.

3. Note that K_n term is a ridge regression like term which helps to handle the collinearity among the variables.

4. Note that if we set $\lambda_2 = 0$, we obtain the limit for adaptive Lasso GMM estimator. In that case $K_n = I_{p_A}$, and

$$\delta' \left(\hat{G}(\hat{\beta}_{alasso,A})' \hat{\Omega}_*^{-1} \hat{G}(\hat{\beta}_{alasso,A}) \right)^{1/2} n^{-1/2} (\hat{\beta}_{alasso,A} - \beta_{A,0}) \xrightarrow{d} N(0, 1),$$

There will be discussions on how to choose the tuning parameters $\lambda_1, \lambda_2, \lambda_1^*$, and how to set the small parameter estimates to zero in finite samples in the simulation section.

5. Instead of Liapounov Central Limit Theorem, we can use Central Limit Theorem for stationary time series data. These already exist in Davidson (1994). Theorem 4 will proceed

as before in the independent data case. When we define the GMM objective function, use sample moments as weighted in time. We conjecture that this result in same proofs for Theorems 1-3. This technique of weighting sample moments by time is used in Otsu (2006) and Guggenberger and Smith (2008).

6. After obtaining the adaptive elastic net GMM results, one can run the unpenalized GMM with nonzero parameters and conduct inference.

7. First from Remark 1, we have $\|K_n - I_{p_A}\|_2^2 = o_p(1)$. Then $\|\delta\|_2 = (\delta_1^2 + \dots + \delta_{p_A}^2)^{1/2} = 1$, and δ is a p_A vector. And then by Assumption 1 and consistency of the adaptive elastic net we have $\|\hat{G}(\hat{\beta}_{aenet,A})\|_2 = O_p(n^{1/2})$. These provide the rate of $\sqrt{n/p_A}$ for the adaptive elastic net estimators.

4 Simulation

In this section we analyze the finite sample properties of the adaptive elastic net estimator for GMM. Namely, we evaluate its bias, the root mean squared error as well as the correct number of redundant versus relevant parameters. We have the following simultaneous equations for all $i = 1, \dots, n$

$$\begin{aligned} y_i &= x_i' \beta_0 + \epsilon_i, \\ x_i &= z_i' \pi + \eta_i, \\ \epsilon_i &= \rho \iota' \eta_i + \sqrt{1 - \rho^2} \iota' v_i, \end{aligned}$$

where the number of instruments q is set to be equal to the number of parameters p , x_i is a $p \times 1$ vector, z_i is a $p \times 1$ vector, $\rho = 0.5$, and π is a square matrix of dimension p . Furthermore, η_i is iid $N(0, I_p)$, v_i is iid with $N(0, I_p)$, and I is a $p \times 1$ vector of ones.

The model that is estimated:

$$E z_i \epsilon_i = 0,$$

for all $i = 1, \dots, n$.

We have two different designs for the parameter vector β_0 . In the first case $\beta_0 = \{3, 3, 0, 0, 0\}$ (Design 1), and in the second one $\beta_0 = \{3, 3, 3, 3, 0\}$ (Design 2). We have $n = 100$, and $z_i \approx N(0, \Omega_z)$ for all $i = 1, \dots, n$, and

$$\Omega_z = \begin{bmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

So there is correlation between z_i 's and this affects the correlation between x_i 's since two equations are correlated. In this section, we compare three methods: GMM-BIC of Andrews and Lu (2001), Bridge-GMM of Caner (2009), and the adaptive elastic net GMM. We use four different measures to compare them here. First, we look at the percentage of correct models selected. Then we evaluate the following summary Mean Squared Error (MSE)

$$E \left[\left(\hat{\beta} - \beta_0 \right)' \Sigma_{\epsilon} \left(\hat{\beta} - \beta_0 \right) \right], \quad (10)$$

where $\Sigma_{\epsilon} = E \epsilon_i \epsilon_i'$, and $\hat{\beta}$ represents the estimated coefficient vector given by three different methods. This measure is commonly used in statistics literature; see Zou and Zhang (2009). The other two measures are concerned about individual coefficients. First, the bias of each individual coefficient estimate is measured. Then the root mean squared error of each coefficient is computed. We use 10,000 iterations.

Truncation of small coefficient estimates is set to zero via $|\hat{\beta}_{Bridge}| < 2/\lambda$ for Bridge-GMM as suggested in Caner (2009). For the adaptive elastic net, we use the modified shooting algorithm given in Appendix 2 of Zhang and Lu (2007). Least Angle Regression (LAR) is not used because it is not clear whether it is useful in the GMM context.

This modified shooting algorithm amounts to using Kuhn-Tucker conditions for a corner solution. First, the absolute value of the partial derivative of the GMM objective (unpenalized one) with respect to the parameter of interest is evaluated at zero for that parameter, and for the rest at the current adaptive elastic net estimates. If this is less than $\lambda_1^*/|\hat{\beta}_{enet}|^{4.5}$, then we set that parameter to zero. We have also tried slightly larger exponents than 4.5, and observed that the results are not affected much. Note that the reason for a large γ comes from Assumption 3(iii). This is similar to the adaptive lasso case used in Zhang and Lu (2007).

The choice of λ 's in both Bridge-GMM and the adaptive elastic net GMM is done via BIC. This is suggested by Zou, Hastie, and Tibshirani (2007), and as well as by Wang, Li, and Tsai (2007). Specifically, we use the following BIC from Wang, Li, Leng (2009). For each pair of $\lambda_s = (\lambda_1^*, \lambda_2) \in \Lambda$

$$BIC(\lambda_s) = \log(SSE) + |A| \frac{\log n}{n},$$

where $|A|$ is the cardinality of the set A . $SSE = \left[n^{-1} \sum_{i=1}^n g_i(\hat{\beta}) \right]' W_n \left[n^{-1} \sum_{i=1}^n g_i(\hat{\beta}) \right]$. Basically, given a specific λ_s , we analyze how many nonzero coefficients are in the estimator and use this to calculate the cardinality of A , and for that choice compute SSE . The final λ is chosen as

$$\hat{\lambda}_s = \arg \min_{\lambda_s \in \Lambda} BIC(\lambda_s),$$

where Λ represents a finite number of possible values of λ_s .

The Bridge-GMM estimator in Caner (2009) is $\hat{\beta}$ that minimizes $U_n(\beta)$, where

$$U_n(\beta) = \left[\sum_{i=1}^n g_i(\beta) \right]' W_n \left[\sum_{i=1}^n g_i(\beta) \right] + \lambda \sum_{j=1}^p |\beta_j|^\gamma, \quad (11)$$

for a given positive regularization parameter λ and $0 < \gamma < 1$.

We now describe the model selection by GMM-BIC proposed in Andrews and Lu (2001). Let $b \in R^p$ denote a model selection vector. By definition, each element of b is either zero or one. If the j th element of b is one, the corresponding β_j is to be estimated; if the j th element of b is zero we set β_j to be zero. We set $|b|$ as the number of parameters to be estimated, or in the equivalent form, $|b| = \sum_{j=1}^p |b_j|$. We then set $\beta_{[b]}$ as the $p \times 1$ vector representing the element by the element (Hadamard) product of β and b . The model selection will be based on the GMM objective function and a penalty term. The objective function in BIC benefits from:

$$J_n(b) = \left[\sum_{i=1}^n g_i(\beta_{[b]}) \right]' W_n \left[\sum_{i=1}^n g_i(\beta_{[b]}) \right], \quad (12)$$

where in the simulation

$$g_i(\beta_{[b]}) = z_i (y_i - x_i' \beta_{[b]}).$$

The model selection vectors “ b ” in our case represent 31 different possibilities (excluding the all-zero case). The following are the possibilities for all “ b ” vectors

$$M = [M_{11}, M_{12}, M_{13}, M_{14}, M_{15}],$$

where M_{11} is the identity matrix of dimension 5, I_5 , which represents all the possibilities with only one nonzero coefficient. M_{12} represents all the possibilities with two nonzero coefficients,

$$M_{12} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}. \quad (13)$$

In the same way, M_{13} represents all possibilities with 3 nonzero coefficients, M_{14} represents all the possibilities with four nonzero coefficients, and M_{15} is the vector of ones, showing all nonzero coefficients. The true model in Design 1 is the first column vector in M_{12} . For design 2, the true model is in M_{14} and that is $(1, 1, 1, 1, 0)'$.

The GMM-BIC selects the model based on minimizing the following criterion among the 31 possibilities

$$J_n(b) + |b| \log(n). \quad (14)$$

The penalty term penalizes larger models more. Denote the optimal model selection vector by b^* . After selecting the optimal model in (14), the vector b^* , we then estimate the model parameters by the GMM. Next we present the results on Tables 1-4 for these three techniques that are examined in the simulation section. In Table 1, we provide the correct model selection percentages for Designs 1 and 2. We see that both Bridge and Adaptive Elastic Net are doing very well. The Bridge-GMM selects the correct model 100%, and the Adaptive Elastic Net 91 – 95% of the time, whereas the GMM-BIC selects only 0 – 6.9%. This is due to lots of possibilities in the case of GMM-BIC, and with a large number of

parameters the performance of GMM-BIC tends to deteriorate. Table 2 provides a summary of MSE measure results. This clearly shows that the Adaptive Elastic Net estimator is the best among the three, since its MSE figures are the smallest. The GMM-BIC is much worse in terms of MSE, due to its wrong model selection, and after the model selection estimating the zero coefficients with nonzero and large magnitudes. Tables 3 and 4 provide the bias and root mean squared error (MSE) of each coefficient in Designs 1 and 2. Comparing the Bridge with the Adaptive Elastic Net, we observe that the bias of the nonzero coefficients are generally smaller for the Adaptive Elastic Net. The same is true for generally in the case of root mean squared errors, which are smaller for the nonzero coefficients in the Adaptive Elastic Net estimator.

To get confidence intervals for nonzero parameters, one can run the adaptive elastic net first and find the zero and nonzero coefficients. Then for those nonzero estimates, we have the standard GMM standard errors by Theorem 4. By using that we can calculate confidence intervals for nonzero coefficient parameters.

5 Application

In this part we go through a useful application of the new estimator. The following is the external habit specification model considered by Chen and Ludvigson (2009) (also equation (2.7) in Chen (2007)):

$$E \left(\iota_0 \left(\frac{C_{t+1}}{C_t} \right)^{-\phi_0} \frac{\left(1 - h_0 \left(\frac{C_t}{C_{t+1}} \right) \right)^{-\phi_0}}{\left(1 - h_0 \left(\frac{C_{t-1}}{C_t} \right) \right)^{-\phi_0}} R_{l,t+1} - 1 | z_t \right) = 0,$$

where C_t represents the consumption at time t , and ι_0 and ϕ_0 are both positive and they represent time discount and curvature of the utility function respectively. $R_{l,t+1}$ is the l th asset return at time $t + 1$, $h_0(\cdot) \in [0, 1)$ is an unknown habit formation function, and z_t is the information set and this will be linked to valid instruments. We took only one lag in consumption ratio, rather than several of them. The possibility of this specific model is mentioned in p.1069 of Chen and Ludvigson(2009). Chen and Ludvigson (2009) use the sieve estimation to estimate the unknown h_0 function. They setup the dimension of the sieve as a given number. In this paper we use the adaptive elastic net GMM to automatically select the dimension of sieve and estimate the structural parameters at the same time. The parameters and the unknown habit function that we try to estimate are δ_0 , γ_0 , h_0 . Now denote

$$\rho(C_t, R_{l,t+1}, \iota_0, \phi_0, h_0) = \iota_0 \left(\frac{C_{t+1}}{C_t} \right)^{-\phi_0} \frac{\left(1 - h_0 \left(\frac{C_t}{C_{t+1}} \right) \right)^{-\phi_0}}{\left(1 - h_0 \left(\frac{C_{t-1}}{C_t} \right) \right)^{-\phi_0}} R_{l,t+1} - 1.$$

Before setting up the orthogonality restrictions, set $s_{0j}(z_t)$ is a sequence of known basis functions that can approximate any square integrable function. Then for each $l = 1, \dots, N$, $j = 1, \dots, J_T$, the restrictions are

$$E [\rho(C_t, R_{l,t+1}, \iota_0, \phi_0, h_0) s_{0j}(z_t)] = 0.$$

In total we have NJ_T restrictions, and N is fixed, but $J_T \rightarrow \infty$, as $T \rightarrow \infty$, and $NJ_T/T \rightarrow 0$, as $T \rightarrow \infty$. The main issue is the approximation of the unknown function h_0 . Chen and Ludvigson (2009) use sieves to approximate that function. In theory the dimension of sieve

$K_T \rightarrow \infty$, but $K_T/T \rightarrow 0$, as $T \rightarrow \infty$. As Chen and Ludvigson (2009) we use an artificial neural network sieve approximation,

$$h\left(\frac{C_{t-1}}{C_t}\right) = \zeta_0 + \sum_{j=1}^{K_T} \zeta_j \Psi\left(\tau_j \frac{C_{t-1}}{C_t} + \kappa_j\right),$$

where $\Psi(\cdot)$ is an activation function, and this is chosen as a logistic function $\Psi(x) = (1 + e^{-x})^{-1}$. This implies that in order to estimate the habit function, we need $3K_T + 1$ parameters. The parameters are $\zeta_0, \zeta_j, \tau_j, \kappa_j, j = 1, \dots, K_T$. Chen and Ludvigson (2009) use $K_T = 3$. In our paper, along with parameters δ_0, γ_0 , estimation of h_0 through selection of correct sieve dimension will be done. So if true dimension of sieve is K_{T0} , with $0 < K_{T0} < K_T$, then the Adaptive Elastic Net GMM aims to estimate that dimension (through estimation of parameters in the habit function). The total number of parameters to be estimated is $p = 3(K_T + 1)$, since we also estimate I_0, φ_0 in addition to the habit function parameters. The number of orthogonality restrictions is $q = NJ_T$, and we assume $q = NJ_T - 3(K_T + 1) = p$. Chen and Ludvigson (2009) use sieve minimum distance estimator, and Chen (2007) uses sieve GMM estimator to estimate the parameters. Specifically, equation (2.16) in Chen (2007) use unpenalized sieve GMM estimation for this problem. Instead we will assume that true dimension of the sieve is unknown and will estimate the parameters along with habit function (parameters in that function) with the adaptive elastic net GMM estimator. Set $\beta = (I, \tau, h)$ and the compact sieve space is $B_p = B_\delta \times B_\gamma \times H_T$. The compactness assumption is discussed in p.1067 of Chen and Ludvigson (2009), which is mainly needed so that sieve parameters do not generate tail observations on $\Psi(\cdot)$. Also set the approximating known basis functions as $s(z_t) = (s_{0,1}(z_t), \dots, s_{0,J_T}(z_t))'$, which is a $J_T \times 1$ vector. So $J_T = 3$, there are three instruments. These are a constant, lagged consumption growth, and its square¹. There are seven asset returns that are used in the study so $N = 7$. The detailed explanations can be found in Chen and Ludvigson (2009).

Implementation Details

1. First, we run the elastic net GMM to obtain the adaptive weights w_j 's. The elastic net GMM has the same objective function as the adaptive version but with $w_j = 1$ for all j . The enet-GMM estimator is obtained by setting the weights as 1 in the estimator in the third step given as follows.
2. Then for the weights, since a priori it is known that nonzero coefficients cannot be large positive, we use $\gamma = 2.5$ in the exponent. For specifying the weights, $w_j = 1/|\hat{\beta}_{enet,j}|^{2.5}$ is chosen for all j . We have also experimented with $\gamma = 4.5$ as in the simulations, and the results were mildly different but qualitatively very similar, so those are not reported.
3. Our adaptive elastic net GMM estimator is:

¹There are more instruments that are used by Chen and Ludvigson (2009), but only these three are available to us. Also we thank Sydney Ludvigson to remind us the discrepancy in the unused instruments in her website and the Journal of Applied Econometrics website.

$$\begin{aligned} \hat{\beta} &= (1+\lambda_2/T) \arg \min_{\beta \in B_p} \left\{ \left[\sum_{t=1}^T \rho(C_t, R_{t+1}, \beta) \otimes \mathbf{s}(\mathbf{w}_t) \right]' \widehat{W} \left[\sum_{t=1}^T \rho(C_t, R_{t+1}, \beta) \otimes \mathbf{s}(\mathbf{w}_t) \right] \right. \\ &\quad \left. + \lambda_1^* \sum_{j=1}^{3(K_T+1)} \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^{3(K_T+1)} \beta_j^2 \right\}, \end{aligned}$$

where $\beta_1 = 1$, $\beta_2 = \phi$, and the remaining $3K_T + 1$ parameters correspond to the habit function estimation by sieves. We use the following weight to make the comparison with Chen and Ludvigson (2009) in a fair way:

$$\widehat{W} = I \otimes (S' S)^{-},$$

where S is $(s(z_1), \dots, s(z_t), \dots, s(z_T))'$, which is $T \times 3$ matrix, where we use Moore-Penrose inverse as described in (2.16) of Chen (2007). Note that $\rho(C_t, R_{l,t+1}, \beta)$ is an $N \times 1$ vector, and

$$\rho(C_t, R_{t+1}, \beta) = (\rho(C_t, R_{1,t+1}, \beta), \dots, \rho(C_t, R_{l,t+1}, \beta), \dots, \rho(C_t, R_{N,t+1}, \beta))'.$$

After the implementation steps we explain the data here. The data points start from the second quarter of 1953 and end at the second quarter of 2001. This is a slightly shorter span than Chen and Ludvigson (2009) since we did not want to use missing data cells for certain variables. So $NJ_T = 21$ (number of orthogonality restrictions). At first we try $K_T = 3$ like Chen and Ludvigson (2009), but this is the maximal number of sieve dimension, and we estimate sieve parameter with structural parameters and select the model unlike Chen and Ludvigson (2009). We also try $K_T = 5$. When $K_T = 3$, the total number of parameters to be estimated is 12, and if $K_T = 5$, then this number is 18. We also use BIC to choose from three possible tuning parameter choices, $\lambda_1 = \lambda_1^* = \lambda_2 = \{0.000001, 1, 10\}$. The tuning parameters are taking the same value for ease of computation. So here we will compare our results with unpenalized sieve GMM of Chen (2007). As discussed above we apply a certain subset of the instruments, since the remainder are unavailable, and do not use missing data in Chen and Ludvigson (2009). So our results corresponding to unpenalized sieve GMM will be slightly different than Chen and Ludvigson (2009).

We provide the estimates for our adaptive elastic net GMM method. This is for the case of $K_T = 5$. The time discount estimate $\hat{\tau} = 0.88$ and the curvature of the utility curve parameter is $\hat{\phi} = 0.66$. The sieve parameter estimates are $\hat{\zeta}_0 = 0$, $\hat{\zeta}_j = 0$, $0, 0, 0, 0$, $\hat{\tau}_j = 0.086, 0.078, 0.084, 0.073, 0.083$, $\hat{\kappa}_j = 0.082, 0.072, 0.087, 0.086, 0.080$, for $j = 1, 2, 3, 4, 5$ respectively. To compare if we use sieve GMM with imposing $K_T = 5$ as the true dimension of the sieve, we get $\hat{\tau}_{sg} = 0.86$, $\hat{\phi}_{sg} = 0.73$ for time discount, and curvature parameter, where subscript sg denotes Chen and Ludvigson (2009), Chen (2007) with $\lambda_1 = \lambda_1^* = \lambda_2 = 0$. So the results are basically the same as ours for these two parameters. However, the estimates of sieve parameters in unpenalized sieve GMM case is $\hat{\zeta}_{sg,0} = 0$, $\hat{\zeta}_{sg,j} = 0$ for all $j = 1, \dots, 5$, and $\hat{\tau}_{sg,j} = 0.083, 0.079, 0.079, 0.079, 0.076$, $\hat{\kappa}_{sg,j} = 0.089, 0.086, 0.086, 0.084, 0.082$ respectively for $j = 1, \dots, 5$. So the habit function in the sieve GMM with $K_T = 5$ is estimated as zero (on the boundary), our method gives the same result.

In Chen and Ludvigson (2009), they fit $KT = 3$ for the sieve. We provide the estimates for our adaptive elastic net GMM method in this case as well. We also reestimate Chen and Ludvigson (2009). In the adaptive elastic net GMM, the time discount estimate is $\hat{\gamma}=0.93$ and the curvature of the utility curve parameter is $\hat{\phi}=0.64$. The sieve parameter estimates are $\hat{\zeta}_0=0$, $\hat{\zeta}_j=0$ for $j = 1, 2, 3$, $\hat{\tau}_j=0.057, 0.054, 0.064$, $\hat{\kappa}_j=0.067, 0.066, 0.058$, for $j = 1, 2, 3$ respectively. To compare with our method we use the sieve GMM of Chen and Ludvigson (2009) with imposing $K_T = 3$ as the true dimension of the sieve, we get $\hat{\gamma}_{sg}=0.94$, $\hat{\phi}_{sg}=0.71$ for time discount, and curvature parameter, where subscript sg denotes Chen and Ludvigson (2009), Chen (2007) with $\lambda_1=\lambda_1^*=\lambda_2=0$. So the results are again basically the same as ours for these two parameters. However, the estimates of sieve parameters in unpenalized sieve GMM case is $\hat{\zeta}_{sg,0}=0$, $\hat{\zeta}_{sg,j}=0.022, 0.025, 0.019$ for all $j = 1, \dots, 3$, and $\hat{\tau}_{sg,j}=0.051, 0.055, 0.056$, $\hat{\kappa}_{sg,j}=0.076, 0.075, 0.075$, respectively for $j = 1, \dots, 3$. So the habit function in the adaptive elastic net GMM with $K_T = 3$ is estimated as zero (on the boundary), but with the sieve GMM, the estimate of the habit function is positive. In Chen and Ludvigson article (2009), with a larger instrument set than we used for their case, they find the time parameter estimate to be 0.99, and curvature to be 0.76, and the habit function was positive at $K_T = 3$. Note that we use time series data, and as we suggest after our theorems, this is plausible given our technique and the structure of the proofs.

We now discuss how our assumptions fit this application. First, all of our parameters are uniformly bounded in this application, which is discussed in p.1067 of Chen and Ludvigson (2009). Then the second issue is whether the uniform convergence of partial derivative is plausible. This will be satisfied in iid case through Condition 3.5M in Chen (2007). This amounts to uniformly bounded partial derivatives, Lipschitz continuity of partial derivatives, and log of covering numbers to be growing less than rate T .

Assumption 2 is related to convergence of the weight matrix, which is not restrictive, and it shows a relation between q and T , so q cannot grow fast. In our case $q = NJ_T$, where N is fixed, so this restricts the growth of J_T . Assumption 3(i) is also similar to Assumption 2. For Assumption 3(ii), we have $p = 3(K_T + 1)$, since $q = NJ_T$, $3(K_T + 1) = p$, given that $NJ_T/T \rightarrow 0$ provides us the assumption. Assumption 3(iii) is satisfied here by imposing a value between 0 and 1 (including one) for the exponent in the elastic net based weights. Assumption 4 is concerned about the sample moment functions, since all of our variables are stationary, in terms of ratios, and bounded variables such as returns, we do not expect its $2 + l$ moment to grow larger than $T^{1/2}$. Assumption 5 is a penalty function, and this is satisfied with using T in place of n .

6 Conclusion

In the paper here we analyze the adaptive elastic net GMM estimator. It can simultaneously select the model and estimate that. The new estimator also allows for a diverging number of parameters. The estimator is shown to have the oracle property, so we can estimate nonzero parameters with the standard GMM limit and the redundant parameters are set to zero by a data-dependent technique. Commonly used AIC, BIC methods as well as our estimator face some non uniform consistency issues of the estimators. If we have to select the model with AIC or BIC and then use GMM, this also has the non uniform consistency issues and also does much worse than the adaptive elastic net estimator considered here. The issue with model selection (i.e. including AIC, BIC) is that all of them are uniformly inconsistent (from model selection perspective). So any arbitrarily local to zero coefficients cannot be selected as nonzero. Leeb and Pötscher (2005) establish that, in order to get selected, the order of the magnitude of local to zero coefficients should be larger than $n^{-1/2}$. Between 0 and the

magnitude of $n^{-1/2}$, the model selection is indeterminate. We study the selection issue of local to zero coefficients in the GMM and extend the results of Leeb and P tscher (2005) to the GMM with a diverging number of parameters.

Acknowledgments

Zhang's research is supported by National Science Foundation Grants DMS-0654293, 1347844, 1309507, and National Institutes of Health Grant NIH/NC1 R01 CA-08548. We thank the co-editor Jonathan Wright, an associate editor and two anonymous referees for their comments which have substantially improved the paper. Mehmet Caner also thanks Andres Bredahl Kock for advice on consistency proof.

Appendix

Proof of Theorem 1(i)

Huang, Horowitz and Ma (2008) analyzed the least squares with Bridge penalty with a diverging number of parameters. Here we extend that to a diverging number of moment conditions with parameters and the Adaptive Elastic Net penalty in nonlinear GMM.

Starting with $\hat{\beta}_{enet}$ definition

$$(1+\lambda_2/n) S_n(\hat{\beta}_{enet}) \leq (1+\lambda_2/n) S_n(\beta_0)$$

which is

$$\begin{aligned} & \left[\sum_{i=1}^n g_i(\hat{\beta}_{enet}) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_{enet}) \right] + \lambda_1 \sum_{j=1}^p |\hat{\beta}_{j,enet}| + \lambda_2 \|\hat{\beta}_{enet}\|_2^2 \leq \left[\sum_{i=1}^n g_i(\beta_0) \right]' W_n \left[\sum_{i=1}^n g_i(\beta_0) \right] \\ & + \lambda_1 \sum_{j=1}^p |\beta_{j0}| \\ & + \lambda_2 \|\beta_0\|_2^2. \end{aligned} \quad (15)$$

Then setting $\iota_n = \lambda_1 \sum_{j=1}^p |\beta_{j0}| + \lambda_2 \|\beta_{j0}\|_2^2 = \lambda_1 \sum_{j \in A} |\beta_{j0}| + \lambda_2 \sum_{j \in A} \beta_{j0}^2$, and by (15)

$$\iota_n \geq \left[\sum_{i=1}^n g_i(\hat{\beta}_{enet}) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_{enet}) \right] - \left[\sum_{i=1}^n g_i(\beta_0) \right]' W_n \left[\sum_{i=1}^n g_i(\beta_0) \right] \quad (16)$$

$$= -2 \left[\hat{G}_n(\bar{\beta}) (\beta_0 - \hat{\beta}_{enet}) \right]' W_n \left[\sum_{i=1}^n g_i(\beta_0) \right] + \left[\hat{G}_n(\bar{\beta}) (\beta_0 - \hat{\beta}_{enet}) \right]' W_n \left[\hat{G}_n(\bar{\beta}) (\beta_0 - \hat{\beta}_{enet}) \right], \quad (17)$$

via the mean value theorem and $\bar{\beta} \in (\beta_0, \hat{\beta}_{enet})$. We now try to simplify (17). In this way, set

$$\begin{aligned} \Delta_n &= n W_n^{1/2} \left[\frac{\sum_{i=1}^n G_i(\bar{\beta})}{n} \right] (\beta_0 - \hat{\beta}_{enet}) \quad \text{and} \quad D_n = n^{1/2} W_n^{1/2} \left[\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right] \text{ and} \\ G_i(\bar{\beta}) &= \partial g_i(\bar{\beta}) / \partial \beta' \text{ which is a } q \times p \text{ matrix. The next two equations (18)-(20) are from} \end{aligned}$$

P.603 Huang, Horowitz and Ma (2008). We use them to illustrate our point. Then it is clear that (17) can be written as

$$\Delta_n' \Delta_n - 2D_n' \Delta_n - \iota_n \leq 0, \quad (18)$$

which provides us

$$\|\Delta_n - D_n\|_2^2 - \|D_n\|_2^2 - \iota_n \leq 0.$$

Using the last inequality, we deduct

$$\|\Delta_n - D_n\|_2 \leq \|D_n\|_2 + \iota_n^{1/2},$$

and by triangle inequality

$$\|\Delta_n\|_2 \leq \|\Delta_n - D_n\|_2 + \|D_n\|_2 \leq 2\|D_n\|_2 + \iota_n^{1/2}. \quad (19)$$

By using (18)(19) with simple algebra, we have

$$\|\Delta_n\|_2^2 \leq 6\|D_n\|_2^2 + 3\iota_n. \quad (20)$$

Note

$$nE \left| \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right)' W_n \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right) - nE \left| \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right)' \Omega^{-1} \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right) \right| = o(1), \quad (21)$$

with probability approaching one, by substituting $W = \Omega^{-1}$ and by Assumptions 2-3. Then by the definition of D_n and (21)

$$E\|D_n\|_2^2 - nE \left| \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right)' \Omega^{-1} \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right) \right| = o(1).$$

Next,

$$\begin{aligned} E \left| \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right)' \Omega^{-1} \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right) \right| &= \text{tr} \left(E \left[\left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right) \left(\frac{\sum_{i=1}^n g_i(\beta_0)}{n^{1/2}} \right)' \right] \Omega^{-1} \right) \\ &= \text{tr}(I_q) = q, \end{aligned} \quad (22)$$

Using (22) in (21) with the definition of D_n , with probability approaching one

$$E\|D_n\|_2^2 = O(nq). \quad (23)$$

Next by the definition of $\|\Delta_n\|$ and Assumption 2, we have

$$\begin{aligned}
\|\Delta_n\|_2^2 &= n^2 \left[(\beta_0 - \hat{\beta}_{enet})' \left[\left(\frac{\sum_{i=1}^n G_i(\bar{\beta})}{n} \right)' \Omega^{-1} \left(\frac{\sum_{i=1}^n G_i(\bar{\beta})}{n} \right) \right] (\beta_0 - \hat{\beta}_{enet}) \right] \\
&\geq n^2 \text{Eig}_{\min} \left[\left(\frac{\sum_{i=1}^n G_i(\bar{\beta})}{n} \right)' \Omega^{-1} \left(\frac{\sum_{i=1}^n G_i(\bar{\beta})}{n} \right) \right] \|\hat{\beta}_{enet} - \beta_0\|_2^2 \\
&\geq n^2 b \|\hat{\beta}_{enet} - \beta_0\|_2^2,
\end{aligned} \tag{24}$$

with probability approaching one and seeing (6) with remembering that elastic net is a subcase of $\hat{\beta}_w$. Next use (23)(24) in (20) to have

$$n^2 b E \|\hat{\beta}_{enet} - \beta_0\|_2^2 \leq O(nq) + O(\lambda_1 p_A + \lambda_2 p_A), \tag{25}$$

by seeing that $\lambda_1 \sum_{j \in A} |\beta_{j0}| + \lambda_2 \sum_{j \in A} \beta_{j0}^2 = O(\lambda_1 p_A + \lambda_2 p_A)$, given that nonzero parameters can be all constants at most. So by Assumptions 3 and 5, since $\lambda_1/n \rightarrow 0$, $\lambda_2/n \rightarrow 0$, $p_A/n \rightarrow 0$, $q/n \rightarrow 0$, we have

$$\|\hat{\beta}_{enet} - \beta_0\|_2^2 \xrightarrow{P} 0.$$

Q.E.D.

Proof of Theorem 1(ii)

The proof for the consistency of the estimator $\hat{\beta}_w$ is the same as in the elastic net case in Theorem 1(i). The only difference is the definition of $\iota_n = \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_{j0}| + \lambda_2 \sum_{j=1}^p \beta_{j0}^2$. Note that when $\beta_{j0} = 0$, we can write $\iota_n = \lambda_1 \sum_{j \in A} \hat{w}_j |\beta_{j0}| + \lambda_2 \sum_{j \in A} \beta_{j0}^2$. In other words, only nonzero coefficient weights play a role in term I_n .

As in (25)

$$n^2 b E \|\hat{\beta}_w - \beta_0\|_2^2 = O(nq) + O \left(\lambda_1 \sum_{j \in A} \hat{w}_j |\beta_{j0}| + \lambda_2 \sum_{j \in A} \beta_{j0}^2 \right). \tag{26}$$

The key issue is the order of the penalty terms. We are only interested in nonzero parameter weights. Defining $\hat{\eta} = \min_{j \in A} |\hat{\beta}_{j,enet}|$, we have for $j \in A$,

$$\hat{w}_j = \frac{1}{|\hat{\beta}_{j,enet}|^\gamma} \leq \frac{1}{\hat{\eta}^\gamma}.$$

Then

$$\frac{\lambda_1 \sum_{j \in A} \hat{w}_j |\beta_{j0}|}{n^2} \leq \frac{\lambda_1 \hat{\eta}^{-\gamma} p_A C}{n^2}, \quad (27)$$

where we use $\max_{j \in A} |\beta_{j0}| \leq C$, where C is a generic positive constant. So we can write (27)

$$\frac{\lambda_1 \hat{\eta}^{-\gamma} p_A C}{n^2} = \left[\frac{\lambda_1 \eta^{-\gamma} p_A C}{n^2} \right] \left[\left(\frac{\hat{\eta}}{\eta} \right)^{-\gamma} \right]. \quad (28)$$

In the above equation we first show

$$E \left(\frac{\hat{\eta}}{\eta} \right)^2 = O(1). \quad (29)$$

First see that by simple algebraic inequality as in (6.13) of Zou and Zhang (2009),

$$\begin{aligned} E \left(\frac{\hat{\eta}}{\eta} \right)^2 &\leq 2 + \frac{2}{\eta^2} E(\hat{\eta} - \eta)^2 \\ &\leq 2 + \frac{2}{\eta^2} E \|\hat{\beta}_{enet} - \beta_0\|_2^2. \end{aligned} \quad (30)$$

Next by Theorem 1(i) (equation (25))

$$\frac{1}{\eta^2} E \|\hat{\beta}_{enet} - \beta_0\|_2^2 = O \left(\frac{qn + \lambda_1 p_A + \lambda_2 p_A}{n^2 \eta^2} \right),$$

since $\lambda_1/n \rightarrow 0$, $\lambda_2/n \rightarrow 0$ and $p_A = q$, the largest stochastic order is $qn/n^2 \eta^2 = q/n \eta^2$. By $q = n^\nu$, and $0 < \nu < 1$ with $n^{1-\nu} \eta^2 \rightarrow \infty$ by Assumption 5(iii) clearly

$$\frac{1}{\eta^2} E \|\hat{\beta}_{enet} - \beta_0\|_2^2 = o(1), \quad (31)$$

Next by (30)(31) we have

$$E \left(\frac{\hat{\eta}}{\eta} \right)^2 \leq 2 + o(1).$$

So (29) is shown, and by Markov's inequality

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

clearly $\left(\frac{\hat{\eta}}{\eta} \right)^2 \neq o_p(1)$. Note that

$$\left(\frac{\hat{\eta}}{\eta} \right)^{-\gamma} = \left[\left(\frac{\hat{\eta}}{\eta} \right)^2 \right]^{-\gamma/2}. \quad (32)$$

Then by (32) we have

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

Next on the right hand side of (28)

$$\frac{\lambda_1}{n} \frac{p_A C}{n \eta^\gamma} \rightarrow 0,$$

since $\lambda_1/n \rightarrow 0$ by Assumption 5(i), and by Assumption 5(iv) $n^{1-a} \eta^\gamma \rightarrow \infty$. The last two equations illustrate that (28)

$$\frac{\lambda_1 \hat{\eta}^{-\gamma} p_A C}{n^2} = \left[\frac{\lambda_1 \eta^{-\gamma} p_A C}{n^2} \right] \left[\left(\frac{\hat{\eta}}{\eta} \right)^{-\gamma} \right] \xrightarrow{P} 0. \quad (33)$$

Next in (26) we have

$$\frac{\lambda_2 \sum_{j \in A} \beta_{j0}^2}{n^2} = O\left(\frac{\lambda_2 p_A}{n^2}\right) = o(1), \quad (34)$$

by $\lambda_2/n \rightarrow 0$, $p_A/n \rightarrow 0$. Also note that by $q/n \rightarrow 0$ through Assumption 3 combining (33) (34) in (26) above provides us

$$E\|\hat{\beta}_w - \beta_0\|_2^2 = O\left(\frac{nq}{n^2}\right) + O\left(\frac{\lambda_1 \sum_{j \in A} \hat{w}_j |\beta_{j0}| + \lambda_2 \sum_{j \in A} \beta_{j0}^2}{n^2}\right) = o(1). \quad (35)$$

Q.E.D.

Proof of Theorem 2

In this proof we start by analyzing the GMM-Ridge estimator that is defined as follows:

$$\hat{\beta}_R + \arg \min_{\beta} \left[\sum_{i=1}^n g_i(\beta) \right]' W_n \left[\sum_{i=1}^n g_i(\beta) \right] + \lambda_2 \|\beta\|_2^2.$$

Note that this estimator is similar to the elastic net estimator, if we set $\lambda_1 = 0$, in elastic net estimator, then we obtain the GMM-Ridge estimator. So since the elastic net estimator is consistent, GMM-Ridge will be consistent as well. Define also the following $q \times p$ matrix

$$\hat{G}_n(\hat{\beta}_R) = \frac{\sum_{i=1}^n \partial g_i(\hat{\beta}_R)}{\partial \beta'}. \text{ Then set } \bar{\beta} \in (\beta_0, \hat{\beta}_R). \text{ Note that } \hat{G}_n(\bar{\beta}) \text{ is the value of } \hat{G}_n(\cdot) \text{ evaluated at } \bar{\beta}. \text{ A mean value theorem around } \beta_0 \text{ applied to first order conditions provides, with } g_n(\beta_0) = \sum_{i=1}^n g_i(\beta_0),$$

$$\hat{\beta}_R = - \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right]. \quad (36)$$

Also using the mean value theorem with first order conditions, adding and subtracting $\lambda_2 \beta_0$ from first order conditions yields

$$\hat{\beta}_R - \beta_0 = - \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) + \lambda_2 \beta_0 \right]. \quad (37)$$

We need the following expressions by using (36)

$$\begin{aligned} \hat{\beta}_R' \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right] &= - \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right]' \\ &\times \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \\ &\times \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right]. \end{aligned} \quad (38)$$

$$\begin{aligned} \hat{\beta}_R' \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_R &= \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right]' \\ &\times \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \\ &\times \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right]. \end{aligned} \quad (39)$$

Next the aim is to rewrite the following GMM-Ridge objective function via a mean value expansion

$$\begin{aligned} \left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right] &+ \lambda_2 \|\hat{\beta}_R\|_2^2 = g_n(\beta_0)' W_n g_n(\beta_0) \\ &+ g_n(\beta_0)' W_n \hat{G}_n(\bar{\beta}) (\hat{\beta}_R - \beta_0) + (\hat{\beta}_R - \beta_0)' \hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) \\ &+ (\hat{\beta}_R - \beta_0)' \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) (\hat{\beta}_R - \beta_0) + \lambda_2 \|\hat{\beta}_R\|_2^2. \end{aligned} \quad (40)$$

When $\lambda_1 = 0$, using Theorem 1 we see that $\|\hat{\beta}_R - \bar{\beta}\|_2 \xrightarrow{P} 0$. Then use Assumption 1 to have

$$\begin{aligned} \hat{\beta}_R' \left[\left(\frac{\hat{G}_n(\hat{\beta}_R)}{n} \right)' W_n \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right) + \lambda_2 I_p \right] \hat{\beta}_R &= \hat{\beta}_R' \left[\left(\frac{\hat{G}_n(\bar{\beta})}{n} \right)' W_n \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right) + \lambda_2 I_p \right] \hat{\beta}_R \\ &+ \hat{\beta}_R' \left[o_P(1) W_n \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right) + \lambda_2 I_p \right] \hat{\beta}_R, \end{aligned} \quad (41)$$

where the $o_P(1)$ term comes from the uniform law of large numbers. Clearly the stochastic order of the second term is smaller than the first one on the right hand side of (41). By using the same argument to get (41), we have

$$\begin{aligned} \tilde{\beta}'_R \left[\left(\frac{\hat{G}_n(\hat{\beta}_R)}{n} \right)' W_n g_n(\beta_0) - \left(\frac{\hat{G}_n(\hat{\beta}_R)}{n} \right)' W_n \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right) \beta_0 \right] &= \tilde{\beta}'_R \left[\left(\frac{\hat{G}_n(\bar{\beta})}{n} \right)' W_n g_n(\beta_0) - \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right)' W_n \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right) \beta_0 \right] \\ &+ \tilde{\beta}'_R \left[o_P(1) W_n g_n(\beta_0) - o_P(1) W_n \left(\frac{\hat{G}_n(\bar{\beta})}{n} \right) \beta_0 \right]. \end{aligned}$$

Again the second term's stochastic order is smaller than the first one in (42).

Furthermore we can rewrite the right hand side of (40)

$$\begin{aligned} g_n(\beta_0)' W_n g_n(\beta_0) &+ \tilde{\beta}'_R \left[\hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right] \\ &+ \left[\hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right]' \hat{\beta}_R \\ &+ \tilde{\beta}'_R \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_R \\ &- g_n(\beta_0)' W_n \hat{G}_n(\bar{\beta}) \beta_0 - \beta_0' \hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) + \beta_0' \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \quad (43) \\ &= g_n(\beta_0)' W_n g_n(\beta_0) - \tilde{\beta}'_R \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_R \\ &- g_n(\beta_0)' W_n \hat{G}_n(\bar{\beta}) \beta_0 - \beta_0' \hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) + \beta_0' \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \\ &- \tilde{\beta}'_R so_1 \hat{\beta}_R, \end{aligned}$$

where so_1 represents the small order terms mentioned in (41)(42). The equality is obtained through (38)(39) via Assumption 1 and the consistency of GMM-Ridge since ridge is a subcase of elastic net with $\lambda_1 = 0$ imposed on the elastic net. Note that the right hand side expression in (38) is just the negative of the right hand side of the expression in (39). As in

(43), for the estimator $\hat{\beta}_w$ we have the following where $\bar{\beta}_w \in (\beta_0, \hat{\beta}_w)$,

$$\begin{aligned}
& \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right] + \lambda_2 \|\hat{\beta}_w\|_2^2 = g_n(\beta_0)' W_n g_n(\beta_0) \\
& + \hat{\beta}_w' \left[\hat{G}_n(\bar{\beta}_w)' W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta}_w)' W_n \hat{G}_n(\bar{\beta}_w) \beta_0 \right] \\
& + \left[\hat{G}_n(\bar{\beta}_w)' W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta}_w)' W_n \hat{G}_n(\bar{\beta}_w) \beta_0 \right]' \hat{\beta}_w \\
& + \hat{\beta}_w' \left[\hat{G}_n(\bar{\beta}_w)' W_n \hat{G}_n(\bar{\beta}_w) + \lambda_2 I_p \right] \hat{\beta}_w \\
& - g_n(\beta_0)' W_n \hat{G}_n(\bar{\beta}_w) \beta_0 - \beta_0' \hat{G}_n(\bar{\beta}_w)' W_n g_n(\beta_0) \\
& + \beta_0' \hat{G}_n(\bar{\beta}_w)' W_n \hat{G}_n(\bar{\beta}_w) \beta_0.
\end{aligned} \tag{44}$$

Then see that by Theorem 1 $\|\hat{\beta} - \bar{\beta}_w\|_2^2 \xrightarrow{P} 0$ and using (36)

$$\begin{aligned}
\hat{\beta}_w' \left[\hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right] &= \hat{\beta}_w' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \\
&\times \left[\hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \right] \\
&= -\hat{\beta}_w' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_R - \hat{\beta}_w' so_2 \hat{\beta}_R.
\end{aligned} \tag{45}$$

The term so_2 above comes from the same type of analysis done for the second and small stochastic order terms in (41)(42). Next substitute (45) into (44) to have

$$\begin{aligned}
& \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right] + \lambda_2 \|\hat{\beta}_w\|_2^2 = g_n(\beta_0)' W_n g_n(\beta_0) - \hat{\beta}_w' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_R \\
& - \hat{\beta}_R' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_w + \hat{\beta}_w' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] \hat{\beta}_w \\
& - g_n(\beta_0)' W_n \hat{G}_n(\bar{\beta}) \beta_0 \\
& - \beta_0' \hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) + \beta_0' \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \beta_0 \\
& - \hat{\beta}_w' so_2 \hat{\beta}_R - \hat{\beta}_R' so_3 \hat{\beta}_w + \hat{\beta}_w' so_4 \hat{\beta}_w.
\end{aligned} \tag{46}$$

Term so_3 is the transpose of so_2 , and term so_4 comes from approximation error between $\bar{\beta}$ and $\bar{\beta}_w$. Specifically note that $\hat{\beta}_w' so_2 \hat{\beta}_R$, $\hat{\beta}_R' so_3 \hat{\beta}_w$, $\hat{\beta}_w' so_4 \hat{\beta}_w$ are smaller order terms than the second, third, and fourth terms on the right hand side of (46) respectively. Denote $so_5 = \min(so_1, so_2, so_3, so_4)$. Now subtract (43) from (46) to have

$$\begin{aligned}
\left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right] &+ \lambda_2 \|\hat{\beta}_w\|_2^2 - \left(\left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right] + \lambda_2 \|\hat{\beta}_R\|_2^2 \right) \\
&\geq (\hat{\beta}_w - \hat{\beta}_R)' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] (\hat{\beta}_w - \hat{\beta}_R) \\
&+ (\hat{\beta}_w - \hat{\beta}_R)' so_5(\hat{\beta}_w - \hat{\beta}_R).
\end{aligned} \quad (47)$$

The next analysis is very similar to equations (6.1)-(6.6) of Zou and Zhang (2009). After this important result see that by the definitions of $\hat{\beta}_w$ and $\hat{\beta}_R$

$$\begin{aligned}
\lambda_1 \sum_{j=1}^p \hat{w}_j (|\hat{\beta}_{j,R}| - |\hat{\beta}_{j,w}|) &\geq \left(\sum_{i=1}^n g_i(\hat{\beta}_w) \right)' W_n \left(\sum_{i=1}^n g_i(\hat{\beta}_w) \right) + \lambda_2 \|\hat{\beta}_w\|_2^2 \\
&- \left[\left(\sum_{i=1}^n g_i(\hat{\beta}_R) \right)' W_n \left(\sum_{i=1}^n g_i(\hat{\beta}_R) \right) + \lambda_2 \|\hat{\beta}_R\|_2^2 \right].
\end{aligned} \quad (48)$$

Then also see that

$$\sum_{j=1}^p \hat{w}_j (|\hat{\beta}_{j,R}| - |\hat{\beta}_{j,w}|) \leq \sqrt{\sum_{j=1}^p (\hat{w}_j)^2} \|\hat{\beta}_R - \hat{\beta}_w\|_2. \quad (49)$$

Next benefit from (48), with (47)(49) to have

$$\begin{aligned}
(\hat{\beta}_w - \hat{\beta}_R)' \left[\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right] (\hat{\beta}_w - \hat{\beta}_R) &+ (\hat{\beta}_w - \hat{\beta}_R)' so_5(\hat{\beta}_w - \hat{\beta}_R) \\
&\leq \lambda_1 \sqrt{\sum_{j=1}^p (\hat{w}_j)^2} \|\hat{\beta}_R - \hat{\beta}_w\|_2.
\end{aligned} \quad (50)$$

See that $\text{Eig}_{\min} \left(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right) = \text{Eig}_{\min} \left(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \right) + \lambda_2$. Use this in (50) to have

$$\|\hat{\beta}_w - \hat{\beta}_R\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^p \hat{w}_j^2}}{\text{Eig}_{\min} \left(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \right) + \lambda_2 + so_5}, \quad (51)$$

where $\text{Eig}_{\min} \left(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) \right)$ is a term of larger stochastic order than so_5 which is explained in (46) (43). We also want to modify the last inequality. By the consistency of $\hat{\beta}_w, \hat{\beta}_R, \bar{\beta} \xrightarrow{P} \beta_0$. Then with the uniform law of large numbers on the partial derivative we have by Assumptions 1-3

$$\left\| \left[\frac{\hat{G}_n(\bar{\beta})}{n} \right]' W_n \left[\frac{\hat{G}_n(\bar{\beta})}{n} \right] - G(\beta_0)' W G(\beta_0) \right\|_2 \xrightarrow{P} 0.$$

The last equation is true also for $\hat{\beta}_w, \hat{\beta}_R$ replacing $\bar{\beta}$. Then

$$\left\| \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) - n^2 [G(\beta_0)' W G(\beta_0)] - o(n^2) \right\|_2 \xrightarrow{P} 0. \quad (52)$$

Using Lemma A0 of Newey and Windmeijer (2009b), modify (51) in the following way given the last equality, set $W = \Omega^{-1}$ (since this is the efficient limit weight as shown by Hansen (1982))

$$\|\hat{\beta}_w - \hat{\beta}_R\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^p \hat{w}_j^2}}{n^2 \text{Eig}_{\min}(G(\beta_0)' \Omega^{-1} G(\beta_0)) + \lambda_2 + o_P(n^2)}. \quad (53)$$

Now we consider the second part of the proof of this theorem. We use the GMM ridge formula. Note that from (37)

$$\begin{aligned} \hat{\beta}_R - \beta_0 &= -\lambda_2 \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \beta_0 \\ &\quad - \left[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p \right]^{-1} \left[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) \right]. \end{aligned} \quad (54)$$

We try to modify the equation above a little.

In the same way we obtain (52)

$$\left\| \hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - n [G(\beta_0)' W g_n(\beta_0)] - o_P(n) \right\|_2 \xrightarrow{P} 0. \quad (55)$$

Second, see that by $g_i(\beta)$ being independent $E_{g_n(\beta_0)} g_n(\beta_0)' - n\Omega \rightarrow 0$.

$$\begin{aligned} E \left[g_n(\beta_0)' \Omega^{-1} G(\beta_0) G(\beta_0)' \Omega^{-1} g_n(\beta_0) \right] &= \text{tr} \left\{ G(\beta_0)' \Omega^{-1} E \left[g_n(\beta_0) g_n(\beta_0)' \right] \Omega^{-1} G(\beta_0) \right\} \\ &= n \text{tr} \left\{ G(\beta_0)' \Omega^{-1} G(\beta_0) \right\} + o(1) \\ &\leq n p \text{Eig}_{\max}(\Sigma) + o(1), \end{aligned} \quad (56)$$

where we use $\Sigma = G(\beta_0)' \Omega^{-1} G(\beta_0)$. Now we modify (54) using (55) (52), $W = \Omega^{-1}$.

$$\begin{aligned} \hat{\beta}_R - \beta_0 &= -\lambda_2 \left[n^2 G(\beta_0)' \Omega^{-1} G(\beta_0) + \lambda_2 I_p + o_P(n^2) \right]^{-1} \beta_0 \\ &\quad - \left[n^2 G(\beta_0)' \Omega^{-1} G(\beta_0) + \lambda_2 I_p + o_P(n^2) \right]^{-1} \left[n G(\beta_0)' \Omega^{-1} g_n(\beta_0) + o_P(n) \right]. \end{aligned} \quad (57)$$

Then see that

$$\begin{aligned}
E \left(\|\hat{\beta}_R - \beta_0\|_2^2 \right) &\leq 2\lambda_2^2 \left[n^2 \text{Eig}_{\min} \left(G(\beta_0)' \Omega^{-1} G(\beta_0) \right) + \lambda_2 + o(n^2) \right]^{-2} \|\beta_0\|_2^2 \\
&\quad + 2 \left[n^2 \text{Eig}_{\min} \left(G(\beta_0)' \Omega^{-1} G(\beta_0) \right) + \lambda_2 + o(n^2) \right]^{-2} \\
&\quad \times n^2 E \left[g_n(\beta_0)' \Omega^{-1} G(\beta_0) G(\beta_0)' \Omega^{-1} g_n(\beta_0) \right] + o(n^2) \\
&\leq 2 \left[n^2 \text{Eig}_{\min} \left(G(\beta_0)' \Omega^{-1} G(\beta_0) \right) + \lambda_2 + o(n^2) \right]^{-2} \\
&\quad \times \left[\lambda_2^2 \|\beta_0\|_2^2 + n^3 p \text{Eig}_{\max}(\Sigma) + o(n^2) \right],
\end{aligned} \tag{58}$$

where the last inequality is by (56). Note that we do not use orders smaller than $o(n^2)$ in (58) since this will not make any difference in the proofs of the theorems below. Now use (53) and (58) to have

$$\begin{aligned}
E \left(\|\hat{\beta}_w - \beta_0\|_2^2 \right) &\leq 2E \left(\|\hat{\beta}_R - \beta_0\|_2^2 \right) + 2E \left(\|\hat{\beta}_w - \hat{\beta}_R\|_2^2 \right) \\
&\leq \frac{4\lambda_2^2 \|\beta_0\|_2^2 + 4n^3 p B + o(n^2) + 2\lambda_1^2 E \sum_{j=1}^p \hat{w}_j^2}{[n^2 b + \lambda_2 + o(n^2)]^2}.
\end{aligned} \tag{59}$$

See that $b = \text{Eig}_{\min}(G(\beta_0)' \Omega^{-1} G(\beta_0)) = \text{Eig}_{\min}(\Sigma)$, $B = \text{Eig}_{\max}(\Sigma)$.

Q.E.D

Proof of Theorem 3(i)

The proof is similar to proof of Theorem 3.2 in Zou and Zhang (2009). The differences are due to nonlinear nature of the problem. Our upper bounds in Theorem 2 converge to zero at a different rate than Zou and Zhang (2009). To prove the theorem, we need to show the following (Note that by Kuhn-Tucker conditions of (1)),

$$P \left[\forall j \in A^c, |2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| \leq \lambda_1^* \hat{w}_j \right] \rightarrow 1,$$

where $\hat{G}_n(\tilde{\beta}) = \frac{\sum_{i=1}^n \partial g_i(\tilde{\beta})}{\partial \beta'}$, and $A^c = \{j : \beta_{j0} = 0, j = 1, 2, \dots, p\}$. $\hat{G}_{n,j}(\tilde{\beta})$ denotes the j th column of the partial derivative matrix which corresponds to irrelevant parameters, evaluated at $\tilde{\beta}$. Or we need to show

$$P \left[\exists j \in A^c, |2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* \hat{w}_j \right] \rightarrow 0,$$

Now set $\eta = \min_{j \in A} |\beta_{j0}|$, $\hat{\eta} = \min_{j \in A} |\hat{\beta}_{j,enet}|$, and $A = \{j : \beta_{j0} > 0, j = 1, 2, \dots, p\}$. So

$$\begin{aligned}
P \left[\exists j \in A^c, |2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* \hat{w}_j \right] &\leq \sum_{j \in A^c} P \left[|2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* \hat{w}_j, \hat{\eta} > \eta/2 \right] \\
&\quad + P[\hat{\eta} \leq \eta/2].
\end{aligned} \tag{60}$$

Then as in equation (6.7) of Zou and Zhang (2009), we can show that

$$\begin{aligned}
 P[\hat{\eta} \leq \eta/2] &\leq \frac{E\|\hat{\beta}_{enet}-\beta_0\|_2^2}{\eta^2/4} \\
 &\leq 16 \frac{\lambda_2^2 \|\beta_0\|_2^2 + n^3 p B + \lambda_1^2 p + o(n^2)}{[n^2 b + \lambda_2 + o(n^2)]^2 \eta^2}, \quad (61)
 \end{aligned}$$

where the second inequality is due to Theorem 2. Then we can also have

$$\begin{aligned}
 \sum_{j \in A^c} P \left[|2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* \hat{w}_j, \hat{\eta} > \eta/2 \right] &> \lambda_1^* \hat{w}_j, \hat{\eta} > \eta/2 \\
 &\leq \sum_{j \in A^c} P \left[|2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* \hat{w}_j, \hat{\eta} > \eta/2, |\hat{\beta}_{j,enet}| \leq M \right] \\
 &\quad + \sum_{j \in A^c} P \left[|\hat{\beta}_{j,enet}| > M \right] \\
 &\leq \sum_{j \in A^c} P \left[|2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* M^{-\gamma}, \hat{\eta} > \eta/2 \right] \\
 &\quad + \sum_{j \in A^c} P \left[|\hat{\beta}_{j,enet}| > M \right], \quad (62)
 \end{aligned}$$

where $M = \left(\frac{\lambda_1^*}{n^{3+\alpha}} \right)^{\frac{1}{2\gamma}}$. Compared to Zou and Zhang (2009), M converges to zero faster.

In (62) we consider the second term on the right hand side. Via inequality (6.8) of Zou and Zhang (2009) and Theorem 2, we have

$$\begin{aligned}
 \sum_{j \in A^c} P \left[|\hat{\beta}_{j,enet}| > M \right] &\leq \frac{E\|\hat{\beta}_{enet}-\beta_0\|_2^2}{M^2} \\
 &\leq 4 \frac{\lambda_2^2 \|\beta_0\|_2^2 + 4n^3 p B + \lambda_1^2 p + o(n^2)}{[n^2 b + \lambda_2 + o(n^2)]^2 M^2}. \quad (63)
 \end{aligned}$$

Next we can consider the first term on the right hand side of (62)

$$\begin{aligned}
 \sum_{j \in A^c} P \left[|2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* M^{-\gamma}, \hat{\eta} > \eta/2 \right] &> \lambda_1^* M^{-\gamma}, \hat{\eta} > \eta/2 \\
 &\leq \frac{4M^{2\gamma}}{\lambda_1^{*2}} E \left[\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)|^2 I_{\{\hat{\eta} > \eta/2\}} \right]. \quad (64)
 \end{aligned}$$

So we try to simplify the term on the right hand side of (64). Now we evaluate

$$\begin{aligned}
 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)|^2 &\leq 2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\beta_{A,0}) \right)|^2 \\
 &\quad + 2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \hat{G}_n(\tilde{\beta}) (\tilde{\beta} - \beta_{A,0})|^2, \quad (65)
 \end{aligned}$$

where we have $\tilde{\beta} \in (\beta_{A,0}, \tilde{\beta})$, and

$$g_i(\tilde{\beta}) = g_i(\beta_{A,0}) + \left[\frac{\partial g_i(\tilde{\beta})}{\partial \beta'} \right] (\tilde{\beta} - \beta_{A,0}).$$

Analyze each term in (65). Note that $\tilde{\beta}$ is consistent if we go through the same steps as in Theorem 1 by Assumption 5 on λ_1^* . Then applying Assumption 1 with Assumption 2 (Uniform Law of Large Numbers) in the same way as in equation (6.9) of Zou and Zhang (2009), we have

$$2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'| W_n \left(\sum_{i=1}^n g_i(\beta_{A,0}) \right) \leq 2n^2 \|G(\beta_{A,0})'\Omega_*^{-1} \sum_{i=1}^n g_i(\beta_{A,0})\|_2^2 + o_P(n^2), \quad (66)$$

where we put $W = \Omega_*^{-1}$ and use Assumption 3(i),

$$\left\| \left[n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})' \right]^{-1} - \Omega_*^{-1} \right\|_2^2 \rightarrow 0 \text{ as the definition of the efficient limit weight matrix for the case of nonzero parameters to have}$$

$$E \left[\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'| W_n \left(\sum_{i=1}^n g_i(\beta_{A,0}) \right) \right]^2 \leq n^3 B + o(n^3), \quad (67)$$

where we use

$$B \geq \text{Eig}_{\max}(\Sigma) \geq \text{Eig}_{\max} \left(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right). \quad (68)$$

See that $\sum_{i=1}^n g_i(\beta_{A,0}) = O_P(n^{1/2})$.

In the same manner as in equation (6.9) of Zou and Zhang (2009), we have

$$\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'| W_n \hat{G}_n(\tilde{\beta}) (\tilde{\beta} - \beta_{A,0}) \leq n^4 |G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0})| (\tilde{\beta} - \beta_{A,0})^2 + o_P(n^4). \quad (69)$$

Then by (68) and taking into account (69), we have

$$\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'| W_n \hat{G}_n(\tilde{\beta}) (\tilde{\beta} - \beta_{A,0}) \leq n^4 B^2 \|\tilde{\beta} - \beta_{A,0}\|_2^2 + o_P(n^4). \quad (70)$$

Now substitute (67)-(70) into the term on the right hand side of (64), we get

$$E \left[\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'| W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right) \right]^2 I_{\{\hat{\eta} > \eta/2\}} \leq 2B^2 n^4 E \left(\|\tilde{\beta} - \beta_{A,0}\|_2^2 I_{\{\hat{\eta} > \eta/2\}} \right) + 2Bn^3 + o(n^4). \quad (71)$$

Define the ridge based version of $\tilde{\beta}$ with imposing $\lambda_1^* = 0$

$$\tilde{\beta}(\lambda_2, 0) = \arg \min_{\beta} \left\{ \left(\sum_{i=1}^n g_i(\beta_A)' W_n \left(\sum_{i=1}^n g_i(\beta_A) \right) + \lambda_2 \sum_{j \in A} \beta_j^2 \right) \right\}. \quad (72)$$

Then use the arguments in the proof of Theorem 2 (equation (53)), we have

$$\begin{aligned} \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2 &\leq \frac{\lambda_1^* \max_{j \in A} \hat{w}_j \sqrt{p}}{n^2 \text{Eig}_{\min} \left(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right) + \lambda_2 + o_P(n^2)} \\ &\leq \frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2 b + \lambda_2 + o_P(n^2)}, \end{aligned} \quad (73)$$

$$\text{where } \text{Eig}_{\min} \left(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right) \geq \text{Eig}_{\min} \left(G(\beta_0)' \Omega_*^{-1} G(\beta_0) \right) \geq b.$$

Then follow the proof of Theorem 2 (i.e. equation (59)), for the right hand side term in (71),

$$E \left(\|\tilde{\beta} - \beta_{A,0}\|_2^2 I_{\{\hat{\eta} > \eta/2\}} \right) \leq 4 \frac{\lambda_2^2 \|\beta_0\|_2^2 + n^3 p B + \lambda_1^{*2} (\eta/2)^{-2\gamma} p + o(n^2)}{(bn^2 + \lambda_2 + o(n^2))^2}. \quad (74)$$

Now combine (61) (62)(63)(64)(71)(74) into (60),

$$\begin{aligned} P \left[\exists j \in A^c, |2\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta}) \right)| > \lambda_1^* \hat{w}_j \right] &\leq \frac{16}{\eta^2} \left[\frac{\lambda_2^2 \beta_0 \|_2^2 + B n^3 p + \lambda_1^2 p + o(n^2)}{[n^2 b + \lambda_2 + o(n^2)]^2} \right] \\ &+ \frac{4}{M^2} \left[\frac{\lambda_2^2 \beta_0 \|_2^2 + B n^3 p + \lambda_1^2 p + o(n^2)}{[n^2 b + \lambda_2 + o(n^2)]^2} \right] \\ &+ \frac{4M^{2\gamma}}{\lambda_1^{*2}} \left[(2n^3 B + o(n^3)) + (2B^2 n^4 + o(n^4)) \right] \\ &\times \left(\frac{\lambda_2^2 \|\beta_0\|_2^2 + n^3 p B + \lambda_1^{*2} (\eta/2)^{-2\gamma} p + o(n^2)}{(bn^2 + \lambda_2 + o(n^2))^2} \right). \end{aligned} \quad (75)$$

Now we need to show that each square bracketed term on the right hand side of equation (75) converges in probability to zero. We consider each of the right hand side elements in (75). We start with the first square bracketed term. The orders of the expressions are

$$\begin{aligned} \frac{p}{n\eta^2} \frac{\lambda_2^2}{n^3} &\rightarrow 0, \\ \frac{p}{n\eta^2} &\rightarrow 0, \\ \frac{p}{n\eta^2} \frac{\lambda_1^2}{n^3} &\rightarrow 0, \end{aligned}$$

by $\lambda_1^2/n^3 \rightarrow 0$, $\lambda_2^2/n^3 \rightarrow 0$ via Assumption 5, and

$$\frac{p}{n\eta^2} = \frac{1}{n^{1-\alpha}\eta^2} \rightarrow 0, \quad (76)$$

by Assumption 5(iv) and 3 with $n^{1-\nu}\eta^2 \rightarrow \infty$ with $\alpha < \nu$. Note that $\|\beta_0\|_2^2 = O(p)$ by Assumption 1.

Next we consider the second square bracketed term on the right hand side of (75). See that the dominating term in that expression is stochastic order of

$$O\left(\frac{p}{n M^2}\right) \rightarrow 0.$$

The above is true since by Assumption 5, $\frac{\lambda_1^*}{n^{3+\alpha}} n^{\gamma(1-\alpha)} \rightarrow \infty$, and $M = \left[\frac{\lambda_1^*}{n^{3+\alpha}} \right]^{1/2\gamma}$ since $\gamma > 0$

$$\frac{M^2 n}{p} = \left(\frac{\lambda_1^*}{n^{3+\alpha}} \right)^{1/\gamma} n^{1-\alpha} \rightarrow \infty.$$

The other terms in the second term on the right hand side of (75) are

$$\frac{\lambda_2^2 \|\beta_0\|_2^2}{n^3 n M^2} \rightarrow 0,$$

by $\lambda_2^2/n^3 \rightarrow 0$, $\|\beta_0\|_2^2 = O(p)$, $\frac{p}{n M^2} \rightarrow 0$ by the analysis of the dominating term above. Then also in the same way

$$\frac{\lambda_1^2 p}{M^2 n^4} = \frac{\lambda_1^2}{n^3} \frac{p}{n M^2} \rightarrow 0,$$

where we use $\lambda_1^2/n^3 \rightarrow 0$, and the analysis of the dominating term $\frac{p}{n M^2} \rightarrow 0$ above.

Now we consider the last square bracketed term in (75). The orders of the expression are

$\frac{M^{2\gamma} n^3}{\lambda_1^{*2}}$, $\frac{M^{2\gamma} \lambda_2^2 p}{\lambda_1^{*2}}$, $\frac{M^{2\gamma} n^3 p}{\lambda_1^{*2}}$, $\frac{M^{2\gamma} \lambda_1^{*2} p}{\lambda_1^{*2} \eta^{2\gamma}}$ respectively. Clearly the order of the third expression dominates that of the first and second expressions, given $\lambda_2^2/n^3 \rightarrow 0$. Now the order of the third term is

$$\frac{M^{2\gamma} n^3 p}{\lambda_1^{*2}} = \frac{1}{\lambda_1^*} \rightarrow 0,$$

given $M = \left(\frac{\lambda_1^*}{n^{3+\alpha}} \right)^{2\gamma}$ definition. For the order of the last expression

$$\frac{M^{2\gamma} \lambda_1^{*2}}{\lambda_1^{*2} \eta^{2\gamma}} p = M^{2\gamma} \frac{p}{\eta^{2\gamma}} = \frac{\lambda_1^*}{n} \frac{1}{n^2 \eta^{2\gamma}} \rightarrow 0,$$

since $\lambda_1^*/n \rightarrow 0$, $p = n^\alpha$, $p = n^\alpha$, M definition, and by Assumption 5(iv).

Q.E.D.

Proof of Theorem 3(ii)

The proof technique is very similar to the proof of Theorem 3.3 in Zou and Zhang (2009), given Theorems 2 and 3(i). After Theorem 3(i) result, it suffices to prove

$$P \left(\min_{j \in A} |\tilde{\beta}_j| > 0 \right) \rightarrow 1.$$

Then we can write the following with $\tilde{\beta}(\lambda_2, 0)$ defined in (72)

$$\min_{j \in A} |\tilde{\beta}_j| > \min_{j \in A} |\tilde{\beta}(\lambda_2, 0)_j| - \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2. \quad (77)$$

Also see that

$$\min_{j \in A} |\tilde{\beta}(\lambda_2, 0)_j| > \min_{j \in A} |\beta_{j0}| - \|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2. \quad (78)$$

Combine (77)(78) to have

$$\min_{j \in A} |\tilde{\beta}_j| > \min_{j \in A} |\beta_{j0}| - \|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2 - \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2. \quad (79)$$

Now we consider the last two terms on the right hand side of (79). Similar to derivation of (57)(58) we have

$$\begin{aligned} E\|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2^2 &\leq \frac{4\lambda_2^2\|\beta_{A,0}\|_2^2 + 4n^3pB + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2} \\ &= O(p/n) = o(1), \end{aligned} \quad (80)$$

by $\lambda_2^2\|\beta_{A,0}\|_2^2/n^4 = (\lambda_2^2/n^2)(\|\beta_{A,0}\|_2^2/n^2) \rightarrow 0$, and $\|\beta_{A,0}\|_2^2 = O(p)$. Then by (73)

$$\|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2 \leq \frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2b + \lambda_2 + o_p(n^2)}. \quad (81)$$

See that

$$\frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2b + \lambda_2 + o_p(n^2)} = O\left((\hat{\eta}/\eta)^{-\gamma} \frac{\lambda_1^* \sqrt{p}}{n^2} \eta^{-\gamma}\right). \quad (82)$$

$$\frac{\lambda_1^* \sqrt{p}}{n^2} \eta^{-\gamma} = \frac{1}{p^{1/2}} \frac{\lambda_1^*}{n} \left(\frac{p}{n} \frac{1}{\eta^\gamma}\right) = \frac{1}{n^{\alpha/2}} o(1), \quad (83)$$

by (76) ($p = n^\alpha$), Assumptions 5(i),(iv). Then by Theorem 2

$$\begin{aligned} E[(\hat{\eta}/\eta)^2] &\leq 2 + \frac{1}{\eta^2} E[\hat{\eta} - \eta]^2 \\ &\leq 2 + \frac{2}{\eta^2} E\|\tilde{\beta}(\lambda_2, \lambda_1) - \beta_0\|_2^2 \\ &\leq 2 + \frac{8}{\eta^2} \left[\frac{\lambda_2^2\|\beta_0\|_2^2 + n^3pB + \lambda_1^2p + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2} \right] \\ &= O(1), \end{aligned} \quad (84)$$

by $\lambda_1/n \rightarrow 0$, $\lambda_2/n \rightarrow 0$, $p/n \rightarrow 0$, $\lambda_1/n \rightarrow 0$, $\lambda_2/n \rightarrow 0$, $p/n \rightarrow 0$, $\|\beta_0\|_2^2 = O(p)$, by (76).

Using the same technique in the proof of Theorem 1(ii), substitute (83)(84) into (82) to have

$$\frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2b + \lambda_2 + o_p(n^2)} \xrightarrow{P} 0, \quad (85)$$

Now use (80) and (83) with $\eta = \min_{j \in A} |\beta_{j0}|$ to have

$$\min_{j \in A} |\tilde{\beta}_j| > \eta - \sqrt{\frac{p}{n}} O_P(1) - \sqrt{\frac{1}{n^\alpha}} o_P(1). \quad (86)$$

Then we obtain the desired result since by Assumption 5(v) and 5(iv) the last two terms on right hand side of (86) converges to zero faster than η .

Q.E.D

Proof of Theorem 4

We now prove the limit result. First, define

$$z_n = \delta' \left[\frac{I + \lambda_2 \left(\hat{G}(\hat{\beta}_{acnet,A})' W_n \hat{G}(\hat{\beta}_{acnet,A}) \right)^{-1}}{1 + \lambda_2/n} \right] \left(\hat{G}(\hat{\beta}_{acnet,A})' W_n \hat{G}(\hat{\beta}_{acnet,A}) \right)^{1/2} n^{-1/2} (\hat{\beta}_{acnet,A} - \beta_{A,0}).$$

Then we need the following result. Following the proof of Theorem 1 and using Theorem 3,

$$\|\hat{\beta}_{acnet,A} - \beta_{A,0}\|_2^2 \xrightarrow{P} 0,$$

and by (80)

$$\|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2^2 \xrightarrow{P} 0.$$

Next by Assumption 1 and Assumption 2, and considering the above results about consistency, we have

$$\left\| \frac{\hat{G}(\hat{\beta}_{acnet,A})'}{n} W_n \frac{\hat{G}(\hat{\beta}_{acnet,A})}{n} - \frac{\hat{G}(\tilde{\beta}(\lambda_2, 0))'}{n} W_n \frac{\hat{G}(\tilde{\beta}(\lambda_2, 0))}{n} \right\|_2^2 \xrightarrow{P} 0. \quad (87)$$

Next note that as in p.18 of Zou and Zhang (2009)

$$\begin{aligned} \delta' \left[I + \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} \right] & \times \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{1/2} n^{1/2} \left(\tilde{\beta} - \frac{\beta_{A,0}}{1 + \lambda_2/n} \right) \\ & = \left[\delta \left(I + \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} \right) \left(\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right)^{1/2} \right. \\ & \times \left. n^{-1/2} \frac{\lambda_2 \beta_{A,0}}{n + \lambda_2} \right] \\ & + \left[\delta \left(I + \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} \right) \left(\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right)^{1/2} \right. \\ & \times \left. n^{-1/2} (\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)) \right] \\ & + \left[\delta \left(I + \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} \right) \left(\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right)^{1/2} \right. \\ & \times \left. n^{-1/2} (\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}) \right]. \end{aligned}$$

The last term on the right hand side of (88) can be rewritten as (by (37))

$$\begin{aligned}
& \left\{ \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} + I \right\} \times \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{1/2} n^{-1/2} [\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}] \\
& = -\lambda_2 n^{-1/2} \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1/2} \beta_{A,0} \\
& - n^{-1/2} \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1/2} \\
& \times \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n g_n(\beta_{A,0}) \right] + o_P(1),
\end{aligned} \tag{89}$$

where $g_n(\beta_{A,0}) = \sum_{i=1}^n g_i(\beta_{A,0})$. Note that in the above equation we have an asymptotically negligible term due to using (37) where we have $\bar{\beta}$ instead of $\tilde{\beta}(\lambda_2, 0)$.

Via Theorem 3, and also using (87), with probability tending to one, $z_n = T_1 + T_2 + T_3$, where

$$\begin{aligned}
T_1 &= \left[\delta' \left[\left(I + \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} \right) \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{1/2} \right. \right. \\
&\quad \times \left. \left. n^{-1/2} \lambda_2 \frac{\beta_{A,0}}{n + \lambda_2} \right] \right. \\
&\quad \left. - \left[\delta' \lambda_2 n^{-1/2} \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1/2} \beta_{A,0} \right] \right] \\
T_2 &= \delta' \left[I + \lambda_2 \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1} \right] \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{1/2} \\
&\quad \times n^{-1/2} (\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)). \\
T_3 &= \delta' \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)) \right]^{-1/2} \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \sum_{i=1}^n \frac{g_i(\beta_{A,0})}{n^{1/2}} \right].
\end{aligned}$$

Consider T_1 , use Assumptions 1, 2, 5 with $W_n = \hat{\Omega}_*^{-1}$, $W = \Omega_*^{-1}$

$$\begin{aligned}
T_1^2 &\leq \frac{2}{n} \left\| \left(I + \lambda_2 \left(G(\beta_{A,0})' W G(\beta_{A,0}) \right)^{-1} \right) \left(G(\beta_{A,0})' W G(\beta_{A,0}) \right)^{1/2} \frac{\lambda_2 \beta_{A,0}}{n + \lambda_2} \right\|_2^2 \\
&\quad + \frac{2}{n} \left\| \lambda_2 \left(G(\beta_{A,0})' W G(\beta_{A,0}) \right)^{-1/2} \beta_{A,0} \right\|_2^2 + o_P(1) \\
&\leq \frac{2}{n} \frac{\lambda_2^2 B n}{(n + \lambda_2)^2} \left(1 + \frac{\lambda_2}{b n} \right)^2 \|\beta_{A,0}\|_2^2 + \frac{2}{n} \lambda_2^2 \|\beta_{A,0}\|_2^2 \frac{1}{b n^2} + o_P(1) \\
&= o_P(1),
\end{aligned}$$

via $\lambda_2 = o(n)$, $\|\beta_{A,0}\|_2^2 = O(p)$, and $\lambda_2^2 \|\beta_{A,0}\|_2^2 / n^2 \rightarrow 0$. Consider T_2 similar to the above analysis and (73)

$$\begin{aligned}
T_2^2 &\leq \frac{1}{n} \left(1 + \frac{\lambda_2}{b n} \right)^2 (B n) \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2^2 \\
&\leq B \left(1 + \frac{\lambda_2}{b n} \right)^2 \left[\frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{[n^2 b + \lambda_2 + o_P(n^2)]} \right]^2 \\
&= o_P(1),
\end{aligned}$$

by (85) and $\lambda_2 = o(n)$.

For the term T_3 we benefit from Liapunov Central Limit Theorem. By Assumptions 1 and 2

$$\begin{aligned} T_3 &= \frac{\sum_{i=1}^n \delta' \left[\widehat{G}(\tilde{\beta}(\lambda_{2,0}))' \widehat{\Omega}_*^{-1} \widehat{G}(\tilde{\beta}(\lambda_{2,0})) \right]^{-1/2} \left[\widehat{G}(\tilde{\beta}(\lambda_{2,0}))' \widehat{\Omega}_*^{-1} \right] g_i(\beta_{A,0})}{\sum_{i=1}^n \delta' \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} \left[G(\beta_{A,0})' \Omega_*^{-1} \right] g_i(\beta_{A,0})} + o_P(1). \\ &= \frac{\sum_{i=1}^n \delta' \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} \left[G(\beta_{A,0})' \Omega_*^{-1} \right] g_i(\beta_{A,0})}{n^{1/2}} + o_P(1). \end{aligned}$$

Next set $R_i = \delta' \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} \left[G(\beta_{A,0})' \Omega_*^{-1} \right] \frac{g_i(\beta_{A,0})}{n^{1/2}}$. So

$$\begin{aligned} \sum_{i=1}^n ER_i^2 &= n^{-1} \sum_{i=1}^n E \left[\delta' \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} G(\beta_{A,0})' \Omega_*^{-1} g_i(\beta_{A,0}) g_i(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right] \right] \\ &= \delta' \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} G(\beta_{A,0})' \Omega_*^{-1} \left[n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})' \right] \Omega_*^{-1} G(\beta_{A,0}) \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right] \end{aligned}$$

Then use $\|n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})' - \Omega_*\|_2^2 \rightarrow 0$. Take the limit of the term above

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{i=1}^n ER_i^2 &= \delta' \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} G(\beta_{A,0})' \Omega_*^{-1} \Omega_* \Omega_*^{-1} G(\beta_{A,0}) \left[G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right]^{-1/2} \delta. \\ &= \delta' \delta = 1. \end{aligned}$$

Next we show $\sum_{i=1}^n E |R_i|^{2+l} \rightarrow 0$. See that by (6)(7) and 4

$$\begin{aligned} \sum_{i=1}^n \frac{1}{n^{1+l/2}} E \left[\delta' \left(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0}) \right)^{-1/2} G(\beta_{A,0})' \Omega_*^{-1} g_i(\beta_{A,0}) \right]^{2+l} &\leq \left[\frac{B}{b} \right]^{1+l/2} \frac{1}{n^{1+l/2}} \sum_{i=1}^n E \| \Omega_*^{-1/2} g_i(\beta_{A,0}) \|_{2+l}^{2+l} \\ &\rightarrow 0. \end{aligned}$$

So $T_3 \xrightarrow{d} N(0, 1)$. The desired result then follows from $z_n = T_1 + T_2 + T_3$ with probability approaching one.

Q.E.D.

References

- Abadir, KM.; Magnus, JR. Matrix Algebra. Cambridge University Press; 2005.
- Ai C, Chen X. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*. 2003; 71:1795–1843.
- Alfaro L, Kalemli-Ozcan S, Volosoych V. Why does not capital flow from rich to poor countries? An empirical investigation. *Review of Economics and Statistics*. 2008; 90:347–368.
- Andrews DWK, Lu B. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics*. 2001; 101:123–165.
- Caner M. Lasso-type GMM estimator. *Econometric Theory*. 2009; 25:270–290.
- Chen X. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*. 2007; Vol6b:5550–5623. Chapter 76.
- Chen X, Ludvigson C. Land of Addicts? An empirical investigation of habit based asset pricing models. *Journal of Applied Econometrics*. 2009; 24:1057–1093.

- Davidson, J. Stochastic Limit Theory. Oxford University Press; 1994.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*. 2001; 96:1348–1360.
- Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*. 2004; 32:928–961.
- Gao X, Huang J. Asymptotic analysis of high dimensional lad regression with lasso. *Statistica Sinica*. 2010; 20:1485–1506.
- Guggenberger P, Smith RJ. Generalized Empirical Likelihood Tests in time series models with potential identification failure. *Journal of Econometrics*. 2008; 142:134–161.
- Han C, Phillips PCB. GMM with many moment conditions. *Econometrica*. 2006; 74:147–192.
- Hansen LP. Large sample properties of generalized method of moment estimators. *Econometrica*. 1982; 50:1029–1054.
- He X, Shao QM. On parameters of increasing dimensions. *Journal of Multivariate Analysis*. 2000; 75:120–135.
- Huang J, Horowitz J, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*. 2008; 36:587–613.
- Huber P. Robust regression: Asymptotics, conjectures and monte carlo. *Annals of Statistics*. 1988; 1:799–821.
- Knight K, Fu W. Asymptotics for lasso type estimators. *Annals of Statistics*. 2000; 28:1356–1378.
- Knight K. Shrinkage estimation for nearly-singular designs. *Econometric Theory*. 2008; 24:323–338.
- Lam C, Fan J. Profile-kernel likelihood inference with diverging number of parameters. *Annals of Statistics*. 2008; 36:2232–2260. [PubMed: 19173010]
- Leeb H, Pötscher B. Model selection and inference: facts and fiction. *Econometric Theory*. 2005; 21:21–59.
- Liao, Z. Adaptive GMM shrinkage estimation with consistent moment selection. Yale University. Department of Economics; 2011. Working paper
- Newey W, Powell J. Instrumental variable estimation of nonparametric models. *Econometrica*. 2003; 71:1557–1569.
- Newey W, Windmeijer F. GMM with many weak moment conditions. *Econometrica*. 2009a; 77:687–719.
- Newey W, Windmeijer F. Supplement to GMM with many weak moment conditions. *Econometrica*. 2009b; 77:687–719.
- Otsu T. Generalized Empirical Likelihood Inference for nonlinear and time series models under weak identification. *Econometric Theory*. 2006; 22:513–527.
- Portnoy S. Asymptotic behavior of M-estimators of p-regression parameters when p^2/n is large. I. consistency. *Annals of Statistics*. 1984; 12:1298–1309.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J. Royal. Statistical Society Series B*. 1996; 58:267–288.
- Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Series B*. 2009; 71:671–683.
- Wang H, Li R, Tsai C. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. 2007; 94:553–568. [PubMed: 19343105]
- Zhang H, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*. 2007; 37:1–13.
- Zou H. The adaptive lasso and its oracle properties. *Journal of The American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. *Annals of Statistics*. 2007; 35:2173–2192.
- Zou H, Zhang H. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*. 2009; 37:1733–1751. [PubMed: 20445770]

Table 1

Success Percentages of Selecting the Correct Model

Estimators	Design 1	Design 2
Adaptive Elastic Net	91.2	94.9
Bridge-GMM	100.0	100.0
GMM-BIC	6.9	0.0

Note: The GMM-BIC (Andrews and Lu, 2001) represents the models that are selected according to BIC and subsequently we use GMM. The Bridge-GMM estimator is studied in Caner (2009). The Adaptive Elastic Net estimator is the new procedure proposed in this study.

Table 2

Summary Mean Squared Error (MSE)

Estimators	Design 1	Design 2
Adaptive Elastic Net	1.8	1.3
Bridge-GMM	4.2	1.3
GMM-BIC	165848.5	876080.2

Note: The MSE formula is given in (10). Instead of expectations, the average of iterations are used. A small number for summary MSE is desirable for a model. The GMM-BIC (Andrews and Lu, 2001) represents models that are selected according to BIC and subsequently we use GMM. The Bridge-GMM estimator is studied in Caner (2009). The Adaptive Elastic Net is the new procedure proposed in this study.

Table 3

Bias and RMSE Results of Design 1

	Adaptive Elastic Net		Bridge-GMM		GMM-BIC	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
β_1	-0.244	0.272	-0.117	0.126	2.903	159.85
β_2	-0.244	0.272	-0.667	0.669	-4.082	261.32
β_3	0.013	0.042	0.000	0.000	-0.859	158.839
β_4	0.000	0.009	0.000	0.000	0.612	188.510
β_5	0.013	0.041	0.000	0.000	1.162	62.240

Note: The GMM-BIC (Andrews and Lu, 2001) represents models that are selected according to BIC and subsequently we use GMM. The Bridge-GMM estimator is studied in Caner (2009). The Adaptive Elastic Net estimator is the new procedure proposed in this study.

Table 4

Bias, RMSE of Design 2

	Adaptive Elastic Net		Bridge-GMM		GMM-BIC	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
β_1	-0.181	0.193	-0.112	0.124	-0.805	158.171
β_2	-0.181	0.193	-0.662	0.665	-0.326	112.970
β_3	0.010	0.061	0.157	0.166	-0.314	120.358
β_4	-0.038	0.071	0.337	0.341	-6.759	659.673
β_5	-0.001	0.007	0.000	0.000	7.740	617.509

Note: The GMM-BIC (Andrews and Lu, 2001) represents models that are selected according to BIC and subsequently we use GMM. The Bridge-GMM estimator is studied in Caner (2009). The Adaptive Elastic Net estimator is the new procedure proposed in this study.