

GENERAL ESTIMATING EQUATIONS: MODEL SELECTION AND ESTIMATION WITH DIVERGING NUMBER OF PARAMETERS

BY MEHMET CANER[†], HAO HELEN ZHANG^{*,†}

North Carolina State University[†]

This paper develops adaptive elastic net estimator for general estimating equations. We allow for number of parameters diverge to infinity. The estimator can also handle collinearity among large number of variables as well. This method has the oracle property, meaning we can estimate nonzero parameters with their standard limit and the redundant parameters are dropped from the equations simultaneously. This paper generalizes 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 solution.

*Zhang's research is supported by National Science Foundation Grant DMS 0645293, and National Institutes of Health Grant NIH/NCI R01 CA-085848.

AMS 2000 subject classifications: Primary 62J05; secondary 62J07

Keywords and phrases: Estimating Equations, Lasso Penalty

1. Introduction. General estimating equations are studied extensively in statistics. They provide a convenient way to describe parameters and statistics. We can estimate the parameters by two-step efficient Generalized Method of Moments (GMM) of Hansen (1982), or by empirical likelihood estimator of Owen (2001), and Qin and Lawless (1994). The other estimators used to are exponential tilting of Kitamura and Stutzer (1997), and exponentially tilted empirical likelihood of Schennach (2007). They all share the first order efficiency of GMM. GMM is also used in economics, finance, accounting, and strategic planning literature as well.

In this paper we are concerned about model selection in GMM when the number of parameters diverge. These situations can arise in labor economics, international finance (see Alfaro, Kalemli-Ozcan, Volosovych (2008)). In linear models when the some of the regressors are correlated with errors and with lots of co-variates, the model selection tools are essential. They can improve the finite sample performance of the estimators.

Model selection techniques are very useful and now widely used in statistics. For example Knight and Fu (2000) derive the asymptotics of lasso, and Fan and Li (2001) propose SCAD estimator. In econometrics, Knight (2008), and Caner (2009) offer Bridge-least squares and Bridge-GMM estimators respectively. But these are all in finite dimensions. Recently model selection with large number of parameters are analyzed in least squares by Huang, Horowitz, and Ma (2008), and Zou and Zhang (2009), where the first article analyzes Bridge estimator, and the second paper is concerned with adaptive elastic net estimator.

Adaptive elastic net estimator has the oracle property when the number of parameters diverge with the sample size. Furthermore, this method can handle the collinearity arising from 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 sparse representation) this estimator can estimate the zero parameters as zero. When we set parameters to zero, we benefit from a data dependent method unlike Bridge estimation.

In this paper we extend the least squares based adaptive elastic net of Zou and Zhang (2009) to GMM. GMM has no closed form solution unlike least squares estimator. This results in a new proof technique compared with least squares case of Zou and Zhang (2009). We need a different and extensive consistency proof. The nonlinear nature of the functions introduce additional difficulties, and this restricts the growth of the number of parameters compared to simple linear case. The estimator also has the oracle property, the nonzero coefficients are estimated converging to a normal distribution. This is their standard limit furthermore, the zero parameters are estimated

as zero.

Earlier work on diverging parameters include Huber (1988), and Portnoy (1984). In recent years, Fan, Peng, Huang (2005) study the a semi-parametric model with a growing number of nuisance parameters, Lam and Fan (2007) analyze the profile likelihood ratio inference with growing number of parameters. As far as we know this is the first paper to estimate and select the model in GMM with diverging number of parameters.

Section 2 presents the model and the estimator. Then in section 3 we derive the asymptotic results for the estimator. Section 4 conducts simulations. 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, 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. When n converges to infinity the parameter space is denoted as B_∞ . For example this may be infinite cube $[0, 1]^\infty$ in certain examples. 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 are q , $q \geq p$. So we also allow q to grow with the sample size, but $q/n \rightarrow 0$ as $n \rightarrow \infty$. From now on we denote $g(X_i, \beta)$ as $g_i(\beta)$. 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 answer the following questions. If we have large number of control variables some of them may be irrelevant (we may have also large number of endogenous variables and control variables) in the structural equation in a simultaneous equation system or 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 large number of variables possibly there may be some correlation among the variables, can this method handle that? Is it also possible that the estimators 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 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 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 adaptive lasso estimator which can handle the first and third questions and does not handle correlation among large number of variables.

First we introduce the notation: $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, and $\|\beta\|_2^2 = \sum_{j=1}^p |\beta_j|^2$. We start by introducing the adaptive elastic net estimator, given the positive and diverging tuning parameters 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)

$$\hat{\beta}_{aenet} = (1 + \lambda_2/n) \{ \underset{\beta}{\operatorname{argmin}} [(\sum_{i=1}^n g_i(\beta))' W_n (\sum_{i=1}^n g_i(\beta)) + \lambda_2 \|\beta\|_2^2 + \lambda_1^* \sum_{j=1}^p \hat{w}_j |\beta_j|] \},$$

(1)

where $\hat{w}_j = \frac{1}{|\hat{\beta}_{enet}|^\gamma}$, $\hat{\beta}_{enet}$ is a consistent estimator immediately explained below, and γ is a positive constant, and with the $p = n^\alpha$, $0 < \alpha < 1/3$,

$$0 < \frac{(2 + 4\alpha)}{1 - \alpha} < \gamma < \frac{2(2 - \alpha)}{(1 - \alpha)}.$$

γ will be explained in detail in Assumption 4iii. 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,

$$\hat{\beta}_{enet} = (1 + \lambda_2/n) \{ \underset{\beta}{\operatorname{argmin}} S_n(\beta) \},$$

where

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

λ_1, λ_2 are positive and diverging sequences that will be defined in Assumption 6.

We now discuss the penalty functions in both estimators and why we need $\hat{\beta}_{enet}$. The elastic net estimator has both l_1, l_2 penalties. The l_1 part performs the automatic variable selection, and l_2 part improves the prediction and handles the collinearity that may arise with large number of variables. However, the elastic net estimator does not provide oracle property. So by introducing an adaptive weight in adaptive elastic net we obtain the oracle property. The adaptive weights are also useful, 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 able to get oracle property, but with large number of parameters/variables that may be collinear an additional ridge-like penalty as in adaptive elastic net is better.

2.2. *The Assumptions.* We now provide the main assumptions.

1. The following uniform law of large number holds

$$\sup_p \sup_{\beta \in B_p} \left[\frac{1}{n} \sum_{i=1}^n |g_i(\beta) - E g_i(\beta)| \right] \xrightarrow{p} 0.$$

2. Define $En^{-1} \sum_{i=1}^n g_i(\beta) = m_{1n}(\beta)$, then

(i). Assume uniformly over β and p that $m_{1n}(\beta) \rightarrow m_1(\beta)$, $m_{1n}(\beta)$ is continuously differentiable in β , $m_1(\beta_0) = 0$, and $m_1(\beta) \neq 0$, for $\beta \neq \beta_0$, $m_1(\beta)$ is continuous in β .

(ii). 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 in a neighborhood \mathcal{N} of β_0 (uniform over p as well), $\hat{G}_n(\beta)/n \xrightarrow{p} G(\beta)$, where $G(\beta_0)$ is of full column rank p , and $G(\beta)$ is continuous in β . The partial derivative of $g_i(\beta)$ is continuous in β .

3. W_n is a positive definite matrix, and $W_n \xrightarrow{p} W$, where W is symmetric, positive definite, and finite matrix.

4. (i). $b \leq \text{Eigmin}(G(\beta_0)' \Omega^{-1} G(\beta_0))$, where $G(\beta_0)' \Omega^{-1} G(\beta_0) = \Sigma$, $\text{Eigmax}(\Sigma) \leq B$, and $\Omega = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n E g_i(\beta_0) g_i(\beta_0)'$, positive definite, and finite. b and B are positive constants.

(ii). $p/n \rightarrow 0$ as $n \rightarrow \infty$, $p = n^\alpha$, $0 < \alpha < 1/3$, $q/n \rightarrow 0$ as $n \rightarrow \infty$, and $q \rightarrow \infty$.

(iii). The coefficient on the weights γ satisfies the following bounds:

$$0 < \frac{(2 + 4\alpha)}{1 - \alpha} < \gamma < \frac{2(2 - \alpha)}{(1 - \alpha)}.$$

5.

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

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

6. $\lambda_1/n^{1/2} \rightarrow 0$, $\lambda_2/n^{1/2} \rightarrow 0$, and

$$(3) \quad \lambda_1^* = o(n^{(2-\alpha) + \frac{\gamma}{2}(\alpha-1)}).$$

$$(4) \quad \frac{\lambda_1^*}{n^{3+\alpha}} n^{\gamma(1-\alpha)} \rightarrow \infty.$$

Note that since B_p is compact, B_∞ is compact through Theorem 6.17 (Tychonoff's Theorem) in Davidson (1994). Assumption 1 simplifies the proofs since the main concern is the limit law and oracle property of the estimators. The uniform law of large number result can be obtained from a triangle array based Donsker result which is used in empirical process theory (Andrews, 1994). The primitives for that functional central limit theorem is the Lipschitz continuity and the uniform boundedness of the functions. Also we can benefit from another functional central limit theorem in van der Vaart (1998), Theorem 19.28 and the primitive of Lipschitz continuous functions are in Lemma 19.31 of van der Vaart (1998). Note that we need additional uniformity over p in these conditions compared to the previous literature cited above.

Assumptions 2-4 are standard in the literature of GMM (Newey and McFadden, 1994). Assumption 5 is a primitive condition for the triangular array central limit theorem. This is also restraining the number of orthogonality conditions q . The rates on λ_1, λ_2 are standard, but λ_1^* rate depends on α, γ . It is easy to see that with $0 < \alpha < 1/3$, and $\frac{(2+4\alpha)}{1-\alpha} < \gamma < \frac{2(2-\alpha)}{(1-\alpha)}$ two conditions on λ_1^* are compatible. With α approaching 0, it is easy to see that by (4) λ_1^* can diverge slower, so we do not need to penalize a lot. But with $\alpha \rightarrow 1/3$, then it is clear that the penalty should be larger. To give a concrete example with $\alpha = 1/5$, we can set $\gamma = 4$, and $\lambda_1^* = o(n^{1/5})$.

3. Asymptotics. Now we introduce two estimators which will be useful in providing the consistency of the adaptive elastic net estimator. We specifically analyze the following estimator, which is connected to elastic net estimator in (2)

$$(5) \quad \hat{\beta}(\lambda_2, \lambda_1) = \operatorname{argmin}_\beta S_n(\beta).$$

We will prove the consistency of $\hat{\beta}(\lambda_2, \lambda_1)$ estimator which will show the consistency of the elastic net estimator. Then we define another estimator which is tied to adaptive elastic net estimator in (1) and also will be used in the risk bound calculations.

$$(6) \quad \hat{\beta}_w = \operatorname{argmin}_\beta \left[\left(\sum_{i=1}^n g_i(\beta) \right)' W_n \left(\sum_{i=1}^n g_i(\beta) \right) + \lambda_2 \|\beta\|_2^2 + \lambda_1^* \sum_{j=1}^p \hat{w}_j |\beta_j| \right],$$

The following theorem provides consistency for both elastic net and an estimator closely connected to adaptive elastic net.

Theorem 1. *Under Assumptions 1-3, 4(ii), 6*

(i).

$$\hat{\beta}_{enet} \xrightarrow{p} \beta_0.$$

(ii).

$$\hat{\beta}_w \xrightarrow{p} \beta_0.$$

Remarks. 1. It is clear from Theorem 1ii that adaptive elastic net estimator in (1) is also consistent.

2. We should note that in Zou and Zhang (2009) least squares adaptive elastic net estimator, there is no explicit consistency proof. This is possible by the closed form solution in least squares estimator. However, the GMM adaptive elastic net estimator has no closed form solution. So we need a new consistency proof compared with the least squares case. The proof here uses empirical process theory.

Theorem 2. *Under Assumptions 1-4,*

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

and

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

Remark. Note that the first bound is related to the estimator in (6). The second bound is related to the estimator in (5). $\hat{\beta}_w$ is related to adaptive elastic net estimator in (1), and $\hat{\beta}(\lambda_1, \lambda_2)$ is related to the estimator in (2).

It is clear from the last result that the elastic net estimator is converging at the rate $\sqrt{n/p}$.

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 represent the zero (redundant) parameters. Let β_A represent the nonzero parameters.

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

Theorem 3. *Given Assumptions 1-4, and 6*

(i). *With probability tending to one $((1 + \frac{\lambda_2}{n})\tilde{\beta}, 0)$ is the solution to (1).*

(ii). *(Consistency in Selection) Also we have*

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

Remarks. 1. Theorem 3i shows that ideal estimator $\tilde{\beta}$ becomes the same as adaptive elastic net estimator in large samples. So GMM elastic adaptive net estimator has the same solution as $((1 + \lambda_2/n)\tilde{\beta}, 0_{p-p_A})$. In other words, GMM adaptive elastic net is like an oracle informing about the true subset model.

2. Theorem 3ii shows that the nonzero adaptive elastic net estimates display the oracle property. This is a sharper result than the one in Theorem 3i.

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 $\hat{\beta}_{aenet,A}$. Define a consistent variance estimator for nonzero parameters that can be derived from elastic net estimators as: $\hat{\Omega}_A$. We also define $\Omega_A = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})'$.

Theorem 4. *Under Assumptions 1-6, given $W_n = \hat{\Omega}_A^{-1}$, where $\hat{\Omega}_A^{-1} - \Omega_A^{-1} \xrightarrow{p} 0$, (we set $W = \Omega_A^{-1}$),*

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

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

Remarks. 1. First see that

$$(K_n - I_{p_A}) \xrightarrow{p} 0,$$

due to Assumptions 2ii, 3, and $\lambda_2 = o(\sqrt{n})$.

2. Since by Assumptions 2ii and 3 $[\hat{G}(\hat{\beta}_{aenet,a})' \hat{\Omega}_A^{-1} \hat{G}(\hat{\beta}_{aenet,a})]^{1/2} n^{-1/2} = O_p(n^{1/2})$, and with δ being a p_A vector of norm one, the rate of convergence of the adaptive elastic net estimator is $\sqrt{n/p_A}$.

3. This theorem clearly extends Zou and Zhang (2009) from least squares case to a GMM context, which involves different penalty factors, and a more restrictive α, γ as it is clear from Assumptions 4 and 6. This result generalizes theirs to a nonlinear functions of endogenous variables which are heavily used in econometrics and finance. This is not a simple or tedious generalization of least squares proof. The limit that we derive also corresponds to the standard GMM limit in Hansen (1982). That result is for fixed number of parameters with a well specified model. In this way we extend his result in the direction of large number of parameters with model selection.

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

5. 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

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

6. There will be discussion of 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.

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

$$y_i = x_i' \beta_0 + \epsilon_i,$$

$$x_i = z_i' \pi + \eta_i,$$

$$\epsilon_i = \rho \iota' \eta_i + \sqrt{1 - \rho^2} \iota' v_i,$$

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

The model that is estimated:

$$Ez_i\epsilon_i = 0,$$

for all $i = 1, \dots, n$. So $g_i(\beta_0) = z_i(y_i - x_i'\beta_0)$.

We have two different designs for parameter vector β_0 . In the first case $\beta_0 = \{3, 3, 0, 0, 0, \}$ (Design 1), then in the second one $\beta_0 = \{3, 3, 3, 3, 0\}$ (Design 2). We have $n = 100$, and $z_i \cong N(0, \Omega_z)$ 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 the x_i 's since two equations are correlated. Using five coefficients out of 100 observations clearly satisfy our Assumption 4ii. There may be more coefficients that are analyzed, but this makes BIC analysis very difficult.

In this section, we compare three methods. Namely they are GMM-BIC of Andrews and Lu (2001), Bridge-GMM of Caner (2009), and the adaptive elastic net GMM here. We have 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

$$(7) \quad E[(\hat{\beta} - \beta_0)'\Sigma_\epsilon(\hat{\beta} - \beta_0)],$$

where $\Sigma_\epsilon = E\epsilon_i\epsilon_i'$, and $\hat{\beta}$ represents the estimated coefficient vector coming from these three different methods. This measure is used in statistics literature, and for that see Zou and Zhang (2009). Next two measures are concerned about individual coefficients. First, the bias of each individual coefficient estimate is measured. Then root mean squared error of each coefficient is computed. We use 10000 iterations. The choice of λ 's in both Bridge-GMM and the adaptive elastic net GMM here is done via BIC. This is suggested by Zou, Hastie, Tibshirani (2007), and Wang, Li, Tsai (2007). Truncation of the small coefficient estimates are set to zero via $|\hat{\beta}_{Bridge}| < 2/\lambda$ for Bridge-GMM as suggested in Caner (2009). For adaptive elastic net, we use the modified shooting algorithm in Appendix 2 of Zhang and Lu (2007). This is possible also in adaptive elastic net since the quadratic penalty does not play a role in setting coefficients to zero. This is true since the partial derivative of that with respect to parameters evaluated at zero is also equal to zero. Least Angle Regression Solution is not used because it is not clear that this is useful and can be applied in GMM context.

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

$$(8) \quad U_n(\beta) = \left[\sum_{i=1}^n g_i(\beta) \right]' W_n \left[\sum_{i=1}^n g_i(\beta) \right] + \lambda \sum_{j=1}^p |\beta_j|^\gamma,$$

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

We now describe now the model selection by BIC in GMM (All subset selection) as 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 the parameters to be estimated or in 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 :

$$(9) \quad J_n(b) = \left[\sum_{i=1}^n g_i(\beta_{[b]}) \right]' W_n \left[\sum_{i=1}^n g_i(\beta_{[b]}) \right],$$

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 , where this represents all the possibilities with only one nonzero coefficient. M_{12} represents all the possibilities with two nonzero coefficients.

$$(10) \quad M_{12} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

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 $(1, 1, 1, 1, 0)'$.

Table 1: Success Percentage of Selecting the Correct Model

| Selection Procedure | Design 1 | Design 2 |
|--------------------------|----------|----------|
| Adaptive Elastic Net | 91.2 | 94.9 |
| Bridge | 100.0 | 100.0 |
| BIC-All Subset Selection | 6.9 | 0.0 |

Note: BIC-All subset selection represents models that are selected according to BIC and subsequently we use GMM. Bridge is the Bridge-GMM estimator in Caner (2009).

Adaptive elastic net is the one that is proposed in this study.

BIC (All subset selection) selects the model based on minimizing the following criterion among the 31 possibilities

$$(11) \quad J_n(b) + |b| \ln(n).$$

The penalty term penalizes the larger models. Denote the optimal model selection vector by b^* . After selecting the optimal model in (11), the vector b^* , we then estimate the parameters of the model by GMM. Next we provide 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 percentage for Designs 1 and 2. We see that both Bridge and Adaptive Elastic Net is doing very well, Bridge selects the correct model 100%, and Adaptive Elastic Net 91-95% of the time, whereas BIC (All subset selection) selects only 0-6.9%. This is due to lots of possibilities in the case of BIC (All subset selection), and with large number of parameters the performance of BIC (All subset selection) deteriorated. Table 2 provides summary MSE measure results. This clearly shows that here Adaptive Elastic Net is the best among the three. Its MSE figures are the smallest, whereas BIC (All subset selection) is terrible due to wrong model selection, and after the model selection estimating the zero coefficients with nonzero and large magnitudes. One of the main reasons that BIC's (All subset selection) bad performance is GMM is susceptible to near singular design and provides large coefficients when there is correlation among instruments. For this point, see Caner (2008). The other two methods are not affected by this problem. Tables 3 and 4 provide the bias and root mean squared error of each coefficient in Designs 1 and 2. Comparing Bridge with Adaptive Elastic Net, the bias of the nonzero coefficients are smaller generally in Adaptive Elastic Net. The same is true for generally in the case of root mean squared errors. These are smaller for the nonzero coefficients in Adaptive Elstic Net.

Table 2: Summary MSE

| Estimators | Design 1 | Design 2 |
|--------------------------|----------|----------|
| Adaptive Elastic Net | 1.8 | 1.3 |
| Bridge | 4.2 | 1.3 |
| BIC-All Subset Selection | 165848.5 | 876080.2 |

Note: Summary MSE is the formula in (7). Instead of expectations, the average of iterations are used. A small number for summary MSE is desirable for a model. BIC-All subset selection represents models that are selected according to BIC and subsequently we use GMM. Bridge is the Bridge-GMM estimator in Caner (2009). Adaptive elastic net is the one that is proposed in this study.

Table 3: Bias, RMSE of Design 1

| | Adaptive Elastic Net | | Bridge | | BIC-All Subset Selection | |
|-----------|----------------------|-------|--------|--------|--------------------------|---------|
| | 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: BIC-All subset selection represents models that are selected according to BIC and subsequently we use GMM. Bridge is the Bridge-GMM estimator in Caner (2009). Adaptive elastic net is the one that is proposed in this study.

Table 4: Bias, RMSE of Design 2

| | Adaptive Elastic Net | | Bridge | | BIC-All Subset Selection | |
|-----------|----------------------|-------|--------|--------|--------------------------|---------|
| | 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: BIC-All Subset selection represents models that are selected according to BIC and subsequently we use GMM. Bridge is the Bridge-GMM estimator in Caner (2009). Adaptive elastic net is the one that is proposed in this study.

Appendix

Proof of Theorem 1(i). Via Assumptions 1-3, uniformly over β

$$(12) \quad [n^{-1} \sum_{i=1}^n g_i(\beta)]' W_n [n^{-1} \sum_{i=1}^n g_i(\beta)] \xrightarrow{p} m_1(\beta)' W m_1(\beta).$$

Now we show why we need the assumptions about the penalty rates in the proofs. First start with the following. If we had assumed the following

$$(13) \quad \frac{\lambda_2}{n^2} \|\beta\|_2^2 \rightarrow m_2(\beta),$$

$$(14) \quad \frac{\lambda_1}{n^2} \|\beta\|_1 \rightarrow m_3(\beta),$$

where $m_2(\beta) > 0, m_3(\beta) > 0$, then uniformly over β , the limit would have been

$$\frac{S_n(\beta)}{n^2} \xrightarrow{p} m_1(\beta)' W m_1(\beta) + m_2(\beta) + m_3(\beta).$$

So in that case,

$$\hat{\beta}(\lambda_2, \lambda_1) \xrightarrow{p} \operatorname{argmin}_{\beta} S_n(\beta).$$

But due to $m_2(\beta), m_3(\beta)$, terms under (13)(14), there is no consistency of the estimator. However, if we modify the assumptions (13)(14) in the following way

$$(15) \quad \frac{\lambda_2}{n^2} \|\beta\|_2^2 = \frac{\lambda_2}{n^2} \sum_{j=1}^p |\beta_j|^2 \rightarrow 0.$$

$$(16) \quad \frac{\lambda_1}{n^2} \|\beta\|_1 = \frac{\lambda_1}{n^2} \sum_{j=1}^p |\beta_j| \rightarrow 0.$$

To get (15)(16) we need $\frac{\lambda_2}{n^2} p \rightarrow 0$ and $\frac{\lambda_1}{n^2} p \rightarrow 0$, where we need $\lambda_2 = o(n^{2-\alpha}), \lambda_1 = o(n^{2-\alpha})$ with the usage of $p = n^\alpha$. This is clear from Assumptions 4ii, and 6.

Then with (15)(16) instead of (13)(14) we have by Assumptions 1-3, 4ii, 6

$$\hat{\beta}(\lambda_2, \lambda_1) \xrightarrow{p} \operatorname{argmin}_{\beta} [m_1(\beta)' W m_1(\beta)].$$

Use Corollary 3.2.3 of van der Vaart and Wellner (1996) to have

$$\beta_0 = \operatorname{argmin}_{\beta} [m_1(\beta)' W m_1(\beta)].$$

This is also a unique minimum by Assumptions 2 and 3. So

$$\hat{\beta}(\lambda_2, \lambda_1) \xrightarrow{p} \beta_0,$$

and by the definition of $\hat{\beta}_{enet}$,

$$\hat{\beta}_{enet} \xrightarrow{p} \beta_0.$$

Q.E.D.

Proof of Theorem 1(ii). This follows exactly in the case of (i), given $\lambda_2 = o(n^{2-\alpha})$, and $\frac{\lambda_1^*}{n^2} \sum_{j=1}^p \hat{w}_j |\beta_j| \rightarrow 0$. For the proof of the last point we need this to be true for the "zero" parameters as well as "nonzero" parameters. The estimated weights behave differently in these two cases.

For the zero parameter case in elastic net estimators, $\hat{w}_j = \frac{1}{|\hat{\beta}_{j,enet}|^\gamma}$ and $\hat{\beta}_{j,enet} = O_p(\sqrt{p/n})$ via Theorem 2i below (independent of the proof here in Theorem 1ii). So for all $j = 1, \dots, p$,

$$\hat{w}_j = O_p(n^{\frac{\gamma}{2}(1-\alpha)}),$$

where we use $p = n^\alpha$. So for the zero parameters in elastic net

$$\frac{\lambda_1^*}{n^2} \sum_{j=1}^p \hat{w}_j |\beta_j| = O_p(\lambda_1^* n^{\alpha-2} n^{\frac{\gamma}{2}(1-\alpha)}) \rightarrow 0,$$

where we use $\sum_{j=1}^p |\beta_j| = O(n^\alpha)$. The last equation provides us with

$$(17) \quad \lambda_1^* = o(n^{(2-\alpha) + \frac{\gamma}{2}(\alpha-1)}).$$

For the nonzero parameters in elastic net estimator, using Theorem 1i, $\hat{\beta}_{j,enet} \xrightarrow{p} \beta_{j,0} \neq 0$. This shows us that $\hat{w}_j \xrightarrow{p} \frac{1}{|\beta_j|^\gamma} < \infty$, where $\beta_j \neq 0$. So

$$\frac{\lambda_1^*}{n^2} \sum_{j=1}^p \hat{w}_j |\beta_j| \xrightarrow{p} 0,$$

if $\frac{\lambda_1^*}{n^2} p \xrightarrow{p} 0$, which is true if

$$(18) \quad \lambda_1^* = o(n^{2-\alpha}).$$

To have the consistency we need the minimum of the rates in (17)(18) which is

$$\min(2-\alpha, (2-\alpha) + \frac{\gamma}{2}(\alpha-1)) = (2-\alpha) + \frac{\gamma}{2}(\alpha-1).$$

This is true since $\gamma > 0, \alpha < 1/3$.

Q.E.D.

Proof of Theorem 2. In this proof we start by analyzing the GMM-Ridge estimator that is defined as follows:

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

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 elastic net estimator is consistent, GMM-Ridge will be consistent as well. Define also the following $q \times p$ matrix $\hat{G}_n(\hat{\beta}_R) = \frac{\sum_{i=1}^n \partial g_i(\hat{\beta}_R)}{\partial \beta'}$. Then set $\bar{\beta} \in (\beta_0, \hat{\beta}_R)$. A mean value theorem around β_0 applied to first order conditions provides, with $g_n(\beta_0) = \sum_{i=1}^n g_i(\beta_0)$,

$$(19) \quad \hat{\beta}_R = -[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p]^{-1} [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0].$$

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

$$(20) \quad \hat{\beta}_R - \beta_0 = -[\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p]^{-1} [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) + \lambda_2 \beta_0].$$

We need the following expressions by using (19)

$$(21) \quad \begin{aligned} \hat{\beta}_R' [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0] &= -[\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0]' \\ &\times [\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p]^{-1} \\ &\times [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0]. \end{aligned}$$

$$(22) \quad \begin{aligned} \hat{\beta}_R' [\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] \hat{\beta}_R &= [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0]' \\ &\times [\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p]^{-1} [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) \beta_0]. \end{aligned}$$

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

$$(23) \quad \begin{aligned} \left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right] + \lambda_2 \|\hat{\beta}_R\|_2^2 &= g_n(\beta_0)' W_n g_n(\beta_0)' \\ &+ g_n(\beta_0)' W_n \hat{G}_n(\bar{\beta}) (\hat{\beta}_R - \beta_0) + (\hat{\beta}_R - \beta_0)' \hat{G}_n(\bar{\beta})' W_n g_n(\beta_0) \\ &+ (\hat{\beta}_R - \beta_0)' \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) (\hat{\beta}_R - \beta_0) + \lambda_2 \|\hat{\beta}_R\|_2^2. \end{aligned}$$

Furthermore we can rewrite the right hand side of (23) as

$$\begin{aligned}
g_n(\beta_0)'W_n g_n(\beta_0) &+ \hat{\beta}'_R[\hat{G}_n(\bar{\beta})'W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta})\beta_0] \\
&+ [\hat{G}_n(\bar{\beta})'W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta})\beta_0]' \hat{\beta}_R \\
&+ \hat{\beta}'_R[\hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] \hat{\beta}_R \\
&- g_n(\beta_0)'W_n \hat{G}_n(\bar{\beta}) - \beta'_0 \hat{G}_n(\bar{\beta})'W_n g_n(\beta_0) + \beta'_0 \hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta})\beta_0 \\
&= g_n(\beta_0)'W_n g_n(\beta_0) - \hat{\beta}'_R[\hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] \hat{\beta}_R \\
(24) \quad &- g_n(\beta_0)'W_n \hat{G}_n(\bar{\beta})\beta_0 - \beta'_0 \hat{G}_n(\bar{\beta})'W_n g_n(\beta_0) + \beta'_0 \hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta})\beta_0.
\end{aligned}$$

The equality is obtained through (21)(22), and note that the right hand side expression in (21) is just the negative of the right hand side of the expression in (22). As in (24), for the estimator $\hat{\beta}_w$ we have the following and the mean value theorem applied to that, where $\bar{\beta}_w \in (\beta_0, \hat{\beta}_w)$,

$$\begin{aligned}
[\sum_{i=1}^n g_i(\hat{\beta}_w)]'W_n [\sum_{i=1}^n g_i(\hat{\beta}_w)] &+ \lambda_2 \|\hat{\beta}_w\|_2^2 = g_n(\beta_0)'W_n g_n(\beta_0) \\
&+ \hat{\beta}'_w[\hat{G}_n(\bar{\beta}_w)'W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta}_w)'W_n \hat{G}_n(\bar{\beta}_w)\beta_0] \\
&+ [\hat{G}_n(\bar{\beta}_w)'W_n g_n(\beta_0) - \hat{G}_n(\bar{\beta}_w)'W_n \hat{G}_n(\bar{\beta}_w)\beta_0]' \hat{\beta}_w \\
&+ \hat{\beta}'_w[\hat{G}_n(\bar{\beta}_w)'W_n \hat{G}_n(\bar{\beta}_w) + \lambda_2 I_p] \hat{\beta}_w \\
&- g_n(\beta_0)'W_n \hat{G}_n(\bar{\beta}_w)\beta_0 - \beta'_0 \hat{G}_n(\bar{\beta}_w)'W_n g_n(\beta_0) \\
(25) \quad &+ \beta'_0 \hat{G}_n(\bar{\beta}_w)'W_n \hat{G}_n(\bar{\beta}_w)\beta_0.
\end{aligned}$$

Then see that by assuming $\bar{\beta}$ to be the same variable in the mean value theorem for both β_w, β_R analysis without losing any generality (i.e. $\bar{\beta} = \bar{\beta}_w$)

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

The asymptotically negligible remainder in the last equality is due to Assumption 2i and Theorem 1i with $\lambda_1 = 0$. Next substitute (26) into (25) to have

$$\begin{aligned}
[\sum_{i=1}^n g_i(\hat{\beta}_w)]'W_n [\sum_{i=1}^n g_i(\hat{\beta}_w)] &+ \lambda_2 \|\hat{\beta}_w\|_2^2 = g_n(\beta_0)'W_n g_n(\beta_0) - \hat{\beta}'_w[\hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] \hat{\beta}_R \\
&- \hat{\beta}'_R[\hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] \hat{\beta}_w + \hat{\beta}'_w[\hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] \hat{\beta}_w \\
&- g_n(\beta_0)'W_n \hat{G}_n(\bar{\beta})\beta_0 \\
(27) \quad &- \beta'_0 \hat{G}_n(\bar{\beta})'W_n g_n(\beta_0) + \beta'_0 \hat{G}_n(\bar{\beta})'W_n \hat{G}_n(\bar{\beta})\beta_0 + o_p(1).
\end{aligned}$$

Now subtract (24) from (27) to have

$$\begin{aligned} & \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_w) \right] + \lambda_2 \|\hat{\beta}_w\|_2^2 - \left(\left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right]' W_n \left[\sum_{i=1}^n g_i(\hat{\beta}_R) \right] + \lambda_2 \|\hat{\beta}_R\|_2^2 \right) \\ (28) \quad & = (\hat{\beta}_w - \hat{\beta}_R)' [\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] (\hat{\beta}_w - \hat{\beta}_R) + o_p(1). \end{aligned}$$

After this important result see that by the definitions of $\hat{\beta}_w, \hat{\beta}_R$

$$\begin{aligned} \lambda_1 \sum_{j=1}^p (\hat{\beta}_{j,R} - \hat{\beta}_{j,w}) & \geq \left(\sum_{i=1}^n g_i(\hat{\beta}_w) \right)' W_n \left(\sum_{i=1}^n g_i(\hat{\beta}_w) \right) + \lambda_2 \|\hat{\beta}_w\|_2^2 \\ (29) \quad & - \left[\left(\sum_{i=1}^n g_i(\hat{\beta}_R) \right)' W_n \left(\sum_{i=1}^n g_i(\hat{\beta}_R) \right) + \lambda_2 \|\hat{\beta}_R\|_2^2 \right]. \end{aligned}$$

Then also see that

$$(30) \quad \sum_{j=1}^p \hat{w}_j (|\hat{\beta}_{j,R}| - |\hat{\beta}_{j,w}|) \leq \sqrt{\sum_{j=1}^p (\hat{w}_j)^2} \|\hat{\beta}_R - \hat{\beta}_w\|_2.$$

Next benefit from (29), with (28)(30) to have

$$\begin{aligned} & (\hat{\beta}_w - \hat{\beta}_R)' [\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p] (\hat{\beta}_w - \hat{\beta}_R) + o_p(1) \leq \lambda_1 \sqrt{\sum_{j=1}^p (\hat{w}_j)^2} \|\hat{\beta}_R - \hat{\beta}_w\|_2. \\ (31) \quad & \end{aligned}$$

See that $Eig_{min}(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p) = Eig_{min}(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta})) + \lambda_2$. Use this in (31) to have

$$\{Eig_{min}(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta})) + \lambda_2\} \|\hat{\beta}_w - \hat{\beta}_R\|_2^2 \leq \lambda_1 \sqrt{\sum_{j=1}^p \hat{w}_j^2} \|\hat{\beta}_R - \hat{\beta}_w\|_2.$$

This results in

$$(32) \quad \|\hat{\beta}_w - \hat{\beta}_R\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^p \hat{w}_j^2} + o_p(1)}{Eig_{min}(\hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta})) + \lambda_2}.$$

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 2ii, 3

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

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

$$(33) \quad \hat{G}_n(\bar{\beta})' W_n \hat{G}_n(\bar{\beta}) = n^2 [G(\beta_0)' W G(\beta_0)] + o_p(n^2).$$

Modify (32) in the following way given the last equality, set $W = \Omega^{-1}$ (since this is the efficient limit weight as shown by Hansen (1982))

$$(34) \quad \|\hat{\beta}_w - \hat{\beta}_R\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^p \hat{w}_j^2} + o_p(1)}{n^2 \text{Eig}_{\min}(G(\beta_0)' \Omega^{-1} G(\beta_0)) + \lambda_2 + o_p(n^2)}.$$

Now we consider the second part of the proof of this theorem. We use GMM ridge formula. Note that from (20)

$$(35) \quad \begin{aligned} \hat{\beta}_R - \beta_0 &= -\lambda_2 [\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}_R) + \lambda_2 I_p]^{-1} \beta_0 \\ &- [\hat{G}_n(\hat{\beta}_R)' W_n \hat{G}_n(\bar{\beta}) + \lambda_2 I_p]^{-1} [\hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0)]. \end{aligned}$$

We try to modify the equation above a little.

In the same way we obtain (33)

$$(36) \quad \hat{G}_n(\hat{\beta}_R)' W_n g_n(\beta_0) = n [G(\beta_0)' W g_n(\beta_0)] + o_p(n).$$

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

$$(37) \quad \begin{aligned} E[g_n(\beta_0)' \Omega^{-1} G(\beta_0) G(\beta_0)' \Omega^{-1} g_n(\beta_0)] &= \text{tr}\{G(\beta_0)' \Omega^{-1} E[g_n(\beta_0) g_n(\beta_0)'] \Omega^{-1} G(\beta_0)\} \\ &= n \text{tr}\{G(\beta_0)' \Omega^{-1} G(\beta_0)\} \\ &\leq n p \text{Eig}_{\max}(\Sigma), \end{aligned}$$

where we use $\Sigma = G(\beta_0)' \Omega^{-1} G(\beta_0)$. Now we modify (35) using (36) (33)

$$(38) \quad \begin{aligned} \hat{\beta}_R - \beta_0 &= -\lambda_2 [n^2 G(\beta_0)' \Omega^{-1} G(\beta_0) + \lambda_2 I_p + o_p(n^2)]^{-1} \beta_0 \\ &- [n^2 G(\beta_0)' \Omega^{-1} G(\beta_0) + \lambda_2 I_p + o_p(n^2)]^{-1} [n G(\beta_0)' \Omega^{-1} g_n(\beta_0)]. \end{aligned}$$

Then see that

$$(39) \quad \begin{aligned} E(\|\hat{\beta}_R - \beta_0\|_2^2) &\leq 2\lambda_2^2 [n^2 \text{Eig}_{\min}(G(\beta_0)' \Omega^{-1} G(\beta_0)) + \lambda_2 + o(n^2)]^{-2} \|\beta_0\|_2^2 \\ &+ 2[n^2 \text{Eig}_{\min}(G(\beta_0)' \Omega^{-1} G(\beta_0)) + \lambda_2 + o(n^2)]^{-2} \\ &\times n^2 E[g_n(\beta_0)' \Omega^{-1} G(\beta_0) G(\beta_0)' \Omega^{-1} g_n(\beta_0)] + o(n^2) \\ &\leq 2[n^2 \text{Eig}_{\min}(G(\beta_0)' \Omega^{-1} G(\beta_0)) + \lambda_2 + o(n^2)]^{-2} \\ &\times [\lambda_2^2 \|\beta_0\|_2^2 + n^3 p \text{Eig}_{\max}(\Sigma) + o_p(n^2)], \end{aligned}$$

where the last inequality is by (37). Now use (34) and (39) to have

$$(40) \quad \begin{aligned} E(\|\hat{\beta}_w - \beta_0\|_2^2) &\leq 2E(\|\hat{\beta}_R - \beta_0\|_2^2) + 2E(\|\hat{\beta}_w - \hat{\beta}_R\|_2^2) \\ &\leq \frac{4\lambda_2^2\|\beta_0\|_2^2 + 4n^3pB + o(n^2) + 2\lambda_1^2E\sum_{j=1}^p\hat{w}_j^2 + o(1)}{[n^2b + \lambda_2 + o_p(n^2)]^2}. \end{aligned}$$

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

Proof of Theorem 3i. To prove the Theorem we need to show the following (Note that by Kuhn-Tucker conditions of (1)),

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

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 zero parameters (i.e. at true values), evaluated at $\tilde{\beta}$. Or we need to show

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

Now set $\eta = \min_{j \in A} |\beta_{j0}|$, $\hat{\eta} = \min_{j \in A} |\hat{\beta}_{j,enet}|$. So

$$(41) \quad \begin{aligned} P[\exists j \in A^c, |2\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^n g_i(\tilde{\beta}))| > \lambda_1^*\hat{w}_j] &\leq \sum_{j \in A^c} P[|2\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^n g_i(\tilde{\beta}))| > \lambda_1^*\hat{w}_j, \hat{\eta} > \eta/2] \\ &+ P[\hat{\eta} \leq \eta/2]. \end{aligned}$$

Then as in p.15 of Zou and Zhang (2009), we can show that

$$(42) \quad \begin{aligned} P[\hat{\eta} \leq \eta/2] &\leq \frac{E\|\hat{\beta}_{enet} - \beta_0\|_2^2}{\eta^2/4} \\ &\leq 16 \frac{\lambda_2^2\|\beta_0\|_2^2 + n^3pB + \lambda_1^2p + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2\eta^2}, \end{aligned}$$

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

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

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

where $M = (\frac{\lambda_1^*}{n^{3+\alpha}})^{\frac{1}{2\gamma}}$.

In (43) we consider the second term on the right hand side. Via inequality (6.8) of Zou and Zhang (2009), and Theorem 2 here

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

Next we can consider the first term on the right hand side of (43)

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

So we try to simplify the term on the right hand side of (45). Now we evaluate

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

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 (46). Note that $\tilde{\beta}$ is consistent if we go through the same steps as in Theorem 1. Then applying Assumption 2ii with Assumption 3 (Uniform Law of Large Numbers)

$$(47) \quad 2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^n g_i(\beta_{A,0}))|^2 \leq 2n^2 \|G(\beta_{A,0})'W \sum_{i=1}^n g_i(\beta_{A,0})\|_2^2 + o_p(n^2).$$

Then via (46)

$$(48) \quad E\left[\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\beta_{A,0})\right)|^2\right] \leq n^3 B + o(n^3),$$

where we use $\lim_{n \rightarrow \infty} (\sum_{i=1}^n g_i(\beta_{A,0})) (\sum_{i=1}^n g_i(\beta_{A,0}))' = n\Omega_*$, and with $W = \Omega_*^{-1}$,

$$(49) \quad B \geq \text{Eigmax}(\Sigma) \geq \text{Eigmax}(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0})).$$

In the same manner we have

$$(50) \quad \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \hat{G}_n(\tilde{\beta})(\tilde{\beta} - \beta_{A,0})|^2 \leq n^4 |G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0})(\tilde{\beta} - \beta_{A,0})|^2 + o_p(n^4).$$

Then by (49), and taking into account (50)

$$(51) \quad \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \hat{G}_n(\tilde{\beta})(\tilde{\beta} - \beta_{A,0})|^2 \leq n^4 B^2 \|\tilde{\beta} - \beta_{A,0}\|_2^2 + o_p(n^4).$$

Now substitute (48)-(51) into

$$(52) \quad E\left[\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})' W_n \left(\sum_{i=1}^n g_i(\tilde{\beta})\right)|^2 I_{\{\hat{\eta} > \eta/2\}}\right] \leq 2B^2 n^4 E(\|\tilde{\beta} - \beta_{A,0}\|_2^2 I_{\{\hat{\eta} > \eta/2\}}) + 2Bn^3 + o(n^4).$$

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

$$(53) \quad \tilde{\beta}(\lambda_2, 0) = \text{argmin}_{\beta} \left\{ \left(\sum_{i=1}^n g_i(\beta_A)\right)' W_n \left(\sum_{i=1}^n g_i(\beta_A)\right) + \lambda_2 \sum_{j \in A} \beta_j^2 \right\}.$$

Then use the arguments in the proof of Theorem 2 (equation (34)) leading to

$$(54) \quad \begin{aligned} \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2 &\leq \frac{\lambda_1^* \max_{j \in A} \hat{w}_j \sqrt{p} + o_p(1)}{n^2 \text{Eigmin}(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0})) + \lambda_2 + o_p(n^2)} \\ &\leq \frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2 b + \lambda_2 + o_p(n^2)}, \end{aligned}$$

where $\text{Eigmin}(G(\beta_{A,0})' \Omega_*^{-1} G(\beta_{A,0})) \geq \text{Eigmin}(G(\beta_0)' \Omega_*^{-1} G(\beta_0)) \geq b$.

Then follow the proof of Theorem 2, for the right hand side term in (52)

$$(55) \quad E(\|\tilde{\beta} - \beta_{A,0}\|_2 I_{\{\hat{\eta} > \eta/2\}}) \leq 4 \frac{\lambda_2^2 \|\beta_0\|_2^2 + n^3 p B + \lambda_1^{*2} (\eta/2)^{-2\gamma} p + o_p(n^2)}{(bn^2 + \lambda_2 + o_p(n^2))^2}.$$

Now combine (42) (43)(44)(45)(52)(55) into (41)

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

Now we have to show that each square bracketed term on the right hand side of the equation (56) converges in probability to zero. We consider each of the right hand side elements in (56). The first square bracketed element converges in probability to zero since by $\lambda_2^2/n \rightarrow 0, \lambda_1^2/n \rightarrow 0, p/n \rightarrow 0$ provides us with η^2 being a constant

$$\begin{aligned}
\frac{\lambda_2^2 \|\beta_0\|_2^2}{n^4} &\rightarrow 0, \\
\frac{Bn^3 p}{n^4} &\rightarrow 0, \\
\frac{\lambda_1^2 p}{n^4} &\rightarrow 0.
\end{aligned}$$

Next we consider the second square bracketed term on the right hand side of (56). See that the dominating term in that expression is stochastic order of

$$O\left(\frac{p}{n} \frac{1}{M^2}\right) \rightarrow 0.$$

The above is true since by Assumption 6 $\frac{\lambda_1^*}{n^{3+\alpha}} n^{\gamma(1-\alpha)} \rightarrow \infty$, and $M = \left[\frac{\lambda_1^*}{n^{3+\alpha}}\right]^{1/2\gamma}$, and 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 (56) are

$$\frac{\lambda_2^2 \|\beta_0\|_2^2}{n} \frac{1}{nM^2} \frac{1}{n^2} \rightarrow 0,$$

by $\lambda_2^2/n \rightarrow 0$, $\frac{p}{nM^2} \rightarrow 0$ by the analysis of the dominating term above. Then also in the same way

$$\frac{\lambda_1^2 p}{M^2 n^4} = \frac{\lambda_1^2}{n} \frac{p}{nM^2} \frac{1}{n^2} \rightarrow 0,$$

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

Now we consider the last square bracketed term in (56). For that term the dominating terms are (the last two terms in (56))

$$O\left(\frac{M^{2\gamma}}{\lambda_1^{*2}} n^3 p\right) \rightarrow 0,$$

$$O(M^{2\gamma} p) \rightarrow 0.$$

The other terms in the last square bracketed term is of smaller order. To show the first result, note that by definition of M

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

since λ_1^* is diverging to infinity. Then for the second result

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

by $\lambda_1^* = o(n^{(2-\alpha)+\frac{\gamma}{2}(\alpha-1)})$, and $0 < \alpha < 1/3$. **Q.E.D.**

Proof of Theorem 3ii. After Theorem 3i result, it suffices to prove

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

Then we can write the following with $\tilde{\beta}(\lambda_2, 0)$ defined in (53)

$$(57) \quad \min_{j \in A} |\tilde{\beta}_j| > \min_{j \in A} |\tilde{\beta}(\lambda_2, 0)_j| - \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2.$$

Also see that

$$(58) \quad \min_{j \in A} |\tilde{\beta}(\lambda_2, 0)_j| > \min_{j \in A} |\beta_{j0}| - \|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2.$$

Combine (57)(58) to have

$$(59) \quad \min_{j \in A} |\tilde{\beta}_j| > \min_{j \in A} |\beta_{j0}| - \|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2 - \|\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)\|_2.$$

Now we consider the last two terms on the right hand side of (59). Similar to derivation of (38)(39) we have

$$(60) \quad \begin{aligned} E\|\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}\|_2 &\leq \frac{4\lambda_2^2\|\beta_{A,0}\|_2^2 + 4n^3pB + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2} \\ &= O(p/n) = o(1). \end{aligned}$$

Then by (54)

$$(61) \quad \|\tilde{\beta}(\lambda_2, 0) - \tilde{\beta}\|_2 \leq \frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2b + \lambda_2 + o_p(n^2)}.$$

See that

$$(62) \quad \frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2b + \lambda_2 + o_p(n^2)} = O(1) (\hat{\eta}/\eta)^{-\gamma} \frac{\lambda_1^* \sqrt{p}}{n^2} \eta^{-\gamma}.$$

Since $p = n^\alpha$, by Assumption 6,

$$(63) \quad \begin{aligned} \frac{\lambda_1^* \sqrt{p}}{n^2} &= o(n^{(2-\alpha) + \frac{\gamma}{2}(\alpha-1) + \alpha/2 - 2}) \\ &= o(n^{-\alpha/2 + \frac{\gamma}{2}(\alpha-1)}) = o(1), \end{aligned}$$

since $0 < \alpha < 1/3, \gamma > 0$.

Then by Theorem 2

$$(64) \quad \begin{aligned} E\left[\left(\hat{\eta}/\eta\right)^2\right] &\leq 2 + \frac{2}{\eta^2} E[\hat{\eta} - \eta]^2 \\ &\leq 2 + \frac{2}{\eta^2} E\|\hat{\beta}(\lambda_2, \lambda_1) - \beta_0\|_2^2 \\ &\leq 2 + \frac{8}{\eta^2} \left[\frac{\lambda_2^2\|\beta_0\|_2^2 + n^3pB + \lambda_1^2p + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2} \right] \\ &= O(1), \end{aligned}$$

by $\lambda_1^2/n \rightarrow 0, \lambda_2^2/n \rightarrow 0, p/n \rightarrow 0$.

Substitute (63)(64) into (62) to have

$$(65) \quad \frac{\lambda_1^* \hat{\eta}^{-\gamma} \sqrt{p}}{n^2b + \lambda_2 + o_p(n^2)} \xrightarrow{p} 0,$$

Now use (65) in (61) and combine this with (60) to have

$$\min_{j \in A} |\tilde{\beta}_j| > \eta - o_p(1).$$

Then we obtain the desired result. **Q.E.D**

Proof of Theorem 4. We now prove the limit result. First, define

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

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

$$\hat{\beta}_{aenet,A} \xrightarrow{p} \beta_{A,0},$$

and by (60)

$$\tilde{\beta}(\lambda_2, 0) \xrightarrow{p} \beta_{A,0}.$$

Next by Assumption 2.ii, and Assumption 3, and considering the results about consistency we have

$$(66) \quad \frac{\hat{G}(\hat{\beta}_{aenet,A})'}{n} W_n \frac{\hat{G}(\hat{\beta}_{aenet,A})}{n} - \frac{\hat{G}(\tilde{\beta}(\lambda_2, 0))'}{n} W_n \frac{\hat{G}(\tilde{\beta}(\lambda_2, 0))}{n} \xrightarrow{p} 0.$$

Next note that

$$\begin{aligned} \delta' [I + \lambda_2 (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{-1}] &\times (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{1/2} n^{-1/2} (\tilde{\beta} - \frac{\beta_{A,0}}{1 + \lambda_2/n}) \\ &= [\delta' (I + \lambda_2 (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{-1}) (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{1/2} \\ &\times n^{-1/2} \frac{\lambda_2 \beta_{A,0}}{n + \lambda_2}] \\ &+ [\delta' (I + \lambda_2 (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{-1}) (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{1/2} \\ &\times n^{-1/2} (\tilde{\beta} - \tilde{\beta}(\lambda_2, 0))] \\ &+ [\delta' (I + \lambda_2 (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{-1}) (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{1/2} \\ (67) \quad &\times n^{-1/2} (\tilde{\beta}(\lambda_2, 0) - \beta_{A,0})]. \end{aligned}$$

To explain the last term in the equation above via (20), and with simple matrix algebra

$$\begin{aligned} \{\lambda_2 [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{-1} + I\} &\times (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{1/2} n^{-1/2} [\tilde{\beta}(\lambda_2, 0) - \beta_{A,0}] \\ &= -\lambda_2 n^{-1/2} [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{-1/2} \beta_{A,0} \\ &- n^{-1/2} [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{-1/2} \\ (68) \quad &\times [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n g_n(\beta_{A,0})], \end{aligned}$$

where $g_n(\beta_{A,0}) = \sum_{i=1}^n g_i(\beta_{A,0})$.

Via Theorem 3, and also using (66), with probability tending to one , $z_n = T_1 + T_2 + T_3$, where

$$T_1 = [\delta' [(I + \lambda_2 (\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0)))^{-1})] [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{1/2}$$

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

$$\begin{aligned} T_2 &= \delta' [I + \lambda_2 [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{-1}] [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{1/2} \\ & \times n^{-1/2} (\tilde{\beta} - \tilde{\beta}(\lambda_2, 0)). \end{aligned}$$

$$T_3 = \delta' [\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \hat{G}(\tilde{\beta}(\lambda_2, 0))]^{-1/2} \left[\hat{G}(\tilde{\beta}(\lambda_2, 0))' W_n \sum_{i=1}^n \frac{g_i(\beta_{A,0})}{n^{1/2}} \right].$$

Consider T_1 , use Assumption 4i, Assumption 4ii

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

via $\lambda_2 = o(n^{1/2})$. Consider T_2 similar to the above analysis and (54)

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

by (65) and $\lambda_2 = o(n^{1/2})$.

For the term T_3 we benefit from Liapunov Central Limit Theorem. By Assumptions 2 and 3

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

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

$$\begin{aligned} \sum_{i=1}^n E R_i^2 &= n^{-1} \sum_{i=1}^n E [\delta' [G(\beta_{A,0})' W G(\beta_{A,0})]^{-1/2} G(\beta_{A,0})' W g_i(\beta_{A,0}) g_i(\beta_{A,0})' W G(\beta_{A,0}) [G(\beta_{A,0})' W G(\beta_{A,0})]^{-1/2} \delta] \\ &= \delta' [G(\beta_{A,0})' W G(\beta_{A,0})]^{-1/2} G(\beta_{A,0})' W [n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})'] W G(\beta_{A,0}) [G(\beta_{A,0})' W G(\beta_{A,0})]^{-1/2} \delta. \end{aligned}$$

Then set $W = \Omega_A^{-1}$, and use the definition $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n E g_i(\beta_{A,0}) g_i(\beta_{A,0})' = \Omega_A$. Take the limit of the term above

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

Next we show $\sum_{i=1}^n E |R_i|^{2+l} \rightarrow 0$, for $l > 0$. See that by Assumptions 4 and 5

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

The desired result then follows from $z_n = T_1 + T_2 + T_3$ with probability one.

Q.E.D.

References

- Alfaro, L. & Kalemli-Ozcan, S, & Volosoych V. (2008) "Why does not capital flow from rich to poor countries? An empirical investigation", *Review of Economics and Statistics* **90**, 347-368.
- Andrews, D. (1994), "Empirical process methods in econometrics" in *Handbook of Econometrics* **4**, 2247-2294.
- Andrews, D.W.K., and B. Lu (2001), "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models", *Journal of Econometrics* **101**, 123-165.
- Caner, M. (2008), "Nearly Singular Design in GMM and Generalized Empirical Likelihood Estimators", *Journal of Econometrics*, 144, 511-523.
- Caner, M., (2009), "Lasso-type GMM estimator" *Econometric Theory*, **25** 270-290.
- Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press.
- Fan, J. & Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties", *Journal of The American Statistical Association* **96**, 1348-1360.
- Fan, J., Peng, H. & Huang, T. (2005), "Semilinear high dimensional model for normalization of microarray data: a theoretical analysis and partial consistency (with discussion)", *Journal of the American Statistical Association* **100**, 781-813.
- Hansen, L. P. (1982), "Large sample properties of generalized method of moment estimators", *Econometrica* **50** , 1029-1054.
- Huang, J. & Horowitz, J. & Ma S. (2008), "Asymptotic properties of bridge estimators in sparse high-dimensional regression models" *The Annals of Statistics* **36**, 587-613.
- Huber, P. (1988), "Robust regression: Asymptotics, conjectures and monte carlo" *The Annals of Statistics* **1**, 799-821.
- Kitamura, Y. & Stutzer, M. (1997), "An information-theoretic alternative to generalized method of moments estimation", *Econometrica* **65**, 861-874.
- Knight, K. & Fu, W. (2000), "Asymptotics for lasso type estimators" *The Annals of Statistics* **28**, 1356-1378.
- Knight, K. (2008), "Shrinkage estimation for nearly-singular designs", *Econometric Theory* **24**, 323-338.
- Lam, C. & Fan, J. (2007), "Profile-kernel likelihood inference with diverging number of parameters" *The Annals of Statistics* to appear.
- Newey, W. & McFadden, D. (1994), *Large Sample estimation and hypothesis testing Handbook of Econometrics* **4**.
- Owen, A. (2001), *Empirical Likelihood*, Chapman Hall.
- Portnoy, S. (1984), "Asymptotic behavior of M-estimators of p-regression

parameters when p^2/n is large. I. consistency" *The Annals of Statistics* **12**, 1298-1309.

Qin, J. & Lawless, J. (1994), " Empirical Likelihood and general estimating equations", *The Annals of Statistics* **22**, 300-325.

Schennach, S. (2007), "Point Estimation with Exponentially Tilted Empirical Likelihood", *The Annals of Statistics* **35**, 634-672.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. & Wellner, J. (1996), *Weak Convergence and Empirical Processes* Springer-Verlag.

Wang, H., Li, R., & Tsai, C., (2007), "Tuning parameter selectors for the smoothly clipped absolute deviation method", *Biometrika* **94**, 553-568.

Zhang, H. & W. Lu (2007), "Adaptive Lasso for Cox's proportional hazards model", *Biometrika* **37** 1-13.

Zou, H., Hastie, T., & Tibshirani, R. (2007), "On the degrees of freedom of the lasso", *The Annals of Statistics* **35**, 2173-2192.

Zou, H. & Zhang, H. (2009), "On the adaptive elastic-net with a diverging number of parameters", *The Annals of Statistics*, to appear.

MEHMET CANER
DEPARTMENT OF ECONOMICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC 27695
E-MAIL: mcaner@ncsu.edu

HAO HELEN ZHANG
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC 27695
E-MAIL: hzhang2@stat.ncsu.edu