Econometric Theory, **29**, 2013, 857–904.
doi:10.1017/S0266466612000783

ADAPTIVE GMM SHRINKAGE ESTIMATION WITH CONSISTENT MOMENT SELECTION

ZHIPENG LIAO
University of California, Los Angeles

This paper proposes a generalized method of moments (GMM) shrinkage method to efficiently estimate the unknown parameters θ_o identified by some moment restrictions, when there is another set of possibly misspecified moment conditions. We show that our method enjoys oracle-like properties; i.e., it consistently selects the correct moment conditions in the second set and at the same time, its estimator is as efficient as the GMM estimator based on all correct moment conditions. For empirical implementation, we provide a simple data-driven procedure for selecting the tuning parameters of the penalty function. We also establish oracle properties of the GMM shrinkage method in the practically important scenario where the moment conditions in the first set fail to strongly identify θ_o. The simulation results show that the method works well in terms of correct moment selection and the finite sample properties of its estimators. As an empirical illustration, we apply our method to estimate the life-cycle labor supply equation studied in MaCurdy (1981, *Journal of Political Economy* 89(6), 1059–1085) and Altonji (1986, *Journal of Political Economy* 94(3), 176–215). Our empirical findings support the validity of the instrumental variables used in both papers and confirm that wage is an endogenous variable in the labor supply equation.

1. INTRODUCTION

The generalized method of moments (GMM) is a popular methodology for estimating structural equations in economics and finance. It is particularly attractive when moment conditions appear naturally in model formulation. The statistical properties of the GMM estimators rely heavily on the quality of these moment conditions. For example, the GMM estimator based on misspecified moment conditions is inconsistent. On the other hand, including strong and valid moment conditions in GMM can help to reduce finite-sample bias and improve efficiency of the GMM estimator. Hence, whenever an empirical researcher has a set of

The author is deeply indebted to Peter Phillips and Xiaohong Chen for guidance, inspiration, and encouragement. He is also grateful to Donald Andrews for valuable suggestion. The author benefited from insightful comments made by the co-editor, two anonymous referees, Joseph Altonji, Xu Cheng, Jinyong Hahn, Yuichi Kitamura, Oliver Linton, Xiaoxia Shi, Edward Vytlacil, and Bin Wang. Financial support from a Carl Arvid Anderson Prize of the Cowles Foundation is acknowledged. Address correspondence to Zhipeng Liao, Department of Economics, UC Los Angeles, 8279 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095; e-mail: zhipeng.liao@econ.ucla.edu.

© Cambridge University Press 2013

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

moment conditions and there is no prior information about their validity, it is important to have some procedure to select the correctly specified moment conditions in that set and include them in estimation. This paper proposes a new method to achieve that goal.

Specifically, we are interested in estimating some unknown parameters θ_o identified by the moment restrictions

$$\mathbb{E}\left[g_q(Z,\theta_o)\right] = 0, \tag{1.1}$$

where $\{Z_i\}_{i\leq n}$ is stationary and ergodic, Z is used generically for Z_i, the subscript q of $g_q(\cdot,\cdot)$ denotes the number of moment conditions in (1.1), and $g_q(\cdot,\cdot) : R^{d_z} \times R^{d_\theta} \to R^q$. The moment conditions in (1.1) can be used to estimate θ_o consistently. Suppose there is another set of possibly misspecified moment conditions

$$\mathbb{E}[g_k(Z,\theta_o)] \stackrel{?}{=} 0, \tag{1.2}$$

where "$\stackrel{?}{=}$" signifies that equality may hold for some elements but not others, the subscript k of $g_k(\cdot,\cdot)$ denotes the number of moment conditions in (1.2), and $g_k(\cdot,\cdot) : R^{d_z} \times R^{d_\theta} \to R^k$.[1] When the moment conditions in set-2 (or some of them) are correctly specified, including them in estimation can improve the asymptotic efficiency of the estimator for θ_o. However, if they are misspecified, then using these moment conditions will lead to inconsistent estimation. The goal of this paper is to consistently select the correct moment conditions in set-2 and automatically include them in GMM estimation to improve the efficiency of estimating θ_o.

To reduce the risk of misspecification, one can employ the Sargan/Hansen test (Sargan, 1958; Eichenbaum, Hansen, and Singleton, 1988) to check the validity of the set-2 moment conditions. In addition to the Sargan/Hansen test, there are other moment selection procedures in the literature for empirical researchers to use. For example, Andrews (1999) defines moment selection criterion (MSC) using the J-test statistic and shows that consistent moment selection can be achieved by choosing the moment selection vector minimizing the MSC.[2] Based on the J-test statistic, Andrews also proposes downward testing (DT) and upward testing (UT) moment selection procedures and shows their consistency. Hong, Preston, and Shum (2003) construct the MSC, DT, and UT procedures using the generalized empirical likelihood (GEL) statistic and show that these procedures are consistent in moment selection. The above methods perform moment selection in a stepwise manner and break the moment selection and efficient estimation into two separate procedures. Moreover, when the number of moment conditions in set-2 is large, there may be too many candidate subsets of moment conditions for these methods to investigate, which makes them computationally intensive in practice. Finally, all of these methods are nonetheless subject to pretest bias and post-model selection inferential problems (Leeb and Pötscher, 2005).

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

This paper provides a new method for moment selection. The new method attaches a penalty function to the GMM criterion, and the resulting estimator of θ_o is called as the GMM shrinkage estimator. Our method embeds the moment selection in GMM estimation, and once a certain moment condition is selected, it will be included automatically in estimating θ_o. Hence, our method not only selects the correct moment conditions in set-2 in one step, but also deals with the moment selection and efficient estimation simultaneously. Under some regularity conditions, we show that the GMM shrinkage estimator of θ_o is root-n consistent and asymptotically normal. Moreover, we show that consistent moment selection is automatically achieved in the penalized GMM estimation, and the GMM shrinkage estimator of θ_o is asymptotically oracle-efficient (i.e., as efficient as the oracle GMM estimator based on all valid moment conditions). While apparently effective asymptotically, our method does not avoid the post-model selection inference issues in finite samples (Leeb and Pötscher, 2008). On the other hand, our method radically simplifies empirical research with many (but a fixed number of) moment conditions where moment selection has to be performed to reduce the variance of the GMM estimator.

There are some recent works in the literature that are related to this paper. Caner (2009) studies the GMM shrinkage estimation with bridge penalty function. Under some regularity conditions, he shows that the GMM shrinkage estimator is consistent and has a nonstandard limiting distribution. However, no model selection results are derived in this paper.[3] Caner and Zhang (2012) study GMM shrinkage estimation with adaptive elastic net penalty function in the scenario where the number of structural coefficients d_θ grows with the sample size. Under the restriction $d_\theta = o(n^{1/3})$ and some other regularity conditions, they show that their GMM shrinkage estimate is consistent in model selection. Our paper is different from the above two papers in the basic setup, because we start with valid and potentially invalid moment conditions, while the moment conditions in Caner and Caner and Zhang are assumed to be valid. This paper is also different from the Lasso-type instrumental variable (IV) selection papers (see, e.g., Belloni, Chen, Chernozhukov, and Hansen, 2012), because all IVs in these works are assumed to be valid, and their goals are to select the IVs with strong information. In a more recent work, Gautier and Tsybakov (2011) propose a Danzig selector-based IV estimator in high-dimensional models. They derive an upper bound of estimation error for the structural coefficients under the assumption that all IVs are valid.

The rest of this paper is organized as follows. Section 2 describes our method and gives some examples from applied econometrics that arise naturally from the framework of this paper. Section 3 establishes the main asymptotic properties of the GMM shrinkage estimators. We also provide simple and data-dependent tuning parameters that are easy to compute in practice. Section 4 studies the GMM shrinkage estimation when the moment conditions in set-1 fail to strongly identify θ_o. Section 5 conducts several Monte Carlo experiments to check the finite sample properties of our method. Section 6 applies our method to a life-cycle labor supply model to illustrate how the GMM shrinkage method works with real data.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Section 7 concludes the paper. Proofs and technical derivations are included in the Appendix.

2. GMM SHRINKAGE ESTIMATION AND SOME EXAMPLES

To incorporate moment selection into estimation, we first introduce a set of auxiliary unknown parameters β_o and reparametrize the moment conditions in set-2 to be

$$\mathbb{E}[g_k(Z,\theta_o)-\beta_o]=0. \tag{2.1}$$

From the above expression, we see that if the jth ($j=1,\ldots,k$) moment condition in (1.2) is correctly specified (or misspecified), then $\beta_{o,j}=0$ (or $\beta_{o,j}\neq 0$). Hence, the zero/nonzero components in β_o can be used to identify the correctly specified/misspecified moment conditions in set-2, and consistent moment selection is equivalent to consistent selection of the zero components in β_o.[4]

We can stack the moment conditions in (1.1) and (2.1) to get

$$\mathbb{E}[\rho(Z,\theta_o,\beta_o)]\equiv\mathbb{E}\begin{bmatrix} g_q(Z,\theta_o) \\ g_k(Z,\theta_o)-\beta_o \end{bmatrix}=0; \tag{2.2}$$

then the GMM shrinkage estimator $\left(\widehat{\theta}_n,\widehat{\beta}_n\right)$ is defined as

$$\left(\widehat{\theta}_n,\widehat{\beta}_n\right)=\underset{(\theta,\beta)\in\Theta\times\mathcal{B}}{\arg\min}\left[\frac{\sum_{i=1}^n\rho(Z_i,\theta,\beta)}{\sqrt{n}}\right]'W_n\left[\frac{\sum_{i=1}^n\rho(Z_i,\theta,\beta)}{\sqrt{n}}\right]+n\sum_{j=1}^k\widehat{P}_{\lambda_n}(\beta_j), \tag{2.3}$$

where $\Theta\times\mathcal{B}$ is the parameter space where (θ_o,β_o) lies, W_n is a $(q+k)\times(q+k)$ weight matrix, and λ_n is the tuning parameter in some general penalty function $\widehat{P}_{\lambda_n}(\cdot)$. The success of our method in simultaneous moment selection and efficient estimation relies on the "oracle properties" of the shrinkage techniques. That is, if $\beta_{o,j}=0$ for some $j\in\{1,\ldots,k\}$, our method will estimate $\beta_{o,j}$ as zero with probability approaching 1 (w.p.a.1). When $\beta_{o,j}$ is estimated as zero w.p.a.1, the information contained in the jth moment condition of (1.2) is automatically used in estimating θ_o w.p.a.1. On the other hand, the nonzero components in β_o are consistently estimated, and their estimators are nonzero w.p.a.1. Hence, our method can consistently distinguish the zero and nonzero components in β_o and is consistent in moment selection. Moreover, it estimates θ_o as if we knew all potentially correct moment conditions in set-2.

There are many popular choices for the penalty function $\widehat{P}_{\lambda_n}(\cdot)$ in the statistics literature. For example, the bridge penalty is defined as $\widehat{P}_{\lambda_n}(\beta)=\lambda_n|\beta|^\gamma$, where $\gamma\in(0,1)$; the adaptive Lasso penalty is defined as $\widehat{P}_{\lambda_n}(\beta)=\lambda_n\widehat{\omega}_\beta|\beta|$, where $\widehat{\omega}_\beta=|\widehat{\beta}_{1st}|^{-\omega}$, $\omega>0$ and $\widehat{\beta}_{1st}$ is some first-step consistent estimator of β_o; and

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

the smoothly clipped absolute deviation (SCAD) penalty is defined as

$$\widehat{P}_{\lambda_n}(\beta) = \begin{cases} \lambda_n|\beta| & |\beta| \le \lambda_n \\ \dfrac{\lambda_n a|\beta|}{a-1} - \dfrac{\beta^2+\lambda_n^2}{2(a-1)} & \lambda_n < |\beta| \le a\lambda_n \\ \dfrac{(a+1)\lambda_n^2}{2} & a\lambda_n < |\beta| \end{cases}, \tag{2.4}$$

where a is some positive real number strictly larger than 2. The above penalty functions differ in their empirical implementations, although the related GMM shrinkage estimators may have the same asymptotic properties (see the results in Section 3). In this regard, the adaptive Lasso penalty is simple and computationally attractive because it is a convex function of β, which together with the convexity of the GMM criterion makes the penalized GMM criterion function convex in β. In the variable selection literature (i.e., when an investigator seeks to select the relevant variables to appear in the statistic model), Knight and Fu (2000) show that least squares (LS) shrinkage estimation based on the bridge penalty has positive possibility of shrinking the estimators of zero regression coefficients toward zero. Zou (2006) shows that the LS shrinkage estimator based on the adaptive Lasso penalty has the oracle properties. In a more general framework, Fan and Li (2001) study penalized maximum likelihood estimation (PMLE) using the SCAD penalty, and they establish the oracle properties of their procedure in variable selection.

Let $\mathcal{S}_\beta \equiv \{j : \beta_{o,j} \neq 0,\ j = 1, \ldots, k\}$ and $\mathcal{S}_{\beta,n} \equiv \{j : \widehat{\beta}_{n,j} \neq 0,\ j = 1, \ldots, k\}$ be the index sets of the nonzero components in β_o and $\widehat{\beta}_n$ respectively. Under some regularity conditions, we show that the GMM shrinkage estimator $\left(\widehat{\theta}_n, \widehat{\beta}_n\right)$ enjoys oracle-like properties in the sense that

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_{\beta,n} = \mathcal{S}_\beta\right) = 1 \tag{2.5}$$

and

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_o\right) \to_d N(0, \Sigma^*), \tag{2.6}$$

where Σ^* is the asymptotic variance-covariance matrix of the optimally weighted oracle GMM estimator based on all valid moment conditions. When the data are independent and identically distributed (i.i.d.), the inverse of Σ^* equals the semiparametric efficiency bound of θ_o, implied by all correct moment conditions. It is clear that the results in (2.5) and (2.6) imply both consistent moment selection and efficient estimation.

We use two well-known examples to illustrate the moment selection issue and show how our method can be applied in practice. The first example is a dynamic panel model with fixed effects. In this model, different specification assumptions give different sets of moment conditions. As noted in Arellano and Bover (1995)

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

and Blundell and Bond (1998), the first-differenced moment conditions may contain weak information about the structural coefficients. On the other hand, the moment conditions based on the initial value are strong, but their validity depends on a stationarity assumption.

Example 2.1 (Dynamic panel model)
Consider the dynamic panel model with fixed effects

$$Y_{i,t} = Y_{i,t-1}\theta_{1,o} + X_{i,t}'\theta_{2,o} + \varepsilon_{i,t} \quad \text{and} \quad \varepsilon_{i,t} = \nu_i + u_{i,t}, \tag{2.7}$$

where $|\theta_{1,o}| < 1$, $Y_{i,t}$ is the dependent variable of individual i at the beginning of period t, $X_{i,t}$ is a set of predetermined control variables, ν_i is an unobservable individual effect, and $u_{i,t}$ is a time-varying error term. Under the assumption

$$\mathbb{E}[u_{i,t}Y_{i,0}] = 0, \qquad \mathbb{E}[u_{i,t}\nu_i] = 0, \qquad \mathbb{E}[u_{i,t}u_{i,s}] = 0 \quad \text{and}$$
$$\mathbb{E}\left[(X_{i,1}, \ldots, X_{i,t-1})u_{i,t}\right] = 0 \tag{2.8}$$

for all i, t, and $t \neq s$, one can estimate the equation in (2.7) using GMM with the moment conditions

$$\mathbb{E}[(Y_{i,0}, \ldots, Y_{i,t-2})\Delta\varepsilon_{i,t}] = 0 \quad \text{and} \quad \mathbb{E}[(X_{i,1}, \ldots, X_{i,t-2})\Delta\varepsilon_{i,t}] = 0. \tag{2.9}$$

However, Arellano and Bover (1995) argue that the GMM estimator based on (2.9) may have large finite-sample bias, because the moment conditions in (2.9) are weak if $\theta_{1,o}$ is close to one. They suggest estimating the equation in (2.7) by GMM with moment conditions in (2.9) and the moment conditions

$$\mathbb{E}[\varepsilon_{i,t}\Delta Y_{i,t-1}] \stackrel{?}{=} 0, \tag{2.10}$$

which is implied by (2.8) and the specification assumptions

$$\mathbb{E}[\nu_i \Delta X_{i,t-1}] = 0 \quad \text{and} \quad \mathbb{E}[\nu_i Y_{i,1}] = \mathbb{E}[\nu_i Y_{i,0}]$$

for all t and i. In this example, the moment conditions in (2.9) and (2.10) are our set-1 and set-2 moment conditions, respectively. We can use our method to pick up the valid moment conditions in (2.10) and automatically include them in the estimation of $(\theta_{1,o}, \theta_{2,o})$.

Our second example is a linear IV model. Empirical researchers sometimes can find credibly valid IVs from a so-called natural experiment. For example, one can use quarter of birth as the IV for education (Angrist and Krueger, 1991) and rainfall as the IV for economic growth (Miguel, Satyanath, and Sergenti, 2004).[5] On the other hand, they may have other candidate IVs that are strongly correlated with the endogenous variables but may not be exogenous.

Example 2.2 (Linear IV model)
Consider the model

$$Y_i = X_i\theta_{1,o} + W_i'\theta_{2,o} + u_i, \tag{2.11}$$
$$X_i = Z_{1,i}\pi_{1,o} + Z_{2,i}'\pi_{2,o} + W_i'\delta_o + v_i, \tag{2.12}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

where Y_i, X_i are scaler endogenous variables, W_i contains a set of exogenous variables, and $Z_{1,i}$ and $Z_{2,i}$ are candidate IVs for X_i. Suppose that we have the following moment conditions:

$$\mathbb{E}[u_i Z_{1,i}] = 0 \quad \text{and} \quad \mathbb{E}[u_i W_i] = 0, \tag{2.13}$$

to identify and consistently estimate $\theta_o = (\theta_{1,o}, \theta_{2,o})$. It is clear that $Z_{2,i}$ could also be valid IVs for estimating θ_o under the condition $\mathbb{E}[u_i Z_{2,i}] \stackrel{?}{=} 0$. Moreover, if X_i is exogenous, then the generalized LS (GLS) estimator is consistent and more efficient. The GLS estimator of $(\theta_{1,o}, \theta_{2,o})$ can be viewed as a GMM estimator based on the moment conditions in (2.13) and the possibly misspecified moment condition $\mathbb{E}[u_i X_i] \stackrel{?}{=} 0$. In this example, set-1 moment conditions are in (2.13), while the moment conditions constructed using $Z_{2,i}$ and X_i are in set-2. Our method can be used to check whether $Z_{2,i}$ are valid IVs. Moreover, our method can automatically detect the endogeneity of X_i. The GMM shrinkage estimator will asymptotically become the GLS estimator if the moment condition $\mathbb{E}[u_i X_i] \stackrel{?}{=} 0$ is valid.

3. ASYMPTOTIC PROPERTIES OF THE GMM SHRINKAGE ESTIMATOR

This section establishes the asymptotic properties of the GMM shrinkage estimator. For ease of notation, we sort the elements in β_o in the following way: $\beta_o' = (\beta_{o,-}', \beta_{o,+}')$, where $\beta_{o,-} = 0$ and the elements in $\beta_{o,+}$ are nonzero. Let k_o denote the number of valid moment conditions in set-2. By definition, we know that $\beta_{o,-}$ and $\beta_{o,+}$ are k_o dimensional and $k - k_o$ dimensional vectors, respectively. We define $\alpha' = (\theta', \beta')$ and

$$m(\alpha) \equiv \mathbb{E}[\rho(Z, \theta, \beta)] \equiv \mathbb{E}\begin{bmatrix} g_q(Z, \theta) \\ g_k(Z, \theta) - \beta \end{bmatrix} \equiv \begin{bmatrix} G_q(\theta) \\ G_k(\theta) - \beta \end{bmatrix}$$

for any $(\theta, \beta) \in \Theta \times \mathcal{B}$. We use $\|\cdot\|$ to denote the Euclidean norm in the Euclidean space.

3.1. Consistency and the Rate of Convergence

We first present and discuss the sufficient conditions for consistency of $\widehat{\alpha}_n$. The assumptions imposed on the moment functions are similar to these ensuring the consistency of the GMM estimator, while some extra conditions are needed to make sure that attaching a penalty function to the GMM criterion function will not lead to inconsistent estimation.

Assumption 3.1. (i) $G_k(\theta)$ is continuous in θ and for any $\varepsilon > 0$, there exists some $\delta_\varepsilon > 0$ such that

$$\inf_{\{\theta \in \Theta : \|\theta - \theta_o\| \geq \varepsilon\}} \left\| G_q(\theta) \right\| > \delta_\varepsilon; \tag{3.1}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

(ii) the uniform law of large numbers (ULLN) holds

$$\sup_{\theta\in\Theta}\left[n^{-1}\sum_{i=1}^{n}\left\{g_l(Z_i,\theta)-\mathbb{E}[g_l(Z_i,\theta)]\right\}\right]=o_p(1) \tag{3.2}$$

for $l=q,k$[6]; (iii) W_n is a symmetric and real matrix and its eigenvalues are bounded from below and above by some fixed finite positive constants for all n; (iv) the penalty function $\widehat{P}_{\lambda_n}(\cdot)$ is nonnegative and $\widehat{P}_{\lambda_n}(\beta_{o,j})=o_p(1)$ for $j=1,\ldots,k$.

Condition (3.1) in Assumption 3.1.(i) is the identifiable uniqueness condition for θ_o. By definition $\beta_o=G_k(\theta_o)$, thus β_o is locally uniquely identified under (3.1) and the continuity of $G_k(\theta)$. Assumption 3.1.(ii) is a high-level condition, because it does not specify the data structure and the properties of the moment functions. The advantage of this high-level condition is that it makes our results applicable to models with a general data structure and general moment functions (e.g., nonsmooth moment functions). Assumption 3.1.(iii) is also a high-level condition, because it does not specify the form of the weight matrix W_n. It is clear that when W_n is an identity matrix, this assumption holds automatically. Assumption 3.1.(iv) implies that the shrinkage effect of the penalty function on the estimators of nonzero moment selection coefficients (i.e., $\beta_{o,+}$) converges in probability to zero as $n\to\infty$. This condition includes the case that $\widehat{P}_{\lambda_n}(\beta_{o,j})=0$ for $j=1,\ldots,k$ as a special example.

LEMMA 3.1. *Under Assumption 3.1, the GMM shrinkage estimator defined in (2.3) is consistent, i.e.,* $\left\|\widehat{\alpha}_n-\alpha_o\right\|=o_p(1)$.

Let $\mathcal{S}_\beta^c$ and $\mathcal{S}_{\beta,n}^c$ denote the complement of the index sets $\mathcal{S}_\beta$ and $\mathcal{S}_{\beta,n}$ respectively. From the consistency of $\widehat{\beta}_n$, we deduce that if $j\in\mathcal{S}_\beta$, then $\widehat{\beta}_{n,j}$ will be estimated as nonzero w.p.a.1 and we have $j\in\mathcal{S}_{\beta,n}$ w.p.a.1. Hence, under Assumption 3.1, the misspecified moment conditions in set-2 will not be selected asymptotically. However, consistent moment selection also requires that if $j\in\mathcal{S}_\beta^c$, then $j\in\mathcal{S}_{\beta,n}^c$ w.p.a.1. The latter result cannot be deduced from the consistency of $\widehat{\beta}_n$, because what we need to show is $\widehat{\beta}_{n,j}$ ($j\in\mathcal{S}_\beta^c$) concentrates on zero w.p.a.1, while the consistency only indicates that $\widehat{\beta}_{n,j}$ ($j\in\mathcal{S}_\beta^c$) concentrates on local neighborhoods of zero w.p.a.1.

Remark 3.1. In Corollary B.1 (in Appendix B), we show that if $\lambda_n=o(1)$, then the bridge penalty, adaptive Lasso penalty and SCAD penalty functions satisfy Assumption 3.1.(iv). Let W_o denote the probability limit of W_n; we next show that $\lambda_n=o(1)$ is also a necessary condition for the consistency of the GMM shrinkage estimator based on these penalty functions. First, note that if $\lambda_n\to\infty$, then $\widehat{P}_{\lambda_n}(\beta)\to_p\infty$ unless $\beta=0$. Hence, if $\lambda_n\to\infty$, we can invoke the epi-convergence theorem in Geyer (1994) or Knight (1999) to deduce that the GMM

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

shrinkage estimator $\widehat{\alpha}_n$ satisfies $\widehat{\beta}_n \to_p 0$ and

$$\widehat{\theta}_n \to_p \arg\min_{\theta\in\Theta} \left\{\mathbb{E}[\rho(Z,\theta,0)]' W_o \mathbb{E}[\rho(Z,\theta,0)]\right\}. \tag{3.3}$$

From (3.3), we see that $\widehat{\theta}_n$ is inconsistent if β_o is a nonzero vector. On the other hand, if $\lambda_n \to \lambda_0 \in (0,\infty)$, then using the argmax continuous mapping theorem (ACMT) we can show that the GMM shrinkage estimator based on the bridge penalty satisfies

$$\widehat{\alpha}_n \to_p \arg\min_{\alpha\in\mathcal{A}} \left\{ m'(\alpha) W_o m(\alpha) + \lambda_0 \sum_{j=1}^{k} |\beta_j|^{\gamma} \right\}, \tag{3.4}$$

where $\alpha' = (\theta', \beta')$ and $\mathcal{A} = \Theta \times \mathcal{B}$, and the GMM shrinkage estimator based on the SCAD penalty satisfies

$$\widehat{\alpha}_n \to_p \arg\min_{\alpha\in\mathcal{A}} \left\{ m'(\alpha) W_o m(\alpha) + \sum_{j=1}^{k} P_{\lambda_0}(\beta_j) \right\}, \tag{3.5}$$

where

$$P_{\lambda_0}(\beta_j) = \begin{cases} \lambda_0 |\beta_j| & |\beta_j| \le \lambda_0 \\ \dfrac{\lambda_0 a |\beta_j|}{a-1} - \dfrac{\beta_j^2 + \lambda_0^2}{2(a-1)} & \lambda_0 < |\beta_j| \le a\lambda_0 \\ \dfrac{(a+1)\lambda_0^2}{2} & a\lambda_0 < |\beta_j| \end{cases}. \tag{3.6}$$

Using the epi-convergence theorem, we can show that the GMM shrinkage estimator based on the adaptive Lasso penalty satisfies

$$\widehat{\alpha}_n \to_p \arg\min_{\alpha\in\mathcal{A}} \begin{cases} m'(\alpha) W_o m(\alpha) + \lambda_0 \sum\limits_{j\in\mathcal{S}_\beta} \frac{|\beta_j|}{|\beta_{o,j}|^{\omega}} & if\ \beta_j = 0\ \forall j \in \mathcal{S}_\beta^c \\ \infty & \text{otherwise} \end{cases}. \tag{3.7}$$

From the results in (3.4), (3.5), and (3.7), we see that if $\lambda_n \to \lambda_0 \in (0,\infty)$ and $\beta_o \neq 0$, then $\widehat{\alpha}_n$ is inconsistent. Thus, for the bridge, SCAD, and adaptive Lasso penalty functions, $\lambda_n = o(1)$ is also a necessary condition for the consistency of $\widehat{\alpha}_n$.

We next present conditions needed to derive the convergence rate of $\widehat{\alpha}_n$.

Assumption 3.2. (i) The following functional central limit theorem (FCLT) holds

$$\sup_{\theta\in\Theta} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \{g_l(Z_i,\theta) - \mathbb{E}[g_l(Z_i,\theta)]\} \right] = O_p(1) \tag{3.8}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

for $l = q, k$; (ii) $G_l(\theta)$ is continuously differentiable in some neighborhood of θ_o for $l = q, k$; (iii) $\partial G_q(\theta_o)/\partial\theta'$ has full column rank; and (iv) the penalty function $\widehat{P}_{\lambda_n}(\cdot)$ satisfies $\widehat{P}_{\lambda_n}(0) = 0$ and is continuously twice differentiable at $\beta_{o,j}$ for any $j \in \mathcal{S}_\beta$ with

$$\max_{j\in\mathcal{S}_\beta} \left|\widehat{P}''_{\lambda_n}(\beta_{o,j})\right| = o_p(1). \tag{3.9}$$

Assumption 3.2.(i) is a high-level condition, which can be verified by applying Donsker theorems in specific models. Assumption 3.2.(ii) imposes a local differentiability condition on the expectation of the moment function $g_l(Z,\theta)$ $(l = q, k)$. Assumption 3.2.(iii) is a local identification condition for θ_o. If this assumption fails, the resulting estimator $\widehat{\alpha}_n$ may not be $\sqrt{n}$-consistent. Assumption 3.2.(iv) imposes some local smoothness conditions on the penalty function $\widehat{P}_{\lambda_n}(\cdot)$. Intuitively, this condition implies that attaching a penalty function to the GMM criterion function does not cause any local identification problem for the unknown parameters (θ_o, β_o). We show that the bridge penalty, adaptive Lasso penalty, and SCAD penalty functions satisfy Assumption 3.2.(iv) in Appendix B.

LEMMA 3.2. *Under Assumptions 3.1 and 3.2, we have*

$$\left\|\widehat{\alpha}_n - \alpha_o\right\| = O_p(\delta_n), \tag{3.10}$$

where $\delta_n = \max\{b_n, n^{-1/2}\}$ *and* $b_n = \max_{j\in\mathcal{S}_\beta} |\widehat{P}'_{\lambda_n}(\beta_{o,j})|$.

It is interesting to see that the convergence rate of $\widehat{\alpha}_n$ may depend on the convergence rate of the tuning parameter λ_n. Intuitively, the finite-sample bias of the GMM shrinkage estimator comes from two sources. The first is the stochastic error, which converges (in probability) to zero with the $n^{-1/2}$ rate. The second bias is due to the shrinkage effect of the penalty function on the estimators of $\beta_{o,+}$. The shrinkage bias converges (in probability) to zero with the rate b_n. As a result, the convergence rate of $\widehat{\alpha}_n$ is of the order δ_n.

Remark 3.2. If $\widehat{P}_{\lambda_n}(\cdot)$ is the bridge or adaptive Lasso penalty function, then $b_n = O_p(\lambda_n)$. The condition imposed on λ_n to show the consistency of $\widehat{\alpha}_n$, i.e., $\lambda_n = o(1)$, is insufficient to deduce that $\widehat{\alpha}_n$ is $\sqrt{n}$-consistent. For example, if $\lambda_n = o(1)$ and $\sqrt{n}\lambda_n \to \infty$, then from Lemma 3.2, we have

$$\left\|\widehat{\alpha}_n - \alpha_o\right\| = O_p(\lambda_n),$$

which implies that the convergence rate of $\widehat{\alpha}_n$ may be slower than $\sqrt{n}$. Hence, we need to impose stronger condition on λ_n to ensure that the GMM shrinkage estimator is $\sqrt{n}$-consistent. However, if $\widehat{P}_{\lambda_n}(\cdot)$ is the SCAD penalty function, then under the condition $\lambda_n = o(1)$, we can deduce that $b_n = \max_{j\in\mathcal{S}_\beta} |\widehat{P}'_{\lambda_n}(\beta_{o,j})| = 0$ when n is sufficiently large. So in this case, $\lambda_n = o(1)$ is a sufficient condition for the $\sqrt{n}$-consistency of $\widehat{\alpha}_n$.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

3.2. Consistent Moment Selection and Asymptotic Normality

In this subsection, we derive the consistent moment selection and the centered limiting distribution of $\widehat{\alpha}_{n,\mathcal{S}}$, where $\widehat{\alpha}'_{n,\mathcal{S}} = (\widehat{\theta}'_n, \widehat{\beta}'_{n,+})$ denotes the GMM shrinkage estimator of $\alpha'_{o,\mathcal{S}} = (\theta'_o, \beta'_{o,+})$. For ease of notation, we define

$$g_{q+k}(Z,\theta) \equiv \begin{bmatrix} g_q(Z,\theta) \\ g_k(Z,\theta) \end{bmatrix} \quad \text{and} \quad G_{q+k}(\theta) \equiv \begin{bmatrix} G_q(\theta) \\ G_k(\theta) \end{bmatrix}.$$

We first present and discuss the assumptions needed to show consistent moment selection.

Assumption 3.3. (i) The tuning parameter λ_n satisfies

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| = o_p(1); \tag{3.11}$$

(ii) for any $j \in \mathcal{S}_\beta^c$ and any random sequence $\{\beta_{j,n}\}_n$ with $\beta_{j,n} \neq 0$ almost everywhere (a.e.) for all n and $\beta_{j,n} = O_p(n^{-1/2})$, there is

$$\liminf_{n\to\infty} \left[\frac{1}{r_n \lambda_n} \left| \widehat{P}'_{\lambda_n}(\beta_{j,n}) \right| \right] > 0 \text{ a.e.}, \tag{3.12}$$

where r_n is some nonnegative sequence such that $n^{1/2}\lambda_n r_n \to \infty$.[7]

Assumption 3.3.(i) indicates that the convergence rate of $|\widehat{P}'_{\lambda_n}(\beta_{o,j})|$ for all $j \in \mathcal{S}_\beta$ is faster than $n^{-1/2}$. Under this assumption, Lemma 3.2 implies that

$$\sqrt{n}\left(\widehat{\alpha}_n - \alpha_o\right) = O_p(1), \tag{3.13}$$

i.e., the convergence rate of $\widehat{\alpha}_n$ is $n^{-1/2}$. Assumption 3.3.(ii) is a generalized version of condition (3.5) in Fan and Li (2001). Intuitively, Assumption 3.3.(ii) implies that the shrinkage estimator $\widehat{\beta}_{n,j}$ of $\beta_{o,j}$ ($j \in \mathcal{S}_\beta^c$) is the minimizer of $\widehat{P}_{\lambda_n}(\cdot)$ w.p.a.1. From Assumptions 3.1.(iv) and 3.2.(iv), we know that $\widehat{P}_{\lambda_n}(\cdot)$ is locally minimized at 0. Hence, Assumption 3.3.(ii) is the key condition needed for showing consistent moment selection. We show that the bridge penalty, adaptive Lasso penalty, and SCAD penalty functions satisfy Assumption 3.3 in Appendix B.

THEOREM 3.3. *Under Assumptions 3.1, 3.2, and 3.3, we have*

$$\lim_{n\to\infty} \Pr\left(\widehat{\beta}_{n,j} = 0\right) = 1 \quad \textit{for any} \quad j \in \mathcal{S}_\beta^c. \tag{3.14}$$

From the consistency of $\widehat{\beta}_n$ and Theorem 3.3, we immediately get

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1, \tag{3.15}$$

i.e., the consistent moment selection. We next provide the condition needed for deriving the limiting distribution of the GMM shrinkage estimator.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Assumption 3.4. (i) There exists a symmetric, nonrandom, and positive definite matrix W_o such that

$$W_n \to_p W_o; \tag{3.16}$$

(ii) the central limit theorem (CLT)

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \left\{ g_{q+k}(Z_i, \theta_o) - \mathbb{E}\left[g_{q+k}(Z_i, \theta_o) \right] \right\} \to_d \Psi(\theta_o), \tag{3.17}$$

holds, where $\Psi(\theta_o)$ is a $q+k$ dimensional nondegenerate Gaussian random vector.

Assumption 3.4.(i) is a regularity condition. Assumption 3.4.(ii) can be verified by applying CLTs in specific models with specific moment functions and data structure. Let $g_{d_{\beta_-}}(Z,\theta)$ and $g_{d_{\beta+}}(Z,\theta)$ denote the potentially valid and misspecified moment functions in set-2, respectively. We define

$$g_{q+d_{\beta_-}}(Z,\theta) \equiv \begin{bmatrix} g_q(Z,\theta) \\ g_{d_{\beta_-}}(Z,\theta) \end{bmatrix} \quad \text{and} \quad \frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} \equiv \begin{bmatrix} \frac{\partial E[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial \theta'} & 0 \\ \frac{\partial E[g_{d_{\beta+}}(Z,\theta_o)]}{\partial \theta'} & -I_{k-k_o} \end{bmatrix},$$

where I_{k-k_o} denotes a $(k-k_o)\times(k-k_o)$ identity matrix. If we define

$$M_{\mathcal{S}} = \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} \right]' W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} \right],$$

then under Assumptions 3.1.(iii) and 3.2.(iii) we know that $M_{\mathcal{S}}$ is a nonsingular matrix.

THEOREM 3.4 ($\sqrt{n}$-Normality). *Under Assumptions 3.1, 3.2, 3.3, and 3.4, we have*

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) \to_d N\left(0, M_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S}} M_{\mathcal{S}}^{-1}\right), \tag{3.18}$$

where

$$\Sigma_{\mathcal{S}} = \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} \right]' W_o E\left[\Psi(\theta_o)\Psi'(\theta_o) \right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} \right],$$

and $\Psi(\theta_o)$ is defined in (3.17).

As illustrated in the above theorem, under consistent moment selection the asymptotic distribution of the GMM shrinkage estimator $\widehat{\theta}_n$ is normal with mean zero. On the other hand, if the tuning parameter λ_n goes to zero faster than the requirement in Assumption 3.3.(ii) such that $n^{1/2}\lambda_n r_n$ converges to some finite positive number, the moment selection result may be conservative, i.e., the invalid moment conditions will not be selected w.p.a.1 but the valid moment conditions will be selected with probability strictly less than 1. In that scenario, the asymptotic distribution of $\widehat{\theta}_n$ will be mixture normal with zero mean.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

3.3. Oracle Properties

The oracle properties state that the GMM shrinkage estimation can consistently identify all potentially valid moment conditions in set-2, and its estimator of θ_o is as efficient as the oracle GMM estimator based on all valid moment conditions. As the consistent moment selection is directly implied by Lemma 3.1 and Theorem 3.3, the oracle properties follow if we can show that the asymptotic variance-covariance matrix of $\widehat{\theta}_n$ coincides with that of the oracle GMM estimator.

If we had prior information about the validity of the moment conditions in set-2, then there would be $q+k_o$ valid moment conditions to estimate θ_o, where k_o is the number of valid moment conditions in set-2. We can stack all valid moment conditions as

$$m_e(\theta_o) \equiv \mathbb{E}\left[g_{q+d_{\beta_-}}(Z,\theta_o)\right] = 0. \tag{3.19}$$

From the moment conditions in (3.19), we can compute the asymptotic variance-covariance matrix of the optimally weighted oracle GMM estimator as

$$\Sigma^* = \left\{\left[\frac{\partial m_e(\theta)}{\partial \theta_o'}\right]' V_{e,o}^{-1}\left[\frac{\partial m_e(\theta)}{\partial \theta_o'}\right]\right\}^{-1}, \tag{3.20}$$

where $V_{e,o}$ is the leading $(q+k_o)\times(q+k_o)$ submatrix of $\mathbb{E}\left[\Psi(\theta_o)\Psi'(\theta_o)\right]$.[8]

In the GMM shrinkage estimation, if we choose a weight matrix W_n^* such that

$$W_n^* \to_p W_o = \left\{\mathbb{E}\left[\Psi(\theta_o)\Psi'(\theta_o)\right]\right\}^{-1}, \tag{3.21}$$

then an interesting question is whether the resulting GMM shrinkage estimator $\widehat{\theta}_n$ of θ_o could be as efficient as the optimally weighted oracle GMM estimator. The answer to the above question is affirmative, as illustrated in the following theorem.

THEOREM 3.5 (Oracle properties). *Under Assumptions 3.1, 3.2, and 3.3, we have*

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1. \tag{3.22}$$

Furthermore, if the weight matrix W_n satisfies (3.21) and Assumption 3.4 holds, then we have

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_o\right) \to_d N(0, \Sigma^*), \tag{3.23}$$

where Σ^* is defined in (3.20).

Remark 3.6. In finite samples, the naive confidence intervals (CIs) for θ_o are constructed using the asymptotic distribution in (3.23) and plug-in estimator of the variance-covariance matrix Σ^* based on the GMM shrinkage estimator $\widehat{\theta}_n$ and

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

the selected moment conditions. It should be noted that the results in (3.22) and (3.23) are pointwise asymptotic properties in the sense that they are derived under the assumption that the unknown parameters α_o and the latent data-generating mechanism are fixed and unchanged with respect to the sample size. Hence, in finite samples, the naive CIs fail to take the moment selection errors into account, though the moment selection errors are integrated into the plug-in estimator of Σ^*. That is to say, the naive CIs may be mistakenly centered if some misspecified moment conditions are selected in finite samples. One should note that the naive post-moment selection inference based on other moment selection procedures also have this problem. Ignoring the finite-sample moment selection errors may lead to poor coverage probabilities of the naive CIs and size distortion of the resulting hypothesis tests, which represents a well-known challenge in the model/moment selection literature as recently overviewed in Leeb and Pötscher (2005). The treatment of the post-moment selection inference is beyond the scope of this paper, and we will address this issue in a different paper.

3.4. Selection of the Tuning Parameter

From the results of previous subsections, we see that the tuning parameter λ_n plays an important role in deriving the asymptotic properties of the GMM shrinkage estimator. Assumptions 3.1.(iv), 3.2.(iv), and 3.3 are sufficient conditions imposed on λ_n for the oracle properties of the GMM shrinkage estimator. However, these conditions do not suggest a straightforward way of selecting the tuning parameter λ_n in finite samples. In this subsection, we provide explicit data-dependent tuning parameters for the GMM shrinkage estimation. The proposed tuning parameters not only satisfy Assumptions 3.1.(iv), 3.2.(iv), and 3.3, but they also take the finite sample performance of the GMM shrinkage estimator into account. We use the adaptive Lasso penalty as the illustrating example, as this penalty function is used in the simulation study and empirical application of this paper. The same idea applies to the other penalty functions.

From the Karush-Kuhn-Tucker (KKT) optimality condition, we know that the nonzero $\beta_{o,j}$ will be estimated as zero only if the following inequality holds,

$$\left| W_n(q+j) \left[n^{-1/2} \sum_{i=1}^{n} \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n) \right] \right| < \frac{\sqrt{n}\lambda_n}{2|\widehat{\beta}_{1st,j}|^{\omega}}, \tag{3.24}$$

where $W_n(q+j)$ denotes the $(q+j)$ th row of the weight matrix W_n. If nonzero $\beta_{o,j}$ is estimated as zero, then the left-hand side of the above inequality will be asymptotically bounded away from zero because an invalid moment condition is used in estimation. To ensure the shrinkage bias and moment selection errors are small in finite samples, one would like to have $n^{1/2}\lambda_n/2$ converge to zero as fast as possible. Next note that the term in the right-hand side of (3.24) can be rewritten as $n^{(1+\omega)/2}\lambda_n|\sqrt{n}\widehat{\beta}_{1st,j}|^{-\omega}/2$. Under the root-n consistent estimation, the left-hand side of (3.24) is asymptotically a linear combination of Gaussian random

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

variables. Hence, one would like to have $n^{(1+\omega)/2}\lambda_n/2$ diverge to infinity as fast as possible to ensure that valid moment conditions are selected with high probability in finite samples. We choose $\lambda_n = 2cn^{-1/2-\omega/4}$ to balance the requirements that $n^{1/2}\lambda_n/2$ converge to zero and $n^{(1+\omega)/2}\lambda_n/2$ diverge to infinity as fast as possible. We next discuss the selection of the loading term c in the above λ_n.

As $\lambda_n = 2cn^{-1/2-\omega/4}$ satisfies Assumptions 3.1.(iv), 3.2.(iv), and 3.3, we can use the arguments in the proof of Theorem 3.4 and Theorem 3.5 to deduce that

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n) &= n^{-1/2}\sum_{i=1}^{n}\left\{\rho(Z_i,\widehat{\alpha}_n)-\rho(Z_i,\alpha_o)-m(\widehat{\alpha}_n)\right\} \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\rho(Z_i,\alpha_o)+n^{1/2}m(\widehat{\alpha}_n) \\ &= \left\{I_{q+k}-\left[\frac{\partial m(\theta_o)}{\partial\alpha_{\mathcal{S}}'}\right]' M_{\mathcal{S}}^{-1}\left[\frac{\partial m(\theta_o)}{\partial\alpha_{\mathcal{S}}'}\right]W_o\right\} W_o^{-1/2}B_{q+k}(1) \\ &\quad + o_p(1), \end{aligned}$$

where W_o satisfies (3.21), $W_o^{-1/2}$ is the inverse of the square root $W_o^{1/2}$ of W_o, and $B_{q+k}(1)$ denotes a $q+k$ dimensional standard Gaussian random vector. The consistency of W_n and the above asymptotic approximation imply that

$$W_n(q+j)\left[n^{-1/2}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n)\right] = W_o^{1/2}(q+j)\Pi_{q+k}B_{q+k}(1)+o_p(1), \quad \textbf{(3.25)}$$

where $W_o^{1/2}(q+j)$ denotes the $(q+j)$th row of the matrix $W_o^{1/2}$ and

$$\Pi_{q+k} = I_{q+k} - W_o^{1/2}\left[\frac{\partial m(\theta_o)}{\partial\alpha_{\mathcal{S}}'}\right]' M_{\mathcal{S}}^{-1}\left[\frac{\partial m(\theta_o)}{\partial\alpha_{\mathcal{S}}'}\right]W_o^{1/2} \quad \textbf{(3.26)}$$

is an idempotent matrix. We propose to choose the loading term $\widehat{c}_{j,n} = ||W_n^{1/2}(q+j)\widehat{\Pi}_{q+k}||$ to normalize the linear combination of the projected Gaussian random vector in (3.25), where $\widehat{\Pi}_{q+k}$ can be estimated by a first step GMM shrinkage estimation with $\lambda_n = 2\sqrt{\log(k)}n^{-1/2-\omega/4}$, for example. To sum up the above discussion, we propose to select the tuning parameter

$$\widehat{\lambda}_{n,j} = 2||W_n^{1/2}(q+j)\widehat{\Pi}_{q+k}||n^{-1/2-\omega/4} \quad \textbf{(3.27)}$$

for the jth moment selection coefficient $\beta_{o,j}$ $(j=1,\ldots,k)$.

4. GMM SHRINKAGE ESTIMATION UNDER WEAK IDENTIFICATION

In this section we study the properties of the GMM shrinkage estimator when the moment conditions in set-1 fail to strongly identify the unknown parameters θ_o. Specifically, we assume that we have the following moment conditions in set-1,

$$\mathbb{E}\left[g_{q,n}(Z,\theta)\right] = n^{-\tau}G_{q,n}(\theta), \quad \textbf{(4.1)}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

where $g_{q,n}(\cdot,\cdot): R^{d_z} \times R^{d_\theta} \to R^q$, $G_{q,n}(\theta_o) = 0$, and $\tau \in [0, \frac{1}{2})$.[9] At the same time, we have another set of possibly misspecified moment conditions,

$$\mathbb{E}\left[g_{k,n}(Z,\theta_o)\right] = G_k(\theta_o) \stackrel{?}{=} 0, \tag{4.2}$$

where $g_{k,n}(\cdot,\cdot): R^{d_z} \times R^{d_\theta} \to R^k$. Hahn and Kuersteiner (2002) study a linear IV model where the moment conditions constructed from the IVs are similar to these in (4.1). They show that if $\tau \in [0, \frac{1}{2})$, then the IV estimators have the convergence rate $n^{\tau-1/2}$. Caner (2009) shows that the GMM bridge estimator based on the moment conditions in (4.1) has the same rate of convergence.

In this section we show that if there are strong and correctly specified moment conditions in (4.2), then these moment conditions can be consistently selected by the shrinkage method. More importantly, we show that the GMM shrinkage estimator $\widehat{\theta}_n$ has a faster rate of convergence and hence better stochastic properties than the estimators only using the moment conditions in (4.1). The intuition is that when the valid and strong moment conditions in (4.2) are consistently selected, the information contained in these moment conditions is automatically included into estimating θ_o w.p.a.1.

Assumption 4.1. (i) Assume $G_k(\theta)$ is continuous in θ and there exists some continuous function $G_q(\theta)$ such that

$$\sup_{\theta\in\Theta} \left\| G_{q,n}(\theta) - G_q(\theta) \right\| = o(1); \tag{4.3}$$

(ii) for any $\varepsilon > 0$, there exists some $\delta_\varepsilon > 0$ such that

$$\inf_{\{\theta\in\Theta:\|\theta-\theta_o\|\geq\varepsilon\}} \left\| G_q(\theta) \right\| > \delta_\varepsilon; \tag{4.4}$$

(iii) for $l = q, k$, the following FCLTs hold,

$$\sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} \left\{ g_{l,n}(Z_i,\theta) - \mathbb{E}\left[g_{l,n}(Z_i,\theta)\right] \right\} \right| = O_p\left(n^{-1/2}\right); \tag{4.5}$$

(iv) let N_{θ_o} denote some local neighborhood of θ_o, then $G_{q,n}(\theta)$, $G_q(\theta)$, and $G_k(\theta)$ are continuously differentiable for all $\theta \in N_{\theta_o}$ and

$$\sup_{\theta\in N_{\theta_o}} \left\| \frac{\partial G_{q,n}(\theta)}{\partial \theta'} - \frac{\partial G_q(\theta)}{\partial \theta'} \right\| = o(1), \tag{4.6}$$

where $\partial G_q(\theta_o) / \partial \theta'$ has full column rank;
(v) the penalty function $\widehat{P}_{\lambda_n}(\cdot)$ is nonnegative and $\widehat{P}_{\lambda_n}(0) = 0$; moreover, $\widehat{P}_{\lambda_n}(\cdot)$ satisfies $n^{2\tau} \max_{j\in\mathcal{S}_\beta} \widehat{P}_{\lambda_n}(\beta_{o,j}) = o_p(1)$ and is continuously twice differentiable at $\beta_{o,j}$ for any $j \in \mathcal{S}_\beta$ with

$$\max_{j\in\mathcal{S}_\beta} \sup_{\beta_j\in N_{\beta_{o,j}}} \left| n^{2\tau} \widehat{P}''_{\lambda_n}(\beta_j) \right| = o_p(1), \tag{4.7}$$

where $N_{\beta_{o,j}}$ denotes some local neighborhood of $\beta_{o,j}$ for $j \in \mathcal{S}_\beta$.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Assumption 4.1.(i) imposes a continuity condition on $G_k(\theta)$ and assumes the existence of uniform limit function $G_q(\theta)$ for $G_{q,n}(\theta)$. The uniform approximation in (4.3) is a regularity condition in the weak moment literature (e.g., Stock and Wright, 2000). Assumption 4.1.(ii) is an identifiable uniqueness condition of θ_o. Assumptions 4.1.(iii) and (iv) are the generalized versions of Assumptions 3.1.(ii) and 3.2.(i)–(iii). Compared with Assumptions 3.1.(iv) and 3.2.(iv), Assumption 4.1.(v) imposes stronger restrictions on the tuning parameter λ_n. When the moment conditions in (4.1) are nearly weak, their information about θ_o is vanishing at the rate $n^{-\tau}$, and the information contained in GMM criterion function is vanishing at the rate $n^{-2\tau}$. Hence, $\widehat{P}_{\lambda_n}(\beta_{o,j})$ $(j \in \mathcal{S}_\beta)$ must converge to zero faster than $n^{-2\tau}$ to ensure the consistency of the GMM shrinkage estimator. Condition (4.7) ensures that the penalty function does not affect the local identification of θ_o, when the moment conditions in set-1 are nearly weak. It is clear that this condition is stronger than condition (3.9) in Assumption 3.2.(iv).

LEMMA 4.1 (Rate of convergence). *Under Assumptions 3.1.(iii) and 4.1, we have*

$$\left\|\widehat{\alpha}_n - \alpha_o\right\| = O_p\left(\max\left\{n^{2\tau} b_n, n^{\tau-1/2}\right\}\right),$$

where $b_n = \max_{j\in\mathcal{S}_\beta} |\widehat{P}'_{\lambda_n}(\beta_{o,j})|$.

It is clear that if $n^{1/2+\tau} b_n = O_p(1)$, then $||\widehat{\theta}_n - \theta_o|| = O_p(n^{\tau-1/2})$, which implies that the convergence rate of the GMM shrinkage estimator $\widehat{\theta}_n$ is at least as fast as the optimal convergence rate of the GMM estimator of θ_o based on the moment conditions in (4.1). However, we next show that if the tuning parameter λ_n converges to zero not very fast, then the valid moment conditions in (4.2) are consistently selected in the GMM shrinkage estimation and, moreover, the GMM shrinkage estimator $\widehat{\theta}_n$ has the rate of convergence faster than $n^{\tau-1/2}$.

Assumption 4.2. (i) The penalty function $\widehat{P}_{\lambda_n}(\cdot)$ satisfies

$$n^{1/2+\tau} \max_{j\in\mathcal{S}_\beta} \left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = o_p(1); \tag{4.8}$$

(ii) for any $j \in \mathcal{S}_\beta^c$ and any random sequence $\{\beta_{j,n}\}_n$ with $\beta_{j,n} \neq 0$ a.e. for all n and $\beta_{j,n} = O_p(n^{-1/2+\tau})$, there is

$$\liminf_{n\to\infty} \left[\frac{1}{r_n \lambda_n} \left|\widehat{P}'_{\lambda_n}(\beta_{j,n})\right|\right] > 0 \text{ a.e.}, \tag{4.9}$$

where r_n is some nonnegative sequence such that $n^{1/2-\tau}\lambda_n r_n \to \infty$.

It is clear that when $\tau = 0$, Assumption 4.2 becomes Assumption 3.3. Intuitively, the sequence r_n represents the magnitude of adaptive shrinkage in the penalty function. When the moment conditions in set-1 are almost uninformative,

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

i.e., τ is close to $1/2$, the condition that Assumption 4.2 imposes on r_n becomes very restrictive. This fact is important for the choice of penalty functions in practice. For example, in Appendix B we show that when $\tau \geq \frac{1}{3}$ (or $\tau > \frac{1}{4}$), the bridge (or SCAD) penalty fails to satisfy Assumption 4.2. The condition for the adaptive Lasso penalty satisfying Assumption 4.2 is $\omega > \frac{4\tau}{1-2\tau}$.[10]

LEMMA 4.2. *Under the conditions of Lemma 4.1 and Assumption 4.2, there is*

$$\lim_{n\to\infty} \Pr\left(\widehat{\beta}_{n,j} = 0\right) = 1 \text{ for all } j \in \mathcal{S}_\beta^c. \tag{4.10}$$

Lemma 4.2 implies that $\beta_{o,j}$ $(j \in \mathcal{S}_\beta^c)$ is estimated as zero w.p.a.1. This result, combined with the following local identification assumption, enables us to improve the convergence rate of $\widehat{\theta}_n$. Let $G_{d_{\beta_-}}(\theta)$ and $G_{d_{\beta_+}}(\theta)$ denote the expectation of potentially valid and misspecified moment functions in (4.2), respectively.

Assumption 4.3. The matrix $\partial G_{d_{\beta_-}}(\theta_o) \Big/ \partial \theta'$ has full column rank.

Assumption 4.3 implies that the unknown parameters θ_o are locally identified by the potentially valid moment conditions in set-2, which is an important condition for showing the $\sqrt{n}$-consistency of $\widehat{\theta}_n$. If this condition does not hold, then the moment conditions in set-1 are needed to achieve the local identification of θ_o. In that case, the convergence rate of $\widehat{\theta}_n$ is not $n^{-1/2}$, but is still faster than $n^{\tau-1/2}$.

LEMMA 4.3. *Under the conditions of Lemma 4.2 and Assumption 4.3, we have*

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) = O_p(1). \tag{4.11}$$

Let $W_{o,kk}$ denote the last $k \times k$ submatrix of W_o,

$$M_+ = \left[\frac{\partial G_k(\theta_o)}{\partial \alpha'_{\mathcal{S}}}\right]' W_{o,kk} \left[\frac{\partial G_k(\theta_o)}{\partial \alpha'_{\mathcal{S}}}\right] \quad \text{and} \quad \frac{\partial G_k(\theta_o)}{\partial \alpha'_{\mathcal{S}}} = \begin{pmatrix} \frac{\partial G_{d_{\beta_-}}(\theta_o)}{\partial \theta'} & 0 \\ \frac{\partial G_{d_{\beta_+}}(\theta_o)}{\partial \theta'} & -I_{k-k_o} \end{pmatrix}.$$

From Assumptions 3.1.(iii) and 4.3, we know that the matrix M_+ is invertible.

COROLLARY 4.1 ($\sqrt{n}$-Normality). *Under the conditions of Lemma 4.3 and Assumption 3.4, we have*

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) \to_d N\left(0, M_+^{-1}\Sigma_+ M_+^{-1}\right), \tag{4.12}$$

where

$$\Sigma_+ = \left[\mathbf{0}, \left(\frac{\partial G_k(\theta_o)}{\partial \alpha'_{\mathcal{S}}}\right)'\right] W_o \mathbb{E}\left[\Psi(\theta_o)\Psi'(\theta_o)\right] W_o \left[\mathbf{0}, \left(\frac{\partial G_k(\theta_o)}{\partial \alpha'_{\mathcal{S}}}\right)'\right]',$$

and $\Psi(\theta_o)$ is defined in (3.17).

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Remark 4.2. If the probability limit of weight matrix W_n satisfies $W_o^{-1} = \mathbb{E}\left[\Psi(\theta_o)\Psi'(\theta_o)\right]$, then it is clear that $\Sigma_+ = M_+$, which together with Corollary 4.1 implies that

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) \to_d N\left(0, M_+^{-1}\right).$$

The marginal distribution of $\widehat{\theta}_n$ can be derived using similar arguments in showing (3.23).

5. SIMULATION STUDY

In this simulation study the data are generated from the linear IV model

$$Y_i = X_i\theta_{1,o} + W_{1,i}\theta_{2,o} + W_{2,i}\theta_{3,o} + u_i, \tag{5.1}$$

$$X_i = Z_{1,i}\pi_{1,o} + W_i'\pi_{2,o} + Z_{2,i}'\pi_{3,o} + \upsilon_i, \tag{5.2}$$

where Y_i is a scaler dependent variable, X_i is a scaler endogenous variable, $Z_{1,i}$ is the IV whose validity is assumed to be known, $W_i = \left(W_{1,i}, W_{2,i}\right)$ contains two exogenous variables, $Z_{2,i}' = \left(Z_{21,i}, \ldots, Z_{25,i}\right)$ is a set of potentially valid IVs, and u_i and υ_i are error terms that are correlated with each other.

Suppose an econometrician specifies the model

$$Y_i = X_i\theta_o + W_{1,i}\theta_{2,o} + W_{2,i}\theta_{3,o} + u_i$$

with the set-1 moment conditions

$$\mathbb{E}\left[u_i(Z_{1,i}, W_{1,i}, W_{2,i})'\right] = 0 \tag{5.3}$$

to identify and consistently estimate $\theta_o = (\theta_{1,o}, \theta_{2,o}, \theta_{3,o})$. The potentially valid IVs in $Z_{2,i}$ are mixed with 20 invalid IVs $F_i' = (F_{1,i}, \ldots, F_{20,i})$ to construct the set-2 moment conditions

$$\mathbb{E}\left[u_i(Z_{2,i}', F_i')'\right] \stackrel{?}{=} 0. \tag{5.4}$$

To generate the simulated data, we first generate $(W_i', Z_{1,i}, Z_{2,i}', u_i, \upsilon_i, F_i^*)$ from a multivariate normal distribution with mean 0 and variance-covariance matrix Σ, where $F_i^* = (F_{1,i}^*, \ldots, F_{20,i}^*)$, $\Sigma = \text{diag}(\Sigma_Z, \Sigma_{u,\upsilon}, I_{20})$, Σ_Z is an 8×8 matrix with the i, jth element being $0.2^{|i-j|}$, $\Sigma_{u,\upsilon}$ is a 2×2 matrix with diagonal elements 1 and off-diagonal elements 0.6, and I_{20} is a 20×20 identity matrix. Let c_l be some value between 0 and 0.8 and l be a 1×20 vector with the jth $(j = 1, \ldots, 20)$ element being $c_l + (0.8 - c_l) * (j-1)/19$. The invalid IVs are generated in the following way,

$$F_i = F_i^* + u_i \times l.$$

It is clear that when c_l is close to zero, the IVs in F_i with smaller index numbers (e.g., $F_{1,i}$ and $F_{2,i}$) behave more like valid IVs in finite samples, and it becomes difficult to distinguish these IVs from the potentially valid ones. We choose

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

different values for c_l ($c_l = 0.2$ or 0.5) to see how our method works in different scenarios. The parameters in the model (5.1)–(5.2) take the values

$$\theta_o = (0.5, 0.4, 0.3), \qquad \pi_{2,o} = (0.1, 0.1), \quad \text{and}$$
$$\pi_{3,o} = (0.15, 0.15, 0.2, 0.2, 0.25).$$

When $\pi_{1,o}$ is close to zero, $Z_{1,i}$ may contain weak information about the unknown parameter $\theta_{1,o}$, which may also affect the performance of our method in moment selections. In the simulation studies, we choose different values for $\pi_{1,o}$ ($\pi_{1,o} = 0.2$ or 0.8) to see how our method is affected by the signal strength of $Z_{1,i}$.

For each specification of $\left(c_l, \pi_{1,o}\right)$, we use the simulated samples with sample sizes $n = 250$, 500, and 1,000, respectively. For each sample size, 10,000 simulated samples are drawn from the data generating mechanism. The adaptive Lasso penalty with $\omega = 2$ is used to construct the criterion of GMM shrinkage estimation. We use the projected scaled subgradient method (active-set variant) proposed in Schmidt (2010) to solve the minimization problem in the GMM shrinkage estimation. It is remarkable that in this simulation study, there are 25 moment conditions in set-2 and hence 2^{25} subsets of moment conditions to be investigated by the traditional moment selection procedures.

With each simulated sample, we calculate four different GMM estimators, which include the oracle GMM estimator, GMM estimator based on set-1 moment conditions, GMM shrinkage estimator and GMM estimator based on set-1 moment conditions and the moment conditions selected by GMM shrinkage estimation. The oracle GMM estimator is based on the moment conditions in set-1 and all valid moment conditions in set-2. We call the second and the last estimators as feasible GMM estimator and post-shrinkage GMM (PS-GMM) estimator, respectively. Given the specification of $\left(c_l, \pi_{1,o}\right)$ and the sample size n, we get 10,000 values for each of the above four estimators using the 10,000 simulated samples. We use these 10,000 values of each estimator to calculate its finite sample bias, standard error, and root of mean square errors, and estimate its finite sample density.

For ease of describing the simulation results, we divide all possible subsets of moment conditions in set-2 into three categories. The first category includes subsets of moment conditions that contain at least one invalid moment condition. The second category has only one subset that contains and only contains all potentially valid moment conditions in set-2. The third category includes the subsets of moment conditions that do not have any invalid moment conditions but fail to include all valid ones in set-2. The GMM estimators based on the subsets of moment conditions from the first category are inconsistent. On the other hand, the GMM estimators based on the subsets of moment conditions from the second or third categories (and the moment conditions in set-1) are consistent, while the subset of moment conditions in the second category (together with the moment conditions in set-1) gives the most efficient estimator.

Table 5.1 presents the finite sample (estimated) probabilities of our method in selecting subsets from the first category (the first number in each bracket), the

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 5.1. Performance of GMM shrinkage method in moment selection

	$n = 250$	$n = 500$	$n = 1{,}000$
		$\pi_{1,o} = 0.2$	
$c_l = 0.2$	(.5429 .3699 .0872)	(.0857 .8182 .0961)	(.0015 .9503 .0482)
$c_l = 0.5$	(.0022 .8145 .1833)	(.0000 .9012 .0988)	(.0000 .9541 .0459)
		$\pi_{1,o} = 0.8$	
$c_l = 0.2$	(.4973 .4709 .0318)	(.0743 .9012 .0245)	(.0010 .9862 .0128)
$c_l = 0.5$	(.0000 .9312 .0688)	(.0000 .9726 .0274)	(.0000 .9872 .0128)

Notes: The three numbers in parentheses are the estimated finite sample probabilities of selecting subsets of moment conditions in set-2 from three different categories, respectively. The first category includes the subsets of moment conditions that contain at least one invalid moment condition. The second category contains only the subset of all potentially valid moment conditions in set-2. The third category includes the subsets that do not have the invalid moment conditions, but fail to contain all valid moment conditions in set-2. The finite sample probabilities are computed based on 10,000 replications.

second category (the second number in each bracket), and the third category (the third number in each bracket), respectively. Within each row, we see that the probability of selecting the subset of moment conditions from the second category increases with the sample size growing. When the sample size is 1,000, this probability is close to 1, which is predicted by the consistent moment selection result established in Section 3. Both the signal strength of the IVs in set-1 and the severity of misspecification of IVs in set-2 affect the finite sample performance of our method in moment selection. When the misspecified moment conditions in set-2 are close to being valid (i.e., $c_l = 0.2$) and the sample size is small (i.e., $n = 250$), the probabilities of selecting the subset of moment conditions from the second category become small. This problem is exacerbated if the IVs in set-1 only contain weak information about $\theta_{1,o}$. On the other hand, when the invalid moment conditions in set-2 are severely misspecified, i.e., $c_l = 0.5$, the probabilities of selecting invalid moment conditions are small. Finally, when the IVs in set-1 are strong (i.e., $\pi_{1,o} = 0.8$) and the invalid IVs in set-2 are severely misspecified (i.e., $c_l = 0.5$), the GMM shrinkage method performs very well even when the sample size is small (i.e., $n = 250$).

Tables 5.2–5.5 describe the finite sample properties of the above four estimators in terms of finite sample bias, standard deviation, and root of mean square errors. We summarize the main results in Tables 5.2–5.5 by comparing the GMM shrinkage estimator with the oracle GMM estimator, the feasible GMM estimator, and the PS-GMM estimator, respectively. In all scenarios, the finite sample bias, standard error, and mean square errors of the GMM shrinkage estimator approximate those of the oracle GMM estimator with the sample size growing. The finite sample properties of these two estimators are almost identical when the sample size is large, i.e., $n = 1{,}000$. Compared with the feasible GMM estimator, the GMM shrinkage estimator has much smaller standard error, particularly when the IV in set-1 only contains weak information about $\theta_{1,o}$; i.e., $\pi_{1,o} = 0.2$. The

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 5.2. Finite sample properties of the GMM and GMM shrinkage estimators with $(c_l, \pi_{1,o}) = (0.2, 0.2)$

	$n = 250$			$n = 500$			$n = 1,000$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
				Oracle estimator					
θ_{1o}	.0316	.1125	.1169	.0158	.0804	.0819	.0066	.0569	.0573
θ_{2o}	−.0035	.0671	.0672	−.0016	.0460	.0460	−.0008	.0329	.0329
θ_{3o}	−.0029	.0673	.0674	−.0024	.0478	.0478	−.0016	.0333	.0334
				GMM shrinkage estimator					
θ_{1o}	.1184	.2467	.2737	.0293	.0931	.0976	.0067	.0588	0592
θ_{2o}	−.0124	.0714	.0725	−.0029	.0462	.0463	−.0008	.0329	.0329
θ_{3o}	−.0160	.0754	.0771	−.0043	.0482	.0484	−.0016	.0334	.0334
				Post-shrinkage GMM estimator					
θ_{1o}	.0539	.2640	.2694	.0155	.0918	.0931	.0057	.0594	.0597
θ_{2o}	−.0061	.0727	.0730	−.0015	.0463	.0463	.0007	.0329	.0329
θ_{3o}	−.0061	.0781	.0783	.0023	.0483	.0483	−.0015	.0335	.0335
				Feasible GMM estimator					
θ_{1o}	−.0633	.4173	.4221	−.0228	.2178	.2190	−.0133	.1462	.1468
θ_{2o}	.0053	.0841	.0843	.0022	.0521	.0521	.0012	.0358	.0358
θ_{3o}	.0110	.0948	.0954	.0034	.0582	.0583	.0012	.0390	.0390

Note: The finite sample bias (BS), standard deviation (SD), and root of mean square error (RE) are computed using the corresponding estimates from 10,000 replications.

finite sample bias of the GMM shrinkage estimator is larger than that of the GMM estimator in some scenarios (e.g., when $c_l = 0.2$). For the GMM shrinkage estimator, one important source of its finite sample bias is the shrinkage bias. To get rid of this shrinkage bias, one can perform GMM estimation using the set-1 moment conditions and the moment conditions in set-2 selected by the GMM shrinkage estimation to get the PS-GMM estimator. Although generally speaking, the PS-GMM estimator enjoys smaller finite sample bias, its finite sample standard error may be larger than the GMM shrinkage estimator. This phenomenon is recovered in Tables 5.2 and 5.3, where the potentially valid moment conditions contain strong information about the endogenous variable. The underlying reason for this phenomenon is that when the moment selection coefficients (i.e., $\beta_{o,+}$) of invalid moment conditions are shrunk toward zero, part of the information contained in these moment conditions is used in the GMM shrinkage estimation, which helps to reduce the variance but at the same time introduces a new bias to the GMM shrinkage estimator.

Figure 1 depicts the finite sample (n = 250) densities of the oracle GMM estimator, the feasible GMM estimator, and the GMM shrinkage estimator of $\theta_{1,o}$. It is clear that when the signal strength of the IVs in set-1 and the severity of misspecification of the IVs in set-2 are increased, the finite sample density of

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 5.3. Finite sample properties of the GMM and GMM shrinkage estimators with $(c_l, \pi_{1,o}) = (0.5, 0.2)$

	$n = 250$			$n = 500$			$n = 1{,}000$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
				Oracle estimator					
θ_{1o}	.0316	.1125	.1169	.0158	.0804	.0819	.0066	.0569	.0573
θ_{2o}	−.0035	.0671	.0672	−.0016	.0460	.0460	−.0008	.0329	.0329
θ_{3o}	−.0029	.0673	.0674	−.0024	.0478	.0478	−.0016	.0333	.0334
				GMM shrinkage estimator					
θ_{1o}	.0247	.2290	.2304	.0134	.0859	.0869	.0058	.0588	.0591
θ_{2o}	−.0030	.0708	.0708	−.0014	.0462	.0462	−.0007	.0329	.0329
θ_{3o}	−.0020	.0731	.0731	−.0020	.0480	.0481	−.0015	.0334	.0335
				Post-shrinkage GMM estimator					
θ_{1o}	.0151	.2514	.2518	.0117	.0888	.0896	.0055	.0594	.0596
θ_{2o}	−.0020	.0719	.0719	−.0012	.0463	.0463	−.0007	.0329	.0329
θ_{3o}	−.0005	.0750	.0750	−.0017	.0482	.0482	−.0014	.0335	.0335
				Feasible GMM estimator					
θ_{1o}	−.0633	.4173	.4221	−.0228	.2178	.2190	−.0133	.1462	.1468
θ_{2o}	.0053	.0841	.0843	.0022	.0521	.0521	.0012	.0358	.0358
θ_{3o}	.0110	.0948	.0954	.0034	.0582	.0583	.0012	.0390	.0390

Note: The finite sample bias (BS), standard deviation (SD), and root of mean square error (RE) are computed using the corresponding estimates from 10,000 replications.

the GMM shrinkage estimator is approaching that of the oracle GMM estimator. Moreover, bimodality shows up in the finite sample density of the GMM shrinkage estimator when $\pi_{1,o} = 0.8$ and $c_l = 0.2$. Bimodality appears because in finite samples, our moment selection method has nonzero probability of selecting the misspecified moment conditions, and in this scenario, the GMM shrinkage estimator has large bias due to the inconsistency. The bimodality in the finite sample density of the GMM shrinkage estimator causes a challenge for conducting stochastic inferences based on the selected moment conditions. The pointwise asymptotic distribution of the GMM shrinkage estimator does not capture the errors of moment selection; hence, in finite samples, the CIs constructed using this limiting distribution may have low coverage probability, and the resulting hypothesis tests may have incorrect size.

6. AN EMPIRICAL EXAMPLE

In this section we apply the GMM shrinkage method to the life-cycle labor supply model studied in MaCurdy (1981) and Altonji (1986). Both papers estimate the labor supply equation

$$\Delta \log(h_{i,t}) = \alpha_{o,t} + \Delta \log(w_{i,t})\delta_o + \varepsilon_{i,t}, \tag{6.1}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 5.4. Finite sample properties of the GMM and GMM shrinkage estimators with $(c_l, \pi_{1,o}) = (0.2, 0.8)$

	$n = 250$			$n = 500$			$n = 1{,}000$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
				Oracle estimator					
θ_{1o}	.0110	.0661	.0670	.0058	.0466	.0470	.0020	.0332	.0333
θ_{2o}	−.0016	.0672	.0672	−.0007	.0457	.0457	−.0004	.0325	.0326
θ_{3o}	−.0012	.0680	.0680	−.0015	.0482	.0482	−.0012	.0334	.0334
				GMM shrinkage estimator					
θ_{1o}	.0378	.0721	.0814	.0094	.0474	.0483	.0020	.0333	.0333
θ_{2o}	−.0043	.0684	.0685	−.0010	.0458	.0458	−.0004	.0326	.0326
θ_{3o}	−.0087	.0694	.0699	−.0024	.0483	.0484	−.0012	.0334	.0334
				Post-shrinkage GMM estimator					
θ_{1o}	.0173	.0692	.0713	.0060	.0470	.0474	.0019	.0333	.0334
θ_{2o}	−.0023	.0677	.0678	−.0006	.0457	.0457	−.0003	.0326	.0326
θ_{3o}	−.0030	.0687	.0688	−.0016	.0483	.0483	−.0012	.0334	.0335
				Feasible GMM estimator					
θ_{1o}	−.0031	.0780	.0781	.0000	.0552	.0552	−.0012	.0391	.0392
θ_{2o}	−.0001	.0667	.0667	.0000	.0455	.0455	.0000	.0325	.0325
θ_{3o}	.0025	.0683	.0683	.0003	.0488	.0488	−.0003	.0337	.0337

Note: The finite sample bias (BS), standard deviation (SD), and root of mean square error (RE) are computed using the corresponding estimates from 10,000 replications.

where $h_{i,t}$ is the annual hours worked for money and $w_{i,t}$ is the hourly wage rate of individual i at period t, $\alpha_{o,t}$ is a time-varying constant (invariant across the individuals), and $\varepsilon_{i,t}$ is the time varying error term. As discussed in MaCurdy, the coefficient δ_o measures the intertemporal substitution elasticity of labor supply with respect to the evolutionary wage changes, and the theoretical prediction of its sign is positive.

Due to the measurement errors in $w_{i,t}$, the ordinary least squares (OLS) estimator of (6.1) may be inconsistent. MaCurdy (1981) proposes to use a set of family background variables (father's education, mother's education, and parents' economic status when individual i was young; education; age; and the interaction between education and age of individual i) as IVs for $\Delta \log(w_{i,t})$. However, Altonji (1986) argues that the family background variables and education may only contain weak information about $\Delta \log(w_{i,t})$, and the age of individual i may not even be a valid IV. Altonji proposes to use an alternative measure of wage $w^*_{i,t}$ to construct an IV for $\Delta \log(w_{i,t})$. However, to ensure that $\Delta \log(w^*_{i,t})$ is a valid IV for $\Delta \log(w_{i,t})$, one needs to impose the strong assumption that the measurement errors in $w_{i,t}$ and $w^*_{i,t}$ are independent.[11]

Instead of using all of the IVs in MaCurdy (1981) to construct the set-1 moment conditions, we only use the parents' economic status as the credibly valid IV

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 5.5. Finite sample properties of the GMM and GMM shrinkage estimators with $(c_l, \pi_{1,o}) = (0.5, 0.8)$

	$n = 250$			$n = 500$			$n = 1{,}000$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
				Oracle estimator					
θ_{1o}	.0110	.0661	.0670	.0058	.0466	.0470	.0020	.0332	.0333
θ_{2o}	−.0016	.0672	.0672	−.0007	.0457	.0457	−.0004	.0325	.0326
θ_{3o}	−.0012	.0680	.0680	−.0015	.0482	.0482	−.0012	.0334	.0334
				GMM shrinkage estimator					
θ_{1o}	.0101	.0663	.0671	.0056	.0466	.0469	.0020	.0332	.0333
θ_{2o}	−.0015	.0671	.0671	−.0006	.0457	.0457	−.0004	.0326	.0326
θ_{3o}	−.0010	.0679	.0679	−.0015	.0482	.0482	−.0012	.0334	.0335
				Post-shrinkage GMM estimator					
θ_{1o}	.0096	.0670	.0677	.0054	.0468	.0472	.0019	.0333	.0334
θ_{2o}	−.0015	.0671	.0672	−.0006	.0457	.0457	−.0003	.0326	.0326
θ_{3o}	−.0008	.0681	.0681	−.0014	.0483	.0483	−.0012	.0334	.0335
				Feasible GMM estimator					
θ_{1o}	−.0031	.0780	.0781	.0000	.0552	.0552	−.0012	.0391	.0392
θ_{2o}	−.0001	.0667	.0667	.0000	.0455	.0455	.0000	.0325	.0325
θ_{3o}	.0025	.0683	.0683	.0003	.0488	.0488	−.0003	.0337	.0337

Note: The finite sample bias (BS), standard deviation (SD), and root of mean square error (RE) are computed using the corresponding estimates from 10,000 replications.

and include the rest of them in set-2. We also include the alternative measure of wage $w^*_{i,t}$ and the wage $w_{i,t}$ itself in set-2. This specification at least enables us to answer the following four questions. First, are the other IVs in MaCurdy, especially the age and education of individual i, valid for $\Delta \log(w_{i,t})$? If they are not, then the results of MaCurdy may contain bias incurred not only by the weak moment conditions but also by the misspecified moment conditions. Second, is the IV $\Delta \log(w^*_{i,t})$, constructed by $w^*_{i,t}$, valid for $\Delta \log(w_{i,t})$? If it is not, then the results of Altonji (1986) may be inconsistent. Third, is there measurement error in $\Delta \log(w_{i,t})$ that causes it to be an endogenous variable? If $\Delta \log(w_{i,t})$ is endogenous, then the OLS estimator of δ_o is inconsistent. On the other hand, if it is exogenous, then the GLS estimator is not only consistent but also efficient. Fourth, are the lagged logarithms of wages rates, i.e., $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$ $(s < t)$, potentially valid IVs for $\Delta \log(w_{i,t})$? These lagged logarithm of wage variables would provide some extra information about $\Delta \log(w_{i,t})$ if the shocks to the wage process were dependent. To investigate the validity of $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$ $(s < t)$ and at the same time avoid the potential pitfall of using weak IVs, we only include the previous three period differenced logarithm of wage variables (i.e., $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$, $s = t-1$, $t-2$, and $t-3$) in set-2.

Our sample is constructed from the Michigan Panel Study of Income Dynamics (PSID) data set from year 1970 to year 1981. The sample is selected according

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

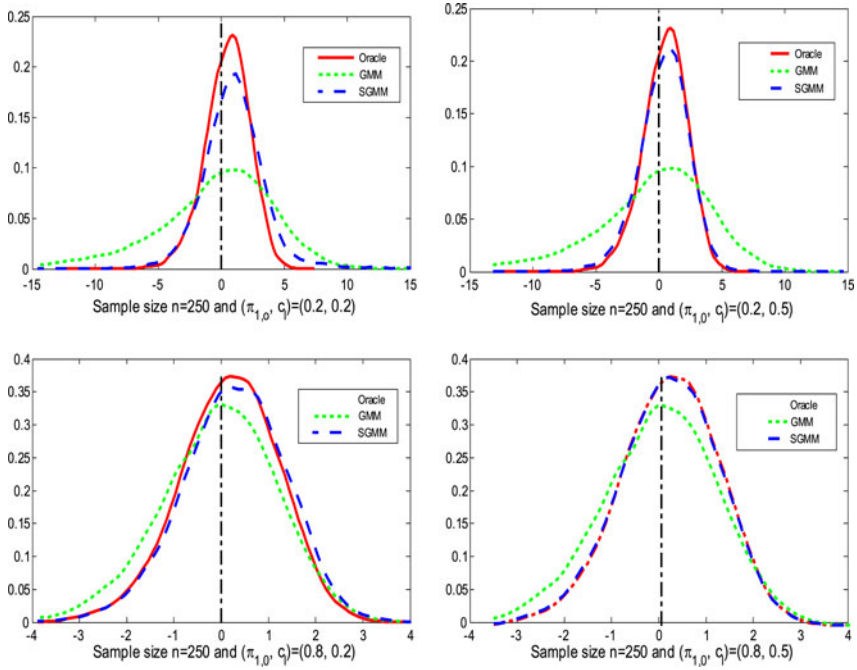


FIGURE 1. Finite sample densities of the GMM and GMM shrinkage estimators. 1. Only the densities of the estimators of $\theta_{1,o}$ are depicted. 2. The adaptive Lasso penalty is used in the GMM shrinkage estimation, and the first-step estimators of moment selection coefficients β_o are the GMM estimators. 3. The oracle estimators are the GMM estimators using both the moment conditions in set-1 and all correct moment conditions in set-2. 4. The GMM estimators only use the moment conditions in set-1. 5. SGMM refers to the GMM shrinkage estimators.

to the following criterion. First, it is limited to men with stable marriage status for the years 1970–1981. Second, individuals below age 25 in 1970 or above age 60 in 1981 are excluded to minimize the complication incurred by schooling and retirement. Third, the observations in certain years are excluded if the data are missing for the variables used in estimation.[12]

Table 6.1 presents the GMM estimators of δ_o based on the moment conditions constructed by $w_{i,t}$, IVs used in MaCurdy (1981), and IV used in Altonji (1986), respectively. The results in Table 6.1 can be summarized as follows. First, the GMM estimation using moment conditions constructed by $w_{i,t}$ gives highly misleading results, because its estimators of δ_o are negative and very significant. Second, the GMM estimation using the IVs in MaCurdy provides reasonable results, as the signs of its estimators of δ_o are positive. But the estimators have large standard errors, which indicates that the IVs used in MaCurdy only contain weak

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 6.1. GMM estimation of the labor supply equation

IV:	$\Delta \log w_{i,t}$		MaCurdy (1981)		Altonji (1986)	
	(1)	(2)	(3)	(4)	(5)	(6)
a^*	−.0008	−.0164	−.0123	−.0233	−.0125	−.0217
	(.0040)	(.0189)	(.0069)	(.0204)	(.0061)	(.0212)
δ_o	−.3937	−.3992	.0703	.2743	.1638	.2032
	(.0276)	(.0278)	(.2730)	(.4396)	(.1967)	(.2234)
$d_t?^\dagger$	No	Yes	No	Yes	No	Yes

Notes: Standard errors are in parentheses; the sample size $n = 3487$. For MaCurdy, The moment conditions are constructed using the following IVs: father's education, parents' economic status when the individual was young, education, education square, age, age square, and the interaction between age and education. For Altonji, The moment condition is constructed using an alternative measure of hour wage.

*a refers to the constant term.

†$d_t?$ denotes whether the set of time dummy variables for the years from 1971 to year 1981 are included or not.

information about the endogenous variable $w_{i,t}$. Third, the GMM estimators using Altonji's IV are reasonable and have smaller standard errors.

We next apply the GMM shrinkage estimation to the labor supply equation (6.1). The estimation results are presented in Table 6.2. As a comparison, we also include the GMM estimators based on the moment condition in set-1 and the post-shrinkage GMM estimators in Table 6.2. Columns (1)–(2) of Table 6.2 present the GMM estimators of δ_o based on the IV parent's economic status when individual was young, which provides the moment condition in set-1. Compared with other estimators in Table 6.2, the GMM estimators in columns (1)–(2) are larger in magnitude and have larger standard errors. Columns (3) and (5) of Table 6.2 contain the results of GMM shrinkage estimation with the rest of IVs (i.e., education, square of education, age and age square, father's education, and the interaction between education and age of individual i) from MaCurdy (1981), the IV (i.e., $\Delta \log(w^*_{i,t})$) from Altonji (1986), the first difference of logarithm of wage $\Delta \log(w_{i,t})$, and the lagged differences of logarithm of wage rates $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$, $s = t-1$, $t-2$, and $t-3$ in set-2. All moment conditions in set-2, except the one constructed using $\Delta \log(w_{i,t})$, are selected in the GMM shrinkage estimation. The GMM shrinkage estimators of the moment selection coefficients of $\Delta \log(w_{i,t})$ are -0.0301 (with standard error 0.0079) and -0.0340 (with standard error 0.0098) in columns (3) and (5), respectively. The GMM shrinkage estimators of δ_o have much smaller standard errors, because some moment conditions in set-2 are selected and automatically included into estimation by the GMM shrinkage method. The post-shrinkage GMM estimators in columns (4) and (6) are free of shrinkage bias, but they are only slightly different from the GMM shrinkage estimators in columns (3) and (5), respectively.

From columns (3)–(6) of Table 6.2, we see that when $\alpha_{o,t}$ is treated as a time-invariant constant term, the standard errors of the GMM estimators are relatively

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

TABLE 6.2. GMM and GMM shrinkage estimation of the labor supply equation

	GMM (1)	GMM (2)	SGMM (3)	PS-GMM (4)	SGMM (5)	PS-GMM (6)	SGMM (7)	PS-GMM (8)
a^*	−.0159	−.0224	−.0135	−.0139	−.0323	−.0313	−.0200	−.0103
	(.0257)	(.0236)	(.0051)	(.0051)	(.0192)	(.0190)	(.0185)	(.0047)
δ_o	.3241	.2682	.1648	.1659	.2395	.2509	.1615	.2096
	(1.225)	(1.143)	(.1454)	(.1450)	(.1826)	(.1814)	(.1356)	(.1663)
$d_t?^\dagger$	No	Yes	No	No	Yes	Yes	Yes	Yes

Notes: Standard errors are in parentheses; sample size $n = 3487$. In columns (3), (5), and (7), the IVs in set-2 are: father's education, education, education square, age, age square, the interaction between age and education, the first difference of alternative measure of hour wage, the first difference of constructed wage, and the lagged (3-period) differences of alternative measure wage and constructed wage. In column (7) the constraints on the constant term are also selected together with the IVs in set-2 in the GMM shrinkage estimation.

*a refers to the constant term.

†$d_t?$ denotes whether the set of time dummy variables for the years from 1971 to year 1981 are included or not.

small, though the differences between the estimators in columns (3) and (5) (or columns (4) and (6)) are nontrivial. The smaller standard errors of the estimators in columns (3) and (4) benefit from the assumption that $\alpha_{o,t}$ is constant across time, which reduces the number of unknown parameters to be estimated. If such an assumption is misspecified, then extra bias will be introduced to the GMM estimators. We next use the shrinkage method to test if $\alpha_{o,t}$ is constant or variant across the time, i.e., we penalize $|\alpha_t - \alpha_{t-1}|$ $(t = 2, \ldots, 11)$ together with the moment selection coefficients in the GMM shrinkage estimation. The results of this GMM shrinkage estimation are included in column (7). The constraints on the constants found in the GMM shrinkage estimation are $\alpha_1 = \cdots = \alpha_5$ and $\alpha_6 = \cdots = \alpha_{11}$, and all IVs, except $\Delta \log(w_{i,t})$, are selected. The GMM shrinkage estimator of δ_o in column (7) is between the GMM shrinkage estimators of δ_o in columns (3) and (5) (or columns (4) and (6)) in the magnitude and standard errors. The post-shrinkage GMM estimator based on the selected moment conditions and the selected constraints on $\alpha_{o,t}$ are included in column (8).

We summarize our findings in this empirical example as follows. First, our method selects the IVs used in MaCurdy (1981), which relieves the concern in Altonji (1986) that age and education may be invalid IVs. Second, our method picks up the IV used in Altonji and hence confirms the validity of $\Delta \log(w^*_{i,t})$ as an IV for $\Delta \log(w_{i,t})$. Third, our method does not pick up the moment condition constructed by $w_{i,t}$, which indicates that $\Delta \log(w_{i,t})$ is an endogenous variable in the labor supply equation. Fourth, our method selects the lagged differences of logarithm of wage rates $\Delta \log(w^*_{i,s})$ and $\Delta \log(w_{i,s})$ as valid IVs and hence provides extra moment conditions to more efficiently estimate δ_o. Fifth, our method does not assume that $\alpha_{o,t}$ is constant across time and detects a structural break in $\alpha_{o,t}$. Hence, our estimator of δ_o is more robust compared with the estimators based on the time-invariant assumption of $\alpha_{o,t}$, and at the same time it is more efficient

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

than the estimators based on the assumption that $\alpha_{o,t}$ changes across the time. Finally, the GMM shrinkage estimator, though consistent and asymptotically efficient, may contain some shrinkage bias in finite samples. We recommend use of the post-shrinkage GMM estimator, which is as efficient as the GMM shrinkage estimator but has smaller finite sample bias.

7. CONCLUSION

This paper proposes a GMM shrinkage method to efficiently estimate the unknown parameters θ_o identified by some moment restrictions, when there is another set of possibly misspecified moment conditions. We show that our GMM shrinkage method enjoys oracle properties, i.e., it consistently selects the correct moment conditions in the second set, and at the same time, its estimator is as efficient as the oracle GMM estimator based on all valid moment conditions. When the moment conditions in the first set fail to strongly identify θ_o, we show that the GMM shrinkage method can still consistently select the correctly specified moment conditions in the second set and, more importantly, the GMM shrinkage estimator has better stochastic properties compared with the estimators that only use the moment conditions in the first set. We provide a simple, data-driven procedure of selecting the tuning parameters in finite samples, which makes our method fully adaptive for empirical implementation. The finite sample properties of the GMM shrinkage method are explored in simulations, and an empirical labor market application illustrates the usefulness of the methods in practical work. The simulation findings show that the method performs well in terms of the correct moment selection as well as the performance characteristics of the resulting estimators.

Many research directions involving GMM shrinkage estimation arise from these findings. First, moment selection problems grow in complexity as the number of the moment conditions in (1.2) increases with the sample size. Some econometric models, such as dynamic panels, have instrument numbers that pass to infinity with the sample size, making moment or instrument selection as well as efficient estimation important practical matters in empirical research. Second, it will be interesting and useful to study the GMM shrinkage estimation when some components in β_o are local to zero. The asymptotic properties derived under the local to zero framework may help us understand the effect of model/moment selection on the post-model/moment selection inference in finite samples (Guggenberger, 2012). Third, we have not touched the issue of constructing robust CIs in GMM shrinkage estimation. As discussed earlier, the limiting distribution of the GMM shrinkage estimator is a pointwise asymptotic result, and in finite samples naive CIs constructed using these asymptotics fail to take the moment selection errors into account. It follows that naive CIs can have poor coverage probabilities, and associated hypothesis tests may suffer size distortion. Some procedures in the literature have been suggested to tackle this problem, such as Andrews, Cheng,

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

and Guggenberger (2011). It is interesting to explore whether versions of these procedures can be applied within the present framework to construct robust CIs and tests. These research directions will be explored in future work.

NOTES

1. Hereafter, the moment conditions in (1.1) and (1.2) are cited as set-1 moment conditions and set-2 moment conditions, respectively. Throughout this paper, the numbers of the moment conditions in set-1 and set-2, i.e., q and k, are fixed positive integers.

2. In this paper, we say a moment selection procedure is consistent if it selects and only selects the set of valid moment conditions in set-2 with probability approaching 1.

3. Theorem 2 of Caner (2009) shows that the centered GMM bridge estimator converges in distribution to some nonstandard random vector at the $n^{-1/2}$ rate. The nonstandard limiting random vector has positive probability measure on the point zero. Hence, Theorem 2 of Caner only implies that the GMM bridge estimators of the zero coefficients converge to zero at the rate faster than $n^{-1/2}$ and therefore does not explain why zero coefficients are estimated as zero with positive probability in finite samples.

4. Andrews (1999) notes that one can specify different β that takes some of its components as zero and the rest as unknown. As a result, θ_o and the unknown components in β can be estimated using GMM. Different specifications of β will give different GMM estimators $(\widehat{\theta}_n, \widehat{\beta}_n)$ and different values of the MSC. Consistent moment selection is indicated by the zero components in $\widehat{\beta}_n$, if $(\widehat{\theta}_n, \widehat{\beta}_n)$ asymptotically minimizes the MSC. Instead of using different specifications of β, this paper treats β_o as unknown nuisance parameters, and we use the shrinkage method to consistently identify the zero components in β_o. Hence, in place of multiple sets of GMM estimations, the shrinkage approach uses only a single-step revised GMM estimation.

5. One may argue that both the quarter of birth and rainfall are not valid IVs in some empirical researches. However, to ensure identification of the structural parameters θ_o, one has to start with a set of valid moment conditions. Hence, the point of our second example is that in practice, one can choose his/her most reliable IVs to construct the set-1 moment conditions and put the less reliable ones in set-2.

6. Formal definitions of the stochastic order symbols are provided in the Appendix.

7. The properties of the penalty function $\widehat{P}_{\lambda_n}(\cdot)$ may depend on the underlying penalized parameter. The adaptive Lasso penalty function is one typical example. We double index the random sequence $\beta_{j,n}$ $(j \in \mathcal{S}_\beta^c)$ to make it clear that Assumption 3.3.(ii) is imposed on the penalty function only when $\widehat{P}_{\lambda_n}(\cdot)$ is defined with respect to the zero components in β_o.

8. It is clear that when the data are i.i.d., the inverse of the matrix on the right-hand side of (3.20) is the semiparametric efficiency bound of θ_o implied by all valid moment conditions.

9. These moment conditions are sometimes called nearly weak moment conditions, as they allow consistent estimation of θ_o but at a rate slower than $n^{-1/2}$. By definition, they are different from the weak moment conditions, which do not allow consistent estimation of θ_o in our framework.

10. This result implies that to achieve consistent moment selection under weak identification, one needs to know a nontrivial upper bound (smaller than 1/2) of τ to design adaptive penalty for the GMM shrinkage estimation.

11. In MaCurdy (1981) and Altonji (1986), $w_{i,t}$ is constructed by dividing the annual labor income of individual i by the product of annual labor supply and gross national product (GNP) price deflator. In Altonji, $w_{i,t}^*$ is the hourly wage rate of individual i if this person is paid based on hours.

12. Following the criterion used in Altonji (1986), the imputed wage $w_{i,t}$ was treated as missing if the wage increased by 250% or more than \$13 or fell by 60% or more than \$13 from one year to another. It was also treated as missing if the real wage was less than \$0.40 in 1972 dollars. The same criterion was applied to $w_{i,t}^*$. The 250%, 60% limits were also used for labor supply. In addition, the labor supply variable was treated as missing if annual hours exceeded 4,860.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

REFERENCES

Altonji, J. (1986) Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy* 94(3), 176–215.

Andrews, D. (1999) Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 67(3), 543–563.

Andrews, D., X. Cheng, & P. Guggenberger (2011) Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests. Cowles Foundation Discussion Paper No.1813.

Angrist, J. & A. Krueger (1991) Does compulsory school attendance affect schooling and earnings. *Quarterly Journal of Economics* 106(4), 979–1014.

Arellano, M. & O. Bover (1995) Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68(1), 29–51.

Belloni, A., D. Chen, V. Chernozhukov, & C. Hansen (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6), 2369–2431.

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

Caner, M. (2009) Lasso-type GMM estimator. *Econometric Theory* 25(1), 270–290.

Caner, M. & H. Zhang (2012) Adaptive Elastic Net for Generalized Methods of Moments, Working paper, North Carolina State University.

Eichenbaum, M., L. Hansen, & K. Singleton (1988) A time series analysis of representative agent models of consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103(1), 51–78.

Fan, J. & R. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.

Gautier, E. & Tsybakov (2011) High-dimensional Instrumental Variables Regression and Confidence Sets. Preprint, arXiv: 1105.2454v2, ENSAE Paris Tech.

Geyer, C. (1994) On the asymptotics of constrained M-estimation. *Annals of Statistics* 22(4), 1993–2010.

Guggenberger, P. (2012) On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory* 28, 387–421.

Hahn, J. & G. Kuersteiner (2002) Discontinuities of weak instrument limiting distributions. *Economics Letters* 75(3), 325–331.

Hong, H., B. Preston, & M. Shum (2003) Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory* 19(6), 923–943.

Knight, K. (1999) Epi-convergence in Distribution and Stochastic Equisemicontinuity. Working paper, University of Toronto.

Knight, K. & W. Fu (2000) Asymptotics for lasso-type estimators. *Annals of Statistics* 28(5), 1356–1378.

Leeb, H. & B.M. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21(01), 21–59.

Leeb, H. & B.M. Pötscher (2008) Sparse estimators and the oracle property, or the return of the Hodges estimator. *Journal of Econometrics* 142(1), 201–211.

MaCurdy, T. (1981) An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 89(6), 1059–1085.

Miguel, E., S. Satyanath, & E. Sergenti (2004) Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* 112(4), 725–753.

Pakes, A. & Pollard, D. (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Sargan, J. (1958) The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3), 393–415.

Schmidt, M. (2010) Graphical Model Structure Learning with L1-Regularization. Thesis, University of British Columbia.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Stock, J. & J. Wright (2000) GMM with weak identification. *Econometrica* 68(5), 1055–1096.
Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

APPENDIX A

We first introduce some notations and definitions. Let $v_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(Z_i) - \mathbb{E}(g(Z_i)]$ denote the empirical process indexed by some function g. Suppose that $\{X_n\}$ is a sequence of random vectors; then for a given sequence of nonnegative constants δ_n, we write $X_n = O_p(\delta_n)$ to mean that for any constant $\varepsilon > 0$, there is a finite constant C_ε such that $\Pr(\|X_n\| > C_\varepsilon \delta_n) < \varepsilon$ eventually; we write $X_n = o_p(\delta_n)$ to mean that for any constants $\varepsilon_1, \varepsilon_2 > 0$, there is $\Pr(\|X_n\| > \varepsilon_1 \delta_n) < \varepsilon_2$ eventually.

A.1. Two Useful Lemmas. In this Appendix, we prove two lemmas that are useful for deriving the asymptotic properties of the GMM shrinkage estimator. Define

$$V_n(\theta, \beta) \equiv \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]' W_n \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right] \tag{A.1}$$

and

$$V_{0,n}(\theta, \beta) \equiv \{\mathbb{E}[\rho(Z, \theta, \beta)]\}' W_n \{\mathbb{E}[\rho(Z, \theta, \beta)]\}. \tag{A.2}$$

LEMMA A.1. *Under Assumption 3.1(iii), we have*

$$\frac{1}{2} V_{0,n}(\theta, \beta) - R_n \leq V_n(\theta, \beta) \leq 2V_{0,n}(\theta, \beta) + 2R_n \tag{A.3}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$*, where*

$$R_n \equiv \sup_{(\theta,\beta) \in \Theta \times \mathcal{B}} n^{-1} \{v_n[\rho(Z, \theta, \beta)]\}' W_n \{v_n[\rho(Z, \theta, \beta)]\}. \tag{A.4}$$

Proof. By Assumption 3.1(iii), we deduce that

$$\left[\frac{2\sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - \mathbb{E}[\rho(Z, \theta, \beta)] \right]' W_n \left[\frac{2\sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - \mathbb{E}[\rho(Z, \theta, \beta)] \right] \geq 0$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$, which implies that

$$\left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right]' W_n \left[n^{-1} \sum_{i=1}^n \rho(Z_i, \theta, \beta) \right] \geq \frac{1}{2} V_{0,n}(\theta, \beta) - R_n. \tag{A.5}$$

Note that Assumption 3.1(iii) also implies

$$\left[\frac{\sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - 2\mathbb{E}[\rho(Z, \theta, \beta)] \right]' W_n \left[\frac{\sum_{i=1}^n \rho(Z_i, \theta, \beta)}{n} - 2\mathbb{E}[\rho(Z, \theta, \beta)] \right] \geq 0$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

for all $(\theta,\beta) \in \Theta \times \mathcal{B}$, which implies that

$$\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_n \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \leq 2V_{0,n}(\theta,\beta)+2R_n. \tag{A.6}$$

From the inequalities in (A.5) and (A.6), we immediately get the claimed results in (A.3). ■

The next lemma establishes the local quadratic approximation of $V_{0,n}(\theta,\beta)$ in terms of $\sqrt{||\theta-\theta_o||^2+||\beta-\beta_o||^2}$ for all (θ,β) in shrinking neighborhoods of (θ_o,β_o), which is useful to derive the convergence rate of the GMM shrinkage estimator.

LEMMA A.2. *Under Assumptions 3.1(iii) and 3.2(ii)–(iii), we have*

$$\left[c_1+o(1)\right]||\alpha-\alpha_o||^2 \leq V_{0,n}(\theta,\beta) \leq \left[c_2+o(1)\right]||\alpha-\alpha_o||^2 \tag{A.7}$$

for all α in shrinking neighborhoods of α_o, where c_1, c_2 are generic positive finite constants.

Proof. Denote

$$g_q(Z,\theta) = \left[g_{q,1}(Z,\theta),\ldots,g_{q,q}(Z,\theta)\right]'$$

and

$$g_k(Z,\theta) = \left[g_{k,1}(Z,\theta),\ldots,g_{k,k}(Z,\theta)\right]'.$$

First, note that by Assumption 3.2(ii),

$$m(\alpha) = \begin{pmatrix} G_q(\theta) \\ G_k(\theta)-\beta \end{pmatrix} = \begin{pmatrix} \frac{\partial G_q(\widetilde{\theta})}{\partial\theta'} & 0 \\ \frac{\partial G_k(\widetilde{\theta})}{\partial\theta'} & -I_k \end{pmatrix} \begin{pmatrix} \theta-\theta_o \\ \beta-\beta_o \end{pmatrix} \equiv \frac{\partial m(\widetilde{\theta})}{\partial\alpha'} \begin{pmatrix} \theta-\theta_o \\ \beta-\beta_o \end{pmatrix}, \tag{A.8}$$

where

$$\frac{\partial G_q(\widetilde{\theta})}{\partial\theta'} = \left[\left(\frac{\partial E\left[g_{q,1}(Z,\widetilde{\theta}_1)\right]}{\partial\theta'}\right)',\ldots,\left(\frac{\partial E\left[g_{q,q}(Z,\widetilde{\theta}_q)\right]}{\partial\theta'}\right)'\right]',$$

$$\frac{\partial G_k(\widetilde{\theta})}{\partial\theta'} = \left[\left(\frac{\partial E\left[g_{k,1}(Z,\widetilde{\theta}_{q+1})\right]}{\partial\theta'}\right)',\ldots,\left(\frac{\partial E\left[g_{k,k}(Z,\widetilde{\theta}_{q+k})\right]}{\partial\theta'}\right)'\right]',$$

$\widetilde{\theta}_j$ $(j=1,\ldots,q+k)$ lies between θ and θ_o, and I_k is a $k\times k$ identity matrix. As θ is in the shrinking neighborhood of θ_o and $\frac{\partial G_l(\theta)}{\partial\theta'}$ $(l=q,k)$ is continuous in θ, we deduce that

$$\frac{\partial G_l(\widetilde{\theta})}{\partial\theta'} = \frac{\partial G_l(\theta_o)}{\partial\theta'}+o(1) \quad \text{for } l=q,k. \tag{A.9}$$

By (A.8), (A.9), and Cauchy-Schwarz inequality, we have

$$m(\alpha) = \frac{\partial m(\theta_o)}{\partial\alpha'}(\alpha-\alpha_o)+o(||\alpha-\alpha_o||). \tag{A.10}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Using Assumption 3.1(iii), the result in (A.10), and Cauchy-Schwarz inequality, we get

$$V_{0,n}(\theta,\beta)=(\alpha-\alpha_o)'\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right]'W_n\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right](\alpha-\alpha_o)+o\big(||\alpha-\alpha_o||^2\big). \tag{A.11}$$

As $\frac{\partial G_q(\theta_o)}{\partial\theta'}$ has full column rank and W_n is strictly positive definite, $\frac{\partial m(\theta_o)}{\partial\alpha'}$ has full rank and $\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right]'W_n\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right]$ is strictly positive definite. Let $\gamma_{1,n}$ and $\gamma_{2,n}$ $(\gamma_{1,n},\gamma_{2,n}>0)$ denote the smallest and largest eigenvalues of $\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right]'W_n\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right]$; then by Assumptions 3.1(iii) and 3.2(iii), we have

$$0<c_1\le\gamma_{1,n}\le\gamma_{2,n}\le c_2<\infty,$$

which together with (A.11) implies that

$$\begin{aligned}&(\alpha-\alpha_o)'\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right]'W_n\left[\frac{\partial m(\theta_o)}{\partial\alpha'}\right](\alpha-\alpha_o)+o\big(||\alpha-\alpha_o||^2\big)\\&\quad\le\gamma_{2,n}||\alpha-\alpha_o||^2+o(||\alpha-\alpha_o||^2)\le\big[c_2+o(1)\big]\,||\alpha-\alpha_o||^2.\end{aligned} \tag{A.12}$$

The right inequality in (A.7) is implied by (A.12). The left inequality in (A.7) can be similarly derived. This finishes the proof. ■

A.2. Proof of the Main Results in Section 3.

Proof of Lemma 3.1. By the definition of $\widehat{\alpha}_n$, we have

$$V_n(\widehat{\theta}_n,\widehat{\beta}_n)+\sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\widehat{\beta}_n)\le V_n(\theta_o,\beta_o)+\sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_{o,j}). \tag{A.13}$$

Applying Lemma A.1 and Assumption 3.1(iv), we deduce from (A.13) that

$$V_{0,n}(\widehat{\theta}_n,\widehat{\beta}_n)\le 2\sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_{o,j})+4R_n, \tag{A.14}$$

where R_n is defined in Lemma A.1.

From Assumption 3.1(ii) and the definition of $\rho(Z,\theta,\beta)$, we get

$$\sup_{(\theta,\beta)\in\Theta\times\mathcal{B}}\frac{v_n\left[\rho(Z,\theta,\beta)\right]}{\sqrt{n}}=\sup_{\theta\in\Theta}n^{-1}\sum_{i=1}^{n}\left\{g_{q+k}(Z_i,\theta)-\mathbb{E}\left[g_{q+k}(Z_i,\theta)\right]\right\}=o_p(1). \tag{A.15}$$

By the triangle inequality, ULLN in (A.15), and Assumptions 3.1(iii)–(iv), we have

$$R_n=o_p(1)\quad\text{and}\quad\sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_{o,j})=o_p(1). \tag{A.16}$$

From Assumption 3.1(iii) and the results in (A.14) and (A.16), we deduce that

$$\left\|G_q(\widehat{\theta}_n)\right\|=o_p(1)\quad\text{and}\quad\left\|G_k(\widehat{\theta}_n)-\widehat{\beta}_n\right\|=o_p(1). \tag{A.17}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Now, the first result in (A.17) together with Assumption 3.1(i) implies that $\widehat{\theta}_n \to_p \theta_o$. From the second result in (A.17), the triangle inequality, the consistency of $\widehat{\theta}_n$, and Assumption 3.1(i), we have

$$\begin{aligned} o_p(1) &= \left\| G_k(\widehat{\theta}_n) - \widehat{\beta}_n \right\| \\ &\geq \left| \left\| G_k(\widehat{\theta}_n) - G_k(\theta_o) \right\| - \left\| \widehat{\beta}_n - \beta_o \right\| \right| \\ &= \left| \left\| \widehat{\beta}_n - \beta_o \right\| + o_p(1) \right|, \end{aligned} \tag{A.18}$$

which implies $\widehat{\beta}_n \to_p \beta_o$. ■

Proof of Lemma 3.2. Using the inequalities in (A.3) and (A.13), we get

$$\frac{1}{2} V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + 2R_n, \tag{A.19}$$

where R_n is defined in Lemma A.1. By Assumptions 3.1(iv), 3.2(iv), and the inequality in (A.19), we have

$$V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) + 2 \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right] \leq 4R_n. \tag{A.20}$$

Next, by Assumption 3.2(iv), Taylor expansion, the triangle inequality, and Cauchy-Schwarz inequality, we get

$$\begin{aligned} &\left| \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n}(\beta_{o,j}) - \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) \right] \right| \\ &\quad = \left| \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}'_{\lambda_n}(\beta_{o,j})(\widehat{\beta}_{n,j} - \beta_{o,j}) + \frac{1}{2} \widehat{P}''_{\lambda_n}(\widetilde{\beta}_j)(\widehat{\beta}_{n,j} - \beta_{o,j})^2 \right] \right| \\ &\quad \leq \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| \left\| \widehat{\alpha}_n - \alpha_o \right\| + \max_{j \in \mathcal{S}_\beta} \left| \frac{\widehat{P}''_{\lambda_n}(\beta_{o,j})}{2} + o_p(1) \right| \left\| \widehat{\alpha}_n - \alpha_o \right\|^2, \end{aligned} \tag{A.21}$$

where $\widetilde{\beta}_j$ lies between $\beta_{o,j}$ and $\widehat{\beta}_{n,j}$ for $j \in \mathcal{S}_\beta$. From Lemma A.2, we obtain

$$V_{0,n}(\widehat{\theta}_n, \widehat{\beta}_n) \geq \left[c_1 + o_p(1) \right] \| \widehat{\alpha}_n - \alpha_o \|^2, \tag{A.22}$$

where $c_1 > 0$ is a finite constant. The inequality in (A.22), together with Assumption 3.2(iv) and the inequalities in (A.20), and (A.21) implies that

$$\left[c_1 + o_p(1) \right] \| \widehat{\alpha}_n - \alpha_o \|^2 - 2 \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| \left\| \widehat{\alpha}_n - \alpha_o \right\| \leq 4R_n. \tag{A.23}$$

By Assumption 3.2(i), we have $R_n = O_p(n^{-1})$. When the sample size n is large enough, by definition the probability that $c_1 + o_p(1) \leq c_1/2$ is strictly smaller than any given small

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

number $\omega/2$, which together with the inequality in (A.23) implies that

$$\begin{aligned}
&\Pr\left(\frac{||\widehat{\alpha}_n-\alpha_o||}{\delta_n}>M\right)\\
&\quad<\Pr\left(\inf_{\{\alpha\in\mathcal{A}:\,||\alpha-\alpha_o||>M\delta_n\}}\left\{||\alpha-\alpha_o||^2-4c_1^{-1}b_n||\alpha-\alpha_o||\right\}\le 8c_1^{-1}R_n\right)+\frac{\omega}{2}\\
&\quad\le\Pr\left(\inf_{\{\alpha\in\mathcal{A}:\,||\alpha-\alpha_o||>M\delta_n\}}||\alpha-\alpha_o||\le 2c_1^{-1}\left(b_n+\sqrt{b_n^2+2R_nc_1}\right)\right)+\frac{\omega}{2}\\
&\quad\le\Pr\left(M\le\frac{2c_1^{-1}\left(b_n+\sqrt{b_n^2+2R_nc_1}\right)}{\delta_n}\right)+\frac{\omega}{2},
\end{aligned}\tag{A.24}$$

where $\delta_n=\max\{b_n,n^{-1/2}\}$. By definition, $\frac{2c_1^{-1}(b_n+\sqrt{b_n^2+2R_nc_1})}{\delta_n}=O_p(1)$. Hence, we can choose some large enough number M_ω such that

$$\Pr\left(M_\omega\le\frac{2c_1^{-1}\left(b_n+\sqrt{b_n^2+2R_nc_1}\right)}{\delta_n}\right)<\frac{\omega}{2}.\tag{A.25}$$

This and the results in (A.24) immediately imply that

$$\Pr\left(\frac{||\widehat{\alpha}_n-\alpha_o||}{\delta_n}>M_\omega\right)<\omega$$

eventually, which gives us $\left\|\widehat{\alpha}_n-\alpha_o\right\|=O_p(\delta_n)$. ■

Proof of Theorem 3.3. On the event $\left\{\widehat{\beta}_{n,j}\neq 0\right\}$ for some $j\in\mathcal{S}_\beta^c$, we have the KKT optimality condition

$$2\left[n^{-1/2}\sum_{i=1}^n\frac{\partial\rho\left(Z_i,\widehat{\theta}_n,\widehat{\beta}_n\right)}{\partial\beta_j}\right]'W_n\left[n^{-1/2}\sum_{i=1}^n\rho\left(Z_i,\widehat{\theta}_n,\widehat{\beta}_n\right)\right]+n\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})=0,\tag{A.26}$$

which implies

$$\left|W_n(q+j)\left[n^{-1/2}\sum_{i=1}^n\rho\left(Z_i,\widehat{\theta}_n,\widehat{\beta}_n\right)\right]\right|=\frac{\sqrt{n}\left|\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\right|}{2},\tag{A.27}$$

where $W_n(q+j)$ denotes the $(q+j)$th row of the weight matrix W_n.

By Assumption 3.2(ii) and the consistency of $\widehat{\theta}_n$, there is

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^n\rho(Z_i,\widehat{\alpha}_n)&=\upsilon_n\left[\rho(Z,\widehat{\alpha}_n)\right]+n^{1/2}m(\widehat{\alpha}_n)\\
&=\upsilon_n\left[\rho(Z,\widehat{\alpha}_n)\right]+\frac{\partial m(\widetilde{\theta}_n)}{\partial\alpha'}\left[n^{1/2}(\widehat{\alpha}_n-\alpha_o)\right],
\end{aligned}$$

where $\frac{\partial m(\theta)}{\partial\alpha'}$ is defined in (A.8), $\widetilde{\theta}_n=(\widetilde{\theta}_{1,n},\ldots,\widetilde{\theta}_{q+k,n})$, and $\widetilde{\theta}_{j,n}$ $(j=1,\ldots,q+k)$ lies between θ_o and $\widehat{\theta}_n$. From Assumption 3.2(i), we have $\upsilon_n\left[\rho(Z_i,\widehat{\alpha}_n)\right]=O_p(1)$. By

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Lemma 3.2 and Assumption 3.3(i), we have $n^{1/2}(\widehat{\alpha}_n - \alpha_o) = O_p(1)$. By the triangle inequality, Assumption 3.2(ii), and the consistency of $\widehat{\theta}_n$, we deduce that

$$\left\| \frac{\partial m(\widetilde{\theta}_n)}{\partial \alpha'} \right\| \leq \left\| \frac{\partial m(\widetilde{\theta}_n)}{\partial \alpha'} - \frac{\partial m(\theta_o)}{\partial \alpha'} \right\| + \left\| \frac{\partial m(\theta_o)}{\partial \alpha'} \right\| = O_p(1).$$

Hence, we have $n^{-1/2}\sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n) = O_p(1)$, which combined with Assumption 3.1.(iii) implies that

$$\left| W_n(q+j) \left[n^{-\frac{1}{2}} \sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n) \right] \right| = O_p(1). \tag{A.28}$$

By Lemma 3.2 and Assumption 3.3(i), we have $n^{1/2}\widehat{\beta}_{n,j} = O_p(1)$ for all $j \in \mathcal{S}_\beta^c$. Hence, conditional on the event $\{\widehat{\beta}_{n,j} \neq 0\}$ for some $j \in \mathcal{S}_\beta^c$, we can invoke Assumption 3.3(ii) to deduce that

$$\frac{n^{1/2}\left|\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\right|}{2} = \frac{n^{1/2} r_n \lambda_n}{2} \frac{\left|\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\right|}{r_n \lambda_n} \to_p \infty \tag{A.29}$$

for $j \in \mathcal{S}_\beta^c$. Now, using (A.27), (A.28), and (A.29), we deduce that $\Pr\left(\widehat{\beta}_{n,j} = 0\right) \to 1$ as $n \to \infty$ for any $j \in \mathcal{S}_\beta^c$. ■

Proof of Theorem 3.4. Recall that $\alpha'_{o,\mathcal{S}} = (\theta'_o, \beta'_{o,+})$ and accordingly $\widehat{\alpha}'_{n,\mathcal{S}} = (\widehat{\theta}'_n, \widehat{\beta}'_{n,+})$. For any compact subset K in $R^{d_\theta + k - k_o}$, we denote any element $u_\mathcal{S} \in K$ as $u'_\mathcal{S} = (u'_\theta, u'_{\beta_+})$, where u_θ are the first d_θ elements in $u_\mathcal{S}$ and u_{β_+} are the last $k - k_o$ elements in $u_\mathcal{S}$. Denote

$$\begin{aligned} V_{2,n}(u_\mathcal{S}) \equiv & \left[n^{-1/2} \sum_{i=1}^n \rho^s \left(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) \right]' W_n \left[n^{-1/2} \sum_{i=1}^n \rho^s \left(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) \right] \\ & - \left[n^{-1/2} \sum_{i=1}^n \rho(Z_i, \alpha_o) \right]' W_n \left[n^{-1/2} \sum_{i=1}^n \rho(Z_i, \alpha_o) \right] \\ & + n \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n} \left(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}} \right) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right] \\ \equiv & V^*_{2,n}(u_\mathcal{S}) + n \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n} \left(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}} \right) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right], \end{aligned}$$

where $\rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) \equiv \rho(Z_i, \theta_o + \frac{u_\theta}{\sqrt{n}}, \beta_{o,+} + \frac{u_{\beta_+}}{\sqrt{n}}, \beta_{o,-})$. From Theorem 3.3, we know that $\widehat{\beta}_{n,-} = 0$ w.p.a.1. Thus, $\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right)$ is the minimizer of $V_{2,n}(u_\mathcal{S})$ w.p.a.1.

If we define

$$\mathcal{F}_n \equiv \left\{ f^n_{u_\mathcal{S}}(Z) = \rho^s \left(Z, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}} \right) - \rho(Z, \alpha_o) : u_\mathcal{S} \in K \right\},$$

then by Assumption 3.2(i) we know that $\mathcal{F}_n$ is a Donsker class. As K is compact, there exists some constant C_k, such that $\sup_{u_\mathcal{S} \in K} \left\| n^{-1/2} u_\mathcal{S} \right\| \leq n^{-1/2} C_k = o(1)$. Now we can

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

use Lemma 2.17 in Pakes and Pollard (1989) to deduce that

$$v_n\left\{\rho^s\left(Z,\alpha_{o,\mathcal{S}}+\frac{u_\mathcal{S}}{\sqrt{n}}\right)-\rho(Z,\alpha_o)\right\}=o_p(1), \tag{A.30}$$

uniformly over $u_\mathcal{S}\in K$.

By Assumption 3.2(ii) and the compactness of K, we have

$$\sqrt{n}\left\{\mathbb{E}\left[\rho^s\left(Z,\alpha_{o,\mathcal{S}}+\frac{u_\mathcal{S}}{\sqrt{n}}\right)\right]-\mathbb{E}[\rho(Z,\alpha_o)]\right\}=\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}u_\mathcal{S}+o(1), \tag{A.31}$$

uniformly over $u_\mathcal{S}\in K$. Thus, (A.30) and (A.31) imply that uniformly over $u_\mathcal{S}\in K$ there is

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^n\rho^s\left(Z_i,\alpha_{o,\mathcal{S}}+\frac{u_\mathcal{S}}{\sqrt{n}}\right)\\
&\quad=v_n\left\{\rho^s\left(Z,\alpha_{o,\mathcal{S}}+\frac{u_\mathcal{S}}{\sqrt{n}}\right)-\rho(Z,\alpha_o)\right\}+v_n[\rho(Z,\alpha_o)]\\
&\qquad+\sqrt{n}\left\{\mathbb{E}\left[\rho^s\left(Z,\alpha_{o,\mathcal{S}}+\frac{u_\mathcal{S}}{\sqrt{n}}\right)\right]-\mathbb{E}[\rho(Z,\alpha_o)]\right\}\\
&\quad=v_n[\rho(Z,\alpha_o)]+\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}u_\mathcal{S}+o_p(1).
\end{aligned} \tag{A.32}$$

Now, we can use the result in (A.32), Assumptions 3.2(i) and 3.4(i), and the compactness of K to deduce that

$$\begin{aligned}
V^*_{2,n}(u_\mathcal{S})=u'_\mathcal{S}&\left[\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}\right]'W_o\left[\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}\right]u_\mathcal{S}\\
&+2u'_\mathcal{S}\left[\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}\right]'W_o\{v_n[\rho(Z,\alpha_o)]\}+o_p(1),
\end{aligned} \tag{A.33}$$

uniformly over $u_\mathcal{S}\in K$. If $j\in\mathcal{S}_\beta$, then by Assumptions 3.2(iv) and 3.3(i) and the compactness of K, we have

$$\begin{aligned}
&n\left[\widehat{P}_{\lambda_n}\left(\beta_{o,j}+\frac{u_{\beta_+,j}}{\sqrt{n}}\right)-\widehat{P}_{\lambda_n}(\beta_{o,j})\right]\\
&\quad=\sqrt{n}\,\widehat{P}'_{\lambda_n}(\beta_{o,j})u_{\beta_+,j}+\left[\widehat{P}''_{\lambda_n}(\beta_{o,j})+o_p(1)\right]u^2_{\beta_+,j}\to 0
\end{aligned} \tag{A.34}$$

uniformly in $u_{\beta_+,j}$.

Using Assumption 3.4(ii) and the results in (A.33) and (A.34), we get

$$V_{2,n}(u_\mathcal{S})\to_d V_2(u_\mathcal{S})=u'_\mathcal{S}M_\mathcal{S}u_\mathcal{S}+2u'_\mathcal{S}\left[\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}\right]'W_o\Psi(\theta_o) \tag{A.35}$$

uniformly over $u_\mathcal{S}\in K$. It is clear that $V_2(u_\mathcal{S})$ is uniquely minimized at

$$u^*_\mathcal{S}=-M^{-1}_\mathcal{S}\left[\frac{\partial m(\theta_o)}{\partial\alpha'_\mathcal{S}}\right]'W_o\Psi(\theta_o). \tag{A.36}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

By Lemma 3.2 and Assumption 3.3(i), there is

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right)=O_p(1). \tag{A.37}$$

Now, the uniform weak convergence in (A.35), the unique minimization in (A.36), and the asymptotic tightness of $\widehat{\alpha}_{n,\mathcal{S}}$ in (A.37) enable us to invoke the ACMT to deduce that

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right)\to_d N\left(0,M_{\mathcal{S}}^{-1}\Sigma_{\mathcal{S}}M_{\mathcal{S}}^{-1}\right). \quad \blacksquare$$

Proof of Theorem 3.5. The first result is implied by Lemma 3.1 and Theorem 3.3, so we only need to show the second claim. First, note that if

$$W_n\to_p W_o=\left\{\mathbb{E}\left[\Psi(\theta_o)\Psi'(\theta_o)\right]\right\}^{-1},$$

then the centered limiting distribution in (3.18) will be simplified to

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right)\to_d N\left(0,M_{\mathcal{S}}^{-1}\right). \tag{A.38}$$

Let Ω_{θ_o} denote the leading $d_{\theta_o}\times d_{\theta_o}$ submatrix of $M_{\mathcal{S}}^{-1}$ and $\frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'}=\frac{\partial E[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial\theta'}$. Note that we can rewrite

$$\begin{aligned}
M_{\mathcal{S}}&=\begin{pmatrix}\frac{\partial\mathbb{E}[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial\theta'} & 0\\ \frac{\partial\mathbb{E}[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial\theta'} & -I_{k-k_o}\end{pmatrix}'W_o\begin{pmatrix}\frac{\partial\mathbb{E}[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial\theta'} & 0\\ \frac{\partial\mathbb{E}[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial\theta'} & -I_{k-k_o}\end{pmatrix}\\
&=\begin{pmatrix}\frac{\partial m_e(\theta_o)}{\partial\theta'} & 0\\ \frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'} & -I_{k-k_o}\end{pmatrix}'\begin{pmatrix}W_{11} & W_{12}\\ W_{21} & W_{22}\end{pmatrix}\begin{pmatrix}\frac{\partial m_e(\theta_o)}{\partial\theta'} & 0\\ \frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'} & -I_{k-k_o}\end{pmatrix}\\
&=\begin{pmatrix}\Sigma_{\mathcal{S}}^{11} & \Sigma_{\mathcal{S}}^{12}\\ \Sigma_{\mathcal{S}}^{21} & W_{22}\end{pmatrix},
\end{aligned} \tag{A.39}$$

where

$$\Sigma_{\mathcal{S}}^{12}=-\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]'W_{12}-\left[\frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'}\right]'W_{22}=\left(\Sigma_{\mathcal{S}}^{21}\right)'$$

and

$$\begin{aligned}
\Sigma_{\mathcal{S}}^{11}&=\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]'W_{11}\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]+\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]'W_{12}\left[\frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'}\right]\\
&\quad+\left[\frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'}\right]'W_{21}\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]+\left[\frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'}\right]'W_{22}\left[\frac{\partial G_{\beta_+}(\theta_o)}{\partial\theta'}\right].
\end{aligned}$$

Using the inverse formula of partitioned matrix and (A.39), we have

$$\begin{aligned}
\Omega_{\theta_o}^{-1}&=\Sigma_{\mathcal{S}}^{11}-\Sigma_{\mathcal{S}}^{12}W_{22}^{-1}\Sigma_{\mathcal{S}}^{21}\\
&=\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]'\left(W_{11}-W_{12}W_{22}^{-1}W_{21}\right)\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]\\
&=\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]'V_{e,o}^{-1}\left[\frac{\partial m_e(\theta_o)}{\partial\theta'}\right]=\left(\Sigma^*\right)^{-1},
\end{aligned} \tag{A.40}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

where the last equality is by the fact that $\left(W_{11} - W_{12}W_{22}^{-1}W_{21}\right)^{-1} = V_{e,o}$. Now, using the results in (A.38), (A.40), and the continuous mapping theorem (CMT), we deduce that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_o\right) \to_d N(0, \Sigma^*),$$

which finishes the proof. ■

A.3. Proof of the Main Results in Section 4.

Proof of Lemma 4.1. Using similar arguments to those in the proof of Lemma 3.1, we get

$$V_{0,n}\left(\widehat{\theta}_n, \widehat{\beta}_n\right) \leq 2\sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + 4R_n, \tag{A.41}$$

where R_n is defined in Lemma A.1. From the above inequality, we can use Assumptions 3.1(iii) and 4.1(iii), (v) to deduce that

$$\left\|n^{-\tau} G_{q,n}(\widehat{\theta}_n)\right\|^2 + \left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\|^2 = O_p\left(n^{-1} + \max_{j\in\mathcal{S}_\beta} \widehat{P}_{\lambda_n}(\beta_{o,j})\right), \tag{A.42}$$

which implies that

$$\left\|G_{q,n}(\widehat{\theta}_n)\right\|^2 = O_p\left(n^{2\tau-1} + n^{2\tau} \max_{j\in\mathcal{S}_\beta} \widehat{P}_{\lambda_n}(\beta_{o,j})\right) = o_p(1).$$

Hence, using the uniform approximation in Assumption 4.1(i), we deduce that $\left\|G_q(\widehat{\theta}_n)\right\| = o_p(1)$, which together with the identification condition in Assumption 4.1(ii) implies the consistency of $\widehat{\theta}_n$. The consistency of $\widehat{\beta}_n$ can be proved using the consistency of $\widehat{\theta}_n$ and similar arguments in the proof of Lemma 3.1.

Next, we derive the convergence rate of $\widehat{\alpha}_n$. From the inequality (A.20) and Assumption 4.1(v), we have

$$V_{0,n}\left(\widehat{\theta}_n, \widehat{\beta}_n\right) \leq 2\sum_{j\in\mathcal{S}_\beta} \left|\widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) - \widehat{P}_{\lambda_n}(\beta_{o,j})\right| + 4R_n. \tag{A.43}$$

Using similar arguments in the proof of Lemma A.2, we can apply Assumptions 3.1(iii) and 4.1(iv) to deduce that

$$\begin{aligned} V_{0,n}\left(\widehat{\theta}_n, \widehat{\beta}_n\right) &\geq cn^{-2\tau}\left\|G_{q,n}(\widehat{\theta}_n) - G_{q,n}(\theta_o)\right\|^2 + c\left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\|^2 \\ &= n^{-2\tau}\left[c + o_p(1)\right]\left\|\widehat{\theta}_n - \theta_o\right\|^2 + c\left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\|^2, \end{aligned} \tag{A.44}$$

where c is a generic finite positive constant. From (A.43) and (A.44), we deduce that

$$n^{-2\tau}\left[1 + o_p(1)\right]\left\|\widehat{\theta}_n - \theta_o\right\|^2 + \left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\| \leq c\sum_{j\in\mathcal{S}_\beta} \left|\widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) - \widehat{P}_{\lambda_n}(\beta_{o,j})\right| + cR_n. \tag{A.45}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

By the consistency of $\widehat{\beta}_n$, we know that when the sample size is large, the probability that $\beta_j \notin N_{\beta_{o,j}}$ is strictly smaller than any prespecified small number. Hence, using Assumption 4.1(v) and the consistency of $\widehat{\beta}_n$, without loss of generality, we have

$$\sum_{j\in\mathcal{S}_\beta}\left|\widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j})-\widehat{P}_{\lambda_n}(\beta_{o,j})\right| \le \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|\left\|\widehat{\beta}_n-\beta_o\right\| + \max_{j\in\mathcal{S}_\beta}\sup_{\beta_j\in N_{\beta_{o,j}}}\left|\widehat{P}''_{\lambda_n}(\beta_j)\right|\left\|\widehat{\beta}_n-\beta_o\right\|^2. \tag{A.46}$$

By Cauchy-Schwarz inequality,

$$\left\|G_k(\widehat{\theta}_n)-\widehat{\beta}_n\right\|^2 \ge \left\|\widehat{\beta}_n-\beta_o\right\|^2 - 2\left\|\widehat{\beta}_n-\beta_o\right\| J_{1,n} + J_{1,n}^2, \tag{A.47}$$

where $J_{1,n} = \left\|G_k(\widehat{\theta}_n)-G_k(\theta_o)\right\| = O_p\left(\left\|\widehat{\theta}_n-\theta_o\right\|\right)$ by Assumption 4.1(iv). Denote $J_{2,n} = n^{-2\tau}\left[c+o_p(1)\right]\left\|\widehat{\theta}_n-\theta_o\right\|^2$; then from Assumptions 4.1(v), (A.45), and (A.46), we get

$$\left[1+o_p(1)\right]\left\|\widehat{\beta}_n-\beta_o\right\|^2 - c\left(b_n+J_{1,n}\right)\left\|\widehat{\beta}_n-\beta_o\right\| \le 4R_n - J_{1,n}^2 - J_{2,n}, \tag{A.48}$$

where $b_n = \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|$. As $J_{1,n}^2 = O_p\left(\left\|\widehat{\theta}_n-\theta_o\right\|^2\right)$ and $J_{2,n} = O_p\left(\left\|\widehat{\theta}_n-\theta_o\right\|^2\right)$, from the inequality in (A.48), we deduce that

$$\left\|\widehat{\beta}_n-\beta_o\right\| = O_p\left(\max\left\{\left\|\widehat{\theta}_n-\theta_o\right\|, b_n, n^{-1/2}\right\}\right). \tag{A.49}$$

Note that if $\left\|\widehat{\theta}_n-\theta_o\right\| = O_p\left(\max\left\{b_n, n^{-1/2}\right\}\right)$, then the claimed result is proved. Hence, we only need to consider the case that $\left\|\widehat{\theta}_n-\theta_o\right\|$ has a convergence rate slower than $\max\left\{b_n, n^{-1/2}\right\}$. In such a case, (A.49) implies that

$$\left\|\widehat{\beta}_n-\beta_o\right\| = O_p\left(\left\|\widehat{\theta}_n-\theta_o\right\|\right). \tag{A.50}$$

Now from (A.45), we have

$$\left[1+o_p(1)\right]\left\|\widehat{\theta}_n-\theta_o\right\|^2 \le cn^{2\tau}\sum_{j\in\mathcal{S}_\beta}\left|\widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j})-\widehat{P}_{\lambda_n}(\beta_{o,j})\right| + cn^{2\tau}R_n,$$

which together with Assumption 4.1(v), (A.46), and (A.50) implies that

$$\left[1+o_p(1)\right]\left\|\widehat{\theta}_n-\theta_o\right\|^2 - O_p\left(n^{2\tau}b_n\right)\left\|\widehat{\theta}_n-\theta_o\right\| - n^{2\tau}cR_n \le 0. \tag{A.51}$$

By the above inequality, Assumption 4.1(iii), and similar arguments in showing (A.24) and (A.25), we deduce that

$$\left\|\widehat{\theta}_n-\theta_o\right\| = O_p\left(\max\left\{n^{2\tau}b_n, n^{-1/2+\tau}\right\}\right).$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

From (A.50), we know that $\left\|\widehat{\beta}_n - \beta_o\right\|$ has the same rate of convergence. This finishes the proof. ■

Proof of Lemma 4.2. On the event $\left\{\widehat{\beta}_{n,j} \neq 0\right\}$ for some $j \in \mathcal{S}_\beta^c$, we have the KKT optimality condition

$$\left| W_n(q+j)\left[n^{-1/2}\sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n)\right]\right| = \frac{n^{1/2}\left|\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\right|}{2}, \tag{A.52}$$

where $W_n(q+j)$ denotes the $(q+j)$th row of the weight matrix W_n. Define

$$m_n(\alpha) \equiv \mathbb{E}\begin{bmatrix} g_{q,n}(Z,\theta) \\ g_{k,n}(Z,\theta) - \beta \end{bmatrix} = \begin{bmatrix} G_{q,n}(\theta) \\ G_k(\theta) - \beta \end{bmatrix}.$$

Then we note that

$$n^{-1/2-\tau}\sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n) = n^{-\tau} v_n\left[\rho(Z, \widehat{\alpha}_n)\right] + n^{1/2-\tau} m_n(\widehat{\alpha}_n), \tag{A.53}$$

By Lemma 4.1 and Assumptions 4.1(iv) and 4.2(i), we get

$$n^{1/2-\tau}\left[G_{q,n}(\widehat{\theta}_n) - G_{q,n}(\theta_o)\right] = \frac{\partial G_{q,n}(\widetilde{\theta}_n)}{\partial \theta'}\left[n^{1/2-\tau}\left(\widehat{\theta}_n - \theta_o\right)\right] = O_p(1), \tag{A.54}$$

where

$$\frac{\partial G_{q,n}(\widetilde{\theta}_n)}{\partial \theta'} = \left[\left(\frac{\partial G_{q,n,1}(\widetilde{\theta}_{1,n})}{\partial \theta'}\right)', \ldots, \left(\frac{\partial G_{q,n,q}(\widetilde{\theta}_{q,n})}{\partial \theta'}\right)'\right]'$$

and $\widetilde{\theta}_{j,n}$ $(j = 1, \ldots, q)$ lies between $\widehat{\theta}_n$ and θ_o. Similarly, we have

$$\begin{aligned} n^{1/2-\tau}\left[G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right] &= n^{1/2-\tau}\left\{\left[G_k(\widehat{\theta}_n) - G_k(\theta_o)\right] - \left(\widehat{\beta}_n - \beta_o\right)\right\} \\ &= \frac{\partial G_k(\widetilde{\theta}_n)}{\partial \theta'}\left[n^{1/2-\tau}\left(\widehat{\theta}_n - \theta_o\right)\right] + O_p(1) = O_p(1), \end{aligned} \tag{A.55}$$

where

$$\frac{\partial G_k(\widetilde{\theta}_n)}{\partial \theta'} = \left[\left(\frac{\partial G_{k,1}(\widetilde{\theta}_{q+1,n})}{\partial \theta'}\right)', \ldots, \left(\frac{\partial G_{k,k}(\widetilde{\theta}_{q+k,n})}{\partial \theta'}\right)'\right]'$$

and $\widetilde{\theta}_{j,n}$ $(j = q+1, \ldots, q+k)$ lies between $\widehat{\theta}_n$ and θ_o. Hence, from the results in (A.53)–(A.55) and Assumptions 3.1(iii) and 4.1(iii), we deduce that

$$\left| W_n(q+j)\left[n^{-1/2-\tau}\sum_{i=1}^n \rho(Z_i, \widehat{\alpha}_n)\right]\right| = O_p(1). \tag{A.56}$$

It is clear that $n^{1/2-\tau}\widehat{\beta}_{n,j} = O_p(1)$ for all $j \in \mathcal{S}_\beta^c$. Hence, conditional on the event $\{\widehat{\beta}_{n,j} \neq 0\}$ for some $j \in \mathcal{S}_\beta^c$, we can invoke Assumption 4.2(ii) to deduce that

$$\frac{n^{1/2-\tau}\left|\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\right|}{2} = \frac{n^{1/2-\tau} r_n \lambda_n}{2}\frac{\left|\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\right|}{r_n \lambda_n} \to_p \infty. \tag{A.57}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Now, (A.52), (A.56), and (A.57) imply that $\Pr\left(\widehat{\beta}_{n,j}=0\right)\to 1$ as $n\to\infty$ for any $j\in\mathcal{S}_\beta^c$. ■

Proof of Lemma 4.3. From the inequality (A.44) in the proof of Lemma 4.1, we have

$$V_{0,n}\left(\widehat{\theta}_n,\widehat{\beta}_n\right)\geq c\left\|G_k(\widehat{\theta}_n)-\widehat{\beta}_n\right\|^2, \tag{A.58}$$

where c is a generic positive constant, and

$$\left\|G_k(\widehat{\theta}_n)-\widehat{\beta}_n\right\|^2=\left\|G_{d_{\beta_-}}(\widehat{\theta}_n)-\widehat{\beta}_{n,-}\right\|^2+\left\|G_{d_{\beta_+}}(\widehat{\theta}_n)-\widehat{\beta}_{n,+}\right\|^2. \tag{A.59}$$

By Lemma 4.2, we know that $\Pr\left(\widehat{\beta}_{n,-}=0\right)\to 1$ as $n\to\infty$, which implies that

$$\begin{aligned}\Pr\left(\left\|n\widehat{\beta}_{n,-}\right\|>M\right)&\leq\Pr\left(\left\|n\widehat{\beta}_{n,-}\right\|>M,\widehat{\beta}_{n,-}=0\right)+\Pr\left(\widehat{\beta}_{n,-}\neq 0\right)\\&=\Pr\left(\widehat{\beta}_{n,-}\neq 0\right)\to 0,\text{ as } n\to\infty\end{aligned} \tag{A.60}$$

for any $M>0$. From (A.60), we get $\widehat{\beta}_{n,-}=o_p(n^{-1})$, which combined with Assumptions 4.1(iv), 4.3, and the equation (A.59) implies that

$$\begin{aligned}\left\|G_k(\widehat{\theta}_n)-\widehat{\beta}_n\right\|^2&=\left\|G_{d_{\beta_-}}(\widehat{\theta}_n)\right\|^2+\left\|G_{d_{\beta_+}}(\widehat{\theta}_n)-\widehat{\beta}_{n,+}\right\|^2+o_p\left(n^{-1}\right)\\&=(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}})'\left[\frac{\partial G_k(\widetilde{\theta}_n)}{\partial\alpha_{\mathcal{S}}'}\right]\left[\frac{\partial G_k(\widetilde{\theta}_n)}{\partial\alpha_{\mathcal{S}}'}\right]'(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}})+o_p\left(n^{-1}\right)\\&=(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}})'\left[\frac{\partial G_k(\theta_o)}{\partial\alpha_{\mathcal{S}}'}\right]\left[\frac{\partial G_k(\theta_o)}{\partial\alpha_{\mathcal{S}}'}\right]'(\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}})\\&\quad+\left\|\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right\|^2o_p(1)+o_p\left(n^{-1}\right)\\&\geq\left[c+o_p(1)\right]\left\|\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right\|^2+o_p\left(n^{-1}\right),\end{aligned} \tag{A.61}$$

where

$$\frac{\partial G_k(\theta_o)}{\partial\alpha_{\mathcal{S}}'}=\begin{pmatrix}\frac{\partial G_{d_{\beta_-}}(\theta_o)}{\partial\theta'} & 0\\ \frac{\partial G_{d_{\beta_+}}(\theta_o)}{\partial\theta'} & -I_{k-k_o}\end{pmatrix}, \tag{A.62}$$

and the last inequality in (A.61) is by the fact that $\frac{\partial G_k(\theta_o)}{\partial\alpha_{\mathcal{S}}'}$ has full rank. Combining the results in (A.58) and (A.61), we obtain

$$V_{0,n}(\widehat{\theta}_n,\widehat{\beta}_n)\geq\left[c+o_p(1)\right]||\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}||^2+o_p\left(n^{-1}\right), \tag{A.63}$$

which together with (A.43) and (A.46) in the proof of Lemma 4.1 implies that

$$\left[c+o_p(1)\right]||\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}||^2-2\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|\left\|\widehat{\alpha}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right\|\leq 4R_n+o_p\left(n^{-1}\right). \tag{A.64}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

From the above inequality, Assumption 4.1.(iii), and similar arguments in showing (A.24) and (A.25), we deduce that

$$||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}|| = O_p\left(\max\left\{\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|, n^{-1/2}\right\}\right). \tag{A.65}$$

By Assumption 4.2.(i), there is $n^{1/2}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = o_p(1)$, which combined with the equation (A.65) gives us $||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}|| = O_p\left(n^{-1/2}\right)$. ■

Proof of Corollary 4.1. For any compact subset K in $R^{d_\theta+k-k_o}$, we denote any element $u_\mathcal{S} \in K$ as $u'_\mathcal{S} = (u'_\theta, u'_{\beta_+})$, where u_θ are the first d_θ elements in $u_\mathcal{S}$, and u_{β_+} are the last $k-k_o$ elements in $u_\mathcal{S}$. Define

$$\begin{aligned} V_{3,n}(u_\mathcal{S}) \equiv & \left[n^{-1/2}\sum_{i=1}^n \rho^s\left(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}\right)\right]' W_n \left[n^{-1/2}\sum_{i=1}^n \rho^s\left(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}\right)\right] \\ & - \left[n^{-1/2}\sum_{i=1}^n \rho(Z_i, \alpha_o)\right]' W_n \left[n^{-1/2}\sum_{i=1}^n \rho(Z_i, \alpha_o)\right] \\ & + n\sum_{j\in\mathcal{S}_\beta}\left[\widehat{P}_{\lambda_n}\left(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}}\right) - \widehat{P}_{\lambda_n}(\beta_{o,j})\right] \\ \equiv & V^*_{3,n}(u_\mathcal{S}) + n\sum_{j\in\mathcal{S}_\beta}\left[\widehat{P}_{\lambda_n}\left(\beta_{o,j} + n^{-1/2}u_{\beta_+,j}\right) - \widehat{P}_{\lambda_n}(\beta_{o,j})\right], \end{aligned} \tag{A.66}$$

where $\rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) \equiv \rho(Z_i, \theta_o + \frac{u_\theta}{\sqrt{n}}, \beta_{o,+} + \frac{u_{\beta_+}}{\sqrt{n}}, \beta_{o,-})$. From Lemma 4.2, we know that $\widehat{\beta}_{n,-} = 0$ w.p.a.1. Hence, $\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right)$ is the minimizer of $V_{3,n}(u_\mathcal{S})$ w.p.a.1.

Using similar arguments in the proof of Lemma A.1, one can deduce that

$$v_n\left\{\left[\rho^s\left(Z, \alpha_{o,\mathcal{S}} + n^{-1/2}u_\mathcal{S}\right) - \rho\left(Z, \alpha_o\right)\right]\right\} = o_p(1) \tag{A.67}$$

uniformly over $u_\mathcal{S} \in K$. By Assumption 4.1(iv) and the compactness of K, we have

$$\mathbb{E}\left[g_{q,n}\left(Z, \theta_o + \frac{u_\theta}{\sqrt{n}}\right)\right] = n^{-\tau-1/2}\left(\frac{\partial G_{q,n}(\theta_o)}{\partial\theta'}, \quad 0\right)u_\mathcal{S} + o\left(n^{-\tau-1/2}\right) = O\left(n^{-\tau-1/2}\right). \tag{A.68}$$

Recall that $g'_{k,n}(Z,\theta) = \left[g'_{d_{\beta_-,n}}(Z,\theta),\ g'_{d_{\beta_+,n}}(Z,\theta)\right]$. Hence, by Assumption 4.1(iv) and the compactness of K,

$$\mathbb{E}\left[\begin{array}{c} g_{d_{\beta_-,n}}\left(Z, \theta_o + \frac{u_\theta}{\sqrt{n}}\right) \\ g_{d_{\beta_+,n}}\left(Z, \theta_o + \frac{u_\theta}{\sqrt{n}}\right) - \left(\beta_{o,+} + \frac{u_{\beta_+}}{\sqrt{n}}\right) \end{array}\right] = n^{-1/2}\frac{\partial G_k(\theta_o)}{\partial\alpha'_\mathcal{S}}u_\mathcal{S} + o\left(n^{-1/2}\right). \tag{A.69}$$

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Denote $\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'} = \left[\mathbf{0}, \left(\frac{\partial G_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right)'\right]'$, from the results in (A.68) and (A.69), we obtain

$$n^{1/2}\mathbb{E}\left[\rho^s\left(Z, \alpha_{o,\mathcal{S}} + n^{-1/2}u_{\mathcal{S}}\right) - \rho(Z, \alpha_o)\right] = \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}u_{\mathcal{S}} + o(1). \tag{A.70}$$

Thus, (A.67) and (A.70) imply that uniformly over $u_{\mathcal{S}} \in K$, there is

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\rho^s\left(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}\right) &= \upsilon_n\left\{\rho^s\left(Z, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}\right) - \rho(Z, \alpha_o)\right\} + \upsilon_n\left[\rho(Z, \alpha_o)\right] \\ &\quad + n^{1/2}\left\{\mathbb{E}\left[\rho^s\left(Z, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[\rho(Z, \alpha_o)\right]\right\} \\ &= \upsilon_n\left[\rho(Z, \alpha_o)\right] + \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}u_{\mathcal{S}} + o_p(1), \end{aligned} \tag{A.71}$$

which together with the compactness of K and Assumptions 3.4(i) and 4.1(iii) further implies that

$$V_{3,n}^*(u_{\mathcal{S}}) = \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}u_{\mathcal{S}} + 2\upsilon_n\left[\rho(Z, \alpha_o)\right]\right]' W_o\left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}u_{\mathcal{S}}\right] + o_p(1) \tag{A.72}$$

uniformly over K. Note that there is

$$\left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right]' W_o\left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right] = \left[\frac{\partial G_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right]' W_{o,kk}\left[\frac{\partial G_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right] = M_+, \tag{A.73}$$

which combined with (A.72) implies

$$V_{3,n}^*(u_{\mathcal{S}}) = u_{\mathcal{S}}'M_+u_{\mathcal{S}} + 2u_{\mathcal{S}}'\left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right]' W_o\upsilon_n\left[\rho(Z, \alpha_o)\right] + o_p(1). \tag{A.74}$$

Using the same arguments in the proof of Theorem 3.4, we can show that under Assumptions 4.1(v) and 4.2(i) and the compactness of K, there is

$$\begin{aligned} &n\left|\widehat{P}_{\lambda_n}\left(\beta_{o,j} + n^{-\frac{1}{2}}u_{\beta_+,j}\right) - \widehat{P}_{\lambda_n}(\beta_{o,j})\right| \\ &\quad \le \left|\sqrt{n}\widehat{P}_{\lambda_n}'(\beta_{o,j})u_{\beta_+,j}\right| + \max_{j\in\mathcal{S}_\beta}\sup_{\beta_j\in N_{\beta_{o,j}}}\left|\widehat{P}_{\lambda_n}''(\beta_j)\right|u_{\beta_+,j}^2 \to 0 \end{aligned} \tag{A.75}$$

uniformly over $u_{\beta_+,j}$ for any $j \in \mathcal{S}_\beta$.

From the results in (A.66), (A.74), (A.75), the triangle inequality, and Assumption 3.4(ii), we deduce that

$$V_{3,n}(u_{\mathcal{S}}) \to_d V_3(u_{\mathcal{S}}) \equiv u_{\mathcal{S}}'M_+u_{\mathcal{S}} + 2u_{\mathcal{S}}'\left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right]' W_o\Psi(\theta_o), \tag{A.76}$$

uniformly over K. It is clear that $V_3(u_{\mathcal{S}})$ is uniquely minimized at

$$u_{\mathcal{S}} = -M_+^{-1}\left[\mathbf{0}, \left(\frac{\partial G_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right)'\right] W_o\Psi(\theta_o), \tag{A.77}$$

and $\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right)$ is asymptotically tight. Now the result in (4.12) follows by ACMT. ■

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

APPENDIX B

In this Appendix we check the general conditions imposed on $\widehat{P}_{\lambda_n}(\cdot)$ to derive the oracle properties of the GMM shrinkage estimators with the bridge, adaptive Lasso, and SCAD penalty functions, respectively.

COROLLARY B.1. *Suppose that $\lambda_n = o(1)$; then the bridge and SCAD penalty functions satisfy Assumption 3.1(iv). If we further assume that $\widehat{\beta}_n$ is a consistent estimator of β_o, then the adaptive Lasso penalty function also satisfies Assumption 3.1(iv).*

Proof of Corollary B.1. First, if $\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma$, then $\widehat{P}_{\lambda_n}(\beta)$ is nonnegative and $\widehat{P}_{\lambda_n}(0) = 0$. Moreover, by $\lambda_n = o(1)$,

$$\lambda_n \left|\beta_{o,j}\right|^\gamma = o(1) \quad \text{for all } j \in \mathcal{S}_\beta.$$

If $\widehat{P}_{\lambda_n}(\beta)$ is the SCAD penalty function, then $\widehat{P}_{\lambda_n}(\beta)$ is nonnegative and $\widehat{P}_{\lambda_n}(0) = 0$. Moreover, by $\lambda_n = o(1)$, when n is sufficiently large we have

$$\left|\widehat{P}_{\lambda_n}(\beta_{o,j})\right| \leq \frac{(a+1)\lambda_n^2}{2} = o(1) \quad \text{for all } j \in \mathcal{S}_\beta.$$

Finally, for the adaptive Lasso penalty $\widehat{P}_{\lambda_n}(\beta) = \lambda_n \widehat{w}_\beta |\beta|$, it is clear that $\widehat{P}_{\lambda_n}(\beta)$ is nonnegative and $\widehat{P}_{\lambda_n}(0) = 0$. By the consistency of $\widehat{\beta}_{1st}$, $\lambda_n = o(1)$ and the Slutsky theorem, we can deduce that

$$\lambda_n \widehat{w}_{\beta_j} |\beta_{o,j}| \to_p 0 \quad \text{for all } j \in \mathcal{S}_\beta. \qquad \blacksquare$$

Remark B.2. Compared with the results in Knight and Fu (2000) and Caner (2009), where bridge penalty is used, Corollary B.1 imposes the same condition on λ_n to derive the consistency of the shrinkage estimator. When the penalty function is SCAD, $\lambda_n = o(1)$ is also the sufficient condition in Fan and Li (2001) to derive the consistency of the shrinkage estimator. In the adaptive Lasso case, Zou (2006) derives the limit distribution of the adaptive Lasso LS estimator under the conditions $\lambda_n = o(1)$ and $\lambda_n n^{1+\omega/2} \to \infty$. As we will impose the same conditions on λ_n to derive the limit distribution of the GMM shrinkage estimator based on the adaptive Lasso penalty, our condition imposed on λ_n to derive the consistency is not stronger than that of Zou.

COROLLARY B.3. *Suppose that $\lambda_n = o(1)$; then the bridge and SCAD penalty functions satisfy Assumption 3.2.(iv). If we further assume that $\widehat{\beta}_{1st}$ is a consistent estimator of β_o, then the adaptive Lasso penalty function also satisfies Assumption 3.2(iv).*

Proof of Corollary B.3. First, if $\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma$, then for any $\beta > 0$ there is

$$\widehat{P}'_{\lambda_n}(\beta) = \gamma\, \lambda_n \beta^{\gamma-1} \quad \text{and} \quad \widehat{P}''_{\lambda_n}(\beta) = \gamma\,(\gamma-1)\lambda_n \beta^{\gamma-2}.$$

Similarly, we can show that $\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma$ is continuously twice differentiable at any $\beta < 0$. As $\lambda_n = o(1)$, we have

$$\left|\widehat{P}''_{\lambda_n}(\beta_{o,j})\right| = \gamma\,(1-\gamma)\lambda_n \left|\beta_{o,j}\right|^{\gamma-2} = o(1)$$

for any $j \in \mathcal{S}_\beta$.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

Next, if $\widehat{P}_{\lambda_n}(\beta) = \lambda_n \widehat{w}_\beta |\beta|$, then for any $\beta \neq 0$, we have

$$\widehat{P}'_{\lambda_n}(\beta) = \lambda_n \widehat{w}_{\beta,j} \operatorname{sgn}(\beta) \quad \text{and} \quad \widehat{P}''_{\lambda_n}(\beta) = 0.$$

So Assumption 3.2.(iv) is trivially satisfied.

Finally, if $\widehat{P}_{\lambda_n}(\beta)$ is the SCAD penalty function, then $\widehat{P}_{\lambda_n}(\beta) = \frac{(a+1)\lambda_n^2}{2}$ for any β in the local neighborhood of $\beta_{o,j}$ ($j \in \mathcal{S}_\beta$) when n is sufficiently large. Hence, we can deduce that $\widehat{P}_{\lambda_n}(\beta)$ is continuously twice differentiable at $\beta_{o,j}$ ($j \in \mathcal{S}_\beta$) with $\widehat{P}''_{\lambda_n}(\beta_{o,j}) = 0$. ■

COROLLARY B.4. *(i) If $\lambda_n n^{1/2} = o(1)$ and $\lambda_n n^{1-\gamma/2} \to \infty$, then the bridge penalty function satisfies Assumption 3.3; (ii) if $\lambda_n n^{1/2} = o(1)$, $\lambda_n n^{1/2(1+\omega)} \to \infty$, and $\widehat{\beta}_{1st}$ is $\sqrt{n}$ consistent, then the adaptive Lasso penalty function satisfies Assumption 3.3; (iii) if $\lambda_n = o(1)$ and $\lambda_n n^{1/2} \to \infty$, then the SCAD penalty function satisfies Assumption 3.3.*

Proof of Corollary B.4. For the bridge penalty, under the condition $\lambda_n n^{1/2} = o(1)$ it is clear that

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| = \max_{j \in \mathcal{S}_\beta} \gamma\, n^{1/2} \lambda_n \left| \beta_{o,j} \right|^{\gamma - 1} = o(1).$$

If $\beta_{j,n} = O_p\left(n^{-1/2}\right)$ and $\beta_{j,n} \neq 0$ a.e., then we can let $r_n = n^{1/2-\gamma/2}$ such that

$$\liminf_{n\to\infty} \frac{\left| \widehat{P}'_{\lambda_n}(\beta_{j,n}) \right|}{r_n \lambda_n} = \liminf_{n\to\infty} \gamma \left| n^{\frac{1}{2}} \beta_{j,n} \right|^{\gamma - 1} > 0, \quad a.e.$$

and $n^{1/2} \lambda_n r_n = \lambda_n n^{1-\gamma/2} \to \infty$.

For the adaptive Lasso penalty, by the consistency of $\widehat{\beta}_{1st}$, $n^{1/2}\lambda_n = o(1)$ and the Slutsky theorem, we have

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| = \max_{j \in \mathcal{S}_\beta} n^{1/2} \lambda_n \widehat{w}_{\beta,j} = o_p(1).$$

Let $r_n = n^{\omega/2}$, then $n^{1/2}\lambda_n r_n = \lambda_n n^{1/2(1+\omega)} \to \infty$. For any $\beta_{j,n} = O_p\left(n^{-1/2}\right)$ and $\beta_{j,n} \neq 0$ a.e., by the root-n consistency of $\widehat{\beta}_{j,1st}$, we have

$$\liminf_{n\to\infty} \frac{\left| \widehat{P}'_{\lambda_n}(\beta_{j,n}) \right|}{r_n \lambda_n} = \liminf_{n\to\infty} \left| n^{1/2} \widehat{\beta}_{j,1st} \right|^{-\omega} > 0, \quad a.e.$$

for any $j \in \mathcal{S}_\beta^c$.

Finally, if $\widehat{P}_{\lambda_n}(\cdot)$ is the SCAD penalty function, then by the definition of $\widehat{P}'_{\lambda_n}(\cdot)$, it is easy to see that when n is sufficiently large,

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| = 0.$$

As $\beta_{j,n} = O_p\left(n^{-1/2}\right)$ and $\lambda_n n^{1/2} \to \infty$, we have $|n^{1/2}\beta_{j,n}| \leq n^{1/2}\lambda_n$ w.p.a.1. Hence, by definition, we have $\left| \widehat{P}'_{\lambda_n}(\beta_{j,n}) \right| = \lambda_n$, w.p.a.1. Let $r_n = 1$, there is $n^{1/2}\lambda_n r_n = n^{1/2}\lambda_n \to \infty$ and

$$\liminf_{n\to\infty} \frac{\left| \widehat{P}'_{\lambda_n}(\beta_{j,n}) \right|}{r_n \lambda_n} = 1, \quad a.e.$$ ■

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press

We next check Assumptions 4.1(v) and 4.2 using the bridge, adaptive Lasso, and SCAD penalty functions. For the bridge penalty, It is clear that if $n^{1/2+\tau}\lambda_n = o(1)$, then Assumptions 4.1(v) and 4.2(i) are satisfied. Let $r_n = n^{(1/2-\tau)(1-\gamma)}$; then for any $\beta_{j,n} = O_p(n^{-1/2+\tau})$ and $\beta_{j,n} \neq 0$ a.e.,

$$\liminf_{n\to\infty} \frac{\left|\widehat{P}'_{\lambda_n}(\beta_{j,n})\right|}{r_n\lambda_n} = \liminf_{n\to\infty} \gamma \left|n^{1/2-\tau}\beta_{j,n}\right|^{\gamma-1} > 0, \text{ a.e.}$$

Hence Assumption 4.2.(ii) is also satisfied if

$$n^{1/2-\tau}\lambda_n r_n = n^{1/2+\tau}\lambda_n \times n^{(1/2-\tau)(1-\gamma)-2\tau} \to \infty.$$

The above condition and $n^{1/2+\tau}\lambda_n = o(1)$ require that $(\frac{1}{2}-\tau)(1-\gamma)-2\tau > 0$, which means $\gamma < \frac{1-6\tau}{1-2\tau}$. It is clear that when $\tau \geq \frac{1}{3}$, then there may be no such sequence λ_n that makes Assumptions 4.1(v) and 4.2 hold simultaneously.

For the adaptive Lasso penalty, it is tedious to show that $\widehat{\beta}_{j,1st} = O_p(n^{-1/2+\tau})$ for any $j \in \mathcal{S}_\beta^c$. If $n^{1/2+\tau}\lambda_n = o(1)$, then Assumptions 4.1(v) and 4.2(i) are satisfied. Let $r_n = n^{(1/2-\tau)\omega}$; then for any $\beta_{j,n} = O_p(n^{-1/2+\tau})$ and $\beta_{j,n} \neq 0$ a.e., there is

$$\liminf_{n\to\infty} \frac{\left|\widehat{P}'_{\lambda_n}(\beta_{j,n})\right|}{r_n\lambda_n} = \liminf_{n\to\infty} \gamma \left|n^{1/2-\tau}\widehat{\beta}_{j,1st}\right|^{-\omega} > 0, \quad a.e.$$

for any $j \in \mathcal{S}_\beta^c$. Hence, if we take $\omega > \frac{4\tau}{1-2\tau}$, then there exists λ_n such that

$$n^{1/2-\tau}\lambda_n r_n = n^{1/2+\tau}\lambda_n \times n^{\omega(1/2-\tau)-2\tau} \to \infty.$$

Finally, for the SCAD penalty, Assumptions 4.1(v) and 4.2(i) require $n^\tau\lambda_n = o(1)$. Note that for any $\beta_n = O_p(n^{-1/2+\tau})$ and $\beta_n \neq 0$,

$$\frac{\widehat{P}'_{\lambda_n}(\beta_n)}{\lambda_n} = I(\beta_n \leq \lambda_n) + \frac{(a\lambda_n - \beta_n)_+}{(a-1)\lambda_n} I(\beta_n > \lambda_n).$$

Hence, $r_n = a$ and to get

$$n^{1/2-\tau}\lambda_n r_n = an^\tau\lambda_n \times n^{1/2-2\tau} \to \infty,$$

we need $\tau < \frac{1}{4}$. It is clear that when $\tau > \frac{1}{4}$, there may be no such sequence λ_n that makes Assumptions 4.1(v) and 4.2 hold simultaneously.

https://doi.org/10.1017/S0266466612000783 Published online by Cambridge University Press