Ridge-regularization for moment-based estimation in high-dimensional settings

[Click here for the latest version]

Marine Carrasco[*] and Eugène Dettaa[†]

November 6, 2024

Abstract

This paper provides a user-friendly approach for moment-based estimation in high-dimensional contexts where many moments/instruments are available. Several economic applications involve many instruments. In this paper, we address the issue of efficient estimation in such frameworks. In fact, using many moment conditions can improve the efficiency of GMM-type estimators, as is well known, but can also lead to heavily biased estimates due to strong overidentification. We consider a specific setting where the large set of moments is derived from a single conditional moment restriction. The benchmark estimator we consider is the continuous updating GMM estimator (CUE) due to its relatively low bias under many moment conditions. We introduce a Ridge-type regularized version of CUE (RRCUE) to address the singularity problem of the weighting matrix under many moments. We show that the RRCUE estimator is consistent, asymptotically normal, and reaches the semi-parametric efficiency bound under an asymptotic regime where the regularization parameter goes to zero at a certain rate. We propose a data-driven approach for selecting the optimal regularization parameter based on cross-validation criteria. We evaluate the performance of the RRCUE through Monte Carlo simulations. Our results reveal that regularization reduces the dispersion problem of CUE and improves efficiency, although it introduces some bias that remains relatively low with a moderately large number of moments. In the specific linear instrumental variables framework, our estimator is shown to be competitive with some state-of-the-art estimators in the field. We apply our method to revisit the Hall and Jones (1999) empirical application, which aims to estimate the effect of the quality of institutions and government policies, the so-called social infrastructure, on output per worker. Our empirical results are consistent with simulation results, providing estimates with better precision.

Keywords: Conditional moment restriction, many moments/instruments, continuous updating estimator, efficient estimation, regularization, semiparametric efficiency bound.

JEL classification: C13, C26, C36.

[*] University of Montreal, CIREQ, CIRANO, Canada. E-mail: marine.carrasco@umontreal.ca

[†] University of Montreal, CIREQ, Canada. E-mail: eugene.delacroix.dettaa.mboudjiho@umontreal.ca

1 Introduction

The issue of efficient estimation of models involving many instruments/moments is an important part of the econometric literature. This paper considers the efficient estimation of a parameter of interest defined by a single conditional moment restriction. This framework is compatible with several applications in microeconomic data where a large set of valid instruments is available. Examples include an influential instrumental variable study that estimates the economic return to schooling. Angrist and Frandsen (2022) revisiting the Angrist and Krueger (1991)'s use up to $1,530$ instruments for schooling by interacting quarter of birth dummies, year of birth dummies, and state of birth dummies. Altonji et al. (2013) examine a joint model of earnings, employment, job changes, wage rates, and work hours over a career with a full specification of $2,429$ moments. Eaton et al. (2011) explore the sales of French manufacturing firms in 113 destination countries with $1,360$ moments. Han et al. (2005) investigate the cost efficiency of the Spanish saving banks in a time-varying coefficient model with 872 moments.

Efficient estimation of models with conditional moment restrictions poses a challenging problem. Indeed, many unconditional moment restrictions can be obtained from nonlinear transformations of an exogenous variable or by using interactions between various exogenous variables. Selecting a small number of moments may lead to a loss of efficiency in finite dimension since a single conditional moment restriction is equivalent to an infinite countable sequence of unconditional moments under certain conditions (see, Chamberlain, 1987; Donald et al., 2003). However, all the information in the conditional moment restriction will eventually be accounted for by allowing the number of unconditional moments to grow with the sample size, thus achieving asymptotic efficiency (Donald et al., 2003). This suggests that, rather than using the infinite set of moments available, one can use a reasonably large number of them (even larger than the sample size) to gain efficiency in applications.

GMM-type estimation with a large number of moments presents certain challenges. Firstly, while it can improve asymptotic efficiency, using an excessive number of moments can deteriorate the finite sample properties of the GMM estimator. Specifically, the standard two-step GMM estimator can exhibit significant bias and/or inaccuracy, even in applied work with a large number of observations (see, Hahn and Hausman, 2003; Hansen et al., 2008; Newey and Smith, 2004; Newey and Windmeijer, 2009). This trade-off between variance and bias is known in the literature as the "many moments problem". The benchmark estimator we consider is the Hansen et al. (1996)'s continuous updating estimator (CUE), due to its relatively low bias when dealing with a large number of moments. Secondly, the weighting matrix used in the CUE's objective function may become singular when dealing with a substantial number of moment conditions. Consequently, the CUE estimator may become infeasible or exhibit poor finite sample properties. Lastly, the CUE estimator is not guaranteed to have finite moments of any order, resulting in the undesirable property of significant dispersion in the estimates (see, Guggenberger, 2005; Hausman et al., 2011, 2012). This property is referred to in the literature as the "no-moments problem" of CUE.

We introduce ridge-type regularization in the weighting matrix to address the singularity problem. The resulting estimator is referred to as the regularized CUE (RRCUE). We demonstrate that the RRCUE estimator can be derived through L2 penalization of the generalized empirical likelihood (GEL) representation of CUE. We establish that RRCUE is consistent and asymptotically normal, subject to certain restrictions on the convergence rate of the regularization parameter. Furthermore, its asymptotic variance achieves the Chamberlain (1987)'s semiparametric efficiency bound. To assess the benefits of regularization, we conduct a Monte Carlo simulation. Our findings demonstrate that regularization has the potential to alleviate the no-moments problem observed in CUE by reducing the dispersion of the CUE estimator. Additionally, regularization aids in improving the efficiency of the CUE estimatorin finite sample, albeit at the cost of introducing some bias. However, this bias remains smaller than the GMM overidentification bias in almost all settings we consider.

This article contributes to the extensive literature on many instruments/moments. This literature can be divided into two parts. The first part, which dates back to Bekker (1994), focuses on an asymptotic framework where the number of moments, denoted by K, grows with the sample size, denoted by n, but remains relatively small (see, e.g., Chao and Swanson, 2005; Donald et al., 2003, 2009; Donald and Newey, 2001; Hansen et al., 2008; Hausman et al., 2012; Newey and Windmeijer, 2009, among others). Specifically, Hansen et al. (2008) derived asymptotics properties of the limited information maximum likelihood (LIML) and Fuller (1977) estimators under a variety of many instrument asymptotics, including the many instrument sequence of Bekker (1994) and the many weak instruments sequence of Chao and Swanson (2005) and Stock and Yogo (2005). Newey and Windmeijer (2009) derived properties of CUE under many weak moment conditions as those of Hansen et al. (2008). Donald et al. (2003) demonstrated that the generalized empirical likelihood (GEL) class, including CUE, offers consistent and asymptotically normal estimators for models with conditional moment restrictions, which attain the semiparametric efficiency bound under a stringent condition on the growth rate of K relative to n. Donald and Newey (2001) and Donald et al. (2009) proposed a method to select the optimal number of moments by minimizing an approximate mean square error derived from higher-order expansion. This paper falls into the second strand of literature that does not require selecting moments. Papers in this strand allow for the sample size to grow faster than the number of moments (see, for example, Belloni et al., 2012; Carrasco, 2012; Carrasco and Tchuente, 2015; Shi, 2016, among others). In particular, Belloni et al. (2012) recommended using LASSO in the first step to construct the optimal instrument when assuming the approximate sparsity of the first stage equation. Carrasco (2012) proposed three regularized estimators to improve the small sample properties of the standard two-stage least squares (2SLS) estimator when dealing with a large number of instruments. Carrasco and Tchuente (2015) extended this regularization approach to the limited information maximum likelihood (LIML) estimator and demonstrated that the regularized LIML estimator has finite first moments provided that the sample size is sufficiently large. This paper can be seen as an extension of Carrasco and Tchuente (2015) from efficient estimation of linear homoscedastic models using LIML to nonlinear heteroscedastic models using the continuous updating estimator (CUE). Other papers in the literature address specific issues such as heteroscedasticity and the no-moments problem. Hausman et al. (2012) addressed the problem of many instruments in heteroskedastic data and recommended using a jackknife version of the Fuller (1977) estimator in applications that align with this framework. Hansen and Kozbur (2014) proposed a ridge-regularized version of the jackknife instrumental variable estimator (JIVE) that is robust to heteroscedasticity in the presence of many instruments. Hausman et al. (2011) proposed a modification of the continuous updating estimator (CUE) to address the no-moments problem and improve the finite sample properties of the standard CUE in time series settings with many weak moment conditions. In a more recent work, Angrist and Frandsen (2022) investigated the performance of machine learning (ML) for instrument selection. They argued that the optimal way to leverage the ML toolkit is to combine it with a sample splitting procedure.

The rest of the paper is structured as follows. Section 2 introduces the framework. Section 3 presents the Ridge regularized continuous updating estimator (RRCUE). In Section 4, we derive first-order asymptotic properties of the RRCUE. In Section 5, we suggest a data-driven procedure for selecting the optimal regularization parameter using cross-validation. We conduct a large-scale Monte Carlo experiment to evaluate the benefits of regularization in Section 6. Section 7 applies our method to estimate the impact of institutions and government policies on productivity. Section 8 concludes our findings, while technical proofs and additional lemmas are provided in the Appendix.

2 The framework and moment restrictions

We consider an environment where there are many unrestricted moment conditions generated by a single conditional moment restriction (CMR) like Chamberlain (1987) and Donald et al. (2003, 2009). To describe this setting let w denote a single observation from an i.i.d. sequence $(w_1, w_2, \ldots)$, β a $p \times 1$ parameter vector, and $\rho(w, \beta)$ a scalar that can be seen as a residual. z is a subvector acting as conditioning variables such that for a value β_0 of the parameters

$$E\left[\rho(w, \beta_0)|z\right] = 0, \tag{1}$$

where $E\left[\cdot\right]$ is the expectation taken with respect to the distribution of w.

We rely on GMM-type estimator to address the issue of efficient estimation of the parameter β_0. We need a vector of unrestricted moment conditions for this purpose. It is well known that that a conditional moment restriction as in Eq. (1) is equivalent to a countable number of unconditional moment restrictions under certain conditions like the one in Assumption 1(b) below, see e.g., Chamberlain (1987). Following Donald et al. (2003, DIN03 hereafter), our unconditional moment conditions are based on splines or other approximating functions like power series. For each positive integer K, let $q^K(z) = (q_{1K}(z), \ldots, q_{KK}(z))'$ be a $K \times 1$ vector of approxamating functions. Under Assumption 1(b) below DIN03 showed that the conditional moment restriction of Eq. (1) is equivalent to a sequence of unconditional moment restrictions of the following form

$$E\left[g(x, \beta_0)\right] = 0, \tag{2}$$

where $g(x, \beta) = q^K(z)\rho(w, \beta)$ and $x = (w', z')'$. The immediate consequence of this result is that an efficient estimation of β_0 under CMR in Eq. (1) can be obtained from the sequence of unconditional moments of Eq. (2) by letting K grows with the sample size n. Indeed, all the information in the CMR will be eventually accounted for by letting K goes to infinity with n. For notational convenience, we omit the K superscript on $g(w, \beta)$ which denotes its dependence on the number K of approximating functions.

In the literature, there are sevaral choices of the approximating functions, $q_{kK}(z)$, including power series, splines and Fourier series. In this article, we will focus on the first two. They both have faster approximation rates for smoother functions (up to the order of the spline or the power serie). Unlike power series, spline approximations are not severely affected by singularities (e.g. discontinuities) in the function being approximated. To describe $q^K(z)$ in detail, consider the simple case where z is a scalar. In this case the vector of power series approximating functions is given by

$$q^K(z) = \left(1, z, z^2, \dots, z^{K-1}\right)'. \tag{3}$$

For splines, let s be a positive scalar giving the order of the spline. Let $t_1, \dots, t_{K-s-1}$ denote knots and let $\xi(z) = z1_{\{z>0\}}$, where 1_A denotes the indicator function for the event A. Then a vector of spline approximating functions is given by

$$q^K(z) = (1, z, \dots, z^s, \xi\,(z-t_1)^s, \dots, \xi\,(z-t_{K-s-1})^s)'. \tag{4}$$

The most common specification is $s = 3$. In practice, it is recommended to choose the knots t_j in the observed data range of z, see e.g. Donald et al. (2003). We impose the following conditions on the sequence $q^K(z)$ and the distribution of z. Let $\mathscr{Z}$ denotes the support of z.

Assumption 1. *(a) For each K there is a constant $\zeta(K) \geq \sqrt{K}$ and a positive constant C such that: $E\left[q^K(z)'q^K(z)\right]$ is finite and $\sup_{z\in\mathscr{Z}} \left\|q^K(z)\right\| \leq C\zeta(K)$, (b) for any $a(z)$ with $E\left[a(z)^2\right] < \infty$ there are $K \times 1$ vectors γ_K such that as $K \longrightarrow \infty$, $E\left[\left\{a(z) - \tilde{q}^K(z)'\gamma_K\right\}^2\right] \longrightarrow 0$, where $\tilde{q}^K(z) = q^K(z)/\zeta(K)$, (c) for each K, $E\left[\tilde{q}^K(z)\tilde{q}^K(z)'\right]$ has only nonzero eigenvalues and there is $\theta \geq 1/2$ such that for any $b(z)$ with $E\left[b(z)^2\right] < \infty$ and for any nonnegative scalar function $U(z)$ bounded away fron zero[1],*

$$\sum_{j=1}^{\infty} \frac{\left(E\left[b(z)\tilde{q}^K(z)\right]'\phi_j\right)^2}{\lambda_j^{2\theta+1}} < \infty, \tag{5}$$

where $\left(\lambda_j, \phi_j : j = 1, 2, \dots K\right)$ are eigenvalues and orthonormal eigenvectors of the $K \times K$ symmetric and positive semidefinite matrix $L = E\left[U(z)\tilde{q}^K(z)\tilde{q}^K(z)'\right]$.

Assumption 1(b) is similar to Assumption 1 of DIN03. DIN03 argue that its specific role is to obtain estimators that achieve the Chamberlain (1987)'s semiparametric effiency bound by ensuring that linear combinations of $q^K(z)$ can approximate certain square integrable function of z. Assumption 1(a) is similar to a normalization of the approximating functions like that adopted by Newey (1997) and Donald et al. (2003, 2009). The bound $\zeta(K)$ plays an important role in the asymptotic theory for GMM and the generalized empirical likelihood (GEL) class of estimators developed by DIN03. DIN03 showed under some mild conditions that Assumption 1(b) is sufficient to obtain the asymptotic efficiency of the continuous updating estimator (CUE), known as

[1] The sum in Eq. (5) is defined by

$$\lim_{K\to\infty} \sum_{j=1}^{K} \frac{\left(E\left[b(z)\tilde{q}^K(z)\right]'\phi_j\right)^2}{\lambda_j^{2\theta+1}}.$$

an element of the GEL class (Newey and Smith, 2004), if the condition $\zeta(K)^2K^2/n \longrightarrow 0$ holds. Such a condition on the growth rate of K restrict the number K of moment conditions that can be used in applications. We have shown that Assumption 1(b) is no longer sufficient to obtain the asymptotic efficiency of the regularized estimator that we will introduce in the next section. An additional condition given by Assumption 1(c) is required. On the one hand, Assumption 1(c) implies that eigenvalues of the matrix are all non-zero for fixed K although they can converge to zero if K grows with the sample size[2]. On the other hand, the condition in Eq. (5) is similar to that used by Carrasco (2012) and Carrasco and Tchuente (2015). This condition is important to obtain asymptotic efficiency of the regularized estimator. More pricisely, as pointed out by Carrasco et al. (2007), this regularity condition will facilitate the calculation of the rate of convergence of the regularization bias. The value of θ in Eq. (5) measures how well the vector of instruments $\tilde{q}^K(z)$ approximates a certain square integrable function, $b(z)$. The larger θ, the better the approximation. For $\theta = 1/2$, condition (5) implies that $E\left[b(z)\tilde{q}^K(z)\right]$ belongs to the range of L (Carrasco et al., 2007).

We show under Assumption 1 and some regularity conditions that a regularization of the second moment of $g(w,\beta)$ allows to free the number of moments K from any constraint as the one imposed by DIN03 to obtain asymptotic efficiency. Indeed, our rates of convergence do no longer depend on K but only on the sample size n and the regularization parameter α. For example we show that the regularized version of CUE is consistent under the asymptotic where K goes to infinity and α goes to zero as the sample size goes to infinity with the following restriction on the convergence rate of α relative to n: $\alpha^{-2}n^{-1/2+1/\gamma} \rightarrow 0$. Asymptotic normality of the regularized estimator requires a stronger restriction, that is, $\alpha^{-5/2}n^{-1/2+1/\gamma} \rightarrow 0$. The parameter $\gamma > 2$ is specified as in Assumption 2 bellow.

An explicit formula for $\zeta(K)$ is available for a number of cases. For example it has been shown under some unrestricted conditions that $\zeta(K) = \sqrt{K}$ for splines and $\zeta(K) = K$ for power series, see e.g., Newey (1997) among others. Under Assumption 1(a), $E\left[q^K(z)'q^K(z)\right] = O(\zeta(K)^2)$ and therefore the second moment matrix of the approximating functions is a trace-class matrix[3] (even for large K) if they are normalized by the bound $\zeta(K)$. We show in this paper that this type of normalization is useful to obtain convergence rates that do not depend on the number K of moment conditions. DIN03 used an additional condition, that is $A \overset{def}{=} E\left[\tilde{q}^K(z)\tilde{q}^K(z)'\right]$ has a smallest eigenvalue bounded away from zero uniformly in K. This restriction limits the number of moments K that can be used in practice and does not make it possible to deal with the cases where the matrix A is neraly singular due to the numerosity of approximating functions. Regularization will allow us to get rid of such a condition by allowing A nearly singular as K grows with the sample size[4].

[2] Which is less restrictive than the condition used by Donald et al. (2003), that is, $E\left[\tilde{q}^K(z)\tilde{q}^K(z)'\right]$ has smallest eigenvalue bounded away from zero uniformly in K.

[3] A matrix is said to be trace-class matrix if its trace is a finite number. If a trace-class matrix is symmetric, then its maximum eigenvalue is bounded above.

[4] The nearly singularity of A refers here to the cases where its minimal eigenvalue, although different from zero, goes to zero as K goes to infinity.

3 Ridge-regularized version of CUE

One of the standard approaches to efficiently estimate the parameter of interest β_0 defined by the conditional moment restriction of Eq. (1) is to use GMM-type estimators based on unconditional moment restrictions as specified in Eq. (2). The role of the approximating functions in this specification of the moment conditions is to make the CMR approximately satisfied in the sample. Asymptotic efficiency is then obtained by letting K grow with n at a certain rate (Donald et al., 2003). To gain efficiency in finite sample, the use of many approximating functions may be necessary. As is well known, one of the costs of using many approximating functions is that the sample counterpart of the second moment matrix of the moment function, $g(w,\beta)$, might be ill-conditioned. Following Carrasco and Florens (2000), Carrasco (2012), and Carrasco and Tchuente (2015) among others, we use regularization to fix this problem.

The benchmark estimator we consider is the continuous updating GMM estimator (CUE) due to its relatively low bias with many moment conditions. Before presenting the regularized CUE, we begin by recalling the standard CUE estimator. We introduce here some notations to ease the presentation of estimators. Let $q_i = q^K(z_i)/\zeta(K)$, $\rho_i(\beta) = \rho(w_i,\beta)$, $g_i(\beta) = q_i\rho_i(\beta)$, $\hat{g}(\beta) = n^{-1}\sum_{i=1}^n g_i(\beta)$, $\widehat{\Omega}(\beta) = n^{-1}\sum_{i=1}^n g_i(\beta)g_i(\beta)' = n^{-1}\sum_{i=1}^n \rho_i(\beta)^2 q_i q_i'$, and $\Omega(\beta) = E\left[\rho_i(\beta)^2 q_i q_i'\right]$. The Hansen et al. (1996)'s CUE uses $\widehat{\Omega}(\beta)^{-1}$ as weighting matrix without replacing β by a first step estimator as it is the case for the Hansen (1982)'s conventional two-step GMM estimator. It is defined by

$$\hat{\beta}_{CUE} = \arg\min_{\beta\in\mathscr{B}} \hat{g}(\beta)'\widehat{\Omega}(\beta)^{-1}\hat{g}(\beta). \tag{6}$$

In the sequel, we will refer to this estimator as the standard CUE.

Note that the standard GMM estimator and the standard CUE are specialized for the cases with $K < n$. They can be infeasible when the number of moment conditions is larger than the sample size, restricting their use in empirical applications. Indeed, when K is larger than n or smaller than n but close to n, the weighting matrix used in the CUE's objective function is singular or nearly singular. To illustrate this fact, assume that the error term $\rho_i(\beta_0)$ is conditional homoskedastic with $E\left[\rho_i(\beta_0)^2|z_i\right] = \sigma^2$. Then, by the law of iterated expectations, $\Omega \overset{def}{=} \Omega(\beta_0) = \sigma^2 E\left[q_i q_i'\right]$. If $q := [q_1,\dots,q_n]'$ then the sample counterpart of Ω, $\widehat{\Omega} := \widehat{\sigma}^2 q'q/n$ is singular when $K > n$. It is natural to think that in general the matrix $\widehat{\Omega}(\beta)$ suffers from this singularity problem when $K > n$. Even when K is smaller than n but large, the naive inverse of $\widehat{\Omega}(\beta)$ will be unstable in the sense that a seemingly innocuous change in the sample moment function may induce a large variation of $\widehat{\Omega}(\beta)^{-1}\hat{g}(\beta)$. A matrix with such a property is said to be ill-conditioned.

A Monte Carlo experiment by Hausman et al. (2012) reveals that using an heteroskedasticity consistent weighting matrix can degrade the finite sample performance of the continuously updated estimators (CUE) with many moments. We suspect that the instability of the inverse of the weighting matrix plays an important role in the deterioration of the finite sample properties of CUE under many moment restrictions. This instability of the naive inverse of $\widehat{\Omega}(\beta)$ for a large K is caused by the fact that its condition number is large. Since the condition number is defined as the ratio of the maximum eigenvalue (λ_{max}) and the minimum eigenvalue (λ_{min}), it is large if λ_{max} is very large or λ_{min} is close to zero. Before regularizing $\widehat{\Omega}(\beta)$ we first normalize the approximating functions $q^K(z_i)$ by the upper bound $\zeta(K)$ so that it becomes a trace-class matrix. Thus, the only source of instability of the inverse of $\widehat{\Omega}(\beta)$ is the fact that its minumum eigenvalue may be close to zero when K is large compared to n. Another gain of this normalization is to obtain convergence rates for the regularized estimator which do not depend on K.

We propose to stabilize the inverse of $\widehat{\Omega}(\beta)$ using Ridge[5] type regularization. This consists in replacing, in the CUE objective function, the naive inverse $\widehat{\Omega}(\beta)^{-1}$ by the regularized inverse $\left(\widehat{\Omega}(\beta)^{\alpha}\right)^{-1} = \left(\widehat{\Omega}(\beta) + \alpha I\right)^{-1}$, where $\alpha > 0$ is the regularization parameter and I is the $K \times K$ identity matrix. The resulting estimator that we refer to as the Ridge regularized CUE (RRCUE) depends on the regularization parameter α and is defined by

$$\hat{\beta} = arg \min_{\beta\in\mathscr{B}} \hat{g}(\beta)' \left(\widehat{\Omega}(\beta)^{\alpha}\right)^{-1} \hat{g}(\beta). \tag{7}$$

The main purpose of this paper is to study properties of RRCUE.

Newey and Smith (2004) showed that the standard CUE is part of a class of estimators introduced by Smith (1997, 2001) called generalized empirical likelihood (GEL) estimators. The GEL representation of CUE facilitates theoretical derivation of asymptotic properties of CUE. To describe GEL let $s(v)$ be a function of a scalar v that is concave on its domain, an open interval $\mathscr{V}$ containing zero with $s_0 = 0$ and $s_1 = s_2 = -1$ where $s_j(v) = \partial^j s(v)/\partial v^j$. Let $\widehat{\Lambda}(\beta) = \{\lambda : \lambda' g_i(\beta) \in \mathscr{V}, i = 1, \ldots, n\}$. The GEL estimator associated with the concave function s is the solution to a saddle point problem

$$\hat{\beta}_{\text{GEL}} = \arg\min_{\beta\in\mathscr{B}} \sup_{\lambda\in\widehat{\Lambda}(\beta)} n^{-1} \sum_{i=1}^{n} s\left(\lambda' g_i(\beta)\right). \tag{8}$$

Newey and Smith (2004) showed that $\hat{\beta}_{CUE} = \hat{\beta}_{\text{GEL}}$ if $s(v)$ is quadratic, e.g., if $s(v) = -v - v^2/2$. The following theorem establishes a similar result for the RRCUE.

Theorem 3.1. *If Assumption 1 (a) is satisfied, then for* $s(v) = -v - v^2/2$,
$\hat{\beta} = \underset{\beta\in\mathscr{B}}{\operatorname{argmin}} \sup_{\lambda\in\widehat{\Lambda}(\beta)} \hat{P}(\beta, \lambda)$ *where* $\hat{P}(\beta, \lambda) = n^{-1} \sum_{i=1}^{n} s\left(\lambda' g_i(\beta)\right) - \dfrac{\alpha}{2}\lambda'\lambda.$

This result shows that the regularized CUE can be obtained alternatively by penalizing the L^2 norm of λ in the GEL criterion. This is an important result that will be useful for deriving asymptotic properties of RRCUE.

Here we give first-order conditions for RRCUE which are useful for deriving first-order asymptotic properties. We need some notations. Let $\hat{\pi}_i$ $(i = 1, \cdots, n)$ denote the empirical probabilities associated with $\hat{\beta}$. They are defined by

$$\hat{\pi}_i = s_1\left(\hat{\lambda}'\hat{g}_i\right) \Big/ \sum_{j=1}^{n} s_1\left(\hat{\lambda}'\hat{g}_j\right) = \frac{1 + \hat{v}_i}{\sum_j \left(1 + \hat{v}_j\right)} \quad (i = 1, \cdots, n). \tag{9}$$

where $\hat{\lambda} = \arg\max_{\lambda\in\hat{\Lambda}(\hat{\beta})} P\left(\hat{\beta}, \lambda\right)$, $\hat{g}_i = g_i\left(\hat{\beta}\right)$, and $\hat{v}_i = \hat{\lambda}'\hat{g}_i$. These empirical probabilities sum to one by construction and satisfy the sample moment condition $\sum_{i=1}^{n} \hat{\pi}_i \hat{g}_i = 0$ when the first

[5] There are other regularization methods. For example, Carrasco (2012) considered three other regularization schemes: Tikhonov, Landweber–Fridman and Principal components.

order conditions for $\hat{\lambda}$ hold. For any function $a(w,\beta)$, these probabilities can be used to form an efficient estimator $\sum_{i=1}^{n}\hat{\pi}_i a\left(w_i,\hat{\beta}\right)$ of $E\left[a\left(w,\beta_0\right)\right]$, as in Newey and Smith (2004). The following result gives first-order conditions for RRCUE.

Theorem 3.2. *If Assumption 1 (a) is satisfied, then the RRCUE first order conditions imply*

$$\left[\sum_{i=1}^{n}\hat{\pi}_i G_i\left(\hat{\beta}\right)\right]'\left[\hat{\Omega}\left(\hat{\beta}\right)+\alpha I\right]^{-1}\hat{g}\left(\hat{\beta}\right)=0, \tag{10}$$

where $G_i(\beta)=\partial g_i(\beta)/\partial\beta'$.

4 First order asymptotic properties of the RRCUE

In this section we establish first order asymptotic properties of the regularized estimator. We show that RRCUE is consistent and asymptotically normal, and achieves the Chamberlain (1987)'s semiparametric efficiency bound under some standard assumptions. We first give some regularity conditions for consistency of RRCUE.

Assumption 2. *The data are i.i.d. and (a) β_0 is unique value of β in $\mathscr{B}$ satisfying $E[\rho(w,\beta)|z]=0$; (b) $\mathscr{B}$ is compact; (c) $E\left[\sup_{\beta\in\mathscr{B}}|\rho(w,\beta)|^2\left|z\right.\right]$ is bounded and there is $\delta_1(w)$ and $r>0$ such that for all $\tilde{\beta},\beta\in\mathscr{B}$, $|\rho(w,\tilde{\beta})-\rho(w,\beta)|\leqslant\delta_1(w)\|\tilde{\beta}-\beta\|^r$ and $E\left[\delta_1(w)^2\right]<\infty$; (d) there are $\delta_2(w)$ and a neighborhood $\mathscr{N}$ of β_0 such that $E\left[\sup_{\beta\in\mathscr{N}}|\rho(w,\beta)|^4\left|z\right.\right]$ is bounded and for all $\beta\in\mathscr{N}$ $|\rho(w;\beta)-\rho\left(w,\beta_0\right)|\leq\delta_2(w)\left\|\beta-\beta_0\right\|$ and $E\left[\delta_2(w)^2|z\right]$ is bounded; (e) $\sigma(z)^2\overset{def}{=}E\left[\rho\left(w,\beta_0\right)^2|z\right]$ is bounded away from zero; (f) there is $\gamma>2$ with $E\left[\sup_{\beta\in\mathscr{B}}|\rho(w,\beta)|^\gamma\right]<\infty$.*

Assumption 2(a) is the minimal identification condition that β_0 is the unique value where the conditional moment restriction is satisfied. The stronger condition that there is a known K such that the unconditional moment restrictions $E\left[q^K(z)\rho(w,\beta)\right]=0$ serve to identify β_0 is not required. As K grows with n, the weak condition in Assumption 2(a) is sufficient to identified β_0 as justified in Lemma 2.1 of DIN03. Assumption 2(b) is the usual compacity assumption. Assumption 2(c) imposes a bounded second conditional moment and Lipschitz condition, that is used to apply the uniform convergence result of Newey (1991). Assumption 2(d) plays an important role in obtaining a convergence rate for the sample second moment matrix $\hat{\Omega}\left(\hat{\beta}\right)$. Assumption 2(f) requires the exitence of slightly higher moments than consistency for GMM, as in Hansen (1982).

Let $\gamma>2$ be as defined in Assumption 2(f). For any $\varepsilon>0$ such that $1/2-1/\gamma-\varepsilon>0$, we obtain the following consistency result:

Theorem 4.1. *If Assumptions 1 and 2 are satisfied, $K\to\infty$, $\alpha\to0$ and $\alpha n^{1/2-1/\gamma-\varepsilon}\to\infty$, then $\hat{\beta}\xrightarrow{p}\beta_0$.*

The restriction, $\alpha n^{1/2-1/\gamma-\varepsilon}\to\infty$, on the rate of convergence of the regularization parameter α is the counterpart of the restriction on the growth rate of K imposed by DIN03 to obtain consistency of the standard CUE, that is $\zeta(K)^2K/n^{1-2/\gamma}\to0$. This restriction on the rate of convergence of α is weaker when there are more moments of $\rho(z_i, \beta)$. It implies in particular that α goes to zero slower that $1/\sqrt{n}$.

We need some additional conditions for asymptotic normality. Let $\rho_\beta(w,\beta) = \partial\rho(w,\beta)/\partial\beta'$, $D(z) = E\left[\rho_\beta(w,\beta_0)\,|z\right]$, and $\rho_{\beta\beta}(w,\beta) = \partial^2\rho(w,\beta)/\partial\beta\partial\beta'$.

Assumption 3. *(a) $\beta_0 \in int(\mathscr{B})$; (b) $\rho(w,\beta)$ is twice continuously differentiable in a neightorhood $\mathscr{N}$ of β_0; (c) $E\left[\sup_{\beta\in\mathscr{N}} \left\|\rho_\beta(w,\beta)\right\|^2 |z\right]$ and $E\left[\left\|\rho_{\beta\beta}(w,\beta_0)\right\| |z\right]$ are bounded; (d) $E\left[D(z)'D(z)\right]$ is nonsingular.*

These assumptions are quite standard regularity conditions used by DIN03. Part (d) is the local identification condition that is essential for asymptotic normality. Parts (b) and (c) are standard smoothness conditions.

We need to introduce some notions before stating the asymptotic normality result. Let $\hat{g}_i = g_i(\hat\beta)$, $\widehat{G}_i = G_i(\hat\beta)$, $\widehat{G} = \sum_{i=1}^n \hat\pi_i \widehat{G}_i$, $\widehat{\Omega} = \sum_{i=1}^n \hat{g}_i\hat{g}_i'/n$, and $\widehat{V} = \left(\hat{G}'\left(\hat\Omega + \alpha I\right)^{-1}\hat{G}\right)^{-1}$.

Theorem 4.2. *If Assumptions 1, 2 and 3 are satisfied, $K \to \infty$, $\alpha \to 0$, $\alpha^{3/2}n^{1/2-1/\gamma-\varepsilon} \to \infty$, and $\alpha^{5/2}n^{1/2} \to \infty$ then*

$$\sqrt{n}\left(\hat\beta - \beta_0\right) \xrightarrow{d} \mathscr{N}(0,V), \quad \hat{V} \xrightarrow{p} V, \quad V = \left(E\left[D(z)'\sigma(z)^{-2}D(z)\right]\right)^{-1}.$$

This result gives the restrictions on the convergence rate of α for the regularized CUE estimator to reach the semiparametric efficiency bound of Chamberlain (1987). These restrictions imply, in particular, that α goes to zero slower than $1/n^{1/5}$. This condition is the counterpart of the restriction on the growth rate of K imposed by DIN03 to obtain asymptotic efficiency of the standard CUE, that is, $\zeta(K)^2K^2/n \to 0$. The main advantage is that our rate no longer depends on K, so more instruments can be used to gain efficiency in finite samples. Although the standard CUE and its regularized version are both asymptotically efficient, their small sample properties can differ, as shown in Monte Carlo simulations. Theorem 4.2 also provides an efficient estimator of the asymptotic variance of the regularized CUE. To the best of our knowledge, these asymptotic results are new and extend DIN03 to the asymptotic framework where no restriction is imposed on the growth rate of K relative to n.

Alternative variance estimator: more robust to many moments

Newey and Windmeijer (2009) argued that in the presence of many moments (potentially weak), the standard textbook variance estimator $\widehat{V}_0/n$, where $\widehat{V}_0 = \left(\widehat{G}'\widehat{\Omega}^{-1}\widehat{G}\right)^{-1}$, does not provide a good approximation to the finite sample distribution of the standard CUE estimator. They suggest instead approximating the finite sample variance of $\hat\beta_{CUE}$ by $\widetilde{V}/n$, where $\widetilde{V} = \hat{H}^{-1}\hat{D}'\hat\Omega^{-1}\hat{D}\hat{H}^{-1}$. They argued that, in the presence of many moment conditions, the former underestimates the finite sample variance of $\hat\beta_{CUE}$, the latter then adjusts for the presence of many moments. In the same spirit, we suggest using the following estimator of the asymptotic variance of the regularized CUE estimator,

$$\tilde{V} = \hat{H}^{-1}\hat{D}'(\hat\Omega + \alpha I)^{-1}\hat{D}\hat{H}^{-1}, \tag{11}$$

where $\hat{H} = \left.\frac{\partial^2 \hat{Q}(\beta)}{\partial \beta \partial \beta'}\right|_{\beta=\hat{\beta}}$, $\quad \hat{D} = \hat{D}(\hat{\beta}), \quad \hat{\Omega} = \hat{\Omega}(\hat{\beta})$, with

$$\hat{Q}(\beta) = \hat{g}(\beta)'(\hat{\Omega}(\beta) + \alpha I)^{-1}\hat{g}(\beta)/2, \quad \hat{D}(\beta) = \sum_i \hat{\pi}_i(\beta)G_i(\beta) \quad \hat{\Omega}(\beta) = \frac{1}{n}\sum_i g_i(\beta)g_i(\beta)', \quad \text{and}$$

$$\hat{\pi}_i(\beta) = \frac{1 - \hat{g}(\beta)'(\hat{\Omega}(\beta) + \alpha I)^{-1}g_i(\beta)}{\sum_j \left(1 - \hat{g}(\beta)'(\hat{\Omega}(\beta) + \alpha I)^{-1}g_j(\beta)\right)}.$$

We will not investigate consistency of $\tilde{V}$ in this paper, but we believe that strategies used to obtain consistency of $\hat{V}$ (see the proof of Theorem 4.2 in the Appendix) can be used to show consistency of $\tilde{V}$ under certain restrictions on the convergence rate of the regularization parameter. We see in simulation that a Wald test of $H_0 : \beta = \beta_0$ based on $\tilde{V}$ performs better (in terms of size) in the presence of many moment conditions compared to the test constructed from $\hat{V}$, in almost all simulation frameworks we considered.

In the next section, we propose a data-driven method for choosing the regularization parameter in practice.

5 Data-driven selection of the regularization parameter

This section is devoted to the selection of the optimal regularization parameter. We propose to employ cross-validation to compute the distance of sample moments from zero, which is then used as a criterion to select the regularization parameter. This approach aims to choose, from the family of RRCUEs indexed by α, the estimator that best satisfies the sample counterpart of the moment condition (2). Indeed, if $\hat{\beta}$ is a '*good*' estimator of β_0, the sample moment function $\sum_{i=1}^n g(x_i, \beta_0)/n$ would be '*close*' to zero, in the sense of a certain norm, if β_0 were replaced by $\hat{\beta}$. Instead of using the simple l_2 norm to measure the distance of the sample moment function from zero, we propose an alternative distance[6]. We use L-fold cross-validation to construct a '*suitable*' distance of the sample mean $\sum_{i=1}^n g\left(x_i, \hat{\beta}\right)/n$ from zero. Using this distance as criteria allows us to choose α such that the corresponding estimator best satisfies the moment condition of Eq.(2) even out-of-sample.

To elaborate, let $\{J_l, l = 1, \cdots, L\}$ denote a partition of the set of data indices $[n] := \{1, 2, \cdots, n\}$. For each $l = 1, \ldots, L$, let J_{-l} denote the set of all indices in $[n]$ except those in J_l, and let n_l denote the cardinality of J_l. Let $\hat{\beta}_{-l}$ denote the version of the regularized CUE estimator obtained using the part of the sample indexed by J_{-l}. The following algorithm describes the general procedure for choosing the optimal α using L-fold cross-validation.

Algorithm (L-fold CV approach for selecting the optimal α).

1. *Consider a grid Δ_K of values of α.*

[6] As the entire sample was used to obtain $\hat{\beta}$ by minimizing a quadratic function of the sample moment function $\sum_{i=1}^n g(x_i, \beta_0)/n$, the simple l_2 norm of $\sum_{i=1}^n g\left(x_i, \hat{\beta}\right)/n$ may provide a misleading measure of how well $\hat{\beta}$ satisfies the sample counterpart of the moment condition (2).

2. *For each $\alpha \in \Delta_K$ and for each $l = 1, \cdots, L$, compute $\hat{\beta}^{\alpha}_{-l}$ (resp. $\hat{\beta}^{\alpha}_{l}$) , the version of the RRCUE obtained using the part of the sample indexed by J_{-l} (resp. J_l). Let $\widetilde{\Omega}_l = diag\left(\Omega(\hat{\beta}^{\alpha}_{l})\right)$.*

3. *For each $\alpha \in \Delta_K$ and for each $l = 1, \cdots, L$, compute the distance $\mathscr{I}_{nl}(\alpha)$ (it measures how well $\hat{\beta}^{\alpha}_{-l}$ satisfies the moment condition $E\left[g(x_i, \beta_0)\right] = 0, i \in J_l$.*

$$\mathscr{I}_{nl}(\alpha) = \left(\frac{1}{n_l}\sum_{i \in J_l} g\left(x_i, \hat{\beta}^{\alpha}_{-l}\right)\right)' \widetilde{\Omega}_l^{-1} \left(\frac{1}{n_l}\sum_{i \in J_l} g\left(x_i, \hat{\beta}^{\alpha}_{-l}\right)\right) \tag{12}$$

4. *The optimal α is obtained by*

$$\hat{\alpha} = \arg\min_{\alpha \in \Delta_K} \left(\mathscr{I}_n(\alpha) := \sum_{l=1}^{L} \mathscr{I}_{nl}(\alpha)\right) \tag{13}$$

Remark 5.1.

- *The dependence of the grid Δ_K in K is motivated by the fact that the regularization of the covariance matrix of $\hat{\Omega}(\beta)$ is preceded by the normalization of q_i by the upper bound, $\zeta(K)$, of the sup-norm of the approximating functions. Our simulation exercise suggests choosing the grid Δ_K to be inversely proportional to K, so that the grid shrinks to zero as K increases. This consideration seems counter-intuitive, as one might think that the regularization parameter would be higher for larger values of K. However, this intuition could be misleading here because the normalization performed prior to regularization tends to considerably reduce the eigenvalues of the matrix $\hat{\Omega}(\beta)$. As a consequence, large values of α will introduce substantial regularization bias. With this normalization, one expects the optimal choice of α to be a decreasing function of K. This is why we anticipated this by choosing a grid that narrows towards zero as K increases. Even when considering a grid that does not depend on K, we observed in simulations that the optimal choice of α, according to our procedure, decreases with K.*

- *The number of folds L has to be chosen. We see in simulations that our result is not very sensitive to a small number of folds (ranging from 2 to 10). In the application, we choose $L = 5$.*

6 Monte Carlo study

This section aims to examine the small sample properties of our RRCUE estimator in order to evaluate the gain of regularization. The baseline setup for our simulation is given by the following system,

$$\begin{cases} y_i = h(x_i, \beta_0) + e_i \\ x_i = f(z_i) + u_i \end{cases} \quad \text{with } \beta_0 = 0.1, \tag{14}$$

where the first equation is the main structural model and the second is the reduced-form equation for the right-hand side endogenous regressor. Both y_i and x_i are univariate, but the underlying vector of instruments z_i is potentially high-dimensional. Our parameter of interest is β_0. We generate data in a way that $E[e_i|z_i] = 0$, given Eq. (1) with $\rho(w_i, \beta_0) = y_i - h(x_i, \beta_0)$ with $w_i = (x_i, y_i)'$.

We consider different experiments. Each framework is designed to mimic a specific situation that can arise in application. The data-generating processes differ: (i) by the type of specification of the main equation (the fonction h is either linear or nonlinear with respect to β); (ii) by the nature of the structural disturbance (homoskedastic or heteroskedastic); (iii) by the number of relevant instruments that enter the first-stage equation (small number or large number); (iv) or by the way the explanatory power, captured by the concentration parameter, is distributed among instruments.

In experiments 1 and 2 below, we assume that the reduced-form error term $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$, with $\sigma_u^2 = 1$. Following Hausman et al. (2012), we suppose that the structural disturbance e_i, which is allowed to be heteroskedastic, is given by

$$e_i = \rho u_i + \sqrt{\frac{1-\rho^2}{\phi^2 + \psi^4}} (\phi v_{1i} + \psi v_{2i}),$$

where $\rho = 0.3$, $\psi = 0.86$ and conditional on z_{i1} (where z_{ij} is the jth component of z_i), $v_{1i} \overset{i.i.d.}{\sim} N\left(0, z_{i1}^2\right)$ and $v_{2i} \overset{i.i.d.}{\sim} N(0, \psi^2)$ are independent of u_i. $\phi = 0$ or 1.38072 is chosen so that the R-squared for the regression of e^2 on the instruments[7], $\mathscr{R}^2_{e^2|z}$, is 0 or 0.2, corresponding to homoskedastic and heterosckedastic cases respectively.

We compare the performance of RRCUE to that of some state-of-the-art estimators in the literature, including the standard two-step GMM of Hansen (1982) and the standard CUE of Hansen et al. (1996). In the case of a linear homoskedastic structural equation, we consider four additional alternative estimators: 2SLS, LIML, the Tikhonov regularized 2SLS estimator (T2SLS) of Carrasco (2012), and the Tikhonov regularized LIML estimator (TLIML) of Carrasco and Tchuente (2015). In the case of linear heteroskedastic structural equation, we consider two more alternative estimators: Hausman et al. (2012)'s heteroskedasticity-robust version of Fuller (1977)'s estimator (HFUL) and heteroskedasticity-robust LIML (HLIM).

We set $n = 500$, and the number of instrumental variables K is chosen from the set $\{2, 30, 50, 100\}$ for experiment 1 and from the set $\{15, 30, 50, 100\}$ for experiment 2. With two specifications for the structural equation (linear or nonlinear), two specifications for the corresponding error term (homoskedastic or heteroskedastic), and four different choices for the number of instruments, there are a total of 16 specifications for experiment 1. Similarly, with two different sets of first-stage coefficients and four different choices for the number of instruments, there are a total of 8 specifications for experiment 2. Therefore, the two experiments total 24 specifications.

For each specification, we performed 10,000 Monte Carlo simulations. For each draw, we compute the optimal RRCUE estimator and the alternative comparison estimators. The computation of the optimal RRCUE requires a suitable choice of the grid Δ_K of values for α. We search for the optimal α in the following 100-point grid,

$$\Delta_K = \left\{ \frac{1}{\sqrt{K}} \times \left(0.0001 + (i-1) \times \frac{0.0499}{99} \right) \middle| i = 1, 2, \ldots, 100 \right\},$$

[7] $R^2_{e^2|z} = \text{var}\left\{\text{E}\left(e^2 \mid z\right)\right\} / \left[\text{var}\left\{\text{E}\left(e^2|z\right)\right\} + \text{E}\left\{\text{var}\left(e^2 \mid z\right)\right\}\right]$.

which is the set of 100 points uniformly distributed between 0.0001 and 0.05, where each element is multiplied by $1/\sqrt{K}$. This normalization allows the grid to approach zero as K increases.

At the end of the 10,000 replications, we calculate several performance measures for each estimator. We consider two measures of bias: the mean bias (Mean.bias) and the median bias (Med.bias); four measures of dispersion: the variance of estimates (Var), the median absolute deviation (MAD, defined as the median of the absolute value of the difference between simulated estimates and the median simulation estimate), and the nine decile range (Ndr$(0.95-0.05)$, defined as the range between the 0.05 and 0.95 quantiles of the distribution of simulated estimates). We also compute the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$. To compute the Wald statistic for CUE, we rely on the many-instruments robust standard error of Newey and Windmeijer (2009), and for RRCUE, we use its regularized counterpart presented in section 4.

6.1 Experiment 1: Small number of relevant instruments

We first consider a setup with a small number of relevant instruments. In particular, we suppose that $z_i \overset{i.i.d.}{\sim} N(0,1)$ and $f(z_i) = \pi z_i$, where π is a scalar chosen so that the concentration parameter $n\pi^2 = \mu^2 = 32$. Also, we consider the following set of instruments in the spirit of Hausman et al. (2012),

$$q^K(z_i) = \begin{cases} (1, z_i) & \text{if} \quad K = 2 \\ \left(1, z_i, z_i^2, z_i^3, z_i^4, z_i D_{i1}, \cdots, z_i D_{i,K-5}\right)' & \text{if} \quad K \in \{30, 50, 100\} \end{cases}$$

where $D_{ik} \in \{0,1\}, \Pr(D_{ik} = 1) = 1/2$. This instrument set consists of powers of z up to fourth power plus interactions with dummy variables. Note that only z affects the reduced-form. Since the exact specification of the reduced-form is unknown in practice, this framework will also help evaluate the effect of including more power series than necessary.

In this experiment, we consider two types of specifications for the structural equation:

(i) **Model 1a.** linear specification, $h(x_i, \beta) = \beta x_i$,

(ii) **Model 1b.** nonlinear specification, $h(x_i, \beta) = \exp(\beta x_i)$.

For each of these frameworks, we consider cases where the structural disturbance is either homoskedastic ($\phi = 0$) or heteroskedastic ($\phi = 1.38072$.)

Table 1 presents results for experiment 1 with linear structural equation (Model 1a). It offers a detailed comparison of the RRCUE estimator's performance against several competitors, including CUE, GMM, T2SLS, TLIML, HFUL, and HLIM, under both homoskedastic and heteroskedastic disturbances across varying numbers of instruments (K). The focus will be on bias (Mean.bias, Med.bias) and dispersion (Var, MAD, Ndr$(0.95-0.05)$), particularly highlighting RRCUE's performance relative to CUE, GMM, and other estimators.

Panel A - Homoskedasticity: For small K ($K = 2$, only instruments that enter the reduced-form equation are used), regularization is not needed. In fact, the median bias for RRCUE is close to that of CUE and slightly better than GMM. In terms of dispersion, RRCUE and CUE report almost the same variance and MAD, while GMM shows slightly lower variance, though this comes at the cost of higher median bias. As K increases ($K = 30, 50, 100$), RRCUE outperforms GMM in terms of median bias, although regularization introduces a certain amount of bias compared to CUE. RRCUE remains comparable to competitors like T2SLS and TLIML in terms of median bias. In terms of dispersion, RRCUE exhibits substantially lower variance compared to CUE. This pattern persists as K increases, with RRCUE maintaining a much lower variance, MAD, and nine-decile range than CUE. Compared to T2SLS and TLIML, RRCUE consistently shows comparable performance in terms of bias and dispersion as the number of instruments grows.

Panel B - Heteroskedasticity: Under heteroskedastic disturbances, RRCUE continues to demonstrate strong performance relative to CUE, GMM, HFUL, and HLIM with some exceptions. For $K = 2$, RRCUE's median bias is comparable to that of CUE, GMM, HFUL, and HLIM. Similarly, RRCUE is almost equivalent to standard competitors in terms of dispersion, denoting again the fact that regularization is not needed for small K. As K increases, the effect of regularization is more effective. RRCUE keeps dispersion under control while maintaining a reasonable level of bias that remains smaller than the GMM over-identification bias. Against HFUL and HLIM, RRCUE shows clear advantages in terms of dispersion. For example, at $K = 30, 50$, RRCUE substantially outperforms HFUL and HLIM in terms of dispersion (variance, MAD, and nine-decile range), while keeping the regularization bias at a manageable level.

Moreover, RRCUE's rejection frequency remains close to the 5% nominal level in both homoskedastic and heteroskedastic settings, indicating accurate hypothesis testing. In contrast, CUE and GMM show deteriorating rejection frequencies for larger K, while HFUL and HLIM also deviate substantially from the nominal level, especially at $K = 100$ (HFUL = 0.604, HLIM = 0.070).

Table 2 presents results for experiment 1 with nonlinear structural equation (Model 1b). The results highlight that RRCUE generally performs well compared to CUE and GMM in both homoskedastic and heteroskedastic settings. In terms of bias, RRCUE shows low median bias across different numbers of moments (K), particularly it maintains comparable or slightly better performance than CUE and GMM. Regarding dispersion, RRCUE exhibits smaller variance than CUE, especially as K increases, while maintaining lower MAD compared to CUE, indicating tighter concentration around the median. Notably, RRCUE's Ndr (0.95-0.05) remains relatively stable across all K values, outperforming CUE and GMM in controlling over-dispersion for larger sets of moments ($K = 30, 50, 100$), particularly under heteroskedastic disturbances. Finally, in terms of rejection frequency at the 5% nominal level, RRCUE maintains competitive performance, consistently rejecting less often than GMM, particularly when K is large, which suggests better size control in finite samples.

Overall, RRCUE demonstrates a balance between bias and dispersion across different numbers of instruments and disturbance structures. Results in Tables 1 & 2 reveal that even if a small number of instruments enter the reduced-form equation, using power series and splines as additional instruments, together with regularization, can help improve the efficiency of the CUE while maintaining the regularization bias at a manageable level if the regularization parameter is chosen in a suitable manner.

Table 1: Simulations results: Experiment 1 - Small number of relevant instruments and linear structural equation (Model 1a)

		Panel A: Homoskedasticity							**Panel B: Heteroskedasticity**				
	Estimator	**RRCUE**	**CUE**	**GMM**	**2SLS**	**T2SLS**	**LIML**	**TLIML**	**RRCUE**	**CUE**	**GMM**	**HFUL**	**HLIM**
K=2	Mean.bias	-0.010	-0.010	0.000	0.000	0.000	-0.010	-0.010	-0.008	-0.009	0.010	0.013	0.003
	Med.bias	0.002	0.002	0.010	0.010	0.010	0.001	0.001	-0.001	-0.001	0.019	0.021	0.013
	Var	0.029	0.029	0.027	0.027	0.027	0.029	0.029	0.072	0.072	0.062	0.057	0.062
	MAD	0.106	0.106	0.103	0.102	0.102	0.106	0.106	0.170	0.171	0.158	0.153	0.158
	Ndr (0.95-0.05)	0.553	0.556	0.528	0.530	0.529	0.553	0.552	0.870	0.877	0.813	0.779	0.810
	0.05 rej.Freq	0.040	0.042	0.047	0.045	0.045	0.041	0.040	0.045	0.046	0.049	0.049	0.046
K=30	Mean.bias	0.048	1.211e+09	0.120	0.143	0.046	-0.014	-0.007	0.046	6.147e+08	0.122	-0.014	-0.048
	Med.bias	0.054	0.019	0.122	0.147	0.050	0.004	0.004	0.055	0.014	0.123	0.022	0.012
	Var	0.034	6.063e+23	0.017	0.012	0.021	0.838	0.033	0.061	2.078e+23	0.037	0.197	6.672
	MAD	0.112	0.186	0.084	0.074	0.095	0.147	0.112	0.148	0.226	0.121	0.205	0.215
	Ndr (0.95-0.05)	0.575	1.054	0.419	0.361	0.472	0.839	0.579	0.788	1.343	0.613	1.307	1.510
	0.05 rej.Freq	0.054	0.098	0.327	0.281	0.096	0.035	0.041	0.041	0.095	0.285	0.060	0.058
K=50	Mean.bias	0.074	2.870e+09	0.136	0.182	0.061	-5.927e+07	-0.009	0.078	2.651e+10	0.143	-0.005	-3.248
	Med.bias	0.080	0.044	0.136	0.182	0.064	0.003	0.005	0.087	0.043	0.142	0.023	0.013
	Var	0.031	3.464e+23	0.014	0.009	0.018	3.693e+20	0.033	0.051	1.421e+25	0.034	0.270	6.854e+04
	MAD	0.101	0.219	0.077	0.064	0.087	0.173	0.112	0.133	0.266	0.116	0.241	0.255
	Ndr (0.95-0.05)	0.525	1.315	0.387	0.311	0.443	1.048	0.582	0.708	1.664	0.593	1.593	2.047
	0.05 rej.Freq	0.052	0.145	0.463	0.495	0.106	0.037	0.038	0.041	0.145	0.424	0.065	0.062
K=100	Mean.bias	0.102	-1.806e+09	0.132	0.226	0.096	-2.19e+09	-0.013	0.114	1.016e+09	0.143	0.021	0.019
	Med.bias	0.110	0.088	0.131	0.227	0.096	0.008	0.002	0.123	0.093	0.139	0.040	0.027
	Var	0.030	1.919e+22	0.012	0.005	0.013	1.451e+23	0.057	0.055	1.054e+23	0.033	0.420	107.265
	MAD	0.093	0.239	0.072	0.050	0.075	0.225	0.115	0.120	0.282	0.114	0.303	0.329
	Ndr (0.95-0.05)	0.498	1.543	0.356	0.241	0.377	1.694	0.601	0.677	1.821	0.589	2.159	3.250
	0.05 rej.Freq	0.061	0.235	0.620	0.864	0.154	0.038	0.038	0.064	0.268	0.604	0.074	0.070

Note: Simulation results based on 10,000 replications. We report six (06) measures of performance: the mean bias (Mean.bias), the median bias (Med.bias), the variance of estimates (Var), the median absolute deviation (MAD), the nine decile range (Ndr$(0.95-0.05)$), and the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$.

Table 2: Simulations results: Experiment 1 - Small number of relevant instruments and nonlinear structural equation (Model 1b)

		Panel A: Homoskedasticity			**Panel B: Heteroskedasticity**		
		RRCUE	CUE	GMM	RRCUE	CUE	GMM
K=2	Mean.bias	-0.020	-0.020	-0.014	-0.038	-0.038	-0.033
	Med.bias	0.001	0.000	0.007	-0.011	-0.011	-0.004
	Var	0.022	0.022	0.021	0.033	0.033	0.032
	MAD	0.094	0.094	0.091	0.122	0.122	0.117
	Ndr (0.95-0.05)	0.485	0.486	0.476	0.587	0.587	0.585
	0.05 rej.Freq	0.128	0.132	0.119	0.205	0.209	0.176
K=30	Mean.bias	0.019	-0.026	0.086	-0.006	-0.041	0.064
	Med.bias	0.044	0.011	0.095	0.033	0.003	0.082
	Var	0.023	0.041	0.010	0.028	0.037	0.016
	MAD	0.092	0.138	0.062	0.101	0.131	0.071
	Ndr (0.95-0.05)	0.494	0.644	0.318	0.543	0.606	0.410
	0.05 rej.Freq	0.114	0.177	0.329	0.134	0.179	0.284
K=50	Mean.bias	0.045	-0.020	0.103	0.020	-0.045	0.091
	Med.bias	0.066	0.030	0.110	0.056	0.011	0.103
	Var	0.019	0.054	0.008	0.025	0.046	0.015
	MAD	0.081	0.153	0.057	0.090	0.144	0.071
	Ndr (0.95-0.05)	0.455	0.737	0.293	0.523	0.677	0.391
	0.05 rej.Freq	0.107	0.224	0.455	0.115	0.196	0.407
K=100	Mean.bias	0.072	-0.003	0.106	0.053	-0.043	0.108
	Med.bias	0.090	0.065	0.111	0.085	0.034	0.113
	Var	0.017	0.078	0.008	0.024	0.073	0.018
	MAD	0.071	0.163	0.057	0.076	0.157	0.080
	Ndr (0.95-0.05)	0.415	0.873	0.284	0.493	0.814	0.437
	0.05 rej.Freq	0.127	0.310	0.614	0.120	0.268	0.597

Note: Simulation results based on 10,000 replications. We report six (06) measures of performance: the mean bias (Mean.bias), the median bias (Med.bias), the variance of estimates (Var), the median absolute deviation (MAD), the nine decile range (Ndr$(0.95-0.05)$), and the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$.

6.2 Experiment 2: Large number of relevant instruments

Our second experiment design involves a large number of relevant instruments in a framework where both structural and reduced-form models are linear. In particular, we assume that $h(x_i, \beta) = \beta x_i$ and $f(z_i) = \pi' z_i$, where π is a high-dimensional vector of first-stage coefficients that satisfies $n\sigma_u^{-2}\pi'\Sigma_z\pi = \mu^2$, with the concentration parameter μ^2 measuring the strength of the instruments and $\Sigma_z = E\left[z_i z_i'\right]$. Following Hansen and Kozbur (2014) we consider Gaussian instruments that are correlated with one another. Under this Gaussian instrument design, all instruments are drawn with mean 0 and variance $\text{var}\left(z_{ij}\right) = \Sigma_{zjj} = 0.3$. Dependence between instruments is given by the Pearson correlation coefficient $\text{corr}\left(z_{ij}, z_{ik}\right) = 0.5^{|j-k|}$. As z_i is already a large vector, we consider it as the instrument set without adding power series; that is, $q^K(z_i) = z_i$. In this experiment, we focus on heteroskedastic structural disturbance and consider two different set of first stage coefficients in the spirit of Donald and Newey (2001) and Carrasco (2012):

(i) **Model 2a.** $\pi_l = d\left(1 - \frac{l}{K+1}\right)^4$, for $l = 1, \ldots, K$, where the constant $d := \sqrt{\frac{\sigma_u^2 \mu^2}{n \tilde{\pi}' \Sigma_z \tilde{\pi}}}$, with $\tilde{\pi} = \pi/d$, is chosen so that $\mu^2 = 32$. The instruments are ranked in decreasing order of importance. This specification is relevant for applications where there is some prior information about which instruments are more important.

(ii) **Model 2b.** $\pi_l = d$, for $l = 1, \ldots, K$, where the constant $d := \sqrt{\frac{\sigma_u^2 \mu^2}{n \iota_K' \Sigma_z \iota_K}}$, with ι_K a $K \times 1$ vector of ones, is chosen so that $\mu^2 = 32$. This framework is relevant for applications where the instruments are equally important.

Table 3 presents results for the case where a large number of instruments enter the reduced-form equations. In Panel A, which considers the case of instruments ranked in decreasing order (Model 2a), the performance of RRCUE is generally superior to that of its competitors, particularly CUE and GMM. With a median absolute deviation (MAD) of 0.103 and a variance of 0.027 at $K = 15$, RRCUE demonstrates lower dispersion compared to CUE, which exhibits erratic behavior with a MAD of 0.126 and an excessively high variance. Moreover, RRCUE maintains a favorable Ndr (0.95-0.05) of 0.529, indicating a balanced performance across different deciles. In terms of rejection frequency, RRCUE shows a consistent and controlled rejection frequency, which is significantly lower than that of GMM, further underscoring its robustness in maintaining type I error rates. In comparison to HFUL and HLIM, RRCUE exhibits comparable MAD and variance, indicating its competitiveness in terms of dispersion. The same pattern is observed when the number of instruments increases. We have similar results in Panel B, which evaluates the scenario with equally important instruments (Model 2b). Overall, RRCUE consistently demonstrates superior performance in terms of bias (compared to GMM) and dispersion (compared to CUE, HFUL, and HLIM). Moreover, the role of regularization seems to be much more important when there is a large number of relevant instruments that enter the reduced-form equation.

In summary, regularization allows solving the moment problem of CUE by reducing its dispersion. RRCUE can take advantage of a bunch of moments/instruments and gain efficiency while maintaining the regularization bias at a relatively low and reasonable level.

7 Empirical application: Institutions and growth

This section revisits the empirical work of Hall and Jones (1999), aiming to answer the famous question: *Why do some countries produce so much more output per worker than others?*. This question is primarily motivated by the simple fact that output per worker varies enormously across countries. Hall and Jones (1999) argue that the differences in capital accumulation, productivity, and therefore output per worker are driven by differences in institutions and government policies, which they call social infrastructure.

Table 3: Simulations results: Experiment 2 - Large number of relevant instruments and heteroskedastic structural disturbance

		Panel A: Instruments ranked in decreasing order (Model 2a)					**Panel B: Instruments are equally important (Model 2b)**				
	Estimator	**RRCUE**	**CUE**	**GMM**	**HFUL**	**HLIM**	**RRCUE**	**CUE**	**GMM**	**HFUL**	**HLIM**
K=15	Mean.bias	0.024	2.494e+10	0.102	0.003	-0.005	0.014	-1.23e+10	0.087	-0.002	-0.018
	Med.bias	0.030	0.003	0.106	0.014	0.005	0.023	0.001	0.091	0.010	0.001
	Var	0.027	4.577e+24	0.013	0.033	0.319	0.021	1.509e+24	0.011	0.029	0.224
	MAD	0.103	0.126	0.077	0.110	0.115	0.091	0.106	0.070	0.099	0.103
	Ndr (0.95-0.05)	0.529	0.685	0.375	0.585	0.621	0.470	0.579	0.350	0.528	0.564
	0.05 rej.Freq	0.042	0.049	0.215	0.052	0.049	0.038	0.043	0.192	0.050	0.046
K=30	Mean.bias	0.054	1.033e+11	0.144	-0.002	-0.022	0.044	-3.576e+10	0.135	-0.006	-0.071
	Med.bias	0.057	0.006	0.145	0.009	-0.000	0.047	0.005	0.136	0.011	0.001
	Var	0.022	2.760e+25	0.009	0.042	0.196	0.020	7.573e+24	0.008	0.039	10.418
	MAD	0.095	0.141	0.062	0.117	0.123	0.088	0.123	0.058	0.111	0.116
	Ndr (0.95-0.05)	0.481	0.800	0.307	0.643	0.692	0.456	0.724	0.291	0.618	0.674
	0.05 rej.Freq	0.046	0.056	0.449	0.052	0.048	0.043	0.048	0.426	0.054	0.050
K=50	Mean.bias	0.099	1.318e+10	0.174	-0.006	0.051	0.091	2.973e+10	0.168	-0.010	-0.095
	Med.bias	0.097	0.015	0.174	0.011	0.001	0.090	0.013	0.168	0.011	0.001
	Var	0.018	7.978e+25	0.006	0.065	137.613	0.015	5.630e+25	0.006	0.059	31.513
	MAD	0.079	0.162	0.051	0.131	0.138	0.075	0.151	0.050	0.126	0.133
	Ndr (0.95-0.05)	0.410	1.025	0.256	0.754	0.839	0.391	0.948	0.250	0.732	0.815
	0.05 rej.Freq	0.054	0.072	0.716	0.052	0.049	0.051	0.068	0.701	0.050	0.047
K=100	Mean.bias	0.160	-1.969e+10	0.207	-0.000	-0.531	0.157	5.229e+09	0.204	-0.001	-0.004
	Med.bias	0.161	0.073	0.207	0.020	0.008	0.155	0.065	0.205	0.018	0.006
	Var	0.010	5.630e+25	0.004	0.122	1.258e+03	0.009	1.749e+25	0.004	0.121	60.889
	MAD	0.064	0.199	0.041	0.167	0.177	0.063	0.193	0.041	0.167	0.177
	Ndr (0.95-0.05)	0.317	1.297	0.203	1.072	1.288	0.308	1.286	0.200	1.051	1.265
	0.05 rej.Freq	0.095	0.138	0.965	0.067	0.062	0.090	0.128	0.965	0.065	0.061

Note: Simulation results based on 10,000 replications. We report six (06) measures of performance: the mean bias (Mean.bias), the median bias (Med.bias), the variance of estimates (Var), the median absolute deviation (MAD), the nine decile range (Ndr$(0.95-0.05)$), and the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$.

To quantify the effect of social infrastructure on per capita income, they treat social infrastructure as endogenous. Identification is then based on the idea that social infrastructure is determined historically by location and other factors captured in part by language, all of which are exogenous. More precisely, they use 2SLS with four instruments for social infrastructure: the fraction of population speaking English at birth (EnL), the fraction of population speaking one of the five major European languages at birth (EuL), the distance from the equator (latitude, Lt), and Frankel and Romer (1999) geography-predicted trade intensity (FR). The linear IV regression model is given by

$$y_i = c + \delta S_i + \varepsilon_i \quad i = 1, 2 \ldots, n = 79,$$

where y_i is country i's log income per capita, S_i is country i's proxy for social infrastructure, c is a constant, and δ is the scalar parameter of interest.

The baseline $n \times 4$ matrix of instruments is given by $z = [EnL, EuL, Lt, FR]$. Recently, Dmitriev (2013) pointed out that these instruments are weak. To boost their identification strength, we increase the set of instruments from 4 to 18, as suggested by Carrasco and Tchuente (2016)[8].

Our enriched set of instruments is given by[9]

$$\begin{aligned} q(z) =& \big[z, z.^2, z.^3, z(:,1) \times z(:,2), z(:,1) \times z(:,3), z(:,1) \times z(:,4), \\ & z(:,2) \times z(:,3), z(:,2) \times z(:,4), z(:,3) \times z(:,4)\big], \end{aligned}$$

where all instruments are divided by their standard deviation prior to regularization. Our sample consists of $n = 79$ countries for which no data were imputed[10]. Results are collected in Table 4.

Table 4 presents estimates of the effect of social infrastructure on growth using various estimators for two different numbers of instruments: $K = 4$ (benchmark) and $K = 18$ (many instruments). Across the estimators, RRCUE demonstrates a notable balance between the size of the estimates and the precision, particularly in the case of $K = 18$. For $K = 4$, RRCUE produces an estimate of 5.704, comparable to LIML and HLIM, but with a standard error of 1.059, which is slightly higher than other methods and might denote the fact that regularization is not needed in this low-dimensional case. However, as the number of instruments increases to $K = 18$, RRCUE achieves a precise estimate of 4.813 with a relatively low standard error of 0.639, maintaining stability in the presence of many instruments.

Compared to other estimators, RRCUE performs robustly with many instruments, showing superior precision compared to LIML, HLIM, and HFUL, which all experience substantial increases in standard errors. For instance, HLIM and HFUL report estimates of 6.828 and 6.561, respectively, but with much larger standard errors (1.610 and 1.497), indicating less precision which is consistent with the simulation results. In contrast, RRCUE's regularized approach effectively controls for many instruments, offering both a reasonable estimate and greater reliability, making it a competitive choice for empirical analysis in such settings.

[8] Carrasco and Tchuente (2016) argued that the use of many instruments allows the concentration parameter to increase from $\hat{\mu}_n^2 = 28.6$ for 4 instruments to $\hat{\mu}_n^2 = 51.48$ for 18 instruments, resulting in a moderately strong set of instruments.

[9] $z.^k = \left[z_{ij}^k\right]$, $z(:,j)$ is the j th column of z, and $z(:,j) \times z(:,l)$ is a vector of interactions between columns j and l.

[10] Data used are collected from https://web.stanford.edu/~chadj/HallJones400.asc.

Table 4: Estimates of the effect of social infrastructure on growth (NW09 variance)

	OLS	2SLS	LIML	GMM	CUE	HLIM	HFUL	T2SLS	TLIML	RRCUE
$K = 4$	3.074	5.412	5.938	5.367	5.728	5.982	5.812	5.628	5.958	5.704
	(0.296)	(0.777)	(1.012)	(0.695)	(1.017)	(0.936)	(0.874)	(0.807)	(0.948)	(1.059)
								$\alpha = 0.500$	$\alpha = 0.500$	$\alpha = 0.025$
$K = 18$	3.074	3.986	6.093	3.603	4.764	6.828	6.561	4.438	5.523	4.813
	(0.296)	(0.427)	(1.597)	(0.169)	(0.778)	(1.610)	(1.497)	(0.507)	(1.002)	(0.639)
								$\alpha = 0.020$	$\alpha = 0.010$	$\alpha = 0.012$

Note: The sample consists of $n = 79$ countries for which no data were imputed. We present results for both $K = 4$ instruments (benchmark) and for many instruments ($K = 18$), as well as for alternative estimators for comparison purposes. Standard errors are in parentheses. For CUE with 18 instruments, we report the many instruments robust standard error of Newey and Windmeijer (2009), and for RRCUE, we report its regularized counterpart.

8 Conlusion

This paper introduces the Ridge-regularized Continuous Updating Estimator (RRCUE) to address the challenges posed by using a large number of instruments/moments to improve efficiency in moment-based estimation in a framework where the parameter of interest is defined by a single conditional moment restriction. Through theoretical analysis and Monte Carlo simulations, we demonstrate that RRCUE offers a significant reduction in dispersion and improves efficiency compared to standard CUE. Despite introducing a small bias, the estimator remains robust in high-dimensional settings where the number of moments increases with the sample size, providing consistent and precise estimates. We show that RRCUE is competitive and sometimes outperforms state-of-the-art estimators like HLIM and HFUL of Hausman et al. (2012), specifically in the linear instrumental variable framework with heteroskedasticity and many instruments. These findings make RRCUE a promising tool for empirical research, particularly in econometric applications where a large set of instruments/moments is available and there is either no rule to select a subset of them or they have almost the same explanatory power. Promising future research may explore the extension of our regularization scheme to the generalized empirical likelihood (GEL) class of estimators. This research path is particularly interesting as Newey and Smith (2004) justified that the empirical likelihood (EL) estimator enjoys good performance in terms of higher-order bias within the GEL class. Therefore, EL estimator might be a good candidate to address overidentification bias in the many moments setting. Further promising future work includes the investigation of higher-order expansion to derive an approximate mean squared error for optimal selection of the regularization parameter, as suggested by Carrasco (2012) and Carrasco and Tchuente (2015).

Appendix

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses, and M, CS and T the Markov, Cauchy-Schwarz and triangle inequalities respectively. Also, with probability approaching one will be abbreviated as w.p.a.1., p.d. and p.s.d. will be the abbreviation for positive definite and positive semidefinite matrix respectively, and CLT will refer to the Lindeberg-Lévy central limit theorem.

Proof of Theorem 3.1. By $s(v)$ quadratic, a second order Taylor expansion is correct giving

$$\begin{aligned}\widehat{P}(\beta,\lambda) &= \sum_{i=1}^{n} s\left(\lambda' g_i(\beta)\right)/n - \frac{\alpha}{2}\lambda'\lambda \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{s(0)+s'(0)\lambda' g_i(\beta)+\frac{s''(0)}{2}\left(\lambda' g_i(\beta)\right)^2\right\}-\frac{\alpha}{2}\lambda'\lambda \\ &= -\hat{g}(\beta)'\lambda - \frac{1}{2}\lambda'\hat{\Omega}(\beta)\lambda - \frac{\alpha}{2}\lambda'\lambda \\ &= -\hat{g}(\beta)'\lambda - \frac{1}{2}\lambda'\left(\hat{\Omega}(\beta)+\alpha I_K\right)\lambda.\end{aligned}$$

By concavity of $\widehat{P}(\beta,\lambda)$ in λ, any solution $\hat{\lambda}(\beta)$ to the FOCs $0=\hat{g}(\beta)+(\widehat{\Omega}(\beta)+\alpha I)\lambda$ will maximize $\widehat{P}(\beta,\lambda)$ with respect to λ holding β fixed. That is $\hat{\lambda}(\beta)=-\left[\hat{\Omega}(\beta)+\alpha I\right]^{-1}\hat{g}(\beta)$ maximizes $\widehat{P}(\beta,\lambda)$ holding holding β fixed. Then the penalized GEL objective function is given by

$$\widehat{P}\left(\beta,\hat{\lambda}(\beta)\right)=\frac{1}{2}\hat{g}(\beta)'\left[\hat{\Omega}(\beta)+\alpha I_K\right]^{-1}\hat{g}(\beta).$$

Therefore, the penalized GEL objective function is a monotonic increasing transformation of the regularized CUE objective function so that the result follows. □

Proof of Theorem 3.2. As justified in the proof of Theorem 3.1, $\hat{\lambda}(\beta)=-\left[\hat{\Omega}(\beta)+\alpha I\right]^{-1}\hat{g}(\beta)$ maximizes $\widehat{P}(\beta,\lambda)$ holding β fixed.

By the envelope theorem, the FOCs for the regularized CUE $\hat{\beta}$ are given by

$$\begin{aligned}0 &= \left.\frac{\partial\widehat{P}(\beta,\lambda(\beta))}{\partial\beta}\right|_{\beta=\hat{\beta}} \\ &= \left.\frac{\partial\widehat{P}(\beta,\lambda)}{\partial\beta}\right|_{\lambda=\hat{\lambda}(\beta),\,\beta=\hat{\beta}} \\ &= n^{-1}\sum_{i=1}^{n}s_1\left(\lambda' g_i(\beta)\right)G_i(\beta)'\lambda\Bigg|_{\lambda=\hat{\lambda}(\beta),\,\beta=\hat{\beta}} \\ &= n^{-1}\sum_{i=1}^{n}s_1\left(\hat{v}_i\right)G_i\left(\hat{\beta}\right)'\hat{\lambda},\end{aligned}$$

where $\hat{v}_i=\hat{\lambda}\hat{g}_i$. Multiplying by $-n\left(\sum_{i=1}^{n}s_1\left(\hat{v}_i\right)\right)^{-1}$ and using $\hat{\lambda}=-\left[\hat{\Omega}\left(\hat{\beta}\right)+\alpha I\right]^{-1}\hat{g}\left(\hat{\beta}\right)$ give the result. □

We now give some preliminary lemmas for the proof of Theorem 4.1 and of Theorem 4.2. The proof of Theorem 4.1 will be based on the following lemma, borrowed from DNI03, with suitable choices of functions $\hat{R}(\beta)$ and $R(\beta)$. The proof of this lemma can be found in the DIN03's Appendix.

Lemma 1. *(Lemma A1 of Donald et al. (2003)) Suppose that (i) $R(\beta)$ has a unique minimum at $\beta_0 \in \mathscr{B}$; (ii) B is compact; (iii) $R(\beta)$ is continuous; and (iv) $\sup_{\beta\in\mathscr{B}} |\hat{R}(\beta) - R(\beta)| \xrightarrow{p} 0$. Then for any $\tilde{\beta} \in \mathscr{B}$, if $\hat{R}(\tilde{\beta}) \xrightarrow{p} R(\beta_0)$ then $\tilde{\beta} \xrightarrow{p} \beta_0$.*

Lemma 2. *If Assumption 1(a) is satisfied and $\sigma(z)^2 \overset{def}{=} E\left[\rho(w,\beta_0)^2|z\right]$ is bounded then $\|\hat{g}(\beta_0)\| = O_p(n^{-1/2})$*

Proof. Let $q_i = \tilde{q}^K(z_i) = q^K(z_i)/\zeta(K)$ and $\rho_i = \rho(w_i,\beta)$. By Assumption 1(a) $\sup_{\beta\in\mathscr{B}} \left\|\tilde{q}^K(z)\right\| \leq C$ so that $E\left[\|q_i\|^2\right] = O(1)$. It follows from i.i.d. data and the law of iterated expectations that

$$\begin{aligned} E\left[\|\hat{g}\,(\beta_0)\|^2\right] &= E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} q_i\rho_i\right\|^2\right] \\ &= E\left[\frac{1}{n^2}\sum_{i,j}(q_i\rho_i)'\left(q_j\rho_j\right)\right] \\ &= E\left[\rho_i^2\|q_i\|^2\right]/n \\ &= E\left[E\left[\rho_i^2|z_i\right]\|q_i\|^2\right]/n \\ &= E\left[\sigma_i^2\,\|q_i\|^2\right]/n \leq C/n. \end{aligned}$$

The conclusion then follows by M. □

Lemma 3. *If Assumption 1(a) is satisfied, $U_i \overset{def}{=} U(z_i)$ a nonnegative scalar function bounded away from zero, $P_i = q_i U_i^{1/2}$, $P = [P_1,\cdots,P_n]'$, $Q^\alpha = P\left[P'P/n + \alpha I\right]^{-1} P'/n$ with $\alpha > 0$, and $\left(\lambda_j, \phi_j : j = 1,2,\ldots,K\right)$ the eigenvalues and orthonormal eigenvectors $E\left[U_i q_i q_i'\right]$ then*

(i) $\operatorname{tr}(E\,[Q^\alpha]) = O(1/\alpha)$;

(ii) For all x and y, $x'Q^\alpha y \leq \|x\|\|y\|$ so that $\lambda_{\max}(Q^\alpha) \leq 1$;

(iii) $x'(I - Q^\alpha)^2 x \leq x'(I - Q^\alpha)x$ for all x;

(iv) If $\bar{x}$ is an n-dimensional vector such that $\|\bar{x}\|/\sqrt{n} = O_p(1)$, for eack K there is a K-dimensional vector γ_K such that $\|\bar{x} - P\gamma_K\|/\sqrt{n} = o_p(1)$, and there is $\beta \geq 1/2$ such that

$$\sum_{j=1}^{\infty} \frac{\left(E\,[\bar{x}_i P_i]'\,\phi_j\right)^2}{\lambda_j^{2\beta+1}} < \infty,$$

then $\bar{x}'(I - Q^\alpha)\,\bar{x}/n \xrightarrow{p} 0$ as $n \to \infty$ and $\alpha \to 0$.

Proof. To prove (i) remark that $P'P/n+\alpha I \geqslant \alpha I$ so that $\left(P'P/n+\alpha I\right)^{-1}P'P \leqslant P'P/\alpha$ and therefore $\mathrm{tr}\left(Q^{\alpha}\right) \leqslant \mathrm{tr}\left(P'P\right)/(n\alpha)$. Also, note that $P'P=\sum_{i=1}^{n}U_i q_i q_i'$ so that $\mathrm{tr}\left(P'P\right)/n = n^{-1}\sum_{i=1}^{n}U_i\,\mathrm{tr}\left(q_i q_i'\right) \leqslant Cn^{-1}\sum_{i=1}^{n}\|q_i\|^2 \le C$ by $U(z)$ bounded and Assumption 1(a).

To prove (ii) we make use of singular value decomposition (SVD). Recall that $T_n = P/\sqrt{n}$ is an $n\times K$ matrix. Let $T_n = \hat{\Psi}D\widehat{\Phi}$ denotes its SVD, where $\hat{\Psi}$ is an $n\times n$ orthogonal matrix, D is an $n\times K$ rectangular diagonal matrix with nonnegative real numbers on the diagonal, $\widehat{\Phi}$ is an $K\times K$ matrix. Let $\sqrt{\hat{\lambda}_i} = D_{ii}$ denote the diagonal entries of D known as singular values of P. The number of nonzero singular values is equal to the rank r of P. The columns of $\widehat{\Psi}$ and $\widehat{\Phi}$ form two sets of orthonormal bases $\hat{\psi}_1,\ldots,\hat{\psi}_n$ and $\hat{\phi}_1,\ldots,\hat{\phi}_K$. If singular values $\sqrt{\hat{\lambda}_i}$ are sorted in decreasing order such that $\sqrt{\hat{\lambda}_1}\ge\sqrt{\hat{\lambda}_2}\ge\cdots\ge\sqrt{\hat{\lambda}_r}>\sqrt{\hat{\lambda}_{r+1}}=0$ then the SVD can be written as $T_n=\sum_{i=1}^{r}\sqrt{\hat{\lambda}_i}\hat{\psi}_i\hat{\phi}_i'$, where $r\leqslant\min(n,K)$. It follows that for all $j=1,\ldots,K$, $T_n'T_n\hat{\phi}_j=\hat{\lambda}_j\hat{\phi}_j$ so that $\left(\hat{\lambda}_j,\hat{\phi}_j : j=1,2,\ldots,K\right)$ is the system of eigenvalues and orthonormal eigenvectors of $T_n'T_n$. Since $T_n'T_n$ is the sample counterpart of $L=E\left[U_i q_i q_i'\right]$, then $\hat{\lambda}_j$ and $\hat{\phi}_j$ are consistent estimators of the corresponding eigenvalues and eigenvectors of L, λ_j and ϕ_j.

Also, note that $Q^{\alpha}=T_n\left[T_n'T_n+\alpha I\right]^{-1}T_n'=\sum_{j=1}^{n}\frac{\hat{\lambda}_j}{\hat{\lambda}_j+\alpha}\hat{\psi}_j\hat{\psi}_j'$ with $\hat{\lambda}_j=0$ for all $j>r$. $(\hat{\psi}_1,\ldots,\hat{\psi}_n)$ being an orthonormal basis $Q^{\alpha}\hat{\psi}_j=\frac{\hat{\lambda}_j}{\hat{\lambda}_j+\alpha}\hat{\psi}_j$ for all $j=1,\ldots,n$. It follows that eigenvalues of Q^{α} are $\frac{\hat{\lambda}_j}{\hat{\lambda}_j+\alpha}$ for $j=1,\ldots,n$ and the associated eigenvectors are respectively $\hat{\psi}_j$, $j=1,\ldots,n$. Therefore $\lambda_{\max}\left(Q^{\alpha}\right)\leqslant 1$ so that for all vectors x and y, $x'Q^{\alpha}y\le\lambda_{\max}\left(Q^{\alpha}\right)\|x\|\|y\|\le\|x\|\|y\|$.

To prove (iii) note that $I-Q^{\alpha}=\sum_{j=1}^{n}\frac{\alpha}{\hat{\lambda}_j+\alpha}\hat{\psi}_j\hat{\psi}_j'$ so that $\lambda_{\max}\left(I-Q^{\alpha}\right)\le 1$. Also, for $\alpha>0$, $I-Q^{\alpha}$ is a symmetric positive semidefinite matrix. Let $(I-Q^{\alpha})^{1/2}$ be a symmetric square root of $I-Q^{\alpha}$. Then,

$$x'\left(I-Q^{\alpha}\right)^2x=x'\left(I-Q^{\alpha}\right)^{1/2}\left(I-Q^{\alpha}\right)\left(I-Q^{\alpha}\right)^{1/2}x\leqslant\left\|\left(I-Q^{\alpha}\right)^{1/2}x\right\|^2=x'\left(I-Q^{\alpha}\right)x,$$

giving the result.

It remains to prove (iv). Note that for $\alpha=0$, Q^{α} coincide with $Q\overset{def}{=}P(P'P)^{-}P'=\sum_{j=1}^{r}\hat{\psi}_j\hat{\psi}_j'$, where $(\cdot)^{-}$ is the Moore-Penrose generalized inverse. By definition, $I-Q$ is idempotent and satisfied $QP=P$ so that

$$\bar{x}'\left(I-Q\right)\bar{x}/n=\left(\bar{x}-P\gamma_K\right)'\left(I-Q\right)\left(\bar{x}-P\gamma_K\right)/n\leqslant\frac{\|\bar{x}-P\gamma_K\|^2}{n}\xrightarrow{p}0.$$

Also, by $\beta\ge 1/2$, the function $\lambda^{2\beta}/(\alpha+\lambda)$ is increasing in λ and reaches it maximum for the maximal eigenvalue (which is bounded by $\mathrm{tr}(P'P)/n\le C$) an therefore $\sup_{\lambda}\lambda^{2\beta}/(\alpha+\lambda^2)\le C$. Also, by the SVD, $\hat{\psi}_j=T_n\hat{\phi}_j/\sqrt{\lambda_j}=\frac{1}{\sqrt{\hat{\lambda}_j}}\left[P_1'\hat{\phi}_j,\cdots,P_n'\hat{\phi}_j\right]'/\sqrt{n}$ so that $\left(\bar{x}'\hat{\psi}_j\right)^2=\frac{\sqrt{n}}{\hat{\lambda}_j}\left(\hat{E}\left[\bar{x}_iP_i\right]'\hat{\phi}_j\right)^2$

where $\hat{E}\left[\bar{x}_i P_i\right] = \sum_{i=1}^{n} \bar{x}_i P_i/n$. It follows by $Q - Q^{\alpha} = \sum_{j=1}^{r} \frac{\alpha}{\hat{\lambda}_j + \alpha} \hat{\psi}_j \hat{\psi}_j'$ that

$$
\begin{aligned}
\bar{x}'\left(Q - Q^{\alpha}\right)\bar{x}/n &= \sum_{j=1}^{r} \frac{\alpha}{\alpha + \hat{\lambda}_j}\left(\bar{x}'\hat{\psi}_j\right)^2/n \\
&= \sum_{j=1}^{r} \frac{\alpha\hat{\lambda}_j^{2\beta}}{\alpha + \hat{\lambda}_j} \frac{\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2}{\hat{\lambda}_j^{2\beta+1}} \\
&\leqslant \sup_{\lambda}\left(\frac{\alpha\lambda^{2\beta}}{\alpha + \lambda}\right)\sum_{j=1}^{r} \frac{\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2}{\hat{\lambda}_j^{2\beta+1}} \\
&\leq C\alpha \sum_{j=1}^{r} \frac{\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2}{\hat{\lambda}_j^{2\beta+1}}. \qquad \text{(A.15)}
\end{aligned}
$$

At the limit, the sum in (A.15) is finite by hypothesis in the statement of Lemma 3 so that $\bar{x}'\left(Q - Q^{\alpha}\right)\bar{x}/n = O_p(\alpha)$. It follows that $\bar{x}'\left(I - Q^{\alpha}\right)\bar{x}/n = \bar{x}'\left(I - Q\right)\bar{x}/n + \hat{x}'\left(Q - Q^{\alpha}\right)\bar{x}/n \xrightarrow{p} 0$ as $n \to \infty$ and $\alpha \to 0$, giving the conclusion in (iv). □

Lemma 4. *If Assumption 1 is satisfied, (i)* $\hat{\beta} \xrightarrow{p} \bar{\beta}$, *(ii)* $a_i(\beta) \overset{def}{=} a\left(w_{i,}, \beta\right)$ *and* $b_i(\beta) \overset{def}{=} b\left(w_i, \beta\right)$ *are scalar functions that are continuous at* $\bar{\beta}$ *w.p.1 and there is a neighborhood* $\mathcal{N}$ *of* $\bar{\beta}$ *such that* $E\left[\sup_{\beta\in\mathcal{N}} |a_i(\beta)|^2\right] < \infty$ *and* $E\left[\sup_{\beta\in\mathcal{N}} |b_i(\beta)|^2\right] < \infty$, $E\left[a_i(\bar{\beta})^2|z_i\right]$ *and* $E\left[b_i(\bar{\beta})^2|z_i\right]$ *are bounded; (iii)* $U_i \overset{def}{=} U(z_i)$ *is a nonnegative scalar function bounded away from zero; (iv)* $K \to \infty$, $\alpha \to 0$, *and* $n\alpha \to \infty$ *as* $n \to \infty$, *then*

$$
\widehat{\Lambda}^{\alpha} \overset{def}{=} \left(\frac{1}{n}\sum_{i=1}^{n} a_i(\hat{\beta})q_i\right)'\left(\frac{1}{n}\sum_{i=1}^{n} U_i q_i q_i' + \alpha I\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} b_i(\hat{\beta})q_i\right) \xrightarrow{p} \Lambda,
$$

where $\Lambda \overset{def}{=} E\left[E\left[a_i(\bar{\beta})|z_i\right]U_i^{-1}E\left[b_i(\bar{\beta})|z_i\right]\right]$.

Proof. Let $P_i = q_i U_i^{1/2}$, $P = [P_1, \ldots, P_n]'$, $A_i(\beta) = U_i^{-1/2}a_i(\beta)$, $A(\beta) = [A_1(\beta), \cdots, A_n(\beta)]'$, $\hat{A} = A(\hat{\beta})$, $A = A(\bar{\beta})$, $B_i(\beta) = U_i^{-1/2}b_i(\beta)$, $B(\beta) = [B_1(\beta), \ldots, B_n(\beta)]'$, and $\hat{B} = B(\hat{\beta})$, and $B = B(\bar{\beta})$. Note that $\sum_{i=1}^{n} U_i q_i q_i' = P'P$, $\sum_{i=1}^{n} a_i(\hat{\beta})q_i' = \hat{A}'P$, and $\sum_{i=1}^{n} b_i(\hat{\beta})q_i = P\hat{B}$ so that for $Q^{\alpha} = P\left(P'P/n + \alpha I\right)P'/n$, $\hat{\Lambda}^{\alpha} = \hat{A}'P\left(P'P/n + \alpha I\right)^{-1}P'\hat{B}/n^2 = \hat{A}'Q^{\alpha}\hat{B}/n$.

Let $\Delta(w, \beta) = U(z)^{-1}[b(w, \beta) - b(w, \bar{\beta})]^2$. By hypothesis in the statement of Lemma 4, $\Delta(w, \beta)$ is continuous with respect β in a neighborhood $\mathcal{N}$ of $\bar{\beta}$. Moreover, $E\left[\sup_{\beta\in\mathcal{N}} |\Delta(w, \beta)|\right] \leq CE\left[\sup_{\beta\in\mathcal{N}} |b(w, \beta)|^2\right] < \infty$. It follows by Lemma 4.3 of Newey and McFadden (1994) with $a(z, \theta)$ there equal to $\Delta(w, \beta)$ that $\|\hat{B} - B\|^2/n = \sum_{i=1}^{n} \Delta\left(\omega_i, \hat{\beta}\right)/n \xrightarrow{p} E\left[\Delta\left(\omega_i, \bar{\beta}\right)\right] = 0$. Therefore by Lemma 3(ii)

$$
\hat{T}_B^{\alpha} \overset{def}{=} (\hat{B} - B)'Q^{\alpha}(\hat{B} - B)/n \leqslant \lambda_{\max}\left(Q^{\alpha}\right)\|\hat{B} - B\|^2/n \leq \|\hat{B} - B\|^2/n \xrightarrow{p} 0.
$$

For $Z = [z_1, \cdots, z_n]'$, let $a_i = a_i(\bar{\beta})$, $\bar{a}_i = E\left[a_i|z_i\right]$, and note that $\bar{A} \overset{\text{def}}{=} E[A|Z] = \left(U_1^{-1/2}\bar{a}_1, \cdots, U_n^{-1/2}\bar{a}_n\right)'$. Note that from i.i.d. observations,

$$
E\left[(A - \bar{A})(A - \bar{A})'|Z\right] = \operatorname{Diag}\left(U_1^{-1}V\left[a_1|z_1\right], \cdots, U_n^{-1}V\left[a_n|z_n\right]\right) \leq CI,
$$

by $E\left[|a_i|^2\,|z_i\right]$ bounded and U_i bounded away from zero.

Let Σ be the Cholesky factor of the symmetric and positive definite matrix $(P'P+\alpha I)^{-1}$. Then $Q^\alpha = P\Sigma'\Sigma P$. By the law of iterated expectations and Lemma 3(i) it follows for $\widetilde{T}_A^\alpha \overset{def}{=} (A–\bar{A})'Q^\alpha(A–\bar{A})/n$ that

$$\begin{aligned}
E\left[\widetilde{T}_A^\alpha\right] &= \operatorname{tr} E\left[E\left[\widetilde{T}_A^\alpha|Z\right]\right] \\
&= E\left[E\left[\operatorname{tr}\left((A-\bar{A})(A-\bar{A})'Q^\alpha\right)|Z\right]\right]/n \\
&= E\left[\operatorname{tr}\left(\Sigma P' E\left[(A-\bar{A})(A-\bar{A})'|Z\right]P\Sigma'\right)\right]/n \\
&\le CE\left[\operatorname{tr}\left(\Sigma P'P\Sigma'\right)\right]/n \\
&\le CE\left[\operatorname{tr}\left(P\Sigma'\Sigma P'\right)\right]/n \\
&\le CE\left[\operatorname{tr}(Q^\alpha)\right]/n \le C/(n\alpha) \to 0 \text{ as } n \to \infty.
\end{aligned}$$

It then follows by M that $\widetilde{T}_A^\alpha \xrightarrow{p} 0$. Also the same result holds for $\widetilde{T}_B^\alpha$.

By Assumption 1(b) there exists a $K \times 1$ vector, γ_K such that $E\left[\left\{U_i^{-1}\bar{a}_i - q_i'\gamma_K\right\}^2\right] \to 0$ as $K \to \infty$. Then by M

$$\begin{aligned}
\|\bar{A}-P\gamma_K\|^2/n &= \sum_{i=1}^n \left|U_i^{-1/2}\bar{a}_i - P_i'\gamma_K\right|^2/n \\
&= \sum_{i=1}^n U_i^{1/2}\left|U_i^{-1}\bar{a}_i - q_i'\gamma_K\right|^2/n \\
&\le C\sum_{i=1}^n \left|U_i^{-1}\bar{a}_i - q_i'\gamma_K\right|^2/n \xrightarrow{p} 0 \text{ as } n,K \to \infty.
\end{aligned}$$

Also by M $\bar{A}'\bar{A}/n = O_p(1)$. By Assumption 1(c), hypothesis in Lemma 3 is satisfied for $\bar{x} = \bar{A}$ so that part (iv) of Lemma 3 gives $\bar{T}_A^\alpha \overset{def}{=} \bar{A}'(I-Q^\alpha)\bar{A}/n \xrightarrow{p} 0$. The analogous result holds for B replacing A.

Next note that by CS

$$\begin{aligned}
T_A^\alpha &\overset{def}{=} (\hat{A}-\bar{A})'Q^\alpha(\hat{A}-\bar{A}) = (\hat{A}-A+A-\bar{A})'Q^\alpha(\hat{A}-A+A-\bar{A}) \\
&\le \widehat{T}_A^\alpha + \tilde{T}_A^\alpha + 2\sqrt{\hat{T}_A^\alpha}\sqrt{\tilde{T}_A^\alpha} \xrightarrow{p} 0.
\end{aligned}$$

Then by CS and T,

$$\begin{aligned}
|\hat{A}'Q^\alpha\hat{B}/n - \bar{A}'\bar{B}/n| =& |(\hat{A}-\bar{A})'Q^\alpha(\hat{B}-\bar{B}) + (\hat{A}-\bar{A})'Q^\alpha\bar{B} + \bar{A}'Q^\alpha(\hat{B}-\bar{B}) - \bar{A}'(I-Q^\alpha)\bar{B}|/n \\
\le& \sqrt{T_A^\alpha}\sqrt{T_B^\alpha} + \sqrt{T_A^\alpha}\sqrt{\bar{B}'\bar{B}/n} + \sqrt{\bar{A}'\bar{A}/n}\sqrt{T_A^\alpha} + \sqrt{\bar{T}_A^\alpha}\sqrt{\bar{T}_B^\alpha} \xrightarrow{p} 0.
\end{aligned}$$

Nothing that $\bar{A}'\bar{B}/n = \sum_{i=1}^n \bar{a}_i U_i^{-1}\bar{b}_i/n$ the conclusion follows by the standard law of large numbers. □

In the sequel we will use the following notations

$$\hat{R}(\beta) = \hat{g}(\beta)'\tilde{W}\hat{g}(\beta), \quad R(\beta) = E\left[(E\left[\rho(w,\beta)|z\right])^2\right], \tag{A.16}$$

where $\widetilde{W}^\alpha = \left(\widehat{A}+\alpha I\right)^{-1}$, $\hat{A} = \sum_{i=1}^n q_i q_i'/n$, and $q_i = q^K(z_i)/\zeta(K)$.

Lemma 5. *If Assumptions 1 and 2(a)-(c) are satisfied, $K \to \infty$, $\alpha \to 0$, and $n\alpha \to \infty$ as $n \to \infty$, then $R(\beta)$ has a unique minimun at β_0, $R(\beta)$ is continuous on $\mathscr{B}$ and $\sup_{\beta\in\mathscr{B}} |\hat{R}(\beta) - R(\beta)| \xrightarrow{\text{p}} 0$.*

Proof. For any $\beta \neq \beta_0$ it follows by Assumption 2(a) that $E[\rho(w,\beta)|z] \neq 0$ so that $R(\beta) = E\left[(E[\rho(w,\beta)|z])^2\right] > 0 = R(\beta_0)$ giving the first result.

To show the continuity of $R(\beta)$, note that by Assumption 2(c), for all $\beta, \tilde{\beta} \in \mathscr{B}$

$$
\begin{aligned}
\left|R(\tilde{\beta}) - R(\beta)\right| &= \left|E\left[\left(E[\rho(w,\tilde{\beta})|z]\right)^2 - (E[\rho(w,\beta)|z])^2\right]\right| \\
&\leq E\left[(E[\rho(w,\tilde{\beta}) - \rho(w,\beta)|z])^2\right] \\
&\leq E\left[E\left[\left(\rho(w,\tilde{\beta}) - \rho(w,\beta)\right)^2 \Big| z\right]\right] \\
&\leqslant E\left[\left(\rho(w,\tilde{\beta}) - \rho(w,\beta)\right)^2\right] \\
&\leqslant CE\left[\delta_1(w)^2\right] ||\tilde{\beta} - \beta||^{2r} \\
&\leq C||\tilde{\beta} - \beta||^{2r}.
\end{aligned}
$$

Therefore $R(\beta)$ is continuous being uniformly continuous.

To obtain the last conclusion, $\sup_{\beta\in B} |\hat{R}(\beta) - R(\beta)| \xrightarrow{P} 0$, it suffices, by Corollary 2.2 of Newey (1991), to show that: (i) $\mathscr{B}$ is compact (it is the case by Assumption 2(b)); (ii) $\widehat{R}(\beta) \xrightarrow{p} R(\beta)$ for all $\beta \in \mathscr{B}$; and (iii) there is $\hat{D} = O_p(1)$ with $\left|\hat{R}(\tilde{\beta}) - \hat{R}(\beta)\right| \leq \hat{D}||\tilde{\beta} - \beta||^r$ for all $\beta, \tilde{\beta} \in \mathscr{B}$.

To show (ii), apply Lemma 4 with $a(w,\beta) = b(w,\beta) = \rho(w,\beta)$ and $U(z) = 1$ with β fixed. Hypothesis in the statement of Lemma 4 are satisfied by Assumption 2(c). The conclusion of Lemma 4 implies that $\widehat{R}(\beta) \xrightarrow{p} R(\beta)$ for all $\beta \in \mathscr{B}$ giving (ii).

To show (iii), let $\rho = (\rho_1, \ldots, \rho_n)'$ and $\tilde{\rho} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_n)$ with $\rho_i = \rho\left(w_i, \beta\right)$ and $\tilde{\rho}_i = \rho\left(\omega_i, \tilde{\beta}\right)$. Also, note that $\hat{R}(\beta) = \rho' Q^\alpha \rho / n$ where Q^α is defined as in the statement of Lemma 3 for $U_i = 1$. It follows by Lemma 3(ii) and by CS that

$$
\begin{aligned}
\left|\hat{R}(\tilde{\beta}) - \hat{R}(\beta)\right| &= \left|\tilde{\rho}' Q^\alpha \tilde{\rho} - \rho' Q^\alpha \rho\right| / n \\
&= \left|(\tilde{\rho} - \rho)' Q^\alpha \tilde{\rho} + \rho' Q^\alpha (\tilde{\rho} - \rho)\right| / n \\
&\leq \lambda_{max}(Q^\alpha) ||\tilde{\rho} - \rho|| (||\tilde{\rho}|| + ||\rho||) / n \\
&\leqslant ||\tilde{\rho} - \rho|| (||\tilde{\rho}|| + ||\rho||) / n.
\end{aligned}
$$

Note by Assumption 2(c) and M that $n^{-1/2}||\tilde{\rho} - \rho|| = \left[\sum_{i=1}^n (\tilde{\rho}_i - \rho_i)^2 / n\right]^{1/2} \leq \hat{D}_{\delta_1} ||\tilde{\beta} - \beta||^r$, where $\hat{D}_{\delta_1} = \left[\sum_{i=1}^n \delta_1(w_i)^2 / n\right]^{1/2} = O_p(1)$. Also, for any fixed $\bar{\beta} \in \mathscr{B}$, by Assumption 2(b),

$$
\begin{aligned}
\sup_{\beta\in\mathscr{B}} |\rho(w_i, \beta)| / \sqrt{n} &\leq \sup_{\beta\in\mathscr{B}} \left|\rho(w_i, \beta) - \rho\left(w_i, \bar{\beta}\right)\right| / \sqrt{n} + \bar{D} \\
&\leq \hat{D}_{\delta_1} \sup_{\beta\in\mathscr{B}} ||\beta - \bar{\beta}||^r + \bar{D} \leq C\hat{D}_{\delta_1} + \bar{D},
\end{aligned}
$$

where $\bar{D} = \left[\sum_{i=1}^n \rho\left(w_i, \bar{\beta}\right)^2 / n\right]^{1/2} = O_p(1)$ by Assumption 2(c) and M. Therefore $\left|\hat{R}(\tilde{\beta}) - \hat{R}(\beta)\right| \leq \hat{D}||\tilde{\beta} - \beta||^r$, where $\hat{D} = 2\hat{D}_{\delta_1}\left(C\hat{D}_{\delta_1} + \bar{D}\right) = O_p(1)$, giving (iii). □

Lemma 6. *If Assumptions 1(a) and 2(f) are satisfied, for $\Lambda_n = \{\lambda : ||\lambda|| \leq \delta_n\}$, where δ_n is a sequence of nonnegative real numbers such that $\delta_n n^{1/\gamma} \to 0$ as n goes to infinity, then we have*

$\max_{\substack{\beta\in\mathscr{B},\lambda\in\Lambda_n\\ 1\le i\le n}} |\lambda' g_i(\beta)| \xrightarrow{p} 0$ *and w.p.a.1,* $\Lambda_n \subseteq \hat\Lambda(\beta)$ *for all* $\beta\in\mathscr{B}$.

Proof. Let $b_i = \sup_{\beta\in\mathscr{B}} |\rho\left(w_i,\beta\right)|$. By Assumption 2(f), $E\left[b_i^\gamma\right] < C$. If follows by M that

$$\max_{1\le i\le n} b_i = \left\{\max_i b_i^\gamma\right\}^{1/\gamma} \le \left\{\sum_{i=1}^n b_i^\gamma\right\}^{1/\gamma} = n^{1/\gamma}\left\{\sum_{i=1}^n b_i^\gamma/n\right\}^{1/\gamma} \le n^{1/\gamma} O_p\left(\left\{E\left[b_i^\gamma\right]\right\}^{1/\gamma}\right) = O_p\left(n^{1/\gamma}\right).$$

It then follows by Assumption 1(a) that

$$X_n \overset{def}{=} \max_{\substack{\beta\in\mathscr{B},\lambda\in\Lambda_n\\ 1\le i\le n}} \left|\lambda' g_i(\beta)\right| = \max_{\substack{\beta\in\mathscr{B},\lambda\in\Lambda_n\\ 1\le i\le n}} \left|\lambda' q_i \rho(w_i,\beta)\right| \le \delta_n \max_{1\le i\le n} b_i = \delta_n O_p\left(n^{1/\gamma}\right) \xrightarrow{p} 0,$$

giving the first conclusion.

Also, since $\mathscr{V}$ is a neighborhood of 0, by the first conclusion $X_n \in \mathscr{V}$ w.p.a.1. Equivalently, $|\lambda' g_i(\beta)| \in \mathscr{V}$ for all $\beta\in\mathscr{B}$, $\lambda\in\Lambda_n$, and $i = 1,\ldots,n$. It follows that $\Lambda_n \subseteq \hat\Lambda(\beta)$ for all $\beta\in\mathscr{B}$, giving the second conclusion. □

Lemma 7. *If Assumptions 1(a) and 2(f) are satisfied,* δ_n *a sequence of nonnegative real numbers such that* $\delta_n n^{1/\gamma} \to 0$, a_n *a sequence of real numbers such that* $\alpha\delta_n a_n \to \infty$ *as n goes to infinity,* $\tilde\beta$ *an estimator of* β_0 *with* $\left\|\hat g\left(\tilde\beta\right)\right\| = O_p\left(a_n^{-1}\right)$ *then* $\sup_{\lambda\in\hat\Lambda(\tilde\beta)} \hat P(\tilde\beta,\lambda) = O_p\left(\alpha^{-1}a_n^{-2}\right)$, $\tilde\lambda = \arg\min_{\lambda\in\hat\Lambda(\tilde\beta)} \hat P(\tilde\beta,\lambda)$ *exists w.p.a.1 and* $\left\|\tilde\lambda\right\| = O_p\left(\alpha^{-1}a_n^{-1}\right)$.

Proof. Let $\tilde g = \hat g(\tilde\beta)$ and $\tilde\Omega = \hat\Omega(\tilde\beta)$. Also let Λ_n be as defined in Lemma 6. It is obvious that $\hat P(\beta,\lambda)$ is twice continuously differentiable on Λ_n (as it is a quadratic function of λ). Then by compacity of Λ_n, $\bar\lambda \overset{def}{=} \arg\max_{\lambda\in\Lambda_n} \hat P(\tilde\beta,\lambda)$ exists. Furthermore, by $\lambda_{\min}\left(\tilde\Omega + \alpha I\right) \ge \alpha$, the following inequalities hold

$$0 = \hat P\left(\tilde\beta,0\right) \le \hat P\left(\tilde\beta,\bar\lambda\right) = -\bar\lambda'\tilde g - \frac{1}{2}\bar\lambda'\left(\tilde\Omega + \alpha I\right)\bar\lambda \le \left\|\bar\lambda\right\| \|\tilde g\| - \alpha C \left\|\bar\lambda\right\|^2. \tag{A.17}$$

Adding $\alpha C\left\|\bar\lambda\right\|^2$ from both sides and dividing by $C\left\|\bar\lambda\right\|$ we find that $\alpha\left\|\bar\lambda\right\| \leqslant C\|\tilde g\|$. Then by the hypothesis in the statement of Lemma 7, $\left\|\bar\lambda\right\| = O_p\left(\alpha^{-1}a_n^{-1}\right) = \delta_n O_p\left(\alpha^{-1}\delta_n^{-1}a_n^{-1}\right) = \delta_n o_p(1)$. Therefore $\lim_{n\to\infty} P\left(\left\|\bar\lambda\right\| < \delta_n\right) = 1$ and then $\bar\lambda \in \operatorname{int}(\Lambda_n)$ w.p.a.1. It follows that $\bar\lambda = \arg\max_{\lambda\in\Lambda_n} \hat P(\tilde\beta,\lambda)$ satisfies the first order conditions, $\partial\hat P\left(\tilde\beta,\lambda\right)/\partial\lambda\big|_{\lambda=\bar\lambda} = 0$. By Lemma 6 $\bar\lambda \in \Lambda_n \subseteq \hat\Lambda(\tilde\beta)$ w.p.a.1. Then by concavity of $\hat P\left(\tilde\beta,\lambda\right)$ with respect of λ, and convexity of $\hat\Lambda(\tilde\beta)$ it follows that $\hat P\left(\tilde\beta,\bar\lambda\right) = \max_{\lambda\in\widehat\Lambda(\tilde\beta)} \hat P(\tilde\beta,\lambda)$, giving the second and the third conclusions with $\tilde\lambda = \bar\lambda$. The last inequality of Eq. (A.17) gives

$$\hat P\left(\tilde\beta,\tilde\lambda\right) \le \left\|\tilde\lambda\right\| \|\tilde g\| - \alpha C\|\tilde\lambda\|^2 \leqslant O_p\left(\frac{1}{\alpha a_n}\right) O_p\left(\frac{1}{a_n}\right) - \alpha O_p\left(\frac{1}{\alpha^2 a_n^2}\right) = O_p\left(\alpha^{-1}a_n^{-2}\right),$$

giving the first result. □

Lemma 8. *If Assumptions 1(a) and 2(f) are satisfied,* $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$ *as* $n\to\infty$, *where* $\varepsilon > 0$ *is such that* $1/2 - 1/\gamma - \varepsilon > 0$, *then for any* $\bar\lambda \in \hat\Lambda\left(\hat\beta\right)$ *it is the case that w.p.a.1* $\hat P\left(\hat\beta,\bar\lambda\right) \leqslant \sup_{\lambda\in\hat\Lambda(\hat\beta)} \hat P\left(\hat\beta,\lambda\right) = O_p\left(\alpha^{-1}n^{-1}\right)$.

Proof. The inequality is obvious by the fact that $\bar{\lambda} \in \hat{\Lambda}\left(\hat{\beta}\right)$. Let $a_n = n^{1/2}$, then $\|\hat{g}\left(\beta_0\right)\| = O_p\left(a_n^{-1}\right)$ by Lemma 2. For $\delta_n = n^{-1/\gamma-\varepsilon}$, $\delta_n n^{1/\gamma} = n^{-\varepsilon} \to 0$, and $\alpha\delta_n a_n = \alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$, so that the hypotheses in the statement of Lemma 7 is satisfied for $\tilde{\beta} = \beta_0$. The conclusion of Lemma 7 gives $\sup_{\lambda\in\hat{\Lambda}(\beta_0)} \hat{P}\left(\beta_0, \lambda\right) = O_p\left(\alpha^{-1}n^{-1}\right)$. By Theorem 3.1,

$$\sup_{\lambda\in\hat{\Lambda}(\hat{\beta})} \hat{P}(\hat{\beta},\lambda) \leq \sup_{\lambda\in\hat{\Lambda}(\beta_0)} \hat{P}\left(\beta_0,\lambda\right) = O_p\left(\alpha^{-1}n^{-1}\right),$$

giving the second result. □

We need the following notations for the next result. Let $g_i = g_i\left(\beta_0\right)$, $\Omega = E\left[g_i g_i'\right]$, $\hat{\Omega}(\beta) = \sum_{i=1}^n g_i(\beta)g_i(\beta)'/n$, $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, $\tilde{\Omega} = \sum_{i=1}^n g_i g_i'/n$ and $\bar{\Omega} = n^{-1}\sum_{i=1}^n \sigma_i^2 q_i q_i'$.

Lemma 9. *If Assumptions 1(a) and 2(b)-(e) are satisfied then for any $\hat{\beta} \in \mathscr{B}$*

(i) If $\alpha \to 0$ as $n \to 0$ then for n large enough $\lambda_{\max}\left(\hat{\Omega}+\alpha I\right) \leq C$ w.p.a.1;

(ii) If $\hat{\beta} = \beta_0 + O_p\left(\tau_n\right)$ with $\tau_n \to 0$, then

$$\left\|\hat{\Omega}-\tilde{\Omega}\right\| = O_p\left(\tau_n\right), \quad \left\|\tilde{\Omega}-\bar{\Omega}\right\| = O_p\left(n^{-1/2}\right), \quad \textit{and} \quad \left\|\bar{\Omega}-\Omega\right\| = O_p\left(n^{-1/2}\right).$$

Moreover $\lambda_{\max}\left(\Omega\right) \leq C$ and w.p.a.1 $\lambda_{\max}\left(\bar{\Omega}\right) \leq C$, and $\lambda_{\max}\left(\hat{\Omega}\right) \leq C$. If in addition $\left\|\tilde{\lambda}\right\| = O_p\left(\kappa_n\right)$ then for $\check{\Omega} = -\sum_{i=1}^n s_1\left(\tilde{\lambda}'g_i\right)g_i g_i'/n$, we have $\left\|\check{\Omega}-\bar{\Omega}\right\| = O_p\left(\kappa_n+\tau_n+n^{-1/2}\right)$.

Proof. Let $\hat{\rho}_i = \rho\left(w_i,\hat{\beta}\right)$ and $\rho_i = \rho\left(w_i,\beta_0\right)$. For $b_i = \sup_{\beta\in\mathscr{B}}\|\rho\left(w_i,\beta\right)\|$ we have $\hat{\Omega} \leqslant \sum_{i=1}^n b_i^2 q_i q_i'/n \overset{def}{=} \dot{\Omega}$. Also, by Assumptions 1(a) and 2(d),

$$\begin{aligned}
E\left[\left\|\dot{\Omega}-E\left[\dot{\Omega}\right]\right\|^2\right] &= E\left[\left\|\sum_{i=1}^n b_i^2 q_i q_i'/n - E\left[b_i^2 q_i q_i'\right]\right\|^2\right] \\
&= \operatorname{tr} E\left[\left(b_i^2 q_i q_i' - E\left[b_i^2 q_i q_i'\right]\right)^2\right]/n \\
&\leq \operatorname{tr} E\left[b_i^4\left\{q_i q_i'\right\}^2\right]/n \\
&\leqslant \operatorname{tr} E\left[E\left[b_i^4|z_i\right]\left\{q_i q_i'\right\}^2\right]/n \\
&\leqslant \operatorname{tr} E\left[\|q_i\|^4\right]/n \leqslant C/n.
\end{aligned}$$

It follows by M that $\left\|\dot{\Omega}-E\left[\dot{\Omega}\right]\right\| = O_p(n^{-1/2})$. Also, by Assumptions 1(a) and 2(c),

$$\begin{aligned}
\lambda_{\max}\left(E\left[\dot{\Omega}\right]\right) &= \lambda_{\max}\left(E\left[b_i^2 q_i q_i'\right]\right) \\
&= \lambda_{\max}\left(E\left[E\left[b_i^2|z_i\right]q_i q_i'\right]\right) \\
&\leqslant C \operatorname{tr} E\left[q_i q_i'\right] \\
&\leq CE\left[\|q_i\|^2\right] \leq C.
\end{aligned}$$

It follows that $\left\|\lambda_{\max}\left(\dot{\Omega}\right)-\lambda_{\max}\left(E\left[\dot{\Omega}\right]\right)\right\| \leqslant \left\|\dot{\Omega}-E\left[\dot{\Omega}\right]\right\| \xrightarrow{p} 0$ and therefore $\lambda_{\max}\left(\dot{\Omega}\right) \leq C$ w.p.a.1. Moreover $\alpha \to 0$ as $n \to \infty$, then $\alpha \leqslant C$ for n large enough; giving the result (i) by $\hat{\Omega} \leq \dot{\Omega}$.

Also, by $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{\beta} \in N$ w.p.a.1 so that by Assumption 2(d) $|\hat{\rho}_i - \rho_i| \le \delta_i \left\|\hat{\beta} - \beta_0\right\|$ for all $i = 1, \dots, n$ w.p.a.1, where $\delta_i = \delta_2(w_i)$. $M_i = \delta_i^2 + 2\delta_i \|\rho_i\|$ has $E[M_i|z_i]$ bounded by CS and Assumption 2(d) so that $E\left[M_i \|q_i\|^2\right] = E\left[E[M_i|z_i] \|q_i\|^2\right] \le C$. It follows by Assumption 2(b), CS and M that

$$\begin{aligned}
\left\|\hat{\Omega} - \tilde{\Omega}\right\| &= \left\| n^{-1} \sum_{i=1}^{n} \left(\hat{\rho}_i^2 - \rho_i^2\right) q_i q_i' \right\| \\
&\leqslant n^{-1} \sum_{i=1}^{n} \left|\hat{\rho}_i^2 - \rho_i^2\right| \|q_i\|^2 \\
&\leqslant n^{-1} \sum_{i=1}^{n} \left[(\hat{\rho}_i - \rho_i)^2 + 2|\hat{\rho}_i - \rho_i||\rho_i|\right] \|q_i\|^2 \\
&\leqslant n^{-1} \sum_{i=1}^{n} \left[\delta_i^2 \|\hat{\beta} - \beta\|^2 + 2\delta_i \, |\rho_i| \left\|\hat{\beta} - \beta_0\right\|\right] \|q_i\|^2 \\
&\leqslant \left\|\hat{\beta} - \beta_0\right\| \sum_{i=1}^{n} M_i \|q_i\|^2 / n \\
&\le O_p(\tau_n) O_p\left(E\left[M_i \|q_i\|^2\right]\right) = O_p(\tau_n),
\end{aligned}$$

giving the first result in (ii).

Note that by Assumptions 1(a) and 2(d),

$$\begin{aligned}
E\left[\left\|\tilde{\Omega} - \bar{\Omega}\right\|^2\right] &= E\left[\left\| \sum_{i=1}^{n} \left(\rho_i^2 - \sigma_i^2\right) q_i q_i' / n \right\|^2\right] \\
&= \operatorname{tr} E\left[\left(\rho_i^2 - \sigma_i^2\right)^2 \left\{q_i q_i'\right\}^2\right] / n \\
&\le \operatorname{tr} E\left[\rho_i^4 \left\{q_i q_i'\right\}^2\right] / n \\
&\leqslant \operatorname{tr} E\left[E\left[\rho_i^4 | z_i\right] \{q_i q_i'\}^2\right] / n \\
&\le C E\left[\|q_i\|^4\right] / n \le C/n,
\end{aligned}$$

so that the second result in (ii) follows by M.

The third result follows by M and

$$\begin{aligned}
E\left[\left\|\bar{\Omega} - \Omega\right\|^2\right] &= E\left[\left\| \sum_{i=1}^{n} \sigma_i^2 q_i q_i' / n - E\left[\rho_i^2 q_i q_i'\right] \right\|^2\right] \\
&= \operatorname{tr} E\left[\left(\sigma_i^2 q_i q_i' - E\left[\rho_i^2 q_i q_i'\right]\right)^2\right] / n \\
&\leqslant \operatorname{tr} E\left[\sigma_i^4 \left\{q_i q_i'\right\}^2\right] / n \\
&\le \operatorname{tr} E\left[E\left[\rho_i^2 | z_i\right]^2 \left\{q_i q_i'\right\}^2\right] / n \\
&\le C E\left[\|q_i\|^4\right] / n \le C/n.
\end{aligned}$$

As in the proof of (i), $\lambda_{\max}(\Omega) \le C$ by Assumptions 1(a) and 2(c). Similarly to the proof of (i), it follows by $|\lambda_{\max}(A) - \lambda_{\max}(B)| \le \|A - B\|$ that w.p.a.1 $\lambda_{\max}\left(\bar{\Omega}\right) \le C$ and $\lambda_{\max}\left(\hat{\Omega}\right) \le C$.

It remains to show that $\left\|\check{\Omega} - \bar{\Omega}\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right)$. Given previous results, it will be sufficient to show that $\left\|\breve{\Omega}-\hat{\Omega}\right\| = O_p(\kappa_n)$. Note that $s_1(v) = -(1+v)$ so that by Assumptions 1(a) and 2(d), and CS,

$$\begin{aligned}
\left\|\breve{\Omega}-\hat{\Omega}\right\| &= \left\| n^{-1}\sum_{i=1}^{n}\left(\lambda'\hat{g}_i\right)\hat{b}_i^2 q_i q_i' \right\| \\
&\leqslant \left(\sum_{i=1}^{n}\left|\tilde{\lambda}'\hat{g}_i\right|^2/n\right)^{1/2}\left(\sum_{i=1}^{n} b_i^4 \|q_i\|^4/n\right)^{1/2} \\
&\leqslant \sqrt{\tilde{\lambda}'\hat{\Omega}\tilde{\lambda}}\sqrt{\sum_{i=1}^{n} b_i^4\,\|q_i\|^4/n} \\
&\leq \lambda_{\max}\left(\hat{\Omega}\right)\|\tilde{\lambda}\| O_p\left(\left\{E\left[E\left[b_i^4|z_i\right]\|q_i\|^4\right]\right\}^{1/2}\right) \\
&\leq C\|\tilde{\lambda}\|O_p(1) = O_p(\kappa_n).
\end{aligned}$$

□

Lemma 10. *If Assumptions 1 and 2 are satisfied, $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$, where $\varepsilon > 0$ is such that $1/2 - 1/\gamma - \varepsilon > 0$, then $\left\|\hat{g}\left(\hat{\beta}\right)\right\| = O_p\left((n\alpha)^{-1/2}\right)$.*

Proof. Let $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, $\hat{g} = \hat{g}(\hat{\beta})$, and Λ_n be as defined in Lemma 6. Also, let $\delta_n = n^{-1/\gamma-\varepsilon}$ and $\bar{\lambda} = -\delta_n \hat{g}/\|\hat{g}\|$ so that $\bar{\lambda}'\hat{g} = -\delta_n\|\hat{g}\|$ and $\|\bar{\lambda}\| = \delta_n$. Then $\bar{\lambda} \in \Lambda_n$ and by Lemma 6 $\bar{\lambda} \in \Lambda_n \subseteq \hat{\Lambda}(\widehat{\beta})$ w.p.a.1. Also, by Lemma 9(i), $\lambda_{\max}\left(\hat{\Omega}+\alpha I\right) \leq C$ so that Lemma 8 applied to $\bar{\lambda}$ gives

$$O_p(1/(\alpha n)) = \hat{P}(\hat{\beta},\bar{\lambda}) = -\bar{\lambda}'\hat{g} - \frac{1}{2}\bar{\lambda}'\left(\hat{\Omega}+\alpha I\right)\bar{\lambda} \geqslant \delta_n\|\hat{g}\| - C\delta_n^2,$$

or equivalently $\delta_n\|\hat{g}\| - C\delta_n^2 \leq O_p(1/(n\alpha))$. Adding $C\delta_n^2$ from both sides and dividing by δ_n gives

$$\|\hat{g}\| \leq O_p(1/(n\alpha\delta_n)) + C\delta_n = \frac{1}{\alpha n^{\frac{1}{2}-\frac{1}{\gamma}-\varepsilon} n^{\frac{1}{2}}} O_p(1) + C\delta_n = o(1)O_p(1) + C\delta_n = O_p(\delta_n).$$

Now consider any $\varepsilon_n \to 0$. Let $\tilde{\lambda} = -\varepsilon_n\hat{g}$. Then $\|\tilde{\lambda}\| = |\varepsilon_n|\,\|\hat{g}\| = o(1)O_p(\delta_n)$ so that w.p.a.1. $\tilde{\lambda} \in \Lambda_n \subseteq \hat{\Lambda}(\hat{\beta})$ by Lemma 6. For n large enough,

$$\hat{P}(\hat{\beta},\tilde{\lambda}) \geqslant -\tilde{\lambda}'\hat{g} - C\|\tilde{\lambda}\|^2 = \left(\varepsilon_n - C\varepsilon_n^2\right)\|\hat{g}\|^2 \geqslant \|\hat{g}\|^2\varepsilon_n/2.$$

It then follows by Lemma 8 $\|\hat{g}\|^2\varepsilon_n = O_p(1/(n\alpha))$. Since ε_n is any sequence converging to zero, it follows that $\|\hat{g}\|^2 = O_p(1/(n\alpha))$ giving the result. □

Proof of Theorem 4.1. By $\hat{A} \overset{def}{=} n^{-1}\sum_{i=1}^{n} q_i q_i'$ being positive semidefinite, $\hat{A}+\alpha I \geqslant \alpha I$ so that $\lambda_{\min}(\hat{A}+\alpha I) \geqslant \alpha$ and therefore $\lambda_{\max}\left(\tilde{W}\right) \leqslant 1/\alpha$ for $\tilde{W} = (\hat{A}+\alpha I)^{-1}$. By CS and Lemma 10, $\hat{R}(\hat{\beta}) = \hat{g}\tilde{W}\hat{g} \leqslant \alpha^{-1}\|\hat{g}\|^2 = O_p\left(\alpha^{-2}n^{-1}\right) \xrightarrow{p} 0$ since $\alpha^2 n = \left(\alpha n^{\frac{1}{2}-\frac{1}{\gamma}-\varepsilon}\right)^2 n^{\frac{2}{\gamma}+2\varepsilon} \to \infty$ as $n \to \infty$. It follows that $\hat{R}(\hat{\beta}) \xrightarrow{p} 0 = R(\beta_0)$. By Lemma 5, hypotheses in the statement of Lemma 1 are satisfied. The conclusion then follows from Lemma 1. □

Lemma 11. *If Assumptions 1, 2, and 3 are satisfied, $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$, where $\varepsilon > 0$ is such that $1/2 - 1/\gamma - \varepsilon > 0$, then $\hat{\beta} = \beta_0 + O_p\left((\alpha\sqrt{n})^{-1}\right)$.*

Proof. For notational convenience, let $\hat\beta = \hat\beta$. By an expansion $\hat g = \bar g + \dot G\left(\hat\beta - \beta_0\right)$ for $\dot G = n^{-1}\sum_{i=1}^n q_i\rho_\beta\left(w_i, \bar\beta\right)$, where $\bar\beta$ is on the line joining $\hat\beta$ and β_0. Therefore,

$$\begin{aligned}\hat R(\hat\beta) \overset{def}{=} \hat g'\tilde W\hat g &= \left(\bar g + \dot G\left(\hat\beta - \beta_0\right)\right)'\tilde W\left(\bar g + \dot G\left(\hat\beta + \hat\beta_0\right)\right)\\ &= \hat R\left(\beta_0\right) + 2\bar g'\tilde W\dot G\left(\hat\beta - \beta_0\right) + \hat D^2\end{aligned}$$

where $\hat D = \left[\left(\hat\beta - \beta_0\right)\dot G'\tilde W\dot G\left(\hat\beta - \beta_0\right)\right]^{1/2}$. Then for $\hat F = \left[\hat R(\hat\beta) + \hat R\left(\beta_0\right)\right]^{1/2}$, it follows by T, CS and $\hat R\left(\beta_0\right)^{1/2} \leq \hat F, \hat D \geqslant 0$ that

$$\begin{aligned}\hat D^2 = \hat R(\hat\beta) - \hat R\left(\beta_0\right) - 2\bar g'\tilde W\dot G\left(\hat\beta - \beta_0\right) &= \left|\hat R(\hat\beta) - \hat R\left(\beta_0\right) - 2\bar g'\tilde W\dot G\left(\hat\beta - \beta_0\right)\right|\\ \leq \hat R(\hat\beta) + \hat R\left(\beta_0\right) + 2\left|g'\tilde W\dot G\left(\hat\beta - \beta_0\right)\right| &\leq \hat R(\hat\beta) + \hat R\left(\beta_0\right) + 2\hat R\left(\beta_0\right)^{1/2}\hat D \leq \hat F^2 + 2\hat F\hat D.\end{aligned}$$

Subtracting $2\hat F\hat D$, adding $\hat F^2$ from both sides and the taking square roots gives $|\hat D - \hat F| \leq \sqrt{2}\hat F$. Also, by T, $|\hat D - \hat F| \geqslant \hat D - \hat F$, so that $\hat D \leq (\sqrt{2} + 1)\hat F = C\hat F$. By Lemma 2, $\|\bar g\| = O_p\left(n^{-1/2}\right)$ and by Lemma 10 $\|\hat g\| = O_p\left((n\alpha)^{-1/2}\right)$. Also, as in the proof of Theorem 4.1 $\lambda_{\max}\left(\tilde W\right) \leq 1/\alpha$ w.p.a.1 so that by T

$$\hat F^2 \leq \hat R(\hat\beta) + \hat R\left(\beta_0\right) \leq \frac{1}{\alpha}\left(\|\hat g\|^2 + \|\bar g\|^2\right) \leq \frac{1}{\alpha}\left(O_p(1/n) + O_p\left(1/(\alpha n)\right)\right) = O_p\left(1/\alpha^2 n\right).$$

Also, Lemma 4 applied to $\hat\beta = \bar\beta, \bar\beta = \beta_0, U(z) = 1$, $a(w,\beta) = \partial\rho(w,\beta)/\partial\beta$, and $b(w,\beta) = \partial\rho(w,\beta)/\partial\beta_l$ for $k, l = 1, \ldots, p$ gives $\left(\dot G'\tilde W\dot G\right)_{kl} \xrightarrow{p} E\left[(D(z)'D(z))_{kl}\right]$ for all k, l so that $\dot G'\tilde W\dot G \xrightarrow{p} E\left[D(z)'D(z)\right]$ which is non singular by Assumption 3(d). It then follows that $\lambda_{\min}\left(\dot G'\tilde W\dot G\right) \geqslant C$ w.p.a.1 and then $\hat D^2 = \left(\hat\beta - \beta_0\right)\dot G'\tilde W\dot G\left(\hat\beta - \beta_0\right) \geq C\left\|\hat\beta - \beta_0\right\|$. Therefore, $C\left\|\hat\beta - \beta_0\right\|^2 \leq \hat D^2 \leq C\hat F^2 = O_p\left(1/(n\alpha^2)\right)$, giving the result. □

Some useful notations are needed for the next result. Let $D_i = D\left(z_i\right)$, $\hat G = n^{-1}\sum_{i=1}^n q_i\rho_\beta\left(w_i, \hat\beta\right)$, $\bar G = n^{-1}\sum_{i=1}^n q_i D_i$, $G = E\left[q_i D_i\right]$, and $\tilde G = n^{-1}\sum_{i=1}^n q_i\rho_\beta\left(\omega_i, \beta_0\right)$.

Lemma 12. *If Assumptions 1, 2(b)-(e), and 3(b)-(c) are satisfied and $\hat\beta = \beta_0 + O_p\left(\tau_n\right)$ with $\tau_n \to 0$, then*

(i) $\left\|\hat G - \bar G\right\| = O_p\left(\tau_n + n^{-1/2}\right)$ *and* $\left\|\bar G - G\right\| = O_p\left(n^{-1/2}\right)$;

(ii) If in addition $\alpha \to 0$ and $\alpha n \to \infty$ as $n \to \infty$, then $\left\|\bar G'\left(\bar\Omega + \alpha I\right)^{-1}\left(\widehat\Omega - \bar\Omega\right)\right\| = O_p\left(\tau_n + n^{-1/2}\right)$;

(iii) For $\|\tilde\lambda\| = O_p\left(\kappa_n\right)$ and $\check G = -n^{-1}\sum_{i=1}^n s_1\left(\tilde\lambda'\hat g_i\right)\partial g_i(\hat\beta)/\partial\beta'$ then $\left\|\check G - \hat G\right\| = O_p\left(\kappa_n\right)$ *and* $\left\|\check G - \bar G\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right)$.

Proof. Let $\rho_{\beta i} = \rho_\beta\left(w_i, \beta_0\right)$, then $E\left[\rho_{\beta i}\right] = E[D_i]$ by iterated expectation. Also, by Assumption

3(c)

$$\begin{aligned}
E\left[\left\|\tilde{G}-\bar{G}\right\|^2\right] &= E\left[\left\|n^{-1}\sum_{i=1}^{n} q_i\left(\rho_{\beta_i}-D_i\right)\right\|^2\right] \\
&= n^{-1}\operatorname{tr} E\left[\left(\rho_{\beta_i}-D_i\right)'\left(\rho_{\beta i}-D_i\right)q_i'q_i\right] \\
&\le n^{-1}\operatorname{tr} E\left\{E\left[\rho_{\beta i}'\rho_{\beta i}|z_i\right]\|q_i\|^2\right\} \\
&\le n^{-1}CE\left[E\left[\left\|\rho_{\beta i}\right\|^2|z_i\right]\|q_i\|^2\right] \le C/n.
\end{aligned}$$

It follows by M that $\left\|\tilde{G}-\bar{G}\right\| = O_p\left(n^{-1/2}\right)$.

Also, by the mean value theorem for vector-valued functions, $\left\|\rho_\beta(w,\beta)-\rho_\beta\left(w,\beta_0\right)\right\| \le \delta_3\left(w\right)\|\beta-\beta_0\|$, for all $\beta\in\mathscr{N}$, where $\delta_3(w)=\sup_{\beta\in\mathscr{N}}\left\|\rho_{\beta\beta}(w,\beta)\right\|$ with $E\left[\delta_3(w)\right]$ bounded by Assumption 3(c). For $\hat{\rho}_{\beta i}=\rho_\beta\left(w_i,\hat{\beta}\right)$, it follows by T, CS, and M that

$$\begin{aligned}
\left\|\hat{G}-\tilde{G}\right\| &= \left\|n^{-1}\sum_{i=1}^{n} q_i\left(\hat{\rho}_{\beta i}-\rho_{\beta i}\right)\right\| \\
&\le n^{-1}\sum_{i=1}^{n}\left\|\hat{\rho}_{\beta i}-\rho_{\beta i}\right\|\|q_i\| \\
&\le C\left\|\hat{\beta}-\beta_0\right\|\sum_{i=1}^{n}\delta_3\left(w_i\right)/n = O_p\left(\tau_n\right)O_p\left(E\left[\delta_3\left(w_i\right)\right]\right) \le O_p\left(\tau_n\right).
\end{aligned}$$

It then follows by T that $\left\|\hat{G}-\bar{G}\right\| \leqslant \left\|\hat{G}-\tilde{G}\right\| + \left\|\tilde{G}-\bar{G}\right\| = O_p\left(\tau_n+n^{-1/2}\right)$, giving the first result in (i).

Also, by the Jensen's Inequality and Assumption 3(c), $\|D_i\|^2=\left\|E\left[\rho_{\beta i}|z_i\right]\right\|^2 \le E\left[\left\|\rho_{\beta i}\right\|^2|z_i\right] \leqslant C$. The second conclusion in (i), then follows by M and

$$E\left[\left\|\bar{G}-G\right\|^2\right] = E\left[\left\|n^{-1}\sum_{i=1}^{n} q_iD_i-G\right\|^2\right] \le n^{-1}E\left[\|D_i\|^2\|q_i\|^2\right] \le C/n.$$

For the proof of (ii) let $D(z)=\left[D^1(z),\cdots,D^p(z)\right]$ where $D^k(z)=E\left[\partial\rho\left(w,\beta_0\right)/\partial\beta|z\right]$. Then by Assumptions 2(e) and 3(c), the hypothesis in the statement of Lemma 4 are satisfied for $a(w,\beta)=D^k(z), b(w,\beta)=D^l(z)$ and $U(z)=\sigma(z)^2$, for $k,l=1,\cdots,p$. It follows by the conclusion of Lemma 4 that $\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right)_{kl} \xrightarrow{p} E\left[\sigma(z)^{-2}\left(D(z)'D(z)\right)_{kl}\right]$ for all $k,l=1,\ldots,p$ so that $\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}$ $\xrightarrow{p} E\left[\sigma(z)^{-2}D(z)'D(z)\right]$ and hence $\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}=O_p(1)$. Let $H_i=\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}q_i$. Then by Σ_i bounded away from zero and $\bar{\Omega}\le\bar{\Omega}+\alpha I$,

$$\begin{aligned}
\sum_{i=1}^{n}\|H_i\|^2/n &= \operatorname{tr}\left(\sum_{i=1}^{n}H_iH_i'/n\right) = \operatorname{tr}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}q_iq_i'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right) \\
&\le \operatorname{tr}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{\Omega}\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right) \\
&\le \operatorname{tr}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right) = O_p\left(1\right).
\end{aligned}$$

Next, let $M_i = \delta_i^2 + 2\delta_i\left|\rho_i\right|$, where $\delta_i = \delta_2\left(w_i\right)$ and $\rho_i = \rho\left(w_i,\beta_0\right)$. Also, let $Z = (z_1,\cdots,z_n)$. It is well known that if $E\left[\hat{R}_n|Z\right] = O_p\left(\nu_n\right)$ for some ν_n then $\hat{R}_n = O_p\left(\nu_n\right)$. For $\hat{R}_n = \sum_{i=1}^n M_i\|H_i\|\|q_i\|/n$, it follows by CS and M that

$$\begin{aligned} E\left[\hat{R}_n|Z\right] &= \sum_{i=1}^n E\left[M_i|Z\right]\|H_i\|\,\|q_i\|/n \le C\sum_{i=1}^n \|H_i\|\,\|q_i\|/n \\ &\le C\left(\sum_{i=1}^n \|H_i\|^2/n\right)^{1/2}\left(\sum_{i=1}^n \|q_i\|^2/n\right)^{1/2} = O_p\left(1\right), \end{aligned}$$

so that $\hat{R}_n = O_p\left(1\right)$. Therefore by Assumption 2(d), CS and T

$$\begin{aligned} \left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}-\tilde{\Omega}\right)\right\| &= \left\|n^{-1}\sum_{i=1}^n \bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left\{\left(\hat{\rho}_i^2-\rho_i^2\right)q_i q_i'\right\}\right\| \\ &= \left\|n^{-1}\sum_{i=1}^n\left(\hat{\rho}_i^2-\rho_i^2\right)H_i q_i'/\zeta(K)^2\right\| \\ &\le n^{-1}\sum_{i=1}^n\left|\hat{\rho}_i^2-\rho_i^2\right|\,\|H_i\|\,\|q_i\| \\ &\leqslant \left\|\hat{\beta}-\beta_0\right\|\hat{R}_n = O_p\left(\tau_n\right). \end{aligned}$$

Also by Assumptions 1(a) and 2(d)

$$\begin{aligned} E\left[\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\tilde{\Omega}-\bar{\Omega}\right)\right\|^2\middle|Z\right] &= E\left[\left\|\sum_{i=1}^n\left(\rho_i^2-\sigma_i^2\right)H_i q_i'/n\right\|^2\middle|Z\right] \\ &= \frac{1}{n^2}\operatorname{tr}E\left[\sum_{i=1}^n\left(\rho_i^2-\sigma_i^2\right)^2 H_i q_i' q_i H_i\middle|Z\right] \\ &\le \frac{1}{n^2}E\left[\sum_{i=1}^n E\left[\rho_i^4|z_i\right]\operatorname{tr}\left\{H_i\,\|q_i\|^2\,H_i'\right\}\right] \\ &\le \frac{C}{n}\sum_{i=1}^n\|H_i\|^2/n = O_p(1/n), \end{aligned}$$

so that $\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\tilde{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(n^{-1/2}\right)$. The conclusion in (ii) follows by T, that is, $\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(\tau_n + n^{-1/2}\right)$.

To prove (iii), note that $s_1(v) = -(1+v)$ so that $\check{G} = \sum_{i=1}^n\left(1+\tilde{\lambda}'\hat{g}_i\right)q_i\hat{\rho}_{\beta i}/n$, where $\hat{\rho}_{\beta i} =$

$\rho_\beta\left(w_i,\hat\beta\right)$. Let $b_i = \sup_{\beta\in\mathcal{N}}\left\|\rho_\beta\left(w_i,\beta\right)\right\|$, then by 3(c), T, CS, and M

$$\begin{aligned}
\left\|\check G-\hat G\right\| &= \left\|n^{-1}\sum_{i=1}^n\left(\tilde\lambda'\hat g_i\right)q_i\hat\rho_{\beta i}\right\| \\
&\leqslant \sum_{i=1}^n\left|\tilde\lambda'\hat g_i\right|\left|b_i\right|\left\|q_i\right\|/n \\
&\leqslant \sqrt{\sum_{i=1}^n\left|\tilde\lambda'\hat g_i\right|^2/n}\sqrt{\sum_{i=1}^n b_i^2\left\|q_i\right\|^2/n} \\
&\leqslant \sqrt{\tilde\lambda'\hat\Omega\tilde\lambda}\sqrt{\sum_{i=1}^n b_i^2\left\|q_i\right\|^2/n} \\
&\leq C\|\tilde\lambda\|O_p\left(\left\{E\left[E\left[b_i^2|z_i\right]\|q_i\|^2\right]\right\}^{1/2}\right)=O_p\left(\kappa_n\right),
\end{aligned}$$

giving the first result in (iii). The second result in (iii) follows by T. □

Lemma 13. *If Assumption 1 is satisfied, ε_i, Y_i are random variables with $E\left[\varepsilon_i|z_i\right]=0$, $E\left[\|\varepsilon_i\|^2|z_i\right]\leq C$, $E\left[\|Y_i\|^2|z_i\right]\leq C$, $U_i=U(z_i)$ is a nonnegative scalar function that is bounded away from zero, $K\to\infty,\alpha\to 0,\alpha\sqrt{n}\to\infty$ then,*

$$\left(\frac{1}{n}\sum_{i=1}^n q_iY_i\right)'\left(\frac{1}{n}\sum_{i=1}^n U_iq_iq_i'+\alpha I\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n q_i\varepsilon_i\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^n E\left[Y_i|z_i\right]'U_i^{-1}\varepsilon_i\xrightarrow{p}0.$$

Proof. Let P and Q^α be as defined in the proof of Lemma 4, $A_i=U_i^{-1/2}Y_i$, $\bar A_i=E\left[A_i|z_i\right]=U_i^{-1/2}E\left[Y_i|z_i\right]$, $A=(A_1,\cdots,A_n)'$, $\bar A=\left(\bar A_1,\cdots,\bar A_n\right)'$, $B_i=U_i^{-1/2}\varepsilon_i$, and $B=(B_1,\ldots,B_n)'$. Then, similarly to the proof of Lemma 4

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{i=1}^n q_iY_i\right)'\left(\frac{1}{n}\sum_{i=1}^n U_iq_iq_i'+\alpha I\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n q_i\varepsilon_i\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^n E\left[Y_i|z_i\right]'U_i^{-1}\varepsilon_i \\
&=A'Q^\alpha B/\sqrt{n}-\bar A'B/\sqrt{n}=(A-\bar A)'Q^\alpha B/\sqrt{n}-\bar A'\left(I-Q^\alpha\right)B/\sqrt{n}.
\end{aligned}$$

It follows as in the proof of Lemma 4 that $(A-\bar A)'Q^\alpha(A-\bar A)=O_p\left(\mathrm{tr}\left(Q^\alpha\right)\right)=O_p(1/\alpha)$. Also, ly$\bar B_i\overset{\mathrm{def}}{=}E\left[B_i|z_i\right]=U_i^{-1/2}E\left[\varepsilon_i|z_i\right]=0$ for all i, $B'Q^\alpha B=(B-\bar B)'Q^\alpha(B-\bar B)=O_p(1/\alpha)$ as in the proof of Lemma 4. It follows by CS that

$$\left|(A-\bar A)'Q^\alpha B/\sqrt{n}\right|\leq\sqrt{(A-\bar A)'Q^\alpha(A-\bar A)}\sqrt{B'Q^\alpha B}/\sqrt{n}=O_p(1/\alpha\sqrt{n})\xrightarrow{p}0$$

Also, as in Lemma 4, $E\left[\bar A'(I-Q)\bar A\right]/n\to 0$, so by the law of iterated expectations, and Lemma 3,

$$\begin{aligned}
E\left[\left|\bar A'\left(I-Q^\alpha\right)B/\sqrt{n}\right|^2\right] &= E\left[\bar A'\left(I-Q^\alpha\right)E\left[BB'\mid Z\right]\left(I-Q^\alpha\right)\bar A\right]/n \\
&= E\left[\bar A'\left(I-Q^\alpha\right)U_i^{-1}E\left[\varepsilon_i^2|z_i\right]\left(I-Q^\alpha\right)\bar A\right]/n \\
&\leq CE\left[\bar A'\left(I-Q^\alpha\right)^2\bar A\right]/n \\
&\leq CE\left[\bar A'\left(I-Q^\alpha\right)\bar A\right]/n\to 0.
\end{aligned}$$

The conclusion follows by M and T. □

Proof of Theorem 4.2. Let $a_n = (n\alpha)^{1/2}$, then $\|g(\hat{\beta})\| = O_p(a_n^{-1})$ by Lemma 10. For $\delta_n = n^{-1/\gamma-\varepsilon}$, $\delta_n n^{1/\gamma} \to 0$ and $\alpha\delta_n a_n = \alpha^{3/2} n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$. It follows that hypotheses of Lemma 7 are satisfied with $\tilde{\beta} = \hat{\beta}$ so that $\hat{\lambda} = \arg\max_{\lambda \in \hat{\Lambda}(\hat{\beta})} \hat{P}(\hat{\beta}, \lambda)$ exists w.p.a.1, and $\|\hat{\lambda}\| = O_p\left(\alpha^{-3/2} n^{-1/2}\right)$. It follows that $\|\hat{\lambda}\|/\delta_n = O_p\left(\delta_n^{-1}\alpha^{-3/2}n^{-1/2}\right) \xrightarrow{p} 0$ by $\delta_n \alpha^{3/2} n^{1/2} = \alpha^{3/2} n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$. It follows that $\hat{\lambda} \in \Lambda_n$ w.p.a.1 so that $\max_{1\le i\le n}\left|\hat{\lambda}'\hat{g}_i\right|$
$\xrightarrow{p} 0$ by Lemma 6.

Also, by consistency of $\hat{\beta}$ (Theorem 4.1), it will be an element of $int(\mathscr{B})$. It follows by Assumption 3(b) that w.p.a.1, $\hat{P}(\beta, \lambda)$ is twice continuously differentiable in a neighborhood of $(\hat{\beta}, \hat{\lambda})$. Then by the first-order condition for $\hat{\lambda}$, $\partial\hat{P}(\hat{\beta}, \lambda)/\partial\lambda\big|_{\lambda=\hat{\lambda}} = 0$. Also, by the implicit function theorem, for all β in a neighborhood of $\hat{\beta}$ there is $\hat{\lambda}(\beta)$ such that $\partial\hat{P}(\beta, \lambda)/\partial\lambda\big|_{\lambda=\hat{\lambda}(\beta)} = 0$ and $\hat{\lambda}(\beta)$ is continuously differentiable in β with $\hat{\lambda}(\hat{\beta}) = \hat{\lambda}$. By concavity of $\hat{P}(\beta, \lambda)$ it maximize at $\hat{\lambda}(\beta)$ holding β fixed. Then the first order conditions for $\hat{\beta}$ and the envelope theorem give $0 = \partial\hat{P}(\beta, \lambda(\beta))/\partial\beta\big|_{\beta=\hat{\beta}} = \partial\hat{P}(\hat{\beta}, \hat{\lambda})/\partial\beta = \check{G}'\hat{\lambda}$, with $\check{G} = -n^{-1}\sum_{i=1}^n s_1\left(\hat{\lambda}'\hat{g}_i\right) g_i(\hat{\beta})/\partial\beta'$. Also, by $\hat{P}(\hat{\beta}, \lambda) = -\hat{g}'\lambda - \frac{1}{2}\lambda'\left(\hat{\Omega} + \alpha I\right)\lambda$, FOC for $\hat{\lambda}$ implies $0 = -\hat{g} - \left(\hat{\Omega} + \alpha I\right)\hat{\lambda} = 0$ so that $\hat{\lambda} = -\left(\hat{\Omega} + \alpha I\right)^{-1}\hat{g}$. Plugging the equation for $\hat{\lambda}$ in the first order conditions for $\hat{\beta}$ gives $\check{G}'\left(\hat{\Omega} + \alpha I\right)^{-1}\hat{g} = 0$.

Expanding $\hat{g}$ around β_0 gives, for a mean value $\dot{\beta}$, $\dot{G} = n^{-1}\sum_{i=1}^n q_i \rho_\beta\left(w_i, \dot{\beta}\right)$, and $\bar{g} = \hat{g}\left(\beta_0\right)$

$$\check{G}'\left(\hat{\Omega} + \alpha I\right)^{-1}\dot{G}\left(\hat{\beta} - \beta_0\right) + \dot{G}'\left(\hat{\Omega} + \alpha I\right)^{-1}\bar{g} = 0. \tag{A.18}$$

Note by Assumptions 2(e) and 3(d) that $E\left[D(z)'\sigma(z)^{-2}D(z)\right] \geqslant CE\left[D(z)'D(z)\right]$ so that $V = \left\{E\left[D(z)'\sigma(z)^{-2}D(z)\right]\right\}^{-1}$ exists. Now successively apply Lemma 4 with $\theta = (\beta', \lambda')'$, $\tilde{\theta} = \left(\beta_0', 0'\right)'$, $\hat{\theta} = \left(\hat{\beta}', \hat{\lambda}'\right)'$, $a\left(w_i, \theta\right) = s_1\left(\lambda' g_i(\beta)\right)\partial\rho\left(w, \beta\right)/\partial\beta_r$, $b\left(w_i, \theta\right) = \partial\rho\left(w_i, \beta\right)/\partial\beta_s$, and $U(z) = \sigma(z)^2 = E\left[\rho\left(w, \beta_0\right)^2 | z\right]$ for $r, s = 1, \ldots, p$, we obtain $\check{G}'\left(\bar{\Omega} + \alpha I\right)^{-1}\dot{G} \xrightarrow{p} V^{-1}$. Also, by Lemma 11 $\hat{\beta} = \beta_0 + O_p\left(\tau_n\right)$ with $\tau_n = \alpha^{-1}n^{-1/2}$ so that the conclusion of Lemma 9(ii) gives $\left\|\hat{\Omega} - \bar{\Omega}\right\| = O_p\left(\tau_n + n^{-1/2}\right) = O_p\left(\alpha^{-1}n^{-1/2}\right) = o_p(1)$ by $\alpha n^{1/2} \to \infty$. It follows that

$$\check{G}'\left(\hat{\Omega} + \alpha I\right)^{-1}\dot{G} = \check{G}'\left(\bar{\Omega} + \alpha I\right)^{-1}\dot{G} + o_p(1) \xrightarrow{p} V^{-1}. \tag{A.19}$$

Also as previously justified, $\|\hat{\lambda}\| = O_p\left(\kappa_n\right)$, where $\kappa_n = \alpha^{-3/2}n^{-1/2}$. By Lemma 12(ii)&(iii) applied to $\tilde{\lambda} = \hat{\lambda}$, we have $\left\|\bar{G}'\left(\bar{\Omega} + \alpha I\right)^{-1}\left(\hat{\Omega} - \bar{\Omega}\right)\right\| = O_p\left(\tau_n + n^{-1/2}\right) = O_p\left(\alpha^{-1}n^{-1/2}\right)$ and $\left\|\check{G} - \bar{G}\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right) = O_p\left(\alpha^{-3/2}n^{-1/2}\right)$ so that $\hat{A} \overset{def}{=} \left\|\check{G} - \bar{G}\right\| + \left\|\bar{G}'\left(\bar{\Omega} + \alpha I\right)^{-1}\left(\hat{\Omega} - \bar{\Omega}\right)\right\| = O_p\left(\alpha^{-3/2}n^{-1/2}\right)$. By $\hat{\Omega}$ p.s.d. $\lambda_{\min}\left(\hat{\Omega} + \alpha I\right) \geq \alpha$ and therefore $\lambda_{\max}\left\{\left(\hat{\Omega} + \alpha I\right)^{-1}\right\} \leq 1/\alpha$. By Lemma 2 $\|\bar{g}\| = O_p\left(n^{-1/2}\right)$ so that by CS

$$\begin{aligned}\left\|\left(\hat{\Omega} + \alpha I\right)^{-1}\bar{g}\right\| &= \left\{\bar{g}'\left(\hat{\Omega} + \alpha I\right)^{-1}\left(\hat{\Omega} + \alpha I\right)^{-1}\bar{g}\right\}^{1/2} \\ &\leq \frac{1}{\alpha}\left\{\bar{g}'\bar{g}\right\}^{1/2} = \frac{1}{\alpha}\|\bar{g}\| = O_p\left(\alpha^{-1}n^{-1/2}\right).\end{aligned}$$

It follows that $\hat{A}\left\|\left(\hat{\Omega} + \alpha I\right)^{-1}\bar{g}\right\| = b_n O_p\left(n^{-1/2}\right)$ with $b_n = \alpha^{-5/2}n^{-1/2} = o(1)$ by the hypothesis in

Theorem 4.2, so that $\hat{A}\left\|\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| = o_p(n^{-1/2})$. Then by T and CS

$$\begin{aligned}
&\left\|\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}-\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}\right\|\\
&=\left\|\left(\check{G}-\bar{G}+\bar{G}\right)'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}-\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}\right\|\\
&\le\left\|\left(\check{G}-\bar{G}\right)\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\|+\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left[\bar{\Omega}-\hat{\Omega}\right]\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\|\\
&\leqslant\hat{A}\left\|\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\|=o_p(1/\sqrt{n}).
\end{aligned}$$

It follows that $\bar{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}=\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}\ +o_p(1/\sqrt{n})$.

Furthermore, Lemma 13 applied to $Y_i' = D\left(z_i\right)\overset{def}{=}D_i$, $\varepsilon_i=\rho\left(w_i,\beta_0\right)\overset{def}{=}\rho_i$, and $U_i=\sigma_i^2\overset{def}{=}\sigma(z_i)^2$, leads to $\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}-n^{-1}\sum_{i=1}^n D_i'\sigma_i^2\rho_i=o_p(1/\sqrt{n})$. Also, by the Lindbergh-Levy central limit theorem $\sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i/\sqrt{n}\to\xrightarrow{d}N(0,\Lambda)$, where, by iterated expectation, $\Lambda\overset{def}{=}E\left[D_i'\sigma_i^{-2}\rho_i^2\sigma_i^{-2}D_i\right]=E\left[D_i'\sigma_i^{-2}D_i\right]=V^{-1}$. Therefore,

$$\begin{aligned}
\sqrt{n}\bar{G}'(\bar{\Omega}+\alpha I)^{-1}\bar{g}&=\sqrt{n}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}-\sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i\Big/n\right)+\sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i\Big/\sqrt{n}\\
&=o_p(1)+\sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i/\sqrt{n}\xrightarrow{d}\mathcal{N}\left(0,V^{-1}\right),
\end{aligned}$$

and thus

$$\sqrt{n}\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}=\sqrt{n}\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}+o_p(1)\xrightarrow{d}\mathcal{N}\left(0,V^{-1}\right)\tag{A.20}$$

By Eq. (A.18)

$$\sqrt{n}(\hat{\beta}-\beta_0)=-\left[\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\dot{G}\right]^{-1}\sqrt{n}\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g},$$

so that by Eq. (A.19), Eq. (A.20), the Slutzky's theorem, and the continuous mapping theorem, $\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d}\mathcal{N}(0,V)$, giving the first result.

We now establish the consistency of the variance estimator. First, applying Lemma 4 gives $\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}$
$\xrightarrow{p}V^{-1}$. Also, by $\bar{\Omega}$ p.s.d. $\lambda_{\max}\left[\left(\bar{\Omega}+\alpha I\right)^{-1}\right]\leqslant 1/\alpha$ so that for $\hat{B}=\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}$, it follows by CS that

$$\begin{aligned}
\|\hat{B}\|^2=\operatorname{tr}\left(\hat{B}'\hat{B}\right)&=\operatorname{tr}\left(\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}\right)\\
&\le\frac{1}{\alpha}\operatorname{tr}\left(\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}\right)\le O_p(1/\alpha).
\end{aligned}$$

Also by Lemma 9 applied to $\kappa_n=\alpha^{-3/2}n^{-1/2}$ and $\tau_n=\alpha^{-1}n^{-1/2}$ we have $\left\|\check{\Omega}-\bar{\Omega}\right\|=O_p\left(\kappa_n+\tau_n+n^{-1/2}\right)=O_p\left(\alpha^{-3/2}n^{-1/2}\right)$ for $\check{\Omega}=-\sum_{i=1}^n s_1\left(\hat{\lambda}'\hat{g}_i\right)\hat{g}_i\hat{g}_i'\big/n$. By T, CS,

$\lambda_{\max}\left\{(\check{\Omega}+\alpha I)^{-1}\right\} \leq 1/\alpha$, and $A^{-1}-B^{-1}=B^{-1}\left[(B-A)+(B-A)A^{-1}(B-A)\right]B^{-1}$,

$$\begin{aligned}
\left\|\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G}-\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}\right\| &= \left\|\check{G}'\left[\left(\check{\Omega}+\alpha I\right)^{-1}-\left(\bar{\Omega}+\alpha I\right)^{-1}\right]\check{G}\right\| \\
&= \left\|\hat{B}'\left\{\bar{\Omega}-\check{\Omega}+\left(\bar{\Omega}-\check{\Omega}\right)\left(\check{\Omega}+\alpha I\right)^{-1}\left(\bar{\Omega}-\check{\Omega}\right)\right\}\hat{B}\right\| \\
&\leq \|\hat{B}\|^2\left(\left\|\bar{\Omega}-\check{\Omega}\right\|+\alpha^{-1}\left\|\bar{\Omega}-\check{\Omega}\right\|^2\right) \\
&\leq C\alpha^{-1}\left(O_p\left(\alpha^{-3/2}n^{-1/2}\right)+\alpha^{-1}O_p\left(\alpha^{-3}n^{-1}\right)\right) \\
&\leq O_p\left(\alpha^{-5/2}n^{-1/2}\right)+O_p\left(\left(\alpha^{-5/2}n^{-1/2}\right)^2\right)\xrightarrow{p}0.
\end{aligned}$$

It follows that $\check{G}'(\check{\Omega}+\alpha I)^{-1}\check{G} = \check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}+o_p(1)\xrightarrow{p}V^{-1}$. Also, by $s_1(v)=-(1+v)$, we have $\left|1+\sum_{i=1}^n s_1\left(\hat{v}_i\right)/n\right| = \left|\sum_{i=1}^n \hat{v}_i/n\right| \leq \max_{1\leq i\leq n}\left|\hat{\lambda}'\hat{g}_i\right| \xrightarrow{p} 0$ so that $\sum_{i=1}^n s_1\left(\hat{v}_i\right)/n \xrightarrow{p} -1$. Note that $\check{\Omega} = -n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)\hat{\Omega}$ and $\check{G} = -n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)\hat{G}$ so that $\hat{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{G} = -\left(n\big/\sum_{i=1}^n s_1\left(\hat{v}_i\right)\right)\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G}$. It follows by continuous mapping that,

$$\begin{aligned}
\hat{V}^{-1} &\overset{def}{=} \hat{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{G} \\
&= -\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G}\left[\frac{1+n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)}{n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)}\right]+\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G}\xrightarrow{p}V^{-1},
\end{aligned}$$

giving the consistency result for the variance estimator. □

References

Altonji, J. G., Smith Jr, A. A., and Vidangos, I. (2013). Modeling earnings dynamics. *Econometrica*, 81(4):1395–1454.

Angrist, J. D. and Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1):S97–S140.

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

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Carrasco, M. and Florens, J.-P. (2000). Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751.

Carrasco, M. and Tchuente, G. (2015). Regularized liml for many instruments. *Journal of Econometrics*, 186(2):427–442.

Carrasco, M. and Tchuente, G. (2016). Efficient estimation with many weak instruments using regularization techniques. *Econometric Reviews*, 35(8-10):1609–1637.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

Dmitriev, A. (2013). Institutions and growth: evidence from estimation methods robust to weak instruments. *Applied Economics*, 45(13):1625–1635.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1):28–36.

Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69(5):1161–1191.

Eaton, J., Kortum, S., and Kramarz, F. (2011). An anatomy of international trade: Evidence from french firms. *Econometrica*, 79(5):1453–1498.

Frankel, J. A. and Romer, D. H. (1999). Does trade cause growth? *American Economic Review*, 89(3):379–399.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society*, pages 939–953.

Guggenberger, P. (2005). Monte-carlo evidence suggesting a no moment problem of the continuous updating estimator. *Economics Bulletin*, 3(13):1–6.

Hahn, J. and Hausman, J. (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review*, 93(2):118–125.

Hall, R. E. and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The quarterly journal of economics*, 114(1):83–116.

Han, C., Orea, L., and Schmidt, P. (2005). Estimation of a panel data model with parametric temporal variation in individual effects. *Journal of Econometrics*, 126(2):241–267.

Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422.

Hansen, C. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Hausman, J., Lewis, R., Menzel, K., and Newey, W. (2011). Properties of the cue estimator and a modification with moments. *Journal of Econometrics*, 165(1):45–57.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Newey, W. K. and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719.

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics*, 195(1):104–119.

Stock, J. and Yogo, M. (2005). *Asymptotic distributions of instrumental variables statistics with many instruments*, volume 6. Chapter.