econometrics

MDPI

Article

On the Asymptotic Distribution of Ridge Regression Estimators Using Training and Test Samples

Nandana Sengupta [1,†] **and Fallaw Sowell** [2,*,†]

1 School of Public Policy, Indian Institute of Technology Delhi, Delhi 110016, India; nandana.sengupta@sopp.iitd.ac.in
2 Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, USA
* Correspondence: fs0v@andrew.cmu.edu; Tel.: +1-(412)-268-3769
† These authors contributed equally to this work.

Received: 25 November 2019; Accepted: 11 September 2020; Published: 1 October 2020

Abstract: The asymptotic distribution of the linear instrumental variables (IV) estimator with empirically selected ridge regression penalty is characterized. The regularization tuning parameter is selected by splitting the observed data into training and test samples and becomes an estimated parameter that jointly converges with the parameters of interest. The asymptotic distribution is a nonstandard mixture distribution. Monte Carlo simulations show the asymptotic distribution captures the characteristics of the sampling distributions and when this ridge estimator performs better than two-stage least squares. An empirical application on returns to education data is presented.

Keywords: ridge regression; instrumental variables; regularization; training and test samples; generalized method of moments framework

JEL Classification: C13; C18

1. Introduction

This paper concerns the estimation and inference on the structural parameters in the linear instrumental variables (IV) model estimated with ridge regression. This estimator differs from previous ridge regression estimators in three important areas. First, the regularization tuning parameter is selected using a randomly selected test sample from the observed data. Second, the empirically selected tuning parameter's impact on the estimates of the parameters of interest is accounted for by deriving their asymptotic joint distribution, which is a mixture. Third, the traditional Generalized Method of Moments (GMM) framework is used to characterize the asymptotic distribution.

The ridge estimator belongs to a family of estimators which utilize regularization, see (Bickel et al. 2006) and (Hastie et al. 2009) for overviews. Regularization requires tuning parameters and procedures to select them can be split into three broad areas. (1) Plugin-type. For a given criteria function, an optimal value is determined in terms of the model's parameters. These model's parameters are then estimated with the data set and plugged into the formula. Generalization include adjustments to reduce bias and iterative procedures until a fixed point is achieved. (2) Test sample. The data set is randomly split into a training and a test sample. The tuning parameter and estimates from the training sample are used to evaluate some criteria function on the test sample to determine the optimal tuning parameter and model. Generalizations include k-fold cross-validation and generalized cross-validation. (3) Rate of convergence restriction. The tuning parameter must converge at an appropriate rate to guarantee consistency and a known asymptotic distribution for the estimates of the parameters of interest. A key feature is that the tuning parameter only converges to zero asymptotically and is restricted from

zero in finite samples. Previous ridge estimators have relied on plugin-type and rate of convergence restrictions. We study the test sample approach.

This estimator builds on a large literature. The ridge regression estimator was proposed in (Hoerl and Kennard 1970) to obtain smaller MSE relative to the OLS estimates in the linear model when the covariates are all exogenous but have multicollinearity. It was shown that for fixed tuning parameter, α, the bias and variance of the ridge estimator implied that there exists an α value with lower MSE than for the OLS estimator. Assuming that α is fixed, Hoerl et al. (1975) proposed selecting α to minimizing the MSE of the ridge estimator. This resulting formula became a plugin selection for α. Subsequent research has followed the same approach of selecting the tuning parameter to minimize MSE where α is assumed fixed, see (Dorugade 2014) and the papers cited there. A shortcoming of this plugin-type approach is that the form of the MSE assumes the tuning parameter is fixed, however it is then selected on the observed sample and hence is stochastic, see (Theobald 1974) and (Montgomery et al. 2012). Instead of focusing on reducing the MSE, we focus on the large sample properties of the estimates of the parameters of interest. In this literature the work closest to ours is (Firinguetti and Bobadilla 2011) where the sampling distribution is considered for a ridge estimator. However, this estimator is built on minimizing the MSE where the tuning parameter is assumed fixed, see (Lawless and Wang 1976). Our ridge estimator is derived knowing that the tuning parameter is stochastic. This leads to the joint asymptotic distribution for the estimates of the parameters of interest and the tuning parameter.

The supervised learning (machine learning) literature focuses on the ability to generalize to new data sets by selecting the tuning parameters to minimize the prediction error for a test (or holdout or validation) sample. Starting with (Larsen et al. 1996), a test sample is used to select optimal tuning parameters for a neural network model. The problem reduces to finding a local minimum of the criteria function evaluated on the test sample. Extensions of the test sample approach include backward propagation, see (Bengio 2000; Habibnia and Maasoumi 2019), but as the number of parameters increases the memory requirements become too large. This has led to the use of stochastic gradient decent, see (Maclaurin et al. 2015). Much research in this area has focused on effecient ways to optimally select the hyperparameters to minimize prediction errors. We select the tuning parameter to address its impact on the estimates of the model's coefficients and do not focus on the model's predictive power.

A number of papers extend the linear IV model with tuning parameters. Structural econometrics (Carrasco 2012) allows the number of instruments to grow with the sample size and (Zhu 2018) considers models where the number of covariates and instruments is larger than the sample size. In genetics the linear IV model is widely used to model gene regulatory networks, see (Chen et al. 2018; Lin et al. 2015). In this setting the number of covariates and instruments can be larger than the number of observations and the tuning parameter is restricted from being zero in finite samples. In contrast to these models, we fix the number of covariates and instrument to determine the asymptotic distribution and permit the tuning parameter to take the value zero.

Within the structural econometrics literature, ridge type regularization concepts are not new. Notable contributions are (Carrasco and Tchuente 2016; Carrasco and Florens 2000; Carrasco et al. 2007) which allow for a continuum of moment conditions. The authors use ridge regularization to find the inverse of the optimal weighting operator (instead of optimal weighting matrix in traditional GMM). In these papers and in (Carrasco 2012) the rate of convergence restriction is used to select the tuning parameter.

Several types of identification and asymptotic distributions can occur with linear IV models e.g., strong instruments, nearly-strong instruments, nearly-weak instrument and weak instruments, see (Antoine and Renault 2009) for a summary. For this taxonomy, this paper and estimator is in the strong instruments setting. The models considered in this paper are closest to the situation considered in (Sanderson and Windmeijer 2016). However, unlike (Sanderson and Windmeijer 2016) we provide point estimates instead of testing for weak instruments and restrict attention to fixed parameters that

do not drift to zero. The models we study are explicitly strongly identified, however in a finite sample the precision can be low.

This ridge estimator extends the literature in five important dimensions. First, this estimator allows a meaningful prior. When the prior is ignored, or equivalently set to zero, the model penalizes variability about the origin. However, in structural economic models a more appropriate penalty will be variability about some economically meaningful prior. Second, the regularization tuning parameter is selected empirically using the observed data. This removes the internally inconsistent argument about the minimum MSE when the tuning parameters is assumed fixed. Third, the tuning parameter is allowed to take the value zero in finite samples. Fourth, empirically selecting the tuning parameter impacts the asymptotic distribution of the parameter estimates. As stressed in (Leeb and Pötscher 2005), the final asymptotic distribution will depend on empirically selected tuning parameters. We address this directly by characterizing the joint asymptotic distribution that includes both the parameters of interest and the tuning parameter. Fifth, the GMM framework is used to characterize the asymptotic distribution.[1] The GMM framework is used because it is better suited to the social science setting where this estimator will be most useful. Rarely does a social science model imply the actual distribution for an error. Unconditional expectations of zero are more typical in social science theories and are the foundation for the GMM estimator. Adding a regularization penalty term and splitting the observed data into a training and test samples, takes the estimator out of the traditional GMM framework. We present new moment conditions in the traditional GMM framework which include the first order conditions for the ridge estimator.

Section 2 presents the linear IV framework, describes the precision problem and the ridge estimator. Section 3 characterizes the asymptotic distribution of the ridge estimator in the traditional GMM framework. Small sample properties are analyzed via simulations in Section 4. The procedure is applied to the returns to education data set from Angrist and Krueger (1991) in Section 5. Section 6 summarizes the results and presents directions for future research.

2. Ridge Estimator for Linear Instrumental Variables Model

This section presents the ridge regression estimator where regularization tuning parameter is empirically determined by splitting the data into training and test samples. This estimator is then fit into the traditional GMM framework to characterize its asymptotic distribution. Consider the model

$$\begin{aligned} Y &= X\beta_0 + \varepsilon \quad (1) \\ X &= Z\Gamma_0 + u \quad (2) \end{aligned}$$

where Y is $n \times 1$, X is $n \times k$, $Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}'$ is $n \times m$, $m \geq k$, $z_i \sim iid$, $R_z = E[z_i z_i']$ full rank, and conditional on Z, $\begin{bmatrix} \varepsilon_i \\ u_i \end{bmatrix} \sim iid \left(0, \begin{bmatrix} \sigma_\varepsilon^2 & \Sigma_{\varepsilon u} \\ \Sigma_{u\varepsilon} & \Sigma_U \end{bmatrix} \right)$. This model allows for both endogenous X's that are correlated with ε and exogenous X's that are uncorrelated with ε. Endogenous regressors imply OLS will be inconsistent. The Z instruments allow consistent estimates with the IV estimator that minimizes the residual sum of squares projected onto the instruments and has the closed form

$$\begin{aligned} \hat{\beta}_{IV} &= \arg\min_\beta \frac{1}{2n}(Y - X\beta)'Z(Z'Z)^{-1}Z'(Y - X\beta) \quad (3) \\ &= (X'P_Z X)^{-1}X'P_Z Y \end{aligned}$$

[1] Relative the likelihood based approaches, the GMM framework is better suited to the social science setting where this estimator will be most useful. Rarely does a social science model imply the actual distribution for an error. Unconditional expectations of zero are more typical in social science theories and are the foundation for the GMM estimator.

where P_Z is the projection matrix for Z. The well known asymptotic distribution is

$$\sqrt{n}\left(\hat{\beta}_{IV}-\beta_0\right) \sim_a N\left(0, \sigma_\varepsilon^2\left(\Gamma_0' R_z \Gamma_0\right)^{-1}\right)$$

and the covariance can be consistently estimated with

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}\left[\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)\right]^{-1} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}\left[\frac{X'P_ZX}{n}\right]^{-1} \tag{4}$$

where $\hat{\varepsilon} = Y - X\hat{\beta}_{IV}$. Let $S_0 = E[z_i x_i'] = R_z\Gamma_0$.

For a finite sample let[2] $\frac{X'P_ZX}{n}$ have the spectral decomposition $C\Lambda C'$, where Λ is a positive definite diagonal $k \times k$ matrix, and C is orthonormal, $C'C = I_k$. A precision problem occurs when some of the eigenvectors explain *very little* variation, as represented by the magnitude of the corresponding eigenvalues. This occurs when the objective function is relatively flat along these dimensions and the resulting covariance estimates are large because as Equation (4) shows, the variance of $\hat{\beta}_{IV}$ is proportional to $\left(\frac{X'P_ZX}{n}\right)^{-1} = (C\Lambda C')^{-1} = C\Lambda^{-1}C'$. The flat objective function, or equivalently large estimated variances, leads to a relatively large MSE. The ridge estimator addresses this problem by shrinking the estimated parameter toward a prior. The IV estimate still has low bias (it is consistent) and has the asymptotically minimum variance. However, accepting a little higher bias can have a dramatic reduction in the variance and thus provide a point estimate with lower MSE.

The ridge objective function augments the usual IV objective function (3) with a quadratic penalty centered at a prior value, β^p, weighted by a regularization tuning parameter α

$$Q_n(\beta) = \frac{1}{2n}(Y - X\beta)'P_Z(Y - X\beta) + \frac{1}{2}\alpha(\beta - \beta^p)'(\beta - \beta^p). \tag{5}$$

The objective function's second derivative is $\left(\frac{X'P_ZX}{n} + \alpha I_k\right) = C(\Lambda + \alpha I_k)C'$. The regularization parameter injects stability since $\left(\frac{X'P_ZX}{n} + \alpha I_k\right)^{-1} = C\left(\Lambda + \alpha I_k\right)^{-1}C'$ has eigenvalues $1/(\lambda_i + \alpha)$ for $i = 1, \ldots, k$ which are decreasing in α. This results in smaller variance but higher bias.

Denote the ridge solution given α as

$$\begin{aligned}
\hat{\beta}_{IV}(\alpha) &= \left(\frac{X'P_ZX}{n} + \alpha I_k\right)^{-1}\left(\frac{X'P_ZY}{n} + \alpha\beta^p\right) \\
&= C\left(\Lambda + \alpha I_k\right)^{-1}C'\frac{X'P_ZY}{n} + C\left(\Lambda + \alpha I_k\right)^{-1}C'\alpha\beta^p \\
&= C\left(\Lambda + \alpha I_k\right)^{-1}C' \cdot \left[C\Lambda C' \cdot C\Lambda^{-1}C'\right]\frac{X'P_ZY}{n} + C\left(\frac{\Lambda}{\alpha} + I_k\right)^{-1}C'\beta^p \\
&= C\left(I_k + \alpha\Lambda^{-1}\right)^{-1}C'\hat{\beta}_{IV} + C\left(\frac{\Lambda}{\alpha} + I_k\right)^{-1}C'\beta^p.
\end{aligned} \tag{6}$$

Equation (6) shows how the tuning parameter, α creates a smooth curve in the parameter space between the low bias-high variance IV estimate, $\hat{\beta}_{IV}$, (when $\alpha = 0$) to the high bias-no variance prior, β^p, (when $\alpha \to \infty$).

Different values of α result in different estimated values for β_0. The optimal value of α is determined empirically by splitting the data into training and test samples. The training sample is a randomly drawn sample of $[\tau n]$ observations, denoted, $Y_{\tau n}$, $X_{\tau n}$, and $Z_{\tau n}$, and are used to calculate a

[2] This term is both the second derivative of the objective function (3) and the matrix being inverted in the last term of the covariance (4).

path between the IV estimate and the prior as in Equation (6). The estimate using the training sample, conditional on α, is

$$\hat{\beta}_{tr}(\alpha) \equiv \arg\min_{\beta} \frac{1}{2[\tau n]} (Y_{\tau n} - X_{\tau n}\beta)' P_{Z_{\tau n}} (Y_{\tau n} - X_{\tau n}\beta) + \frac{\alpha}{2}(\beta - \beta^p)'(\beta - \beta^p) \quad (7)$$

where $P_{Z_{\tau n}}$ is the projection matrix onto $Z_{\tau n}$ and $[\cdot]$ is the greatest integer function. The first order conditions for an internal solution are

$$-\frac{1}{\tau n} X'_{\tau n} P_{Z_{\tau n}} (Y_{\tau n} - X_{\tau n}\hat{\beta}_{tr}) + \alpha(\hat{\beta}_{tr} - \beta^p) = 0$$

or alternatively

$$-\frac{1}{[\tau n]} \sum_{i=1}^{[\tau n]} \left\{ \left(\frac{X'_{\tau n} Z_{\tau n}}{[\tau n]} \right) \left(\frac{Z'_{\tau n} Z_{\tau n}}{[\tau n]} \right)^{-1} \right\} z_i (y_i - x'_i \hat{\beta}_{tr}) + \alpha(\hat{\beta}_{tr} - \beta^p) = 0. \quad (8)$$

The closed form solution is

$$\hat{\beta}_{tr}(\alpha) = \left(\frac{X'_{\tau n} P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I \right)^{-1} \left(\frac{X'_{\tau n} P_{Z_{\tau n}} Y_{\tau n}}{[\tau n]} + \alpha \beta^p \right). \quad (9)$$

As α goes from 0 towards infinity, this gives a path from the IV estimator, $\hat{\beta}_{tr}$ (at $\alpha = 0$), to the prior, β^p (the limit as $\alpha \to \infty$). Following this path, the optimal α is selected to minimize the IV least squares objective function (3) over the remaining $(n - [\tau n])$ observations, the test sample, denoted $Y_{n(1-\tau)}$, $X_{n(1-\tau)}$ and $Z_{n(1-\tau)}$. The optimal value for the tuning parameter is defined by $\hat{\alpha} = \arg\min_{\alpha \in [0,\infty)} Q_{n(1-\tau)}(\alpha)$ where

$$Q_{n(1-\tau)}(\alpha) = \frac{1}{2(n - [n\tau])} \left(Y_{n(1-\tau)} - X_{n(1-\tau)} \hat{\beta}_{tr}(\alpha) \right)' P_{Z_{n(1-\tau)}} \left(Y_{n(1-\tau)} - X_{n(1-\tau)} \hat{\beta}_{tr}(\alpha) \right) \quad (10)$$

where $P_{Z_{n(1-\tau)}}$ is the projection matrix onto $Z_{n(1-\tau)}$.

$$\frac{1}{(n - [\tau n])} (\beta^p - \hat{\beta}_{tr}(\hat{\alpha}))' \left(\frac{X'_{\tau n} P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \hat{\alpha} I_k \right)^{-1} X'_{n(1-\tau)} P_{Z_{n(1-\tau)}} \left(Y_{n(1-\tau)} - X_{n(1-\tau)} \hat{\beta}_{tr}(\hat{\alpha}) \right) = 0$$

or alternatively

$$\frac{1}{n - [\tau n]} \sum_{i=[\tau n]+1}^{n} \left\{ (\beta^p - \hat{\beta}_{tr}(\hat{\alpha}))' \left(\left(\frac{X'_{\tau n} Z_{\tau n}}{[\tau n]} \right) \left(\frac{Z'_{\tau n} Z_{\tau n}}{[\tau n]} \right)^{-1} \left(\frac{X'_{\tau n} Z_{\tau n}}{[\tau n]} \right) + \hat{\alpha} I_k \right)^{-1} \left(\frac{X'_{\tau n} Z_{\tau n}}{n - [\tau n]} \right) \left(\frac{Z'_{\tau n} Z_{\tau n}}{n - [\tau n]} \right)^{-1} \right\} z_i (y_i - x'_i \hat{\beta}_{tr}(\hat{\alpha})) = 0. \quad (11)$$

The ridge regression estimate $\hat{\beta}_{\hat{\alpha}} \equiv \hat{\beta}_{IV}(\hat{\alpha})$ is then characterized by

$$-\frac{1}{n} X' P_Z (Y - X\hat{\beta}_{\hat{\alpha}}) + \hat{\alpha}(\hat{\beta}_{\hat{\alpha}} - \beta^p) = 0$$

or alternatively

$$-\frac{1}{n} \sum_{i=1}^{n} \left\{ \left(\frac{X'Z}{n} \right) \left(\frac{Z'Z}{n} \right)^{-1} \right\} z'_i (y_i - x_i \hat{\beta}_{\hat{\alpha}}) + \hat{\alpha}(\hat{\beta}_{\hat{\alpha}} - \beta^p) = 0. \quad (12)$$

The first order conditions that characterize the ridge estimator, Equations (8), (11), and (12), are $2k+1$ equations in the $2k+1$ parameters and have the structure of sample averages being set to zero. However, the functions being averaged do not fit into the traditional GMM framework. In Equations (8), (11), and (12) the terms in the curly brackets depend on the entire sample and not just the data for index i and the parameters. The terms in the curly brackets will converge at $O_p\left(n^{-1/2}\right)$ and must be considered jointly with the asymptotic distributions of $(\hat{\beta}'_{\hat{\alpha}}, \hat{\alpha})'$.

The asymptotic distribution of the ridge estimator can be determined with the GMM framework using the parameterization

$$\theta = \begin{bmatrix} \text{vech}(R_\tau)' & \text{vech}\left(R_{(1-\tau)}\right)' & \text{vec}(S_\tau)' & \text{vec}\left(S_{(1-\tau)}\right)' & \beta'_{tr} & \alpha & \beta' \end{bmatrix}'$$

where $\text{vec}(\cdot)$ stacks the elements from a matrix into a column vector and $\text{vech}(\cdot)$ stacks the unique elements from a symmetric matrix into a column vector. The population parameter values are

$$\theta_0 = \begin{bmatrix} \text{vech}(R_z)' & \text{vech}(R_z)' & \text{vec}(R_z\Gamma_0)' & \text{vec}(R_z\Gamma_0)' & \beta'_0 & 0 & \beta'_0 \end{bmatrix}'.$$

The ridge estimator is part of the parameter estimates defined by the just identified system of equations $H_n(\theta) = \frac{1}{n}\sum_{i=1}^n h_i(\theta) = 0$ where

$$h_i(\theta) = \begin{bmatrix} \mathbf{1}_\tau(i)\text{vech}(R_\tau - z_i z_i') \\ (1-\mathbf{1}_\tau(i))\text{vech}(R_{(1-\tau)} - z_i z_i') \\ \mathbf{1}_\tau(i)\text{vec}(S_\tau - z_i x_i') \\ (1-\mathbf{1}_\tau(i))\text{vec}(S_{(1-\tau)} - z_i x_i') \\ \mathbf{1}_\tau(i)\left(-S'_\tau R_\tau^{-1} z_i(y_i - x_i'\beta_{tr}) + \alpha(\beta_{tr} - \beta^p)\right) \\ (1-\mathbf{1}_\tau(i))(y_i - x_i'\beta_{tr}) z_i' R_{(1-\tau)}^{-1} S_{(1-\tau)} \left(S'_\tau R_\tau^{-1} S_\tau + \alpha I_k\right)^{-1}(\beta^p - \beta_{tr}) \\ -(\tau S_\tau + (1-\tau)S_{1-\tau})'(\tau R_\tau + (1-\tau)R_{1-\tau})^{-1} z_i(y_i - x_i'\beta) + \alpha(\beta - \beta^p) \end{bmatrix} \tag{13}$$

and the training and test samples are determined with the indicator function

$$\mathbf{1}_\tau(i) = \begin{cases} 1, & i \le [\tau n] \\ 0, & [\tau n] < i. \end{cases}$$

Using the structure of Equation (13), the system $H_n(\theta) = \frac{1}{n}\sum_{i=1}^n h_i(\theta) = 0$ can be seen as seven sets of equations. The first four sets are each self-contained systems of equal numbers of equations and parameters. The fifth set has k equations and introduces k new parameters, β_{tr}. The six is a single equation with the new parameter α. The seventh set has k equations and introduces the final k parameters, β. Identification occurs because the expectation of the gradient is invertible. This is presented in the Appendices A and B.

3. Asymptotic Behavior

Three assumptions are sufficient to obtain asymptotic distribution for the ridge estimator.

Assumption 1. *z_i is iid with finite fourth moments and $E[z_i z_i'] = R_z$ has full rank.*

Assumption 2. *Conditional on Z, $\begin{bmatrix} \varepsilon_i & u_i' \end{bmatrix}'$ are iid vectors with zero mean, full rank covariance matrix with possibly nonzero off-diagonal elements.*

Assumptions 1 and 2 imply $E[h_i(\theta_0)] = 0$ and $\sqrt{n}H_n(\theta_0)$ satisfies the CLT.

Assumption 3. *The parameter space Θ is defined by: R_z is restricted to a symmetric positive definite matrix with eigenvalues $1/B_1 \leq \tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \ldots \leq \tilde{\lambda}_m \leq B_1$, $|\beta_j| \leq B_2$ for $j = 1, 2, \ldots, k$, $\Gamma_0 = [\gamma_{\ell,j}]$ is of full rank with $|\gamma_{\ell,j}| \leq B_3$ for $\ell = 1, \ldots, m$, $j = 1, 2, \ldots, k$ and $\alpha \in [0, B_4]$ where B_1, B_2, B_3 and B_4 are positive and finite.*

First consider the tuning parameter. Even though it is empirically selected using the training and testing samples, its limiting value and rate of convergence are familiar.

Lemma 1. *Assumptions 1, 2 and 3 imply*

1. $\hat{\alpha} \to 0$ *and*
2. $\sqrt{n}\hat{\alpha} = O_p(1)$.

Proofs are given in the Appendix A.

Lemma 1 implies that the population parameter value for the tuning parameter is zero, $\alpha_0 = 0$, which is on the boundary of the parameter space. This results in a nonstandard asymptotic distribution which can be characterized by appealing to Theorem 1 in (Andrews 2002). The approach in (Andrews 2002) requires the root-n convergence of the parameters. Lemma 1, traditional TSLS and method of moments establishes this for all the parameters in θ. Equation Equation (13) puts the ridge estimator in the form of the first part of (14) from (Andrews 2002). Because the system is just identified, the weighting matrix does not affect the estimator and is set to the identity matrix. The scaled GMM objective function can be expanded into a quadratic approximation about the centered and scaled population parameter values

$$\begin{aligned} nH_n(\theta)'H_n(\theta) &= nH_n(\theta_0)'H_n(\theta_0) + nH_n(\theta_0)\frac{\partial H_n(\theta_0)}{\partial \theta'}(\theta - \theta_0) \\ &\quad + \frac{n}{2}(\theta - \theta_0)'\left\{\frac{\partial H_n(\theta_0)'}{\partial \theta}\frac{\partial H_n(\theta_0)}{\partial \theta'}\right\}(\theta - \theta_0) + o_p(1) \\ &= \frac{n}{2}H_n(\theta_0)'H_n(\theta_0) + \frac{n}{2}\left(H_n(\theta_0) + \frac{\partial H_n(\theta_0)}{\partial \theta'}(\theta - \theta_0)\right)'\left(H_n(\theta_0) + \frac{\partial H_n(\theta_0)}{\partial \theta'}(\theta - \theta_0)\right) + o_p(1) \end{aligned}$$

$$\begin{aligned} &= \frac{n}{2}H_n(\theta_0)'H_n(\theta_0) + \frac{1}{2}\left(\left(-\frac{\partial H_n(\theta_0)}{\partial \theta'}\right)^{-1}\sqrt{n}H_n(\theta_0) - \sqrt{n}(\theta - \theta_0)\right)'\left\{\frac{\partial H_n(\theta_0)'}{\partial \theta}\frac{\partial H_n(\theta_0)}{\partial \theta'}\right\} \\ &\quad \times\left(\left(-\frac{\partial H_n(\theta_0)}{\partial \theta'}\right)^{-1}\sqrt{n}H_n(\theta_0) - \sqrt{n}(\theta - \theta_0)\right) + o_p(1). \end{aligned}$$

The first term does not depend on θ and the last term converges to zero in probability. This suggests that selecting $\hat{\theta}$ to minimize $H_n(\theta)'H_n(\theta)$ will result in the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ being the same as the distribution of $\lambda \in \Lambda \equiv \left\{\lambda \in R^{m(m+1)+2km+2k+1} : \lambda_{m(m+1)+2km+k+1} \geq 0\right\}$ where $(\mathcal{Z} - \lambda)'M_0'M_0(\mathcal{Z} - \lambda)$ takes its minimum, where the random variable is defined as

$$\mathcal{Z} = \lim_{n\to\infty}\left(E\left[-\frac{\partial H_n(\theta_0)}{\partial \theta'}\right]\right)^{-1}\sqrt{n}H_n(\theta_0)$$

and

$$M_0 = E\left[\frac{\partial H_n(\theta_0)}{\partial \theta'}\right].$$

This indeed is the result by Theorem 1 of (Andrews 2002). The needed assumptions are given in (Andrews 2002). The estimator is defined as

$$\hat{\theta} = \underset{\theta\in\Theta}{\arg\min}\quad H_n(\theta)'H_n(\theta).$$

Theorem 1. *Assumptions 1–3 imply the asymptotic distribution of* $\sqrt{n}(\hat{\theta} - \theta_0)$ *is equivalent to the distribution of*

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \quad (\mathcal{Z} - \lambda)' M_0' M_0 (\mathcal{Z} - \lambda).$$

The objective function can be minimized at a value of the tuning parameter in $(0, \infty)$ or possibly at $\alpha = 0$. The asymptotic distribution of the tuning parameter will be composed of two parts, a discrete mass at $\alpha = 0$ and a continuous function over $(0, \infty)$. The asymptotic distribution over the other parameters can be thought of as being composed of two parts, the distribution conditional on $\alpha = 0$ and the distribution over $\alpha > 0$.

In terms of the framework presented in (Andrews 2002), the random sample is used to create a random variable. This is then projected onto the parameter space, which is a cone. The projection onto the cone results in the discrete mass at $\alpha = 0$ and the continuous mass over $(0, \infty)$. As noted in (Andrews 2002), this type of a characterization of the asymptotic distribution can be easily programmed and simulated.

4. Small Sample Properties

To investigate the small sample performance, linear IV models are simulated and estimated using TSLS and the ridge estimator. The model is given in Equations (1) to (4) with $k = 2$ and $m = 3$. To standardize the model, set $z_i \sim \text{iid}N(0, I_3)$ and $\beta_0 = (0, 0)'$. Endogeneity is created with

$$\begin{bmatrix} \varepsilon_i \\ u_i \end{bmatrix} \sim \text{iid}N\left(0, \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0 \\ 0.7 & 0 & 1 \end{bmatrix}\right).$$

The strength of the instrument signal is controlled by the parameter[3] δ in

$$\Gamma_0 = \begin{bmatrix} 1 & 0 \\ 0 & \delta \\ 1 & 0 \end{bmatrix}.$$

To judge the behavior of the estimator, three different dimensions of the model are adjusted.

1. Sample size. For smaller sample sizes, the ridge estimator should have better properties whereas for larger sample sizes, TSLS should perform better. We consider sample sizes of $n = 25$, 50, 250 and 500.
2. Precision. Signal strength of the instruments is one way to vary precision. The instrument signal strength decreases with the value of δ above, conditional on holding the other model parameters fixed. For lower precision settings or smaller signal strengths the ridge estimator should perform better. We consider values of $\delta = 0.1$, 0.25, 0.5 and 1. Note that while $\delta = 1$ leads to a high precision setting for all sample sizes considered, $\delta = 0.1$ leads to a low precision setting in smaller samples and a high precision setting in larger samples.
3. Prior value relative to β_0. For the prior closer to the population parameter values the ridge estimator should perform relatively better. We consider values of β^p which were (a) one standard deviation[4] from the true value $\beta^p = (1/\sqrt{2}, 1/\sqrt{2})'$, (b) two standard deviations

[3] Similar results are obtained via other specifications of Γ_0. These are included as part of Supplementary Materials for the paper, available from the authors on request.
[4] Each individual error term is standard normal.

from the true value $\beta^p = (\sqrt{2}, \sqrt{2})'$, and (c) three standard deviations from the true value[5] $\beta^p = (3/\sqrt{2}, 3/\sqrt{2})'$.

We simulate a total of 48 model specifications corresponding to 4 sample sizes n, 4 values of the precision parameter δ and 3 values of the prior β^p. Each specification is simulated 10,000 times and both TSLS and ridge estimator are estimated. We compare estimated β_0 values on bias, variance and MSE. For the ridge estimator we use $\tau = 0.7$ to split the sample between training and test samples.[6]

The regularization parameter α is selected in two steps—first, we search in the log-space going from 10^{-5} to 10^6; second, we perform a grid search[7] in a linear space around the value selected in the first step. A final selected value of $\hat{\alpha} = 0$ in the second step corresponds to a "no regularization" scenario which implies the ridge estimator ignores the prior in favor of the data and the value $\hat{\alpha} = 10^7$ corresponds to an "infinite regularization" scenario which implies the ridge estimator ignores the data in favor of the prior.

Tables 1 and 2 compare the performance of the TSLS estimator with the ridge estimator for different precision levels and sample sizes when the prior is fixed at $\beta^p = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$ and $\beta^p = (\frac{3}{\sqrt{2}}, \frac{3}{\sqrt{2}})'$ respectively. Recall, our parameter of interest is $\beta_0 = (\beta_1, \beta_2)' = (0, 0)'$. We compare the estimators based on (a) bias, (b) standard deviation of the estimates, (c) MSE values of the estimates and (d) sum of MSE values of $\hat{\beta_1}$ and $\hat{\beta_2}$. In both tables, the TSLS estimator performs as expected—both bias and standard deviation of estimates fall as sample size increases and as instrument signal strength increases. In smaller samples, the TSLS estimators exhibit some bias, which confirms that TSLSL estimators are consistent but not unbiased. Table 1 presents a scenario where the prior for the ridge estimator is one standard deviation away from the true parameter estimate. We note that in the low precision setting of $\delta = 0.1$ the ridge estimator has lower MSE for all sample sizes considered in the simulations. However as precision improves, we note that for larger sample sizes the TSLS estimator has lower MSE. Table 2 describes a scenario where the ridge estimator does not have any particular advantage since it is biased to a prior which is 3 standard deviations away from the true parameter value. However, even when prior values are far from true parameter values, there are a number of scenarios where the ridge estimator outperforms the TSLS estimator in terms of MSE. In particular, in small samples and low precision settings, the ridge estimator leads to smaller MSE. When $\delta = 0.1$, the ridge estimator leads to lower MSE values for all sample sizes except $n = 500$. When $\delta = 1$ and the model has high precision, the ridge estimator has higher MSE than TSLS. Thus as the signal strength improves and low precision issues subside, TSLS dominates. The bias-variance trade-off is at work here. Consider the results corresponding to $n = 25$ and $\delta = 0.25$. The ridge estimator has *higher* bias compared to the TSLS estimator for both parameters, however this is compensated by considerably smaller standard deviation values leading to smaller MSE. This table also demonstrates scenarios where for a given δ value, as the sample size increases the estimator with lower MSE changes from ridge to TSLS. For $\delta = 0.25$, the ridge estimator performs better for sample sizes $n \leq 50$ whereas TSLS performs better for $n \geq 250$. Similarly, for $\delta = 0.50$, the ridge estimator outperforms TSLS only for the smallest sample size of $n = 25$.

Figures 1–4 present scatter plots of the estimates from TSLS and ridge estimator with different priors for the following cases: (a) low precision, small sample size; (b) low precision, large sample size; (c) high precision, small sample size; (d) high precision, large sample size. These figures demonstrate the influence of the priors. The prior pulls the ridge estimates away from the population parameter values. For low precision models ($\delta = 0.1$), the variance associated with TSLS estimates is larger than the ridge estimates, even in larger sample sizes. The ridge estimator is biased towards the prior

---

[5] Other specifications of prior values also led to similar results. These are included as part of Supplementary Materials for the paper, available from the authors on request.

[6] The best value of τ is unclear. All the simulations reported in this section were also performed with $\tau = 0.5$ and $\tau = 0.9$. The results were similar to $\tau = 0.7$. The full set of simulations is available in the Supplementary Materials.

[7] We consider a linear grid of 10,000 points in the the second step.

which is demonstrated by the estimates not being distributed symmetrically around the true value. On the other hand, for high precision models ($\delta = 1$) the variance reduction from TSLS for the ridge estimator is not as dramatic. In fact, while the variance reduction appears substantial for the prior value of $\beta^p = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$, it is unclear at least visually if there is a reduction in variance for a poorly specified prior at $\beta^p = (\frac{3}{\sqrt{2}}, \frac{3}{\sqrt{2}})'$. In larger samples with high precision (Figure 4) the TSLS estimates outperform the ridge estimators which is demonstrated by larger clouds which are slightly off-center from the true parameter values. However, ridge estimators using different priors are still competitive and don't lead to a drastically worse performance (as a reference compare the performance of the TSLS estimates to the ridge estimates in Figure 1).

Table 1. Estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ using TSLS and ridge estimator for $\beta^p = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$. The ridge estimator leads to smaller combined MSE (highlighted in bold) when precision is low ($\delta = 0.10$). This drop in MSE is driven primarily by large reductions in standard deviations of the estimates. The TSLS estimator leads to smaller combined MSE when precision is high ($\delta = 1.00$). For intermediate precision models the ridge estimator leads to smaller combined MSE in small samples.

δ	n	Estimator	$\hat{\beta}_1$			$\hat{\beta}_2$			$(\hat{\beta}_1, \hat{\beta}_2)$
			Bias	**SD**	**MSE**	**Bias**	**SD**	**MSE**	**MSE**
0.10	25	TSLS	0.013	0.232	0.054	0.630	1.514	2.690	2.744
		Ridge	0.100	0.138	0.029	0.677	0.403	0.621	**0.650**
	50	TSLS	0.006	0.189	0.036	0.542	1.428	2.333	2.368
		Ridge	0.062	0.100	0.014	0.646	0.548	0.717	**0.731**
	250	TSLS	−0.000	0.081	0.007	0.204	1.511	2.323	2.330
		Ridge	0.021	0.044	0.002	0.498	0.385	0.396	**0.398**
	500	TSLS	−0.000	0.041	0.002	0.060	0.763	0.585	0.587
		Ridge	0.014	0.032	0.001	0.401	0.349	0.282	**0.283**
0.25	25	TSLS	0.008	0.216	0.047	0.322	1.156	1.440	1.486
		Ridge	0.101	0.145	0.031	0.549	0.467	0.519	**0.550**
	50	TSLS	0.002	0.147	0.022	0.179	1.085	1.209	1.231
		Ridge	0.062	0.098	0.013	0.461	0.342	0.329	**0.343**
	250	TSLS	−0.001	0.047	0.002	−0.002	0.297	0.088	**0.091**
		Ridge	0.016	0.045	0.002	0.215	0.263	0.116	0.118
	500	TSLS	0.000	0.032	0.001	0.000	0.188	0.035	**0.036**
		Ridge	0.009	0.032	0.001	0.143	0.211	0.065	0.066
0.50	25	TSLS	0.002	0.198	0.039	0.053	0.751	0.566	0.605
		Ridge	0.099	0.158	0.035	0.338	0.362	0.245	**0.280**
	50	TSLS	−0.000	0.113	0.013	0.005	0.402	0.162	0.175
		Ridge	0.057	0.104	0.014	0.239	0.264	0.127	**0.141**
	250	TSLS	−0.001	0.045	0.002	−0.000	0.130	0.017	**0.019**
		Ridge	0.014	0.045	0.002	0.088	0.147	0.029	0.032
	500	TSLS	0.000	0.032	0.001	−0.000	0.091	0.008	**0.009**
		Ridge	0.009	0.032	0.001	0.057	0.108	0.015	0.016
1.0	25	TSLS	−0.002	0.163	0.026	−0.004	0.244	0.060	**0.086**
		Ridge	0.090	0.164	0.035	0.144	0.221	0.070	0.105
	50	TSLS	0.000	0.106	0.011	0.001	0.153	0.023	**0.035**
		Ridge	0.054	0.107	0.014	0.095	0.155	0.033	0.047
	250	TSLS	−0.001	0.045	0.002	−0.000	0.064	0.004	**0.006**
		Ridge	0.018	0.048	0.003	0.035	0.073	0.007	0.009
	500	TSLS	0.000	0.032	0.001	−0.000	0.045	0.002	**0.003**
		Ridge	0.013	0.034	0.001	0.024	0.053	0.003	0.005

Table 2. Estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ using TSLS and ridge estimator for $\beta^p = (\frac{3}{\sqrt{2}}, \frac{3}{\sqrt{2}})'$. The prior is 3 standard deviations away from the true parameter value. The ridge estimator outperforms the TSLS estimator in terms of MSE values in a number of cases. In particular, in small samples and low precision settings, the ridge estimator leads to smaller MSE values.

δ	n	Estimator	$\hat{\beta}_1$			$\hat{\beta}_2$			$(\hat{\beta}_1, \hat{\beta}_2)$
			Bias	**SD**	**MSE**	**Bias**	**SD**	**MSE**	**MSE**
0.10	25	TSLS	0.012	0.232	0.054	0.629	1.516	2.693	2.747
		Ridge	0.090	0.222	0.058	1.050	0.895	1.903	**1.961**
	50	TSLS	0.006	0.189	0.036	0.546	1.425	2.328	2.363
		Ridge	0.051	0.152	0.026	1.000	0.972	1.944	**1.970**
	250	TSLS	−0.000	0.081	0.007	0.205	1.511	2.325	2.332
		Ridge	0.012	0.074	0.006	0.686	1.196	1.902	**1.908**
	500	TSLS	−0.000	0.041	0.002	0.058	0.764	0.588	**0.589**
		Ridge	0.006	0.035	0.001	0.489	0.614	0.615	0.617
0.25	25	TSLS	0.008	0.216	0.047	0.324	1.160	1.451	1.498
		Ridge	0.085	0.217	0.054	0.786	0.850	1.340	**1.394**
	50	TSLS	0.002	0.147	0.022	0.178	1.082	1.202	1.223
		Ridge	0.046	0.127	0.018	0.629	0.718	0.910	**0.928**
	250	TSLS	−0.001	0.047	0.002	−0.003	0.297	0.088	**0.091**
		Ridge	0.008	0.042	0.002	0.225	0.318	0.151	0.153
	500	TSLS	0.000	0.032	0.001	−0.000	0.188	0.035	**0.036**
		Ridge	0.005	0.030	0.001	0.136	0.215	0.065	0.066
0.50	25	TSLS	0.002	0.199	0.040	0.053	0.748	0.562	0.602
		Ridge	0.077	0.190	0.042	0.412	0.561	0.485	**0.527**
	50	TSLS	−0.000	0.113	0.013	0.005	0.402	0.161	**0.174**
		Ridge	0.041	0.108	0.013	0.263	0.364	0.201	0.215
	250	TSLS	−0.001	0.045	0.002	−0.001	0.130	0.017	**0.019**
		Ridge	0.011	0.044	0.002	0.086	0.149	0.030	0.032
	500	TSLS	0.000	0.032	0.001	−0.000	0.091	0.008	**0.009**
		Ridge	0.008	0.031	0.001	0.056	0.108	0.015	0.016
1.0	25	TSLS	−0.002	0.162	0.026	−0.003	0.244	0.060	**0.086**
		Ridge	0.076	0.171	0.035	0.142	0.256	0.086	0.121
	50	TSLS	0.000	0.106	0.011	0.001	0.153	0.023	**0.035**
		Ridge	0.048	0.108	0.014	0.093	0.162	0.035	0.049
	250	TSLS	−0.001	0.045	0.002	−0.000	0.064	0.004	**0.006**
		Ridge	0.017	0.048	0.003	0.034	0.074	0.007	0.009
	500	TSLS	0.000	0.032	0.001	−0.000	0.045	0.002	**0.003**
		Ridge	0.012	0.034	0.001	0.024	0.053	0.003	0.005

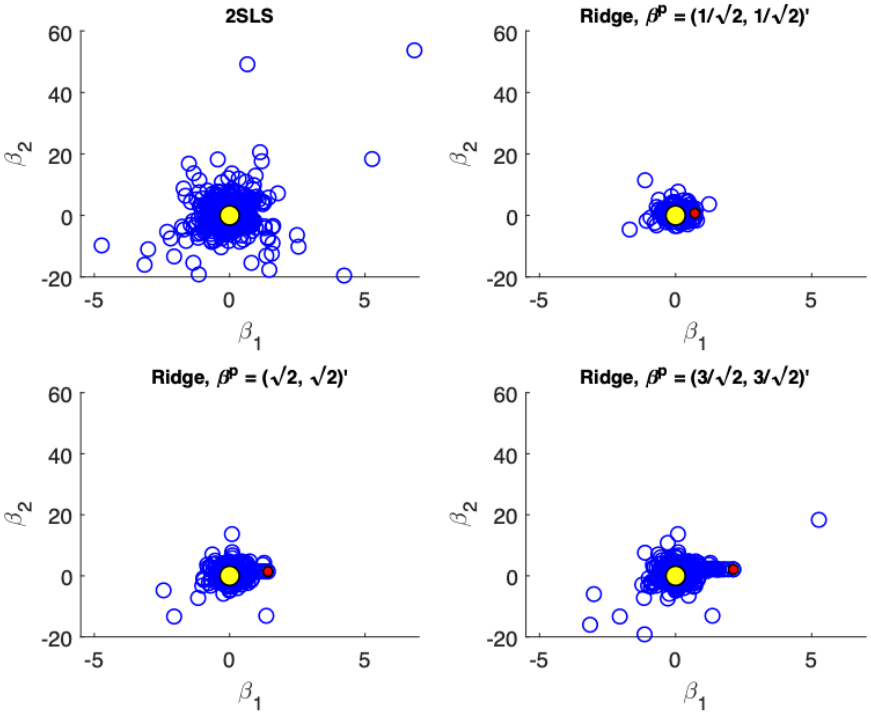

Figure 1. Scatter plots of the estimates from TSLS and ridge estimator with different priors when precision is low ($\delta = 0.1$) and sample size is small ($n = 25$). Estimates, the true parameter value and prior values are represented by blue, yellow and red points respectively. The variance associated with TSLS estimates is much larger than the ridge estimates. The ridge estimator is biased toward the prior.

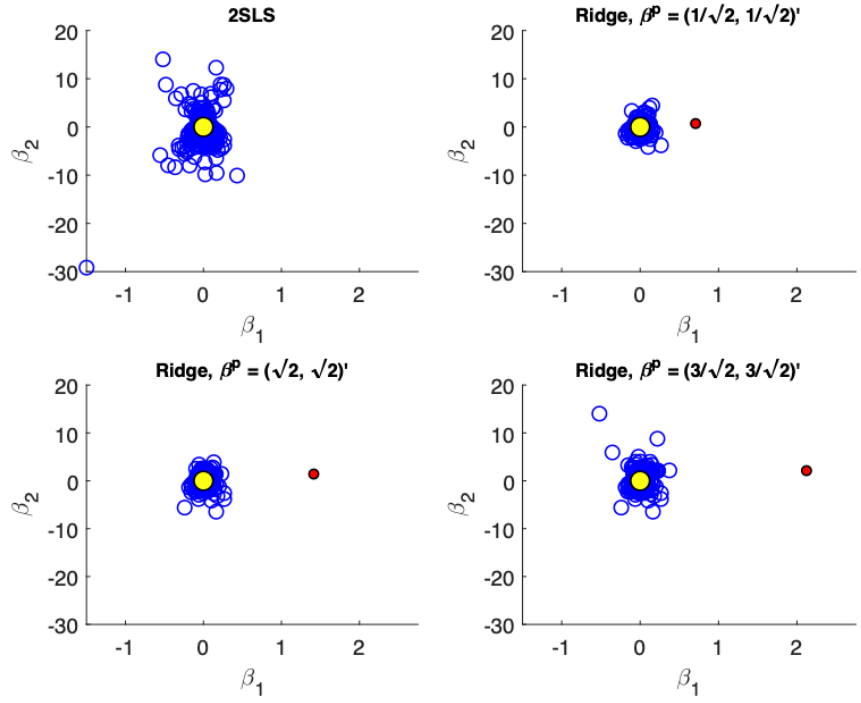

Figure 2. Scatter plots of the estimates from TSLS and ridge estimator with different priors when precision is low ($\delta = 0.1$) and sample size is large ($n = 500$). Estimates, the true parameter value and prior values are represented by blue, yellow and red points respectively. The variance associated with TSLS estimates is much larger than the ridge estimates. The ridge estimator is less biased towards the prior in the larger samples.

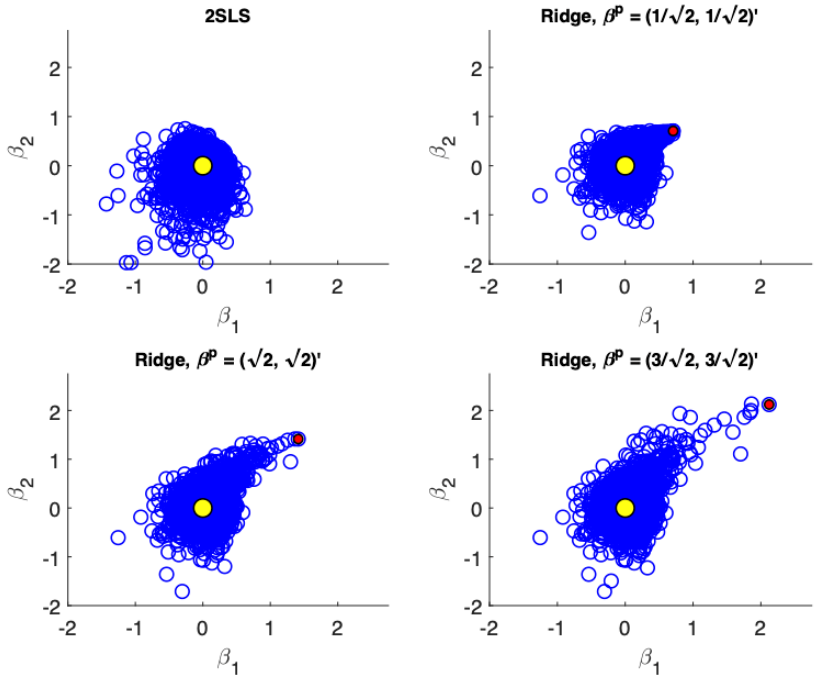

Figure 3. Scatter plots of the estimates from TSLS and ridge estimator with different priors when precision is high ($\delta = 1$) and sample size is small ($n = 25$). Estimates, the true parameter value and prior values are represented by blue, yellow and red points respectively. TSLS performance is much better in this setting. The variance reduction for the ridge estimator is not as dramatic.

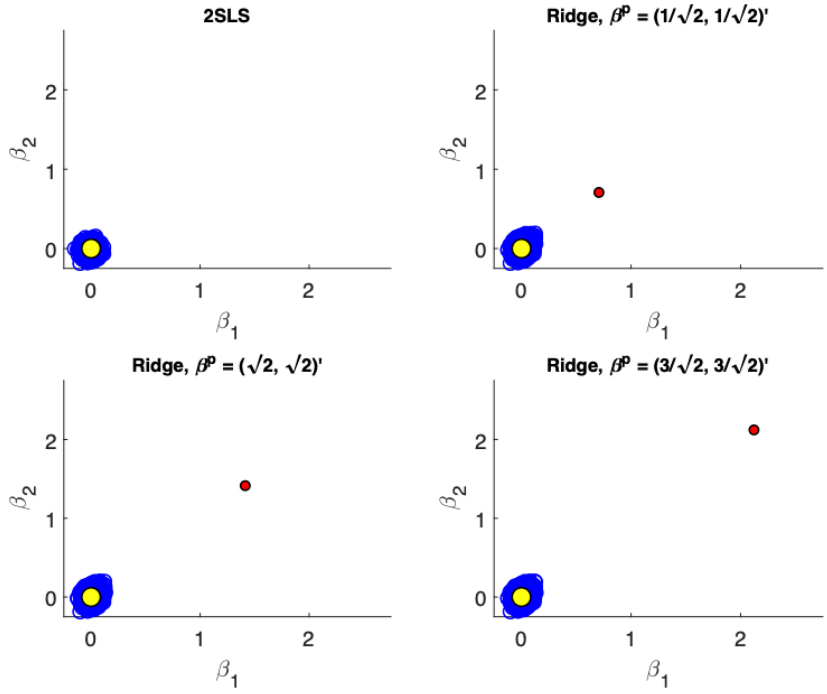

Figure 4. Scatter plots of the estimates from TSLS and ridge estimator with different priors when precision is high ($\delta = 1$) and sample size is large ($n = 500$). Estimates, the true parameter value and prior values are represented by blue, yellow and red points respectively. The TSLS estimates outperform the ridge estimators which is demonstrated by marginally larger clouds which are slightly off-center from the true parameter values for the ridge estimators. However, the ridge estimator using different priors is still competitive.

Table 3, summarizes the distribution of the estimated regularization parameter $\hat{\alpha}$ for different precision levels, sample sizes and prior values. Recall Theorem 1 implies the asymptotic distribution will be a mixed distribution with some discrete mass at $\alpha = 0$. Table 3 reports the proportion of cases which correspond to "no regularization" ($\hat{\alpha} = 0$), "infinite regularization" ($\hat{\alpha} = 10^7 \approx \infty$) and "some regularization" ($\hat{\alpha} \in (0, 10^7)$). In all cases, there is a substantial mass of the distribution concentrated at $\hat{\alpha} = 0$. On the other hand we note that except in the cases where the prior is located at the true parameter value, there is no mass concentrated at $\hat{\alpha} \approx \infty$. We see some interesting variations corresponding to different prior values. In low precision settings (particularly $\delta = 0.1$), keeping sample size fixed, as the prior moves away from the true value, the proportion of cases with "no regularization" increases whereas the proportion of cases with "some regularization" falls. Similarly for high precision settings (particularly $\delta = 1$), as the sample size increases, the proportion of cases with "no regularization" increases whereas the proportion of cases with "some regularization" falls. In this table we also present results for large sample sizes of $n = 10{,}000$, which demonstrate that the mass at $\hat{\alpha} = 0$ approaches 50% asymptotically, as predicted by Theorem 1. Distributions of $\hat{\alpha}$ for large sample sizes of $n = 10{,}000$ via histograms are presented in Figure 5.

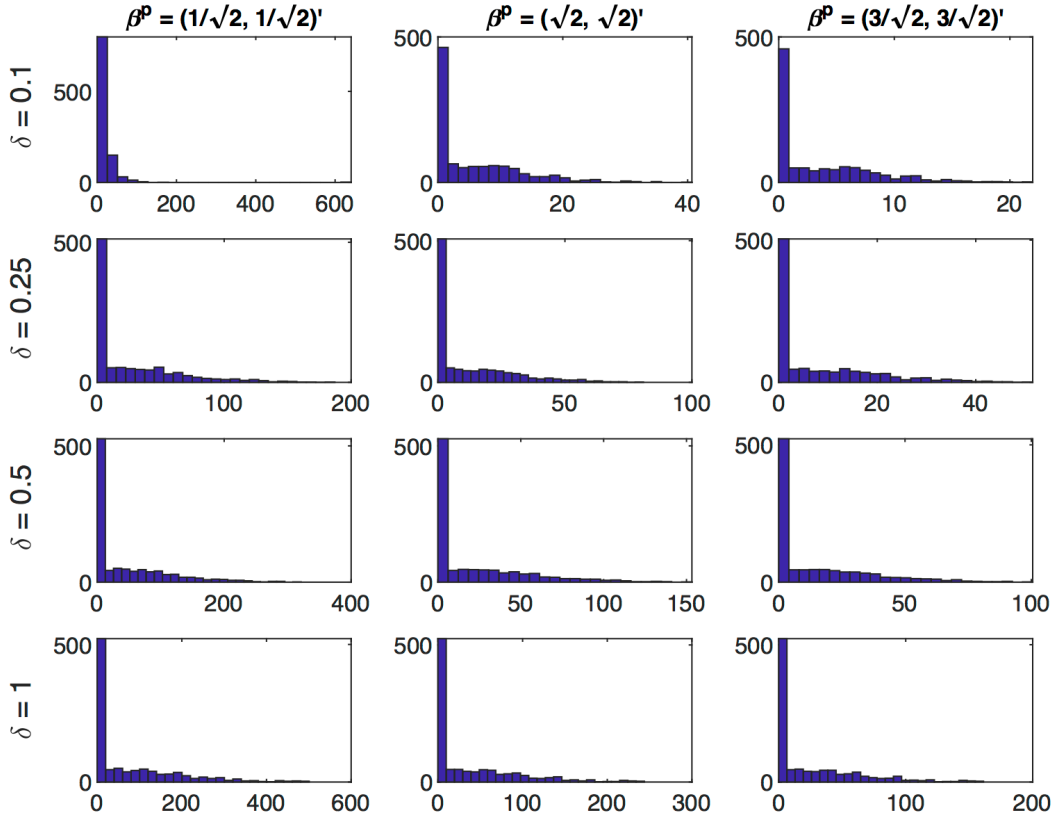

Figure 5. This figure plots the histogram of estimated regularization parameter $\hat{\alpha}$ when $n = 10{,}000$ for all precision parameters and all priors considered in the simulations. The total number of simulations to generate each of these plots is $N = 1000$. As predicted by Theorem 1, the mass at $\hat{\alpha} = 0$ is approaching 50% asymptotically. Distributions of $\hat{\alpha}$ values for all cases considered are presented in Table 3.

Table 3. Distribution of regularization parameter $\hat{\alpha}$. The mixed distribution associated with the finite samples is in agreement with the nonstandard asymptotic distribution given in Theorem 1. The proportion of cases with "no regularization" ($\hat{\alpha} = 0$), "some regularization" ($\hat{\alpha} \in (0, 10^7)$) and "infinite regularization" ($\hat{\alpha} = 10^7 \approx \infty$) are presented. For all cases, there is a substantial mass of the distribution concentrated at $\hat{\alpha} = 0$. On the other hand, there is no mass concentrated at $\hat{\alpha} \approx \infty$ *except* in very small samples of $n = 25$. As predicted by Theorem 1, the mass at $\hat{\alpha} = 0$ is approaching 50% asymptotically. Histograms for the large sample cases of $n = 10{,}000$ are presented in Figure 5.

δ	n	$\beta^p = (1/\sqrt{2}, 1/\sqrt{2})'$			$\beta^p = (\sqrt{2}, \sqrt{2})'$			$\beta^p = (3/\sqrt{2}, 3/\sqrt{2})'$		
		$\hat{\alpha} = 0$	$\hat{\alpha} \in (0, 10^7)$	$\hat{\alpha} = 10^7 \approx \infty$	$\hat{\alpha} = 0$	$\hat{\alpha} \in (0, 10^7)$	$\hat{\alpha} = 10^7$	$\hat{\alpha} = 0$	$\hat{\alpha} \in (0, 10^7)$	$\hat{\alpha} = 10^7$
0.01	25	0.164	0.834	0.003	0.262	0.738	0.001	0.339	0.661	0.000
	50	0.166	0.834	0.000	0.275	0.725	0.000	0.354	0.646	0.000
	250	0.190	0.810	0.000	0.281	0.719	0.000	0.319	0.681	0.000
	500	0.220	0.780	0.000	0.285	0.715	0.000	0.302	0.698	0.000
	10,000	0.413	0.587	0.000	0.411	0.589	0.000	0.411	0.589	0.000
0.25	25	0.176	0.822	0.002	0.262	0.737	0.001	0.315	0.684	0.001
	50	0.184	0.816	0.000	0.263	0.737	0.000	0.299	0.701	0.000
	250	0.293	0.707	0.000	0.309	0.691	0.000	0.314	0.686	0.000
	500	0.346	0.654	0.000	0.354	0.646	0.000	0.354	0.646	0.000
	10,000	0.461	0.539	0.000	0.465	0.535	0.000	0.465	0.535	0.000
0.50	25	0.216	0.780	0.004	0.262	0.737	0.001	0.284	0.716	0.000
	50	0.255	0.745	0.000	0.284	0.716	0.000	0.294	0.706	0.000
	250	0.369	0.631	0.000	0.374	0.626	0.000	0.376	0.624	0.000
	500	0.412	0.588	0.000	0.415	0.585	0.000	0.417	0.583	0.000
	10,000	0.463	0.537	0.000	0.467	0.533	0.000	0.466	0.534	0.000
1.00	25	0.287	0.708	0.005	0.310	0.689	0.001	0.318	0.681	0.000
	50	0.333	0.667	0.000	0.346	0.654	0.000	0.351	0.649	0.000
	250	0.413	0.587	0.000	0.418	0.582	0.000	0.419	0.581	0.000
	500	0.439	0.561	0.000	0.443	0.557	0.000	0.442	0.558	0.000
	10,000	0.474	0.526	0.000	0.478	0.522	0.000	0.477	0.523	0.000

Table 4 presents summaries of the smallest singular value of the matrix[8] $\left(\frac{-X'Z}{n}\right)$ for different values of δ and n. The estimated asymptotic standard deviation is inversely related to the smallest singular value, or equivalently smaller singular values are associated with flatter objective functions at their minimum values. As the precision parameter increases from $\delta = 0.1$ to $\delta = 1$, the mean of the smallest singular value increases. As the sample size increases, the variance of the smallest singular values decreases.

Table 4. Summary statistics of the smallest singular value for the matrix $\left(-\frac{X'Z}{n}\right)$ corresponding to different precision parameter values δ and sample sizes n, using 10,000 samples each. As the precision parameters increase from $\delta = 0.1$ to $\delta = 1$, the mean of the smallest singular value increases. As sample sizes increase from $n = 25$ to $n = 10{,}000$, the spread in the smallest singular value decreases.

δ	n	Mean	Std Dev	1st Quartile	Median	3rd Quartile
0.10	25	0.25	0.14	0.14	0.23	0.33
	50	0.19	0.10	0.12	0.18	0.26
	250	0.12	0.05	0.08	0.12	0.16
	500	0.11	0.04	0.08	0.11	0.14
	2500	0.10	0.02	0.09	0.10	0.12
	5000	0.10	0.01	0.09	0.10	0.11
	10,000	0.10	0.01	0.09	0.10	0.11
0.25	25	0.32	0.17	0.19	0.30	0.42
	50	0.28	0.13	0.19	0.27	0.37
	250	0.26	0.07	0.21	0.26	0.30
	500	0.25	0.05	0.22	0.25	0.29
	2500	0.25	0.02	0.24	0.25	0.26
	5000	0.25	0.01	0.24	0.25	0.26
	10,000	0.25	0.01	0.24	0.25	0.26
0.50	25	0.50	0.21	0.35	0.48	0.63
	50	0.50	0.17	0.39	0.49	0.61
	250	0.50	0.08	0.45	0.50	0.55
	500	0.50	0.05	0.46	0.50	0.54
	2500	0.50	0.02	0.48	0.50	0.52
	5000	0.50	0.02	0.49	0.50	0.51
	10,000	0.50	0.01	0.49	0.50	0.51
1.00	25	0.86	0.27	0.67	0.85	1.03
	50	0.92	0.21	0.77	0.91	1.05
	250	0.98	0.10	0.91	0.98	1.05
	500	0.99	0.08	0.94	0.99	1.04
	2500	1.00	0.03	0.98	1.00	1.02
	5000	1.00	0.02	0.98	1.00	1.02
	10,000	1.00	0.02	0.99	1.00	1.01

5. Returns to Education

This section revisits the question of returns to schooling in Angrist and Krueger (1991). The key insight of that paper was the use of quarter of birth indicator variables as instruments to uniquely identify the impact of years of education on wages. The data are Public use Micro Sample of the 1980 U.S. Census and includes men born between 1920 and 1949 with positive earnings in 1979 and no missing observations. The sample is divided into three data sets, one for each decade. The empirical results are summarized in Table 5.

[8] This corresponds to the estimate of $E\left[\frac{\partial g_i(\beta)}{\partial \beta'}\right]$ where $g_i(\beta) = (y_i - x_i\beta)z_i$.

Table 5. Effect of years of education on the log of weekly earnings.

	1920–1929	**1930–1939**	**1940–1949**
OLS	0.070	0.063	0.052
TSLS	0.058	0.099	−0.073
Ridge	0.027	0.066	0.001
First stage F-test	38.3	30.5	26.3
Overidentification, Basmann	0.776	2.321	9.693
{p-value}	{0.679}	{0.313}	{0.008}
$\lambda_{\min}$	3.3×10^{-6}	1.3×10^{-5}	6.6×10^{-7}
Condition Number	4.1×10^{7}	1.2×10^{7}	2.8×10^{8}
Sample size, n	247,199	329,509	486,926

The specification explains the log of weekly wages with the years of education; with race, standard metropolitan statistical area, marital status, region dummies, and year of birth dummies as controls[9]. The same data set was used in (Staiger and Stock 1997) which focused on weak instruments using the quarter of birth interacted with year of birth to create 30 instruments. To avoid weak instruments, we restrict attention to the quarter of birth dummy variables as instruments. This gives three instruments for years of education. The resulting first stage F-tests give no indication of weak instruments. The appropriateness of the specification is tested with the Basmann test for overidentify restrictions. The specification is not rejected for the 1920s and 1930s. However, the specification is rejected for the 1940s.

Each sample is split into a training sample with $\tau = 0.7$ and a testing sample. The empirical results will be sensitive to this randomization. For each decade, the data was read into R from the Stata data set from the Angrist Data Archive at MIT, dplyr was used to filter the data for each decade and the random seed was set in R with "set.seed(12345)". The samples have hundreds of thousands of observations, n, with 22 parameters estimated. However, there are precision problems with these estimates. As noted above in Section 2, the precision can be judged by the magnitude of the smallest eigenvalue of the second derivative of the objective function at the TSLS estimate for the entire sample, these are denoted $\lambda_{\min}$. The precision can also be judged with the condition number for the second derivatives which varies between 12 million and 282 million.

The prior for the ridge estimates was set empirically to judge the variability in the parameter estimates in the least informative dimension of the parameter space. The prior was set at four times the eigenvector associated with the smallest eigenvalue of the second derivative of the TSLS objective function for the training sample. This is moving four unit lengths away from the TSLS estimate in the flattest (least informative) direction. The ridge estimator is then defined by the value of α that minimizes the TSLS objective function using the test sample.

Even with the large sample sizes, the prior impacts the estimates. For the 1930s and 1940s the ridge estimate is between the OLS and the TSLS estimates. For the 1920s the ridge estimate is below both the TSLS estimate and the OLS estimate.

A final simulation exercise compares TSLS estimates with ridge estimates using the Angrist and Krueger (1991) data. As above the specification explains the log of weekly wages with the years of education; with race, standard metropolitan statistical area, marital status, region dummies,

---

[9] The specification with exogenous variables W, (Staiger and Stock 1997)

$$\begin{aligned} Y &= \tilde{X}a + Wb + u \\ \tilde{X} &= \tilde{Z}c + Wd + \tilde{\varepsilon} \end{aligned}$$

fits into the specification of Equations (1) and (2) with $X = \begin{bmatrix} \tilde{X} & W \end{bmatrix}$, $Z = \begin{bmatrix} \tilde{Z} & W \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \tilde{\varepsilon} & 0 \end{bmatrix}$, $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$, and $\Gamma = \begin{bmatrix} c & 0 \\ d & I \end{bmatrix}$.

and year of birth dummies as controls. The quarter of birth dummies provide three instruments for years of education. Using the parameter estimates obtained by running the TSLS estimation on the entire sample from 1930s we obtain residuals $\hat{u}$ and $\hat{\varepsilon}$ and their estimated covariance matrix. We then draw $N = 10{,}000$ random samples of size n with replacement for the instruments and control variables and use these along with the full sample parameter estimates and error covariance to obtain simulated values of years of education and log of weekly wages.

TSLS and ridge estimates are obtained for each random sample and compared on the basis of bias, standard deviation and root mean square error (RMSE). For the ridge estimates each sample is split into a training sample with $\tau = 0.7$ and a testing sample. Priors for all parameters are set to $\hat{\beta}_{IV,train}$ i.e., the TSLS parameter estimate using only the training sample, except the parameter corresponding to returns to education. Three priors are considered for the returns to education parameter: $\beta^p_{edu} = \frac{1}{2}\left(\hat{\beta}^{IV}_{edu,train}\right)$, $\beta^p_{edu} = \left(\hat{\beta}^{IV}_{edu,train}\right)$ and $\beta^p_{edu} = 2\left(\hat{\beta}^{IV}_{edu,train}\right)$. Results presented in Table 6 demonstrate that for the smallest sample size of n = 1000 ridge estimates corresponding to all three priors produce lower RMSE values for the parameter of interest compared to TSLS. On the other hand for the larger sample size of $n = 100{,}000$ TSLS estimates produce lower RMSE values compared to ridge estimates for all three priors.

Table 6. Simulation results using returns to education data.

Prior	Sample Size	Estimator	Bias	SD	RMSE
$\beta^p_{edu} = \frac{1}{2}\left(\hat{\beta}^{IV}_{edu,train}\right)$	1000	TSLS	−0.032	0.185	0.188
		Ridge	−0.027	0.164	**0.166**
	100,000	TSLS	−0.001	0.039	**0.039**
		Ridge	−0.0003	0.058	0.058
$\beta^p_{edu} = \left(\hat{\beta}^{IV}_{edu,train}\right)$	1000	TSLS	−0.032	0.192	0.194
		Ridge	−0.026	0.166	**0.168**
	100,000	TSLS	−0.002	0.039	**0.039**
		Ridge	−0.004	0.062	0.062
$\beta^p_{edu} = 2\left(\hat{\beta}^{IV}_{edu,train}\right)$	1000	TSLS	−0.036	0.181	0.184
		Ridge	−0.023	0.134	**0.136**
	100,000	TSLS	−0.001	0.038	**0.038**
		Ridge	−0.007	0.050	0.051

6. Conclusions

The asymptotic distribution of the ridge estimator when the tuning parameter is selected with a test sample has been characterized. This estimator incorporates a non-zero prior and allows the non-negative tuning parameter to be zero. The resulting asymptotic distribution is a mixture with discrete mass on zero for the tuning parameter, a novel result, which follows from the true value of the tuning parameter lying on the boundary of the parameter space.

Simulations demonstrate where the ridge estimator produced lower MSE than the TSLS estimator, specifically when model precision is low, particularly in smaller samples and often even in larger samples. Where the TSLS estimator has lower MSE, particularly when precision is high, the ridge estimator remains competitive. As an empirical application, we have applied this procedure to the returns from education dataset used in (Angrist and Krueger 1991). Importantly, even with over 200,000 observations, the prior still influences the point estimates.

The ridge estimator will be particularly useful in addressing applied empirical questions where TSLS is appropriate but the available data suffers from low precision or where the sample size available is small. The characterization the asymptotic distribution of the ridge estimator also provides a useful framework for other estimators that involve tuning parameters.

Extensions and improvements of the approach are worthwhile to pursue. Different loss functions can be applied to the test sample (e.g., k-fold cross validation), we can allow for multiple tuning parameters, consider models without a closed form solution, allow the number of covariates to grow with the sample size, allow the number of tuning parameters to grow with the sample size, consider situation with weak instruments or nearly weak instruments, allow other penalty terms such as the LASSO or elastic-net.

Supplementary Materials: The following are available online at http://www.mdpi.com/2225-1146/8/4/39/s1, File S1: Supplementary Material: On the Asymptotic Distribution of Ridge Regression Estimators using Training and Test Samples.

Author Contributions: The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Lemma 1

The objective function that determines the optimal tuning parameter is given in Equation (10). As the sample size grows the objective function uniformly converges to a deterministic function that takes a unique local minimum at $\alpha = 0$. The parameter space is bounded and the law of large numbers implies

$$\lim_{n\to\infty} Q_{n(1-\tau)}(\alpha) = \frac{1}{2}(\beta_0 - \beta^p)' \left(\frac{\Gamma_0' R_z \Gamma_0}{\alpha} + I_k\right)^{-1} \Gamma_0' R_z \Gamma_0 \left(\frac{\Gamma_0' R_z \Gamma_0}{\alpha} + I_k\right)^{-1} (\beta_0 - \beta^p)$$

which is minimized at $\alpha = 0$. Hence $\alpha_0 = 0$. When $\alpha = 0$ then $\hat{\beta}_{tr}(0) \to \beta_0$.

The root-n consistency of $\hat{\alpha}$ follows from the standard approach of Lemma 5.4 in Ichimura (1993). The needed results are that $\frac{dQ_{n(1-\tau)}(\alpha_0)}{d\alpha}$ satisfies a CLT and $\frac{d^2Q_{n(1-\tau)}(\alpha)}{d\alpha^2}$ is continuous (from the right hand side) at α_0 and $\frac{d^2Q_{n(1-\tau)}(\alpha_0)}{d\alpha^2}$ limits to a positive value. These derivatives reduce to the derivatives of $\hat{\beta}_{tr}(\alpha) = \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I\right)^{-1} \left(\frac{X_{\tau n}' P_{Z_{\tau n}} Y_{\tau n}}{[\tau n]} + \alpha\beta^p\right)$ wrt α. The first derivative is

$$\begin{aligned}
\frac{d\hat{\beta}_{tr}(\alpha)}{d\alpha} &= \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-1} \beta^p - \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-2} \left(\frac{X_{\tau n}' P_{Z_{\tau n}} Y_{\tau n}}{[\tau n]} + \alpha\beta^p\right) \\
&= \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-1} \left(\beta^p - \hat{\beta}_{tr}(\alpha)\right).
\end{aligned}$$

The second derivative is

$$\begin{aligned}
\frac{d^2\hat{\beta}_{tr}(\alpha)}{d\alpha^2} &= -\left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-1} \frac{d\hat{\beta}_{tr}(\alpha)}{d\alpha} - \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-2} \left(\beta^p - \hat{\beta}_{tr}(\alpha)\right) \\
&= -\left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-2} \left(\beta^p - \hat{\beta}_{tr}(\alpha)\right) - \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-2} \left(\beta^p - \hat{\beta}_{tr}(\alpha)\right) \\
&= -2\left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-2} \left(\beta^p - \hat{\beta}_{tr}(\alpha)\right).
\end{aligned}$$

Now determine the derivatives of $Q_{n(1-\tau)}(\alpha)$. The first derivative is

$$\begin{aligned}
\frac{dQ_{n(1-\tau)}(\alpha)}{d\alpha} &= \frac{-1}{(n-[\tau n])} \left(Y_{n(1-\tau)} - X_{n(1-\tau)}\hat{\beta}_{tr}(\alpha)\right)' P_{Z_{n(1-\tau)}} X_{n(1-\tau)} \frac{d\hat{\beta}_{tr}(\alpha)}{d\alpha} \\
&= \frac{-1}{(n-[\tau n])} \left(Y_{n(1-\tau)} - X_{n(1-\tau)}\hat{\beta}_{tr}(\alpha)\right)' P_{Z_{n(1-\tau)}} X_{n(1-\tau)} \left(\frac{X_{\tau n}' P_{Z_{\tau n}} X_{\tau n}}{[\tau n]} + \alpha I_k\right)^{-1} \left(\beta^p - \hat{\beta}_{tr}(\alpha)\right).
\end{aligned}$$

Evaluate at $\alpha_0 = 0$

$$\begin{aligned}\frac{dQ_{n(1-\tau)}(0)}{d\alpha} &= \frac{-1}{(n-\lfloor\tau n\rfloor)}\left(Y_{n(1-\tau)} - X_{n(1-\tau)}\hat{\beta}_{tr}(0)\right)' P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\left(\beta^p - \hat{\beta}_{tr}(0)\right)\\ &= \frac{-1}{(n-\lfloor\tau n\rfloor)}\left(\left(Y_{n(1-\tau)} - X_{n(1-\tau)}\beta_0\right) - X_{n(1-\tau)}\left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)'\\ &\quad\times P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)\\ &= \frac{-1}{(n-\lfloor\tau n\rfloor)}\left(\varepsilon'_{n(1-\tau)} - \frac{\varepsilon'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1} X'_{n(1-\tau)}\right)\\ &\quad\times P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\left(\beta^p - \beta_0 - \left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\frac{X'_{\tau n}P_{Z_{\tau n}}\varepsilon_{\tau n}}{\lfloor\tau n\rfloor}\right).\end{aligned}$$

The CLT applies to the $\varepsilon'_{n(1-\tau)}Z_{n(1-\tau)}$ and $\varepsilon'_{\tau n}Z_{\tau n}$ terms. The others converge by LLN. Hence

$$\begin{aligned}\sqrt{(n-\lfloor\tau n\rfloor)}\frac{dQ_{n(1-\tau)}(0)}{d\alpha} &= \frac{-1}{\sqrt{(n-\lfloor\tau n\rfloor)}}\left(\varepsilon'_{n(1-\tau)} - \frac{\varepsilon'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1} X'_{n(1-\tau)}\right)\\ &\quad\times P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\left(\beta^p - \beta_0\right) + o_p(1).\end{aligned}$$

The second derivative is

$$\begin{aligned}&\frac{d^2Q_{n(1-\tau)}(\alpha)}{d\alpha^2}\\ &= \frac{-1}{(n-\lfloor\tau n\rfloor)}\left(Y_{n(1-\tau)} - X_{n(1-\tau)}\hat{\beta}_{tr}(\alpha)\right)' P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\frac{d^2\hat{\beta}_{tr}(\alpha)}{d\alpha^2}\\ &\quad+\frac{1}{(n-\lfloor\tau n\rfloor)}\left(X_{n(1-\tau)}\frac{d\hat{\beta}_{tr}(\alpha)}{d\alpha}\right)' P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\frac{d\hat{\beta}_{tr}(\alpha)}{d\alpha}\\ &= \frac{2}{(n-\lfloor\tau n\rfloor)}\left(Y_{n(1-\tau)} - X_{n(1-\tau)}\hat{\beta}_{tr}(\alpha)\right)' P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor} + \alpha I_k\right)^{-2}\left(\beta^p - \hat{\beta}_{tr}(\alpha)\right)\\ &\quad+\frac{1}{(n-\lfloor\tau n\rfloor)}\left(\beta^p - \hat{\beta}_{tr}(\alpha)\right)'\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor} + \alpha I_k\right)^{-1} X'_{n(1-\tau)}\\ &\qquad\times P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor} + \alpha I_k\right)^{-1}\left(\beta^p - \hat{\beta}_{tr}(\alpha)\right).\end{aligned}$$

This is a bounded continuous function. Now evaluate at $\alpha_0 = 0$

$$\begin{aligned}\frac{d^2Q_{n(1-\tau)}(0)}{d\alpha^2} &= \frac{2}{(n-\lfloor\tau n\rfloor)}\left(\left(Y_{n(1-\tau)} - X_{n(1-\tau)}\beta_0\right) - X_{n(1-\tau)}\left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)'\\ &\qquad\times P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-2}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)\\ &\quad+\frac{1}{(n-\lfloor\tau n\rfloor)}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)'\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\\ &\qquad\times X'_{n(1-\tau)}P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)\\ &= \frac{2}{(n-\lfloor\tau n\rfloor)}\left(\varepsilon_{n(1-\tau)} - X_{n(1-\tau)}\left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)'\\ &\qquad\times P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-2}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)\\ &\quad+\frac{1}{(n-\lfloor\tau n\rfloor)}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right)'\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\\ &\qquad\times X'_{n(1-\tau)}P_{Z_{n(1-\tau)}} X_{n(1-\tau)}\left(\frac{X'_{\tau n}P_{Z_{\tau n}}X_{\tau n}}{\lfloor\tau n\rfloor}\right)^{-1}\left(\beta^p - \beta_0 - \left(\hat{\beta}_{tr}(0) - \beta_0\right)\right).\end{aligned}$$

The first term will converge to zero and the second term converges to the positive value

$$\left(\beta^p - \beta_0\right)'\left(\Gamma'_0 R_z \Gamma_0\right)\left(\beta^p - \beta_0\right).$$

Now follow the standard approach (Lemma 5.4, Ichimura 1993) to show that $\sqrt{n}(\hat{\alpha} - \alpha_0) = O_p(1)$. Expand $Q_{n(1-\tau)}(\alpha)$ about α_0 and evaluate at $\hat{\alpha}$.

$$Q_{n(1-\tau)}(\hat{\alpha}) = Q_{n(1-\tau)}(\alpha_0) + \frac{dQ_{n(1-\tau)}(\alpha_0)}{d\alpha}(\hat{\alpha} - \alpha_0) + \frac{1}{2}\frac{d^2Q_{n(1-\tau)}(\bar{\alpha})}{d\alpha^2}(\hat{\alpha} - \alpha_0)^2$$

where $0 \leq \bar{\alpha} \leq \hat{\alpha}$. Because $\hat{\alpha} = \arg\min_{[0,\infty)} Q_{n(1-\tau)}(\alpha)$, $0 \geq Q_{n(1-\tau)}(\hat{\alpha}) - Q_{n(1-\tau)}(\alpha_0)$, hence

$$0 \quad \geq \quad \frac{dQ_{n(1-\tau)}(\alpha_0)}{d\alpha}(\hat{\alpha}-\alpha_0) + \frac{1}{2}\frac{d^2Q_{n(1-\tau)}(\bar{\alpha})}{d\alpha^2}(\hat{\alpha}-\alpha_0)^2.$$

Multiply both sides by $\frac{n}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)^2}$.

$$\begin{aligned} 0 &\geq \frac{dQ_{n(1-\tau)}(\alpha_0)}{d\alpha}(\hat{\alpha}-\alpha_0)\frac{n}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)^2} + \frac{1}{2}\frac{d^2Q_{n(1-\tau)}(\bar{\alpha})}{d\alpha^2}(\hat{\alpha}-\alpha_0)^2\frac{n}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)^2} \\ &= \sqrt{n}\frac{dQ_{n(1-\tau)}(\alpha_0)}{d\alpha}\left(\frac{\sqrt{n}(\hat{\alpha}-\alpha_0)}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)}\right)\frac{1}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)} + \frac{1}{2}\frac{d^2Q_{n(1-\tau)}(\bar{\alpha})}{d\alpha^2}\left(\frac{\sqrt{n}(\hat{\alpha}-\alpha_0)}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)}\right)^2 \end{aligned} \tag{A1}$$

Suppose $\sqrt{n}|\hat{\alpha}-\alpha_0|$ diverged to infinity. As noted above $\sqrt{n}\frac{dQ_{n(1-\tau)}(\alpha_0)}{d\alpha} = O_p(1)$. In addition, $\left(\frac{\sqrt{n}(\hat{\alpha}-\alpha_0)}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)}\right) = O_p(1)$. However, $\frac{1}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)} = o_p(1)$ and hence the first term on the LHS of Equation (A1) goes to zero. This means

$$o_p(1) \quad \geq \quad \frac{1}{2}\frac{d^2Q_{n(1-\tau)}(\bar{\alpha})}{d\alpha^2}\left(\frac{\sqrt{n}(\hat{\alpha}-\alpha_0)}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)}\right)^2.$$

However, $\frac{d^2Q_{n(1-\tau)}(\bar{\alpha})}{d\alpha^2}$ limits to $\frac{d^2Q_{n(1-\tau)}(\alpha_0)}{d\alpha^2}$, a positive value, and the RHS can satisfy this only if

$$\frac{\sqrt{n}(\hat{\alpha}-\alpha_0)}{(1+\sqrt{n}|\hat{\alpha}-\alpha_0|)} = o_p(1).$$

This occurs only if $\sqrt{n}|\hat{\alpha}-\alpha_0| = o_p(1)$ which is a contradiction of the assumption that $\sqrt{n}|\hat{\alpha}-\alpha_0|$ diverges. Hence $\sqrt{n}(\hat{\alpha}-\alpha_0) = O_p(1)$. □

Appendix B. Proof of Theorem 1

This is a direct application of Theorem 1 from (Andrews 2002). Assumptions A1–A5 (GMM1*–GMM5*) in (Andrews 2002) are satisfied for the linear model by Assumptions 1–3. To show how the assumptions in (Andrews 2002) are satisfies, we first use Assumtions 1–3 to demonstrate three useful results for the system of Equation (13). The useful results are: $E[h_i(\theta_0)] = 0$, $\sqrt{n}H_n(\theta_0)$ satisfies a central limit theorem and $\left(\lim_{n\to\infty}\frac{\partial H_n(\theta_0)}{\partial\theta'}\right)^{-1}$ exists, which requires showing that LLN leads to a matrix which is invertible. In the statement of the Theorem, the limiting random variable, Z, is composed of two terms: $\sqrt{n}H_n(\theta_0)$ and $\left(-E\left[\frac{\partial h_i(\theta_0)}{\partial\theta'}\right]\right)^{-1}$.

Evaluate the moment condition, Equation (13), at θ_0, to show that $E[h_i(\theta_0)] = 0$ and that $\sqrt{n}H_n(\theta_0)$ satisfies a central limit theorem.

$$H_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{1}_\tau(i)\text{vech}(R_z - z_iz_i') \\ (1-\mathbf{1}_\tau(i))\text{vech}(R_z - z_iz_i') \\ \mathbf{1}_\tau(i)\text{vec}(R_z\Gamma_0 - z_ix_i') \\ (1-\mathbf{1}_\tau(i))\text{vec}(R_z\Gamma_0 - z_ix_i') \\ -\mathbf{1}_\tau(i)\left(\Gamma_0'R_zR_z^{-1}z_i(y_i - x_i'\beta_0)\right) \\ (1-\mathbf{1}_\tau(i))(y_i - x_i'\beta_0)z_i'R_z^{-1}R_z\Gamma_0\left(\Gamma_0'R_zR_z^{-1}R_z\Gamma_0\right)^{-1}(\beta^p - \beta_0) \\ -\left(\Gamma_0'R_zR_z^{-1}z_i(y_i - x_i'\beta_0)\right) \end{bmatrix}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{1}_\tau(i)\text{vech}(R_z - z_i z_i') \\ (1-\mathbf{1}_\tau(i))\text{vech}(R_z - z_i z_i') \\ \mathbf{1}_\tau(i)\text{vec}\left(R_z\Gamma_0 - z_i u_i' - z_i z_i'\Gamma_0\right) \\ -(1-\mathbf{1}_\tau(i))\text{vec}(R_z\Gamma_0 - u_i z_i' - z_i z_i'\Gamma_0) \\ \mathbf{1}_\tau(i)\left(\Gamma_0' z_i \varepsilon_i\right) \\ (1-\mathbf{1}_\tau(i))\varepsilon_i z_i'\Gamma_0\left(\Gamma_0' R_z \Gamma_0\right)^{-1}(\beta^p - \beta_0) \\ -\left(\Gamma_0' z_i \varepsilon_i\right) \end{bmatrix}$$

Each element of $h_i(\theta_0)$ has expectation zero and bounded covariance, hence the iid assumption implies the central limit theorem

$$\sqrt{n}H_n(\theta_0) \sim^A N\left(0, \begin{bmatrix} \tau\chi & 0 & \tau\xi & 0 & 0 & 0 & 0 \\ 0 & (1-\tau)\chi & 0 & (1-\tau)\xi & 0 & 0 & 0 \\ \tau\xi' & 0 & \tau\zeta & 0 & \tau\Psi & 0 & \tau\Psi \\ 0 & (1-\tau)\xi' & 0 & (1-\tau)\zeta & 0 & (1-\tau)\Pi & (1-\tau)\Psi \\ 0 & 0 & \tau\Psi' & 0 & \tau\Xi & 0 & \tau\Xi \\ 0 & 0 & 0 & (1-\tau)\Pi' & 0 & (1-\tau)\Upsilon & (1-\tau)\Phi' \\ 0 & 0 & \tau\Psi' & (1-\tau)\Psi' & \tau\Xi & (1-\tau)\Phi & \Xi \end{bmatrix}\right)$$

where

$$\chi = E\left[\text{vech}(R_z - z_i z_i')\text{vech}(R_z - z_i z_i')'\right],$$
$$\xi = E\left[\text{vech}(R_z - z_i z_i')\text{vec}(R_z\Gamma_0 - z_i z_i'\Gamma_0)'\right],$$
$$\zeta = E\left[\ \text{vec}(R_z\Gamma_0 - z_i x_i')\text{vec}(R_z\Gamma_0 - z_i x_i')'\right],$$
$$\Psi = E\left[\text{vec}(z_i u_i')\left(\varepsilon_i z_i'\Gamma_0\right)\right],$$
$$\Pi = E\left[\text{vec}(-u_i z_i')\varepsilon_i z_i'\Gamma_0(\Gamma_0' R_z\Gamma_0)^{-1}(\beta_0 - \beta^p)\right],$$
$$\Xi = (\Gamma_0' R_z\Gamma_0)\sigma_\varepsilon^2,$$
$$\Upsilon = \sigma_\varepsilon^2(\beta_0 - \beta^p)'(\Gamma_0' R_z\Gamma_0)^{-1}(\beta_0 - \beta^p), \text{ and}$$
$$\Phi = -\sigma_\varepsilon^2(\beta_0 - \beta^p)'.$$

The expectation of the first derivative of the moment conditions evaluated at θ_0 is

$$E\left[\frac{\partial h_i(\theta_0)}{\partial\theta'}\right] =$$
$$\begin{bmatrix} \tau I_{\frac{m(m+1)}{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & (1-\tau)I_{\frac{m(m+1)}{2}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \tau I_{mp} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (1-\tau)I_{mp} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \tau(\Gamma_0' R_z\Gamma_0) & \tau(\beta_0 - \beta^p) & 0 \\ 0 & 0 & 0 & 0 & -(1-\tau)(\beta_0 - \beta^p)' & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & (\beta_0 - \beta^p) & (\Gamma_0' R_z\Gamma_0) \end{bmatrix}$$

The general structure of the last matrix is

$$\begin{bmatrix} A & B \\ \kappa B' & C \end{bmatrix}$$

where

$$A = \begin{bmatrix} \tau I_{\frac{m(m+1)}{2}} & 0 & 0 & 0 & 0 \\ 0 & (1-\tau) I_{\frac{m(m+1)}{2}} & 0 & 0 & 0 \\ 0 & 0 & \tau I_{mp} & 0 & 0 \\ 0 & 0 & 0 & (1-\tau) I_{mp} & 0 \\ 0 & 0 & 0 & 0 & \tau(\Gamma_0' R_z \Gamma_0) \end{bmatrix},$$

$B = \begin{bmatrix} 0 & 0 \\ \tau(\beta_0 - \beta^p) & 0 \end{bmatrix}$, $C = \begin{bmatrix} 0 & 0 \\ (\beta_0 - \beta^p) & (\Gamma_0' R_z \Gamma_0) \end{bmatrix}$, and $\kappa = \frac{-(1-\tau)}{\tau}$. The inverse of the general structure is

$$\begin{bmatrix} A^{-1} + A^{-1}B(C - \kappa B' A^{-1} B)^{-1} B A^{-1} & -A^{-1}B(C - \kappa B' A^{-1} B)^{-1} \\ -(C - \kappa B' A^{-1} B)^{-1} B A^{-1} & (C - \kappa B' A^{-1} B)^{-1} \end{bmatrix}$$

which is well defined if the inverses of A and $(C - \kappa B' A^{-1} B)$ exist. The matrix A^{-1} is well defined because Assumption 3 implies $(\Gamma_0' R_z \Gamma_0)$ is full rank. The matrix

$$\left(C - \kappa B' A^{-1} B\right) = \begin{bmatrix} (1-\tau)(\beta_0 - \beta^p)' \left(\Gamma_0' R_z \Gamma_0\right)^{-1} (\beta_0 - \beta^p) & 0_{1\times p} \\ (\beta_0 - \beta^p) & (\Gamma_0' R_z \Gamma_0) \end{bmatrix}$$

with inverse

$$\left(C - \kappa B' A^{-1} B\right)^{-1} = \begin{bmatrix} \frac{1}{(1-\tau)(\beta_0 - \beta^p)'(\Gamma_0' R_z \Gamma_0)^{-1}(\beta_0 - \beta^p)} & 0_{1\times p} \\ -\frac{(\Gamma_0' R_z \Gamma_0)^{-1}(\beta_0 - \beta^p)}{(1-\tau)(\beta_0 - \beta^p)'(\Gamma_0' R_z \Gamma_0)^{-1}(\beta_0 - \beta^p)} & (\Gamma_0' R_z \Gamma_0)^{-1} \end{bmatrix}.$$

Hence $\left(-E\left[\frac{\partial h_i(\theta_0)}{\partial \theta'}\right]\right)^{-1}$ is well defined. Now verify Assumptions A1–A5 (GMM1*–GMM5*) in Andrews (2002).

Assumption A1 (GMM1*). *This parameter space is bounded. Because z_i has finite fourth moments and $\begin{bmatrix} \varepsilon_i & u_i' \end{bmatrix}'$ has a finite second moment there exists a dominating function with a finite expectation. This implies that $H_n(\theta)' H_n(\theta)$ will uniformly converge to its limiting function, $E[H_n(\theta)']E[H_n(\theta)]$. Identification follows from $E[H_n(\theta_0)] = 0$ and the invertibility of $E\left[\frac{\partial h_i(\theta_0)}{\partial \theta'}\right]$.*

Assumption A2 (GMM2*). *The data are iid. The GMM structure is presented above. The expectation of the first derivative of the moment conditions is evaluated at θ_0 and inverted, hence demonstrating it is full rank. $E[H_n(\theta_0)] = 0$ is demonstrated above. The system is just identified, so an identity weighting matrix is used.*

Assumption A3 (GMM3*). *The CLT applies because the data are iid and z_i has finite fourth moments, $\begin{bmatrix} \varepsilon_i & u_i' \end{bmatrix}'$ has a finite second moment and the z_i and $\begin{bmatrix} \varepsilon_j & u_j' \end{bmatrix}'$ are independent for all i and j.*

Assumption A4 (GMM4*). *Because the eigenvalues of R_z are bounded above zero and below infinity each element of R_z and R_z^{-1} is bounded above. Hence all the parameters in Θ are bounded and Equation (27) of (Andrews 2002) is satisfied with $c = \max(B_1, B_2, B_3, B_4)$.*

Assumption A5 (GMM5*). *The cone for this problem is $\Lambda = \left\{\lambda \in R^{m(m+1)+2mk+2k+1} : \lambda_{m(m+1)+2mk+k+1} \geq 0\right\}$ which is convex.*

□

References

Andrews, Donald W. K. 2002. Generalized method of moments estimation when a parameter is on a boundary. *Journal of Business & Economic Statistics* 20: 530–44.

Angrist, Joshua D., and Alan B. Krueger. 1991. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* 106: 979–1014. [CrossRef]

Antoine, Bertille, and Eric Renault. 2009. Efficient gmm with nearly-weak instruments. *The Econometrics Journal* 12: S135–S171. [CrossRef]

Bengio, Yoshua. 2000. Gradient-based optimization of hyperparameters. *Neural Computation* 12: 1889–900. [PubMed]

Bickel, Peter J., Bo Li, Alexandre B. Tsybakov, Sara A. van de Geer, Bin Yu, Teófilo Valdés, Carlos Rivero, Jianqing Fan, and Aad van der Vaart. 2006. Regularization in statistics. *Test* 15: 271–344. [CrossRef]

Carrasco, Marine, and Guy Tchuente. 2016. Efficient estimation with many weak instruments using regularization techniques. *Econometric Reviews* 35: 1609–37. [CrossRef]

Carrasco, Marine, and Jean-Pierre Florens. 2000. Generalization of gmm to a continuum of moment conditions. *Econometric Theory* 16: 797–834. [CrossRef]

Carrasco, Marine, Jean-Pierre Florens, and Eric Renault. 2007. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics* 6: 5633–751.

Carrasco, Marine. 2012. A regularization approach to the many instruments problem. *Journal of Econometrics* 170: 383–98. [CrossRef]

Chen, Chen, Min Ren, Min Zhang, and Dabao Zhang. 2018. A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research* 19: 40–73.

Dorugade, Ashok Vithoba. 2014. New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences* 15: 94–99. [CrossRef]

Firinguetti, Luis, and Gladys Bobadilla. 2011. Asymptotic confidence intervals in ridge regression based on the edgeworth expansion. *Statistical Papers* 52: 287–307. [CrossRef]

Habibnia, Ali, and Esfandiar Maasoumi. 2019. Forecasting in big data environments: An adaptable and automated shrinkage estimation of neural networks (aashnet). *arXiv*. arXiv:1904.11145.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. Unsupervised learning. In *The Elements of Statistical Learning*. New York: Springer, pp. 485–585.

Hoerl, Arthur E., and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12: 55–67. [CrossRef]

Hoerl, Arthur E., Robert W. Kannard, and Kent F. Baldwin. 1975. Ridge regression: Some simulations. *Communications in Statistics* 4: 105–23. [CrossRef]

Ichimura, Hidehiko. 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58: 71–120. [CrossRef]

Larsen, Jan, Lars Kai Hansen, Claus Svarer, and Børje Ola Mattias Ohlsson. 1996. Design and regularization of neural networks: the optimal use of a validation set. Papaer presented at the 1996 IEEE Signal Processing Society Workshop, Kyoto, Japan, September 4–6, pp. 62–71. [CrossRef]

Lawless, J. F., and P. Wang. 1976. A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods 5*: 307–23. [CrossRef]

Leeb, Hannes, and Benedikt M. Pötscher. 2005. Model selection and inference: Facts and fiction. *Econometric Theory* 21: 21–59. [CrossRef]

Lin, Wei, Rui Feng, and Hongzhe Li. 2015. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association 110*: 270–88. [CrossRef]

Maclaurin, Dougal, David Duvenaud, and Ryan P. Adams. 2015. Gradient-based hyperparameter optimization through reversible learning. Paper presented at the 32nd International Conference on International Conference on Machine Learning—Volume 37, ICML'15, Lille, France, July 7–9, pp. 2113–22.

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. 2012. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. New York: Wiley.

Sanderson, Eleanor, and Frank Windmeijer. 2016. A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics* 190: 212–21. [CrossRef]

Staiger, Douglas, and James H. Stock. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica* 65: 557–86. [CrossRef]

Theobald, Chris M. 1974. Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)* 36: 103–6. [CrossRef]

Zhu, Ying. 2018. Sparse linear models and l1-regularized 2SLS with high-dimensional endogenous regressors and instruments. *Journal of Econometrics* 202: 196–213. [CrossRef]