THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 15-22

Nam-Hyun Kim
University of Konstanz

Winfried Pohlmeier
University of Konstanz
The Rimini Centre for Economic Analysis, Italy

A REGULARIZATION APPROACH TO BIASED TWO-STAGE LEAST SQUARES ESTIMATION

A Regularization Approach to Biased Two-Stage Least Squares Estimation*

Nam-Hyun Kim†
University of Konstanz

Winfried Pohlmeier‡
University of Konstanz, COFE, RCEA

May 1, 2015

Abstract

We propose to apply L_2-norm regularization to address the problem of weak and/or many instruments. We observe that the presence of weak instruments, or weak and many instruments is translated into a nearly singular problem in a control function representation. Hence, we show that mean squares error-optimal L_2-norm regularization with a small sample size reduces the bias and variance of the regularized 2SLS estimators with the presence of weak and/or many instruments. A number of different strategies for choosing a regularization parameter are introduced and compared in a Monte Carlo study.

Keywords: weak instruments, control function approach, ridge regression

JEL classification: C26, C36, C52

*Financial support by the German Research Foundation (DFG) through research unit FOR 1882 "Psychoeconomics" is gratefully acknowledged. Authors like to thank Patrik Guggenberger for his thoughtful comments on the paper. Authors also thank Sebastian Bayer for helpful research assistance. Early versions of the paper have been presented at workshops and conferences in Rimini, Pisa, Paderborn and Lancaster. All remaining errors are ours.

†Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2204, fax: -4450, email: nam-hyun.kim@uni-konstanz.de.

‡Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2660, fax: -4450, email: winfried.pohlmeier@uni-konstanz.de.

1 Introduction

A substantial number of studies have examined the performance of the two-stage least squares (2SLS) estimator with the presence of weak and/or many instruments over the past few decades. The presence of weak and/or many instruments is known to cause the 2SLS estimator to be biased toward the ordinary least squares (OLS) estimator in small samples (see Bekker (1994), Bound et al. (1995), Staiger and Stock (1997), and Stock et al. (2002), for details). A number of studies, such as Hahn and Hausman (2002), Chao and Swanson (2005), and Bun and Windmeijer (2011), analytically examine the higher-order approximations of small sample biases of the 2SLS estimator. On the other hand, studies such as that of Nelson and Startz (1990a), Nelson and Startz (1990b), and Cruz and Moreira (2005) investigate the performance of the estimator in small samples by using Monte Carlo simulation exercises.

There are number of alternative estimation methods for addressing the problem of weak and/or many instruments. Hahn et al. (2004) show that the Fuller's adjusted limited information maximum likelihood (FLIML) estimator (see Fuller (1977), for details) reduces the small sample variance of limited information maximum likelihood (LIML) estimator, whereas the biased-corrected version of the 2SLS estimator based on the Jackknife principle reduces the small sample bias. However, Arel-Bundock (2013) addresses the problem by using indirect inference to reduce the small sample bias of the 2SLS estimator. Furthermore, Donald and Newey (2001) address the problem of many instruments by using truncation techniques; alternatively, Carrasco (2012) addresses the same problem by using L_2-norm regularization of parameters in the reduced form. As the results of higher-order approximation show, the small sample bias of the 2SLS estimator in Carrasco (2012) is constant where the number of instruments does not affect the bias. Donald and Newey (2001) also present similar results where the number of truncated instruments plays a similar role to that played by the regularization parameter in Carrasco (2012).

Contrary to Carrasco (2012), we propose to employ L_2-norm regularization of the parameters in a structural equation to address the problem of weak and/or many instruments. Our analysis is based on a control function (CF) representation of a linear model with an endogenous explanatory variable. This CF representation enables us to translate the presence of weak, or weak and many instruments into the nearly singular problem caused by the multicollinearity between the endogenous explanatory variable and the control variable for controlling endogeneity in the structural equation. Additionally, the presence of many instruments causes similar problems to the problems of the nearly singularity such as the poor determination and high variances of the CF estimators. Hence, we implement the L_2-norm regularization of the structural parameters in the CF representation to address the weak and/or many instruments problem. Furthermore, by using the diagonal matrix design conveniently shows the equivalence between the regularized CF 2SLS (CF-2SLS) and the regularized 2SLS estimators. We also show that the mean squares error (MSE)-optimal L_2-norm regularization of the 2SLS estimators substantially reduces the small sample bias and variance.

We first outline the L_2-norm regularized CF-2SLS estimator in the two-stage CF representation with the presence of weak and/or many instruments and show its equivalence with the regularized 2SLS estimator. In Section 3, we further examine the small sample properties of the MSE-optimal regularized 2SLS estimator. In Section 4, we introduce a number of different strategies for choosing a regularization parameter. We also provide evidence of the satisfactory performance of our proposed regularized 2SLS estimators in a comparison with the existing one, particularly FLIML, by using a Monte Carlo study. We finally conclude the paper with the issues needing further investigation in Section 5.

2 A Regularization Approach to Biased 2SLS Estimation

We consider a simple simultaneous equation model with a single endogenous variable and k number of instruments as follows (see Hahn and Hausman (2002), Hahn and Kuersteiner (2002), Bun and Windmeijer (2011), Arel-Bundock (2013), for example):

$$Y = X\beta + \varepsilon \tag{1}$$

$$X = Z\pi + u, \tag{2}$$

where (X, Y, ε, u) are n-vectors and Z is a $n \times k$ matrix of instruments. We assume throughout the paper that Z is a fixed design with $\lim_{n\to\infty} \frac{1}{n} Z'Z$, and is finite and non-singular. We also assume the iid joint normality of the error terms in (1) and (2) as follows:

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{u\varepsilon} & \sigma_u^2 \end{bmatrix} \right).$$

The 2SLS or instrumental variables (IV) estimator of the structural parameter is given below:

$$\hat{\beta}_{2SLS} - \beta = \frac{\hat{X}'\zeta}{\hat{X}'\hat{X}} = \frac{\hat{X}'\varepsilon}{\hat{X}'\hat{X}}, \tag{3}$$

where $\hat{X} = P_Z X$ with $P_Z = Z(Z'Z)^{-1}Z'$, and $\zeta = (X - \hat{X})\beta + \varepsilon$.

The (infeasible) control function representation of (1) is obtained by incorporating the error term from (2) into (1) by using the linear projection of the error term of (1) on the error term of (2). The error term representation of linear projection is $\varepsilon = u\gamma + \nu$, where $\gamma = \frac{\sigma_{\varepsilon u}}{\sigma_u^2}$ and $\nu_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_\nu^2\right)$, which yields the CF representation of (1) as:

$$Y = X\beta + u\gamma + \nu \equiv W\delta + \nu, \tag{4}$$

where $W = \begin{bmatrix} X & u \end{bmatrix}$ and $\delta = \begin{pmatrix} \beta & \gamma \end{pmatrix}'$. By this construction of the new error term in (4) is uncorrelated with X and u (see Chapter 6 of Wooldridge (2010), for details). However, the presence of weak and/or many instruments in (2) causes the problem of near singularity in the CF representation of (4).

In the conventional CF approach, u is replaced by its fitted value in (4). As an alternative, we consider the two-stage CF representation where both u and X are replaced by their fitted values. Hence, the CF-2SLS estimator is given below:

$$\hat{\delta}_{CF-2SLS} - \delta = \left(\hat{W}'\hat{W}\right)^{-1} \hat{W}'\xi,$$

where $\hat{W} = \begin{bmatrix} \hat{X} & \hat{u} \end{bmatrix}$ and $\xi = (X - \hat{X})\beta + (u - \hat{u})\gamma + \nu = \hat{u}(\beta - \gamma) + \varepsilon$. The regressors in the two-stage CF representation are orthogonal by the construction so that $\hat{W}'\hat{W}$ is a diagonal matrix. Hence the CF-2SLS estimator of β is numerically identical to the 2SLS estimator as shown below:

$$\hat{\beta}_{CF-2SLS} - \beta = \iota_1' \left(\hat{\delta}_{CF-2SLS} - \delta \right) = \frac{\hat{X}'\varepsilon}{\hat{X}'\hat{X}}, \tag{5}$$

where $\iota_1 = (1 \quad 0)'$. The higher-order approximation of the two components of the MSE of $\hat{\beta}_{CF-2SLS}$ is well reported in the literature and is reproduced as follows (see Donald and Newey (2001), and Hahn et al. (2004) for details):

$$\text{Bias}(\hat{\beta}_{CF-2SLS}) \approx \gamma \frac{k}{\mu} \equiv \Delta \quad \text{and} \quad \text{Var}(\hat{\beta}_{CF-2SLS}) \approx \frac{\gamma^2 + \sigma_\nu^2/\sigma_u^2}{\mu}, \tag{6}$$

where $\mu = \frac{\pi' Z' Z \pi}{\sigma_u^2}$ denotes the concentration parameter.

The L_2-norm regularized (ridged) CF-2SLS estimator is given by:

$$\hat{\delta}_{CF-2SLS}(\lambda) = \left(\hat{W}'\hat{W} + \lambda I_2 \right)^{-1} \hat{W}'Y, \tag{7}$$

where I_2 is a two dimensional identity matrix and $\lambda \geq 0$ is a ridging parameter controlling the amount of regularization. For a given λ, it is straightforward to show that the regularized CF-2SLS estimator is equivalent to the regularized 2SLS estimator because of the diagonality of $\hat{W}'\hat{W}$ shown below:

$$\hat{\beta}_{2SLS}(\lambda) = \hat{\beta}_{CF-2SLS}(\lambda) = \iota_1' \left(\hat{\delta}_{CF-2SLS}(\lambda) \right) = \frac{\hat{X}'Y}{\hat{X}'\hat{X} + \lambda}. \tag{8}$$

A crucial issue when performing the proposed ridging estimation is the selection of λ. We discuss the selection of an optimal λ minimizing the MSE in a small sample in the following section. A number of different strategies of choosing λ are introduced in the context of a Monte Carlo study in Section 4.

3 MSE-Optimal Regularization

In this section, we first discuss an optimal ridging parameter that minimizes a small sample MSE, λ^*. Next, we analyze the properties of our proposed estimators given λ^*. For convenience, we rewrite the ridged 2SLS estimator in (8) in terms of the conventional 2SLS estimator as

Table 1: Implications of $\lambda^* \geq 0$

Possible cases	γ	β	$\sigma_{\varepsilon u}$
1.	$\gamma \geq 0$	$\beta \geq 0$	$\sigma_{\varepsilon u} \geq 0$
2.	$\gamma \leq 0$	$\beta \geq -\Delta$	$-\sigma_\nu^2 \leq \sigma_{\varepsilon u} \leq 0$
3.	$\gamma \leq 0$	$\beta \leq 0$	$-(\sigma_\nu^2 + \Delta^2 \mu \sigma_u^2) \leq \sigma_{\varepsilon u} \leq 0$
4.	$\gamma \leq 0$	$0 \leq \beta \leq -\Delta$	$\sigma_{\varepsilon u} \leq -\sigma_\nu^2$

follows:

$$\hat{\beta}_{2SLS}(\lambda) = \frac{\hat{X}'\hat{X}}{\hat{X}'\hat{X} + \lambda}\,\hat{\beta}_{2SLS}. \tag{9}$$

Hereafter, we assume the orthonormality of the higher order approximation of the expectation of $\hat{X}$ without loss of generality for notational simplicity. We then obtain the approximations of the two components of the MSE of $\hat{\beta}_{2SLS}(\lambda)$ in (9) below:

$$\text{Bias}(\hat{\beta}_{2SLS}(\lambda)) \approx \frac{\Delta - \lambda\beta}{1+\lambda} \quad \text{and} \quad \text{Var}(\hat{\beta}_{2SLS}(\lambda)) \approx \left(\frac{1}{1+\lambda}\right)^2 \frac{\gamma^2 + \sigma_\nu^2/\sigma_u^2}{\mu}. \tag{10}$$

It is well known that the variance of a ridged estimator decreases as λ increases (i.e. $\text{Var}(\hat{\beta}_{2SLS}) \geq \text{Var}(\hat{\beta}_{2SLS}(\lambda))$ for any value of $\lambda \geq 0$). We, show, however that the bias of our proposed ridged 2SLS estimator (the first approximation in (10)) is smaller than the bias of the un-ridged 2SLS estimator given λ^*. This is contrary to the conventional ridging estimation in a linear model.

We choose λ^* by minimizing the approximation of the MSE of $\hat{\beta}_{2SLS}(\lambda)$ as follows:

$$\lambda^* = \frac{\Delta^2\mu\sigma_u^2 + \Delta\mu\beta\sigma_u^2 + \sigma_{\varepsilon u} + \sigma_\nu^2}{\mu\beta^2\sigma_u^2 + \Delta\mu\beta\sigma_u^2} \geq 0. \tag{11}$$

The result of $\lambda^* \geq 0$ in (11) suggests a few possible cases, which are used to show our claim about the smaller finite sample bias of $\hat{\beta}_{2SLS}(\lambda^*)$ compared to the bias of $\hat{\beta}_{2SLS}$. The implications of (11) are summarized in Table 1 above. Given that $\lambda^* \geq 0$, we now show that $\hat{\beta}_{2SLS}(\lambda^*)$ has a smaller bias than $\hat{\beta}_{2SLS}$ as follows. Firstly, we show that $\text{Bias}^2(\hat{\beta}_{2SLS})$ is greater than $\text{Bias}^2(\hat{\beta}_{2SLS}(\lambda^*))$. Furthermore, we examine the convexity of $\text{Bias}^2(\hat{\beta}_{2SLS}(\lambda))$.

First, we compare the bias squares, as shown below:

$$\text{Bias}^2(\hat{\beta}_{2SLS}) - \text{Bias}^2(\hat{\beta}_{2SLS}(\lambda^*)) = \frac{\lambda^{*2}\Delta^2 + 2\lambda^*\Delta^2 + 2\lambda^*\Delta\beta - \lambda^{*2}\beta^2}{1 + 2\lambda^* + \lambda^{*2}} \geq 0.$$

We now examine the convexity of $\text{Bias}^2(\hat{\beta}_{2SLS}(\lambda))$ as follows:

$$(1-\theta)\text{Bias}^2(\hat{\beta}_{2SLS}) + \theta\text{Bias}^2(\hat{\beta}_{2SLS}(\lambda^*)) \geq \text{Bias}^2((1-\theta)\hat{\beta}_{2SLS} + \theta\hat{\beta}_{2SLS}(\lambda^*)), \quad (12)$$

for all $0 \leq \theta \leq 1$. Let us now consider each term in (12) as shown below:

$$\begin{aligned}(1-\theta)\text{Bias}^2(\hat{\beta}_{2SLS}) + \theta\text{Bias}^2(\hat{\beta}_{2SLS}(\lambda^*)) &\approx (1-\theta)\Delta^2 + \theta\left(\frac{\Delta - \lambda^*\beta}{1+\lambda^*}\right)^2 \\ &= \frac{\Delta^2 + \lambda^{*2}(\theta\beta^2 + \Delta^2 - \theta\Delta^2) + \lambda^*(2\Delta^2 - 2\theta\Delta^2 - 2\theta\Delta\beta)}{\lambda^{*2} + 2\lambda^* + 1}, \quad (13)\end{aligned}$$

and:

$$\begin{aligned}\text{Bias}^2((1-\theta)\hat{\beta}_{2SLS} + \theta\hat{\beta}_{2SLS}(\lambda^*)) &\approx \left(\frac{-\theta\lambda^*\beta + \Delta(1+\lambda^* - \theta\lambda^*)}{1+\lambda^*}\right)^2 \\ &= \frac{\Delta^2 + \lambda^{*2}(\Delta^2 + \theta^2\Delta^2 + \theta^2\beta^2 - 2\theta\beta\Delta + 2\theta^2\beta\Delta - 2\theta\Delta^2) + \lambda^*(2\Delta^2 - 2\theta\Delta\beta - 2\theta\Delta^2)}{\lambda^{*2} + 2\lambda^* + 1} \quad (14)\end{aligned}$$

We obtain the result of (12) by comparing the two approximations in (13) and (14). Figure 1 below illustrates our claim that the bias and variance tradeoff of $\hat{\beta}_{2SLS}(\lambda)$ is contrary to the conventional ridging estimation. At the optimal point where $\lambda = \lambda^*$, $\hat{\beta}_{2SLS}(\lambda^*)$ has a smaller variance and bias than $\hat{\beta}_{2SLS}$. We use a simple data generating process with parameter values such that $\beta = 1$, $n = 100$, $k = 10$, $\mu = 0.1$, $\sigma_\nu^2 = 10$, $\sigma_u^2 = 1$ and $\gamma = 0.5$ to generate Figure 1. [1]

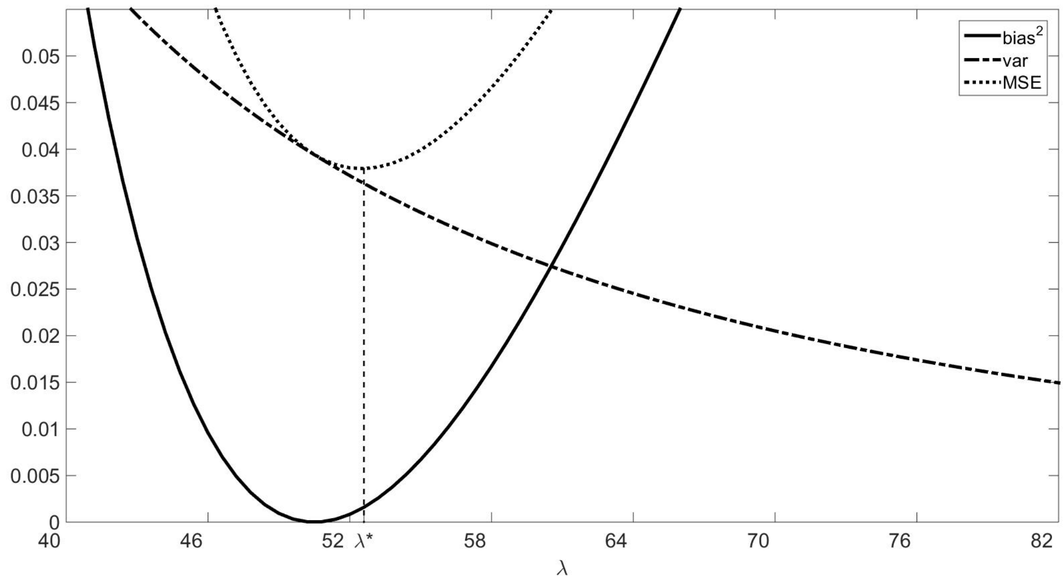
4 Monte Carlo Evidence

In this section, we first introduce and discuss a number of different strategies for the selection of a ridging parameter, then describe our design for a Monte Carlo study. We present the results from the Monte Carlo study based on only selected number of strategies for choosing λ due to space limitations.[2]

A crucial issue when performing our proposed estimation is the selection of λ. We introduce and briefly discuss a number of different strategies for selecting λ and conduct the Monte Carlo study based on the selection strategies below. First we consider the data-driven approach as follows. We propose to compute the ridging parameter as a function of the first-stage of

[1]The figures based on other parameter values can be obtained from the authors on request.
[2]The results based on other strategies can be obtained from the authors on request.

Figure 1: Illustration of the bias and variance tradeoff of L_2-norm regularized 2SLS estimators



F-statistics, $\hat{\lambda}_F = \hat{F}^{-1}$. Furthermore, we compute the ridging parameter minimizes the estimated MSE of the orthonormal design model following the method of Hoerl et al. (1975) such that $\hat{\lambda}_{HKB} = k\hat{\sigma}_u^2/(\hat{\beta}'_{2SLS}\hat{\beta}_{2SLS})$ choosing small values in many settings. Hence we also perform the ridging estimation by using the regularization suggested by Lindley and Smith (1972), which can be derived from the Bayesian framework. Both methods are easy to implement and are superior compared to the 2SLS estimator. Finally, we choose λ by using the mean of a leave-one-out cross-validation for the ridged 2SLS estimator where λ minimizes the mean squared prediction error as shown below:

$$\hat{\lambda}_{CV} = \underset{\lambda \in \mathbb{R}_{++}}{\arg\min} \frac{1}{n}\left(Y - \hat{Y}_{2SLS}^{-i}(\lambda)\right)'\left(Y - \hat{Y}_{2SLS}^{-i}(\lambda)\right), \tag{15}$$

where $\hat{Y}_{2SLS}^{-i}(\lambda) = X\hat{\beta}_{2SLS}^{-i}(\lambda)$ and $\hat{\beta}_{2SLS}^{-i}(\lambda)$ denotes the leave-one-out ridged 2SLS estimator of β.

On the other hand, we consider the ridging parameter written as a function of $\sqrt{n}$ such that $\lambda = \sqrt{n}$. Despite the slower convergence rate compared to the conventional case where a ridge parameter is constant, this simple approach does not suffer from additional estimation noise when estimating λ.

Our Monte Carlo study consists of 24 different designs for the data generating process with

$N = 5000$ replications per design. The design closely follows the one used by Hahn et al. (2004) for an easy comparison. We distinguish among cases with a single instrument ($k = 1$) and many instruments ($k = 10$), two different sample sizes ($n = (100, 500)$), and two correlations of ε and u such that $\rho = (0.5, 0.9)$. The two error terms are drawn from the bivariate normal distribution with $\sigma_\varepsilon = \sigma_u = 1$. The strength of the instruments is determined by fixing the theoretical R^2 and by assuming that the reduced-form parameters are of the same size, $\pi_j = \psi, \forall j = 1, \ldots, k$. This gives a relationship between the theoretical R^2 and the reduced form parameters of the form $R^2 = \frac{k\psi^2}{k\psi^2+1}$, given $z_i \sim N(0, I_k)$. We further distinguish among a weak, semi-weak and strong instrument cases with a corresponding theoretical $R^2 = (0.01, 0.05, 0.3)$, by following Hahn et al. (2004). Note that the concentration parameter can be conveniently approximated by R^2 such that $\mu \approx nR^2/(1 - R^2)$ (see Hahn et al. (2004), for details). The true parameter value of the structural equation is $\beta = 1$.

The results in Table 2 suggest interesting implications for the implementation of our proposed regularized 2SLS estimators with the presence of weak and/or many instruments. The selection strategy based on the ridge parameter, particularly $\lambda = \sqrt{n}$, demonstrates the most preferable performance compared to all other selection strategies in all 24 settings under investigation. The excellent performance of the ridged estimator with $\lambda = \sqrt{n}$, a notably small MSE (0.0621), is observed even in the case of severe endogeneity ($\rho = 0.9$) with the presence of weak ($R^2 = 0.01$) and many instruments ($k = 10$), given that the MSE of FLIML is 0.6755. Overall, ridged estimation with $\hat{\lambda}_F$ provides moderately good performance among the three approaches in Table 2 $\left(\lambda = \sqrt{n}, \hat{\lambda}_F \text{ and } \hat{\lambda}_{CV(1)}\right)$. Furthermore, the regularized 2SLS estimator with $\hat{\lambda}_{CV(1)}$ performs as well as the existing FLIML estimator with the presence of weak instruments. In particular, in the case of severe endogeneity ($\rho = 0.9$) with the presence of many weak instruments ($R^2 = 0.01$ and $k = 10$), the MSE of ridged 2SLS with $\lambda = \sqrt{n}$ is 0.641 compared to the MSE of FLIML (0.6755). In the cases of the moderate ($R^2 = 0.05$) and strong ($R^2 = 0.3$) instruments, the ridged 2SLS estimator with $\lambda = \sqrt{n}$ provides slightly smaller MSEs than the FLIML estimator, unlike ridged 2SLS estimators with $\hat{\lambda}_F$ and $\hat{\lambda}_{CV(1)}$.

In addition to the results reported in Table 2, we investigate a number of alternatives including the LIML estimator and the biased-corrected versions of the 2SLS estimators based on bootstrapping and four different Jackknife sets proposed by Angrist et al. (1999) and Blomquist

Table 2: Simulation Results for Comparing Alternative Estimators

$\rho = 0.5$			Bias($\hat{\beta}$)						MSE($\hat{\beta}$)					
n	k	R^2	LS	2SLS	FLIML	2SLS($\hat{\lambda}_F$)	2SLS($\lambda_{\sqrt{n}}$)	2SLS($\hat{\lambda}_{CV(1)}$)	LS	2SLS	FLIML	2SLS($\hat{\lambda}_F$)	2SLS($\lambda_{\sqrt{n}}$)	2SLS($\hat{\lambda}_{CV(1)}$)
100	1	0.01	0.4937	-0.7153	0.4293	-0.0825	-0.2416	-0.0340	0.2517	1746.8569	0.2226	0.1660	0.0659	0.3610
100	10	0.01	0.4959	0.4546	0.4446	0.3120	-0.0274	0.3847	0.2534	0.2950	0.3234	0.1596	0.1596	0.1992
500	1	0.01	0.4943	-0.1197	0.2523	-0.0464	-0.2279	-0.0743	0.2458	10.6155	0.1061	0.1408	0.0556	0.2297
500	10	0.01	0.4955	0.3259	0.2812	0.2570	-0.1177	0.2998	0.2470	0.1690	0.1759	0.1149	0.0198	0.1352
100	1	0.05	0.4744	-0.2109	0.2465	-0.0527	-0.2108	-0.0871	0.2325	20.2937	0.1058	0.1441	0.0574	0.2293
100	10	0.05	0.4733	0.3189	0.2676	0.2517	-0.0276	0.2915	0.2315	0.1671	0.1763	0.1130	0.0163	0.1324
500	1	0.05	0.4739	-0.0265	-0.0574	-0.0263	-0.1669	-0.0261	0.2261	0.0505	0.0296	0.0480	0.0352	0.0488
500	10	0.05	0.4751	0.1227	0.0585	0.1159	-0.0905	0.1225	0.2272	0.0412	0.0377	0.0387	0.0151	0.0409
100	1	0.3	0.3505	-0.0079	0.0412	-0.0081	-0.0911	-0.0079	0.1287	0.0269	0.0211	0.0268	0.0221	0.0269
100	10	0.3	0.3537	0.0880	0.0371	0.0852	-0.0142	0.0878	0.1309	0.0263	0.0249	0.0256	0.0114	0.0262
500	1	0.3	0.3502	-0.0032	0.0063	-0.0032	-0.0478	-0.0032	0.1238	0.0048	0.0046	0.0048	0.0059	0.0048
500	10	0.3	0.3501	0.0174	0.0059	0.0173	-0.0291	0.0174	0.1237	0.0049	0.0049	0.0049	0.0043	0.0049
$\rho = 0.9$			Bias($\hat{\beta}$)						MSE($\hat{\beta}$)					
n	k	R^2	LS	2SLS	FLIML	2SLS($\hat{\lambda}_F$)	2SLS($\lambda_{\sqrt{n}}$)	2SLS($\hat{\lambda}_{CV(1)}$)	LS	2SLS	FLIML	2SLS($\hat{\lambda}_F$)	2SLS($\lambda_{\sqrt{n}}$)	2SLS($\hat{\lambda}_{CV(1)}$)
100	1	0.01	0.8906	-1.6118	0.7734	0.0957	-0.0506	0.1040	0.7952	3727.1598	0.6208	0.1352	0.0081	0.4645
100	10	0.01	0.8921	0.8171	0.7935	0.6431	0.2235	0.7894	0.7979	0.6967	0.6755	0.4487	0.0621	0.6415
500	1	0.01	0.8911	-0.2852	0.4529	-0.0048	-0.0476	-0.1120	0.7945	48.9170	0.2254	0.0921	0.0049	0.2698
500	10	0.01	0.8912	0.5841	0.4740	0.5091	0.0934	0.5806	0.7946	0.3736	0.2611	0.2888	0.0137	0.3667
100	1	0.05	0.8548	-0.3537	0.4407	-0.0276	-0.0561	-0.1406	0.7329	85.3748	0.2172	0.0971	0.0123	0.2781
100	10	0.05	0.8557	0.5776	0.4480	0.5020	0.1902	0.5748	0.7344	0.3641	0.2392	0.2804	0.0476	0.3585
500	1	0.05	0.8547	-0.0393	0.1082	-0.0380	-0.0412	-0.0390	0.7310	0.0543	0.0302	0.0521	0.0073	0.0537
500	10	0.05	0.8550	0.2271	0.1063	0.2176	0.0652	0.2217	0.7315	0.0672	0.0309	0.0648	0.0095	0.0672
100	1	0.3	0.6301	-0.0252	0.0649	-0.0250	-0.0372	-0.0252	0.4001	0.0292	0.0202	0.0290	0.0137	0.0292
100	10	0.3	0.6293	0.1464	0.0550	0.1450	0.0825	0.1464	0.3989	0.0360	0.0209	0.0354	0.0157	0.0360
500	1	0.3	0.6300	-0.0041	0.0130	-0.0041	-0.0144	-0.0041	0.3976	0.0048	0.0045	0.0048	0.0036	0.0048
500	10	0.3	0.6300	0.0322	0.0120	0.0322	0.0161	0.0322	0.3975	0.0052	0.0045	0.0052	0.0033	0.0052

and Dahlberg (1999). None of alternatives, however, outperforms the proposed ridged 2SLS estimator based on $\hat{\lambda}_F$ and $\lambda = \sqrt{n}$ in the 24 settings under the investigation.[3] Finally, there is no improvement seen by implementing the ridged CF-2SLS estimation compared to the ridged 2SLS approach with the leave-one-out cross-validation.

5 Conclusions and Discussion

In this paper, we show that the ridged 2SLS estimators are preferable to the un-ridged 2SLS ones for smaller finite sample bias and variance with the presence of weak and/or many instruments. This provides a convenient alternative for practitioners when addressing the finite sample bias problem. However, further issues would be interesting to investigate. In particular, the asymptotic properties of the proposed estimators, including derivations of their asymptotic distributions, need to be studied for more accurate statistical inferences. Furthermore, it would be also interesting to study the asymptotic properties of an estimator of the ridging parameter, given that there are number of data-driven alternatives for the selection of λ.

[3] The results for these estimators can be obtained from the authors on request.

References

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14, 57–67.

Arel-Bundock, V. (2013): "A solution to the weak instrument bias in 2SLS estimation: Indirect inference with stochastic approximation," *Economics Letters*, 120, 495–498.

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

Blomquist, S. and M. Dahlberg (1999): "Small sample properties of LIML and jackknife IV estimators: experiments with weak instruments," *Journal of Applied Econometrics*, 14, 69–88.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American Statistical Association*, 90, 443–450.

Bun, M. J. and F. Windmeijer (2011): "A comparison of bias approximations for the two-stage least squares (2SLS) estimator," *Economics Letters*, 113, 76–79.

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

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

Cruz, L. M. and M. J. Moreira (2005): "On the validity of econometric techniques with weak instruments: inference on returns to education using compulsory school attendance laws," *Journal of Human Resources*, XL, 393–410.

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

Fuller, W. A. (1977): "Some properties of a modification of the limited information estimator," *Econometrica*, 45, 939–954.

Hahn, J. and J. Hausman (2002): "Notes on bias in estimators for simultaneous equation models," *Economics Letters*, 75, 237–241.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with weak instruments:

accuracy of higher order bias and MSE approximations," *The Econometrics Journal*, 7, 272–306.

HAHN, J. AND G. KUERSTEINER (2002): "Discontinuities of weak instrument limiting distributions," *Economics Letters*, 75, 325–331.

HOERL, A., R. KENNARD, AND K. BALDWIN (1975): "Ridge regression: some simulations." *Communications in Statistics, Theory and Methods*, 4, 105–123.

LINDLEY, D. AND A. SMITH (1972): "Bayes estimate for the linear model (with discussion) part 1," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 1–41.

NELSON, C. R. AND R. STARTZ (1990a): "The distribution of the instrumental variables estimator and its T-ratio when the instrument is a poor one," *Journal of Business and Economic Statistics*, 63, 125–140.

——— (1990b): "Some further results on the exact small sample properties of the instrumental variable estimator," *Econometrica*, 54, 967–976.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557–586.

STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics*, 20, 518–529.

WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2nd. ed.