Shrinkage methods for instrumental variable estimation*

Ryo Okui†

Department of Economics, University of Pennsylvania

This version: August, 2004

Preliminary and incomplete

Abstract

This paper proposes shrinkage methods for instrumental variable estimation to solve the "many instruments" problem. Even though using a large number of instruments reduces the asymptotic variances of the estimators, it has been observed both in theoretical works and in practice that in finite samples the estimators may behave very poorly if the number of instruments is large. This problem can be addressed by shrinking the influence of a subset of instrumental variables. That is, we reconstruct the estimating equation of an instrumental variable estimator, which is a weighted sum of sample moment conditions, by shrinking some elements of the weighting vector. This procedure can also be understood by a two-step process of shrinking some of the OLS coefficient estimates from the regression of the endogenous variables on the instruments then using the predicted values of the endogenous variables based on the shrunk coefficient estimates as the instruments. The shrinkage parameter is chosen to minimize the asymptotic MSE. We find that the optimal shrinkage parameter has a closed form which leads to easy implementation. The Monte Carlo result shows that the shrinkage methods work well and moreover perform better than the instrument selection procedure in Donald and Newey (2001) in several situations relevant to applications.

Keywords: TSLS, LIML, Shrinkage estimator, Instrumental variables.

*I thank my advisor, Yuichi Kitamura for his help, patience and encouragement. I am grateful to Gregory Kordas and Petra Todd for their helpful supervision. I also obtained valuable comments from Dylan Small, Yoshihiko Nishiyama, and seminar participants of the Fareastern Meeting of the Econometric Society at Yonsei Univerisity, and Kyoto University. I am also indebted to Shalini Roy and Jason Harburger. Of course, all errors are mine.

†442 McNeil building, 3718 Locust Walk, Philadelphia PA, U.S.A. 19104, tel:+1-215-898-7032, e-mail: okui@sas.upenn.edu

1 Introduction

This paper proposes a new solution to the problem of instrumental variable (IV, hereafter) estimation in the presence of many instruments. In this situation, we can estimate the model and perform some inference using a minimal subset of instruments. However, with a small number of instruments, we lose an efficiency which results in relatively large standard errors. We might try to increase the number of instruments in order to reduce the standard error of the estimate. It turns out that this approach may be misleading in finite samples. An IV estimator with many instruments may behave very poorly and can be sensitive to the number of instruments. In particular the two-stage least square (TSLS, hereafter) estimator generates a bias whose order is proportional to the number of the instruments (e.g., see Kunitomo (1980), Morimune (1983) or Bekker (1994)).

An example where this problem occurs in empirical work is the paper of Angrist and Krueger (1991). They use quarter of birth as an instrument to estimate returns to education. Even though quarter of birth seems to be a valid instrument, the authors obtain a relatively large standard error when they estimated the parameter with only the quarter of birth variables as instruments. By increasing the number of instruments we find a reduction in the standard error and the estimates becomes close to the OLS estimate. This is a typical phenomenon of having a large number of instruments and Bound, Jaeger and Baker (1996) illustrates how the problem of many instruments arose in Angrist and Krueger (1991)[1].

Existing solutions for the "many instruments" problem usually involve selection of instruments. Donald and Newey (2001) proposes minimizing the asymptotic mean squared error as a criterion to choose the number of instruments. Small (2002) also proposes a criterion function motivated by the Akaike Information Criteria to choose instruments.

This paper introduces a new procedure for IV estimation based on shrinkage methods. That is, we reconstruct the estimating equation of an instrumental variable estimator, which

[1]Even though Bound, Jaeger and Baker (1996) emphasizes the weak instrument problem, Table 1 in their paper indicates that in the data of Angrist and Krueger (1991), we do not suffer from the bias of the TSLS estimator if we use the minimal subset of instruments. Actually there are two problems arisen: One is the "many instruments" problem and another problem is that the additional instruments are weak. This paper focuses on the "many instruments" problem. Chao and Swanson (2003) and Stock and Yogo (2003) discuss consequences of a large number of weak instruments.

is a weighted sum of sample moment conditions, by shrinking some elements of the weighting vector. This idea can also be interpreted as shrinking a part of the OLS coefficient estimates from the regression of the endogenous variables on the instruments, and then using the predicted values of the endogenous variables, based on the shrunk coefficient estimates, as the instruments.

In the statistical literature, it has been observed that shrinkage methods perform very well and moreover they usually work better than selection methods (e.g., see Hastie, Tibshirani and Friedman (2001) section 3.4.5). The key decision involved in selection methods is to select which instruments to discard. Even though we ease the many instruments problem by doing so, we also ignore the information the discarded instruments might bring. On the other hand, shrinkage methods not only ease the many instruments problem but also we can utilize the information which is lost in discarding variables. Shrinkage procedures can become excellent alternatives to selection methods in IV estimation.

Another advantage of the shrinkage method proposed here is that we don't need to know the order of instruments. To implement selection methods such as Donald and Newey (2001), if we don't know the order, one must try $2^K - 1$ combinations of instruments, where K is the number of instruments, and it is practically impossible. While we need to try only $K-1$ combinations with knowledge of the order. The order may not be so clear in practice and the shrinkage methods can avoid this problem.

Even though there is hardly any literature which explicitly considers an application of shrinkage methods in IV estimations, there is one important paper by Chamberlain and Imbens (2004) whose procedure, the random effect quasi-maximum likelihood (REQML), could be categorized as a shrinkage method. They impose a random effect structure on the coefficients in the regression of the endogenous variable on instruments. However their approach depends on a specific structure: a single endogenous variable and a linear simultaneous equation model. The procedure presented here can be extended to more general models. The kernel weighted GMM in ARMA models by Kuersteiner (2000, 2001) is also related to the ideas explored here[2].

[2]The difference of the kernel weighted GMM and the shrinkage methods would be clear, however we note that we can find a pair of kernel function and bandwidth under which the kernel weighted GMM and the shrinkage TSLS are equivalent. They are $K(u) = 1$, for $|u| < c$ and $K(u) = s$ for $|u| \geq c$ where s is the shrinkage parameter and c is equal to the ratio of the number of main instruments and the total number of

One nontrivial question is how to choose the shrinkage parameter. We propose to choose the shrinkage parameter by minimizing the Nagar (1959)-type approximation of the mean squared error. We find that the optimal shrinkage parameter has a closed form which leads to easy implementation. One may consider to choose the shrinkage parameter in a similar way to the James-Stein estimator. However the James-Stein shrinkage rule is not optimal and in shrinkage TSLS estimation, there is a crucial difference between these two; The optimal shrinkage parameter has an order K^2 term while the James-Stein shrinkage rule has just an order K term where K is the number of instruments. The James-Stein shrinkage rule shrinks less than desired when the number of instruments is large. This shows the importance of the asymptotic MSE calculation in choosing the shrinkage parameter.

This article is organized as follows. The next section introduces the shrinkage TSLS estimator, explains the motivation and presents the theoretical results. Section 3 proposes the shrinkage limited information maximum likelihood estimator. The Monte Carlo experiments are included in Section 4. Discussion and possible extensions are in Section 5.

We use the following notations throughout the paper. For a sequence of vectors $\{A_i\}$, define A as $A = (A_1', A_2', \ldots, A_n')'$. For a matrix A, $||A|| = \sqrt{tr(A'A)}$, the usual Euclidean norm, and $P_A = A(A'A)^{-1}A'$. "wpa1" stands for "with probability approaching one".

2 The shrinkage TSLS estimator.

2.1 Model and Procedure

Following Donald and Newey (2001), I consider the model:

$$\begin{aligned} y_i &= Y_i'\gamma + x_{1i}'\beta + \epsilon_i = W_i'\delta + \epsilon_i \\ W_i &= \begin{pmatrix} Y_i \\ x_{1i} \end{pmatrix} = f(x_i) + u_i = \begin{pmatrix} E(Y_i|x_i) \\ x_{1i} \end{pmatrix} + \begin{pmatrix} \eta_i \\ 0 \end{pmatrix}, \quad i = 1, \ldots, N \end{aligned}$$

where y_i is a scalar outcome variable, Y_i is a $d_1 \times 1$ vector of endogenous variable, $x_i = (X_i', \chi_i')$ is a vector of exogenous variables and x_{1i} is a part of X_i. The set of instruments has the

instruments and the bandwidth is equal to the total number of instruments. This kernel function is not a standard one at all and the choice of the bandwidth and that of the shrinkage parameter are not equivalent. We should regard the kernel weighted GMM as a way to exploit all information from the order of instruments which is clear in ARMA models while this paper implicitly considers situations where the order is not clear.

following form; $(X_i', \psi_1(x_i), \cdots, \psi_K(x_i)) \equiv (X_i', Z_i')$ where ψ_ks are some functions of x_i. X_i is a $m \times 1$ vector of main IVs and Z_i is a $K \times 1$ vector of IVs. ϵ_i and u_i are unobserved random variables. f is an unknown function of x and f would be the best instrument. I employ this semiparametric structure because it allows us to easily analyze the model with many instruments. Another reason is that this paper intends to compare instrument selection methods and shrinkage methods, and to this end, it would be better to have the same structure as used in Donald and Newey (2001) to present a selection method which will be compared with shrinkage procedures in the Monte Carlo section.

We consider the situation similar to Chamberlain and Imbens (2004) where we have two sets of instruments, X and Z. Among the IVs, we typically have "main" instruments and we denote these instruments by X. We consider to shrink the effect of Z on the estimation of δ. The meaning of "main" can differ among situations. For example, suppose that we have a conditional moment restriction model and use a polynomial series as instruments. The main instruments in this case would be the first l-th polynomial series where l is the number of the endogenous variables. Another example could be the case where a number of instruments are generated by multiplying main instruments by regional dummies or time dummies. For instance, Angrist and Krueger (1991) shows that quarter of birth can be an instrument to estimate the return of education and uses quarter-of-birth times year-of-birth or state-of-birth interactions in their TSLS estimation. In this case, the quarter of birth variables are considered as "main" instruments.

Note that we are able to estimate δ with only using those main instruments if the number of the main instruments is larger than the number of the endogenous variables. However such an estimate may have a large standard error. Even though using more instruments is a way to reduce the standard error of the estimate, it is commonly observed that IV estimators with many instruments behave poorly (e.g., Morimune (1983) and Bound, Jaeger and Baker (1996)). The shrinkage TSLS (or LIML) estimator is introduced to address this "many instruments" problem. In this section the shrinkage TSLS estimator is discussed. The shrinkage LIML estimator is discussed in the next section.

Now we describe the procedure. First we assume $X'Z = 0$ without loss of generality. This can be achieved by regressing Z on X and taking the residuals. It is important to note that Z in our discussion may not be the matrix of the instruments itself but the orthogonalized

one in applications. Under $X'Z = 0$, the TSLS estimator of δ is the solution to

$$W'P_X(y - W\delta) + W'P_Z(y - W\delta) = 0.$$

The shrinkage TSLS estimator $\hat{\delta}_s$ is defined as the solution to

$$W'P_X(y - W\delta) + sW'P_Z(y - W\delta) = 0$$

and it is:

$$\hat{\delta}_s = (W'P^sW)^{-1}W'P^sy$$

for some shrinkage parameter s where $P^s = P_X + sP_Z$. By introducing the shrinkage parameter, s, we can reduce the effect of adding Z into the set of instruments. s lies between 0 and 1; $s = 0$ gives the TSLS estimator using only X and $s = 1$ gives the TSLS estimator using all of instruments. A more detailed discussion will be found in the next subsection.

To operationalize this procedure, a method for choosing s is needed. We recommend the following choice of s based on a higher order asymptotic result discussed in Section 2.3.

$$\hat{s}^* = 1 - \left(1 + \frac{\hat{\sigma}_\epsilon^2}{\hat{\lambda}'\hat{H}^{-1}\hat{\sigma}_{u\epsilon}\hat{\sigma}'_{u\epsilon}\hat{H}^{-1}\hat{\lambda}} \frac{\hat{\lambda}'\hat{H}^{-1}W'P_ZW\hat{H}^{-1}\hat{\lambda}}{K^2}\right), \tag{1}$$

where $\hat{\lambda}$ is the (possibly estimated) weighing vector chosen by the researcher, $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_{u\epsilon}$ are the estimates of $\sigma_\epsilon^2 = E(\epsilon_i^2)$ and $\sigma_{u\epsilon} = E(u_i\epsilon_i)$ based on the residuals from a preliminary estimation and $\hat{H} = W'(P_X + P_Z)W/N$ which is an estimate of the first order asymptotic variance.

2.2 Motivation and Discussion

It would be helpful to review shrinkage estimation in simple situations in order to acquire some intuition on how they work and to motivate us to use the shrinkage methods to solve the "many instruments" problem. First we look at the famous James-Stein estimator. Consider the situation where we have the model: $X \sim N(\theta, I_K)$. X is a K-dimensional random vector and θ is the K-dimensional estimated parameter. Suppose our sample is just X (that is, sample size is 1). A natural estimator of θ is $\hat{\theta} = X$ which is the maximum likelihood estimator. This estimator has several nice properties: It is unbiased ($E(\hat{\theta}) = \theta$) and it

achieves the Fisher's lower bound ($Var(\hat{\theta}) = I_K$ which is the inverse of sample size (1) times the Fisher information matrix (I_K)). However if $K \geq 3$, one can do better using a shrinkage estimator in terms of mean squared error. The mean squared error of $\hat{\theta}$ is $E((\hat{\theta} - \theta)'(\hat{\theta} - \theta)) = K$. The James-Stein estimator $\hat{\delta}_{JS}$ is defined as

$$\hat{\theta}_{JS} = \left(1 - \frac{K-2}{X'X}\right) X.$$

The mean squared error of the James-Stein estimator is:

$$\begin{aligned} E((\hat{\theta}_{JS}-\theta)'(\hat{\theta}_{JS}-\theta)) &= E((X-\theta)'(X-\theta)) - 2E\left((X-\theta)'\frac{K-2}{X'X}X\right) + E\left(\frac{(K-2)^2}{(X'X)^2}X'X\right) \\ &= K - 2E\left(\frac{(K-2)^2}{X'X}\right) + E\left(\frac{(K-2)^2}{X'X}\right) = K - E\left(\frac{(K-2)^2}{X'X}\right) < K. \end{aligned}$$

The equality $E\left((X-\theta)'[(K-2)/(X'X)]X\right) = E\left((K-2)^2/(X'X)\right)$ comes from integration by parts. (See Lehmann (1983) section 4.6 for more details.) This is a remarkable result: $\hat{\theta}_{JS}$ has a strictly smaller mean squared error than the MLE $\hat{\theta}$. Note that if we are interested in estimating a just one value θ_i, then we cannot do better than $\hat{\theta}_i = X_i$. The improved performance of the James-Stein estimator occurs because our interest is not concentrated on a particular single element but the overall performance of the estimation and in that sense each parameter θ_i becomes "nuisance".

We can extend this idea into standard regression models. Suppose that we have the linear model $y = X\pi + u$ where y is an $N \times 1$ random vector, X is a $N \times K$ matrix of regressors, π is a $K \times 1$ vector and $u \sim N(0, I_N)$ which is uncorrelated with X. Let $\hat{\pi}$ is the OLS estimator of π. Then $\hat{\pi} \sim N(\pi, \sigma^2(X'X)^{-1})$. Rewriting this, we have $\sigma^{-1}(X'X)^{1/2}\hat{\pi} \sim N(\sigma^{-1}(X'X)^{1/2}\pi, I_K)$. A shrinkage estimator of $\sigma^{-1}(X'X)^{1/2}\pi$ is $\left(1 - (K-2)/(\sigma^{-2}\hat{\pi}'(X'X)\hat{\pi})\right)\sigma^{-1}(X'X)^{\frac{1}{2}}\hat{\pi}$. A shrinkage estimator of the regression coefficient is therefore

$$\hat{\pi}_s = \left(1 - \frac{\sigma^2(K-2)}{\hat{\pi}'(X'X)\hat{\pi}}\right)\hat{\pi}.$$

It can be shown that this estimator has a smaller mean squared error than the OLS estimator under conditions (e.g. see Takada (1979)). Note that the conventional first order asymptotic analysis doesn't tell us how the shrinkage estimator performs better than OLS since $\sqrt{n}\sigma^2(K-2)(\hat{\pi}'(X'X)\hat{\pi})^{-1} \to_p 0$ if $\pi \neq 0$ and $X'X/N$ converges to some positive definite

matrix, and consequently the shrinkage estimator has the same asymptotic distribution as the OLS estimator. The effect of shrinkage arises in higher order terms. This observation motivates us to investigate a higher order asymptotics.

Now we can proceed to the situation of interest. Consider a simple linear model with only one endogenous variable and no exogenous regressors:

$$y = Y\delta + u,$$

and a set of instruments $x = (X, Z)$ where y, Y and u are $N \times 1$ vectors, δ is a scalar parameter to be estimated and X, Z are $N \times m$ and $N \times K$ matrix. As before we assume $X'Z = 0$ without loss of generality. The TSLS estimator is the solution to:

$$\hat{\pi}_X' X'(y - \delta Y) + \hat{\pi}_Z' Z'(y - \delta Y) = 0$$

which is a weighted sum of sample moment conditions where the weighting vector, $(\hat{\pi}_X, \hat{\pi}_Z)$, is the OLS coefficient estimate from the regression of Y on (X, Z). The "many instruments" problem occurs because we need to estimate a large dimension of the weighting vector in the presence of a large number of instruments. The observation given above would indicate that the problem could be addressed by some shrinkage techniques. Let $s \in [0, 1]$ be some shrinkage parameter. We consider to use $(\hat{\pi}_X, s\hat{\pi}_Z)$ instead of $(\hat{\pi}_X, \hat{\pi}_Z)$ as the weighting vector and obtain the shrinkage TSLS estimator which is the solution to: $\hat{\pi}_X' X'(y - \delta Y) + s\hat{\pi}_Z' Z'(y - \delta Y) = 0$.

We are also able to provide another useful insight of the shrinkage TSLS in this simple case. The TSLS estimator is obtained by the following procedure. First we regress Y on x to obtain $\hat{\pi}_x = (x'x)^{-1}x'Y$. Then we estimate δ by $\hat{\delta} = (\hat{Y}'Y)^{-1}\hat{Y}'y$ where $\hat{Y} = x\hat{\pi}_x$. We attempt to shrink $\hat{\pi}_Z$ in order to obtain a better estimate. The shrinkage TSLS is written as:

$$\hat{\delta}_s = (\tilde{Y}'Y)^{-1}\tilde{Y}'y$$

where $\tilde{Y} = X\hat{\pi}_X + sZ\hat{\pi}_Z$. Therefore the shrinkage TSLS can be understood by a two-step process of shrinking some of the OLS coefficient estimates from the regression of the endogenous variables on the instruments then using the predicted values of the endogenous variables based on the shrunk coefficient estimates as the instruments.

Note that if we shrink all elements of $\hat{\pi}$ then there is no difference from the TSLS since $x(s\hat{\pi}) = s\hat{Y}$ and $(s\hat{Y}'Y)^{-1}s\hat{Y}'y = \hat{\delta}$. What matters in the IV estimation is just a ratio of importance of instruments while the magnitude of π_i is not important. On the other hand, the relative weight π_i/π_j matters. Shrinking all does not change the relative scale of π_is. This is a difference from usual OLS regression where the magnitude of π is important.

Choosing the shrinkage parameter is an important and nontrivial decision. One choice can be:

$$s_{JS} \equiv \left(1 - \frac{\hat{\sigma}^2(K-2)}{\hat{\pi}_Z' Z'Z \hat{\pi}_Z}\right),$$

motivated by the James-Stein estimator. However, this choice is naive. IV estimation is different from the standard regression problems. This deviation from standard regression models implies that the best first stage estimation doesn't necessarily imply the best method to estimate δ. The optimal choice is given in equation (1) which is derived to minimize the asymptotic MSE and it is different from the James-Stein shrinkage rule. The Monte Carlo experiment presented later shows the importance of the asymptotic MSE calculation in choosing the shrinkage parameter.

2.3 Theoretical results

We will show the asymptotic properties of the shrinkage TSLS under the following assumptions. These assumptions are similar to those imposed in Donald and Newey (2001).

Assumption 1. $\{y_i, W_i, x_i\}$ *are i.i.d,* $E(\epsilon_i^2|x_i) = \sigma_\epsilon^2 > 0$ *and* $E(||\eta_i||^4|x_i)$ *and* $E(|\epsilon_i|^4|x_i)$ *are bounded.*

Assumption 2. *(i)* $\bar{H} \equiv E(f_i f_i')$ *exists and is nonsingular. (ii) there exists* π_K *such that* $E(||f(x) - \pi_K(X', Z')||) \to 0$ *as* $K \to \infty$

Assumption 3. *(i)* $E((\epsilon, u')'(\epsilon, u')|x_i)$ *is constant: (ii)* $X'X$ *and* $Z'Z$ *are nonsingular wpa1. (iii)* $X'Z = 0$ *wpa1. (iv)* $\max_{i \le N} P_{X,ii} \to_p 0$ *and* $\max_{i \le N} P_{Z,ii} \to_p 0$. *(v)* f_i *is bounded.*

Assumption 3(iii) is satisfied for $(X, \bar{Z})$ where $\bar{Z} = Z - X(X'X)X'Z$. Therefore this is not a real restriction but it is important to keep it in mind that Z in our discussion may not be the matrix of the instruments itself but the orthogonalized one in applications.

The first theorem is on the consistency and the asymptotic normality of the shrinkage TSLS estimator.

Theorem 1. *Suppose Assumption 1-3 are satisfied. If* $(sK)^2/N \to_p 0$ *and either* $s \to_p 1$ *or* $f'P_Z f/N \to_p 0$*, then* $\hat{\delta}_s - \delta \to_p 0$ *and* $\sqrt{N}(\hat{\delta}_s - \delta) \to_d N(0, \sigma_\epsilon^2 \bar{H}^{-1})$.

This justifies the use of the shrinkage TSLS estimator. Unfortunately, this result also indicates that the conventional first order asymptotic analysis is not enough to investigate the effect of shrinkage and we cannot have some guidance to choose the shrinkage parameter s. This is similar to the case of selecting the number of instruments. The first order asymptotic results are not useful to see how many instruments should be used and we have to look at a higher order expansion. Given this observation, we look at the higher order asymptotic expansion and propose to choose the shrinkage parameter to minimize the asymptotic mean squared error. The notion of asymptotic MSE employed here is similar to the Nagar-type asymptotic expansion (Nagar 1959). Following Donald and Newey (2001), we approximate the MSE, $E((\hat{\delta} - \delta_0)(\hat{\delta} - \delta_0)')$, by $\sigma_\epsilon^2 H^{-1} + S(s)$ where

$$N(\hat{\delta} - \delta_0)(\hat{\delta} - \delta_0)' = \hat{Q}(s) + \hat{r}(s), \quad E(\hat{Q}(s)|x) = \sigma_\epsilon^2 H^{-1} + S(s) + T(s),$$

$H = f'f/N$ and $(\hat{r}(s) + T(s))/tr(S(s)) = o_p(1)$ as $K \to \infty, N \to \infty$. First we divide the $N(\hat{\delta} - \delta_0)(\hat{\delta} - \delta_0)'$ into two parts, $\hat{Q}(s)$ and $\hat{r}(s)$, and discard $\hat{r}(s)$ which goes to zero faster than $S(s)$. Then we take the expectation of $\hat{Q}(s)$ conditional on the exogenous variable, x, and ignore the term $T(s)$ which goes to zero faster than $S(s)$. $\sigma_\epsilon^2 H^{-1}$ corresponds to the first order asymptotic variance. Hence $S(s)$ is the nontrivial and dominant term in the MSE and our goal is to find $S(s)$.

This Nagar-type approximation is popular in IV estimation literature but not common in shrinkage literature which mainly focuses on exact finite sample properties. We have several reasons to investigate the Nagar asymptotic MSE even though usual shrinkage literature do not use that. First this approach makes comparison with Donald and Newey (2001) easier since they also use the Nagar expansion. Second, a finite sample parametric approach may not be so convincing as it relies on a distributional assumption. Lastly, the exact finite sample approach usually gives us too complicated results to be meaningful. The application of the Nagar approximation provides a clear result and that leads to an optimal shrinkage parameter selection procedure that can be implemented easily in practice.

The next theorem shows the form of the MSE under $K \to \infty$, $N \to \infty$ and that the shrinkage parameter is exogenous.

Theorem 2. *Suppose that Assumption 1-3 are satisfied. Under* $(sK)^2/N \to 0$ *and either* $(1-s) = O_p(K^2/N)$ *or* $E(f_i Z_i') = 0$,

$$S(s) = H^{-1}\left[\sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(sK)^2}{N} + \sigma_\epsilon^2\frac{f'(I-P^s)(I-P^s)f}{N}\right]H^{-1}.$$

Appendix contains the proof. Note that the formula in Donald and Newey (2001) is given by setting $s = 1$ as $s = 1$ corresponds to the standard TSLS estimator.

Given this formula, our task is to find an s which minimizes $S(s)$. $S(s)$ is a matrix so we need to introduce a weighting vector; λ (λ may be estimated) and solve a minimization problem with objective function $\lambda' S(s)\lambda$. The optimal shrinkage parameter is:

$$s^* = 1 - \left(1 + \frac{\sigma_\epsilon^2}{\lambda' H^{-1}\sigma_{u\epsilon}\sigma_{u\epsilon}' H^{-1}\lambda}\frac{\lambda' H^{-1} f' P_Z f H^{-1}\lambda}{K^2}\right)^{-1}.$$

This form is very intuitive: the optimal shrinkage parameter is a increasing function of a measure of the strength of the instruments, $f'P_Z f/N$, and a decreasing function of the number of instruments, K. Surprisingly, this suggests that we should shrink always since $s = 1$ occurs only if $f'P_Z f = \infty$ however $f'P_Z f$ goes to infinity in at most a smaller order than that of the mean squared error. $0 \leq s < 1$ also means that the shrinkage TSLS does better than TSLS with the same number of instruments.

The standard case is $f'P_Z f/N \to_p c > 0$ and $s \to_p 1$. This means that if Z is a valid instrument, then asymptotically we don't shrink and achieve the semiparametric efficiency. On the other hand, if $f'P_Z f/K^2 \to_p 0$ which occurs when Z is an irrelevant instrument, $s \to_p 0$. Introducing the shrinkage parameter we can defend against the use of completely weak instruments. The weak instruments case in Staiger and Stock (1997) sense occurs when $f'P_Z f/K^2 \to_p c > 0$. Then $s \to_p \bar{s}$ where $0 < \bar{s} < 1$. Even though we do not consider this case formally, we can conjecture that the shrinkage TSLS can even utilize the information from the weak instruments.

The main difference of this shrinkage rule from the James-Stein shrinkage rule is that we have an order K^2 term but the James-Stein shrinkage rule has just an order K term. This might imply that the James-Stein shrinkage rule shrink less than desired when the number of instruments is large.

If we only have one endogenous variable or in other words Y_i is a scalar, we do not need to care about the choice of λ and the optimal shrinkage parameter is given by:

$$s^* = 1 - \left(1 + \frac{\sigma_\epsilon^2}{\sigma_{u\epsilon}^2}\frac{\bar{Y}' P_Z \bar{Y}}{K^2}\right)^{-1}.$$

where $\bar{Y} = (E(Y_1|x_1), \ldots, E(Y_N|x_N))'$.

The optimal shrinkage parameter depends on the unknown parameters which has to be estimated to implement the procedure. A natural estimator of the optimal shrinkage parameter is given by equation (1) and the following theorem justifies the use of it.

Theorem 3. *Assumption 1-3 are satisfied and* $\hat{\sigma}_\epsilon^2 \to_p \sigma_\epsilon^2$, $\hat{\sigma}_{u\epsilon} \to_p \sigma_{u\epsilon}$ *and* $\hat{\lambda} - \lambda \to_p 0$. *Then* $(S(\hat{s}^*) - S(s^*))/S(s^*) = o_p(1)$

The estimation error of s^* can be negligible so that the estimated shrinkage parameter attains the minimum of the MSE asymptotically.

Note that in principle we can choose s and K simultaneously to minimize the asymptotic MSE. However this paper does not consider such a procedure here even though it may be worth a further investigation. The main purpose of this paper is to see the performance of instrumental variables estimators with shrinkage and a shrinkage-selection hybrid method lies beyond the focus of this paper.

3 The shrinkage LIML estimator

We can extend our idea of the shrinkage TSLS into the limited information maximum likelihood estimator (LIML). The LIML estimator is the minimizer of $(y - W\delta)'(P_X + P_Z)(y - W\delta)/((y - W\delta)'(y - W\delta))$. The shrinkage LIML estimator $\hat{\delta}$ is defined as:

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}} \frac{(y - W\delta)' P^s (y - W\delta)}{(y - W\delta)'(y - W\delta)}.$$

Let $v_i = u_i - \epsilon_i \sigma_{u\epsilon}/\sigma_\epsilon^2$ and define $\Sigma_v = E(v_i v_i')$. The next theorem derives the asymptotic MSE of the shrinkage LIML estimator. The third moment condition $E(\epsilon_i^2 v_i) = 0$ is assumed to simplify the formula.

Theorem 4. *Assumptions 1-3 are satisfied,* $\Sigma_v \neq 0$, $E(||\eta_i||^5|x_i)$ *and* $E(|\epsilon|^5|x_i)$ *are bounded and* $E(\epsilon_i^2 v_i) = 0$. *Then under* $sK/N \to_p 0$ *and* $1 - s = O_p(sK/N)$ *or* $E(f_i Z_i') = 0$, *we have* $\hat{\delta} \to_p \delta$, $\sqrt{N}(\hat{\delta} - \delta) \to_d N(0, \sigma_\epsilon^2 \bar{H}^{-1})$ *and*

$$S(s) = H^{-1}\left[\sigma_\epsilon^2 \Sigma_v \frac{s^2 K}{N} + \sigma_\epsilon^2 \frac{f'(I - P^s)(I - P^s)f}{N}\right] H^{-1}.$$

The shrinkage parameter is chosen to minimize $\lambda' S(s)\lambda$ with respect to s and the optimal shrinkage parameter is

$$s^* = 1 - \left(1 + \frac{1}{\lambda' H^{-1} \Sigma_v H^{-1} \lambda} \frac{\lambda' H^{-1} f' P_Z f H^{-1} \lambda}{K}\right)^{-1}.$$

If there is only one endogenous variable, the minimizer doesn't depend on λ and that is

$$s^* = 1 - \left(1 + \frac{\sigma_\epsilon^2}{\sigma_\eta^2 \sigma_\epsilon^2 - \sigma_{\eta\epsilon}^2} \frac{\bar{Y}' P_Z \bar{Y}}{K}\right)^{-1},$$

where $\sigma_\eta^2 = E(\eta_i^2)$ and $\sigma_{\eta\epsilon} = E(\eta_i \epsilon_i)$

The optimal shrinkage parameter has the order K term while that of the shrinkage TSLS has the order K^2 term. This means that we should shrink less in the shrinkage LIML than in the shrinkage TSLS when K is large. This observation is consistent with the established result that the LIML estimator is more robust against the number of instruments than the TSLS estimator. Still we have $0 \leq s^* < 1$ always and we can do always better by shrinking even in the LIML.

4 Monte Carlo Simulation

This section reports the result of the Monte Carlo experiments[3]. The aims of this experiments are to see how the shrinkage estimators behave in a moderate sample size and to compare the shrinkage methods with the other estimation methods. Comparison with the instrument selection procedure in Donald and Newey (2001) is one of the main purposes of this study. To make this comparison easier, we borrowed their experimental design.

[3]This Monte Carlo simulation was conducted with Ox 3.20 (Doornik 2002) for Linux and I really thank the provider.

4.1 Design

Our data generating process is the following model

$$\begin{aligned} y_i &= \delta Y_i + \epsilon_i \\ Y_i &= \pi' Z_i + u_i \end{aligned}$$

for $i = 1, \ldots, n$ where Y_i is a scalar, δ is a scalar parameter of interest, $Z_i \sim i.i.d.N(0, I_{\bar{K}})$ and

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \right).$$

$\bar{K}$ is the total number of instruments. Among Zs, the first instrument is the "main" instrument which corresponds to X in the theoretical part of this paper. That is, $m = 1$ and $K = \bar{K} - 1$ in the notation used before. We fix the true value of δ, $\delta = 0.1$ and we examine how well each estimator estimates δ.

In this framework, each experiment is indexed by the vector of specifications: $(n, \bar{K}, c, \{\pi\})$. n represents the sample size and we use $n = 100$ and $n = 500$. We set $\bar{K} = 20$ if $n = 100$ and $\bar{K} = 25$ if $n = 500$. The degree of endogeneity is summarized in c and we set $c = 0.1$, 0.5 and 0.9.

Hahn and Hausman (2002) observes that the theoretical R^2 of the first stage regression is given by $R_f^2 = \pi'\pi/(\pi'\pi + 1)$. We keep this R_f^2 fixed over the experiments. This means that even though we tried the three different specifications of π which are stated later, π always satisfies $\pi'\pi = R_f^2/(1 - R_f^2)$. We tried $R_f^2 = 0.1$ and 0.01.

The first specification of π is a case where the instruments are all equally important.

$$\text{Model (a):} \quad \pi_k = \sqrt{\frac{R_f^2}{\bar{K}(1 - R_f^2)}}, \forall k.$$

This case is difficult since not only they are all equally important, but also all of them are weak. Using only the first instrument is by no mean good. Using all instruments might cause a "many instruments" problem. Since there is no reason to prefer some to others, the situation is hard for selection methods. It is also problematic for shrinkage methods as the main instrument itself is weak and other instruments are as important as the main one.

The second model considered is

$$\text{Model (b):} \quad \pi_1 = c, \pi_k = \frac{c}{\sqrt{\bar{K}-1}} \forall k > 1$$

where c is chosen to satisfy $\pi'\pi = R_f^2/(1-R_f^2)$ is satisfied. This is a situation which motivates shrinkage methods. The first instrument is strong but others are poor instruments. This data generating process seems relevant for applications. Often we know that the instruments in our hand guarantee the identification of the parameter of interest. However the estimate using only those instruments shows a relatively large standard error and we cannot make a sharp conclusion with this estimates. In this case even if we are aware that other possible instruments are relatively weak, we may want to increase the number of instruments to obtain a smaller standard error. For example, see Angrist and Krueger (1991). Therefore, it is important to see how each method performs especially compared to the IV estimator using only one instrument.

Finally we considered the data generating process which Donald and Newey (2001) used:

$$\text{Model (c):} \quad \pi_k = c(\bar{K}) \left(1 - \frac{k}{\bar{K}+1}\right)^4.$$

The strength of instruments decreases in k moderately. An instrumental selection procedure such as Donald and Newey (2001) proposed would be suitable in this situation. We executed an experiment assuming that we know the order of strength of instruments. Note that this information is only relevant in this model. In model (a) and (b), the order (except that we know the first instrument in model(b)) is not important.

The number of replication is 1000.

We will compare the following estimators. The first one is the ordinally least square estimator (OLS), which is is inconsistent but can be an attractive choice if the degree of endogeny is small. The second is the TSLS estimator with all available instruments (TSLS). The third estimator examined is the IV estimator with only the first instrument (IV,1). This would be the one we try first in an empirical research and typically we find a relatively large standard error and think of using more instruments. The fourth is the TSLS estimator with Donald and Newey's (2001) optimal selection of the number of instruments (DNTSLS).

The next three estimators are the shrinkage TSLS estimators with the different choices of shrinkage parameters. The first choice is the James-Stein shrinkage rule (JSTSLS). The point

of including this procedure is to see the performance of the shrinkage TSLS estimator with a naive way to shrink. The next one is the true optimal shrinkage parameter (OSTSLS), which is infeasible in practice. The performance of OSTSLS can be seen as the upper bound of shrinkage procedures. The last choice of shrinkage parameter is the estimated (i.e., feasible) optimal shrinkage parameter (STSLS).

Other three estimators are the limited maximum likelihood estimator (LIML) with all instruments, the LIML estimator with Donald and Newey (2001)'s optimal selection of the number of instruments (DNLIML), and the (feasible) shrinkage LIML estimator (SLIML).

For the selection methods and the shrinkage methods, we need some caution to see the result. The result depends on how to estimate the selection criteria or the shrinkage parameters. Following Donald and Newey (2001), the preliminary estimate is obtained with the number of instruments chosen by the first stage cross-validation for both methods and the cross-validation criteria is used for $\hat{R}(K)$ (see Donald and Newey (2001)) in the selection criteria.

4.2 Result

The results of the experiments are summarized in Table 1-6. '$*$' indicates that the number is more than 1,000.

For each estimator, we computed the median bias (bias), the median absolute deviation (MAD), the difference between the 0.1 and 0.9 quantile (Dec. Rge.)[4] and the coverage rate (Cov. Rate) of a 95% confidence interval. To construct the confidence intervals to compute the coverage probabilities, we used the following estimate of asymptotic variance respecting Donald and Newey (2001). The estimators examined here have the common form: $\hat{\delta} = (\hat{X}'X)^{-1}\hat{X}'Y$ ($\hat{X} = X$ in OLS, $\hat{X} = Z(Z'Z)^{-1}Z'X$ in TSLS and so on). And the estimates of variance $\hat{V}$ is given by

$$\hat{V} = \frac{(y - X\hat{\delta})'(y - X\hat{\delta})}{N}(\hat{X}'X)^{-1}\hat{X}'\hat{X}(X'\hat{X})^{-1}.$$

"MAD" would represent the performance of each estimator most well and the discussions stated below are based on "MAD".

[4]We use these "robust" measures because of concerns on existence of moments of estimators.

First we summarize the performance of the established procedures: OLS, TSLS, IV,1 and LIML. If the endogeneity is small, then OLS performed well and sometimes it dominated all others. The relative performance between TSLS and LIML depends on the degree of endogeneity. TSLS did much better than LIML in the cases with $c = 0.1$ and LIML outperformed TSLS in the cases with $c = 0.9$. This is consistent with the MSE formula; the MSE of TSLS increases with the degree of endogeneity and that of LIML decreases. IV,1 performed well in model (b), $c = 0.9$ cases and was best in the model (b), $R_f^2 = 0.01$, $c = 0.9$ cases. In the model (c), $R_f^2 = 0.1$, $c = 0.9$ cases, the performance of TSLS was much worse than that of IV,1 and this illustrates the "many instruments" problem.

Now we compare the selection methods and the shrinkage methods. The first comparison is between DNTSLS and STSLS. Generally STSLS performed well in model (a) and (b) and DNTSLS did well in model (c), even though if the endogeneity is small, STSLS is better in model (c). A remarkable phenomenon is that there are cases where DNTSLS performed substantially worse than TSLS especially when the endogeneity is small. On the other hand, STSLS was usually better than TSLS. Even in the cases where TSLS was better, the difference is small. A similar phenomenon is observed when we compare DNLIML and SLIML. Contrary to that SLIML could achieve improvement on LIML generally, the relative performance of DNLIML to LIML is not stable; in some cases DNLIML did much better than LIML but also there are cases where DNLIML did much worse than LIML. The good performance of DNLIML usually occurred in the low endogeneity cases where TSLS type estimators or OLS performed well. In the high endogeneity cases which are suitable for LIML type estimators, SLIML was usually best in model (a) and (b). In model (c) with $c = 0.9$, DNLIML was usually best though the differences between DNLIML and SLIML are small.

Next, we compare JSTSLS and STSLS to see the effect of the choice of the shrinkage parameter. In most cases, STSLS dominates JSTSLS. Take the case of model (b), $R_f^2 = 0.1$, $c = 0.9$, $n = 100$ for example. In this case the difference between these two is quite remarkable and JSTSLS did much worse than all others but OLS, which is inconsistent, and TSLS. Though there are a few of cases where JSTSLS did better than STSLS, the differences are small in those cases. This result demonstrates the importance of investigating an appropriate way to shrink.

We conclude that the selection methods and the shrinkage methods are compliment and

neither of them is dominating. However we observed that the shrinkage methods could achieve an improvement on the baseline methods generally while there are cases where the selection methods might perform substantially worse than using all instruments. It would be beneficial if applied researchers utilize shrinkage methods as alternative approaches to IV selection. Also it would be an interesting future research to investigate how to hybridize these two approaches.

5 Discussion

The idea of shrinkage stated in this paper can be extended into general moment restriction models easily, though how to find an optimal way to shrink might be demanding. As linear instrumental variable estimation, the generalized method of moments estimator is a solution to the equation of a weighted sum of sample moment conditions and the shrinkage type estimators can be obtained by shrinking some elements of the weighting vector. The shrinkage parameter would be chosen to minimize the asymptotic mean squared error which might be difficult. On the other hand, the idea of REQML by Chamberlain and Imbens (2004), which is closely related to the shrinkage methods, is hard to extend into more general models since it is based on the likelihood function though REQML has several attractive features, such as being interpretable as a Bayes procedure.

Another useful extension is to handle multiple groups of instruments. Note that this article focuses on the situation where we have only 2 groups of instruments, main instruments and others. If we have more than 2 groups of instruments, we just need to shrink group by group. The optimal shrinkage parameter would be calculated with a similar way as the one presented here. The crucial assumption is that we know which group some particular instrument belongs to. We may also think of hybrid methods of adaptively partitioning instruments and shrinking group by group. For estimation of a multivariate normal mean, George (1986) provides an interesting discussion on a method to handle a situation with several candidates of partition. We consider hybrid methods as a promising direction to go.

A Proofs

This appendix contains the proofs of the theorems. Hereafter all expectations are conditional on x. We will employ Lemma A.1 in Donald and Newey (2001) to show Proposition 2 and 3. The

estimator examined has the form of $\sqrt{N}(\hat{\delta}-\delta)=\hat{H}^{-1}\hat{h}$. We define $h=f'\epsilon/\sqrt{N}$ and $H=f'f/N$.

Lemma 1 (Donald and Newey (2001) Lemma A.1). *If there is a decomposition* $\hat{h}=h+T^h+Z^h$, $\hat{H}=H+T^H+Z^H$,

$$(h+T^h)(h+T^h)'-hh'H^{-1}T^{H\prime}-T^HH^{-1}hh'=\hat{A}(s)+Z^A(s),$$

such that $T^h=o_p(1)$, $h=O_p(1)$, $H=O_p(1)$, *the determinant of* H *is bounded away from zero with probability 1,* $\rho_{K,N}=o_p(1)$,

$$||T^H||^2=o_p(\rho_{K,N}),\quad ||T^h||||T^H||=o_p(\rho_{K,N}),\quad ||Z^h||=o_p(\rho_{K,N}),\quad ||Z^H||=o_p(\rho_{K,N}),$$
$$Z^A(s)=o_p(\rho_{K,N}),\quad E(\hat{A}(s)|X)=\sigma^2H+HS(s)H+o_p(\rho_{K,N})$$

then

$$\begin{aligned}N(\hat{\delta}-\delta_0)(\hat{\delta}-\delta_0)' &= \hat{Q}(s)+\hat{r}(s)\\ E(\hat{Q}(s)|X) &= \sigma_\epsilon^2H^{-1}+S(s)+T(s)\\ (\hat{r}(s)+T(s))/tr(S(s)) &= o_p(1), K\to\infty, N\to\infty\end{aligned}$$

We state two technical lemmas and their proofs. Those lemmas will be used to prove the theorems. First recall that $Z'X=0$ and $P^s=P_X+sP_Z$ where $P_X=X(X'X)^{-1}X'$ and $P_Z=Z(Z'Z)^{-1}Z'$.

Lemma 2. *Suppose Assumption 1-3 are satisfied. Then we have*

1. $tr(P^s)=m+sK$
2. $\sum_i(P^s_{ii})^2=o_p(sK)$
3. $\sum_{i\neq j}P^s_{ii}P^s_{jj}=(m+sK)^2+o_p(sK)$
4. $\sum_{i\neq j}P^s_{ij}P^s_{ij}=(m+s^2K)+o_p(sK)$
5. $h=f'\epsilon/\sqrt{N}=O_p(1)$ *and* $H=f'f/N=O_p(1)$

Proof. First note that $(sK)^{-1}=O_p(1)$. For 1,

$$tr(P^s)=tr(P_X)+s\cdot tr(P_Z)=m+sK.$$

Assumption 3 and Lemma 2.1 implies

$$\sum_i(P^s_{ii})^2\leq\max_i(P^s_{ii})tr(P^s)=o_p(1)(m+sK)=o_p(sK).$$

This is the proof of 2. Also these results imply

$$\sum_{i\neq j}P^s_{ii}P^s_{jj}=\sum_iP^s_{ii}\sum_jP^s_{jj}-\sum_i(P^s_{ii})^2=(m+sK)^2+o_p(sK)$$

which is 3. To show 4, first we observe that

$$\sum_{i\neq j}P^s_{ij}P^s_{ij}=tr(P^{s\prime}P^s)-\sum_i(P^s_{ii})^2.$$

Now $P^{s\prime}P^s = (P_X + sP_Z)(P_X + sP_Z) = P_X + s^2P_Z$ and $tr(P^{s\prime}P^s) = m + s^2K$. Since we know $\sum_i (P^s_{ii})^2 = o_p(sK)$ from 2 in this Lemma,

$$\sum_{i\neq j} P^s_{ij}P^s_{ij} = m + s^2K + o_p(sK).$$

This is 4.

5 is Lemma A.2 (v) in Donald and Newey (2001). □

Let $e^s_f = f'(I - P^s)(I - P^s)f/N$ and $\Delta_s = tr(e^s_f)$

Lemma 3. *Suppose Assumption 1-3 are satisfied and $s \to_p 1$ or $f'P_Zf/N \to_p 0$. Then we have*

1. $\Delta_s = o_p(1)$,
2. $f'(I - P^s)\epsilon/\sqrt{N} = O(\Delta_s^{1/2})$,
3. $u'P^s\epsilon = O_p(sK)$,
4. $E(u'P^s\epsilon\epsilon'P^su) = \sigma_{u\epsilon}\sigma'_{u\epsilon}(m + sK)^2 + (\sigma^2_\epsilon\Sigma_u + \sigma_{u\epsilon}\sigma'_{u\epsilon})(m + s^2K) + o_p(sK)$,
5. $E(f'\epsilon\epsilon'P^su) = \sum_i f_iP^s_{ii}E(\epsilon^2_iu'_i) = O_p(sK)$,
6. $\Delta_s^{1/2}/\sqrt{N} = o_p(sK/N + \Delta_s)$,
7. $E(hh'H^{-1}u'f/N) = \sum_i f_if'_iH^{-1}E(\epsilon^2u_i)f'_i/(N^2) = O_p(1/N)$,
8. $E(f'(I - P^s)\epsilon\epsilon'P^su/N) = o_p(\Delta_s^{1/2}\sqrt{sK}/\sqrt{N})$.

Proof. Let start the proof from 1. Since $(I - P^s)(I - P^s) = I - P + (s-1)^2P_Z$ by a simple algebra,

$$\frac{f'(I - P^s)(I - P^s)f}{N} = \frac{f'(I - P)f}{N} + (s-1)^2\frac{f'P_Zf}{N}.$$

The first term is $o_p(1)$ by Lemma A.3(i) in Donald and Newey (2001) and the second term converges to 0 if $s \to_p 1$ or $f'P_Zf/N \to_p 0$. Therefore $\Delta_s = o_p(1)$.

Next we observe that $E(f'(I - P^s)\epsilon/\sqrt{N}) = 0$ and

$$E\left(\frac{f'(I - P^s)\epsilon}{\sqrt{N}}\frac{\epsilon'(I - P^s)f}{\sqrt{N}}\right) = \sigma^2_\epsilon\frac{f'(I - P^s)(I - P^s)f}{N} = \sigma^2_\epsilon e^s_f$$

Therefore $f'(I - P^s)\epsilon/\sqrt{N} = O_p(\Delta_s^{1/2})$ by the Chebyshev inequality. This is 2.

For 3, Cauchy-Schwartz inequality says that all element of $u'P^s\epsilon$ is less than $[tr(u'P^su)(\epsilon'P^s\epsilon)]^{1/2}$, then since $E(u'P^su) = \sigma^2_u(m+sK) = O_p(sK)$ and similarly $E(\epsilon'P^s\epsilon) = O_p(sK)$, Therefore Markov inequality implies that $u'P^s\epsilon/\sqrt{N} = O_p(sK/\sqrt{N})$.

To give 4, observe that $E(u_iP^s_{ij}\epsilon_j\epsilon_kP^s_{kl}u'_l) = 0$ if one of (i,j,k,l) is different from all the rest. Also $E(\epsilon^2_iu_iu'_i)$ is bounded by Assumption 1. Therefore we have

$$\begin{aligned}
E(u'P^s\epsilon\epsilon'P^su) &= \sum_i (P^s_{ii})^2E(\epsilon^2_iu_iu'_i) + \sum_{i\neq j} E(u_iP^s_{ii}\epsilon_i\epsilon_jP^s_{jj}u'_j) \\
&\quad + \sum_{i\neq j} E(u_iP^s_{ij}\epsilon_j\epsilon_iP^s_{ij}u'_j) + \sum_{i\neq j} E(u_iP^s_{ij}\epsilon^2_jP^s_{ji}u'_i) \\
&= O_P(1)\sum_i (P^s_{ii})^2 + \sigma_{u\epsilon}\sigma'_{u\epsilon}\sum_{i\neq j} P^s_{ii}P^s_{jj} + (\sigma_\epsilon\Sigma_u + \sigma_{u\epsilon}\sigma'_{u\epsilon})\sum_{i\neq j} P^s_{ij}P^s_{ij} \\
&= o_p(sK) + \sigma_{u\epsilon}\sigma'_{u\epsilon}(m + sK)^2 + (\sigma^2_\epsilon\Sigma_u + \sigma_{u\epsilon}\sigma'_{u\epsilon})(m + s^2K)
\end{aligned}$$

by Lemma 2.2, 2.3 and 2.4.

Assumption 1 also implies

$$E(f'\epsilon\epsilon' P^s u) = \sum_{i.j.k} f_i P^s_{jk} E(\epsilon_i \epsilon_j u'_k) = \sum_i f_i P^s_{ii} E(\epsilon_i^2 u'_i).$$

and furthermore together with Assumption 3 and Lemma 2.1,

$$\left| \sum_i f_i P^s_{ii} E(\epsilon_i^2 u'_i) \right| \leq \sum_i P^s_{ii} \cdot ||f_i|| \cdot ||E(\epsilon_i^2 u'_i)|| = O_p(sK)$$

which gives 5.

To prove 6, first we consider the function of a: $sK/a + a$ which is convex and whose minimum value is $2\sqrt{sK}$ with the minimizer $a = \sqrt{sK}$. If $\Delta_s = 0$, then $(\Delta_s/\sqrt{N})((sK)/N + \Delta_s) = 0$ and for $\Delta_s \neq 0$, $(\Delta_s/\sqrt{N})/((sK)/N + \Delta_s) = (sK/\sqrt{\Delta_s N} + \sqrt{\Delta_s N})^{-1} \leq 1/\sqrt{sK} \to 0$ since $sK \to \infty$.

7 is Lemma A.3(vii) in Donald and Newey (2001).

Now we are in the last step, 8. Let $Q = I - P^s$ and for some a and b let $\zeta_i = f_a(x_i, z_i)$ and $\mu_i = E(\epsilon_i^2 u_{ib}) P^s_{ii}$. Now the a, b th element of $E(f'(I - P^s)\epsilon\epsilon' P^s u)$ satisfies

$$\left| E(\sum_{i.j.k.l} \zeta_i Q_{ij} \epsilon_j \epsilon_k P^s_{kl} u_{lb}) \right| = \left| \sum_{i,j} \zeta_i Q_{ij} E(\epsilon_j^2 u_{jb}) P^s_{jj} \right| = |\zeta' Q \mu| \leq |\zeta' Q Q \zeta|^{1/2} |\mu' \mu|^{1/2}$$

where the inequality is the Cauchy-Schwartz inequality. Now $|\zeta' Q Q \zeta|^{1/2} = O_p((N\Delta_s)^{1/2})$ by the definition of Δ_s. $|\mu'\mu| \leq C \sum_i (P^s_{ii})^2$ for some constant C by Assumption 1 and applying Lemma 2(2) we have $|\mu'\mu| = o_p(sK)$. Therefore we have

$$E(f'(I - P^s)\epsilon\epsilon' P^s u/N) = O_p((N\Delta_s)^{1/2}) o_p(\sqrt{sK}) O_p(1/N) = o_p(\Delta_s^{1/2}\sqrt{sK}/\sqrt{N}).$$

□

A.1 Proof of Theorem 1 and 2

Proof. The shrinkage TSLS estimator has the following form.

$$\sqrt{N}(\hat{\delta} - \delta_0) = \hat{H}^{-1}\hat{h}, \quad \hat{H} = W' P^s W/N, \quad \hat{h} = W' P^s \epsilon/\sqrt{N}.$$

Also $\hat{H}$ and $\hat{h}$ are decomposed as

$$\begin{aligned}
\hat{h} &= h + T_1^h + T_2^h, \\
T_1^h &= -f'(I - P^s)\epsilon/\sqrt{N}, \quad T_2^h = u' P^s \epsilon/\sqrt{N} \\
\hat{H} &= H + T_1^H + T_2^H + Z^H \\
T_1^H &= -f'(I - P^s) f/N, \quad T_2^H = (u' f + f' u)/N \\
Z^H &= (u' P^s u + u'(I - P^s) f + f'(I - P^s) u)/N.
\end{aligned}$$

Now we will show that the conditions of Lemma 1 are satisfied and $S(s)$ has the form given in the theorem. Note that in our case $o_p((sK)^2/N + \Delta_s) = o_p(\rho_{K,N})$. So it is enough to show that the term is $o_p((sK)^2/N + \Delta_s)$ in order to show that a term is $o_p(\rho_{K,N})$.

Now $h = O_p(1)$ and $H = O_p(1)$ by Lemma 2(5). Since $T^h = T_1^h + T_2^h = -f'(I - P^s)\epsilon/\sqrt{N} + u'P^s\epsilon/\sqrt{N}$, Lemma 3(2) and 3(3) say that $T_1^h = O_p(\Delta_s^{1/2})$ and $T_2^h = O_p(sK/\sqrt{N})$ so $T^h = O_p(\Delta_s^{1/2}) + O_p(sK/\sqrt{N})$. $\Delta_s = o_p(1)$ by Lemma 3(1) and $sK/\sqrt{N} = o_p(1)$ by $(sK)^2/N = o_p(1)$. Therefore $T^h = o_p(1)$.

Next,

$$T_1^H = -\frac{f'(I - P^s)f}{N} = -e_f^s - s(1-s)\frac{f'P_Z f}{N} = -e_f^s + O_p(\frac{K^2}{N}) = O_p(\frac{(sK)^2}{N} + \Delta_s).$$

$T_2^H = O_p(1/\sqrt{N})$ by CLT. Note that $(sK)^2/N + \Delta_s = o_p(1)$. Then each of $((sK)^2/N + \Delta_s)^2$, $\frac{1}{N}$ and $((sK)^2/N + \Delta_s)/N$ are $o(\rho_{K,N})$ which implies $||T^H||^2 = o_p(\rho_{K,N})$.

Now we analyze $||T^h||||T^H||$. We have seen that $T^h = O_p(\Delta_s^{1/2}) + O_p(sK/\sqrt{N})$. and $T^H = O_p(\Delta_s) + O_p(1/\sqrt{N})$. Now $O_p(\Delta_s^{3/2}) = o_p(\Delta_s) = o_p(\rho_{K,N})$ by Lemma 3(1), $O_p(\Delta_s^{1/2}/\sqrt{N}) = o_p(sK/N + \Delta_s) = o_p(\rho_{K,N})$ by Lemma 3(6), $O_p(sK\Delta_s/\sqrt{N}) = o_p(\rho_{K,N})$ since $sK\Delta_s^{1/2}/\sqrt{N} \leq (sK)^2/N + \Delta_s$ and $\Delta_s^{1/2} = o_p(1)$ by Lemma 3(1), and $O_p(sK/N) = o_p(\rho_{K,N})$. Therefore $||T^h||||T^H|| = o_p(\rho_{K,N})$.

Since $||Z^h|| = 0$ in our case, $||Z^h|| = o_p(\rho_{K,N})$. The last part which we need to show $o_p(\rho_{K,N})$ is $||Z^H||$. Now $Z^H = u'P^s u/N + u'(I-P^s)f/N + f'(I-P^s)u/N$ where the first term is $O_p(sK/N) = o_p(\rho_{K,N})$ and the second and third term are $O_p(\Delta^{1/2}/\sqrt{N}) = o_p(sK/N + \Delta_s) = o_p(\rho_{K,N})$ by Lemma 3(6). Therefore we have $||Z^H|| = o_p(\rho_{K,N})$.

Note that we have shown $\hat{H} = H + o_p(1)$ and $\hat{h} = h + o_p(1)$. Then Proposition 1 holds by the LLN, the CLT and the Slutzky's Lemma.

We are going to the last part. Here we have $Z^A(s) = 0$ and $\hat{A}(s) = (h + T_1^h + T_2^h)(h + T_1^h + T_2^h)' - hh'H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}hh'$.

Now we calculate the expectation of each term in $A(s)$. First of all, $E(hh') = E(f\epsilon\epsilon' f'/N) = \sigma_\epsilon^2 H$. Second,

$$E(hT_1^{h\prime}) = E(-\frac{f\epsilon\epsilon(I - P^s)f'}{N}) = -\sigma_\epsilon^2\frac{f(I - P^s)f'}{N}.$$

Similarly $E(T_1^h h') = \sigma_\epsilon^2 f(I - P^s)f'/N$. Third,

$$E(hT_2^{h\prime}) = E(\frac{f\epsilon\epsilon' P^s u}{N}) = O_p(\frac{sK}{N})$$

by Lemma 3(5). This implies that $E(T_2^h h') = O_p(sK/N)$ also. Fourth,

$$E(T_1^h T_1^{h\prime}) = E(\frac{f'(I - P^s)\epsilon\epsilon'(I - P^s)f}{N}) = \sigma_\epsilon^2\frac{f'(I - P^s)(I - P^s)f}{N}.$$

Fifth,

$$E(T_1^h T_2^{h\prime}) = -E(\frac{f'(I - P^s)\epsilon\epsilon' P^s u}{N}) = o_p(\frac{\Delta_s^{1/2}\sqrt{sK}}{\sqrt{N}}).$$

by Lemma 3(8). Again we have $E(T_2^h T_1^{h\prime}) = o_p(\Delta_s^{1/2}\sqrt{sK}/\sqrt{N})$. Sixth,

$$E(T_2^h T_2^{h\prime}) = E(\frac{u'P^s\epsilon\epsilon' P^s u}{N}) = \sigma_{u\epsilon}\sigma_{u\epsilon}'\frac{(sK)^2}{N} + o_p(\frac{(sK)^2}{N})$$

by Lemma 3(4). Seventh,

$$E(hh'H^{-1}T_1^{H\prime}) = -E(\frac{f'\epsilon\epsilon' fH^{-1}f'(I-P^s)f}{N^2}) = -\sigma_\epsilon^2\frac{f'(I-P^s)f}{N}$$

also we have $E(T_1^H H^{-1}hh') = -\sigma_\epsilon^2 f'(I-P^s)f/N$. Finally, Lemma 3(7) implies

$$E(hh'H^{-1}T_2^{H\prime}) = E(\frac{hh'H^{-1}(u'f+f'u)}{N}) = O_p(\frac{1}{N}).$$

and $E(T_2^H H^{-1}hh') = O_P(1/N)$. Therefore we have

$$\begin{aligned}
E(\hat{A}(K)) &= \sigma_\epsilon^2 H - \sigma_\epsilon^2\frac{f'(I-P^s)f}{N} + O_p(\frac{sK}{N}) \\
&\quad -\sigma_\epsilon^2\frac{f'(I-P^s)f}{N} + \sigma_\epsilon^2\frac{f'(I-P^s)(I-P^s)f}{N} + o_p(\frac{\Delta_s^{1/2}\sqrt{sK}}{\sqrt{N}}) \\
&\quad +O_p(\frac{sK}{N}) + o_p(\frac{\Delta_s^{1/2}\sqrt{sK}}{\sqrt{N}}) + \sigma_{u\epsilon}\sigma'_{u\epsilon}\frac{(sK)^2}{N} + o_p(\frac{sK}{N}) \\
&\quad +\sigma_\epsilon^2\frac{f'(I-P^s)f}{N} + O_p(\frac{1}{N}) + \sigma_\epsilon^2\frac{f'(I-P^s)f}{N} + O_p(\frac{1}{N}) \\
&= \sigma_\epsilon^2 H + \sigma_{u\epsilon}\sigma'_{u\epsilon}\frac{(sK)^2}{N} + \sigma_\epsilon^2\frac{f'(I-P^s)(I-P^s)f}{N} + o_p(\rho_{K,N})
\end{aligned}$$

where the last equality holds because $1/N = o_p(\rho_{K,N})$, $sK/N = o_p(\rho_{K,N})$ and $o_p(\Delta_s^{1/2}\sqrt{sK}/\sqrt{N}) = o_p(\rho_{K,N})$ by the fact that $\Delta_s^{1/2}\sqrt{sK}/\sqrt{N} \leq sK/N + \Delta_s$. □

A.2 Proof of Theorem 3

Proof. Under the assumptions, $W'(P_X+P_Z)W/N - H \to_p 0$ and $W'P_ZW/N - f'P_Zf/N \to_p 0$. Let

$$V \equiv \frac{\sigma_\epsilon^2}{\lambda'H^{-1}\sigma_{u\epsilon}\sigma'_{u\epsilon}H^{-1}\lambda}\frac{\lambda'H^{-1}f'P_ZfH^{-1}\lambda}{N}, \quad \hat{V} \equiv \frac{\hat{\sigma}_\epsilon^2}{\hat{\lambda}'\hat{H}^{-1}\hat{\sigma}_{u\epsilon}\hat{\sigma}'_{u\epsilon}\hat{H}^{-1}\hat{\lambda}}\frac{\hat{\lambda}'\hat{H}^{-1}W'P_ZW\hat{H}^{-1}\hat{\lambda}}{N}.$$

Then $\hat{V} - V = o_p(1)$ and s^* and $\hat{s}^*$ can be written as $1-(1+VN/K^2)^{-1}$ and $1-(1+\hat{V}N/K^2)^{-1}$ respectively. Suppose $f'P_Zf/N \to c > 0$ for some c, then

$$\hat{s}^* - s^* = \frac{\hat{V}N/K^2 - VN/K^2}{(1+\hat{V}N/K^2)(1+VN/K^2)} = \frac{\hat{V}-V}{(\frac{K^2}{N}+\hat{V})(\frac{K^2}{N}+V)}\frac{K^2}{N} = o_p(K^2/N).$$

This implies that

$$H(S(\hat{s}^*) - S(s^*))H = (\hat{s}^{*2} - s^{*2})K^2/N + ((1-\hat{s}^*)^2 - (1-s^*)^2)f'P_Zf/N = o_p(K^2/N)$$

by the continuous mapping theorem. Since $S(s^*)$ is at least $O_p(K^2/N)$ in this case, the result holds.

Suppose $f'P_Zf = O_p(K)$ which occurs when Z is an irrelevant instrument, then $s^* = O_p(1/K)$ and $S(s^*) = O_p(1/N)$. Since $\mathrm{N}(\hat{V}-V)/K = o_p(1)$, we have

$$\hat{s}^* - s^* = \frac{\hat{V}N/K^2 - VN/K^2}{(1+\hat{V}N/K^2)(1+VN/K^2)} = \frac{N(\hat{V}-V)/K}{(1+\hat{V}N/K^2)(1+VN/K^2)}\frac{1}{K} = o_p(1/K).$$

It follows therefore that $S(\hat{s}^*) - S(s^*) = o_p(1/N)$.

□

A.3 Proof of Theorem 4

First we show that the consistent of shrinkage LIML and derive the asymptotic distribution of it under $sK/N \to 0$. Now our $\hat{\delta}$ is $\hat{\delta} = \text{argmin}_\delta (y - W\delta)' P^s (y - W\delta)/(y - W\delta)'(y - W\delta)$.

Lemma 4. *Assumption 1-3 are satisfied. Then under $sK/N \to 0$ and $s \to 1$ or $f' P_Z f/N \to_p 0$, $\hat{\delta} \to_p \delta_0$.*

Proof. Define $\bar{W} \equiv (y, W)$ and $D_0 \equiv (\delta, I)$. $\bar{W}$ can be written as $\bar{W} = W D_0 + \epsilon e_1$. where e_1 is the first unit vector. Let $\hat{A} = \bar{W}' P^s \bar{W}/N$ and $A = D_0' \bar{H} D_0$.

Observing Lemma A.4 and the proof of Lemma A.5 in Donald and Newey (2001), it is enough to show that $\hat{A} \to_p A$.

$\hat{A}$ has the following decomposition.

$$\hat{A} = D_0' \left(\frac{f'f}{N} - \frac{f'(I - P^s)f}{N} + \frac{u' P^s f}{N} + \frac{f' P^s u}{N} \right) D_0 + e_1 \frac{\epsilon' P^s W}{N} D_0 + D_0' \frac{W P^s \epsilon}{N} e_1' + \frac{\epsilon' P^s \epsilon}{N} e_1 e_1'.$$

First we have $f'f/N \to_p \bar{H}$ by the LLN. $f'(I - P^s)f/N = f'(I - P)f/N + (1 - s) f' P_Z f/N \to_p 0$ by Lemma A.2(1) in Donald and Newey (2001) and that $s \to_p 1$ or $f' P_Z f/N \to_p 0$. $E(\epsilon' P^s \epsilon) = tr(P^s E(\epsilon\epsilon')) = \sigma_\epsilon^2 (m + sK)$ which implies that $\epsilon' P^s \epsilon/N \to_p 0$ by Markov's inequality, Similarly we can show that $u' P^s u/N \to_p 0$. Let W_j be the jth column of W. Then

$$\left| \frac{W_j' P^s \epsilon}{N} \right| \leq \sqrt{\frac{W_j' P^s W_j}{N}} \sqrt{\frac{\epsilon' P^s \epsilon}{N}} \leq \sqrt{\frac{W_j' W_j}{N}} o_p(1) = o_p(1).$$

The first inequality is the Cauchy-Schwartz inequality and the second inequality comes from the fact that $I - P^s$ is positive definite which is because $I - P^s = I - P + (1 - s) P_Z$ and $I - P$ and P_Z are is positive definite and $1 - s \geq 0$. It follows therefore that $W' P^s \epsilon/N \to_p 0$. $f' P^s u/N \to_p 0$ similarly.

Summing up we have $\hat{A} \to_p A$. □

Lemma 5. *Assumption 1-3 are satisfied, $sK/N \to_p 0$ and $s \to_p 1$ or $f' P_Z f/N \to_p 0$, then $\sqrt{N}(\hat{\delta} - \delta_0) \to_d N(0, \sigma_\epsilon^2 \bar{H}^{-1})$.*

Proof. Let $A^s(\delta) \equiv (y - W\delta)' P^s (y - W\delta)/N$ and $B(\delta) \equiv (y - W\delta)'(y - W\delta)/N$. Define $\Lambda(\delta) \equiv A^s(\delta)/B(\delta)$ so that $\hat{\delta} = \text{argmin}_\delta \Lambda(\delta)$.

Let $\Lambda_\delta(\delta)$ and $\Lambda_{\delta\delta}(\delta)$ be the gradient and Hessian of $\Lambda(\delta)$ respectively. A standard Taylor expansion shows that

$$\sqrt{N}(\hat{\delta} - \delta_0) = -\Lambda_{\delta\delta}(\tilde{\delta})^{-1} \sqrt{N} \Lambda_\delta(\delta_0) = (\frac{\tilde{\sigma}_\epsilon^2 \Lambda_{\delta\delta}(\tilde{\delta})}{2})^{-1} (-\frac{\tilde{\sigma}_\epsilon^2 \sqrt{N} \Lambda_\delta(\delta_0)}{2})$$

for some mean value $\tilde{\delta}$. Now we have

$$\begin{aligned} \Lambda_\delta(\delta) &= B(\delta)^{-1}(A_\delta(\delta) - \Lambda(\delta) B_\delta(\delta)) \\ \Lambda_{\delta\delta}(\delta) &= B(\delta)^{-1}(A_{\delta\delta}(\delta) - \Lambda(\delta) B_{\delta\delta}(\delta)) - B(\delta)^{-1}(B_\delta(\delta)\Lambda_\delta(\delta)' + \Lambda_\delta(\delta) B_\delta'). \end{aligned}$$

Since $\hat{\delta} \to_p \delta_0$ by the Lemma 4, $\tilde{\delta} \to_p \delta_0$, which implies that $B(\tilde{\delta}) \to_p \sigma_\epsilon^2$, $B_\delta(\tilde{\delta}) \to_p -2\sigma_{u\epsilon}$. As in before $A(\tilde{\delta}) \to_p 0$ and therefore $\Lambda(\tilde{\delta}) \to_p 0$. Also $A_\delta(\tilde{\delta}) \to_p 0$ and therefore $\Lambda_\delta(\tilde{\delta}) \to_p 0$. $A_{\delta\delta}(\delta) = 2W' P^s W/N \to_p 2\bar{H}$. $B_{\delta\delta}(\delta) = 2W'W/N \to_p 2E(W_i W_i')$. Therefore we have $\tilde{\sigma}_\epsilon^2 \Lambda_{\delta\delta}(\tilde{\delta}) \to_p \bar{H}$.

Consider the gradient term. First define $\hat{\alpha} = W'\epsilon/\epsilon'\epsilon$ and $\alpha = \sigma_{u\epsilon}/\sigma_\epsilon^2$. $\hat{\alpha} - \alpha = O_p(1/N)$ by the CLT. We have the following decomposition.

$$-\frac{\tilde{\sigma}_\epsilon^2\sqrt{N}\Lambda_\delta(\delta_0)}{2} = \frac{W'P^s\epsilon}{\sqrt{N}} - \frac{\epsilon'P^s\epsilon W'\epsilon}{\sqrt{N}\epsilon'\epsilon} = h - \frac{f'(I-P^s)\epsilon}{\sqrt{N}} + \frac{v'P^s\epsilon}{\sqrt{N}} - (\hat{\alpha}-\alpha)\frac{\epsilon'P^s\epsilon}{\sqrt{N}}.$$

$h \to_d N(0, \sigma^2\bar{H})$ by the CLT. Lemma 2(1) and Chebyshev inequality says $f'(I-P^s)\epsilon/\sqrt{N} = o_p(1)$. A similar argument as in the proof of Lemma 2(4) with $E(v_i\epsilon_i) = 0$ implies that $v'P^s\epsilon/\sqrt{N} = O_p(\sqrt{sK/N}) = o_p(1)$. $\epsilon'P^s\epsilon = O_p(sK)$ as we see in the proof of Lemma 4. It follows therefore $(\hat{\alpha}-\alpha)\epsilon'P^s\epsilon/\sqrt{N} = O_p(sK/N) = o_p(1)$. Hence it holds $-\tilde{\sigma}_\epsilon^2\sqrt{N}\Lambda_\delta(\delta_0)/2 \to_d N(0, \sigma_\epsilon^2\bar{H})$.

In conclusion, we have $\sqrt{N}(\hat{\delta}-\delta) \to_d \bar{H}^{-1}N(0, \sigma_\epsilon^2\bar{H}) = N(0, \sigma_\epsilon^2\bar{H}^{-1})$ □

Define $\hat{\Lambda} = \min_\delta (y-W\delta)'P^s(y-W\delta)/(y-W\delta)'(y-W\delta)$ and $\tilde{\Lambda} = \epsilon'P^s\epsilon/(N\sigma_\epsilon^2)$. Also note that in LIML case, to show $o_p(\rho_{K,N})$, it is enough to show $o_p(sK/N + \Delta_s)$.

Lemma 6. *Assumption 1-3 are satisfied, $sK/N \to_p 0$ and $1 - s = o_p(K/N)$ or $f'P_Z f/N \to_p 0$, then*

$$\hat{\Lambda} = \tilde{\Lambda} - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda} - \frac{h'H^{-1}h}{2N\sigma_\epsilon^2} + \hat{R}_\Lambda = \tilde{\Lambda} + o_p(\frac{sK}{N}).$$

And $\sqrt{N}\hat{R}_\Lambda = o_p(\rho_{K,N})$.

Proof. We expand $\hat{\Lambda} = \Lambda(\hat{\delta})$ around the true value δ_0. Then

$$\begin{aligned}\hat{\Lambda} &= \Lambda(\delta_0) - \frac{\Lambda_\delta(\delta_0)'(\Lambda_{\delta\delta}(\delta_0))^{-1}\Lambda_\delta(\delta_0)}{2} + O_p(\frac{1}{N^{3/2}}) \\ &= \tilde{\Lambda} - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda} + \frac{(\tilde{\sigma}_\epsilon^2 - \sigma_\epsilon^2)^2}{\tilde{\sigma}_\epsilon^2\sigma_\epsilon^2}\tilde{\Lambda} - \frac{\Lambda_\delta(\delta_0)'(\Lambda_{\delta\delta}(\delta_0))^{-1}\Lambda_\delta(\delta_0)}{2} + O_p(\frac{1}{N^{3/2}}).\end{aligned}$$

We can see from the proof of Lemma 5,

$$-\frac{\tilde{\sigma}_\epsilon^2\sqrt{N}\Lambda_\delta(\delta_0)}{2} = h + O_p(\Delta_s^{1/2} + \frac{sK}{N}).$$

This also implies that $\Lambda_\delta = O_p(1/\sqrt{N})$. Then,

$$\frac{\tilde{\sigma}_\epsilon^2\Lambda_{\delta\delta}(\tilde{\delta})}{2} = \frac{W'P^sW}{N} - \Lambda(\delta_0)\frac{W'W}{N} + O_p(\frac{1}{\sqrt{N}}).$$

by $B_\delta(\delta_0) = O_p(1)$. It follows that

$$\frac{\tilde{\sigma}_\epsilon^2\Lambda_{\delta\delta}(\tilde{\delta})}{2} = H - \frac{f'(I-P^s)f}{N} + \frac{u'P^sf}{N} + \frac{f'P^su}{N} + \frac{u'P^su}{N} + O_p(\sqrt{\frac{sK}{N}}).$$

by $\Lambda(\delta_0) = O_p(\sqrt{sK/N})$. As in the proof of proposition 2, we have $f'(I-P^s)f/N = O_p(\Delta_s + sK/N)$. It holds also that $u'P^sf/N = O_p(1/\sqrt{N})$ and $u'P^su/N = O_p(sK/N)$. Summing up we have $\tilde{\sigma}_\epsilon^2\Lambda_{\delta\delta}(\tilde{\delta})/2 = H + O_p(\Delta_s^{1/2} + \sqrt{sK/N})$. Then we have

$$\frac{\Lambda_\delta(\delta_0)'(\Lambda_{\delta\delta}(\delta_0))^{-1}\Lambda_\delta(\delta_0)}{2} = \frac{h'H^{-1}h}{N\sigma_\epsilon^2} + O_p(\frac{\Delta_s^{1/2}}{N} + \sqrt{\frac{sK}{N^3}}).$$

Also it follows that $\left(\tilde{\sigma}_\epsilon^2/\sigma_\epsilon^2 - 1\right) = O_p(1/\sqrt{N})$ by the CLT and the Delta method. These results give the first equation of the lemma since

$$\hat{\Lambda} = \tilde{\Lambda} - \left(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2} - 1\right)\tilde{\Lambda} - \frac{h'H^{-1}h}{2N\sigma_\epsilon^2} + O_p(\frac{\Delta_s^{1/2}}{N} + \sqrt{\frac{sK}{N^3}}) + O_p(\frac{sK}{N^2}) + O_p(\frac{1}{N^{3/2}})$$

and all of remainder terms are $o_p(\rho_{K,N})$.

The second equation in the lemma is given by the fact that $\tilde{\Lambda} = O_p(sK/N)$. □

Lemma 7. *Assumption 1-3 are satisfied, $sK/N \to_p 0$ and $s \to_p 1$, then*

1. $u'P^s u/N - \tilde{\Lambda}\Sigma_u = o_p(sK/N)$,
2. $E(h\tilde{\Lambda}\epsilon' v/\sqrt{N}) = (m + sK)/N \cdot \sum_i f_i E(\epsilon_i^2 v_i')/N + O_p(sK/N^2)$,
3. $E(hh'H^{-1}h/\sqrt{N}) = O_p(1/N)$.

Proof. We begin with the proof of (1). $E(\tilde{\Lambda}) = tr(P^s E(\epsilon\epsilon'))/(N\sigma_\epsilon^2) = (m + sK)/N$ and

$$\begin{aligned} E((\tilde{\Lambda} - \frac{m+sK}{N})^2) &= \frac{E(\epsilon' P^s \epsilon \epsilon' P^s \epsilon)}{N^2\sigma_\epsilon^4} - (\frac{m+sK}{N})^2 \\ &= \frac{\sigma_\epsilon^4(m+sK)^2 + o_p((sK)^2)}{N^2\sigma_\epsilon^4} - (\frac{m+sK}{N})^2 = o_p((\frac{sK}{N})^2) \end{aligned}$$

by Lemma 3(4) with replacing u by ϵ. This gives $(\tilde{\Lambda} - (m+sK)/N)\Sigma_u = o_p(sK/N)$. We also have $E(u'P^s u) = (m + sK)\Sigma_u$ and $u'P^s u/N - ((m+sK)/N)\Sigma_u = o_p(sK/N)$. Therefore 1 is proved.

We observe

$$\begin{aligned} E(\frac{h\tilde{\Lambda}\epsilon' v}{\sqrt{N}}) &= \frac{\sum_{i,j,k,l} E(f_i \epsilon_i \epsilon_j P_{jk}^s \epsilon_k \epsilon_l v_l')}{N^2\sigma_\epsilon^2} \\ &= \frac{\sum_i f_i P_{ii}^s E(\epsilon_i^4 v_i')}{N^2\sigma_\epsilon^2} + 2\frac{\sum_{i\neq j} f_i P_{ij}^s E(\epsilon_j^2 v_j')}{N^2} + \frac{\sum_{i\neq j} f_i P_{jj}^s E(\epsilon_i^2 v_i')}{N^2} \\ &= O_p(\frac{sK}{N^2}) + o_p(\frac{sK}{N^2}) + \frac{m+sK}{N}\frac{\sum_i f_i E(\epsilon_i^2 v_i')}{N} \end{aligned}$$

which gives 2.

3 is Lemma A.8(iii) in Donald and Newey (2001). □

Proof of Theorem 4. The consistency and the asymptotic normality of the shrinkage estimator stems from Lemma 4 and Lemma 5. The shrinkage estimator LIML estimator has the following representation.

$$\sqrt{N}(\hat{\delta} - \delta_0) = \hat{H}^{-1}\hat{h}, \quad \hat{H} = \frac{W'P^s W}{N} - \hat{\Lambda}\frac{W'W}{N}, \quad \hat{h} = \frac{W'P^s\epsilon}{\sqrt{N}} - \hat{\Lambda}\frac{W'\epsilon}{\sqrt{N}}.$$

As in the case of TSLS, we are going to verify the assumption of Lemma 1. First $\hat{H}$ and $\hat{h}$ have

the decomposition:

$$\begin{aligned}
\hat{h} &= h + \sum_{i=1}^{5} T_i^h + Z^h, \\
T_1^h &= -\frac{f'(I-P^s)\epsilon}{\sqrt{N}} = O_p(\Delta_s^{1/2}), \quad T_2^h = \frac{v'P^s\epsilon}{\sqrt{N}} = O_p(\sqrt{\frac{sK}{N}}), \\
T_3^h &= -\tilde{\Lambda}h = O_p(\frac{sK}{N}), \quad T_4^h = -\tilde{\Lambda}\frac{v'\epsilon}{\sqrt{N}} = O_p(\frac{sK}{N}), \\
T_5^h &= \frac{h'H^{-1}h}{2\sqrt{N}\sigma_\epsilon^2}\sigma_{u\epsilon} = O_p(\frac{1}{\sqrt{N}}), \\
Z^h &= -(\hat{\Lambda}-\tilde{\Lambda})h + \sqrt{N}(\frac{\tilde{\sigma}_\epsilon^2}{\sigma_\epsilon^2}-1)\tilde{\Lambda}(\frac{u'\epsilon}{N}-\sigma_{u\epsilon}) + \frac{h'H^{-1}h}{2\sqrt{N}\sigma_\epsilon^2}(\frac{u'\epsilon}{N}-\sigma_{u\epsilon}) - \sqrt{N}\hat{R}_\Lambda\frac{u'\epsilon}{N}
\end{aligned}$$

and

$$\begin{aligned}
\hat{H} &= H + \sum_{i=1}^{3} T_i^H + Z^H, \\
T_1^H &= -\frac{f'(I-P^s)f}{N} = O_p(\frac{sK}{N}+\Delta_s), \quad T_2^H = \frac{u'f+f'u}{N} = O_p(\frac{1}{\sqrt{N}}), \quad T_3^H = -\tilde{\Lambda}H = O_p(\frac{sK}{N}), \\
Z^H &= -\frac{u'(I-P^s)f}{N} - \frac{f'(I-P^s)u}{N} + \frac{uP^su}{N} - \tilde{\Lambda}\frac{u'u}{N} - \tilde{\Lambda}(\frac{u'f+f'u}{N}) - (\hat{\Lambda}-\tilde{\Lambda})\frac{W'W}{N}.
\end{aligned}$$

$h = O_p(1)$ and $H = O_p(1)$ by Lemma 3(8). $T^h = o_p(1)$ since all of $\Delta_s^{1/2}$, $\sqrt{sK/N}$, sK/N and $1/\sqrt{N}$ are $o_p(1)$.

$||T_1^H||^2$ consists of terms of order $(sK/N+\Delta_s)^2$, $1/N$, $(sK/N)^2$, $(sK/N+\Delta_s)/\sqrt{N}$, $(sK/N+\Delta_s)\cdot sK/N$ and $sK/N^{3/2}$. It is easy to see that all of them are $o_p(\rho_{K,N})$. It follows that $||T_1^H||^2 = o_p(\rho_{K,N})$.

Similarly $||T^h||\cdot||T^H||$ consists of terms of order $(sK/N+\Delta_s)o_p(1)$, $\Delta_s^{1/2}/\sqrt{N}$, $\sqrt{sK}/N$, $1/N$ and $sK/N\cdot o_p(1)$. A simple inspection and Lemma 3(6) say that all of them are $o_p(\rho_{K,N})$. That gives $||T^h||\cdot||T^H|| = o_p(\rho_{K,N})$.

To show $Z^h = o_p(\rho_{K,N})$, we investigate each term of Z^h. $(\hat{\Lambda}-\tilde{\Lambda})h = o_p(sK/N)O_p(1) = o_p(\rho_{K,N})$ by Lemma 6. $\sqrt{N}(\tilde{\sigma}_\epsilon^2/\sigma_\epsilon^2-1)\tilde{\Lambda}(u'\epsilon/N-\sigma_{u\epsilon}) = O_p(1)O_p(sK/N)O_p(1/\sqrt{N}) = O_p(sK/N^{3/2}) = o_p(\rho_{K,N})$ by the CLT and the delta method. $h'H^{-1}h/(2\sqrt{N}\sigma_\epsilon^2)\cdot(u'\epsilon/N-\sigma_{u\epsilon}) = O_p(1/\sqrt{N})O_p(1/\sqrt{N}) = O_p(1/N) = o_p(\rho_{K,N})$ by the CLT. $\sqrt{N}\hat{R}_\Lambda u'\epsilon/N = o_p(\rho_{K,N})O_p(1) = o_p(\rho_{K,N})$ by the LLN and Lemma 6. Therefore $Z^h = o_p(\rho_{K,N})$.

Similarly, the each term of Z^H is shown to be $o_p(\rho_{K,N})$. $u'(I-P^s)f/N = O_p(\Delta_s^{1/2}/\sqrt{N}) = o_p(\rho_{K,N})$ where the first equality can be verified as in the proof of Lemma 3(2) and the second equality is Lemma 3(6). $uP^su/N - \tilde{\Lambda}u'u/N = uP^su/N - \tilde{\Lambda}\Sigma_u - \tilde{\Lambda}(u'u/N-\Sigma_u) = o_p(sK/N) + O_p(sK/N)o_p(1) = o_p(\rho_{K,N})$ by Lemma 7(1) and the LLN. The CLT implies $\tilde{\Lambda}(u'f+f'u)/N = O_p(sK/N)O_p(1/\sqrt{N}) = o_p(\rho_{K,N})$. Finally $(\hat{\Lambda}-\tilde{\Lambda})W'W/N = o_p(sK/N)O_p(1) = o_p(\rho_{K,N})$ by the LLN and Lemma 6. Hence we have $Z^H = o_p(\rho_{K,N})$.

Consider the decomposition

$$(h+\sum_{i=1}^{5}T_i^h)(h+\sum_{i=1}^{5}T_i^h)' - hh'H^{-1}\sum_{i=1}^{3}T_i^{H\prime} - \sum_{i=1}^{3}T_i^H H^{-1}hh' = A(s) + Z^A(s)$$

where

$$A(s) \equiv hh' + \sum_{i=1}^{5} hT_i^{h\prime} + \sum_{i=1}^{5} T_i^h h' + (T_1^h + T_2^h)(T_1^h + T_2^h)' - hh'H^{-1}\sum_{i=1}^{3} T_i^{H\prime} - \sum_{i=1}^{3} T_i^H H^{-1} hh'$$

$$Z^A(s) \equiv (\sum_{i=3}^{5} T_i^h)(\sum_{i=3}^{5} T_i^h)' + (\sum_{i=3}^{5} T_i^h)(T_1^h + T_2^h)' + (T_1^h + T_2^h)(\sum_{i=3}^{5} T_i^h)'$$

$Z^A(s)$ consists of terms of order $(sK/N)^2$, $sK/N^{3/2}$, $1/N$, $\Delta_s^{1/2} sK/N$, $\Delta_s^{1/2}/\sqrt{N}$, $(sK/N)^{3/2}$ and $\sqrt{sK}/N$. All of them are $o_p(\rho_{K,N})$ by a simple inspection and Lemma 3(6). $Z^A(s) = o_p(\rho_{K,N})$.

What is remained to be shown is the expectation of $A(s)$. As we saw in the TSLS case, we have $E(hh') = \sigma_\epsilon^2 H$, $E(hT_1^{h\prime}) = E(T_1^h h') = -\sigma_\epsilon^2 f'(I - P^s)f/N$, $E(T_1^h T_1^{h\prime}) = \sigma_\epsilon^2 f'(I - P^s)(I - P^s)f/N$, $E(T_1^h T_2^{h\prime}) = o_p(\Delta_s^{1/2}\sqrt{sK/N}) = o_p(\rho_{K,N})$, similarly $E(T_2^h T_1^{h\prime}) = o_p(\rho_{K,N})$, $E(hh'H^{-1}T_1^{H\prime}) = E(T_1^H H^{-1} hh') = -\sigma_\epsilon^2 f'(I - P^s)f/N$, $E(hh'H^{-1}T_2^{H\prime}) = O_p(1/N) = o_p(\rho_{K,N})$ and similarly $E(T_2^H H^{-1} hh') = o_p(\rho_{K,N})$.

A similar argument as in the proof of Lemma 3(6) noting that $E(v_i\epsilon_i) = 0$ gives

$$E(T_2^h T_2^{h\prime}) = \sigma_\epsilon^2 \Sigma_v \frac{s^2 K}{N} + o_p(\rho_{K,N}).$$

Lemma 7(3) shows

$$E(hT_5^{h\prime}) = E(\frac{hh'H^{-1}h}{2N\sigma_\epsilon^2}\sigma_{u\epsilon}) = O_p(\frac{1}{N}) = o_p(\rho_{K,N}).$$

Similarly, $E(T_5^h h) = o_p(\rho_{K,N})$.

Lemma 7(2) gives

$$E(hT_4^{h\prime}) = E(\frac{h\tilde{\Lambda}\epsilon' v}{\sqrt{N}}) = -\frac{sK}{N}\frac{\sum_i f_i E(\epsilon_i^2 v_i')}{N} + o_p(\rho_{K,N}).$$

Also we have $E(hT_2^{h\prime}) = \sum_i f_i P_{ii}^s E(\epsilon_i^2 v_i')/N$. Letting $\hat{\zeta} \equiv \sum_i f_i P_{ii}^s E(\epsilon_i^2 v_i')/N - sK/N \cdot \sum_i f_i E(\epsilon_i^2 v_i')/N$, $E(hT_2^{h\prime}) + E(hT_4^{h\prime}) = \hat{\zeta} + o_p(\rho_{K,N})$ and $E(T_2^h h') + E(T_4^h h') = \hat{\zeta}' + o_p(\rho_{K,N})$.

Summing up, we have

$$\begin{aligned} E(A(s)) &= \sigma_\epsilon^2 H - 2\sigma_\epsilon^2 \frac{f'(I - P^s)f}{N} + \hat{\zeta} + \hat{\zeta}' \\ &\quad + \sigma_\epsilon^2 \frac{f'(I - P^s)(I - P^s)f}{N} + \sigma_\epsilon^2 \Sigma_v \frac{s^2 K}{N} + 2\sigma_\epsilon^2 \frac{f'(I - P^s)f}{N} + o_p(\rho_{K,N}) \\ &= \sigma_\epsilon^2 H + \sigma_\epsilon^2 \Sigma_v \frac{s^2 K}{N} + \sigma_\epsilon^2 \frac{f'(I - P^s)(I - P^s)f}{N} + \hat{\zeta} + \hat{\zeta}' + o_p(\rho_{K,N}). \end{aligned}$$

Note that under $E(\epsilon_i^2 v_i) = 0$, $\hat{\zeta} = 0$. □

References

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

[2] P. A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 62(3):657–681, 1994.

[3] J. Bound, D. A. Jaeger, and R. M. Baker. Problems with instrumental variables estimation when correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1996.

[4] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306, 2004.

[5] J. C. Chao and N. R. Swanson. Consistent estimation with a large number of weak instruments. unpublished manuscript, 2003.

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

[7] J. A. Doornik. *Object-Oriented Matrix Programming Using Ox*. Timberlake Consultants Press and Oxford, London, 3rd edition, 2002. www.nuff.ox.ac.uk/Users/Doornik.

[8] E. I. George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81(394):437–445, 1986.

[9] J. Hahn and J. Hausman. A new specification test for the validity of instrumental variables. *Econometrica*, 70(1):163–189, 2002.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; data mining, inference, and prediction*. Springer, New York, 2001.

[11] G. M. Kuersteiner. Mean squared error reduction for GMM estimators of linear time series models. unpublished manuscript, 2000.

[12] G. M. Kuersteiner. Kernel weighted GMM for conditionally heteroskedastic models. unpublished manuscript, 2001.

[13] N. Kunitomo. Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association*, 75:693–700, 1980.

[14] E. L. Lehmann. *Theory of Point Estimation*. John Wiley and Sons, New York, 1983.

[15] K. Morimune. Approximate distributions of k-class estimators when the degree of overidentification is large compared with the sample size. *Econometrica*, 51(3):821–841, 1983.

[16] A. L. Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 27(4):575–595, 1959.

[17] P. C. B. Phillips. Exact small sample theory in the simultaneous equations model. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometris*, volume 1, chapter 8. North-Holland Publishing Company, 1983.

[18] D. Small. *Inference and model selction for instrumental variables regression*. PhD thesis, Stanford University, 2002.

[19] D. Staiger and J. H. Stock. Instrumental variables regression with weak insturments. *Econometrica*, 65(3):557–586, 1997.

[20] J. H. Stock and M. Yogo. Asymptotic distribution of instrumental variables statistics with many weak instruments. unpublished manuscript, 2003.

[21] Y. Takada. A family of minimax estimators in some multiple regression problems. *The Annals of Statistics*, 7(5):1144–1147, 1979.

Table 1: model (a), $R_f^2 = 0.1$

		OLS	TSLS	IV,1	DNTSLS	JSTSLS	OSTSLS	STSLS	LIML	DNLIML	SLIML
$c = 0.1$											
$n = 100$											
	bias	0.085	0.067	0.026	0.046	0.062	0.064	0.064	0.070	0.055	0.064
	MAD	0.092	0.133	0.718	0.270	0.152	0.134	0.161	0.361	0.265	0.349
	Dec. Rge	0.245	0.482	4.164	1.690	0.581	0.479	0.597	1.998	1.195	1.834
	Cov. Rate	0.847	0.934	0.998	0.979	0.945	0.933	0.950	0.983	0.985	0.983
$n = 500$											
	bias	0.092	0.031	0.006	0.030	0.030	0.030	0.031	-0.007	-0.000	-0.005
	MAD	0.092	0.081	0.415	0.084	0.081	0.080	0.081	0.113	0.112	0.113
	Dec. Rge	0.106	0.291	2.367	0.300	0.296	0.292	0.294	0.432	0.426	0.431
	Cov. Rate	0.429	0.950	0.996	0.954	0.952	0.951	0.949	0.965	0.965	0.966
$c = 0.5$											
$n = 100$											
	bias	0.442	0.325	0.243	0.300	0.321	0.304	0.312	0.041	0.238	0.054
	MAD	0.441	0.325	0.726	0.392	0.338	0.309	0.319	0.329	0.321	0.323
	Dec. Rge	0.225	0.433	4.147	1.673	0.526	0.636	0.578	1.161	1.069	1.540
	Cov. Rate	0.000	0.471	0.978	0.733	0.555	0.704	0.550	0.937	0.901	0.933
$n = 500$											
	bias	0.463	0.153	0.073	0.196	0.157	0.144	0.148	-0.007	0.026	-0.004
	MAD	0.443	0.153	0.419	0.219	0.152	0.144	0.150	0.108	0.108	0.108
	Dec. Rge	0.092	0.271	2.328	0.453	0.272	0.284	0.284	0.411	0.397	0.416
	Cov. Rate	0.000	0.665	0.971	0.706	0.676	0.724	0.681	0.961	0.946	0.961
$c = 0.9$											
$n = 100$											
	bias	0.811	0.569	0.410	0.521	0.567	0.511	0.544	0.020	0.302	0.022
	MAD	0.811	0.569	0.704	0.610	0.570	0.511	0.544	0.232	0.385	0.233
	Dec. Rge	0.128	0.305	3.686	2.138	0.390	0.779	0.469	1.189	1.104	1.198
	Cov. Rate	0.000	0.017	0.849	0.589	0.119	0.450	0.165	0.937	0.723	0.936
$n = 500$											
	bias	0.811	0.273	0.104	0.283	0.272	0.249	0.276	0.005	0.030	0.006
	MAD	0.811	0.273	0.403	0.347	0.272	0.250	0.276	0.095	0.096	0.095
	Dec. Rge	0.054	0.213	2.211	1.106	0.212	0.295	0.261	0.365	0.353	0.367
	Cov. Rate	0.000	0.202	0.895	0.706	0.207	0.494	0.310	0.950	0.941	0.952

Table 2: model (a), $R_f^2 = 0.01$

		OLS	TSLS	IV,1	DNTSLS	JSTSLS	OSTSLS	STSLS	LIML	DNLIML	SLIML
$c = 0.1$											
$n = 100$											
	bias	0.093	0.101	0.077	0.052	0.112	0.096	0.096	0.319	0.129	0.285
	MAD	0.099	0.172	0.912	0.561	0.265	0.187	0.219	0.910	0.482	0.858
	Dec. Rge	0.255	0.575	6.418	4.264	1.702	0.654	0.855	15.981	2.460	8.131
	Cov. Rate	0.836	0.935	0.999	0.991	0.972	0.952	0.952	0.994	0.992	0.995
$n = 500$											
	bias	0.101	0.083	0.163	0.117	0.096	0.080	0.084	0.168	0.109	0.129
	MAD	0.101	0.143	0.853	0.465	0.200	0.143	0.173	0.589	0.363	0.553
	Dec. Rge	0.108	0.478	5.204	2.931	0.914	0.487	0.678	4.666	1.528	4.338
	Cov. Rate	0.389	0.936	1.000	0.986	0.951	0.929	0.951	0.991	0.992	0.990
$c = 0.5$											
$n = 100$											
	bias	0.488	0.479	0.428	0.473	0.482	0.462	0.482	0.361	0.450	0.377
	MAD	0.488	0.479	0.948	0.686	0.527	0.582	0.493	0.889	0.592	0.870
	Dec. Rge	0.234	0.538	4.606	3.301	1.575	1.854	0.804	4.726	2.037	4.144
	Cov. Rate	0.000	0.324	0.979	0.829	0.603	0.932	0.459	0.900	0.892	0.900
$n = 500$											
	bias	0.498	0.414	0.367	0.402	0.414	0.402	0.408	0.078	0.388	0.108
	MAD	0.498	0.414	0.851	0.583	0.440	0.418	0.413	0.525	0.468	0.517
	Dec. Rge	0.097	0.437	4.604	2.633	0.763	0.996	0.571	3.459	1.495	3.434
	Cov. Rate	0.000	0.308	0.984	0.777	0.521	0.791	0.405	0.945	0.896	0.942
$c = 0.9$											
$n = 100$											
	bias	0.894	0.851	0.838	0.830	0.844	0.832	0.858	0.184	0.730	0.220
	MAD	0.898	0.851	0.966	0.883	0.853	0.840	0.859	0.825	0.848	0.817
	Dec. Rge	0.116	0.276	2.920	2.059	0.873	1.541	0.457	*	*	*
	Cov. Rate	0.000	0.000	0.771	0.544	0.288	0.701	0.034	0.739	0.506	0.735
$n = 500$											
	bias	0.890	0.745	0.649	0.719	0.732	0.699	0.744	0.025	0.548	0.045
	MAD	0.890	0.745	0.852	0.773	0.741	0.700	0.744	0.382	0.633	0.375
	Dec. Rge	0.050	0.266	3.547	2.173	0.516	1.039	0.427	4.756	1.625	4.247
	Cov. Rate	0.000	0.000	0.794	0.547	0.195	0.471	0.060	0.895	0.579	0.896

Table 3: model (b), $R_f^2 = 0.1$

		OLS	TSLS	IV,1	DNTSLS	JSTSLS	OSTSLS	STSLS	LIML	DNLIML	SLIML
$c = 0.1$											
$n = 100$											
	bias	0.084	0.062	-0.020	0.022	0.038	0.053	0.046	0.065	0.013	0.010
	MAD	0.091	0.134	0.286	0.216	0.160	0.135	0.162	0.380	0.214	0.309
	Dec. Rge	0.254	0.486	1.233	0.970	0.650	0.489	0.634	2.005	0.978	1.412
	Cov. Rate	0.838	0.933	0.990	0.975	0.961	0.944	0.951	0.982	0.984	0.981
$n = 500$											
	bias	0.091	0.026	-0.014	0.017	0.016	0.022	0.003	0.010	0.018	0.013
	MAD	0.091	0.079	0.119	0.090	0.080	0.079	0.085	0.115	0.109	0.107
	Dec. Rge	0.105	0.289	0.498	0.350	0.305	0.291	0.318	0.448	0.425	0.420
	Cov. Rate	0.432	0.949	0.963	0.951	0.950	0.954	0.928	0.956	0.954	0.950
$c = 0.5$											
$n = 100$											
	bias	0.443	0.319	-0.015	0.129	0.213	0.088	0.224	0.040	0.155	0.045
	MAD	0.443	0.319	0.286	0.284	0.260	0.228	0.271	0.356	0.279	0.303
	Dec. Rge	0.223	0.443	1.298	1.056	0.655	0.777	0.718	1.775	1.023	1.403
	Cov. Rate	0.000	0.492	0.956	0.804	0.729	0.914	0.685	0.916	0.905	0.917
$n = 500$											
	bias	0.452	0.146	-0.014	0.082	0.111	0.049	0.089	0.002	0.050	0.001
	MAD	0.452	0.147	0.120	0.142	0.120	0.098	0.121	0.108	0.107	0.103
	Dec. Rge	0.092	0.266	0.505	0.453	0.287	0.347	0.367	0.415	0.371	0.408
	Cov. Rate	0.000	0.689	0.963	0.832	0.798	0.942	0.827	0.968	0.934	0.966
$c = 0.9$											
$n = 100$											
	bias	0.811	0.565	-0.015	0.128	0.410	0.046	0.245	-0.024	0.153	-0.007
	MAD	0.811	0.565	0.265	0.332	0.429	0.238	0.283	0.242	0.267	0.231
	Dec. Rge	0.127	0.308	1.367	1.274	0.671	0.895	0.667	1.340	1.113	1.176
	Cov. Rate	0.000	0.017	0.912	0.760	0.367	0.889	0.689	0.947	0.852	0.938
$n = 500$											
	bias	0.811	0.270	-0.014	0.035	0.212	0.030	0.081	0.005	0.046	0.005
	MAD	0.811	0.269	0.120	0.125	0.212	0.109	0.120	0.094	0.096	0.093
	Dec. Rge	0.056	0.203	0.520	0.501	0.247	0.418	0.367	0.367	0.340	0.360
	Cov. Rate	0.000	0.200	0.950	0.902	0.429	0.938	0.856	0.952	0.921	0.950

Table 4: model (b), $R_f^2 = 0.01$

		OLS	TSLS	IV,1	DNTSLS	JSTSLS	OSTSLS	STSLS	LIML	DNLIML	SLIML
$c = 0.1$											
$n = 100$											
	bias	0.093	0.100	0.055	0.051	0.093	0.089	0.068	0.275	0.108	0.238
	MAD	0.098	0.170	0.757	0.497	0.273	0.210	0.220	0.892	0.460	0.802
	Dec. Rge	0.255	0.576	4.427	3.063	1.677	0.786	0.926	9.802	2.237	6.833
	Cov. Rate	0.830	0.933	0.999	0.992	0.972	0.973	0.953	0.990	0.993	0.991
$n = 500$											
	bias	0.100	0.080	-0.003	0.034	0.062	0.063	0.060	0.105	0.056	0.076
	MAD	0.100	0.140	0.403	0.316	0.199	0.153	0.184	0.573	0.303	0.472
	Dec. Rge	0.108	0.482	2.167	1.575	0.845	0.527	0.749	3.675	1.390	2.775
	Cov. Rate	0.388	0.934	0.994	0.983	0.968	0.951	0.952	0.984	0.987	0.983
$c = 0.5$											
$n = 100$											
	bias	0.488	0.485	0.265	0.365	0.431	0.313	0.458	0.296	0.396	0.301
	MAD	0.488	0.485	0.768	0.623	0.514	0.593	0.464	0.897	0.532	0.820
	Dec. Rge	0.235	0.539	4.572	3.012	1.375	2.161	0.823	5.925	1.970	5.344
	Cov. Rate	0.000	0.320	0.974	0.828	0.633	0.962	0.482	0.921	0.902	0.918
$n = 500$											
	bias	0.498	0.407	0.046	0.199	0.291	0.134	0.303	0.083	0.228	0.102
	MAD	0.498	0.407	0.388	0.390	0.353	0.299	0.337	0.534	0.378	0.482
	Dec. Rge	0.096	0.433	2.014	1.564	0.934	1.187	0.782	3.359	1.382	2.530
	Cov. Rate	0.000	0.313	0.969	0.822	0.686	0.932	0.571	0.939	0.914	0.934
$c = 0.9$											
$n = 100$											
	bias	0.893	0.852	0.460	0.675	0.770	0.490	0.794	0.113	0.549	0.196
	MAD	0.893	0.852	0.716	0.790	0.799	0.654	0.794	0.857	0.778	0.805
	Dec. Rge	0.115	0.283	4.195	2.810	1.292	2.686	0.717	*	*	*
	Cov. Rate	0.000	0.000	0.824	0.604	0.345	0.803	0.159	0.762	0.589	0.748
$n = 500$											
	bias	0.890	0.724	0.057	0.234	0.530	0.116	0.417	0.021	0.280	0.050
	MAD	0.890	0.742	0.375	0.484	0.573	0.325	0.422	0.399	0.429	0.376
	Dec. Rge	0.050	0.268	2.258	1.892	1.069	1.317	0.715	4.667	2.775	3.787
	Cov. Rate	0.000	0.000	0.900	0.711	0.397	0.870	0.518	0.902	0.764	0.901

Table 5: model (c), $R_f^2 = 0.1$

		OLS	TSLS	IV,1	DNTSLS	JSTSLS	OSTSLS	STSLS	LIML	DNLIML	SLIML
$c = 0.1$											
$n = 100$											
	bias	0.083	0.056	-0.021	0.016	0.041	0.051	0.051	0.040	0.013	0.015
	MAD	0.092	0.137	0.346	0.194	0.157	0.140	0.150	0.372	0.208	0.325
	Dec. Rge	0.245	0.484	1.594	0.816	0.644	0.486	0.584	1.871	0.890	1.534
	Cov. Rate	0.847	0.937	0.994	0.965	0.956	0.938	0.941	0.973	0.978	0.975
$n = 500$											
	bias	0.091	0.025	-0.018	0.005	0.020	0.024	0.033	0.008	0.010	0.008
	MAD	0.091	0.077	0.158	0.084	0.077	0.077	0.081	0.113	0.092	0.109
	Dec. Rge	0.108	0.292	0.660	0.319	0.293	0.292	0.285	0.421	0.359	0.416
	Cov. Rate	0.428	0.948	0.980	0.949	0.947	0.950	0.937	0.948	0.957	0.951
$c = 0.5$											
$n = 100$											
	bias	0.440	0.319	-0.009	0.143	0.249	0.141	0.249	0.017	0.099	0.014
	MAD	0.440	0.319	0.352	0.237	0.283	0.225	0.273	0.346	0.218	0.314
	Dec. Rge	0.225	0.463	1.634	0.835	0.616	0.757	0.649	1.613	0.852	1.472
	Cov. Rate	0.000	0.491	0.960	0.818	0.671	0.891	0.652	0.917	0.917	0.921
$n = 500$											
	bias	0.452	0.144	-0.018	0.052	0.132	0.086	0.116	0.012	0.035	0.011
	MAD	0.452	0.145	0.157	0.098	0.135	0.112	0.139	0.105	0.097	0.103
	Dec. Rge	0.093	0.257	0.676	0.352	0.272	0.329	0.354	0.425	0.365	0.412
	Cov. Rate	0.000	0.700	0.964	0.885	0.752	0.910	0.781	0.969	0.950	0.967
$c = 0.9$											
$n = 100$											
	bias	0.809	0.563	0.015	0.212	0.460	0.120	0.327	-0.026	0.112	-0.001
	MAD	0.809	0.563	0.318	0.310	0.482	0.254	0.336	0.243	0.233	0.234
	Dec. Rge	0.128	0.293	1.612	1.057	0.655	0.858	0.602	1.330	0.934	1.231
	Cov. Rate	0.000	0.021	0.906	0.734	0.287	0.849	0.586	0.939	0.866	0.934
$n = 500$											
	bias	0.811	0.272	-0.019	0.084	0.245	0.074	0.129	0.006	0.021	0.007
	MAD	0.811	0.272	0.153	0.116	0.245	0.123	0.149	0.092	0.089	0.093
	Dec. Rge	0.057	0.205	0.714	0.349	0.220	0.408	0.361	0.358	0.341	0.354
	Cov. Rate	0.000	0.195	0.945	0.851	0.312	0.903	0.789	0.943	0.926	0.942

Table 6: model (c), $R_f^2 = 0.01$

		OLS	TSLS	IV,1	DNTSLS	JSTSLS	OSTSLS	STSLS	LIML	DNLIML	SLIML
$c = 0.1$											
$n = 100$											
	bias	0.092	0.098	0.033	0.064	0.086	0.094	0.068	0.297	0.108	0.269
	MAD	0.098	0.170	0.798	0.460	0.275	0.199	0.221	0.882	0.423	0.818
	Dec. Rge	0.252	0.576	5.068	3.070	1.696	0.718	0.920	9.503	2.046	9.091
	Cov. Rate	0.830	0.931	0.999	0.991	0.973	0.958	0.950	0.990	0.995	0.991
$n = 500$											
	bias	0.101	0.082	0.021	0.058	0.072	0.073	0.071	0.130	0.059	0.107
	MAD	0.101	0.139	0.500	0.300	0.193	0.144	0.174	0.576	0.288	0.529
	Dec. Rge	0.109	0.495	3.002	1.421	0.870	0.528	0.685	3.726	1.280	3.107
	Cov. Rate	0.389	0.932	0.996	0.979	0.965	0.936	0.950	0.992	0.987	0.991
$c = 0.5$											
$n = 100$											
	bias	0.487	0.480	0.317	0.357	0.446	0.344	0.472	0.323	0.398	0.317
	MAD	0.487	0.480	0.800	0.585	0.517	0.581	0.483	0.892	0.528	0.836
	Dec. Rge	0.234	0.548	4.750	2.849	1.442	2.073	0.785	6.241	1.930	5.665
	Cov. Rate	0.000	0.324	0.977	0.827	0.612	0.956	0.462	0.923	0.911	0.919
$n = 500$											
	bias	0.498	0.405	0.115	0.254	0.342	0.252	0.358	0.112	0.236	0.123
	MAD	0.498	0.405	0.511	0.363	0.389	0.336	0.370	0.530	0.344	0.499
	Dec. Rge	0.097	0.427	2.912	1.372	0.914	1.147	0.676	3.235	1.228	2.673
	Cov. Rate	0.000	0.329	0.973	0.775	0.623	0.904	0.500	0.930	0.908	0.927
$c = 0.9$											
$n = 100$											
	bias	0.893	0.851	0.563	0.707	0.797	0.609	0.817	0.130	0.541	0.186
	MAD	0.893	0.851	0.794	0.781	0.819	0.701	0.817	0.854	0.744	0.810
	Dec. Rge	0.115	0.271	3.901	2.450	1.291	2.293	0.637	*	*	*
	Cov. Rate	0.000	0.000	0.813	0.569	0.331	0.776	0.114	0.759	0.616	0.754
$n = 500$											
	bias	0.890	0.747	0.166	0.381	0.630	0.255	0.553	-0.032	0.278	0.015
	MAD	0.890	0.747	0.460	0.509	0.651	0.364	0.554	0.381	0.397	0.376
	Dec. Rge	0.050	0.258	3.596	1.608	0.974	1.585	0.674	6.139	1.702	4.704
	Cov. Rate	0.000	0.001	0.878	0.657	0.319	0.782	0.312	0.905	0.770	0.903