arXiv: 0000.0000

Convex Combination of Ordinary Least Squares and Two-stage Least Squares Estimators

Cedric E. Ginestet[1*], **Richard Emsley**[2], **and Sabine Landau**[1]

[1] *Biostatistics Department, Institute of Psychiatry, Psychology and Neuroscience, King's College London*

[2] *Centre for Biostatistics, Institute of Population Health, University of Manchester*

Abstract: In the presence of confounders, the ordinary least squares (OLS) estimator is known to be biased. This problem can be remedied by using the two-stage least squares (TSLS) estimator, based on the availability of valid instrumental variables (IVs). This reduction in bias, however, is offset by an increase in variance. Under standard assumptions, the OLS has indeed a larger bias than the TSLS estimator; and moreover, one can prove that the sample variance of the OLS estimator is no greater than the one of the TSLS. Therefore, it is natural to ask whether one could combine the desirable properties of the OLS and TSLS estimators. Such a trade-off can be achieved through a convex combination of these two estimators, thereby producing our proposed convex least squares (CLS) estimator. The relative contribution of the OLS and TSLS estimators is here chosen to minimize a sample estimate of the mean squared error (MSE) of their convex combination. This proportion parameter is proved to be unique, whenever the OLS and TSLS differ in MSEs. Remarkably, we show that this proportion parameter can be estimated from the data, and that the resulting CLS estimator is consistent. We also show how the CLS framework can incorporate other asymptotically unbiased estimators, such as the jackknife IV estimator (JIVE). The finite-sample properties of the CLS estimator are investigated using Monte Carlo simulations, in which we independently vary the amount of confounding and the strength of the instrument. Overall, the CLS estimator is found to outperform the TSLS estimator in terms of MSE. The method is also applied to a classic data set from econometrics, which models the financial return to education.

AMS 2000 subject classifications: Convex combination, Instrumental variables, Ordinary least squares, Econometrics, Two-stage least squares.

1. Introduction

Instrumental variables (IVs) estimation is one of the cornerstones of modern econometric theory. The use of IVs has been described as "only second to ordinary least squares (OLS) in terms of methods used in empirical economic

*This work was supported by an MRC project grant MR/K006185/1, Landau et al. (2013-2016) entitled "Developing methods for understanding mechanism in complex interventions." We also would like to thank Stephen Burgess, Paul Clarke, Graham Dunn, Andrew Pickles, and Ian White for useful suggestions and discussions.

research" (Wooldridge, 2002, p.89). This ranking of estimation techniques naturally leads to the following methodological questions: When should we prefer IV estimation over OLS? Is it always preferable to use an instrument even though this may substantially increase the variance of the resulting estimator?

In fields including econometrics and the social sciences, and in some medical disciplines such as psychiatry, the direct randomized allocation of subjects to different experimental conditions is rarely possible, thereby preventing such scientists from inferring causal relations. Without adequate experimental manipulation, the model's predictors may be correlated with the errors. When this is case, we say that the predictors are *endogenous*. The absence of experimental manipulation in observational data, however, can be addressed by using IVs to predict the alleged causal variables. In particular, the resulting IV estimators allow to reduce the bias of the estimated effect. The main difficulty in conducting such IV analyses lies in the choice of appropriate *exogenous* instruments. Indeed, instruments are assumed to be solely correlated with the outcome variable through the predictor. This specific assumption is sometimes referred to as the *exclusion criterion*, since it disallows any direct effect of the instrument on the outcome.

The first published use of IVs is commonly attributed to Wright (1928) in the context of microeconometrics, albeit this has been historically disputed (see Stock and Trebbi, 2003). This estimation technique has been widely adopted in econometrics, and in other social sciences, including psychology, epidemiology, public health and political science. In particular, the use of IV methods has now become an integral part of causal inference (Pearl, 2009). The use of IVs in regression has been extended in several directions, allowing two-sample estimation, for instance (Inoue and Solon, 2010), and the selection of instruments using penalized methods such as the LASSO (Ng and Bai, 2009, Belloni et al., 2012). More recently, these methods have become especially popular in the study of genetic variants, thereby demonstrating the wide applicability of IV-based methods (Palmer et al., 2012, Pierce and Burgess, 2013). The reader may consult Wooldridge (2002) and Cameron and Trivedi (2005) for an introduction to the use of instrumental variables in the context of econometrics. A review of the assumptions underlying the use of IVs is provided by Angrist and Krueger (2001), and Heckman (1997); whereas the finite-sample properties of IV estimators have been described by Maddala and Jeong (1992) and Nelson and Startz (1990).

While the asymptotic properties of IV estimators such as the two-stage least squares (TSLS) are well-understood (Staiger and Stock, 1997, Hahn et al., 2004); in practice, it is not always clear whether or not using an IV estimator over a simpler OLS estimator is necessarily beneficial. Intuitively, since every IV is a random variable, its inclusion in the analysis tends to increase the variance of the resulting estimator. The magnitude of this increase in variance is proportional to the correlation of the instrument with the predictor. Poor or *weak* instruments are variables that are weakly correlated with the endogenous variables in the model. Thus, although the use of an IV estimator is likely to lead to a significant decrease in the bias of the OLS estimator, it will also yield a more variable

estimator. Since the true value of the parameters of interest is unknown in practice, it is generally not possible to evaluate whether the benefit of using a given set of instruments outweighs the cost in variance of incorporating them into the analysis. In addition, the use of weak instruments can also lead to a substantial amount of finite-sample bias. Indeed, the use of weak instruments has been studied by Bound et al. (1995), and these authors have shown that the inclusion of instruments with only weak linear relationships with the endogenous variables, tends to inflate the bias of the IV estimator; ultimately yielding an estimator as biased as the original OLS estimator.

In this paper, we address this issue by proposing a sample estimate of the mean squared error (MSE) of the estimators of interest. Since the MSE can be decomposed into a bias and a variance component, it provides us with a natural criterion for combining the OLS and TSLS estimators. Crucially, however, the proportion parameter weighting the relative contributions of the two candidate estimators is adaptive, in the sense that it depends on the properties of the data, and takes into account the strength of the instruments. The idea of combining the OLS and TSLS estimators has been previously discussed in the literature (Angrist et al., 1995). In particular, Sawa (1973) has proposed an "almost unbiased estimator" for simultaneous equations systems, which strikes a balance between two different k-class estimators by weighting their relative contributions using the sample size and the number of variables in the model. Moreover, Angrist et al. (1995) have given an interpretation of the limited information maximum likelihood (LIML) estimator as a combination estimator, which relies on a weighting of the OLS and TSLS estimators. Such combined estimators, however, do not attempt to estimate the respective contributions of each estimator using the data, as we have done in the paper at hand. The main contribution of this article is therefore to provide a framework for estimating such proportions in a data-informed adaptive manner.

The paper is organized as follows. In section 2, we fix the notation, and briefly recall the assumptions behind OLS and TSLS estimation. We then show that these two estimators have complementary properties, in the sense that the OLS has minimal variance, while the TSLS is asymptotically unbiased. In section 2.4, we describe our proposed convex estimator, and study its asymptotic properties, under the assumption that the optimal proportion is known; whereas in section 2.5, we describe a sample estimator of this proportion parameter. This framework is then extended to other asymptotically unbiased estimators in a third section. In section 4, these theoretical results are tested using a range of different synthetic data sets. The proposed methods are also applied to a classic data set from econometrics in section 5, and some conclusions are provided in section 6. Finally, the proofs of all the propositions in the paper are reported in the appendix.

2. Combining OLS and TSLS Estimators

2.1. Ordinary Least Squares (OLS)

The model under scrutiny is described by the following linear relationship,

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

where X is a random row vector of order $1 \times k$, and $\boldsymbol{\beta}$ is a column vector of order $k \times 1$ representing the parameters of interest, while Y and ε are two real-valued random variables. Throughout, we will treat both the error term, ε, and the vector of predictors, X, as random quantities, thereby allowing for possible correlations between the X_j's and ε. For expediency, all random variables, regardless of their dimensions, are simply denoted by upper-case Roman letters. In general, a sample of n draws will be available from the model in equation (1), such that

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \qquad \forall\, i = 1, \ldots, n;$$

where $\mathbf{x}_i$ is again a row vector of order $1 \times k$. This may also be written using matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are column vectors of order $n \times 1$, and $\mathbf{X}$ is a matrix of order $n \times k$.

The estimation of the unknown vector of parameters, $\boldsymbol{\beta}$, can be performed by making some standard assumptions about the moments of the different random variables in (1), as commonly done in econometrics (see Wooldridge, 2002):

(A1) *Exogeneity:* $\mathbb{E}[X'\varepsilon] = \mathbf{0}$,
(A2) *Homoscedastitity:* $\mathbb{E}[\varepsilon^2|X] = \sigma^2$,
(A3) *Identification:* $\text{rank}(\mathbb{E}[X'X]) = k$;

with $\sigma^2 := \mathbb{E}[\varepsilon^2]$, and where $\mathbb{E}[X'X]$ represents a matrix of order $k \times k$. Under assumptions (A2) and (A3), the OLS estimator behaves asymptotically as follows,

$$\widehat{\boldsymbol{\beta}}_n := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \xrightarrow{p} \mathbb{E}[X'X]^{-1}\mathbb{E}[X'Y] =: \widehat{\boldsymbol{\beta}}. \tag{2}$$

If assumption (A1) also holds, we say that the model in (1) is *exogenous*, and it then follows that the OLS estimator is asymptotically unbiased and consistent. That is, the limit, $\widehat{\boldsymbol{\beta}}$, can be shown to be equal to the true parameter, $\boldsymbol{\beta}$. However, if assumption (A1) is violated, then the OLS estimator is inconsistent. Thus, a model in which the vector of predictors has non-zero correlations with the error term, ε, is referred to as an *endogenous* model.

2.2. Two-stage Least Squares (TSLS)

The limitations of the OLS can be addressed by using a vector of IVs, denoted Z. We will here assume that Z is a random row vector of order $1 \times l$, with $l \geq k$.

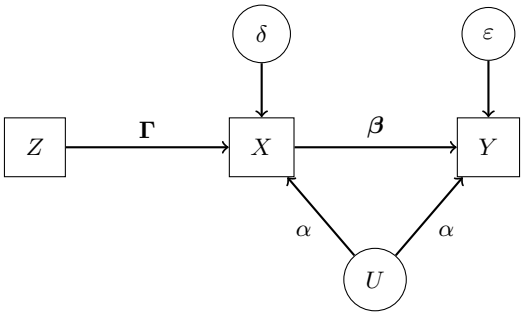

Fig 1. Graphical representation of the IV model described in equations (1) and (3) in the presence of an unmeasured confounder U; where observed and latent variables are denoted by squares and circles, respectively. This graph corresponds to a two-level system of equations composed of $Y = X\boldsymbol{\beta} + U\alpha + \varepsilon$, and $X = Z\boldsymbol{\Gamma} + U\alpha + \delta$. When we assume that $\alpha \neq 0$, condition (A1) is violated, and X becomes endogenous.

This vector of instruments is used in a multivariate linear equation of the form,

$$X = Z\boldsymbol{\Gamma} + \delta, \tag{3}$$

where $\boldsymbol{\Gamma}$ is an unknown matrix of parameters of order $l \times k$, and X and δ are random row vectors of order $1 \times k$. As before, we will usually work with a set of n realizations from this multivariate linear model expressed as follows,

$$\mathbf{x}_i = \mathbf{z}_i\boldsymbol{\Gamma} + \boldsymbol{\delta}_i, \qquad \forall\, i = 1, \ldots, n; \tag{4}$$

where $\mathbf{x}_i$ and $\boldsymbol{\delta}_i$ are $1 \times k$ row vectors, and $\mathbf{z}_i$ is an $1 \times l$ row vector. This can be concisely expressed using matrix notation as

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{D},$$

where $\mathbf{Z}$ and $\mathbf{D}$ are matrices of order $n \times l$ and $n \times k$, respectively. A graphical illustration of the IV model is provided in figure 1.

When using the two-stage least squares (TSLS) estimator, we will make the following additional assumptions about the random row vector of instruments:

(A4) *Exogeneity:* $\mathbb{E}[Z'\varepsilon] = \mathbf{0}$,
(A5) *Homoscedastitity:* $\mathbb{E}[\varepsilon^2|Z] = \sigma^2$,
(A6) *Identification:* $\text{rank}(\mathbb{E}[Z'Z]) = l$, $\text{rank}(\mathbb{E}[Z'X]) = k$;

where, as before, $\sigma^2 := \mathbb{E}[\varepsilon^2]$. The TSLS estimator is then defined as

$$\widetilde{\boldsymbol{\beta}}_n := (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y},$$

with $\widehat{\mathbf{X}} := \mathbf{H}_z\mathbf{X}$ denoting the projection of the matrix of predictors onto the column space of $\mathbf{Z}$, and where $\mathbf{H}_z := \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is the hat matrix of the multivariate regression in equation (4). Under assumptions (A5) and (A6), the

TSLS estimator converges in probability to a non-stochastic vector, $\widetilde{\boldsymbol{\beta}}$, defined as

$$\widetilde{\boldsymbol{\beta}} := \big(\mathbb{E}[X'Z]\mathbb{E}[Z'Z]^{-1}\mathbb{E}[Z'X]\big)^{-1}\big(\mathbb{E}[X'Z]\mathbb{E}[Z'Z]^{-1}\mathbb{E}[ZY]\big), \tag{5}$$

such that $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{p} \widetilde{\boldsymbol{\beta}}$, as described in Wooldridge (2002). Moreover, under assumption (A4), this sequence of estimators can be shown to be asymptotically unbiased and consistent with respect to the true vector of parameters, such that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$. However, this gain in unbiasedness is compensated by a larger variance of the TSLS estimator, as we discuss in the next section.

2.3. Bias/Variance Trade-off

Under assumptions (A2-A6), the TSLS estimator is asymptotically unbiased. By contrast, if assumption (A1) does not hold, then the OLS estimator is asymptotically biased. However, for finite n, the empirical variance of the TSLS estimator can be shown to be larger than the one of the OLS estimator. We make these observations formal by comparing the variance estimators of the OLS and TSLS estimators. These are

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\beta}}_n) := \widehat{\sigma}_n^2(\mathbf{X}'\mathbf{X})^{-1}, \quad \text{and} \quad \widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\boldsymbol{\beta}}_n) := \widetilde{\sigma}_n^2(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}; \tag{6}$$

with the sample residual sums of squares (RSSs), $\widehat{\sigma}_n^2$ and $\widetilde{\sigma}_n^2$, being given by

$$\widehat{\sigma}_n^2 := \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}_n)^2, \quad \text{and} \quad \widetilde{\sigma}_n^2 := \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\widetilde{\boldsymbol{\beta}}_n)^2; \tag{7}$$

for the OLS and TSLS estimators, respectively.

More remarkably, one can also approximate the bias of these two estimators. The theoretical squared bias of a given arbitrary estimator, $\boldsymbol{\beta}_n^\dagger$, is defined as

$$\mathbb{B}\mathrm{ias}^2(\boldsymbol{\beta}_n^\dagger) := (\mathbb{E}[\boldsymbol{\beta}_n^\dagger] - \boldsymbol{\beta})(\mathbb{E}[\boldsymbol{\beta}_n^\dagger] - \boldsymbol{\beta})',$$

for every n. In the sequel, we will assume that the IVs under scrutiny are valid instruments, such that assumption (A4) is true. Therefore, it follows that the TSLS estimator, $\widetilde{\boldsymbol{\beta}}_n$, is known to be consistent, and can be used to construct a consistent approximation of the bias of any arbitrary estimator, $\boldsymbol{\beta}_n^\dagger$. For large n, it follows that the squared bias of any such estimator can be consistently estimated by

$$\widehat{\mathbb{B}\mathrm{ias}}^2(\boldsymbol{\beta}_n^\dagger) := (\boldsymbol{\beta}_n^\dagger - \widetilde{\boldsymbol{\beta}}_n)(\boldsymbol{\beta}_n^\dagger - \widetilde{\boldsymbol{\beta}}_n)'. \tag{8}$$

Observe that this empirical estimate of the bias gives a value of zero for the TSLS estimator, for every n. This particular choice of empirical bias estimate can also be seen to be related to the Hausman test, commonly used in econometrics for testing whether or not the predictors of interest are exogenous (Hausman, 1978). Indeed, the squared bias in equation (8) corresponds to the numerator of the Hausman test statistic.

Combining this empirical estimate of the bias with the standard variance estimators in equation (6), we can formalize our original observation about the trade-off between the superiority of the TSLS estimator in terms of bias, and the superiority of the OLS estimator in terms of variance. This result will motivate our construction of a combined estimator, in which we will exploit the respective strengths of the OLS and TSLS estimators, denoted by $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$, respectively.

Proposition 1. *Under assumptions (A2-A6), for every n, and for every realizations, $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$, if both $\mathbf{X}'\mathbf{X}$ and $\widehat{\mathbf{X}}'\widehat{\mathbf{X}}$ are invertible, then*

(i) $\widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n) \succeq \widehat{\mathbb{B}\mathrm{ias}}^2(\widetilde{\boldsymbol{\beta}}_n)$,
(ii) $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\beta}}_n) \preceq \widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\boldsymbol{\beta}}_n)$;

where $\succeq$ and $\preceq$ denote the positive semidefinite order for $k \times k$ matrices.

Note that, in proposition 1, we have requested both $\mathbf{X}'\mathbf{X}$ and $\widehat{\mathbf{X}}'\widehat{\mathbf{X}}$ to be invertible. Indeed, while assumptions (A3) and (A6) ensures that the stochastic limits of these two matrices are invertible, this does not guarantee that these matrices will be invertible for every n. Although these inequalities appear to be well-known, they do not appear to have been formally proved in standard texts on instrumental variables (see, for instance Wooldridge, 2002, Davidson and MacKinnon, 1993, Cameron and Trivedi, 2005). A full proof of this result is therefore provided in the appendix.

Furthermore, the two statements in proposition 1 can also be shown to hold in the stochastic limit, as described in the following corollary. Note that this corollary is trivially true for the variances of the OLS and TSLS estimators, since both of these quantities converge to a zero matrix. A proof of this result is provided in the appendix.

Corollary 1. *Under assumptions (A2-A6),*

(i) $\operatorname{plim}_n \widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n) \succeq \operatorname{plim}_n \widehat{\mathbb{B}\mathrm{ias}}^2(\widetilde{\boldsymbol{\beta}}_n)$,
(ii) $\operatorname{plim}_n \widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\beta}}_n) \preceq \operatorname{plim}_n \widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\boldsymbol{\beta}}_n)$;

where, as before, $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ denote the OLS and TSLS estimators, respectively.

The inequalities in proposition 1 indicate that it may be fruitful to compare the MSEs of these two estimators for finite n. Clearly, since the bias tends to dominate the MSE asymptotically, it follows that the TSLS should exhibit a smaller level of bias as n goes to infinity. Nonetheless, for finite samples, the OLS may yield a smaller MSE than its two-stage counterpart, due to its greater efficiency. Therefore, one may try to strike a balance between the relative strengths of these two types of estimators, using the sample MSE as a criterion.

2.4. Convex Least Squares (CLS)

In this section and in the rest of this paper, we now assume that (A2-A6) hold. In addition, we also assume that the random vectors, $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$, are well-behaved, in the sense that they are elementwise squared-integrable for every

n. Under these assumptions, we propose an estimator, denoted $\bar{\boldsymbol{\beta}}_n(\pi)$, which is defined as a convex combination of the OLS and TSLS estimators, such that

$$\bar{\boldsymbol{\beta}}_n(\pi) := \pi\widehat{\boldsymbol{\beta}}_n + (1-\pi)\widetilde{\boldsymbol{\beta}}_n, \tag{9}$$

for every $\pi \in [0,1]$. The *proportion parameter*, π, controls the respective contributions of the OLS and TSLS estimators. This parameter is selected in order to minimize the trace of the theoretical MSE of the corresponding CLS estimator,

$$\operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) = \mathbb{E}\big[(\bar{\boldsymbol{\beta}}_n(\pi) - \boldsymbol{\beta})(\bar{\boldsymbol{\beta}}_n(\pi) - \boldsymbol{\beta})'\big],$$

where $\boldsymbol{\beta} \in \mathbb{R}^k$ is the true parameter of interest and the MSE is a $k \times k$ matrix.

The MSE automatically strikes a trade-off between the unbiasedness of the TSLS estimator and the efficiency of the OLS estimator. Indeed, this criterion can be decomposed into a variance and a bias component, such that

$$\operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) = \mathbb{V}\mathrm{ar}(\bar{\boldsymbol{\beta}}_n(\pi)) + \mathbb{B}\mathrm{ias}^2(\bar{\boldsymbol{\beta}}_n(\pi)).$$

Therefore, in the light of proposition 1, this criterion constitutes a natural choice for combining these two types of estimators.

The MSE of the CLS estimator, $\operatorname{MSE}(\pi\widehat{\boldsymbol{\beta}}_n + (1-\pi)\widetilde{\boldsymbol{\beta}}_n)$, can be expressed as the weighted sum of the MSEs of the OLS and TSLS estimators, as well as a *cross-squared-error* (CSE) term between these two estimators,

$$\pi^2 \operatorname{MSE}(\widehat{\boldsymbol{\beta}}_n) + 2\pi(1-\pi)\operatorname{CSE}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) + (1-\pi)^2 \operatorname{MSE}(\widetilde{\boldsymbol{\beta}}_n), \tag{10}$$

where the cross-term is defined as follows,

$$\operatorname{CSE}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) := \mathbb{E}[(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'].$$

By analogy with the MSE, we can also decompose the CSE into a covariance term and a squared *cross-bias* term, denoted $\mathbb{B}\mathrm{ias}^2(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n)$, such that

$$\operatorname{CSE}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) + \mathbb{B}\mathrm{ias}^2(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n),$$

where the squared cross-bias term is $\mathbb{B}\mathrm{ias}^2(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) := (\mathbb{E}[\widehat{\boldsymbol{\beta}}_n] - \boldsymbol{\beta})(\mathbb{E}[\widetilde{\boldsymbol{\beta}}_n] - \boldsymbol{\beta})'$.

The true (or theoretical) proportion parameter, π, is defined as the value that minimizes the trace of the theoretical MSE of the CLS estimator. Note that we are here considering a sequence of parameters, π_n, since this definition may yield a different proportion for different sample sizes. Therefore, for every n, the target proportion parameter is given by

$$\pi_n := \operatorname*{argmin}_{\pi \in [0,1]} \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)). \tag{11}$$

Crucially, this parameter is available in closed-form, and it can also be shown to be unique, since the trace of the theoretical MSE of $\bar{\boldsymbol{\beta}}_n$ is a convex function of π. This statement is made formal in the following proposition, which is proved using the aforementioned decomposition of the MSE of the CLS estimator. The proportion parameter is only non-unique when the square-root of the trace of the MSEs the OLS and TSLS estimators are identical. This quantity, denoted by $(\operatorname{tr} \operatorname{MSE}(\boldsymbol{\beta}_n^\dagger))^{1/2}$ for every estimator $\boldsymbol{\beta}_n^\dagger$, will be referred to as the RMSE of $\boldsymbol{\beta}_n^\dagger$, in the sequel. See appendix A for a proof of this minimization.

Proposition 2. *For every n, the proportion parameter defined in equation (11) is given by*

$$\pi_n = \frac{\operatorname{tr}(\operatorname{MSE}(\widetilde{\boldsymbol{\beta}}_n) - \operatorname{CSE}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n))}{\operatorname{tr}(\operatorname{MSE}(\widetilde{\boldsymbol{\beta}}_n) - 2\operatorname{CSE}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) + \operatorname{MSE}(\widehat{\boldsymbol{\beta}}_n))}.$$

It is unique whenever the RMSEs of the OLS and TSLS estimators are not equal.

Finally, we can verify that the CLS estimator based on the true proportion π_n has an MSE, which is lower or equal to the MSEs of the OLS and TSLS estimators. Note that this inequality is not immediate from our definition of π_n, as we need to control for the additional CSE term in equation 10. A proof of this proposition is also provided in the appendix.

Proposition 3. *The CLS estimator based on the true proportion, π_n, satisfies*

$$\operatorname{tr}\operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi_n)) \leq \operatorname{tr}\min\{\operatorname{MSE}(\widehat{\boldsymbol{\beta}}_n), \operatorname{MSE}(\widetilde{\boldsymbol{\beta}}_n)\},$$

for every n.

Observe that this result holds in greater generality, since the OLS and TSLS estimators could be replaced by other candidate estimators. In the next section, we describe how to estimate the proportion parameter in an adaptive manner for this particular choice of estimators; whereas in section 3, we consider how the CLS can accommodate other estimators.

2.5. CLS Estimation

When evaluating π_n from a particular data set, we estimate this parameter by minimizing the trace of an empirical estimate of the theoretical MSE of the CLS estimator. A consistent estimator of the MSE can be obtained by setting the true parameter, $\boldsymbol{\beta}$, to be equal to the TSLS estimator, $\widetilde{\boldsymbol{\beta}}_n$. Thus, for every $\pi \in [0, 1]$, our proposed empirical MSE is given by

$$\widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(\pi)) = \widehat{\mathbb{V}\operatorname{ar}}(\bar{\boldsymbol{\beta}}_n(\pi)) + \widehat{\mathbb{B}\operatorname{ias}}^2(\bar{\boldsymbol{\beta}}_n(\pi)), \tag{12}$$

where $\widehat{\mathbb{B}\operatorname{ias}}(\bar{\boldsymbol{\beta}}_n(\pi)) := \bar{\boldsymbol{\beta}}_n(\pi) - \widetilde{\boldsymbol{\beta}}_n$. That is, we here use the TSLS estimator as a consistent estimator of the true parameter, $\boldsymbol{\beta}$. To approximate the population variance of the CLS estimator, we can use a combination of the empirical estimates of the variances of the two estimators of interest, such that

$$\widehat{\mathbb{V}\operatorname{ar}}(\bar{\boldsymbol{\beta}}_n(\pi)) = \pi^2\widehat{\mathbb{V}\operatorname{ar}}(\widehat{\boldsymbol{\beta}}_n) + 2\pi(1-\pi)\widehat{\mathbb{C}\operatorname{ov}}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) + (1-\pi)^2\widehat{\mathbb{V}\operatorname{ar}}(\widetilde{\boldsymbol{\beta}}_n), \tag{13}$$

where the empirical variances of the OLS and TSLS estimators have already be given in equation (6); and where the covariance term takes the following form,

$$\widehat{\mathbb{C}\operatorname{ov}}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) := \bar{\sigma}_n^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\widehat{\mathbf{X}})(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1} = \bar{\sigma}_n^2(\mathbf{X}'\mathbf{X})^{-1};$$

with as before, $\widehat{\mathbf{X}} := \mathbf{H}_z\mathbf{X}$, and in which the second equality is obtained by using the idempotency of $\mathbf{H}_z$. Moreover, the cross-RSS, denoted $\bar{\sigma}_n^2$, is given by

$$\bar{\sigma}_n^2 := \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}_n)(y_i - \mathbf{x}_i\widetilde{\boldsymbol{\beta}}_n),$$

which can be compared to the RSSs of the OLS and TSLS estimators in equation (7).

The second term in equation (12) consists of the empirical bias of the CLS estimator. As for the empirical variance, the bias can be estimated by using the TSLS estimator to replace the unknown true parameter, such that

$$\widehat{\mathbb{B}\mathrm{ias}}^2(\bar{\boldsymbol{\beta}}_n(\pi)) = \pi^2\widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n) + 2\pi(1-\pi)\widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) + (1-\pi)^2\widehat{\mathbb{B}\mathrm{ias}}^2(\widetilde{\boldsymbol{\beta}}_n),$$

where the empirical biases of the OLS and TSLS estimators are estimated as in equation (12). Since we have set the bias of $\widetilde{\boldsymbol{\beta}}_n$ to zero, it follows that the cross-bias term also eliminates, Thus, the empirical bias of the CLS estimator becomes proportional to the one of the OLS estimator, such that we obtain

$$\widehat{\mathbb{B}\mathrm{ias}}^2(\bar{\boldsymbol{\beta}}_n(\pi)) = \pi^2\widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n). \tag{14}$$

The empirical estimate of the MSE in equation (12) can be shown to be consistent, as described in the following proposition, which is proved in the appendix. Observe that this statement holds for every arbitrary proportion comprised between 0 and 1.

Proposition 4. *For every π, $\bar{\boldsymbol{\beta}}_n(\pi) \xrightarrow{p} \bar{\boldsymbol{\beta}}(\pi) := \pi\widehat{\boldsymbol{\beta}} + (1-\pi)\widetilde{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are defined as in equations (2) and (5), respectively. Moreover,*

$$\widehat{\mathrm{MSE}}(\bar{\boldsymbol{\beta}}_n(\pi)) \xrightarrow{p} \mathrm{MSE}(\bar{\boldsymbol{\beta}}(\pi)).$$

As for the true proportion parameter, π_n, which minimizes the trace of the theoretical MSE, the proportion estimator, $\widehat{\pi}_n$, which minimizes the trace of the empirical MSE; is also available in closed-form. We thus obtain the following result, as a corollary to proposition 2. Observe that, since the bias of the TSLS estimator is zero under our estimation framework, it follows that the MSE of the TSLS reduces to the variance of that estimator, and that the CSE term reduces to the covariance of the two estimators of interest.

Corollary 2. *The estimator of the proportion parameter, $\pi \in [0,1]$, defined as $\widehat{\pi}_n := \operatorname{argmin} \operatorname{tr} \widehat{\mathrm{MSE}}(\bar{\boldsymbol{\beta}}_n(\pi))$, in which the $\widehat{\mathrm{MSE}}$ is defined as in equation (12); satisfies,*

$$\widehat{\pi}_n = \frac{\operatorname{tr}(\widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\boldsymbol{\beta}}_n) - \widehat{\mathbb{C}\mathrm{ov}}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n))}{\operatorname{tr}(\widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\boldsymbol{\beta}}_n) - 2\widehat{\mathbb{C}\mathrm{ov}}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n) + \widehat{\mathrm{MSE}}(\widehat{\boldsymbol{\beta}}_n))}.$$

We have here emphasized the estimation of the proportion parameter. Our original motivation, however, for constructing the CLS estimator centered on producing an estimator, which would minimize the MSE. This can be achieved

by estimating the CLS estimator, $\bar{\boldsymbol{\beta}}_n(\pi)$, at the value of the estimated proportion, $\widehat{\pi}_n$, thereby producing $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)$. We thus conclude this section by verifying that this particular CLS estimator behaves as expected asymptotically, in the sense that it is both weakly and MSE consistent.

Proposition 5. *Under assumptions (A2-A6), the CLS estimator, $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)$, satisfies (i) $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n) \xrightarrow{p} \boldsymbol{\beta}$, and (ii) $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n) \xrightarrow{L^2} \boldsymbol{\beta}$.*

The proof of this proposition follows from the inequalities reported in proposition 3, combined with the fact that the TSLS estimator is both weakly and MSE consistent. See appendix A for details. Observe that the CLS framework relies on the existence of the first two moments of the TSLS estimator. For finite n, Kinal (1980) has shown that the TSLS estimator only possesses first and second moments when $l \geq k + 2$. Asymptotically, however, such moments always exist. As for the TSLS therefore, we are thus considering an estimator, which is solely asymptotically well-identified. This particular issue is further discussed in section 6.

2.6. Bootstrap CLS Variance

We now turn to the question of estimating the variance of our proposed CLS estimator. In equation (13), we have described the variance of $\bar{\boldsymbol{\beta}}_n(\pi)$, for every π. This quantity was then used in our proposed empirical MSE, in order to obtain a sample estimate of π. However, the variance formula in equation (13) does not take into account the variability associated with the choice of π_n. The derivation of a closed-form estimator for the variance of $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)$ is beyond the scope of this paper. However, in practice, the variance of the CLS estimator can be computed using the bootstrap by sampling with replacement from the triple $(\mathbf{y}, \mathbf{X}, \mathbf{Z})$, and producing B bootstrap samples denoted $(\mathbf{y}_b^*, \mathbf{X}_b^*, \mathbf{Z}_b^*)$, with $b = 1, \ldots, B$. We are here adopting the framework described by previous researchers, who have also used the bootstrap in the context of IV estimation (see for example Wong, 1996).

Specifically, each bootstrap sample is constructed by sampling n cases with replacement from the collection of triples $(y_i, \mathbf{x}_i, \mathbf{z}_i)$, with $i = 1, \ldots, n$. These bootstrap samples are then used to produce the bootstrap distribution of the CLS estimator. That is, for each of these bootstrap samples, we compute the CLS estimator, $\bar{\boldsymbol{\beta}}_{nb} := \bar{\boldsymbol{\beta}}_{nb}(\widehat{\pi}_{nb})$, which leads to the following bootstrap variance estimator,

$$\widehat{\mathbb{V}\mathrm{ar}}^*(\bar{\boldsymbol{\beta}}_n) := \frac{1}{B-1} \sum_{b=1}^{B} (\bar{\boldsymbol{\beta}}_{nb}^* - \mathbb{E}^*[\bar{\boldsymbol{\beta}}_n])(\bar{\boldsymbol{\beta}}_{nb}^* - \mathbb{E}^*[\bar{\boldsymbol{\beta}}_n])',$$

where $\mathbb{E}^*[\bar{\boldsymbol{\beta}}_n] := \sum_b \bar{\boldsymbol{\beta}}_{nb}^*/B$ denotes the bootstrap mean. In our real-world data set application in section 5, we will report the variance of the CLS and its confidence interval using the bootstrap.

One of the limitations of our discussion thus far is the presence of a finite-sample bias in the TSLS estimator. In the next section, we consider other consistent estimators, which could be articulated within our framework by being

substituted to the TSLS estimator. Indeed, every asymptotically unbiased estimator could be used to replace the TSLS estimator in the previous results.

3. Extensions to Other Unbiased Estimators

Our proposed convex combination of least squares estimators essentially relies on the choice of an asymptotically unbiased estimator. Under standard assumptions on the properties of the IVs under scrutiny, the TSLS estimator satisfies this criterion. This choice was mainly motivated by computational considerations. The empirical variance for the TSLS estimator is indeed well-known and can easily be manipulated. We now extend the CLS framework, in order to accommodate other asymptotically unbiased estimators. The corresponding MSE can be empirically estimated using the bootstrap at a greater computational cost, but without additional theoretical complications. The resulting estimator will thus be referred to as the bootstrap CLS.

3.1. Jackknife IV Estimator

An ideal replacement for the TSLS estimator is the jackknife IV estimator (JIVE), which we now describe. This estimator was originally introduced by Angrist et al. (1995) in order to reduce the finite-sample bias of the TSLS estimator, when applied to a large number of instruments. Indeed, the TSLS estimator tends to behave poorly as the number of instruments increases. We briefly outline this method in the present section. See Angrist et al. (1999) for an exhaustive description. Let the estimator of the regression parameter in the first-level equation in model (3) be denoted by

$$\widehat{\mathbf{\Gamma}} := (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}),$$

which is of order $l \times k$. The matrix of predictors, $\mathbf{X}$, projected onto the column space of the instruments is then given by $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\mathbf{\Gamma}}$. The jackknife IV estimator (JIVE) proceeds by estimating each row of $\widehat{\mathbf{X}}$ without using the corresponding data point. That is, the i^{th} row in the jackknife matrix, $\widehat{\mathbf{X}}_J$, is estimated without using the i^{th} row of $\mathbf{X}$.

This is conducted as follows. For every $i = 1, \ldots, n$, we first compute

$$\widehat{\mathbf{\Gamma}}_{(i)} := (\mathbf{Z}'_{(i)}\mathbf{Z}_{(i)})^{-1}(\mathbf{Z}'_{(i)}\mathbf{X}_{(i)}),$$

where $\mathbf{Z}_{(i)}$ and $\mathbf{X}_{(i)}$ denote matrices $\mathbf{Z}$ and $\mathbf{X}$ after removal of the i^{th} row, such that these two matrices are of order $(n-1) \times l$ and $(n-1) \times k$, respectively. Then, the matrix $\widehat{\mathbf{X}}_J$ is constructed by stacking these jackknife estimates of $\widehat{\mathbf{\Gamma}}$, after they have been pre-multiplied by the corresponding rows of $\mathbf{Z}$,

$$\widehat{\mathbf{X}}_J := \begin{vmatrix} \mathbf{z}_1\widehat{\mathbf{\Gamma}}_{(1)} \\ \vdots \\ \mathbf{z}_n\widehat{\mathbf{\Gamma}}_{(n)} \end{vmatrix},$$

where each $\mathbf{z}_i$ is an l-dimensional row vector. The JIVE estimator is then obtained by replacing $\widehat{\mathbf{X}}$ with $\widehat{\mathbf{X}}_J$ in the standard formula of the TSLS, such that

$$\widetilde{\boldsymbol{\beta}}_J := (\widehat{\mathbf{X}}_J{}'\mathbf{X})^{-1}(\widehat{\mathbf{X}}_J{}'\mathbf{y}).$$

In this paper, we have additionally made use of the computational formula suggested by Angrist et al. (1999), in which each row of $\widehat{\mathbf{X}}_J$ is calculated using

$$\mathbf{z}_i\widehat{\boldsymbol{\Gamma}}_{(i)} = \frac{\mathbf{z}_i\widehat{\boldsymbol{\Gamma}} - h_i\mathbf{x}_i}{1-h_i},$$

where $\mathbf{z}_i\widehat{\boldsymbol{\Gamma}}_{(i)}$, $\mathbf{z}_i\widehat{\boldsymbol{\Gamma}}$ and $\mathbf{x}_i$ are k-dimensional row vectors; and with h_i denoting the leverage of the corresponding data point in the first-level equation of our model, such that each h_i is defined as $\mathbf{z}_i(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_i'$.

3.2. Bootstrap CLS Estimation

When replacing the TSLS estimator with an arbitrary estimator, such as the JIVE, some of the quantities required for estimating the proportion, π_n, need not be available in closed-form. However, such quantities can be straightforwardly estimated using the bootstrap, as was done for the variance of the CLS estimator in section 2.6.

We can indeed approximate the unknown joint distribution, $F(Y, X, Z)$, with its bootstrap estimate, F^*, using the straightforward sampling scheme described in section 2.6. As before, we thus generate B bootstrap samples, denoted $(\mathbf{y}_b^*, \mathbf{X}_b^*, \mathbf{Z}_b^*)$, from F^*. These bootstrap samples are then used to produce the bootstrap distributions of the OLS estimator and the unbiased estimator of interest such as the JIVE; and the first and second moments of these estimators are computed. Thus, for every unbiased estimator, $\boldsymbol{\beta}_n^\dagger$, and given the OLS estimator, $\widehat{\boldsymbol{\beta}}_n$, we construct a bootstrap estimate of the MSE of the corresponding CLS estimator, such that for every π, we define

$$\widehat{\mathrm{MSE}}^*(\bar{\boldsymbol{\beta}}_n(\pi)) := \mathbb{E}^*\Big[(\bar{\boldsymbol{\beta}}_n(\pi) - \mathbb{E}^*[\boldsymbol{\beta}_n^\dagger])(\bar{\boldsymbol{\beta}}_n(\pi) - \mathbb{E}^*[\boldsymbol{\beta}_n^\dagger])'\Big].$$

As in section 2.6, the operator, $\mathbb{E}^*$, denotes the expectation over the bootstrap estimate of F. Similarly to the MSE decomposition in equation (10), the bootstrap estimate of the MSE can be decomposed into the following components,

$$\pi^2\widehat{\mathrm{MSE}}^*(\widehat{\boldsymbol{\beta}}_n) + 2\pi(1-\pi)\widehat{\mathrm{CSE}}^*(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_n^\dagger) + (1-\pi)^2\widehat{\mathbb{V}\mathrm{ar}}^*(\boldsymbol{\beta}_n^\dagger),$$

where the bootstrap estimate of the MSE of $\boldsymbol{\beta}_n^\dagger$ was reduced to $\widehat{\mathbb{V}\mathrm{ar}}^*(\boldsymbol{\beta}_n^\dagger)$, since the estimator, $\boldsymbol{\beta}_n^\dagger$, is assumed to be unbiased for every n. Moreover, as in equation (14), the bootstrap estimate of the bias of the CLS estimator is proportional to the bootstrap bias of the OLS estimator, such that we have

$$\widehat{\mathbb{B}\mathrm{ias}}^*(\bar{\boldsymbol{\beta}}_n(\pi)) = \pi^2\widehat{\mathbb{B}\mathrm{ias}}^*(\widehat{\boldsymbol{\beta}}_n).$$

The boostrap estimate of the proportion, denoted $\widehat{\pi}_n^*$, is then given by a formula analogous to the one described in corollary 2, in which each empirical moment is replaced by its bootstrap equivalent. This allows us to show that, for every choice of asymptotically unbiased estimator, $\boldsymbol{\beta}_n^\dagger$, the resulting bootstrap CLS estimator, $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n^*)$, achieves minimal bootstrap MSE amongst its constituent estimators. A proof of this corollary is provided in the appendix. It relies on the same arguments employed in the proof of the optimality of the CLS estimator in proposition 3.

Corollary 3. *For every asymptotically unbiased estimator $\boldsymbol{\beta}_n^\dagger$, the bootstrap CLS estimator of $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_n^\dagger$, based on the bootstrap proportion, $\widehat{\pi}_n^*$, satisfies*

$$\operatorname{tr} \widehat{\operatorname{MSE}}^*(\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n^*)) \leq \operatorname{tr} \min\{\widehat{\operatorname{MSE}}^*(\widehat{\boldsymbol{\beta}}_n), \widehat{\operatorname{MSE}}^*(\boldsymbol{\beta}_n^\dagger)\},$$

for every n.

4. Data Simulations

We here produce synthetic data sets with different number of instruments. All of the models considered in this section are based on a univariate endogenous variable, X, without any additional covariate in the second-level equation. In Model I, we describe a simple Gaussian model with a single valid instrument; whereas in Model II, we consider a similar statistical model comprising $l = 10$ uncorrelated instruments.

4.1. Simulation Model I

Synthetic data sets were created from the following two-level model. We are here focusing on a univariate model composed of a single predictor, X, and a single instrument, Z. For every $i = 1, \ldots, n$, the two levels of the model are

$$\begin{aligned} y_i &= x_i\beta + u_i\alpha + \varepsilon_i, \\ x_i &= z_i\gamma + u_i\alpha + \delta_i; \end{aligned} \tag{15}$$

where α controls the degree of endogeneity of X, and γ controls the amount of covariance between X and the instrument Z, such that γ can be interpreted as the strength of the instrument. We wish to keep the marginal variances of the Y_i's and X_i's constant, while varying the values of α and γ. This is achieved by defining the variances of the error terms, ε_i and δ_i, as functions of α and γ. In doing so, we simplify the interpretation of β, which becomes a standardized regression coefficient, whenever $\gamma = 0$. Throughout these simulations, the true parameter of interest will be set to be $\beta = 1/2$. A graphical representation of this model has been given in figure 1.

The model is thus standardized by setting the marginal variances of the Y_i's and X_i's to one, such that $\mathbb{V}\mathrm{ar}(Y_i) = \mathbb{V}\mathrm{ar}(X_i) = 1$; and by generating the Z_i's and U_i's from a standard normal distribution, such that

$$Z_i, U_i \overset{\text{iid}}{\sim} N(0,1), \qquad \forall\ i = 1, \ldots, n.$$

The marginal variances of these random variables will be denoted by $\sigma_z^2 := \mathbb{V}\text{ar}[Z_i]$ and $\sigma_u^2 := \mathbb{V}\text{ar}[U_i]$, respectively. The two remaining variances can then be defined as functions of the different regression parameters. For the second-level equation, we have

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2(\alpha)), \qquad \sigma_\varepsilon^2(\alpha) := 3/4 - 2\alpha^2, \tag{16}$$

which follows from the constraint $\mathbb{V}\text{ar}[Y_i] = 1$, and from the decomposition,

$$\mathbb{V}\text{ar}(Y_i) = \beta^2\,\mathbb{V}\text{ar}(X_i) + \alpha^2\,\mathbb{V}\text{ar}(U_i) + 2\beta\alpha\,\mathbb{C}\text{ov}(X_i, U_i) + \mathbb{V}\text{ar}(\varepsilon_i).$$

Using the linear independence of Z and U, the covariance term becomes $\mathbb{C}\text{ov}(X_i, U_i) = \alpha\sigma_u^2$. Moreover, from our choice of variances for X_i and U_i, we also obtain $\mathbb{V}\text{ar}(Y_i) = \beta^2 + \alpha^2 + 2\beta\alpha^2 + \sigma_\varepsilon^2$. Fixing the variance of Y_i to unity and using our choice of β, this yields the definition of $\sigma_\varepsilon^2(\alpha)$ given in equation (16). Moreover, observe that the positiveness of σ_ε^2 produces an upper bound for α, which is given by $\alpha < \sqrt{3/8}$.

Similarly, we can ensure that the marginal variances of the X_i's are also constant, irrespective of the choice of α and γ, by controlling the variances of the δ_i's. Thus, we set

$$\delta_i \overset{\text{iid}}{\sim} N(0, \sigma_\delta^2(\alpha, \gamma)), \qquad \sigma_\delta^2(\alpha, \gamma) := 1 - (\gamma^2 + \alpha^2).$$

This specification ensures that the variances of the X_i's are constant with $\sigma_x^2 = 1$. That is, since by assumption $\mathbb{C}\text{ov}(Z_i, U_i) = 0$, and using the fact that for uncorrelated variables, the Bienaymé formula states that $\mathbb{V}\text{ar}(\sum_j X_j) = \sum_j \mathbb{V}\text{ar}(X_j)$; it then follows that for every $i = 1, \ldots, n$, we obtain

$$\mathbb{V}\text{ar}(X_i) = \gamma^2\,\mathbb{V}\text{ar}(Z_i) + \alpha^2\,\mathbb{V}\text{ar}(U_i) + \mathbb{V}\text{ar}(\delta_i),$$

which gives, $\mathbb{V}\text{ar}(\delta_i) = 1 - \gamma^2\sigma_z^2 - \alpha^2\sigma_u^2$, as required. Moreover, note that we must have $\gamma < \sqrt{1 - \alpha^2}$ in order to ensure that $\sigma_\delta^2 > 0$. Using our previous bound for α, which states that $\alpha < \sqrt{3/8}$, it then follows that $\gamma < \sqrt{5/8}$.

Altogether, we have therefore fixed the variances of the Y_i's, X_i's, U_i's, and Z_i's to unity; and by assumption, the instrument is deemed valid in the sense that $\mathbb{C}\text{ov}(Z_i, U_i) = 0$. From these standardizations, it follows that for every $\alpha \in [0, \sqrt{3/8})$, and for every $\gamma \in (0, \sqrt{1 - \alpha^2})$, the correlations of the X_i's with the U_i's and Z_i's are controlled by the two simulation parameters, α and γ:

$$\text{Cor}(X_i, U_i) = \alpha, \qquad \text{and} \qquad \text{Cor}(X_i, Z_i) = \gamma,$$

which respectively represent the *magnitude of the confounding* and the *strength of the instrument*. In addition, the correlations of the Y_i's with the U_i's and the Z_i's are also controlled by a combination of these parameters. These correlations are respectively given by $\text{Cor}(Y_i, U_i) = \beta\alpha + \alpha$, and $\text{Cor}(Y_i, Z_i) = \beta\gamma$. Finally, the correlation between the outcome and the endogenous variable satisfies

$$\text{Cor}(Y_i, X_i) = \beta + \alpha^2.$$

Therefore, in the absence of any confounding effect, β can be interpreted as the correlation coefficient between the Y_i's and the X_i's.

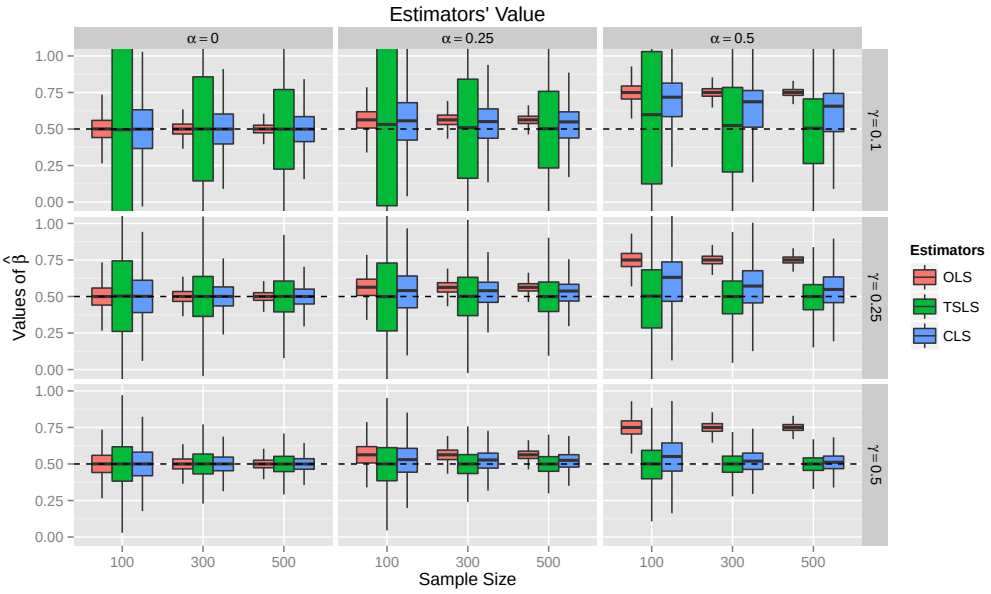

Fig 2. Monte Carlo distributions of the estimators' values under three different levels of confounding, $\alpha = \text{Cor}(X_i, U_i)$; and for three different levels of instrument's strength, $\gamma = \text{Cor}(X_i, Z_i)$. In each panel, the sample size varies between $n = 100$ and $n = 500$. We here compare the OLS, TSLS and CLS estimators with respect to the true parameter $\beta = 1/2$, whose value is indicated by a dashed line. These simulations are based on 10^5 iterations for each scenario. The boxplots are here centered at the median, and the upper and lower hinges correspond to the first and third quartiles.

4.2. Simulation Model II

We extend Model I to the case of several valid instruments. For convenience, these instruments are assumed to be uncorrelated. The second-level equation is taken to be identical to the second-level equation in equation (15). The first-level equation, by contrast, now includes a vector of instruments, such that

$$x_i = \mathbf{z}_i \boldsymbol{\lambda} + u_i \alpha + \delta_i, \tag{17}$$

for every $i = 1, \ldots, n$; and where $\mathbf{z}_i := (z_{i1}, \ldots, z_{il})$ is a row vector of l uncorrelated instruments. The strengths of each of these instruments are controlled by a column vector of parameters denoted by $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_l)'$. We here assume that the λ_j's are held constant such that $\lambda_j := \lambda$, for every $j = 1, \ldots, l$. As for Model I, we fix $\mathbb{V}\text{ar}(Y_i) = \mathbb{V}\text{ar}(X_i) = 1$, and generate the Z_{ij}'s and the U_i's from a standard normal distribution, with $\sigma_z^2 := 1$ and $\sigma_u^2 := 1$, respectively. The formula for the error variance of the second-level equation, $\sigma_\varepsilon^2(\alpha)$, is identical to the one used in Model I.

The error variance for the first-level equation, denoted by σ_δ^2, is also controlled by a parameter γ, which is here defined to be the coefficient of multiple

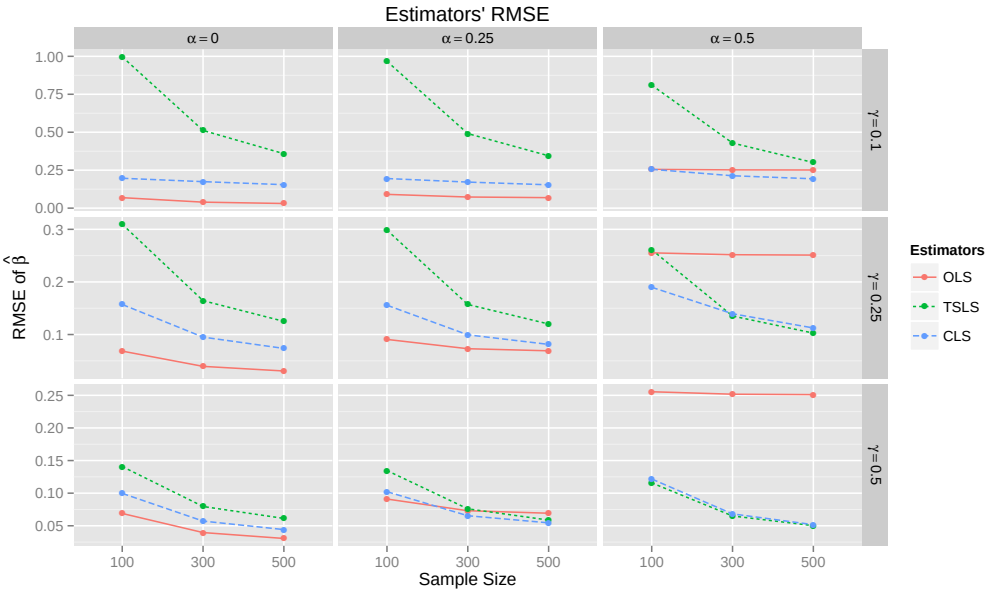

Fig 3. Monte Carlo estimates of the root mean squared errors (RMSEs) of the three estimators of interest under the simulation scenarios described in figure 2. As predicted, the RMSE of the proposed CLS method strikes a trade-off between its two constituent estimators. Indeed, under small α, the CLS's RMSE tends towards the RMSE of the OLS estimator; whereas under large γ, it tends towards the RMSE of the TSLS estimator.

correlation between each X_i and the i^{th} vector of Z_{ij}'s. As before, using the Bienaymé formula, the variance of each X_i can be expanded as follows,

$$\mathbb{V}\text{ar}(X_i) = \sum_{j=1}^{l} \lambda_j^2 \, \mathbb{V}\text{ar}(Z_{ij}) + \alpha^2 \, \mathbb{V}\text{ar}(U_i) + \mathbb{V}\text{ar}(\delta_i),$$

which simplifies to $\mathbb{V}\text{ar}(X_i) = l\lambda^2 + \alpha^2 + \sigma_\delta^2$, by our choice of variances for the Z_{ij}'s and U_i's. When specifying $\mathbb{V}\text{ar}(X_i) = 1$, this gives $\sigma_\delta^2(\alpha, \gamma, l) = 1 - (l\lambda^2 + \alpha^2)$, and moreover, when enforcing the positiveness of σ_δ^2, we obtain $\lambda < \sqrt{(1-\alpha^2)/l}$. Next, if we choose $\lambda := \gamma/\sqrt{l}$, then the parameter γ can be seen to correspond to the multiple correlation coefficient between each X_i and the i^{th} vector of Z_{ij}'s. Indeed, we have $\gamma^2 = \mathbf{r}^T \mathbf{R}^{-1} \mathbf{r}$, in which $\mathbf{r} := (r_{xz}, \ldots, r_{xz})'$, with $r_{xz} := \text{Cor}(X_i, Z_{ij}) = \gamma$; and where $\mathbf{R}$ is the correlation matrix of the i^{th} vector of Z_{ij}'s, such that $R_{ab} := \text{Cor}(Z_{ia}, Z_{ib})$, for every $a, b = 1, \ldots, l$. Thus, as in Model I, we again obtain the upper bound, $\gamma < \sqrt{1-\alpha^2}$, as well as $\sigma_\delta^2(\alpha, \gamma) = 1 - (\gamma^2 + \alpha^2)$.

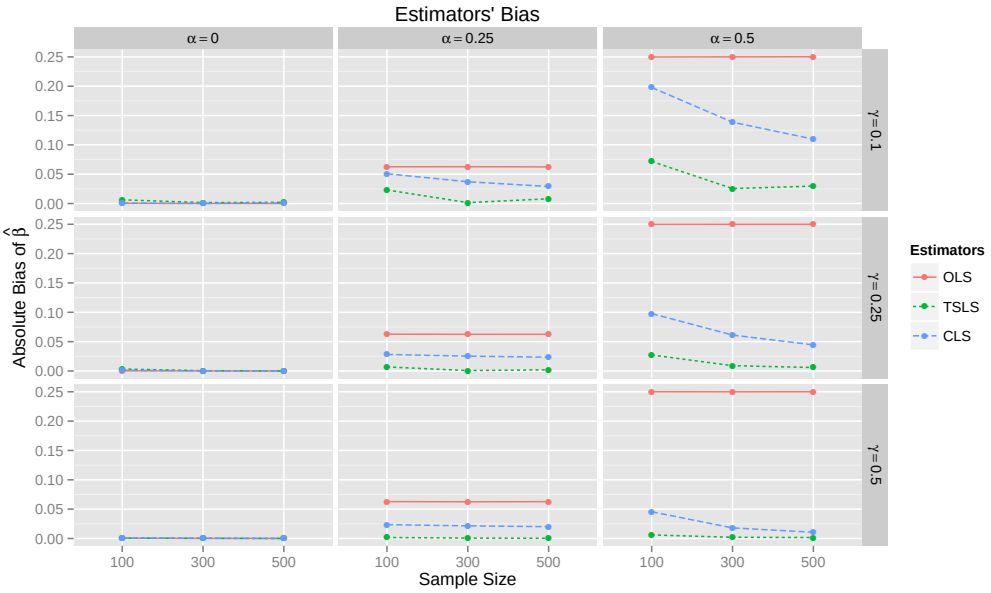

Fig 4. Monte Carlo estimates of the absolute value of the bias of the three estimators of interest, under the simulation scenarios described in figure 2. Observe that the three estimators exhibit no bias, when no confounding is present. That is, the OLS estimator exhibits less bias, when $\alpha = \text{Cor}(X_i, U_i)$ is low. Also, note that the finite-sample bias of the TSLS estimator tends to diminish with large sample sizes. This behavior is especially visible for large α's.

4.3. Monte Carlo Summary Statistics

We evaluate the finite-sample performance of the estimators of interest by comparing the Monte Carlo estimates of three different population statistics. For every candidate estimator, $\boldsymbol{\beta}_n^\dagger$, its Monte Carlo distribution is given by the following empirical distribution function (EDF), $\widehat{F}(b) := T^{-1} \sum_t \mathcal{I}\{\boldsymbol{\beta}_{nt}^\dagger \leq b\}$, where $\mathcal{I}\{f_t\}$ denotes the indicator function taking a value of one, if f_t is true, and zero otherwise. For each simulation scenario, we draw $T := 10^5$ realizations from the two models described in sections 4.1 and 4.2.

Using these simulated samples, we compute the Monte Carlo estimates of the bias, variance, and MSE; denoted by $\mathbb{B}\text{ias}_{\widehat{F}}(\boldsymbol{\beta}_n^\dagger)$, $\mathbb{V}\text{ar}_{\widehat{F}}(\boldsymbol{\beta}_n^\dagger)$, and $\text{MSE}_{\widehat{F}}(\boldsymbol{\beta}_n^\dagger)$, respectively. In figure 2, we have reported the Monte Carlo distribution of the three estimators of interest under T simulations from model I; whereas in figures 3, 4, and 5 we have reported the Monte Carlo MSE, squared bias, and variance, respectively. The quantities in these three figures have been square-rooted in order to facilitate the comparison of these statistics with the estimators' values in figure 2. Similarly, in figure 6, we have reported the Monte Carlo distributions of the TSLS, JIVE, as well as their CLS counterparts; with the Monte Carlo estimates of their absolute value bias being described in figure 7.

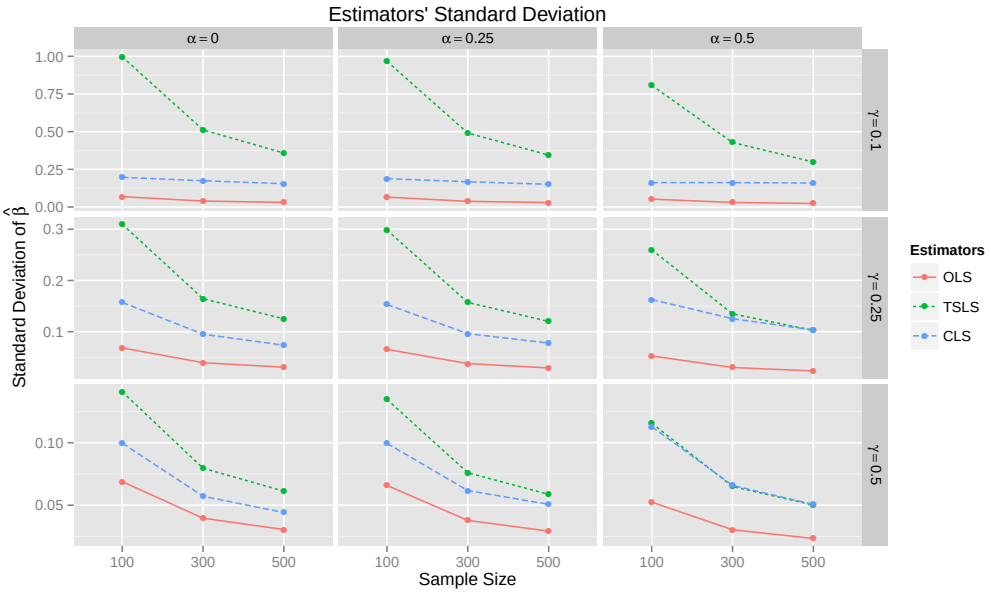

Fig 5. Monte Carlo standard deviation of the three estimators under scrutiny, under the scenarios described in figure 2. By corollary 1, the variance of the OLS estimator is always smaller than its competitors, as verified in these simulations. Moreover, observe that the variance of the TSLS estimator increases as the strength of the instrument, $\gamma = \text{Cor}(X_i, Z_i)$, decreases.

4.4. Results for Model I (Single Instrument)

The behavior of the CLS was found to be mainly controlled by the strength of the instrument, Z. When the instrument was strongly correlated with the predictor X –that is, for large values of $\gamma = \text{Cor}(X_i, Z_i)$; the values of the CLS estimator were close to the ones of the TSLS estimators, as can be observed in the last row of figure 2. By contrast, when the instrument was weak –that is, for small values of γ, the values of the CLS estimator were closer to the ones of the OLS estimator, as can be seen in the first row of figure 2.

Proposition 3 stated that the MSEs of the OLS and TSLS estimators are bounded below by the MSE of the CLS estimator when the true proportion π_n is known. These Monte Carlo simulations appear to support a partial analog of this result when π_n is evaluated from the data. Indeed, on one hand, figure 3 shows that the MSE of the OLS estimator tends to be smaller than the MSE of the CLS estimator, when no confounding is present; thereby showing that proposition 3 does not strictly hold when π_n is estimated from the data. However, on the other hand, one can also observe from figure 3 that the Monte Carlo MSE of the CLS estimator is smaller than or equal to the one of the TSLS estimator under all considered scenarios. Thus, it seems that a weaker version of proposition 3 may hold for estimated π_n, which would solely pertain to a comparison between the

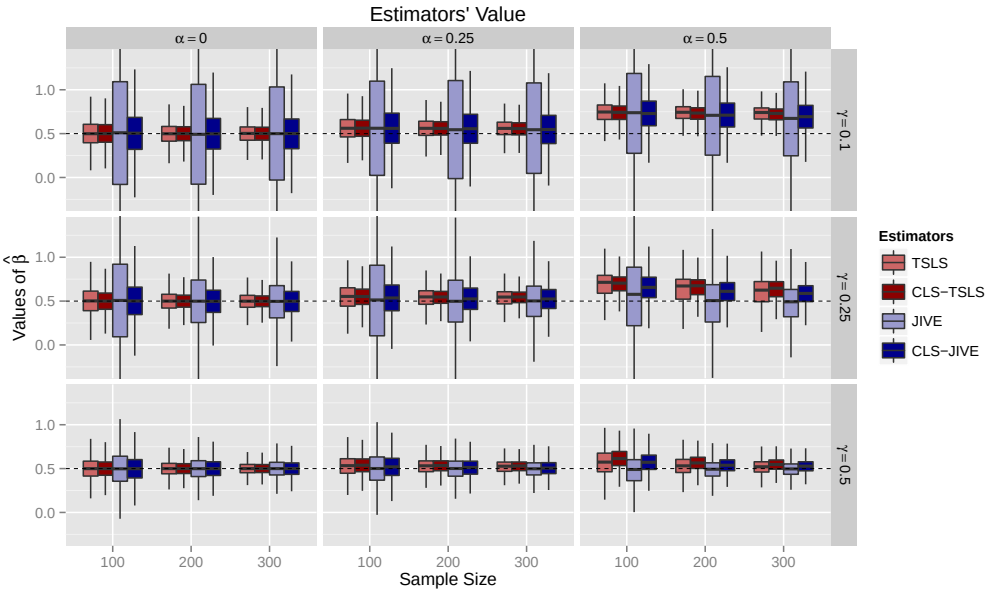

Fig 6. Comparison of the TSLS and JIVE estimators with their CLS counterparts, using $l = 10$ uncorrelated instruments, whose multiple correlation with the outcome is given by γ, and with α measuring the strength of bias, as in Model I. Data have been simulated using Model II from section 4.2. The bootstrap estimate of π_n for the CLS-JIVE is based on $B = 100$ resamples, as described in section 3.2 on bootstrap CLS estimation. All scenarios have been repeated over 10^5 Monte Carlo iterations.

behavior of the CLS and TSLS estimators. In particular, observe that for strong instruments (i.e. for large values of γ), the CLS estimator behave as well as the TSLS estimator, whereas for weak instruments (i.e. small values of γ), the CLS estimator outperforms the TSLS estimator.

The behavior of these estimators can be better understood by separately considering their bias and variance. In figures 4 and 5, we have respectively reported the Monte Carlo estimates of the bias and variance of the OLS, TSLS and CLS estimators. Naturally, the bias of the three estimators tends to increase with the strength of the confounder, which is controlled by $\alpha = \text{Cor}(X_i, U_i)$. In particular, the bias of the OLS estimator becomes larges as α increases. By contrast, the bias of the TSLS estimator remains small for every value of α. In fact, as stated in Corollary 1(i), the bias of the OLS estimator is bounded from below by the bias of the TSLS estimator. Moreover, the finite-sample bias of the TSLS estimator can also be observed to decrease as the sample size increases. As predicted, the bias of the CLS estimator is comprised between the ones of the two other estimators; and the bias of the CLS estimator approaches the one of the TSLS estimator, as the strength of the instrument increases.

Figure 5 describes the behavior of the variance of the estimators of interest under our various simulation scenarios. The variance of the three estimators

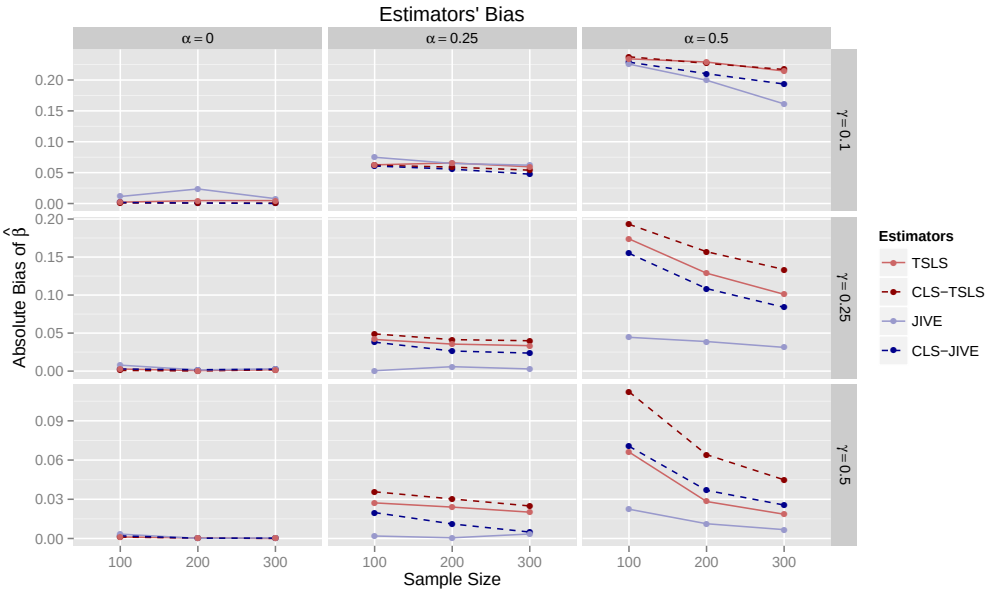

Fig 7. Monte Carlo estimates of the absolute value of the bias of four estimators of interest for Model II with $l = 10$ uncorrelated instruments, whose multiple correlation with the outcome is given by γ, and with the strength of the bias being denoted by α. As in figure 6, all scenarios have been repeated over 10^5 Monte Carlo iterations.

tends to decrease as the sample size increases. This downward trend is especially noticeable for the TSLS estimator, which exhibits a high level of variability, when the instrument is weak (i.e. for small values of γ). As predicted by Corollary 1(ii), the variance of the TSLS estimator can be observed to be bounded from below by the variance of the OLS estimator. In the presence of weak instruments, the CLS estimator's variance is close to the one of the OLS estimator. As γ increases, however, the variance of the CLS estimator converges to the one of the TSLS estimator.

4.5. Results for Model II (Multiple Instruments)

Our second set of simulations aimed to assess whether the use of an estimator possessing better finite-sample properties could be incorporated into the CLS framework. In figure 4, we have already seen that the TSLS estimator suffers from a substantial finite-sample bias. This was found to be especially the case when the instruments of interest are comparatively weak, and the bias is large. In particular, previous authors have shown that the TSLS estimator's bias tends to be especially large, when several instruments are used (Angrist et al., 1995). This limitation of the TSLS estimator has been addressed in the literature by the introduction of the JIVE, which was described in section 3.1. This second set of simulations is thus based on $l = 10$ uncorrelated instruments, and allow us

to compare the relative merits of using either the TSLS estimator or the JIVE within the CLS framework. Consequently, we will refer to using the TSLS and using the JIVE as unbiased estimators within the CLS, as the CLS-TSLS and CLS-JIVE estimators, respectively.

As predicted, the JIVE performs better than the TSLS estimator, when used in conjunction with strong instruments, and in the presence of a substantial amount of confounding (i.e. $\alpha \geq 0.25$), as can be seen from figures 6 and 7. Note, however, that for weak instruments (i.e. when the multiple correlation coefficient is $\gamma = 0.1$), the JIVE's variance is very large. The TSLS estimator should therefore be favored under these scenarios, if one's choice of estimator is motivated by a desire to minimize the MSE.

The benefits of using the JIVE translate into corresponding improvements when using the CLS-JIVE. This relationship is especially visible when considering the bottom right panel of figure 7. Under a set of strong instruments (i.e. with a large multiple correlation coefficient γ), and under a substantial amount of confounding (i.e. large α); one can observe that the JIVE has a smaller finite-sample bias than the TSLS estimator. Similarly, under the same scenario, the CLS-JIVE has a correspondingly smaller finite-sample bias than the CLS-TSLS estimator. This improvement in the CLS-JIVE was particularly remarkable, because the proportion parameter, $\widehat{\pi}_n$, was estimated using only $B = 100$ bootstrap samples. Thus, it appears that a relatively small number of resampling is sufficient to produce a CLS-JIVE estimator that outperforms the CLS-TSLS estimator. One may therefore conjecture that the CLS framework could be used in conjunction with other asymptotically unbiased estimators, even when the proportion parameter is not available in closed-form.

5. Applications to Econometrics

Our proposed methods have been applied to a re-analysis of a classic data set in econometrics, originally published by Angrist and Krueger (1991), which aimed to relate educational attainment with earnings. This particular study has been the subject of numerous replications and re-analysis, and therefore provides us with a well-known example for evaluating the performance of CLS estimation in a real-world data set.

5.1. Quarter-of-Birth as Instrument

Angrist and Krueger (1991) reported a small but persistent seasonal pattern of educational attainment over several decades between the 1920s and the 1950s. They observed that two discrepant regulations in the United States during that period have led to a 'natural experiment', in which individual differences in completed years of education could be predicted by an individual's season of birth. On one hand, nationwide school-entry requirements controlled the age at which a given child began school. Indeed, at that period in the US, all children were expected to reach six years of age by the first of January of their first year

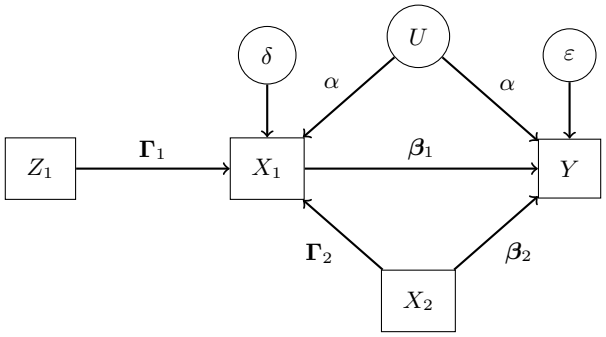

Fig 8. Graphical representation of the IV model described in equations (18) and (19), composed of a vector of endogenous variables, X_1, and a vector of exogenous variables X_2. This graph corresponds to a two-level system of equations composed of $Y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + U\alpha + \varepsilon$, and $X_1 = Z_1\mathbf{\Gamma}_1 + X_2\mathbf{\Gamma}_2 + U\alpha + \delta$. This model can be seen to be a generalization of the simpler model described in figure 1.

at school. Thus, children born early in the year were likely to be older than their peers in the same class. On the other hand, state-specific compulsory schooling laws solely required pupils to remain in school until their sixteenth birthday. Therefore, pupils born in the first quarters of the year, wishing to leave school early, could do so at an earlier stage than their peers born later in the year.

These two regulations –school-entry requirements, and compulsory schooling laws– therefore conspired to enable children born in the early quarters of the year to complete a smaller number of years of education, if they were so inclined. Crucially, Angrist and Krueger (1991) highlights that the randomness of an individual's birth date is unlikely to be related to other events in an individual's life; thereby precluding quarter-of-birth from being a significant predictor of an individual's revenue later in life. Thus, quarter-of-birth could be argued to constitute a legitimate instrument for education attainment, fulfilling the exclusion criterion, in the sense that it is not directly related to earnings. Note, however, that some authors have disputed the validity of quarter-of-birth as an instrument for education (Bound and Jaeger, 1996).

5.2. Model with Extraneous Covariates

The model used by Angrist and Krueger (1991) generalizes the IV model described in sections 2.1 and 2.2. Here, the vector of k predictors, X, is partitioned into k_1 endogenous variables denoted by X_1, and k_2 exogenous variables denoted by X_2. In Angrist and Krueger's model, the sole endogenous variable of interest is the completed years of education of each subject. Therefore, we have $k_1 = 1$. The outcome variable, Y, which represents the log-transformed weekly wage in dollars of each subject, is then modelled as follows,

$$Y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \varepsilon, \tag{18}$$

TABLE 1
Replications and extensions of the IV analysis in Angrist and Krueger (1991).

Covariates (X_2)		OLS	TSLS	JIVE	CLS-TSLS
A.[a]	*Estimator:*	0.0802	0.0769	0.0755	0.0800
	Std. Error:	(0.0004)	(0.0150)	(0.0210)	(0.0126)
	Proportion:	–	–		(0.95)
B.[b]	*Estimator:*	0.0802	0.1398	-0.1276	0.1254
	Std. Error:	(0.0004)	(0.0334)	(1.7233)	(0.0317)
	Proportion:	–	–		(0.24)
C.[c]	*Estimator:*	0.0701	0.0669	0.0650	0.0700
	Std. Error:	(0.0004)	(0.0151)	(0.0234)	(0.0097)
	Proportion:	–	–		(0.96)
D.[d]	*Estimator:*	0.0701	0.1065	0.0224	0.0899
	Std. Error:	(0.0004)	(0.0334)	(2.1380)	(0.0235)
	Proportion:	–	–		(0.46)

[a] This only includes ten dummies for the years of birth.
[b] This includes years of birth, and age with the exclusion of the 1929 dummy.
[c] Covariates include years of birth, and some extraneous covariates described in the text.
[d] Covariates are years of birth, age, and other extraneous covariates, with the exclusion of the 1929 dummy.

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are column vectors of dimension k_1 and k_2, respectively. The endogeneity of X_1 leads to the use of a row vector of instruments, Z_1, of dimension $1 \times l_1$, here denoting a set of dummy variables for the interactions between quarters and years of birth. These instruments are combined with the k_2 exogenous variables from the second-level equation in order to produce the following first-level equation,

$$X = Z_1\boldsymbol{\Gamma}_1 + X_2\boldsymbol{\Gamma}_2 + \delta; \tag{19}$$

where $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are vectors of parameters of order $l_1 \times k$ and $k_2 \times k$, respectively; where $k := k_1 + k_2$. A graphical representation of this model is given in figure 8.

In matrix notation, given a sample of n subjects, this model can be expressed with respect to an n-dimensional column vector of error terms, $\boldsymbol{\varepsilon}$, for the first-level equation; and a matrix, $\mathbf{D}$, of error terms of order $n \times k_1$ for the second-level equation. Altogether, we thus have the following linear system,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \\ \mathbf{X} &= \mathbf{Z}_1\boldsymbol{\Gamma}_1 + \mathbf{X}_2\boldsymbol{\Gamma}_2 + \mathbf{D}. \end{aligned}$$

Moreover, we can construct the following block matrices, $\mathbf{X} := [\mathbf{X}_1 \ \mathbf{X}_2]$ and $\mathbf{Z} := [\mathbf{Z}_1 \ \mathbf{X}_2]$ that are of order $n \times k$ and $n \times l$, respectively; in which we have used $k := k_1 + k_2$ and $l := l_1 + k_2$. In addition, we also define the vectors of parameters $\boldsymbol{\beta} := [\boldsymbol{\beta}_1^T \ \boldsymbol{\beta}_2^T]^T$ and $\boldsymbol{\Gamma} := [\boldsymbol{\Gamma}_1^T \ \boldsymbol{\Gamma}_2^T]^T$, of order $k \times 1$ and $l \times k$,

respectively. Equipped with these block matrices, we can immediately recover the standard TSLS estimator formula described in section 2.2. It also follows that this model is well-identified whenever $l_1 \geq k_1$ is satisfied, as in the model at hand.

5.3. Results of the Re-analysis

The results described in table 4 of Angrist and Krueger (1991) have been replicated and extended. In this portion of their analysis, Angrist and Krueger have considered the cohort of men born between 1920 and 1929. This constitutes a sample of $n = 247,199$ subjects. All data are here based on the 1970 US census. Using the notation introduced in equations (18) and (19), the outcome variable, Y, is defined as the mean log-transformed weekly wages; the endogenous variable, X_1, is the number of completed years of education; and the instrument, Z_1, is composed of a vector of interaction terms between quarter-of-year dummies and year-of-birth dummies, totalling 40 different instruments.

In addition, the authors have also considered different sets of exogenous covariates, denoted by X_2 in equations (18) and (19). The choice of exogenous covariates has been reported as covariate scenarios A to D in table 1. All scenarios include ten dummy variables for each year of birth, except scenarios B and D, in which the 1929 dummy variable has been removed due to multicollinearity, following Angrist and Krueger (1991). In scenarios B and D, this is supplemented by an age covariate. (Note that we have not included age squared in this analysis as was conducted in Angrist and Krueger (1991), since age and age squared were found to be almost perfected correlated.) Finally, scenarios C and D include some further dummy variables for race, marital status, eight different regions of residence, and whether or not the subjects were primarily located in a standard metropolitan statistical area (SMSA).

For scenarios A and C, the OLS and TSLS columns in table 1 are exact replicates of the results described in Angrist and Krueger (1991). The values and standard errors for these estimators are slightly different under scenarios B and D, due to the non-inclusion of age squared in the present analysis. The variance of the CLS estimator was computed using the bootstrap, as described in section 2.6. As expected, one can observe that the CLS strikes a balance between the OLS and the TSLS estimators, such that for all four scenarios, the value of the CLS is comprised between the one of the OLS and the one of the TSLS. The value taken by the estimate of the proportion parameter has also been reported. By comparing scenarios A and C with scenarios B and D in table 1, it can be seen that the inclusion (resp. non-inclusion) of the age variable leads to a decrease (resp. increase) in the value of $\widehat{\pi}_n$.

Thus, this re-analysis suggests that while the use of quarter-of-birth as an IV for education may be justified when age is included as a supplementary exogenous variable in the analysis; it appears that an estimator closer to the OLS is sufficient, when the age variable is not included. The JIVE and its CLS counterpart have also been reported for comparison. The behavior of these two

estimators is comparable to the one of the TSLS and CLS-TSLS estimators. Note that for computational convenience, the variances of the CLS estimators and the proportion parameter of the CLS-JIVE have been here estimated using bootstraps based on solely 10^2 resamples. This real data analysis thus demonstrates that a small number of bootstrap samples suffices to produce reasonable estimates of the standard errors of the CLS estimators.

6. Conclusions

In this paper, we have shown that different IV and non-IV estimators can be the object of convex combinations, and the proportion parameters of these combinations can be consistently estimated from the data. Such CLS estimators are therefore particularly attractive, since they automatically down-weight the influence of weak instruments, when these are not expected to lead to a large reduction in bias. Moreover, this inferential framework bears some similarities with the Hausmann test. Theoretically, our proposed estimator minimizes an empirical MSE over a restricted class of estimators, consisting of all the possible convex combinations of the OLS and TSLS estimators. We have also seen the TSLS estimator in the definition of the CLS can be replaced by other estimators such as the JIVE.

For finite n, the moments of the TSLS estimator and other k-class estimators need not exist, as demonstrated by Kinal (1980). It is common in such situations to assume that at least three instruments are present. This condition ensures that the first two moments of the estimators under scrutiny exist; and the first two moments of the OLS and TSLS estimators are needed to compute our proposed empirical estimator of the MSE. However, note that, asymptotically, all moments exist and that, theoretically, this strategy can thus be applied to any number of instruments. Indeed, irrespective of the number of instruments used, every CLS estimator is guaranteed to be asymptotically consistent. Observe that a similar issue arises for all IV models, in the sense that such models are only asymptotically identifiable. In the present case, our proposed estimation procedure only possesses asymptotic first and second moments. However, just as finite-sample non-identifiability is not a point of concern for the use of IV methods in practice, the reliance of the CLS framework on asymptotic moments does not constitute a significant hindrance to the general application of this method.

The interpretation of the combination of several estimators such as the OLS and TSLS estimators relies on the assumption of *effect homogeneity* [REFs needed]. That is, we are assuming that the causal effect is identical for all subpopulations. This is a strong assumption, which needs not hold in practice. Thus, further research will be needed to clarify the assumptions required for employing the CLS, when effect homogeneity is not expected to hold. Indeed, in such cases, the OLS and the TSLS estimators may represent the local treatment effects in different subpopulations. Therefore, the resulting convex combination of such estimators may be difficult to interpret. Additional assumptions may

be required to ensure that the different estimators of interest are sufficiently comparable to be combined in this fashion.

Observe that the estimators utilized to produce the CLS estimator do not need to share the same data. Indeed, when constructing a combination of the OLS and TSLS estimators, only the TSLS estimator relies on the instrument, Z. Thus, one may also consider how such a framework could be extended to other modelling strategies, such as mixed-effects models for longitudinal models (see Wooldridge, 2002). Similarly, this method could also be extended to models including measurement errors. Calibrated regression is often used in conjunction with explanatory variables, in order to diminish the effect of measurement error. In such cases, the resulting combination utilizes estimators based on distinct data sets. In addition, recall that of the theoretical results derived in this paper are relying one the assumption that the instruments under scrutiny are valid, in the sense that (A4) is assumed to hold. The consequences of relaxing this assumption are difficult to anticipate, and further research should certain consider such situations, as has been done by previous authors in the case of the TSLS estimator (Jackson and Swanson, In press).

Thus far, we have combined the OLS estimator with either the TSLS estimator or the JIVE. Observe, however, that we are not restricted to choosing the OLS as a reference estimator. Within the bootstrap CLS framework described in section 3.2, one could also choose to combine the TSLS and the JIVE, for instance. In general, any pair of estimators could be the object of a convex combination. For such a combination to be useful, it suffices that these estimators are ordered in terms of bias and variance, as in the canonical case of the OLS and TSLS estimators given in proposition 1. Another natural theoretical extension of the current work would be to derive a central limit theorem for the CLS estimator. This would allow researchers to obtain approximate confidence intervals for the CLS estimator, using normal asymptotic theory. Such a central limit theorem would also enable the construction of adequate statistical tests for evaluating whether or not the values of individual parameters are statistically significant. Such extensions are not expected to be too arduous, since under the assumptions stated in this paper, the CLS estimator is consistent; and moreover, estimators such as the TSLS or JIVE are known to be asymptotically normally distributed under standard assumptions.

Appendix A: Proofs of Propositions

Proof of Proposition 1. The proof of (i) immediately follows from the definition of the empirical bias in equation (8), which implies that the empirical bias of the TSLS estimator is identically zero, for every realization. The proof of (ii) can be conducted in two steps. Firstly, one can show that for every pair of matrices $\mathbf{X}$ and $\widehat{\mathbf{X}} := \mathbf{H}_z\mathbf{X}$, we have

$$\mathbf{X}'\mathbf{X} \succeq \widehat{\mathbf{X}}'\widehat{\mathbf{X}}. \tag{20}$$

Observe that we have the following equivalence due to the symmetry of $\mathbf{H}_z$,

$$\widehat{\mathbf{X}}'\mathbf{X} = (\mathbf{H}_z\mathbf{X})'\mathbf{X} = \mathbf{X}'\mathbf{H}_z\mathbf{X} = \mathbf{X}'\widehat{\mathbf{X}}. \tag{21}$$

Secondly, the inner product of $\widehat{\mathbf{X}}$ can also be simplified using the idempotency of $\mathbf{H}_z$, such that

$$\mathbf{X}'\widehat{\mathbf{X}} = \mathbf{X}'\mathbf{H}_z\mathbf{X} = \mathbf{X}'\mathbf{H}_z\mathbf{H}_z\mathbf{X} = \widehat{\mathbf{X}}'\widehat{\mathbf{X}}. \tag{22}$$

Then, expanding the dot product of $\mathbf{X} - \widehat{\mathbf{X}}$, and applying equalities (21) and (22), we obtain

$$(\mathbf{X} - \widehat{\mathbf{X}})'(\mathbf{X} - \widehat{\mathbf{X}}) = \mathbf{X}'\mathbf{X} - 2\mathbf{X}'\widehat{\mathbf{X}} + \widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{X}'\mathbf{X} - \widehat{\mathbf{X}}'\widehat{\mathbf{X}}.$$

Observe that the dot product, $(\mathbf{X}-\widehat{\mathbf{X}})'(\mathbf{X}-\widehat{\mathbf{X}})$, is a Gram matrix, and therefore it is necessarily positive semi-definite. Consequently, this implies that

$$\mathbf{X}'\mathbf{X} - \widehat{\mathbf{X}}'\widehat{\mathbf{X}} = (\mathbf{X} - \widehat{\mathbf{X}})'(\mathbf{X} - \widehat{\mathbf{X}}) \succeq \mathbf{0},$$

and hence $\mathbf{X}'\mathbf{X} \succeq \widehat{\mathbf{X}}'\widehat{\mathbf{X}}$, by the definition of the positive semidefinite order.

Next, observe that the estimates of the error variances under the OLS and TSLS estimation procedures are defined as

$$\begin{aligned}(n-k)\widehat{\sigma}_n^2 &:= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_n)'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_n) \\ &\leq (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_n)'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_n) =: (n-k)\widetilde{\sigma}_n^2,\end{aligned}$$

where the inequality follows from the optimality of the OLS; and therefore,

$$\widehat{\sigma}_n^2(\mathbf{X}'\mathbf{X})^{-1} \preceq \widetilde{\sigma}_n^2(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1},$$

since assumption (A6) guarantees that both sides are invertible. □

Proof of Corollary 1. For inequality (i), observe that the empirical bias of the TSLS estimator is identically zero for every realization and every n. Moreover, by equation 8, the empirical bias of the OLS estimator is given by

$$\widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n) = (\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n)(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n)' \succeq \mathbf{0},$$

since the LHS is a Gram matrix, and is therefore positive semidefinite for every realization. Since it holds for every n, this positive semidefiniteness is preserved in the limit. Moreover, the inequality in (ii) immediately follows from the fact that the variances of these two estimators converge to the zero matrix. □

Proof of Proposition 2. The optimal value of π_n can be found by minimizing the criterion of interest, which will be denoted by $f(\pi) := \mathrm{MSE}(\bar{\boldsymbol{\beta}}_n(\pi))$. For expediency, we will expand this criterion as was done in equation (10) such that

$$\operatorname{tr} f(\pi) = \operatorname{tr}(\pi^2 M_1 + 2\pi(1-\pi)C + (1-\pi)^2 M_2),$$

with $M_1 := \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_n)$, $C := \mathrm{CSE}(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n)$, and $M_2 := \mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_n)$; and where recall that $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ denote the OLS and TSLS estimators, respectively. Since the derivative is a linear operator, it commutes with the trace, and we obtain

$$\operatorname{tr}(\partial f/\partial\pi) = 2\pi \operatorname{tr}(M_2 - 2C + M_1) - 2\operatorname{tr}(M_2 - C),$$

Setting this expression to zero, yields $\pi_n := \operatorname{tr}(M_2 - C)/\operatorname{tr}(M_2 - 2C + M_1)$, as required. Naturally, this minimization holds for every choice of n, thereby proving the first part of proposition 2.

In addition, one can show that this minimizer is unique by performing a second derivative test, such that we obtain

$$\operatorname{tr}(\partial^2 f/\partial \pi^2) = 2\operatorname{tr}(M_1 - 2C + M_2). \tag{23}$$

Since by assumption, the random vectors, $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$, are elementwise squared-integrable, the components, $\mathbb{E}[(\widehat{\beta}_{nj} - \beta_j)^2]$, of M_1 are finite for every $j = 1, \ldots, k$. Hence, using the linearity of the trace, the MSE of $\widehat{\boldsymbol{\beta}}_n$ can be treated as a sum of real numbers, thereby yielding the L^2-norm on $\mathbb{R}^k$, such that

$$\big(\operatorname{tr}\operatorname{MSE}(\widehat{\boldsymbol{\beta}}_n)\big)^{1/2} = \left(\sum_{j=1}^{k} \mathbb{E}\Big[(\widehat{\beta}_{nj} - \beta_j)^2\Big]\right)^{1/2} =: ||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||.$$

The latter quantity will be referred to as the root (trace) MSE of $\widehat{\boldsymbol{\beta}}_n$, and will be denoted by RMSE.

By the same reasoning, it can be shown that C and M_2 corresponds to the inner product, $\langle \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \rangle$, and the squared norm, $||\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||^2$, on $\mathbb{R}^k$. Thus, equation (23) can now be re-expressed as follows,

$$\operatorname{tr}(\partial^2 f/\partial \pi^2) = 2\Big(||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||^2 - 2\langle \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \rangle + ||\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||^2\Big).$$

The Cauchy-Schwarz inequality can here be invoked to produce an upper bound on the cross-term in the latter equation,

$$\langle \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \rangle \leq ||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}|| \cdot ||\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||.$$

It then suffices to complete the square in order to obtain the following lower bound,

$$\operatorname{tr}(\partial^2 f/\partial \pi^2) \geq 2\Big(||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}|| - ||\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}||\Big)^2 \geq 0,$$

for every n, and where equality only holds when the RMSEs of $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ are identical, as required. □

Proof of Proposition 4. Firstly, observe that the stochastic convergence of $\bar{\boldsymbol{\beta}}_n(\pi)$ to $\pi\widehat{\boldsymbol{\beta}} + (1 - \pi)\widetilde{\boldsymbol{\beta}}$ is immediate from the convergence of $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ to their respective limits, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$. Naturally, this holds for every $\pi \in [0, 1]$.

Secondly, recall that the empirical MSE of $\bar{\boldsymbol{\beta}}_n(\pi)$ can be decomposed into a variance and a bias term, as in equation (12), such that for every π, we have

$$\widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(\pi)) = \widehat{\mathbb{V}\mathrm{ar}}(\bar{\boldsymbol{\beta}}_n(\pi)) + \widehat{\mathbb{B}\mathrm{ias}}^2(\bar{\boldsymbol{\beta}}_n(\pi)).$$

The variances of both the OLS and TSLS estimators are known to converge to zero. Moreover, the bias of the TSLS estimator is also known to converge to zero,

and this can be seen to be also true for the cross-bias term. Thus, the stochastic limit of $\widehat{\mathrm{MSE}}(\bar{\boldsymbol{\beta}}_n(\pi))$ reduces to the weighted limit of the empirical bias of the OLS estimator, $\pi^2\widehat{\mathbb{B}\mathrm{ias}}^2(\widehat{\boldsymbol{\beta}}_n)$. Using the consistency of the TSLS estimator, this latter term satisfies

$$\operatorname*{plim}_{n\to\infty} \widehat{\mathbb{B}\mathrm{ias}}(\widehat{\boldsymbol{\beta}}_n) = \operatorname*{plim}_{n\to\infty}(\widetilde{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n) = \mathbb{B}\mathrm{ias}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}),$$

since $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\beta}} = \lim_n \mathbb{E}[\widehat{\boldsymbol{\beta}}_n]$; which is the true bias of the OLS estimator. The result then follows by using the continuous mapping theorem. □

Proof of Corollary 2. The minimization of the empirical MSE follows from the arguments used in the first part of proposition 2, and simplifying the closed-form formula for $\widehat{\pi}_n$ using our adopted definitions for the empirical bias of the TSLS estimator. □

Proof of Proposition 3. This inequality relates the theoretical MSEs of the CLS, OLS and TSLS estimators. This result follows from the convexity of the MSE with respect to its argument, $\boldsymbol{\beta}_n^\dagger$; in which $\boldsymbol{\beta}_n^\dagger$ represents any candidate estimator. The trace of the MSE can indeed be seen to be a convex quadratic form, $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is here the identity matrix and $\mathbf{x} := \boldsymbol{\beta}_n^\dagger - \boldsymbol{\beta}$. That is, for every estimator, $\boldsymbol{\beta}_n^\dagger$, let

$$f(\boldsymbol{\beta}_n^\dagger - \boldsymbol{\beta}) := \operatorname{tr}\left((\boldsymbol{\beta}_n^\dagger - \boldsymbol{\beta})(\boldsymbol{\beta}_n^\dagger - \boldsymbol{\beta})'\right) = (\boldsymbol{\beta}_n^\dagger - \boldsymbol{\beta})'(\boldsymbol{\beta}_n^\dagger - \boldsymbol{\beta}),$$

using the cyclic property of the trace, which shows that this quadratic form is convex. Thus, for every two estimators, $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$, and every $\pi \in [0,1]$; using the fact that $\boldsymbol{\beta}$ can always be expressed as $\pi\boldsymbol{\beta} + (1-\pi)\boldsymbol{\beta}$, we have

$$\begin{aligned} f\Big(\pi\widehat{\boldsymbol{\beta}}_n + (1-\pi)\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\Big) &= f\Big(\pi(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + (1-\pi)(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\Big) \\ &\leq \pi f\Big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\Big) + (1-\pi) f\Big(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\Big). \end{aligned}$$

Since by definition, $\bar{\boldsymbol{\beta}}_n(\pi) := \pi\widehat{\boldsymbol{\beta}}_n + (1-\pi)\widetilde{\boldsymbol{\beta}}_n$, and moreover $\operatorname{tr}(\mathrm{MSE}(\boldsymbol{\beta}_n^\dagger)) = \mathbb{E}[f(\boldsymbol{\beta}_n^\dagger)]$ for every $\boldsymbol{\beta}_n^\dagger$, it then follows that, using the linearity of the expectation,

$$\operatorname{tr}\mathrm{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) \leq \pi \operatorname{tr}\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_n) + (1-\pi)\operatorname{tr}\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_n).$$

Finally, observing that the set of convex combinations of the form, $\pi f(\widehat{\boldsymbol{\beta}}_n) + (1-\pi) f(\widetilde{\boldsymbol{\beta}}_n)$, necessarily includes the endpoints of the corresponding line segment, we hence obtain

$$\begin{aligned} \min_\pi \Big(\operatorname{tr}\mathrm{MSE}(\bar{\boldsymbol{\beta}}_n(\pi))\Big) &\leq \min_\pi \Big(\pi \operatorname{tr}\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_n) + (1-\pi)\operatorname{tr}\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_n)\Big) \\ &\leq \min\Big\{\operatorname{tr}\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_n), \operatorname{tr}\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_n)\Big\}, \end{aligned}$$

as required. □

Proof of Proposition 5. We use a sandwich argument to prove that the CLS estimator is MSE consistent, as stated in (ii). For every $p \in [0,1]$, let

$$T_n(p) := \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(p)) - \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(p)).$$

By proposition 4, we know that $T_n(p) \xrightarrow{p} 0$, for every p. However, the quantity of interest in the case of proposition 5 can be defined as follows,

$$T_n^*(p) := \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(p)) - \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi));$$

with p being chosen as $\widehat{\pi}_n$, and where π denotes the true proportion minimizing the theoretical MSE. Firstly, observe that we can find a lower bound for $T_n^*(\widehat{\pi}_n)$ with respect to $T_n(p)$ by a judicious choice of p. That is,

$$\begin{aligned} T_n(\widehat{\pi}_n) &= \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)) - \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)) \\ &\leq \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)) - \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) = T_n^*(\widehat{\pi}_n), \end{aligned}$$

since, by definition, $\pi := \min_p \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}(p))$. Secondly, one can also derive an upper bound for $T_n^*(\widehat{\pi}_n)$, as follows,

$$\begin{aligned} T_n^*(\pi) &= \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n)) - \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) \\ &\leq \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(\pi))) - \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) = T_n(\pi), \end{aligned}$$

since $\widehat{\pi}_n := \min_p \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}(p))$. Therefore, we obtain the following sandwich inequality,

$$T_n(\widehat{\pi}_n) \leq T_n^*(\widehat{\pi}_n) \leq T_n(\pi),$$

which could be re-expressed as follows,

$$|T_n^*(\widehat{\pi}_n)| \leq \max\{|T_n(\widehat{\pi}_n)|, |T_n(\pi)|\} \xrightarrow{p} 0,$$

where the weak convergence to zero was stated in proposition 4. Thus, we have demonstrated that

$$\min_p \operatorname{tr} \widehat{\operatorname{MSE}}(\bar{\boldsymbol{\beta}}_n(p)) \xrightarrow{p} \min_p \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(p)).$$

By proposition 3, we can use the MSEs of the OLS and TSLS estimators as upper bound for the RHS of the latter equation such that

$$\min_\pi \operatorname{tr} \operatorname{MSE}(\bar{\boldsymbol{\beta}}_n(\pi)) \leq \operatorname{tr} \min\{\operatorname{MSE}(\widehat{\boldsymbol{\beta}}_n), \operatorname{MSE}(\widetilde{\boldsymbol{\beta}}_n)\} \xrightarrow{p} 0,$$

since we know that the TSLS is MSE consistent. Therefore, $\bar{\boldsymbol{\beta}}_n(\widehat{\pi}_n) \xrightarrow{L^2} \boldsymbol{\beta}$, as required. Moreover, the weak consistency of the CLS estimator stated in (i) is a direct consequence of its MSE consistency (see, for example, Bain and Engelhardt, 1992, p.313). □

Proof of Corollary 3. The proof of this inequality is analogous to the proof of proposition 3, in which the theoretical expectation with respect to F, is replaced with the bootstrap expectation with respect to F^*. Here, the true parameter is taken to be $\mathbb{E}^*[\boldsymbol{\beta}_n^\dagger]$, where $\boldsymbol{\beta}_n^\dagger$ is an asymptotically unbiased estimator of interest, and $\bar{\boldsymbol{\beta}}_n(\pi) := \pi\widehat{\boldsymbol{\beta}}_n + (1-\pi)\boldsymbol{\beta}_n^\dagger$ is the BCLS estimator. Letting $f(\mathbf{x}) := \mathbf{x}'\mathbf{x}$, and using the convexity of this quadratic form, we obtain

$$f\Big(\bar{\boldsymbol{\beta}}_n(\pi) - \mathbb{E}^*[\boldsymbol{\beta}_n^\dagger]\Big) \leq \pi f\Big(\widehat{\boldsymbol{\beta}}_n - \mathbb{E}^*[\boldsymbol{\beta}_n^\dagger]\Big) + (1-\pi) f\Big(\boldsymbol{\beta}_n^\dagger - \mathbb{E}^*[\boldsymbol{\beta}_n^\dagger]\Big).$$

for every π. Moreover, by definition, $\operatorname{tr}(\widehat{\operatorname{MSE}}^*(\bar{\boldsymbol{\beta}}_n)) = \mathbb{E}^*[f(\bar{\boldsymbol{\beta}}_n)]$. Thus, using linearity, this gives

$$\operatorname{tr}\widehat{\operatorname{MSE}}^*(\bar{\boldsymbol{\beta}}_n(\pi)) \leq \pi \operatorname{tr}\widehat{\operatorname{MSE}}^*(\widehat{\boldsymbol{\beta}}_n) + (1-\pi)\operatorname{tr}\widehat{\operatorname{MSE}}^*(\boldsymbol{\beta}_n^\dagger).$$

The required inequality then follows by minimizing both sides with respect to $\pi \in [0,1]$, and noticing that every set of convex combinations contains the endpoints of the corresponding line segment, as in proposition 3. □

References

Angrist, J.D., Imbens, G.W., and Krueger, A.B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14**(1), 57–67.

Angrist, J., Imbens, G., and Krueger, A.B. (1995). Jackknife instrumental variables estimation. *Technical Working Paper 172*, National Bureau of Economic Research.

Angrist, J. and Krueger, A.B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *The journal of economic perspective*, **15(4)**, 69–85.

Angrist, J.D. and Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **56**(4).

Bain, L. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. Duxbury, Pacific Grove, CA.

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

Bound, J. and Jaeger, D.A. (1996). On the validity of season of birth as an instrument in wage equations: A comment on Angrist and Krueger's "Does compulsory school attendance affect school attendance affect schooling and earnings?". *Working paper 5835*, National Bureau of Economic Research, Cambridge, MA.

Bound, J., Jaeger, D.A., and Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, **90**(430), 443–450.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. Cambridge University press, Cambridge.

Davidson, R. and MacKinnon, J.G. (1993). Estimation and inference in econometrics. *OUP Catalogue*.

Hahn, J., Hausman, J., and Kuersteiner, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal*, **7**(1), 272–306.

Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica*, 1251–1271.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, 441–462.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, **92**(3), 557–561.

Jackson, J. and Swanson, S. (In press). Toward a clearer portrayal of confounding bias in instrumental variable analyses. *Epidemiology*.

Kinal, T.W. (1980). The existence of moments of k-class estimators. *Econometrica*, **48**(1), 241–249.

Maddala, G.S. and Jeong, J. (1992). On the exact small sample distribution of the instrumental variable estimator. *Econometrica*, **60**(1), 181–183.

Nelson, C. and Startz, R. (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, **58(4)**, 967–976.

Ng, S. and Bai, J. (2009). Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, **1**(1), –.

Palmer, T.M., Lawlor, D.A., Harbord, R.M., Sheehan, N.A., Tobias, J.H., Timpson, N.J., Smith, G.D., and Sterne, J.A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical methods in medical research*, **21**(3), 223–242.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96–146.

Pierce, B.L. and Burgess, S. (2013). Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American journal of epidemiology*, kwt084–.

Sawa, T. (1973). Almost unbiased estimator in simultaneous equations systems. *International Economic Review*, 97–106.

Staiger, D.O. and Stock, J.H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, **65**(3), 557–586.

Stock, J.H. and Trebbi, F. (2003). Retrospectives who invented instrumental variable regression? *Journal of Economic Perspectives*, 177–194.

Wong, K.F. (1996). Bootstrapping Hausman's exogeneity test. *Economics Letters*, **53**, 139–143.

Wooldridge, J. (2002). *Econometric analysis of cross-section and panel data*. MIT press, London.

Wright, P.G. (1928). *Tariff on animal and vegetable oils*. National Agricultural Library.