Bias Reduction in Instrumental Variable Estimation
Through First-Stage Shrinkage

Jann Spiess
Stanford University
January 4, 2021

Abstract

The two-stage least-squares (2SLS) estimator is known to be biased when its first-stage fit is poor. I show that better first-stage prediction can alleviate this bias. In a two-stage linear regression model with Normal noise, I consider shrinkage in the estimation of the first-stage instrumental variable coefficients. For at least four instrumental variables and a single endogenous regressor, I establish that the standard 2SLS estimator is dominated with respect to bias. The dominating IV estimator applies James–Stein type shrinkage in a first-stage high-dimensional Normal-means problem followed by a control-function approach in the second stage. It preserves invariances of the structural instrumental variable equations.

Jann Spiess, Graduate School of Business, Stanford University. jspiess@stanford.edu First version: August 21, 2017. I thank Gary Chamberlain, Maximilian Kasy, and Jim Stock for helpful comments.
The standard two-stage least-squares (2SLS) estimator is known to be biased towards
the OLS estimator when instruments are many or weak. In a linear instrumental
variables model with one endogenous regressor, at least four instruments, and Normal
noise, I propose an estimator that combines James–Stein shrinkage in a first stage
with a second-stage control-function approach. Unlike other IV estimators based
on James–Stein shrinkage, the estimator reduces bias uniformly relative to 2SLS.
Unlike LIML, it is invariant with respect to the structural form and translation of
the target parameter.

I consider the first stage of a two-stage least-squares estimator as a high-dimensional
prediction problem, to which I apply rotation-invariant shrinkage akin to [James and
Stein (1961)]. Regressing the outcome on the resulting predicted values of the en-
dogenous regressor directly would shrink the 2SLS estimator towards zero, which
could increase or decrease bias depending on the true value of the target parameter.
Conversely, shrinking the 2SLS estimator towards the OLS estimator can reduce
risk [Hansen (2017)], but increases bias towards OLS. Instead, the proposed esti-
mator uses the first-stage residuals as controls in the second-stage regression of the
outcome on the endogenous regressor. If no shrinkage is applied, the 2SLS estimator
is obtained as a special case, while a variant of [James and Stein (1961)] shrinkage
that never fully shrinks to zero uniformly reduces bias, but increases variation.

The proposed estimator is invariant to a group of transformations that include
translation in the target parameter. While the limited-information maximum likeli-
hood estimator (LIML) can be motivated rigorously as an invariant Bayes solution
to a decision problem [Chamberlain (2007)], these transformations rotate the (ap-
propriately re-parametrized) target parameter and invariance applies to a loss function
that has a non-standard form in the original parametrization. In particular, unlike
LIML, the invariance of the estimator applies to squared-error loss.

The two-stage linear model is set up in [Section 1] and transformed into a Normal-means problem in [Section 2]. [Section 3] proposes the estimator and establishes bias improvement relative to 2SLS. [Section 4] develops invariance properties of the proposed estimator. [Section 5] discusses the properties of the estimator in a simulation exercise.

1 Two-Stage Linear Regression Setup

I consider estimation of the structural parameter \( \beta \in \mathbb{R} \) in the standard two-stage linear regression model

\[
Y_i = \alpha + X_i' \beta + W_i' \gamma + U_i \\
X_i = \alpha X + Z_i' \pi + W_i' \gamma X + V_i
\]

(1)

from \( n \) iid observations \((Y_i, X_i, Z_i, W_i)\), where \( X_i \in \mathbb{R} \) is the regressor of interest (assumed univariate), \( W_i \in \mathbb{R}^k \) control variables, \( Z_i \in \mathbb{R}^\ell \) instrumental variables, and \((U_i, V_i)' \in \mathbb{R}^2\) is homoscedastic (wrt \( Z_i \)), Normal noise. \( \alpha \) is an intercept\(^1\) and \( \gamma \) and \( \pi \) are nuisance parameters. This model could be motivated by a latent variable present in both outcome and first-stage equation under appropriate exclusion restrictions as in [Chamberlain (2007)]\(^2\). For the noise I use the notation

\[
\begin{pmatrix}
U_i \\
V_i
\end{pmatrix} | Z_i = z_i, W_i = w_i \sim \mathcal{N}
\begin{pmatrix}
0_2, \\
\begin{pmatrix}
\sigma^2 & \rho \sigma \tau \\
\rho \sigma \tau & \tau^2
\end{pmatrix}
\end{pmatrix}
\]

\(^1\)We could alternatively include a constant regressor in \( X_i \) and subsume \( \alpha \) in \( \beta \). I choose to treat \( \alpha \) separately since I will focus on the loss in estimating \( \beta \), ignoring the performance in recovering the intercept \( \alpha \).

\(^2\)In this section, the intercepts \( \alpha, \alpha X \) could be subsumed in the control coefficients \( \gamma, \gamma X \) without loss, but I maintain this notation to keep it consistent.
for some $\rho \in (-1, 1)$. In vector notation, the reduced form is

$$
\begin{pmatrix}
Y \\
X
\end{pmatrix} | Z = z, W = w \sim \mathcal{N}
\left(
\begin{pmatrix}
\alpha + z\pi_Y + w\gamma_Y \\
\alpha_X + z\pi + w\gamma_X
\end{pmatrix}, \Sigma \otimes \mathbb{1}_{2n}
\right)
$$

with

$$
\pi_Y = \pi\beta, \quad \gamma_Y = \gamma + \gamma_X\beta,
\Sigma = \begin{pmatrix}
\sigma^2 + 2\rho\beta\sigma\tau + \beta^2\tau^2 & \rho\sigma\tau + \beta\tau^2 \\
\rho\sigma\tau + \beta\tau^2 & \tau^2
\end{pmatrix}.
$$

Note that there is a one-two-one mapping between reduced-form and structural-form parameters provided that the proportionality restriction $\pi_Y = \pi\beta$ holds. I develop a natural many-means form directly from the structural model, which is thus without loss, yet particularly convenient for my analysis. Throughout, our interest will be in estimating $\beta$ for many instruments (large $\ell$).

2 A Convenient Normal-Means Transformation

Having set up the two-stage linear regression model, I first derive the Normal-means representation (given $Z=z, W=w$)

$$
X^* \sim \mathcal{N}
\left(
\begin{pmatrix}
\mu \\
0_s
\end{pmatrix}, \tau^2\mathbb{1}_{\ell+s}
\right)
$$

$$
Y^*|X^* = x^* \sim \mathcal{N}
\left(
\begin{pmatrix}
x^*\beta + \left(x^* - \begin{pmatrix}
\mu \\
0_s
\end{pmatrix}\right)\frac{\rho\sigma}{\tau}, (1 - \rho^2)\sigma^2\mathbb{1}_{\ell+s}
\end{pmatrix}, \begin{pmatrix}
\sigma^2 + 2\rho\beta\sigma\tau + \beta^2\tau^2 & \rho\sigma\tau + \beta\tau^2 \\
\rho\sigma\tau + \beta\tau^2 & \tau^2
\end{pmatrix}
\right)
$$

3Throughout this document, I write upper-case letters for random variables (such as $Y_i$) and lower-case letters for fixed values (such as when I condition on $X_i = x_i$). When I suppress indices, I refer to the associated vector or matrix of observations, e.g. $Y \in \mathbb{R}^n$ is the vector of outcome variables $Y_i$ and $X \in \mathbb{R}^{n \times m}$ is the matrix with rows $X_i$. 
as follows: Assuming that \((1, w, z)\) has full rank \(1 + k + \ell \leq n - 1\), let \(q = (q_1, q_w, q_z, q_r) \in \mathbb{R}^{n \times n}\) orthonormal where \(q_1 \in \mathbb{R}^n, q_w \in \mathbb{R}^{n \times k}, q_z \in \mathbb{R}^{n \times \ell}\) such that 1 is in the linear subspace of \(\mathbb{R}^n\) spanned by \(q_1 \in \mathbb{R}^n\) (that is, \(q_1 \in \{1/n, -1/n\}\)), the columns of \((1, w)\) are in the space spanned by the columns of \((q_1, q_w)\), and the columns of \((1, w, z)\) are in the space spanned by the columns of \((q_1, q_w, q_z)\). (Such a basis exists, for example, by an iterated singular value decomposition.) Using

\[X^* = \begin{pmatrix} X_z^* \\ X_r^* \end{pmatrix} = \begin{pmatrix} q_z' X \\ q_r' X \end{pmatrix}, \quad Y^* = \begin{pmatrix} Y_z^* \\ Y_r^* \end{pmatrix} = \begin{pmatrix} q_z' Y \\ q_r' Y \end{pmatrix}, \quad \mu = q_z' z \pi,\]

and \(s = n - 1 - k - \ell\).

In this simplified representation, the linear least-squares (OLS) and two-stage least-squares (2SLS) estimators are

\[\hat{\beta}_{\text{OLS}} = \frac{(Y^*)' X^*}{(X^*)' X^*}, \quad \hat{\beta}_{\text{2SLS}} = \frac{(Y_z^*)' X_z^*}{(X_z^*)' X_z^*}.\]

The OLS estimator is biased for \(\beta\) if \(\rho \neq 0\), while the 2SLS estimator is still biased towards OLS due to overfitting in the first stage.

### 3 Control-Function Shrinkage Estimator

An equivalent way of obtaining the two-stage least-square estimator in the linear IV model is the control-function approach which estimates residuals \(V_i\) linearly by \(\hat{V}_i\) in the first stage, and then controls for \(\hat{V}_i\) in the second stage. This corresponds to controlling for the estimate \(X^* - \begin{pmatrix} X_z^* \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ X_r^* \end{pmatrix}\) of \(X^* - \begin{pmatrix} \mu \\ 0 \end{pmatrix}\) in the regression of \(Y^*\) on \(X^*\) in Equation 2. But the control-function approach is not limited to the OLS solution. In general, given an estimator \(\hat{\mu} = \hat{\mu}(X^*)\) of \(\mu\), a feasible implied
estimator for $\beta$ in Equation 2 is the coefficient on $X^*$ in a linear regression of $Y^*$ on $X^*$ and the control function $X^* - \left( \frac{\hat{\mu}}{\hat{\theta}_x} \right)$.

For high-dimensional $\mu$, a natural estimator for $\mu$ is a shrinkage estimator of the form $\hat{\mu}(X^*) = c(X^*) X^*_z$ with scalar $c(X^*)$, where the scalar may be chosen to trade off bias and variance in the estimation of $\mu$. This class contains the linear-regression solution at $c(X^*) = 1$, yielding 2SLS. The conditional bias of the implied control-function estimator $\hat{\beta}$ in Equation 2 with control function $X^* - \left( \frac{c(X^*) X^*_z}{\hat{\theta}_x} \right)$ takes a particularly simple form for this class of estimators:

**Lemma 1 (Conditional bias of CF–shrinkage estimators).** For $x^* \in \mathbb{R}^{\ell+s}$ with $c(x^*) \neq 0$, the shrinkage control-function IV estimator $\beta$ has conditional bias

$$
E[\hat{\beta}|X^* = x^*] - \beta = E \left[ \frac{\hat{\mu}'(\hat{\mu} - \mu)}{\hat{\mu}'} X^* = x^* \right] \frac{\rho \sigma}{\tau} = \left( 1 - \frac{1}{c(x^*)} \frac{(x^*_z)'(x^*_z)}{(x^*_z)'x^*_z} \right) \frac{\rho \sigma}{\tau}.
$$

Shrinkage in the James and Stein (1961) estimator (for unknown $\tau^2$) takes the form $c(x^*) = 1 - \frac{\|x^*_r\|^2}{\|x^*_z\|^2}$. This shrinkage pattern (and its positive-part variant) is unappealing here, as it can cross zero, around which point the estimator diverges. A natural variant that mitigates this problem is

$$
c(x^*) = \frac{1}{1 + p\frac{\|x^*_r\|^2}{\|x^*_z\|^2}} = \frac{\|x^*_z\|^2}{\|x^*_z\|^2 + p\|x^*_r\|^2}, \quad (3)
$$

which behaves as $1 - p\frac{\|x^*_r\|^2}{\|x^*_z\|^2}$ for small $p\frac{\|x^*_r\|^2}{\|x^*_z\|^2}$, but never quite reaches zero. Indeed, this specific choice of shrinkage yields bias dominance:

**Theorem 1 (Bias dominance through shrinkage).** Assume that $\ell \geq 4$ and $p \in (0, 2\frac{\ell-2}{s})$ in Equation 3. Then $|E[\beta]| < |E[\hat{\beta}^{2SLS}]| < |E[\hat{\beta}]|$ provided $p \neq 0$ and $\|\mu\| \neq 0$ (otherwise equality) for the control-function estimator.
\( \hat{\beta} \) in [Equation 2] with control function \( X^* - \left( c(X^*) X^*_0 \right) \).

The requirement \( \ell \geq 4 \) is an artifact of this specific shrinkage pattern and dominance should extend to \( \ell = 3 \) for an appropriate modification.

4 Invariance Properties

The estimator \( \hat{\beta} \) developed in the previous section has invariance properties in a decision problem, where in spirit and notation I follow the treatment of LIML in [Chamberlain (2007)].

First I fix the sample and action spaces, as well as a class of loss functions, for the decision problem of estimating \( \beta \). Starting with [Equation 2] I write \( \mathcal{Z} = (\mathbb{R}^{\ell+s})^2 \) for the sample space from which \((X^*, Y^*)\) is drawn according to \( P_\theta \), where I parametrize \( \theta = (\beta, \mu, \rho, \sigma, \tau) \in \Theta = \mathbb{R} \times \mathbb{R}^\ell \times \mathbb{R}_0^3 \). The action space is \( A = \mathbb{R} \), from which an estimate of \( \beta \) is chosen. I assume that the loss function \( L : \Theta \times A \to \mathbb{R} \) can be written as \( L(\theta, a) = \ell(a - \beta) \) for some sufficiently well-behaved \( \ell : \mathbb{R} \to \mathbb{R} \) (such as squared-error loss \( L(\theta, a) = (a - \theta)^2 \)). The estimator \( \hat{\beta} : \mathcal{Z} \to A \) from the previous section is a feasible decision rule in this decision problem.

For an element \( g = (g_\beta, g_z, g_r) \) in the (product) group \( G = \mathbb{R} \times O(\ell) \times O(s) \), where \( \mathbb{R} \) denotes the group of real numbers with addition (neutral element 0) and \( O(\ell) \) the group of ortho-normal matrices in \( \mathbb{R}^{\ell \times \ell} \) with matrix multiplication (neutral element \( I_\ell \)), consider the following set of transformations (which are actions of \( G \) on \( \mathcal{Z}, \Theta, A \)):

- Sample space: \( m_\mathcal{Z} : G \times \mathcal{Z} \to \mathcal{Z}, \)

\[
(g, (x^*, y^*)) \mapsto \left( \begin{pmatrix} g_z & \emptyset \\ \emptyset & g_r \end{pmatrix} x^*, \begin{pmatrix} g_z & \emptyset \\ \emptyset & g_r \end{pmatrix} (y^* + g_\beta x^*) \right)
\]
- Parameter space: $m_\Theta : G \times \Theta \to \Theta,$

\[(g, \theta) \mapsto (\beta + g_\beta, g_z\mu, \rho, \sigma, \tau)\]

- Action space: $m_A : G \times \mathcal{A} \to \mathcal{A}, (g, a) \mapsto a + g_\beta$

These transformations include translation of the target parameter, as well as rotations of instruments and residuals. They are tied together by leaving model and loss invariant. Indeed, the following result is immediate from Equation 2:

**Proposition 1** (Invariance of model and loss).

1. The model is invariant: $m_Z(g, (X^*, Y^*)) \sim P_{m_\Theta(g, \theta)}$ for all $g \in G.$

2. The loss is invariant: $L(m_\Theta(g, \theta), m_A(g, a)) = L(\theta, a)$ for all $g \in G.$

Since loss and model are invariant to this set of transformations, invariant decision rules are particularly appealing solutions to the estimation problem. A decision rule $d : Z \to \mathcal{A}$ is invariant if, for all $(g, (x^*, y^*)) \in G \times Z,$ $d(m_Z(g, (x^*, y^*))) = m_A(g, d((x^*, y^*))).$ The estimator $\hat{\beta}$ above is included in a class of invariant decision rules:

**Proposition 2** (Invariance of a class of control-function estimators). Consider a control-function decision rule $d((x^*, y^*))$ obtained as the coefficient on $x^*$ in a linear regression of $y^*$ on $x^*$, controlling for $x^* - (c(\|x^*_z\|, \|x^*_r\|)(x^*_z)'0)'$, where $c(\|x^*_z\|, \|x^*_r\|)$ scalar (and measurable). Then $d$ is an invariant decision rule with respect to the above actions of $G.$

Unlike LIML, the shrinkage estimator in Theorem 1 is therefore invariant to translations in the target parameter $\beta.$
In this section, I study the performance of the shrinkage estimator introduced in Section 3 in a simulation exercise. I generate data according to Equation 1 without control variables $W_i$ ($k = 0$), where I normalize the target parameter to $\beta = 1$, the variance of both error terms to one, and set their correlation to $\rho = .5$. The $Z_i$ are drawn independently from a multivariate standard Normal distribution. I fix the sample size to $n = 60$. I vary the size $\|\pi\| \in \{0.5, 1.0\}$ of the first-stage parameter, as well as the number $\ell \in \{5, 10, 20\}$ of instruments. On this data, I compare the performance of estimates of $\beta$ from OLS regression ($Y_i$ on $X_i$), two-stage least squares (2SLS), and the IV shrinkage estimator introduced in Section 3 (SIV). For each parameter setting and each estimator I obtain bias, median bias, standard deviation (SD), and root mean-squared error (RMSE) from 100,000 Monte-Carlo draws.

Table 1 reports the results of the simulation exercise. The two-stage least-squares estimator is biased towards the OLS estimator, which has positive bias. As predicted by the theory, the shrinkage estimator persistently reduces the bias, with higher gains when the control coefficient is small or its dimension high. In the simulation, this pattern carries over to the median bias. Unsurprisingly, the variance of estimates increases, with an ambiguous effect on overall mean-squared error (MSE): while the MSE is consistently below OLS, MSE improves over two-stage least squares only for many instruments.

Figure 1 plots the corresponding Kernel density estimates. The picture shows that the estimator does what it sets out to do – reducing bias – at the cost of increasing variance: While the location of the distribution shifts towards the truth, $\beta = 1$, the spread of estimates increases.
Table 1: Performance in estimating \( \beta \) from 100,000 Monte-Carlo draws for linear least-squares (“OLS”), two-stage least squares (“2SLS”), and IV shrinkage estimator (“SIV”): bias, median bias, standard deviation (“SD”), and root mean-squared error (“RMSE”).

Conclusion

An application of James–Stein shrinkage to instrumental variables in a Normal-means transformation consistently reduces bias, which comes at the cost of an increase in variance. The specific estimator is invariant to a group of transformations of the structural form that involves translation of the target parameter.

This work relates to a recent literature that uses sparse and shrinkage estimators in high-dimensional IV problems (e.g. [Belloni et al. 2014] [Hansen and Kozbur 2014]). Relative to this direction of the literature, I employ a control-function approach and provide finite-sample results. An interesting direction for future research would be to bring machine-learning approaches and control functions together.
Estimate by Dim
OLS
TSLS
JSIV
Method
20105
0.0 0.5 1.0 1.5 2.0
0 0.5 1.0 1.5 2.0
0 0.5 1.0 1.5 2.0
0 1 2 3 4

Figure 1: Kernel-density estimates of the distribution of estimates of $\beta = 1$ from 100,000 Monte-Carlo draws for linear least-squares (grey), two-stage least squares (red), and IV shrinkage estimator (blue) for varying dimension $\ell$.

In a companion article (Spiess 2017), I show how analogous shrinkage in at least three control variables provides consistent loss improvement over the least-squares estimator without introducing bias, provided that treatment is assigned randomly. Together, these results suggests different roles of overfitting in instrumental variable and control coefficients, respectively: while overfitting to instrumental variables in the first stage of a two-stage least-squares procedure induces bias, overfitting to control variables induces variance.

REFERENCES

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2):608–650.
Chamberlain, G. (2007). Decision Theory Applied to an Instrumental Variables Model. *Econometrica*, 75(3):609–652.

Hansen, B. E. (2017). Stein-like 2SLS estimator. *Econometric Reviews*, 36(6-9):840–852.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Fourth Berkeley Symposium*.

Moser, S. M. (2008). Expectations of a noncentral chi-square distribution with application to IID MIMO Gaussian fading. In *International Symposium on Information Theory and Its Applications*. IEEE.

Piegorsch, W. W. and Casella, G. (1985). The existence of the first negative moment. *The American Statistician*, 39(1):60.

Spiess, J. (2017). Unbiased Shrinkage Estimation. *arXiv preprint arXiv:1708.06436*.

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151.
**Appendix**

*Proof of Theorem 1.* For the (rescaled) bias, where $\lambda = ps$ and $M = X_t^*$, we have by Lemma 1 that

$$B(\lambda) = \frac{\tau}{\rho\sigma} E[\hat{\beta} - \beta] = E \left[ 1 - \frac{\|X_t^*\|^2 + p\|X_t^*\|^2 (X_t^*)'\mu}{\|X_t^*\|^2} \left(\frac{X_t^*}{X_t^*}'X_t^*\right) \right]$$

$$= E \left[ 1 - \frac{M'\mu}{\|M\|^2} - \lambda \tau^2 \frac{M'\mu}{\|M\|^4} \right],$$

provided that $E \left| 1 - \frac{M'\mu}{\|M\|^2} - \lambda \tau^2 \frac{M'\mu}{\|M\|^4} \right| < \infty$. By the multi-dimensional version of Stein’s (1981) lemma for $h(M) = \frac{1}{\|M\|^2}$,

$$-2\tau^2 E \left[ \frac{M}{\|M\|^4} \right] = \tau^2 E[\nabla h(M)] = E[(M - \mu)h(M)] = E \left[ \frac{M - \mu}{\|M\|^2} \right],$$

again provided that all moments exist.

For the existence of moments, note that by Cauchy–Schwarz and Jensen it suffices to consider $E\|M/\|M\|^4\| = E[\|M\|^{-3}]$. To establish that this expectation is finite, note that the distribution of $\|M\|^2/\tau^2$, a non-central $\chi^2$ distribution with $\ell$ degrees of freedom and non-centrality parameter $\|\mu\|^2/\tau^2$, is first-order stochastically dominating a central $\chi^2$ distribution with $\ell$ degrees of freedom, so it is sufficient to establish $E[(X^2)^{-3/2}] < \infty$ where $X^2$ has a central $\chi^2$ distribution with $\ell$ degrees of freedom. Now, the density $f(y)$ of $(X^2)^{3/2}$ is proportional to $y^{\ell/3 - 1} \exp(-y^{2/3}/2)$, implying $\lim_{y \to 0} f(y)/y^\alpha = 0$ for $\ell \geq 4$ and, say, $\alpha = 1/4 > 0$. The existence of the inverse moment, i.e. $E[(X^2)^{-3/2}] < \infty$, follows by Piegorsch and Casella (1985).
We thus have
\[
E \left[ \frac{M'\mu}{\|M\|^2} \right] = -\frac{1}{2\tau^2} E \left[ \frac{(M - \mu)'\mu}{\|M\|^2} \right],
\]
which yields
\[
B(\lambda) = E \left[ 1 - \frac{M'\mu}{\|M\|^2} + \frac{\lambda (M - \mu)'\mu}{2\|M\|^2} \right]
= E \left[ 1 - \frac{\|M\|^2 - (M - \mu)'M}{\|M\|^2} + \frac{\lambda \|M\|^2 - \|\mu\|^2 - (M - \mu)'M}{2\|M\|^2} \right]
= \frac{\lambda}{2} - \frac{\lambda}{2} E \left[ \frac{\|\mu\|^2}{\|M\|^2} \right] - \frac{\lambda}{2} E \left[ \frac{(M - \mu)'M}{\|M\|^2} \right].
\]

Denote by $K$ a Poisson random variable with mean $\kappa = \frac{\|\mu\|^2}{2\tau^2} > 0$. ($B(\lambda)$ is constant at 1 for $\|\mu\| = 0$, and there remains nothing to show.) From James and Stein (1961, (9), (16)) we have that
\[
E \left[ \frac{\|\mu\|^2}{\|M\|^2} \right] = E \left[ \frac{2\kappa}{\ell - 2 + 2K} \right] = Q(\ell),
E \left[ \frac{(M - \mu)'M}{\|M\|^2} \right] = E \left[ \frac{\ell - 2}{\ell - 2 + 2K} \right] = P(\ell).
\]
It immediately follows from
\[
B(\lambda) = P(\ell) - \frac{\lambda}{2} (P(\ell) + Q(\ell) - 1)
\]
that the bias for the unshrunk reference estimator ($\lambda = 0$, 2SLS) is $B(0) = P(\ell) > 0$, and that $B(\lambda)$ is decreasing in $\lambda$ since $P(\ell) + Q(\ell) \geq 1$ by Jensen’s inequality (with strict inequality unless $\|\mu\| = 0$). The (infeasible) bias-minimizing choice of $\lambda$ is
given by

\[ \lambda^* = \frac{2P(\ell)}{P(\ell) + Q(\ell) - 1} = \frac{\ell - 2}{\frac{\ell - 2}{2} + \kappa - 1 / \mathbb{E}[(\ell - 2)/2 + K)^{-1}]. \]

To conclude the proof, I assert (and prove below) that, for any \( a \geq 1 \),

\[ \mathbb{E}[(a + K)^{-1}] \leq \frac{1}{a + \nu - 1}. \] (4)

With \( a = \frac{\ell - 2}{2} \) it follows that \( \frac{\ell - 2}{2} + \kappa - 1 / \mathbb{E}[(\ell - 2)/2 + K)^{-1}] \leq 1 \) and thus \( \lambda^* \geq \ell - 2 \).

We obtain \(|B(\lambda)| \leq |B(0)|\) (dominance over 2SLS in terms of bias) for all \( \lambda \in (0, \ell - 2) \) by strict monotonicity of \( B(\lambda) \), which yields the theorem.

To establish Equation 4 fix \( a \in \mathbb{R} \) with \( a \geq 1 \) and note that for \( K \text{ Poisson with parameter } \nu\)

\[ \mathbb{E}\left[ \frac{\nu}{a + K}\right] = \sum_{i=0}^{\infty} \frac{\nu}{a + i} \frac{\nu^i \exp(-\nu)}{i!} = \sum_{i=0}^{\infty} \frac{\nu^i (i + 1) \exp(-\nu)}{(i + 1)!} \]

\[ = \sum_{i=1}^{\infty} \frac{\nu^i \exp(-\nu)}{a + i - 1}. \]

For \( a = 1 \), thus \( \mathbb{E}\left[ \frac{\nu}{a + K}\right] = 1 - \exp(-\nu) \leq 1 \). For \( a > 1 \),

\[ \mathbb{E}\left[ \frac{\nu}{a + K}\right] = \sum_{i=0}^{\infty} \frac{\nu}{a + i - 1} \frac{\nu^i \exp(-\nu)}{i!} = \mathbb{E}\left[ \frac{K}{a + K - 1}\right] \leq \frac{\nu}{a + \nu - 1} \]

by Jensen’s inequality applied to the concave function \( x \mapsto \frac{x}{a - 1 + x} (x \geq 0) \). In both cases, Equation 4 follows by dividing by \( \nu \), yielding a generalization of an inequality in Moser (2008, Theorem 6) to non-integer \( a \).

Proof of Proposition 2 Fix \((x, y) \in \mathcal{Z}\) and consider \( d((x, y)) \). Note first that \( c = \)
\( c(\|x_z\|, \|x_r\|) \) is invariant to the action of \( G \) on \( Z \). The decision rule is

\[
d((x, y)) = \frac{x'a(x)y}{x'a(x)x}
\]

where

\[
a(x) = I - b(x)(b(x)'b(x))^{-1}b(x)' \text{ for } b(x) = \begin{pmatrix} (1 - c)x_z \\ x_r \end{pmatrix}.
\]

Now for any \( g \in G \), where I write \( q_g = \begin{pmatrix} g_z & 0 \\ 0 & g_r \end{pmatrix} \), we have \( b(q_gx) = q_gb(x) \) and thus \( a(q_gx) = q_ga(x)q_g' \). It is immediate that

\[
d(m_Z(g, (x, y))) = d((q_gx, q_\beta g_y + g_\beta g_yx)) = \frac{x'a(x)y}{x'a(x)x} + g_\beta \frac{x'a(x)x}{x'a(x)x}
\]

\[
= d((x, y)) + g_\beta = m_A(g, d((x, y))),
\]

as claimed. \( \square \)