Shrinkage estimators for structural parameters

Tirthankar Chakravarty

UC San Diego

July 27, 2012

Introduction

> From another angle, it is possible to argue that model selection itself is a misguided goal. It is quite common to find that confidence intervals from different plausible models are non-intersecting, raising considerable inferential uncertainty. Fundamentally, the uncertainty concerning the choice of model is not reflected in conventional asymptotic and bootstrap confidence intervals.
>
> –Hansen (2005)

- We introduce `ivshrink` which is part of a trio of Stata commands, `regshrink` and `mvregshrink` which produce Stein-type shrinkage and model averaging estimators.
- We motivate `ivshrink` by considering the theoretical difficulty of uniformly consistent post-model selection inference. Model averaging instead.
- Shrinkage estimators are also well-known to have better risk properties, at the cost of potentially introducing bias.
- `ivshrink` is work in progress, theoretically and Stata-wise. Comments welcome.

The classical linear simultaneous equations model I

- The classical linear simultaneous equations model [CLSEM] which underlies the instrumental variables estimators and inference procedures can be formulated as

$$\begin{aligned} \Upsilon_1 &= \mathbf{Y}_2\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\delta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \end{aligned}$$

for which system, the reduced form for the endogenous variables, in terms of the system exogenous variables $\mathbf{Z} = [\mathbf{Z}_1\mathbf{Z}_2]$ is

$$[\Upsilon_1\mathbf{Y}_2] = \mathbf{Z}\left[\boldsymbol{\pi}_1\boldsymbol{\pi}_2\right] + [\boldsymbol{\vartheta}_1\boldsymbol{\vartheta}_2]$$

In the classical case, a (matrix-)normality assumption is made on the reduced form errors

$$[\boldsymbol{\vartheta}_1\boldsymbol{\vartheta}_2] \sim \mathrm{MN}\left(\mathbf{O}, \boldsymbol{\omega}, \boldsymbol{\iota}_N\right)$$

where

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_{11} & \boldsymbol{\omega}_{12}^{\mathsf{T}} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\omega}_{22} \end{bmatrix}$$

The classical linear simultaneous equations model II

We can specify the (homoskedastic) error structure of the structural error vector in terms of the reduced form covariance matrix,

$$V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\iota}_N$$

where

$$\sigma^2 = \omega_{11} - 2\boldsymbol{\omega}_{12}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\omega}_{22}\boldsymbol{\beta}$$

- The relevance and validity assumptions are

$$\mathbb{E}\left(\mathbf{Z}^{\mathsf{T}}\left[\boldsymbol{\vartheta}_1\boldsymbol{\vartheta}_2\right]\right) = \mathbf{O}$$
$$\mathbb{E}\left(\mathbf{Z}^{\mathsf{T}}\mathbf{Y}_2\right) \neq \mathbf{O}$$

- The dimensions of the various objects

N : sample size
G_1 : number of endogenous explanatory variables
K_1 : number of included exogenous variables
K_2 : number of excluded exogenous variables
K : $K_1 + K_2$, the total number of exogenous variables in the system

Basic estimators I

- The k-class estimator, $\widehat{\gamma}(k)$ of Theil (1958) depends on the choice of a tuning parameter $k \in \mathbb{R}$, is given as the solution to the system of linear equations

$$\begin{bmatrix} \mathbf{Y}_2^\top - k\widehat{\vartheta}_2 \\ \mathbf{Z}_1^\top \end{bmatrix} Y_1 = \begin{bmatrix} \mathbf{Y}_2^\top \mathbf{Y}_2 - k\widehat{\vartheta}_2^\top \widehat{\vartheta}_2 & \mathbf{Y}_2^\top \mathbf{Z}_1 \\ \mathbf{Z}_1^\top \mathbf{Y}_2 & \mathbf{Z}_1^\top \mathbf{Z}_1 \end{bmatrix} \begin{bmatrix} \widehat{\beta}(k) \\ \widehat{\delta}(k) \end{bmatrix}$$

Note that the k-class estimator contains the usual estimators, including OLS ($k = 0$), 2SLS ($k = 1$), LIML ($k = \lambda_0$), where

$$\lambda_0 = \min_{\beta} \frac{(Y_1 - \mathbf{Y}_2\beta)^\top \mathbf{M}_{\mathbf{Z}_1} (Y_1 - \mathbf{Y}_2\beta)}{(Y_1 - \mathbf{Y}_2\beta)^\top \mathbf{M}_{\mathbf{Z}} (Y_1 - \mathbf{Y}_2\beta)}$$

The k-class estimator is the basic combination estimator since it is continuous in the parameter k, and every estimator between the OLS and the 2SLS estimator can be obtained as some choice of $k \in [0, 1]$.

Basic estimators II

- The k-class estimator has been generalized to the double k-class estimator by Nagar (1962), and can be written as

$$\begin{bmatrix} \mathbf{Y}_2^\top - k_2^* \widehat{\boldsymbol{\vartheta}}_2 \\ \mathbf{Z}_1^\top \end{bmatrix} Y_1 = \begin{bmatrix} \mathbf{Y}_2^\top \mathbf{Y}_2 - k_1^* \widehat{\boldsymbol{\vartheta}}_2^\top \widehat{\boldsymbol{\vartheta}}_2 & \mathbf{Y}_2^\top \mathbf{Z}_1 \\ \mathbf{Z}_1^\top \mathbf{Y}_2 & \mathbf{Z}_1^\top \mathbf{Z}_1 \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}}(\boldsymbol{k}^*) \\ \widehat{\boldsymbol{\delta}}(\boldsymbol{k}^*) \end{bmatrix}$$

  where

$$\boldsymbol{k}^* = \begin{bmatrix} k_1^* \\ k_2^* \end{bmatrix} = \begin{bmatrix} k_1 + \zeta(\lambda_0 + 1) \\ k_2 + \zeta(\lambda_0 + 1) \end{bmatrix}$$

  where λ_0 is the LIML eigenvalue. In the case that $\zeta = 0$, the tuning parameters of the estimators are fixed, else, they are sample dependent.

- Most of the estimators implemented in `ivshrink` are implemented using the function returning the double-k class estimator.

Classical combination estimators I

- In the 70s and 80s estimators with better finite sample properties – (almost-)unbiased, minimum MSE, maximum concentration probability – were proposed;
- these estimators tended to have the form of non-stochastic combinations of classical IV estimators;
- a nice unified treatment of the theoretical properties of these estimators is given in Anderson et al. (1986);
- To introduce some of these estimators, we define the class of **finite linear combination estimators**

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}(\boldsymbol{k}, \boldsymbol{w}) \\ \widehat{\boldsymbol{\delta}}(\boldsymbol{k}, \boldsymbol{w}) \end{bmatrix} = \sum_{h=1}^{3} w_h \begin{bmatrix} \widehat{\boldsymbol{\beta}}(k_h) \\ \widehat{\boldsymbol{\delta}}(k_h) \end{bmatrix}$$

  such that

$$\sum_{h=1}^{3} w_h = 1$$

- Using this representation, some examples of *almost unbiased estimators* are given in table 1

Classical combination estimators II

Estimator	weights			k		
	w_1	w_2	w_3	k_1	k_2	k_3
Sawa (1973)	0	$1+\frac{K_2-G_1-1}{N-K}$	$-\frac{K_2-G_1-1}{N-K}$	$\cdot$	1	0
Morimune (1978)	$\frac{K_2-G_1-2}{K_2-G_1+2}$	$\frac{4}{K_2-G_1+2}$	0	λ_0	1	$\cdot$
	$\frac{N-K_1-G_1-1}{N-K_1-G_1}$	0	$\frac{1}{N-K_1-G_1}$	$\cdot$	1	0
Nagar (1959)	1	0	0	$1+\frac{K_2-G_1-1}{N}$	$\cdot$	$\cdot$
Kadane (1971)	1	0	0	$1+\frac{K_2-G_1-1}{N-K_1-K_2}$	$\cdot$	$\cdot$
Fuller (1977)	1	0	0	$\lambda_0+\frac{1}{N-K_1-K_2}$	$\cdot$	$\cdot$

Table: Almost unbiased combination estimators

Stein-type estimators I

- The estimators above are not shrinkage estimators in the sense of Stein (1956).
- Such a class of shrinkage estimators was introduced by Zellner and Vandaele (1974), and the following double g-class estimator of Ullah and Srivastava (1988) is general in this class

$$\widehat{\beta}_U(g_1, g_2) = \left(1 - \frac{g_1 \widehat{\boldsymbol{\vartheta}}_1^\top \widehat{\boldsymbol{\vartheta}}_1}{\mathbf{Y}_1^\top \mathbf{Y}_1 - g_2 \widehat{\boldsymbol{\vartheta}}_1^\top \widehat{\boldsymbol{\vartheta}}_1}\right) \widehat{\beta}_{2SLS}$$

where

$$\widehat{\boldsymbol{\vartheta}}_1 = \mathbf{M_Z Y}_1$$

and the optimal values of g_1 and g_2 lead to the estimator $\widehat{\beta}_U\left(\frac{1}{N-G_1-K_1}, \frac{N-G_1-K_1-1}{N-G_1-K_1}\right)$

Post-model selection size distortion I

- Well known problem that post-model selection estimators do not take the model selection uncertainty into account.
- Literature recently revived in Leeb and Pötscher (2005); Guggenberger (2010); Andrews and Guggenberger (2009), among others.
- We define the asymptotic size of the test based on a test statistic T_N under the null as

$$\text{asy. size}\left(T_N, \gamma_0\right) = \limsup_{N\to\infty} \sup_{\lambda\in\Lambda} \mathbb{P}_{\gamma_0,\lambda}\left[T_N(\gamma_0) > c_{1-\alpha}\right]$$

  where α is the nominal size of the test, $T_n(\gamma_0)$ is the test statistic and $c_{1-\alpha}$ is the critical value of the test.
- Uniformity over $\boldsymbol{\lambda} \in \Lambda$ which is built in to the definition of the asymptotic size of the test is crucial for the asymptotic size to give a good approximation for the finite sample size.
- We recall that the Hausman statistic is given as

$$H_N = N\left(\widehat{\gamma}_{2SLS} - \widehat{\gamma}_{OLS}\right)^{\mathsf{T}}\left(V\left(\widehat{\gamma}_{2SLS} - \widehat{\gamma}_{OLS}\right)\right)^{-1}\left(\widehat{\gamma}_{2SLS} - \widehat{\gamma}_{OLS}\right)$$

Post-model selection size distortion II

- The vector of t-statistics in this context is given by

$$\boldsymbol{T}_N(\widehat{\gamma}, \gamma_0) = \sqrt{N}\frac{\widehat{\gamma} - \gamma_0}{\text{s.e.}\,\widehat{\gamma}}$$

for the appropriate estimator $\widehat{\gamma}$. Also, define

$$\boldsymbol{T}^*_N(\widehat{\gamma}_{2SLS}, \widehat{\gamma}_{OLS}, \gamma_0) = \boldsymbol{T}_N(\widehat{\gamma}_{2SLS}, \gamma_0)1_{\left[H_N > \xi_{\chi^2_1, 1-\alpha_a}\right]} + \boldsymbol{T}_N(\widehat{\gamma}_{OLS}, \gamma_0)1_{\left[H_N \le \xi_{\chi^2_1, 1-\alpha_a}\right]}$$

- The two-stage vector of t-statistics is then given by

$$\boldsymbol{T}^\dagger_N(\widehat{\gamma}_{2SLS}, \widehat{\gamma}_{OLS}, \gamma_0) = \begin{cases} +\boldsymbol{T}^*_N(\widehat{\gamma}_{2SLS}, \widehat{\gamma}_{OLS}, \gamma_0) & \text{if upper one-sided test} \\ -\boldsymbol{T}^*_N(\widehat{\gamma}_{2SLS}, \widehat{\gamma}_{OLS}, \gamma_0) & \text{if lower one-sided test} \\ \left|\boldsymbol{T}^*_N(\widehat{\gamma}_{2SLS}, \widehat{\gamma}_{OLS}, \gamma_0)\right| & \text{if symmetric two-sided test} \end{cases}$$

Post-model selection size distortion III

- Then, in order to be nominal size α_b standard fixed critical value [FCV] test, it must be that the test rejects the null hypothesis if

$$T_N^{\dagger}(\widehat{\gamma}_{2SLS}, \widehat{\gamma}_{OLS}, \gamma_0) > c_{\infty,1-\alpha_b}$$

where

$$c_{\infty,1-\alpha_b} = \begin{cases} \xi_{\Phi,1-\alpha_b} & \text{if lower one-sided test} \\ \xi_{\Phi,1-\alpha_b} & \text{if upper one-sided test} \\ \xi_{\Phi,1-\alpha_b/2} & \text{if two-sided test} \end{cases}$$

- The Hausman test does not have good power to detect local deviations from exogeneity, however, the OLS bias picks up these exogeneity deviations strongly, rejecting the second stage null and leading to over-sized test.
- Simulation results confirm the theoretical findings are reported in fu Wong (1997); Guggenberger (2010).
- Recent research McCloskey (2012); Cornea (2011) has suggested some ways of trying to recover the asymptotic size without completely sacrificing power.

Large sample Stein-type estimators I

- Kim and White (2001) provide shrinkage type estimators where a base (unbiased) estimator is shrunk towards another, possibly biased and correlated estimator using stochastic or non-stochastic weights.
- Under a wide variety of regularity conditions, estimators for parameters γ of a model are (jointly) asymptotically normally distributed. Consider specifically

$$\sqrt{N}\begin{bmatrix}\widehat{\gamma}_{2SLS,N}-\gamma_0\\ \widehat{\gamma}_{OLS,N}-\gamma_0\end{bmatrix}\xrightarrow{d}\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix}\sim \mathrm{N}\left(\boldsymbol{\xi},\boldsymbol{\sigma}\right)$$

Allowing for one of the estimators to be asymptotically biased leads to

$$\boldsymbol{\xi}=\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\theta}\end{bmatrix}$$

and allowing for full correlation between the estimators leads to

$$\boldsymbol{\sigma}=\begin{bmatrix}\boldsymbol{\sigma}_{11} & \boldsymbol{\sigma}_{12}\\ \boldsymbol{\sigma}_{21} & \boldsymbol{\sigma}_{22}\end{bmatrix}$$

Large sample Stein-type estimators II

- Kim and White (2001) define the natural James-Stein estimator where the shrinkage factor is random (c_1 is fixed)

$$\gamma^{JS}\left(\widehat{\gamma}_{2SLS,N}, \widehat{\gamma}_{OLS,N}; c_1\right) = \left(1 - \frac{c_1}{\|\widehat{\gamma}_{2SLS,N} - \widehat{\gamma}_{OLS,N}\|_{\mathbf{Q}_N}}\right)\left(\widehat{\gamma}_{2SLS,N} - \widehat{\gamma}_{OLS,N}\right) + \widehat{\gamma}_{OLS,N}$$

where the norm, $\|x\|_{\mathbf{Q}_N} = \boldsymbol{x}^\top \mathbf{Q}_N \boldsymbol{x}$, where $\mathbf{Q}_N$ is a norming matrix such that it converges to a non-stochastic symmetric p.d. matrix

$$\frac{\mathbf{Q}_N}{N} \xrightarrow{p} \mathbf{q}$$

Here we call the unbiased estimator $\widehat{\gamma}_{2SLS,N}$ the *base estimator* and the asymptotically biased estimator $\widehat{\gamma}_{OLS,N}$ the *data-dependent shrinkage point*.

Large sample Stein-type estimators III

- We also define the following matrices

$$\mathbf{p}\mathbf{p}^{\top} = \boldsymbol{\sigma}$$
$$Z = \mathbf{p}^{-1}\boldsymbol{U}$$
$$\sim \mathrm{N}\left(\underbrace{\mathbf{p}^{-1}\boldsymbol{\xi}}_{\equiv\mu}, \iota_{2(G_1+K_1)}\right)$$

Using these expressions, we can easily write the quadratic forms of of the standardized variates as

$$Z^{\top}\mathbf{m}_1 Z = (\boldsymbol{U}_1 - \boldsymbol{U}_2)^{\top}\,\mathbf{q}\,(\boldsymbol{U}_1 - \boldsymbol{U}_2)$$
$$Z^{\top}\mathbf{m}_2 Z = \boldsymbol{U}_1^{\top}\mathbf{q}\,(\boldsymbol{U}_1 - \boldsymbol{U}_2)$$

Large sample Stein-type estimators IV

where

$$\begin{aligned}
\mathbf{m}_1 &= \mathbf{p}^\top \mathbf{j}_1^\top \mathbf{q} \mathbf{j}_1 \mathbf{p} \\
\mathbf{j}_1 &= \begin{bmatrix} \boldsymbol{\iota}_{K_1+G_1} & -\boldsymbol{\iota}_{G_1+K_1} \end{bmatrix} \\
\mathbf{m}_2 &= \mathbf{p}^\top \mathbf{j}_2^\top \mathbf{q} \mathbf{j}_1 \mathbf{p} \\
\mathbf{j}_2 &= \begin{bmatrix} \boldsymbol{\iota}_{K_1+G_1} & \mathbf{O} \end{bmatrix}
\end{aligned}$$

- *The question is – when does the JSM estimator dominate the base and the data-dependent shrinkage point in terms of asymptotic risk and what is the optimal value of c_1?* The following theorem, which is adapted from (Kim and White, 2001, Theorem 1)

Large sample Stein-type estimators V

Theorem

Define

$$c_1^* \in \underset{c_1}{\operatorname{argmin}} \operatorname{AR}\left(\gamma^{JS}\left(\widehat{\gamma}_{2SLS,N}, \widehat{\gamma}_{OLS,N}; c_1\right)\right)$$

then it must be that

$$c_1^* = \frac{\nu}{\omega}$$

where

$$\omega = \int_0^\infty \left|\mathbf{n}_0(t)\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\mathbf{n}_1(t)\boldsymbol{\mu}\right) dt$$

and

$$\nu = \int_0^\infty \left|\mathbf{n}_0(t)\right|^{-\frac{1}{2}} \left(\operatorname{trace}\left(\mathbf{m}_2\mathbf{n}_0(t)^{-1}\right) + \boldsymbol{\mu}^\mathsf{T}\mathbf{n}_2(t)\boldsymbol{\mu}\right) \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\mathbf{n}_1(t)\boldsymbol{\mu}\right) dt$$

where, the matrix values functions $\{\mathbf{n}_0, \mathbf{n}_1, \mathbf{n}_2\} : \mathbb{R} \longmapsto \mathbb{R}^{d_2(G_1+K_1)\times d_2(G_1+K_1)}$,

$$\mathbf{n}_0(t) = \boldsymbol{\iota}_{K_1+G_1} + 2t\mathbf{m}_1$$
$$\mathbf{n}_1(t) = 2t\mathbf{m}_1\mathbf{n}_0(t)^{-1}$$
$$\mathbf{n}_2(t) = \mathbf{n}_0(t)^{-1}\mathbf{m}_2\mathbf{n}_0(t)^{-1}$$

This infeasible estimator is called the **James-Stein mix** [JSM] estimator.

Large sample Stein-type estimators VI

- We need estimates of the optimal parameters

$$\widehat{\omega} = \int_0^\infty \left|\widehat{\mathbf{n}}_{0N}(t)\right|^{-\frac{1}{2}} \, dt$$

$$\widehat{\nu} = \int_0^\infty \left|\widehat{\mathbf{n}}_{0N}(t)\right|^{-\frac{1}{2}} \operatorname{trace}\left(\widehat{\mathbf{m}}_{2N}\widehat{\mathbf{n}}_{0N}(t)^{-1}\right) dt$$

where

$$\widehat{\mathbf{n}}_{0N}(t) = \iota_{K_1+G_1} + 2t\widehat{\mathbf{m}}_{1N}$$

$$\widehat{\mathbf{m}}_{1N} = \widehat{\mathbf{P}}_N^\top \mathbf{j}_1^\top \mathbf{Q}_N \mathbf{j}_1 \widehat{\mathbf{P}}_N$$

$$\widehat{\mathbf{m}}_{2N} = \widehat{\mathbf{P}}_N^\top \mathbf{j}_2^\top \mathbf{Q}_N \mathbf{j}_1 \widehat{\mathbf{P}}_N$$

This feasible estimator is called the **James-Stein combination** [JSC] estimator.

Asymptotic covariance matrix I

- Both the estimators above belong to the so-called regular consistent second-order indexed [RCASOI] class of estimators proposed by Bates and White (1993), and as such, have a valid first order representations in terms of their scores

$$\sqrt{N}\begin{bmatrix}\widehat{\gamma}_{2SLS,N}-\gamma_0\\ \widehat{\gamma}_{OLS,N}-\gamma_0\end{bmatrix}=\begin{bmatrix}\mathbf{h}_{2SLS,N} & \mathbf{O}\\ \mathbf{O} & \mathbf{h}_{OLS,N}\end{bmatrix}^{-1}\begin{bmatrix}\frac{\sum_{i=1}^{N} s_i^{OLS}(\beta_0)}{N}\\ \frac{\sum_{i=1}^{N} s_i^{2SLS}(\beta_0)}{N}\end{bmatrix}+O_p(1)$$

where

$$s_i^{2SLS}(\boldsymbol{\beta})=\boldsymbol{X}_i\boldsymbol{Z}_i^{\top}\left(\mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}_i^{\top}\right)\right)^{-1}\boldsymbol{Z}_i\left(Y_i-\boldsymbol{X}_i^{\top}\boldsymbol{\beta}\right)$$
$$s_i^{OLS}(\boldsymbol{\beta})=\boldsymbol{X}_i\left(Y_i-\boldsymbol{X}_i^{\top}\boldsymbol{\beta}\right)$$

and

$$\mathbf{h}_{2SLS,N}=\frac{\mathbb{E}\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\right)}{N}$$
$$\mathbf{h}_{OLS,N}=\frac{\mathbb{E}\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{Z}_i^{\top}\left(\mathbb{E}\left(\boldsymbol{Z}_i\boldsymbol{Z}_i^{\top}\right)\right)^{-1}\boldsymbol{Z}_i\boldsymbol{X}_i\right)}{N}$$

Asymptotic covariance matrix II

- By application of a standard CLT (Lindeberg-Feller), we have

$$\begin{bmatrix} \frac{\sum_{i=1}^N s_i^{OLS}(\beta_0)}{N} \\ \frac{\sum_{i=1}^N s_i^{2SLS}(\beta_0)}{N} \end{bmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{bmatrix} \mathbb{E}\left(s_i^{2SLS}(\beta_0)\right) \\ \mathbb{E}\left(s_i^{OLS}(\beta_0)\right) \end{bmatrix}', \begin{bmatrix} \mathbb{E}\left(s_i^{2SLS}(\beta_0) s_i^{2SLS}(\beta_0)^\top\right) & \mathbb{E}\left(s_i^{2SLS}(\beta_0) s_i^{OLS}(\beta_0)^\top\right) \\ \mathbb{E}\left(s_i^{OLS}(\beta_0) s_i^{2SLS}(\beta_0)^\top\right) & \mathbb{E}\left(s_i^{OLS}(\beta_0) s_i^{OLS}(\beta_0)^\top\right) \end{bmatrix} \right)$$

$$= \mathrm{N}(\xi, \boldsymbol{\sigma})$$

- In particular, this expression allows us to compute the sample analogs of the required covariance matrices, $\boldsymbol{\sigma}$, using the plug-in principle.
- We are now in a position to describe choices of the weighting matrix $\mathbf{Q}_N$

$$\mathbf{Q}_N = \begin{cases} \boldsymbol{\iota}_{G_1+K_1} & \text{identity matrix of dimension } K_1 + G_1 \\ \left(\widehat{\boldsymbol{\sigma}}_{22,N}^{-1} - \widehat{\boldsymbol{\sigma}}_{11,N}^{-1}\right) & \text{the non-robust Hausman variance matrix} \\ \left(\widehat{\boldsymbol{\sigma}}_{22,N} + \widehat{\boldsymbol{\sigma}}_{11,N} - \widehat{\boldsymbol{\sigma}}_{12,N} - \widehat{\boldsymbol{\sigma}}_{21,N}\right)^{-1} & \text{the robust Hausman variance matrix} \end{cases}$$

- It is well-known that the Hausman variance matrix is rank-deficient by design. Remedies include generalized inverses, Hausman and Taylor (1981); Wu (1983). Other solutions are explored in Lütkepohl and Burda (1997); Dufour and Valéry (2011). Matrix norm regularization methods are required to get the estimator to behave well.

Bootstrap bias correction & t-statistics I

- Using the results above, the minimum asymptotic risk estimator can be computed. The estimator however is asymptotically biased and the asymptotic distribution does not have a closed form expression.
- The bootstrap can be used to circumvent both of these difficulties. Following Judge and Mittelhammer (2004), the bootstrap procedure for testing null hypotheses of the form

$$\mathfrak{H}_0 : \mathbf{r}\gamma = \boldsymbol{r}$$

can be implemented using a double (or, nested) bootstrap, where the outer bootstrap computes the replicates

$$\boldsymbol{T}^{(b^o)} = \left(\mathbf{r}\left(\gamma^{JS}\left(\widehat{\gamma}^{(b^o)}_{2SLS,N}, \widehat{\gamma}^{(b^o)}_{OLS,N}; \widehat{c}_1^{*(b^o)}\right) - \widehat{\text{bias}}\left(\gamma^{JS}\left(\widehat{\gamma}^{(b^o)}_{2SLS,N}, \widehat{\gamma}^{(b^o)}_{OLS,N}; \widehat{c}_1^{*(b^o)}\right)\right)\right) - \boldsymbol{r}\right)$$
$$\odot \left(\text{diag}\left(\mathbf{r}\widehat{\mathbb{V}}\left(\gamma^{JS}\left(\widehat{\gamma}^{(b^o)}_{2SLS,N}, \widehat{\gamma}^{(b^o)}_{OLS,N}; \widehat{c}_1^{*(b^o)}\right)\right)\mathbf{r}\right)\right)^{-\frac{1}{2}} ; \; b^o = 1, \ldots, B^o$$

- The estimates of the bias and the variance-covariance matrix are computed using an inner bootstrap

$$\widehat{\text{bias}}\left(\gamma^{JS}\left(\widehat{\gamma}^{(b^o)}_{2SLS,N}, \widehat{\gamma}^{(b^o)}_{OLS,N}; \widehat{c}_1^{*(b^o)}\right)\right) = \frac{1}{B^i}\sum_{b^i=1}^{B^i} \gamma^{JS}\left(\widehat{\gamma}^{(b^i)}_{2SLS,N}, \widehat{\gamma}^{(b^i)}_{OLS,N}; \widehat{c}_1^{*(b^i)}\right) - \gamma^{JS}\left(\widehat{\gamma}^{(b^o)}_{2SLS,N}, \widehat{\gamma}^{(b^o)}_{OLS,N}; \widehat{c}_1^{*(b^o)}\right)$$

The covariance matrix is computed using the inner bootstrap resamples in the usual way.

Non-random Stein-type estimator I

- The JSM estimator allows the shrinkage factor to be stochastic;
- we can also derive the optimal (minimum risk) optimal non-stochastic shrinkage parameter, c_2

$$\gamma^{NR}\left(\widehat{\gamma}_{2SLS,N}, \widehat{\gamma}_{OLS,N}; c_2\right) = (1 - c_2)\left(\widehat{\gamma}_{2SLS,N} - \widehat{\gamma}_{OLS,N}\right) + \widehat{\gamma}_{OLS,N}$$

- The optimal value of c_2 is given by the following theorem adapted from (Kim and White, 2001, Theorem 2).

Non-random Stein-type estimator II

Theorem

Define

$$c_2^* \in \underset{c_2}{\operatorname{argmin}} \operatorname{AR}\left(\gamma^{NR}\left(\widehat{\gamma}_{2SLS,N}, \widehat{\gamma}_{OLS,N}; c_2\right)\right)$$

then it must be that

$$c_2^* = \frac{\phi}{\psi}$$

where

$$\phi = \operatorname{trace}\left(\left(\boldsymbol{\omega}_{11} - \boldsymbol{\omega}_{12}\right)\mathbf{q}\right)$$

and

$$\psi = \operatorname{trace}\left(\left(\boldsymbol{\omega}_{11} + \boldsymbol{\omega}_{22} - 2\boldsymbol{\omega}_{12} + \boldsymbol{\theta}^{\top}\boldsymbol{\theta}\right)\mathbf{q}\right)$$

This infeasible estimator is called the **non-random mix** [NRM] estimator.

- We need estimates of the optimal parameters, which are easily had from the asymptotic normality results

$$\widehat{\phi} = \operatorname{trace}\left(\left(\widehat{\boldsymbol{\omega}}_{N,11} - \widehat{\boldsymbol{\omega}}_{N,12}\right)\mathbf{Q}_N\right)$$
$$\widehat{\psi} = \operatorname{trace}\left(\left(\widehat{\boldsymbol{\omega}}_{N,11} + \widehat{\boldsymbol{\omega}}_{N,22} - 2\widehat{\boldsymbol{\omega}}_{N,12} + \widehat{\boldsymbol{\theta}}^{\top}\widehat{\boldsymbol{\theta}}\right)\mathbf{Q}_N\right)$$

This feasible estimator is called the **non-random combination** [NRC] estimator.

Non-random Stein-type estimator III

- Mittelhammer and Judge (2005) define the closely related semiparametric least squares estimator [SLSE], which has the estimate of the optimal nonrandom shrinkage parameter

$$\widehat{c}^{*}_{SLSE} = \frac{\operatorname{trace}\left(\widehat{\boldsymbol{\theta}}^{\top}\widehat{\boldsymbol{\theta}} + \widehat{\boldsymbol{\omega}}_{N,11} - \widehat{\boldsymbol{\omega}}_{N,12}\right)}{\operatorname{trace}\left(\widehat{\boldsymbol{\omega}}_{N,11} + \widehat{\boldsymbol{\omega}}_{N,22} - 2\widehat{\boldsymbol{\omega}}_{N,12} + \widehat{\boldsymbol{\theta}}^{\top}\widehat{\boldsymbol{\theta}}\right)}$$

- Lastly, we need an estimate of the bias, $\boldsymbol{\theta}$ which is provided by

$$\widehat{\boldsymbol{\theta}} = \left(\frac{\mathbf{X}^{\top}\mathbf{X}^{\top}}{N}\right)^{-1}[,1..G_1]\,\frac{\mathbf{Y}_2^{\top}\left(Y_1 - \mathbf{X}\widehat{\gamma}_{OLS}\right)}{N}$$

Estimators implemented in ivshrink

Estimator	***estimator***	***options***
OLS (default)	ols	
2SLS	2sls	
LIML	liml	
Pre-Test Estimator	pte	
Ullah and Srivastava (1988)	ullah	
Sawa (1973) (bias)	sawa	bias
Sawa (1973) (MSE)	sawa	mse
Morimune (1978)(bias)	morimune	bias
Morimune (1978) (MSE)	morimune	mse
Anderson et al. (1986)	anderson	
Zellner and Vandaele (1974)	zellner	
Fuller (1977)	fuller	bias
Mittelhammer and Judge (2005)	slse	
Kim and White (2001)(random)	white	jsc
Kim and White (2001) (nonrandom)	white	nrc
Kim and White (2001) (optimal)	white	ows

`ivshrink` *depvar [inclexogvar (endogvar = exclexogvar)] [if] [in],*
[(*estimator, options*) `vce`(*vcetype*)]

Table: Estimators implemented in ivshrink

Empirical results for Mroz (1987)

Estimator	Estimate	(t-statistic)	Estimator	Estimate	(t-statistic)
OLS	0.1075	(7.5983)	Morimune (1978) (2SLS)	0.0614	$(-0.0091, 0.1234)$
2SLS	0.0614	(1.9622)	Morimune (1978) (LIML)	0.0613	$(-0.0071, 0.1272)$
LIML	0.0612	(1.9524)	Sawa (1973)	0.0614	$(-0.0122, 0.1209)$
k-class ($k = 0.7$)	0.0924	(4.3757)	Anderson et al. (1986)	.0615	$(-.0099, 0.1214)$
Fuller (1977) (MSE)	0.0632	(2.0559)	JSC	0.1038	(·)
Fuller (1977) (bias)	0.0617	(1.9786)	NRC	0.0948	(·)
JIVE	0.0576	(1.7463)	OWS	0.0810	(·)
PTE	0.1075	(7.5983)	SLSE	0.1040	$(0.0779, 0.1317)$
Ullah and Srivastava (1988)	0.0613	$(-0.0123, 0.1161)$			( )
Outcome:			*lwage*		
Included exogenous:			*exper expersq*		
Endogenous:			*educ*		
Excluded exogenous:			*fatheduc motheduc*		

Table: Estimates of the effect of education on female labor market outcomes (Mroz, 1987) (N=428)

Monte Carlo design I

Chmelarova and Hill (2010) construct a very simple just-identified design to assess the properties of the Hausman pre-test estimator. Their design has the simple form:

$$\begin{bmatrix} Y_{2i} \\ Z_{1i} \\ Z_{2i} \\ \varepsilon_i \end{bmatrix} = \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 0 & \rho_2 & \rho_1 \\ 0 & 1 & 0 & 0 \\ \rho_2 & 0 & 1 & 0 \\ \rho_1 & 0 & 0 & 1 \end{bmatrix} \right)$$

The model for outcomes is the just-identified equation

$$Y_i = \beta_0 + \beta_1 Y_{2i} + \beta_2 Z_{1i} + \varepsilon_i$$

where the degree of endogeneity is controlled by the correlation between the single explanatory endogenous regressor and the structural errors, ρ_1. The strength of instruments is controlled by the correlation between the explanatory endogenous variable and the excluded exogenous variable, ρ_2.

In the simplest case, we set the (true) vector of parameters

$$\begin{aligned} \boldsymbol{\beta}_0 &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \end{aligned}$$

Monte Carlo design II

and we set the null hypothesis for computing the rejection frequencies of the testing procedures under the null as

$$\mathfrak{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

And in order to compute the rejection frequencies of the testing procedures under a fixed alternative, we set

$$\mathfrak{H}_a : \boldsymbol{\beta} = \boldsymbol{\beta}_a$$

where

$$\boldsymbol{\beta}_a = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

`ivshrink` does more.

At the cost of stepping on `ivreg2`'s very large shoes, `ivshrink` has features not directly related to shrinkage estimation.

- Since `ivshrink` is modular, it is very easy to build on additional features using already existing features, for example, it implements (not an exhaustive list):
  - Anderson-Rubin tests;
  - Kleibergen K-tests (with & without pretesting)
  - Moreira's conditional likelihood ratio test
  - S-test

  Some of these features are translated from the Ox (Doornik, 2007) code of Marek Jarocinski.

Conclusions

- Post-model selection estimators ignore model uncertainty, leading to improper inference;
- Shrinkage estimators are risk optimal and help take model uncertainty into account;
- principle is general – any finite combination of asymptotically normal estimators;
- more generally, model averaging estimators, which average moment conditions, likelihood equations, estimating equations ...

Anderson, T. W., Kunitomo, N., and Morimune, K. (1986). Comparing single-equation estimators in a simultaneous equation system. *Econometric Theory*, 2(1):pp. 1–32.

Andrews, D. W. and Guggenberger, P. (2009). Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. *Journal of Econometrics*, 152(1):19–27.

Bates, C. E. and White, H. (1993). Determination of estimators with minimum asymptotic covariance matrices. *Econometric Theory*, 9(04):633–648.

Chmelarova, V. and Hill, R. C. (2010). The hausman pretest estimator. *Economics Letters*, 108(1):96 – 99.

Cornea, A. (2011). Bootstrap for shrinkage-type estimators.

Doornik, J. (2007). Ox: An object-oriented matrix language. *Timberlake Consultants Ltd, London*.

Dufour, J. and Valéry, P. (2011). Wald-type tests when rank conditions fail: a smooth regularization approach.

fu Wong, K. (1997). Effects on inference of pretesting the exogeneity of a regressor. *Economics Letters*, 56(3):267 – 271.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, 45(4):939–953.

Guggenberger, P. (2010). The impact of a hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory*, 26(02):369–382.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, 21(01):60–68.

Hausman, J. A. and Taylor, W. E. (1981). A generalized specification test. *Economics Letters*, 8(3):239 – 245.

Judge, G. and Mittelhammer, R. (2004). A semiparametric basis for combining estimation problems under quadratic loss. *Journal of the American Statistical Association*, 99:479–487.

Kadane, J. B. (1971). Comparison of k-class estimators when the disturbances are small. *Econometrica*, 39(5):pp. 723–737.

Kim, T.-H. and White, H. (2001). James-stein-type estimators in large samples with application to the least absolute deviations estimator. *Journal of the American Statistical Association*, 96(454):pp. 697–705.

Leeb, H. and Pötscher, B. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Lütkepohl, H. and Burda, M. M. (1997). Modified wald tests under nonregular conditions. *Journal of Econometrics*, 78(2):315 – 332.

McCloskey, A. (2012). Bonferroni-based size-correction for nonstandard testing problems.

Mittelhammer, R. C. and Judge, G. G. (2005). Combining estimators to improve structural model estimation and inference under quadratic loss. *Journal of Econometrics*, 128(1):1 – 29.

Morimune, K. (1978). Improving the limited information maximum likelihood estimator when the disturbances are small. *Journal of the American Statistical Association*, 73(364):pp. 867–871.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4):pp. 765–799.

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

Nagar, A. L. (1962). Double k-class estimators of parameters in simultaneous equations and their small sample properties. *International Economic Review*, 3(2):pp. 168–188.

Sawa, T. (1973). Almost unbiased estimator in simultaneous equations systems. *International Economic Review*, 14(1):pp. 97–106.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Theil, H. (1958). *Economic forecasts and policy*. North-Holland.

Ullah, A. and Srivastava, V. K. (1988). On the improved estimation of structural coefficients. *Sankhy: The Indian Journal of Statistics, Series B (1960-2002)*, 50(1):pp. 111–118.

Wu, D.-M. (1983). A remark on a generalized specification test. *Economics Letters*, 11(4):365 – 370.

Zellner, A. and Vandaele, W. (1974). *Studies in Bayesian econometrics and statistics*, chapter Bayes-Stein estimators for k-means, regression and simultaneous equation models, pages 628–653. Amsterdam: North Holland.