Inference for high-dimensional instrumental variables regression

David Gold

Department of Statistics
University of Washington

October 2, 2017

Overview

- Paper: "Inference for high-dimensional nested regression" with Johannes Lederer (Statistics & Biostatistics) and Jing Tao (Economics). Available at https://arxiv.org/abs/1708.05499
- Contributions: We develop methods to estimate and conduct statistical inference for low-dimensional components of a high-dimensional regression vector under endogeneity of the regressors.

Endogeneity

- Consider the linear model

$$Y = X\beta + U\,,$$

  where Y is the response, X is a regressor whose effect on Y is of interest, and U is a noise term.

- In many empirical settings (especially in econometrics), X is **endogenous**:

$$\mathrm{E}[XU] \neq 0\,.$$

- Endogeneity prevents us from discerning co-variation between X and Y due to a structural relationship from co-variation that is due to the correlation of random quantities that affect each or both variables.

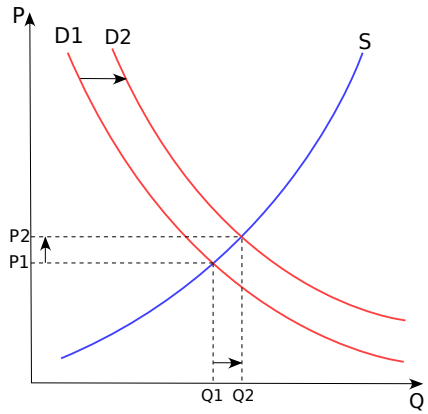
Endogeneity: Sources

Three main sources of endogeneity are:

- **confounding**
- **errors in measurements** of predictor variables, and
- the **interplay of mutually influential processes** that also exhibit random variation.

Endogeneity: Example 1 — Supply & demand

Supply and demand curves
both live in the *quantity* (Q) $\times$ *price* (P) space.



Consider the linear supply and
demand curve equations with random disturbances:

$$\begin{aligned} Q &= \theta_{\mathrm{D}} P + u_{\mathrm{D}} && \text{(Demand)} \\ Q &= \theta_{\mathrm{S}} P + u_{\mathrm{S}} && \text{(Supply)}\,. \end{aligned}$$

Some algebra shows that P has non-trivial
covariance with u_{D} due to the mutual influence of Q
and P on one another.

As a result, OLS estimates of θ_{D} (and of θ_{S}) are biased.

Endogeneity: Example 2 — Returns to schooling

- Suppose we are interested in the question of how educational attainment influences adult wages.
- We adopt the model

$$\underbrace{Y}_{\log(\textit{wage})} = \underbrace{X}_{\textit{education}} \beta + \underbrace{W}_{\substack{\textit{academic}\\ \textit{aptitude}}} \gamma + U\,,$$

but do not observe W. Rather, we can only fit the model

$$Y = X\beta + \tilde{U}\,, \qquad \tilde{U} := W\gamma + U\,.$$

- If $\mathrm{E}[XW] \neq 0$, then X is endogenous relative to $\tilde{U}$.
- See Angrist and Krueger (1991).

Instrumental Variables

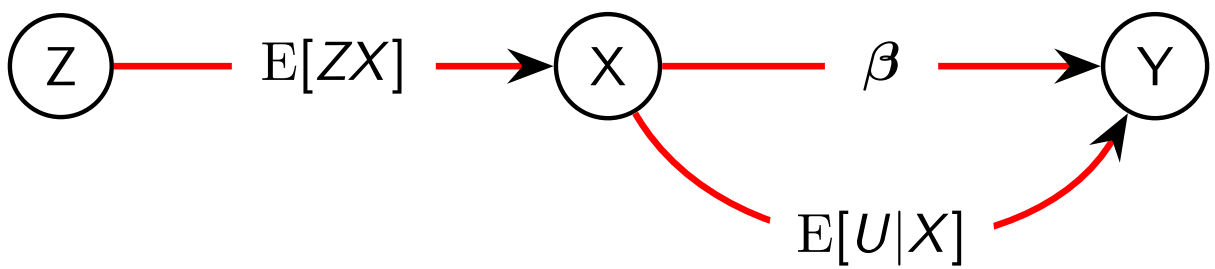
- Consider again the linear model $Y = X\beta + U$
- Suppose that Z is a third variable that satisfies

$$\mathrm{E}[ZU] = 0\,, \qquad \mathrm{E}[ZX] \neq 0\,.$$

- Then, we can write

$$\begin{aligned} \mathrm{E}[ZY] &= \mathrm{E}[ZX]\beta + \mathrm{E}[ZU] \\ \Rightarrow \qquad \beta &= \mathrm{E}[ZY]/\mathrm{E}[ZX] \end{aligned} \quad .$$

- Such a Z is called an **instrumental variable** (IV).



Model

Our model of interest is the *linear instrumental variables model* (Amemiya (1985)):

$$\begin{aligned} y_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + u_i \,, \\ x_{ij} &= \boldsymbol{z}_i^\top \boldsymbol{\alpha}_0^j + v_{ij} =: \underbrace{\boldsymbol{d}_i}_{=\boldsymbol{z}_i^\top \boldsymbol{\alpha}^j} + v_{ij} \,, \end{aligned}$$

where:

- the vector $\boldsymbol{x}_i \in \mathbb{R}^{p_x}$ consists of the *endogenous variables* $x_{i1}, \ldots, x_{ip_x}$;
- the vector $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_x}$ is the parameter of interest;
- the vector $\boldsymbol{z}_i \in \mathbb{R}^{p_z}$ consists of the *instrumental variables* $z_{i1}, \ldots, z_{ip_z}$, which satisfy $\mathrm{E}[\boldsymbol{z}_i] = \boldsymbol{0}$;
- and the vectors $\boldsymbol{\alpha}_0^j \in \mathbb{R}^{p_z}$ are regression parameters up to which the respective conditional means $d_{ij} := \mathrm{E}[x_{ij}|\boldsymbol{z}_i] = \boldsymbol{z}_i^\top \boldsymbol{\alpha}^j$

Model

We assume that the noise elements $u_i, \boldsymbol{v}_i$ satisfy

$$(u_i, \boldsymbol{v}_i)|\boldsymbol{z}_i \sim \mathcal{N}_{1+p_{\boldsymbol{x}}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{u\boldsymbol{v}}), \qquad \boldsymbol{\Sigma}_{u\boldsymbol{v}} := \begin{pmatrix} \sigma_u^2 & \boldsymbol{\sigma}_{u\boldsymbol{v}}^\top \\ \boldsymbol{\sigma}_{u\boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{pmatrix},$$

where

- $\boldsymbol{\sigma}_{u\boldsymbol{v}} := (\sigma_{u\boldsymbol{v}^1}, \ldots, \sigma_{u\boldsymbol{v}^{p_{\boldsymbol{x}}}})^\top$ consists of the noise covariances $\sigma_{u\boldsymbol{v}^j} := \mathrm{cov}(\boldsymbol{u}, \boldsymbol{v}^j)$
- $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is an unstructured covariance matrix with diagonal entries $\sigma^2_{\boldsymbol{v}^j} := \mathrm{var}(\boldsymbol{v}^j)$ for $j \in [p_{\boldsymbol{x}}]$.

Model: Regularity conditions

- The first-stage regression parameters $\boldsymbol{\alpha}_{0,j}$ satisfy

$$\max_{j \in [p_{\boldsymbol{x}}]} \|\boldsymbol{\alpha}_{0,j}\|_1 \leq M_{\mathbf{A}}$$

  for a universal constant $M_{\mathbf{A}}$;
- The second-stage regression parameter $\boldsymbol{\beta}_0$ satisfies

$$\|\boldsymbol{\beta}_0\|_1 \leq M_{\boldsymbol{\beta}}$$

  for a universal constant $M_{\boldsymbol{\beta}}$;
- The variances σ_u^2 and $\sigma_{v^j}^2$ are bounded strictly away from zero and infinity for $j \in [p_{\boldsymbol{x}}]$.
- The minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{z}} = \mathrm{E}[\boldsymbol{z}_i \boldsymbol{z}_i^\top]$ are strictly bounded away from zero and infinity, respectively.

Model: Matrix notation

In matrix notation, we write

$$\boldsymbol{y} = \boldsymbol{X}\beta_0 + \boldsymbol{u}$$

and

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

where:

- the vectors $\boldsymbol{y}, \boldsymbol{u} \in \mathbb{R}^n$ consist of the responses y_i and the noise components u_i, respectively;
- the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p_x}$ has columns $\boldsymbol{x}^j$ given by $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})^\top$;
- the matrix $\boldsymbol{D} = \mathrm{E}[\boldsymbol{X}|\boldsymbol{Z}] \in \mathbb{R}^{n \times p_x}$ has columns $\boldsymbol{d}^j$ given by $\boldsymbol{d}^j = (d_{1j}, \ldots, d_{nj})^\top$;
- the matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times p_z}$ has columns $\boldsymbol{z}^k$ given by $\boldsymbol{z}^k = (z_{1k}, \ldots, z_{nk})^\top$;
- the matrix $\mathbf{A} \in \mathbb{R}^{p_z \times p_x}$ has columns given by $\boldsymbol{\alpha}_0^j$.

Methodology overview

- IV methods were first used by Wright (1928) to estimate supply and demand elasticities for flaxseed (though see Stock and Trebbi (2003)).
- Early work in IV methods considered the *two-stage least squares (2SLS)* estimator (Theil (1953),Basmann (1957),Sargan (1958))

$$\hat{\boldsymbol{\beta}}^{\mathrm{2SLS}} = (\widehat{\boldsymbol{D}}^{\top}\widehat{\boldsymbol{D}})^{-1}\widehat{\boldsymbol{D}}\boldsymbol{y}\,,$$

  where $\widehat{\boldsymbol{D}}$ is the OLS prediction of $\boldsymbol{D} = \mathrm{E}[\boldsymbol{X}|\boldsymbol{Z}]$, and the *limited information maximum likelihood (LIML)* estimator (Anderson and Rubin (1950)); see also Anderson (2005).
- More recently, Belloni et al. (2011) and Belloni et al. (2012) study the use of the Lasso in the first-stage prediction of $\boldsymbol{D}$ when the first-stage regression vectors $\boldsymbol{\alpha}_{0,j}$ are high-dimensional.

Two-stage estimation

Similar to the 2SLS estimator in the low-dimensional linear IV model, we use a two-stage Lasso estimator to contend with both high-dimensionality and endogeneity:

- Definition (first-stage Lasso estimator):

$$\hat{\boldsymbol{\alpha}}_j \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^{p_z}}{\arg\min} \left\{ \|\boldsymbol{x}^j - \boldsymbol{Z}\boldsymbol{\alpha}\|_2^2/(2n) + r_j\|\boldsymbol{\alpha}\|_1 \right\}.$$

- Definition (second-stage Lasso estimator):

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^{p_x}}{\arg\min} \left\{ \|\boldsymbol{y} - \widehat{\boldsymbol{D}}\boldsymbol{\alpha}\|_2^2/(2n) + r_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1 \right\},$$

where $\widehat{\boldsymbol{D}} = \boldsymbol{Z}\hat{\mathbf{A}}$ and $\hat{\mathbf{A}} = (\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{p_z})$.

Compatibility condition

- Definition (Compatibility condition): For a given index set $S \in [p]$ define the set

$$\mathcal{C}(S) := \left\{ \boldsymbol{\delta} \in \mathbb{R}^p \setminus \mathbf{0} : \|\boldsymbol{\delta}\|_1 \leq C, \|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1 \right\}.$$

  We say that the *compatibility condition* holds for the matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$ relative to the index set S and the constant $\phi^2 > 0$ if

$$\phi^2 \leq \inf_{\boldsymbol{\delta} \in \mathcal{C}(S)} \frac{\sqrt{|S|}\|\boldsymbol{M}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_1^2}$$

  holds. We call such a quantity ϕ^2 the *compatibility constant*.
- Related to the *restricted eigenvalue condition* of Bickel et al. (2009); see also (Bühlmann and van de Geer, 2011, Chapter 6) and van de Geer and Bühlmann (2009).

Estimation error bounds: First-stage

- Assumption (First-stage compatibility conditions) For each active set $S_j := \operatorname{supp} \boldsymbol{\alpha}^j$ of the first-stage model, there exists a constant $\phi_j > 0$ such that $\boldsymbol{Z}$ satisfies the compatibility condition relative to S_j and ϕ_j.

- **Lemma** (First-stage estimation error) Write

$$\delta_{\hat{\mathbf{A}}} := \max_{j \in [p_x]} \|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{0,j}\|_1 .$$

Suppose that the above assumption holds. Set $r_j = \sigma_{v^j} c \sqrt{\hat{\sigma}_{\boldsymbol{Z}} \log p_{\boldsymbol{Z}} / n}$, and set $r_{\mathbf{A}} = \max_{j \in [p_x]} r_j$. Then,

$$\mathrm{P}\{\delta_{\hat{\mathbf{A}}} > 4 s_{\mathbf{A}} r_{\mathbf{A}} / \phi_{\mathbf{A}}^2\} \leq 2 p_{\boldsymbol{Z}}^{1 - c_{\mathrm{ep}}} ,$$

where $c_{\mathrm{ep}} := c^2/32 - 1$, where $s_{\mathbf{A}} = \max_{j \in [p_x]} |S_j|$ and $\phi_{\mathbf{A}} = \min_{j \in [p_x]} \phi_j$.

Estimation error bounds: Second-stage

- First-stage estimation error bounds are straightforward union bounds.
- Second-stage bounds are trickier because we need to account for the predictions $\widehat{\boldsymbol{D}}$.
- To take advantage of standard Lasso oracle inequalities for the second-stage Lasso estimator, we must write

$$\boldsymbol{y} = \boldsymbol{X}\beta_0 + \boldsymbol{u} = \widehat{\boldsymbol{D}}\beta_0 + \tilde{\boldsymbol{u}},$$

where $\tilde{\boldsymbol{u}} := \boldsymbol{u} + [(\boldsymbol{D} - \widehat{\boldsymbol{D}}) + \boldsymbol{V}]\beta_0$.

Estimation error bounds: Second-stage

- Using standard oracle inequalities, we find that, on the set $\{4\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \leq r_{\boldsymbol{\beta}}\}$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 4 s_{\boldsymbol{\beta}} r_{\boldsymbol{\beta}} / \phi_{\boldsymbol{\beta}}\,,$$

  where $s_{\boldsymbol{\beta}} = |\operatorname{supp} \boldsymbol{\beta}_0|$.
- Using Hölder's inequality many times, we also find that

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \leq\ & \delta_{\hat{\mathbf{A}}} \hat{\sigma}_{\boldsymbol{Z}} (\delta_{\hat{\mathbf{A}}} \|\boldsymbol{\beta}\|_1 + M_{\mathbf{A}}) \\ & + (\delta_{\hat{\mathbf{A}}} + M_{\mathbf{A}})(\hat{\sigma}_{\boldsymbol{Z},\boldsymbol{V}} \|\boldsymbol{\beta}\|_1 + \hat{\sigma}_{\boldsymbol{Z},\boldsymbol{u}})\,,\end{aligned}$$

  where $\delta_{\hat{\mathbf{A}}} = \max_{j \in [p_x]} \|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{0,j}\|_1$ and $\hat{\sigma}_{\boldsymbol{Z},\cdot} := \|\boldsymbol{Z}^\top \cdot\|_\infty$.

Estimation error bounds: Second-stage

- Assumption (Second-stage compatibility condition): There exists a constant $\phi_{\boldsymbol{\beta}}$ such that $\widehat{\boldsymbol{D}}$ satisfies the compatibility condition with respect to $S_{\boldsymbol{\beta}} = \operatorname{supp} \boldsymbol{\beta}_0$ and $\phi_{\boldsymbol{\beta}}$.
- **Lemma** (Second-stage estimation error): Suppose that the two Assumptions above hold. For each $j \in [p_{\boldsymbol{x}}]$, set r_j and $r_{\mathbf{A}}$ as in the previous Lemma; set

$$\lambda_{\boldsymbol{v}} = \sigma_{\boldsymbol{v}} c \sqrt{\hat{\sigma}_{\boldsymbol{z}} \log p_{\boldsymbol{z}} / n}, \qquad \lambda_{\boldsymbol{u}} = \sigma_u c \sqrt{\hat{\sigma}_{\boldsymbol{z}} \log p_{\boldsymbol{z}} / n},$$

$$r_{\boldsymbol{\beta}} = 16 \psi_{\mathbf{A}} r_{\mathbf{A}} \hat{\sigma}_{\boldsymbol{z}} \big(4 M_{\boldsymbol{\beta}} \psi_{\mathbf{A}} r_{\mathbf{A}} + M_{\mathbf{A}}\big) + \big(4 \psi_{\mathbf{A}} r_{\mathbf{A}} + M_{\mathbf{A}}\big)\big(M_{\boldsymbol{\beta}} \lambda_{\boldsymbol{v}} + \lambda_{\boldsymbol{u}}\big).$$

Then,

$$\mathrm{P}\big\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 > 4 s_{\boldsymbol{\beta}} r_{\boldsymbol{\beta}} / \phi_{\boldsymbol{\beta}}^2\big\} \leq 2 p_{\boldsymbol{z}}^{1 - c_{\mathrm{ep}}} + 2 p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}}.$$

Estimation error bounds: Second-stage compatibility condition

- van de Geer and Bühlmann (2009) and others discuss the tenability of the compatibility condition in the context of the ordinary linear model.
- Since $\widehat{\boldsymbol{D}}$ differs from an ordinary design matrix, we need to do extra work.

Estimation error bounds: Second-stage compatibility condition

Lemma (Second-stage compatibility constant) Let $S \subseteq [p_{\boldsymbol{x}}]$ be an arbitrary index set with $s = |S|$, and let $c_S > 0$ and $C > 0$ be arbitrary. For a given matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$, define the quantity

$$\phi_\dagger^2(\boldsymbol{M}, S, c_S, C) = \inf_{\substack{\delta \in \mathcal{C}^p(S, c_S): \\ \|\boldsymbol{\delta}\|_1 \leq C}} \frac{s\|\boldsymbol{M}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2}.$$

Let $\epsilon_1, \epsilon_2 > 0$ be arbitrary. Then,

$$\begin{aligned} &\mathrm{P}\left\{\phi_\dagger^2(\widehat{\boldsymbol{D}}, S, c_S, C) < (1-\epsilon_1)(\Lambda_{\min}(\boldsymbol{\Sigma_d}) - (1+c_S)^2\epsilon_2)\right\} \\ \leq\ &\mathrm{P}\left\{(2M_{\mathbf{A}}\delta_{\hat{\mathbf{A}}} + \delta_{\hat{\mathbf{A}}}^2)\hat{\sigma}_{\boldsymbol{Z}} C^2 > \epsilon_1\right\} + \mathrm{P}\left\{s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma_d}\|_\infty > \epsilon_2\right\}, \end{aligned}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \boldsymbol{D}^\top \boldsymbol{D}/n$ and $\Lambda_{\min}(\boldsymbol{\Sigma_d})$ denotes the minimal eigenvalue of $\boldsymbol{\Sigma_d}$.

Inference for high-dimensional regression parameters

- Knight and Fu (2000) show that the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ need be neither normal nor pivotal with respect to $\boldsymbol{\beta}_0$ in the low-dimensional case; see also Pötscher and Leeb (2009).
- There exist a number of recent proposals for conducting statistical inference for high-dimensional regression parameters: Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer et al. (2014), Ning and Liu (2014), Neykov et al. (2015).

One-step update: Newton-Raphson method

- Suppose that we have a system of $p_{\boldsymbol{x}}$ equations

$$\boldsymbol{f}(\boldsymbol{\beta}) = \mathbf{0}\,.$$

- The Newton-Raphson method is to take steps

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^{k} - \left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}^{k}}\right]^{-1} \boldsymbol{f}(\boldsymbol{\beta}^{k})$$

  in the direction of the solution in $\boldsymbol{\beta}$.

One-step update

- In the linear model with Gaussian errors, the $p_{\boldsymbol{x}}$-dimensional vector of score functions $\boldsymbol{h}(\boldsymbol{x}_i; \boldsymbol{\beta})$ satisfy

$$\mathrm{E}[\boldsymbol{h}(\boldsymbol{x}_i; \boldsymbol{\beta}_0)] = \mathrm{E}[-\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0)] = \mathbf{0}\,.$$

- The **one-step update** to an initial estimator $\hat{\boldsymbol{\beta}}$

$$\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\,,$$

  where $\widehat{\boldsymbol{\Theta}}$ estimates the inverse of

$$\left.\frac{\partial(-\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n)}{\partial\boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = \boldsymbol{X}^\top\boldsymbol{X}/n =: \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$$

  is one Newton-Raphson step in the direction of the solution in $\boldsymbol{\beta}$ to the empirical analogue

$$\mathbb{E}_n[\boldsymbol{h}(\boldsymbol{x}_i; \boldsymbol{\beta})] = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n = \mathbf{0}\,.$$

One-step update

- If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ is invertible, set $\widehat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$. Then,

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\,.$$

- When $p_{\boldsymbol{x}} > n$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ is not invertible, and one finds

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n \\ &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{X}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + \boldsymbol{u})/n \\ &= \boldsymbol{\beta}_0 + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top\boldsymbol{u}/n + (\widehat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{I})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})\,.\end{aligned}$$

Thus

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top\boldsymbol{u}/\sqrt{n} + \underbrace{\sqrt{n}(\widehat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{I})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})}_{\boldsymbol{\Delta}}\,.$$

One-step update

- The relationship

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_j - \beta_{0,j}) &= \hat{\boldsymbol{\theta}}_j^\top \boldsymbol{X}^\top \boldsymbol{u}/\sqrt{n} + \Delta_j \\ &= \frac{1}{\sqrt{n}} \sum_i \langle \hat{\boldsymbol{\theta}}_j, \boldsymbol{x}_i \rangle u_i + \Delta_j\end{aligned}$$

  suggests a method for conducting statistical inference for $\beta_{0,j}$.

- If $\boldsymbol{\Delta} = o_{\mathrm{P}}(1)$, and if $\hat{\boldsymbol{\theta}}_j, \boldsymbol{x}_i \perp\!\!\!\perp u_i$, then

$$\sqrt{n}(\tilde{\beta}_j - \beta_{0,j})/\omega_j \rightsquigarrow Z_j \sim \mathcal{N}(0,1)\,,$$

  where

$$\omega_j^2 = \mathrm{E}[\langle \hat{\boldsymbol{\theta}}_j, \boldsymbol{x}_i \rangle^2 u_i^2]\,.$$

- This strategy is studied in Javanmard and Montanari (2014); van de Geer et al. (2014), though these authors rather proceed directly from the KKT conditions for the Lasso.

One-step update: With endogeneity

- If the errors are not Gaussian, then $\boldsymbol{h}(\boldsymbol{x}_i; \boldsymbol{\beta}) = -\boldsymbol{x}_i(\boldsymbol{y}_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ is not a score per se but may still satisfy $\mathrm{E}\boldsymbol{h}(\boldsymbol{x}_i; \boldsymbol{\beta}_0) = \mathbf{0}$ as part of an **orthogonality condition**

$$\mathrm{E}[\boldsymbol{x}_i u_i] = \mathbf{0}\,.$$

- In our case, the $\boldsymbol{x}_i$ are endogenous, so the above doesn't hold. Instead,

$$\mathrm{E}[\boldsymbol{z}_i | u_i] = \mathbf{0} \Rightarrow \mathrm{E}[\boldsymbol{z}_i u_i] = \mathbf{0}\,,$$

  and hence

$$\mathrm{E}[\boldsymbol{d}_i u_i] = \mathbf{0}\,.$$

- Can we devise a one-step estimator that takes such information into account?
- (Note that we must substitute a prediction for $\boldsymbol{d}_i$, since the latter is unavailable...)

One-step update: With endogeneity

- The empirical analogue of $\mathbb{E}[\hat{\boldsymbol{d}}_i u_i] = \mathbf{0}$ as a function of $\boldsymbol{\beta}$ is

$$\mathbf{0} = \mathbb{E}_n[-\hat{\boldsymbol{d}}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})] =: \mathbb{E}_n[\tilde{\boldsymbol{h}}(\boldsymbol{\beta})] = -\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n$$

- The one-step update to an initial estimator $\hat{\boldsymbol{\beta}}$ in the direction to the solution of the above is

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbb{E}_n[\tilde{\boldsymbol{h}}(\boldsymbol{\beta})] \\
&= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n \\
&= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}[\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}] + \boldsymbol{u})/n \\
&= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\widehat{\boldsymbol{D}}[\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}] + [\boldsymbol{X} - \widehat{\boldsymbol{D}}][\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}] + \boldsymbol{u})/n \\
&= \boldsymbol{\beta}_0 + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/n + \underbrace{\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})/n}_{\boldsymbol{\Delta}_3/\sqrt{n}} + \underbrace{(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})}_{\boldsymbol{\Delta}_4/\sqrt{n}}.
\end{aligned}$$

One-step update: With endogeneity

- What about $\widehat{\boldsymbol{\Theta}}$? Note that $\frac{\partial \mathbb{E}_n[\tilde{\boldsymbol{h}}(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = \widehat{\boldsymbol{D}}^\top \boldsymbol{X}/n$, but

$$\boldsymbol{\Delta}_4 = (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\beta_0 - \hat{\boldsymbol{\beta}})\,.$$

  Taking $\widehat{\boldsymbol{\Theta}}$ as an estimator of $\boldsymbol{\Theta} := \mathrm{E}[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}]^{-1} =: \boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1}$ works.
- Now write

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\beta}} - \beta_0) &= \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top \boldsymbol{u}/\sqrt{n} + \boldsymbol{\Delta}_3 + \boldsymbol{\Delta}_4 \\ &= \boldsymbol{\Theta}\boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n} + \underbrace{(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}}_{\boldsymbol{\Delta}_1} + \underbrace{\widehat{\boldsymbol{\Theta}}(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n}}_{\boldsymbol{\Delta}_2} \\ &\quad + \boldsymbol{\Delta}_3 + \boldsymbol{\Delta}_4\,.\end{aligned}$$

One-step update: Recap

- Goal: Inference for components $\beta_{0,j}$ under high-dimensionality of $\boldsymbol{\beta}_0$ and endogeneity of $\boldsymbol{x}_i$.
- Method: Adapt recent literature on one-step update for Lasso to present case to define

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbb{E}_n[\tilde{\boldsymbol{h}}(\hat{\boldsymbol{\beta}})] = \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\,,$$

  which also relates to work by Amemiya (1974, 1977); Hansen (1982); Chamberlain (1987); Newey (1990) on *generalized method of moments* (GMM) IV estimators.
- Upshot: We obtained the relationship

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_j - \beta_{0,j}) &= \boldsymbol{\theta}_j\boldsymbol{D}^\top\boldsymbol{u}/\sqrt{n} + \sum_{\ell=1}^{4}\Delta_{\ell,j} \\ &= \frac{1}{\sqrt{n}}\sum_i \langle\boldsymbol{\theta}_i, \boldsymbol{d}_i\rangle u_i + \sum_{\ell=1}^{4}\Delta_{\ell,j}\,.\end{aligned}$$

Inverse estimation

- van de Geer et al. (2014) construct $\widehat{\boldsymbol{\Theta}}$ from the Nodewise Lasso estimates (Meinshausen and Bhlmann (2006))

$$\hat{\gamma}_j := \underset{\gamma \in \mathbb{R}^{p_x - 1}}{\arg\min} \|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\gamma\|_2^2/n + 2\lambda_j\|\gamma\|_1 .$$

- Javanmard and Montanari (2014) construct $\widehat{\boldsymbol{\Theta}}$ with rows $\hat{\boldsymbol{\theta}}_j$ as solutions to

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{p_x}}{\text{minimize}}\, Q(\boldsymbol{\theta}) \qquad \text{subject to } \|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{\theta}\|_\infty \leq \mu$$

  with $Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \boldsymbol{\theta}$.
- We use essentially the CLIME estimator of Cai et al. (2011), which solves the above program with $Q(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$.

Inverse estimation

- The CLIME estimator has established rates for ℓ_1 estimation error, whereas the Javanmard and Montanari (2014) estimator does not.
- In order to obtain the ℓ_1 rates for the CLIME, we must impose additional assumptions on $\widehat{\mathbf{\Theta}}$.

Inverse estimation

- Definition (Uniformity class of Cai et al. (2011)): The *uniformity class* of population inverses $\widehat{\boldsymbol{\Theta}}$ is defined to be

$$\begin{aligned} &\mathcal{U}(M_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}}) \\ &:= \left\{ \boldsymbol{\Theta} = (\theta_{jk})_{j,k=1}^{p_x} \succ \mathbf{0} : \|\boldsymbol{\Theta}\|_{L_1} \leq M_{\boldsymbol{\Theta}}; \max_{j \in [p_x]} \sum_{k \in [p_x]} |\theta_{jk}|^q \leq s_{\boldsymbol{\Theta}} \right\}. \end{aligned}$$

- **Theorem** (ℓ_1 estimation error of $\hat{\boldsymbol{\theta}}_j$): Suppose that: (i) $\widehat{\boldsymbol{\Theta}} \in \mathcal{U}(M_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}})$; (ii) $\hat{\boldsymbol{\theta}}_j$ is the CLIME estimator of $\boldsymbol{\theta}_j$; (iii) $\boldsymbol{\theta}_j$ is feasible for the CLIME program. Then,

$$\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 \leq 2c_q(2M_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}},$$

for each $j \in [p_x]$, where $c_q := 1 + 2^{1-q} + 3^{1-q}$.

Remainder terms

- Under model regularity and rate conditions, the remainder terms $\boldsymbol{\Delta}_\ell$ each satisfy $\boldsymbol{\Delta}_\ell = o_{\mathrm{P}}(1)$.
- For sub-Gaussian $\boldsymbol{Z}$, choosing μ on the order of $\sqrt{\log(p_{\boldsymbol{x}})/n}$ achieves balance between feasibility of $\boldsymbol{\theta}_j$ for inverse program and the needs of the rates for the remainder terms.

Asymptotic normality

- Under model regularity and rate conditions, we then have

$$\sqrt{n}(\tilde{\beta}_j - \beta_{0,j}) = n^{-1/2} \sum_i \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i + o_{\mathrm{P}}(1) .$$

- Now define

$$\begin{aligned} \omega_j^2 &= \mathrm{E}[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 u_i^2] \\ &= \boldsymbol{\theta}_j^\top \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top] \boldsymbol{\theta}_j \mathrm{E}[u_i^2] \\ &= \sigma_u^2 \boldsymbol{\Theta}_{jj} . \end{aligned}$$

Asymptotic normality

Theorem (Asymptotic normality of $\sqrt{n}(\tilde{\beta}_j - \beta_{0,j})/\omega_j$): Let $\tilde{\boldsymbol{\beta}}$ denote the one-step second-stage Lasso estimator of $\boldsymbol{\beta}_0$, and let ω_j be as above. Then, under model regularity and rate conditions, it holds that

$$\sqrt{n}(\tilde{\beta}_j - \beta_{0,j})/\omega_j \rightsquigarrow Z \sim \mathcal{N}(0,1)\,.$$

Moreover, the result continues to hold if ω_j is replaced with

$$\hat{\omega}_j := \hat{\sigma}_u \widehat{\boldsymbol{\Theta}}_{jj}\,,$$

where $\hat{\sigma}_u$ is an estimator of the second-stage noise level that satisfies $\hat{\sigma}_u/\sigma_u - 1 = o_{\mathrm{P}}(1)$.

Design: Data-generating mechanism

- Choose $\alpha_{0,j}, \beta_0$ to have s_{α^j}, s_β non-zero entries equal to 1.
- Draw observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ according to

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{N}_{p_z}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{z}}), \\ (u_i, \boldsymbol{v}_i) \,|\, \boldsymbol{z}_i &\sim \mathcal{N}_{1+p_x}(\boldsymbol{0}, \boldsymbol{\Sigma}_{u\boldsymbol{v}}), \\ x_{ij} &= \langle \boldsymbol{z}_i, \alpha^j \rangle + v_{ij}, \\ y_i &= \langle \boldsymbol{x}_i, \beta \rangle + u_i, \end{aligned}$$

  where $n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}}, \beta, \{\alpha^j\}_{j \in [p_x]}$, and the structure of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ can vary amongst configurations.
- For all configurations, we set

$$\boldsymbol{\Sigma}_{u\boldsymbol{v}} = \begin{pmatrix} \sigma_u & \sigma_{u\boldsymbol{v}} \boldsymbol{1}^\top \\ \sigma_{u\boldsymbol{v}} \boldsymbol{1} & \sigma_v \boldsymbol{I} \end{pmatrix},$$

  where $\sigma_u, \sigma_v = 1$ are held fixed but $\sigma_{u\boldsymbol{v}}$ varies over configurations.

Design: Choices of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$

- The first choice of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is of a Toeplitz (TZ) structure given by

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}}\big|_{jk} = \rho^{|j-k|}\,, \qquad j,k \in [p_{\boldsymbol{z}}]\,, \qquad \rho = 0.8\,.$$

- The second choice of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is a circulant-symmetric (CS) structure given for $j \leq k$ by

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}}\big|_{jk} = \begin{cases} 1 & k = j\,, \\ 0.1 & k \in \{j+1,\ldots,j+5\} \cup \{j+p_{\boldsymbol{z}}-5,\ldots,j+p_{\boldsymbol{z}}-1\}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Design: Metrics

- We generate $N = 100$ trials for each configuration of

$$n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}}, s_{\boldsymbol{\beta}}, s_{\mathbf{A}}, b, a, \boldsymbol{\Sigma}_{\boldsymbol{z}}, \sigma_{\boldsymbol{uv}} .$$

- In each trial, for each component $\tilde{\beta}_j$ of $\tilde{\boldsymbol{\beta}}$ we compute the respective $(1-\alpha)\%$ confidence interval

$$\hat{\mathcal{I}}_{\alpha,j} = \left[\tilde{\beta}_j - z_{0.05}\widehat{\mathrm{SE}}(\tilde{\beta}_j),\ \tilde{\beta}_j + z_{0.05}\widehat{\mathrm{SE}}(\tilde{\beta}_j)\right],$$

  where $z_{0.05} = \Phi^{-1}(1-0.05/2)$ and $\widehat{\mathrm{SE}}(\tilde{\beta}_j)^2 = \mathbb{E}_n[(y_i - \hat{\boldsymbol{\beta}}\boldsymbol{X})^2\langle\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{d}}_i\rangle^2]$

- For each configuration, we calculate the average coverage $\widehat{\mathrm{cvg}}$ for the 95% confidence intervals $\hat{\mathcal{I}}_{\alpha,j}$ about components of $\tilde{\boldsymbol{\beta}}$ and the average interval length $\widehat{\mathrm{len}}$ given by

$$\widehat{\mathrm{cvg}} = \frac{1}{p_{\boldsymbol{x}}}\sum_{j=1}^{p_{\boldsymbol{x}}}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\beta_j \in \hat{\mathcal{I}}_{\alpha,j}\}, \quad \widehat{\mathrm{len}} = \frac{1}{p_{\boldsymbol{x}}}\sum_{j=1}^{p_{\boldsymbol{x}}}\frac{1}{N}\sum_{i=1}^{N}\mathrm{len}(\hat{\mathcal{I}}_{\alpha,j}) .$$

Configurations

We conduct simulations according the design described above for all configurations belonging to

$$\underbrace{\begin{pmatrix} (100, 250, 500) \\ (400, 800, 1000) \\ (1000, 1500, 1750) \end{pmatrix}}_{(n\,, p_{\boldsymbol{x}}\,, p_{\boldsymbol{z}})} \times \underbrace{\begin{pmatrix} (3, 5) \\ (10, 15) \end{pmatrix}}_{(s_{\boldsymbol{\beta}}\,, s_{\mathbf{A}})} \times \underbrace{\begin{pmatrix} .1 \\ 1 \end{pmatrix}}_{b} \times \underbrace{\begin{pmatrix} .1 \\ 1 \end{pmatrix}}_{a} \times \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}} \\ \boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}} \end{pmatrix}}_{\boldsymbol{\Sigma}_{\boldsymbol{z}}} \times \underbrace{\begin{pmatrix} .9 \\ .5 \\ .1 \end{pmatrix}}_{\sigma_{\boldsymbol{uv}}},$$

with the exception of the parameter combination

$$(n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}}, s_{\boldsymbol{\beta}}, s_{\mathbf{A}}) = (1000, 1500, 1750, 20, 25)\,,$$

which we include in place of

$$(n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}}, s_{\boldsymbol{\beta}}, s_{\mathbf{A}}) = (1000, 1500, 1750, 10, 15)$$

due to the magnitude of n, $p_{\boldsymbol{x}}$, and $p_{\boldsymbol{z}}$ in that particular combination.

Results: Simulation results for circulant-symmetric Σ_z

$(s_\beta, s_\mathbf{A})$	σ_{uv}	$\widehat{\text{cvg}}$ $(\widehat{\text{len}})$ β	β_{S_β}	$\beta_{\overline{S}_\beta}$
		$(n,\ p_x,\ p_z) = (100,\ 250,\ 500)$		
(3, 5)	0.9	0.944 (0.59)	0.843 (0.60)	0.946 (0.59)
	0.5	0.944 (0.43)	0.830 (0.43)	0.946 (0.43)
	0.1	0.949 (0.21)	0.843 (0.21)	0.950 (0.21)
(10, 15)	0.9	0.909 (0.32)	0.569 (0.33)	0.923 (0.32)
	0.5	0.919 (0.27)	0.641 (0.27)	0.931 (0.27)
	0.1	0.924 (0.21)	0.727 (0.22)	0.932 (0.21)
		$(n,\ p_x,\ p_z) = (400,\ 800,\ 1000)$		
(3, 5)	0.9	0.952 (0.48)	0.923 (0.48)	0.952 (0.48)
	0.5	0.950 (0.36)	0.950 (0.36)	0.950 (0.36)
	0.1	0.949 (0.16)	0.947 (0.16)	0.949 (0.16)
(10, 15)	0.9	0.949 (0.25)	0.867 (0.25)	0.950 (0.25)
	0.5	0.950 (0.19)	0.891 (0.19)	0.951 (0.19)
	0.1	0.948 (0.09)	0.880 (0.09)	0.949 (0.09)
		$(n,\ p_x,\ p_z) = (1000,\ 1500,\ 1750)$		
(3, 5)	0.9	0.951 (0.39)	0.950 (0.40)	0.951 (0.39)
	0.5	0.950 (0.29)	0.963 (0.29)	0.950 (0.29)
	0.1	0.950 (0.13)	0.950 (0.13)	0.950 (0.13)
(20, 25)	0.9	0.948 (0.16)	0.917 (0.15)	0.949 (0.16)
	0.5	0.950 (0.12)	0.929 (0.12)	0.951 (0.12)
	0.1	0.958 (0.06)	0.929 (0.06)	0.959 (0.06)

Results: Simulation results for Toeplitz Σ_z

$(s_\beta, s_\mathbf{A})$	σ_{uv}	$\widehat{\text{cvg}}$ ($\widehat{\text{len}}$) β	β_{S_β}	$\beta_{\overline{S}_\beta}$
		$(n,\ p_x,\ p_z) = (100,\ 250,\ 500)$		
(3, 5)	0.9	0.946 (0.58)	0.803 (0.59)	0.947 (0.58)
	0.5	0.948 (0.44)	0.780 (0.44)	0.950 (0.44)
	0.1	0.950 (0.21)	0.873 (0.22)	0.951 (0.21)
(10, 15)	0.9	0.913 (0.31)	0.585 (0.31)	0.927 (0.31)
	0.5	0.903 (0.26)	0.521 (0.26)	0.919 (0.26)
	0.1	0.901 (0.19)	0.659 (0.20)	0.911 (0.19)
		$(n,\ p_x,\ p_z) = (400,\ 800,\ 1000)$		
(3, 5)	0.9	0.952 (0.48)	0.897 (0.50)	0.953 (0.48)
	0.5	0.952 (0.36)	0.927 (0.34)	0.952 (0.36)
	0.1	0.952 (0.16)	0.930 (0.16)	0.952 (0.16)
(10, 15)	0.9	0.955 (0.25)	0.861 (0.25)	0.956 (0.25)
	0.5	0.954 (0.18)	0.837 (0.18)	0.956 (0.18)
	0.1	0.956 (0.09)	0.875 (0.09)	0.957 (0.09)
		$(n,\ p_x,\ p_z) = (1000,\ 1500,\ 1750)$		
(3, 5)	0.9	0.951 (0.37)	0.923 (0.35)	0.951 (0.37)
	0.5	0.951 (0.27)	0.940 (0.27)	0.951 (0.27)
	0.1	0.953 (0.12)	0.947 (0.12)	0.953 (0.12)
(20, 25)	0.9	0.952 (0.14)	0.838 (0.14)	0.954 (0.14)
	0.5	0.946 (0.10)	0.810 (0.10)	0.947 (0.10)
	0.1	0.890 (0.05)	0.678 (0.05)	0.892 (0.05)

Limitations & next steps

- Tuning parameters: In practice, cross-validation seems to work fine, but we must develop theory for data-adaptive tuning parameter selection. We will investigate use of results in Cai et al. (2016), for instance, as well as ideas from Chichignoud et al. (2016); Bien et al. (2017).
- Real world applications: We have some ideas for conducting inference for large numbers of interaction effects in the Angrist and Krueger (1991) model that hitherto have been unexplored.
- Noise models: We are currently developing theory to handle non-Gaussian and heteroscedastic noise.
- Asymptotic efficiency: We are currently investigating the use of Jankova and van de Geer (2016) to demonstrate asymptotic efficiency of our method.
- Application to nonlinear regression: We will investigate the use of our method to approximate nonlinear regression functions, for instance under the "approximate sparsity" regime of Belloni et al. (2011, 2012).

Acknowledgments

- Johannes Lederer & Jing Tao
- Joseph Salmon & Mohamed Hebiri
- Donghui Mai
- UW Royalty Research Fund Grant
- Amazon Cloud Credits for Research Grant

References I

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2:105–110.

Amemiya, T. (1977). The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica*, 45(4):955–968.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Anderson, T. W. (2005). Origins of the limited information maximum likelihood estimator and two-stage least squares estimators. *Journal of Econometrics*, 127:1–16.

Anderson, T. W. and Rubin, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 21(4):570–582.

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

References II

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, 25(1):77–83.

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.

Belloni, A., Chernozhukov, V., and Hansen, C. (2011). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bien, J., Gaynanova, I., Lederer, J., and Müller, C. L. (2017). Non-convex Global Minimization and False Discovery Rate Control for the TREX. *J. Comput. Graph. Statist.*

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

References III

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(464):594–607.

Cai, T. T., Liu, W., and Zhou, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.*, 44(2):455–488.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.

Chichignoud, M., Lederer, J., and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *J. Mach. Learn. Res.*, 17:1–17.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Jankova, J. and van de Geer, S. (2016). Semi-parametric efficiency bounds for high-dimensional models. *arXiv:1601.00815*.

References IV

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378.

Meinshausen, N. and Bhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837.

Neykov, M., Ning, Y., Liu, J. S., and Liu., H. (2015). A unified theory of confidence regions and testing for high dimensional estimating equations. *arXiv:1510.08986*.

Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv:1412.8765*.

References V

Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivar. Anal.*, 100(9):2065–2082.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*.

Stock, J. H. and Trebbi, F. (2003). Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194.

Theil, H. (1953). Repeated least-squares applied to complete equation systems. *Centraal Planbureau Memorandum*.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, 3:1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

Wright, P. G. (1928). *The tariff on animal and vegetable oils*. New York, The Macmillan Company.

References VI

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242.