Choosing the Dictionary and Penalty for IV-LASSO

Yukun Ma [1] Manu Navjeevan [2] Bogdan Salagub [3]

[1]Rochester [2]Texas A&M [3]UCLA

November, 2026

Setup

Consider the linear IV model

$$y_i = x_i\beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid z_i] = 0$$

Researcher observes n independent observations $(y_i, x_i, z_i)'$

- Outcome $y_i \in \mathbb{R}$;
  - e.g., log house prices
- Endogenous variable $x_i \in \mathbb{R}$;
  - e.g., number of eminent-domain rulings favoring property owners
- Instruments $z_i \in \mathbb{R}^p$;
  - e.g., composition of randomly assigned judge panels

Setup

Optimal instrument for estimating β under homoskedasticity is

$$\Pi_i = \mathbb{E}[x_i \mid z_i]$$

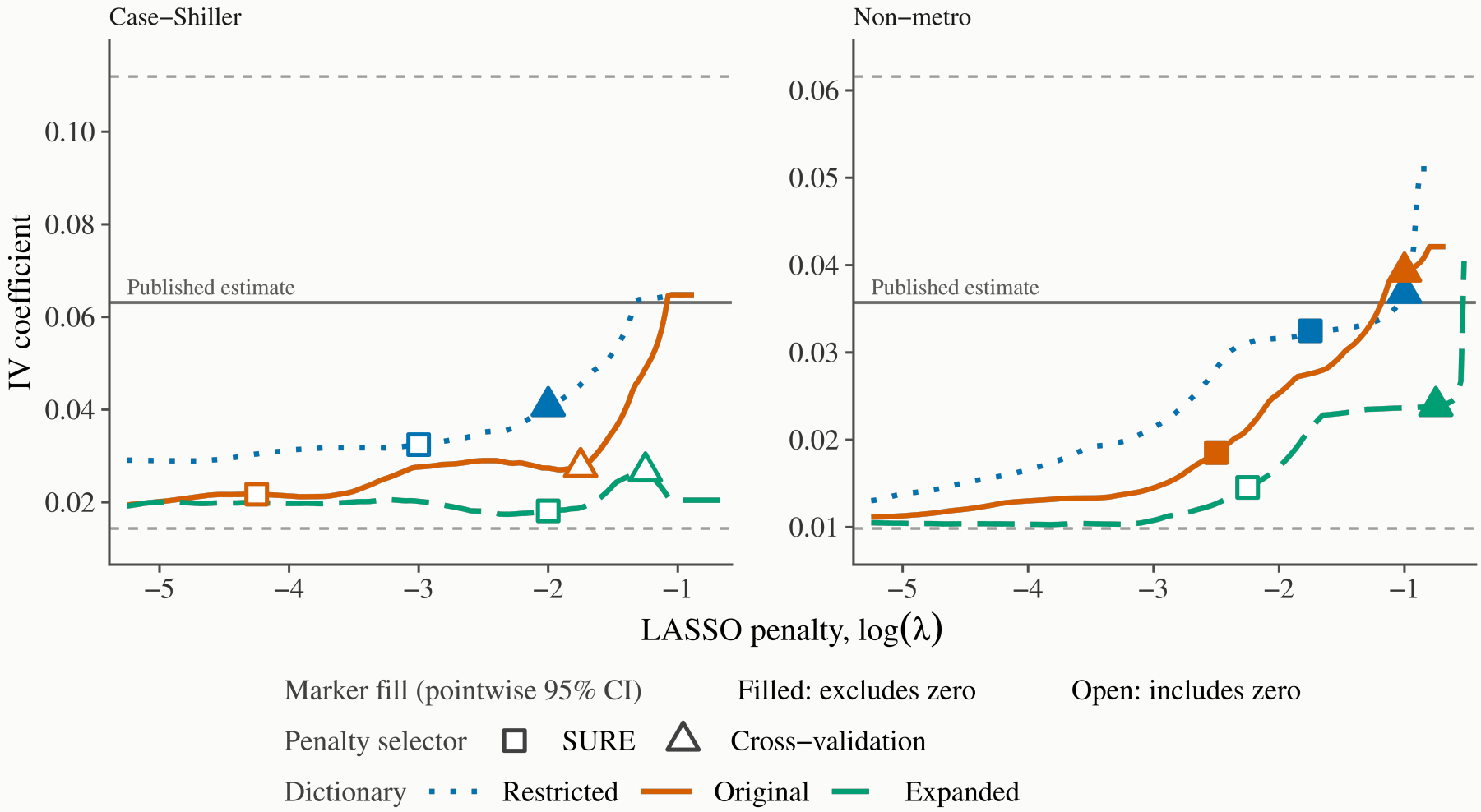
When the researcher has access to a large dictionary of technical instruments, optimal instrument can be estimated using LASSO (Belloni et al., 2012)

$$\widehat{\pi} \in \arg\min_{\pi} \frac{1}{2}\mathbb{E}_n[(x_i - \tilde{z}_i'\pi)^2] + \lambda\|\pi\|_1, \qquad \widehat{\Pi}_i = \widehat{\pi}' z_i$$

Problem: How to select penalty parameter or set of technical instruments to use?

Setup

Any consistent estimate of optimal instrument leads to the same limiting distribution for $\widehat{\beta}$. However…



Existing Selection Rules

Existing selection rules do not target accuracy of structural estimate

- Cross-Validation (Chetverikov et al., 2021) and SURE (Tibshirani and Taylor, 2012) target first-stage prediction
  - More flexible first-stage fits may be more correlated with endogenous variable, biasing $\widehat{\beta}$
- "Plug-in" rules of Belloni et al. (2012) and Chetverikov and Sørensen (2025) are designed to control first-stage estimation error
  - Guarantee first-stage consistency but do not claim optimality

Contribution

This paper develops second order asymptotic theory for the IV-LASSO estimator and uses it to provide a principled method for selecting the first stage implementation.

- Analysis complicated by non-differentiabilty of LASSO solution,
  - Use geometric characterization of Tibshirani and Taylor (2012) as well as second-order Stein identities from Bellec and Zhang (2021) to characterize higher-order moments;
- Estimation of second order terms can be slow,
  - Show that the leading part of estimation error is common to all candidate implementations; feasible criterion still selects optimal implementation;
- In simulations calibrated to Paravisini et al. (2015) improve MSE by up to 50% relative to cross-validation and plug-in rules.

Operate in homoskedastic Gaussian model and then extend results to a class of non-Gaussian errors.

Related Literature

- **Empirical Applications:** Paravisini et al. (2015), Gilchrist and Sands (2016), Reimers and Waldfogel (2021), Bald et al. (2022), Derenoncourt (2022), Ryan and Sudarshan (2022), Agan et al. (2023), Bazzi et al. (2023), Beraja et al. (2023), Do and Jacoby (2025), Bazzi et al. (2026), Cuddy and Currie (2026)
- **Penalty Parameter Selection:** Belloni et al. (2012), Tibshirani and Taylor (2012), Chetverikov et al. (2021), Chetverikov and Sørensen (2025)
- **Second Order Asymptotics:** Nagar (1959), Sargan (1976), Rothenberg (1984), Donald and Newey (2001), Newey and Smith (2004), Hahn et al. (2004), Kuersteiner and Okui (2010), Okui (2011), Carrasco (2012), Velez (2026)
- **High-Dimensional Methods:** Belloni and Chernozhukov (2013), Belloni et al. (2014), Chernozhukov et al. (2018), Wager and Athey (2018), Farrell et al. (2021)

Table of Contents

Theory and Infeasible Criterion

Feasible Selection

Non-Gaussian Extension

Empirically Calibrated Simulations

Framework

Operate in a homoskedastic Gaussian model

$$y_i = x_i\beta + \varepsilon_i$$
$$x_i = \Pi_i + v_i$$

Treat the instruments $\{z_i\}_{i=1}^n$ as fixed.

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{pmatrix}\right)$$

Candidate Implementations

Each candidate $c \in [\mathcal{J}]$ is described by a dictionary-penalty pair

- Dictionary $Z_c \in \mathbb{R}^{n \times p_c}$ consists of a set of technical instruments; i.e transformations of instruments,
  - Allow $p_c \gg n$
- Penalty $\lambda_c > 0$ determines degree of regularization.

Estimate first-stage, $\widehat{\Pi}_c$, via LASSO

$$\widehat{\pi}_c \in \underset{\pi \in \mathbb{R}^{p_c}}{\arg\min} \frac{1}{2}\mathbb{E}_n[(x_i - z_{c,i}'\pi)^2] + \lambda_c \|\pi\|_1, \quad \widehat{\Pi}_{c,i} = z_{c,i}'\widehat{\pi}_c$$

Effective Dimension

For each $c \in [\mathcal{J}]$ define the equicorrelation set

$$E_c := \left\{ j \in [p_c] : \left| \mathbb{E}_n \left[z_{c,ij} (x_i - \widehat{\Pi}_{c,i}) \right] \right| = \lambda_c \right\}$$

i.e, the set of dictionary columns for which LASSO KKT inequalities bind. Effective dimension is the rank of first-stage dictionary associated with these columns:

$$d_c := \operatorname{rank} \left(Z_{c,E_c} \right).$$

Effective dimension is central to our analysis as it provides a measure of complexity for each candidate LASSO fit.

- LASSO solution may not be unique, but d_c is always well defined and observed since LASSO residual is unique;
- When LASSO solution is unique, d_c coincides with number of selected variables.

Structural Estimate

With first-stage estimates in hand, construct structural estimate

$$\widehat{\beta}_c := \frac{\mathbb{E}_n[y_i \widehat{\Pi}_{c,i}]}{\mathbb{E}_n[x_i \widehat{\Pi}_{c,i}]}.$$

So long as each first-stage is consistent

$$\sqrt{n}\left(\widehat{\beta}_c - \beta_c\right) = \frac{\sqrt{n}\mathbb{E}_n[\varepsilon_i \Pi_i]}{\mathbb{E}_n[\Pi_i^2]} + o_p(1)$$

Leading term does not depend on candidate estimate and is distributed

$$N\left(0, \sigma_\varepsilon^2 \big/ \mathbb{E}_n[\Pi_i^2]\right)$$

To distinguish between candidates, need to analyze terms in remainder.

AMSE Criterion

Would like to use MSE to compare candidates, but MSE may not exist for IV estimators. Instead, for each candidate $\widehat{\beta}_c$ we show a decomposition

$$n\left(\widehat{\beta}_c - \beta\right)^2 = C_0 + Q_c + r_c \tag{1}$$

where C_0 is common to all candidates and $\max_{c\in[\mathcal{J}]} |r_c/\mathbb{E}[Q_c]| \to 0$.

- Approach followed in Donald and Newey (2001) and Newey and Smith (2004) for higher order comparasions.

Compare candidates via *Approximate Mean Square Error (AMSE)*, $\mathbb{E}[C_0 + Q_c]$. Goal is to choose a candidate with lowest AMSE.

MSE Nonexistence

Partial Beta-min

Assume first-stage has approximately sparse representation (Belloni et al., 2012):

$$\Pi_i = z_{c,i}'\pi_c^0 + \xi_{c,i}, \qquad T_c := \operatorname{supp}(\pi_c^0), \qquad s_c := |T_c|.$$

Call a coefficient "strong" if magnitude exceeds LASSO estimation error, which vanishes as $n \to \infty$.

Assumption (Partial Beta-min)

Let k_c count the strong coefficients. Assume

$$k_c \to \infty \qquad \text{and} \qquad s_c \lesssim k_c.$$

- Unlike full beta-min (Zhao and Yu, 2006, Wainwright, 2009), does not require the LASSO to recover the full support.
- Mirrors $K \to \infty$ in Donald and Newey (2001).

Infeasible Criterion

Theorem 1 (Infeasible AMSE Criterion)

For all candidates, decomposition in (1) *holds and* $\mathbb{E}[Q_c] = S_c(1 + o(1))$ *where*

$$S_c := \frac{\sigma_\varepsilon^2 \mathcal{A}_c}{h_c^2} + \frac{\sigma_{\varepsilon v}^2 \mathbb{E}[d_c^2 + d_c]}{n h_c^2}$$

for $h_c := \bar{E}[\widehat{\Pi}_{c,i}\Pi_i]$ *and*

$$\mathcal{A}_c := \min_{a \in \mathbb{R}} \bar{\mathbb{E}} \left[\left(\widehat{\Pi}_{c,i} - a\Pi_i \right)^2 \right]$$

Infeasible Criterion

A richer dictionary or lighter penalty may reduce $\mathcal{A}_c$ while increasing the effective dimension. Let $\rho := \text{Corr}(\varepsilon_i, v_i)$ and rewrite

$$S_c = \frac{\sigma_\varepsilon^2}{h_c^2}\left\{\mathcal{A}_c + \rho^2\sigma_v^2\frac{\mathbb{E}[d_c^2 + d_c]}{n}\right\}$$

- Cross validation targets $\bar{E}\left[(\widehat{\Pi}_{c,i} - \Pi_i)^2\right]$ while ignoring complexity cost.
  - Also ignores the fact that optimal instrument is invariant to rescaling
- Candidates that strongly identify the first-stage have lower AMSE.
  - Denominator h_c^2 roughly measures strength of identification (correlation between estimated and true first-stages).
  - AMSE differences also scaled by strength of identification.

Proof Idea: LASSO Geometry

Define the set

$$\mathcal{K}_c := \{q \in \mathbb{R}^n : \|\mathbb{E}_n[z_{c,i} q_i]\|_\infty \leq \lambda\}$$

That is the set of potential LASSO residuals that satisfy the LASSO KKT conditions. Tibshirani and Taylor (2012) show that the fitted LASSO residual is given

$$\widehat{u}_c := x - \widehat{\Pi}_c = P_{\mathcal{K}}(x)$$

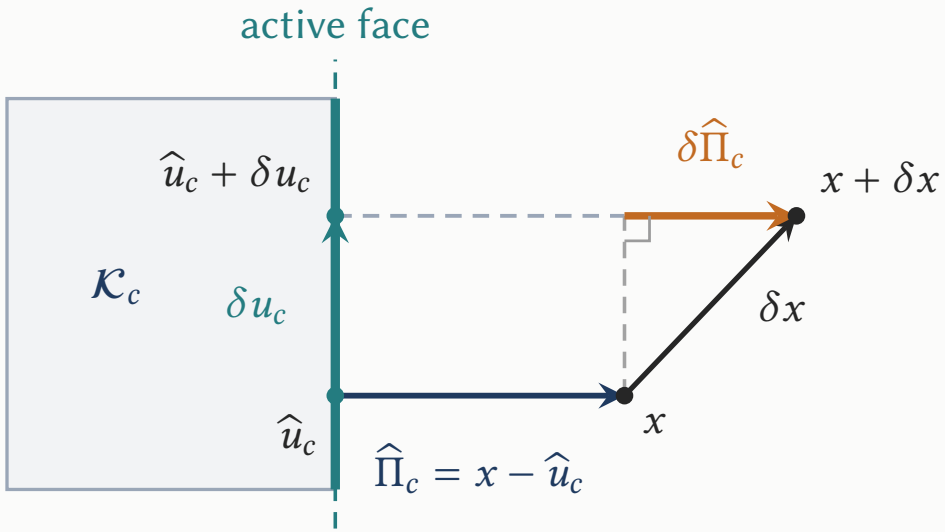
Equicorrelation set E_c describes the "active face" of this projection. As a result

$$\dot{\mu} := D\widehat{\Pi}_c = P_{\text{col}(Z_{c,E_c})} \implies \text{div}(\dot{\mu}) = \text{tr}(D\widehat{\Pi}_c) = d_c$$

Residual Projection

Proof Idea: LASSO Geometry

Locally, the projection stays on the same active face.



Tangent and normal spaces

$$T_c = \ker(Z'_{c,E_c}),$$

$$T_c^{\perp} = \operatorname{col}(Z_{c,E_c}).$$

Residual change

$$\delta u_c = P_{T_c}\delta x.$$

Fitted-value change

$$\delta\widehat{\Pi}_c = \delta x - \delta u_c.$$

$$D\widehat{\Pi}_c(x) = I - DP_{\mathcal{K}_c}(x) = I - P_{T_c} = P_{\operatorname{col}(Z_{c,E_c})}$$

(almost everywhere.)

Proof Idea: Cross-Moments

A key step in the proof is analyzing relationship between first-stage fit and structural error. Specifically, want to compute

$$\mathbb{E}[(\varepsilon'\widehat{\Pi}_c)^2]$$

To do so, write

$$\varepsilon = \frac{\sigma_{\varepsilon v}}{\sigma_v^2} v + \eta, \qquad \eta_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2}\right), \qquad \eta \perp v$$

Decompose

$$\mathbb{E}[(\varepsilon'\widehat{\Pi}_c)^2] = \frac{\sigma_{\varepsilon v}^2}{\sigma_v^4}\mathbb{E}[(v'\widehat{\Pi}_c)^2] + \mathbb{E}[(\eta'\widehat{\Pi}_c)^2]$$

Cross-term drops out since $\eta \perp (\widehat{\Pi}_c, v)$

Proof Idea: Second Order Stein Identity

To characterize $\mathbb{E}[(v'\widehat{\Pi}_c)^2]$, apply the second-order Stein identity from Bellec and Zhang (2021).

Lemma 1 (Stein Identities)

Let $W \sim N(0, \sigma^2 I_n)$. If $m : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz with $A(w) := Dm(w)$ its almost everywhere derivative, then

$$\mathbb{E}[W' m(W)] = \sigma^2 \mathbb{E}[\operatorname{tr}(A(W))]$$

and

$$\mathbb{E}\left[\left\{ W' m(W) - \sigma^2 \operatorname{tr}(A(W)) \right\}^2\right] = \sigma^2 \mathbb{E}[\|m(W)\|_2^2] + \sigma^4 \mathbb{E}[\operatorname{tr}(A(W)^2)].$$

Proof Idea: Second Order Stein Identity

Recall, $D\widehat{\Pi}_c = P_{Z_{c,\text{col}(E_c)}}$ is a projection of rank d_c. The second-order identity gives

$$\mathbb{E}[(v'\widehat{\Pi}_c - \sigma_v^2 d_c)^2] = \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}_c\|_2^2] + \sigma_v^4 \mathbb{E}[d_c].$$

Where we use idempotency of the projection for the second term. One more application of first-order Stein then eliminates the cross term:

$$\mathbb{E}[d_c(v'\widehat{\Pi}_c - \sigma_v^2 d_c)] = 0.$$

Expanding the square then yields

$$\mathbb{E}[(v'\widehat{\Pi}_c)^2] = \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}_c\|_2^2] + \sigma_v^4 \mathbb{E}[d_c^2 + d_c].$$

Cross term

Proof Idea: Cross-Moments

To complete the expression for $\mathbb{E}[(\varepsilon'\widehat{\Pi}_c)^2]$, since $\eta \perp (\widehat{\Pi}_c, v)$ we can write

$$\mathbb{E}[(\eta'\widehat{\Pi}_c)^2] = \left(\sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_v^2}\right)\mathbb{E}[\|\widehat{\Pi}_c\|_2^2].$$

Combining this with the Stein calculation gives

$$\mathbb{E}[(\varepsilon'\widehat{\Pi}_c)^2] = \sigma_\varepsilon^2\mathbb{E}[\|\widehat{\Pi}_c\|_2^2] + \sigma_{\varepsilon v}^2\mathbb{E}[d_c^2 + d_c].$$

Next, expand the structural estimation error to see how this numerator moment enters AMSE.

Proof Idea: Structural Error Expansion

Write $\widehat{h}_c = \mathbb{E}_n[x_i \widehat{\Pi}_{c,i}]$. Under our assumptions, $\widehat{h}_c = h_c + O_p(n^{-1/2})$, so

$$\sqrt{n}(\widehat{\beta}_c - \beta) = \frac{\sqrt{n}\,\mathbb{E}_n[\varepsilon_i \widehat{\Pi}_{c,i}]}{\widehat{h}_c} = \frac{\varepsilon' \widehat{\Pi}_c}{\sqrt{n}\,\widehat{h}_c}$$

$$= \frac{\varepsilon' \widehat{\Pi}_c}{\sqrt{n}\,h_c} \left\{ 1 - \frac{\widehat{h}_c - h_c}{h_c} \right\} + O_p(n^{-1}).$$

Squaring the leading term and taking expectations gives:

$$\mathbb{E}\left[\left(\frac{\varepsilon' \widehat{\Pi}_c}{\sqrt{n}\,h_c} \right)^2 \right] = \frac{\mathbb{E}[(\varepsilon' \widehat{\Pi}_c)^2]}{n h_c^2}.$$

Denominator Correction

Proof Idea: Completing the AMSE Criterion

Recall

$$\mathcal{A}_c = \min_{a \in \mathbb{R}} \bar{\mathbb{E}}[(\widehat{\Pi}_{c,i} - a\Pi_i)^2].$$

The minimizer is $a = h_c / \mathbb{E}_n[\Pi_i^2]$. Substituting gives

$$\mathcal{A}_c = \frac{\mathbb{E}[\|\widehat{\Pi}_c\|_2^2]}{n} - \frac{h_c^2}{\mathbb{E}_n[\Pi_i^2]}.$$

This lets us express the squared-fit term in the numerator moment using $\mathcal{A}_c$.

Proof Idea: Completing the AMSE Criterion

We showed that

$$\mathbb{E}[(\varepsilon'\widehat{\Pi}_c)^2] = \sigma_\varepsilon^2 \mathbb{E}[\|\widehat{\Pi}_c\|_2^2] + \sigma_{\varepsilon v}^2 \mathbb{E}[d_c^2 + d_c].$$

Substitute $\mathbb{E}[\|\widehat{\Pi}_c\|_2^2] = n(\mathcal{A}_c + h_c^2/\mathbb{E}_n[\Pi_i^2])$ and divide by nh_c^2 to obtain

$$\frac{\mathbb{E}[(\varepsilon'\widehat{\Pi}_c)^2]}{nh_c^2} = \underbrace{\frac{\sigma_\varepsilon^2}{\mathbb{E}_n[\Pi_i^2]}}_{\text{common}} + \underbrace{\frac{\sigma_\varepsilon^2 \mathcal{A}_c}{h_c^2} + \frac{\sigma_{\varepsilon v}^2 \mathbb{E}[d_c^2 + d_c]}{nh_c^2}}_{S_c}.$$

Donald–Newey Comparison

Table of Contents

Theory and Infeasible Criterion

Feasible Selection

Non-Gaussian Extension

Empirically Calibrated Simulations

Constructing a Feasible Criterion

The infeasible criterion S_c depends on population moments that are unknown to the researcher. Instead, use feasible criterion

$$\widehat{S}_c := \frac{\widehat{\sigma}_\varepsilon^2 \mathbb{E}_n[\widehat{\Pi}_{c,i}^2] + n^{-1}\widehat{\sigma}_{\varepsilon v}^2(d_c^2 + d_c)}{\widehat{h}_c^2}$$

where $\widehat{\sigma}_\varepsilon^2$ and $\widehat{\sigma}_{\varepsilon v}^2$ are constructed using a pilot estimate $\widehat{\beta}$.

- Approach differs from Donald and Newey (2001) who instead try and directly estimate their analog of $\mathcal{A}_c$.
- Showing this criterion allows for feasible selection requires a different approach.
  - Key idea: Ratio $\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\widehat{h}_c^2}$ allows us to indirectly measure $\mathcal{A}_c$

Estimating the Moments

Constructing a Feasible Criterion

Estimation error in $\widehat{S}_c$ can be of higher order than S_c itself, which is a problem for feasible selection. However, leading part of estimation error is common to all candidates. Let

$$B_n := \widehat{\sigma}_\varepsilon^2 \left(\frac{1}{\mathbb{E}_n[\Pi_i^2]} - \frac{2\mathbb{E}_n[\Pi_i v_i]}{\{\mathbb{E}_n[\Pi_i^2]\}^2} \right)$$

be the (unobserved) common centering term and choose $\hat{c} \in \arg\min_c \widehat{S}_c$.

Feasible Selection

Theorem 2 (Feasible AMSE Selection)

The feasible criterion estimates the infeasible criterion up to the common centering term:

$$\frac{|\widehat{S}_c - B_n - S_c|}{S_c} \to_p 0.$$

Thus the selected candidate $\hat{c}$ satisfies

$$\frac{S_{\hat{c}}}{\min_c S_c} \to_p 1.$$

Since $\mathbb{E}[Q_c] = S_c(1 + o(1))$, the feasible selection is asymptotically AMSE optimal.

Proof Idea: Decomposing the Fitted Moment

Leading estimation error enters through $\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]$, which is used in place of infeasible

$$\widehat{\mathcal{A}}_c = \min_{a \in \mathbb{R}} \mathbb{E}_n[(\widehat{\Pi}_{c,i} - a\Pi_i)^2].$$

Same decomposition as for $\mathcal{A}_c$ applied to sample moments gives

$$\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\widehat{h}_c^2} = \frac{\widehat{\mathcal{A}}_c}{\widehat{h}_c^2} + \frac{1}{\mathbb{E}_n[\Pi_i^2]} \left(\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}\Pi_i]}{\widehat{h}_c} \right)^2.$$

The first part estimates the first term of S_c, the second yields the common centering.

Proof Idea: Deriving the Centering Term

Examine second term further. Since $x_i = \Pi_i + v_i$ and $\widehat{h}_c = \mathbb{E}_n[x_i \widehat{\Pi}_{c,i}]$,

$$\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}\Pi_i]}{\widehat{h}_c} = 1 - \frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}v_i]}{\widehat{h}_c}.$$

The Stein bounds and the first-stage conditions give

$$\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}v_i]}{\widehat{h}_c} = \frac{\mathbb{E}_n[\Pi_i v_i]}{\mathbb{E}_n[\Pi_i^2]} + o_p(S_c).$$

The leading term does not depend on the candidate $c \in [\mathcal{J}]$.

Proof Idea: Deriving the Centering Term

Plugging the above expansion into the second term of the decomposition of $\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]/\widehat{h}_c^2$ and expanding the square gives

$$\frac{\widehat{\sigma}_\varepsilon^2}{\mathbb{E}_n[\Pi_i^2]}\left(\frac{\mathbb{E}_n[\widehat{\Pi}_{c,i}\Pi_i]}{\widehat{h}_c}\right)^2 = \underbrace{\widehat{\sigma}_\varepsilon^2\left(\frac{1}{\mathbb{E}_n[\Pi_i^2]} - \frac{2\mathbb{E}_n[\Pi_i \nu_i]}{\{\mathbb{E}_n[\Pi_i^2]\}^2}\right)}_{B_n} + o_p(S_c).$$

So that in total

$$\begin{aligned}\frac{\widehat{\sigma}_\varepsilon^2\mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\widehat{h}_c^2} &= \frac{\widehat{\sigma}_\varepsilon^2\widehat{\mathcal{A}}_c}{\widehat{h}_c^2} + B_n + o_p(S_c)\\ &= \frac{\sigma_\varepsilon^2\mathcal{A}_c}{h_c^2} + B_n + o_p(S_c)\end{aligned}$$

Proof Idea: Completing Feasible Consistency

To estimate the second part of the infeasible criterion show concentration of effective dimension around its expectation

$$\frac{d_c^2 + d_c}{\mathbb{E}[d_c^2 + d_c]} \to_p 1$$

In total,

$$\begin{aligned}\widehat{S}_c &= \frac{\widehat{\sigma}_\varepsilon^2 \mathbb{E}_n[\widehat{\Pi}_{c,i}^2]}{\hat{h}_c^2} + \frac{\hat{\sigma}_{\varepsilon v}^2 (d_c^2 + d_c)}{n\widehat{h}_c^2} \\ &= B_n + \underbrace{\frac{\sigma_\varepsilon^2 \mathcal{A}_c}{h_c^2} + \frac{\sigma_{\varepsilon v}^2 \mathbb{E}[d_c^2 + d_c]}{n h_c^2}}_{S_c} + o_p(S_c).\end{aligned}$$

Subtract B_n and divide by S_c to obtain Theorem 2.

Table of Contents

Theory and Infeasible Criterion

Feasible Selection

Non-Gaussian Extension

Empirically Calibrated Simulations

Non-Gaussian Errors

Consider smooth-transformation errors, as in Chetverikov et al. (2020):

$$v_i = \psi(e_i), \qquad e_i \overset{iid}{\sim} N(0,1), \qquad \mathbb{E}[\psi(e_i)] = 0.$$

The transformation is twice continuously differentiable, with $0 < c_\psi \leq \psi'(t) \leq C_\psi$. This allows asymmetric, non-Gaussian errors.

For the structural error, assume

$$\varepsilon_i = a v_i + u_i, \qquad \mathbb{E}[u_i \mid v_i] = 0, \qquad \mathbb{E}[u_i^2 \mid v_i] = \sigma_u^2.$$

Pairs (ε_i, v_i) are i.i.d., with $a \neq 0$ and $\sigma_u^2 > 0$.

Exact Conditions

Non-Gaussian Selection

Theorem 3 (Non-Gaussian Extension)

Under the non-Gaussian conditions:

(i) *Decomposition* (1) *holds with* $\mathbb{E}[Q_c] = S_c(1 + o(1))$, *for the same* S_c*;*

(ii) *With a* $\sqrt{n}$*-consistent pilot, the same* $\widehat{S}_c$ *and* B_n *give*

$$\max_c \frac{|\widehat{S}_c - B_n - S_c|}{S_c} \to_p 0, \qquad \frac{S_{\hat{c}}}{\min_c S_c} \to_p 1.$$

Proof Idea: Structural Moments

Fix a candidate and drop the subscript c. Project the structural error on the first-stage error:

$$\varepsilon = av + u, \qquad a = \frac{\sigma_{\varepsilon v}}{\sigma_v^2}, \qquad \mathbb{E}[u \mid v] = 0, \qquad \mathbb{E}[uu' \mid v] = \sigma_u^2 I_n.$$

Since $\widehat{\Pi}$ is a function of v, the cross term still drops out:

$$\mathbb{E}[(\varepsilon'\widehat{\Pi})^2] = a^2\mathbb{E}[(v'\widehat{\Pi})^2] + \sigma_u^2\mathbb{E}[\|\widehat{\Pi}\|_2^2].$$

Need to characterize $\mathbb{E}[(v'\widehat{\Pi})^2]$ but no longer have access to exact Stein identities.

Proof Idea: Weighted Stein Identities

Define the Stein kernel τ that makes integration by parts work

$$\tau(t) := \frac{\int_t^\infty z p_v(z)\, dz}{p_v(t)}, \qquad \mathbb{E}[v_i f(v_i)] = \mathbb{E}[\tau(v_i) f'(v_i)]$$

Collect the weights in $T := \operatorname{diag}(\tau(v_1), \ldots, \tau(v_n))$. If $v_i \sim N(0, \sigma_v^2)$, $T = \sigma_v^2 I_n$.

Lemma 2 (Weighted Stein Identities)

If $m : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz with Jacobian $A = Dm$, then

$$\mathbb{E}[v' m] = \mathbb{E}[\operatorname{tr}(TA)],$$

$$\mathbb{E}\left[\left(v' m - \operatorname{tr}(TA)\right)^2\right] = \mathbb{E}[m' T m] + \mathbb{E}[\operatorname{tr}(TATA)].$$

Stein Kernel

Proof Idea: The Second Moment with a Kernel

Apply the identities to $m = \widehat{\Pi}$, whose Jacobian $P := D\widehat{\Pi}$ is, as before, the projection onto the selected columns:

$$\mathbb{E}[(v'\widehat{\Pi})^2] = \mathbb{E}[\mathrm{tr}(TP)^2] + \mathbb{E}[\widehat{\Pi}' T\widehat{\Pi}] + 2\mathbb{E}[\widehat{\Pi}' TPv] - \mathbb{E}[\mathrm{tr}(TPTP)].$$

We can show $\mathbb{E}[T] = \sigma_v^2 I_n$. Let $W = \sigma_v^2 I_n - T$ and expand:

$$\begin{aligned}\mathbb{E}[(v'\widehat{\Pi})^2] = \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_v^4 \mathbb{E}[d^2 + d] \qquad &\text{(Gaussian)} \\ + \mathbb{E}[\widehat{\Pi}' W\widehat{\Pi}] + 2\mathbb{E}[\widehat{\Pi}' WPv] \\ + \mathbb{E}\left[2\sigma_v^2 d\,\mathrm{tr}(PWP) + \mathrm{tr}(PWP)^2 - \|PWP\|_F^2\right]\end{aligned}$$

Proof Idea: Why W Averages Out

To control the terms in the remainder use two facts

1. $\mathbb{E}[W] = 0$ so $\mathbb{E}[\Pi' W \Pi] = 0$. This controls terms with $\widehat{\Pi}$.
2. W is diagonal with independent, bounded, and mean zero entries. Uniformly over spans B of $\lesssim s$ columns (s the first-stage sparsity),

$$\|P_B W P_B\|_{\text{op}} \lesssim s\sqrt{\frac{\log(pn)}{n}} \to 0.$$

Together, these show that the above remainder terms are $o(n\mathcal{A} + s^2)$.

Term-by-Term Bounds

Proof Idea: Where the Kernel Goes

The kernel therefore only affects the expansion through its mean $\mathbb{E}[\tau(v_i)] = \sigma_v^2$:

$$\mathbb{E}[(v'\widehat{\Pi})^2] = \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_v^4 \mathbb{E}[d^2 + d] + o(n\mathcal{A} + s^2).$$

Returning to the expression for the structural cross-moment and using $a = \frac{\sigma_{v\varepsilon}}{\sigma_\varepsilon^2}$,

$$\mathbb{E}[(\varepsilon'\widehat{\Pi})^2] = \sigma_\varepsilon^2 \mathbb{E}[\|\widehat{\Pi}\|_2^2] + \sigma_{\varepsilon v}^2 \mathbb{E}[d^2 + d] + o(n\mathcal{A} + s^2).$$

This is the Gaussian numerator moment, up to negligible terms, so S_c is unchanged.

Proof Idea: Why a Gaussian Base

The standard normal has Stein kernel one: $\int_z^\infty t\varphi(t)\,dt = \varphi(z)$. Changing variables,

$$\tau(\psi(z)) = \psi'(z) \cdot \underbrace{\frac{\int_z^\infty \psi(t)\varphi(t)\,dt}{\varphi(z)}}_{\in[c_\psi, C_\psi]},$$

so $c_\psi^2 \le \tau \le C_\psi^2$ and W is bounded.

- Gaussian concentration passes through the Lipschitz map ψ, replacing the Gaussian tail and Poincaré bounds in the first-stage arguments;
- Laplace errors, by contrast, have the unbounded kernel $\tau(t) = \frac{1}{2} + |t|/\sqrt{2}$.

Table of Contents

Theory and Infeasible Criterion

Feasible Selection

Non-Gaussian Extension

Empirically Calibrated Simulations

Calibrated Simulation Design

Calibrate to Paravisini et al. (2015), who study the effect of credit supply on exports.

- Hold the observed exposure instruments fixed, with $n = 14{,}208$ export flows;
- Use fitted credit growth as the true first stage and set $\beta = 1$;
- Start with Gaussian errors and vary identification strength and endogeneity:

$$F \in \{2.7,\ 4.4,\ 7.7,\ 14.5,\ 28.0\},$$

$$\rho = c\widehat{\rho}, \qquad c = 0,\ 0.25, \ldots, 2.75, \qquad \widehat{\rho} = -0.218.$$

Two first-stage signals (main and dense), with 1,000 replications per design.

Empirical Setting Calibration Details

Candidate Implementations and Selection Rules

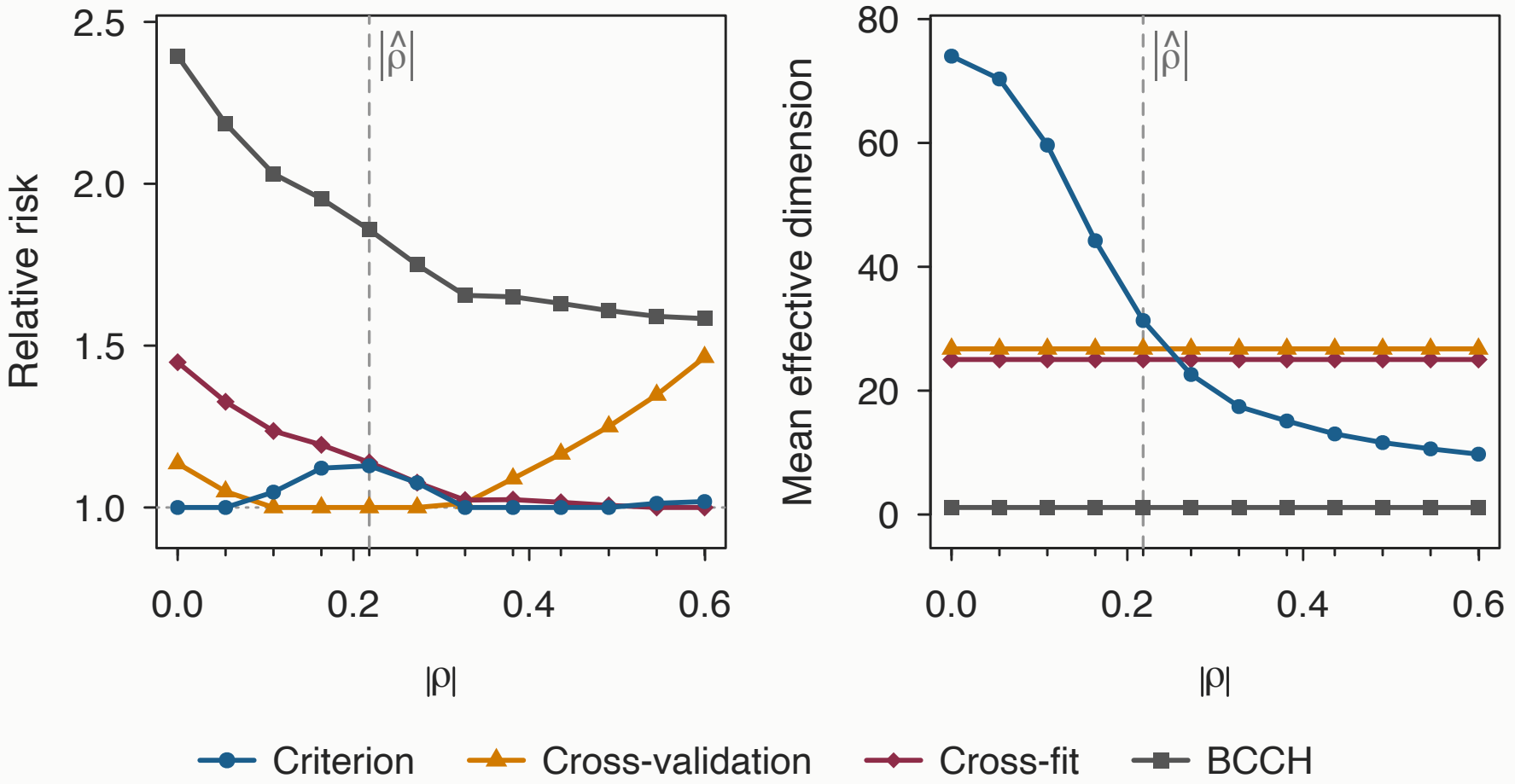
Two dictionaries (10 or 114 instruments), each paired with 37 penalty levels, give 74 candidate implementations.

Rule	Choice of first stage
Criterion	Minimize the feasible structural AMSE criterion
Cross-validation	Minimize ten-fold first-stage prediction error
Cross-fit	Tune within each training sample; use held-out fits as instruments
BCCH	Plug-in penalty on the Original dictionary

Report risk as the average of $n(\widehat{\beta} - \beta)^2$. Relative risk divides this by the lowest risk among these four rules in the same design.

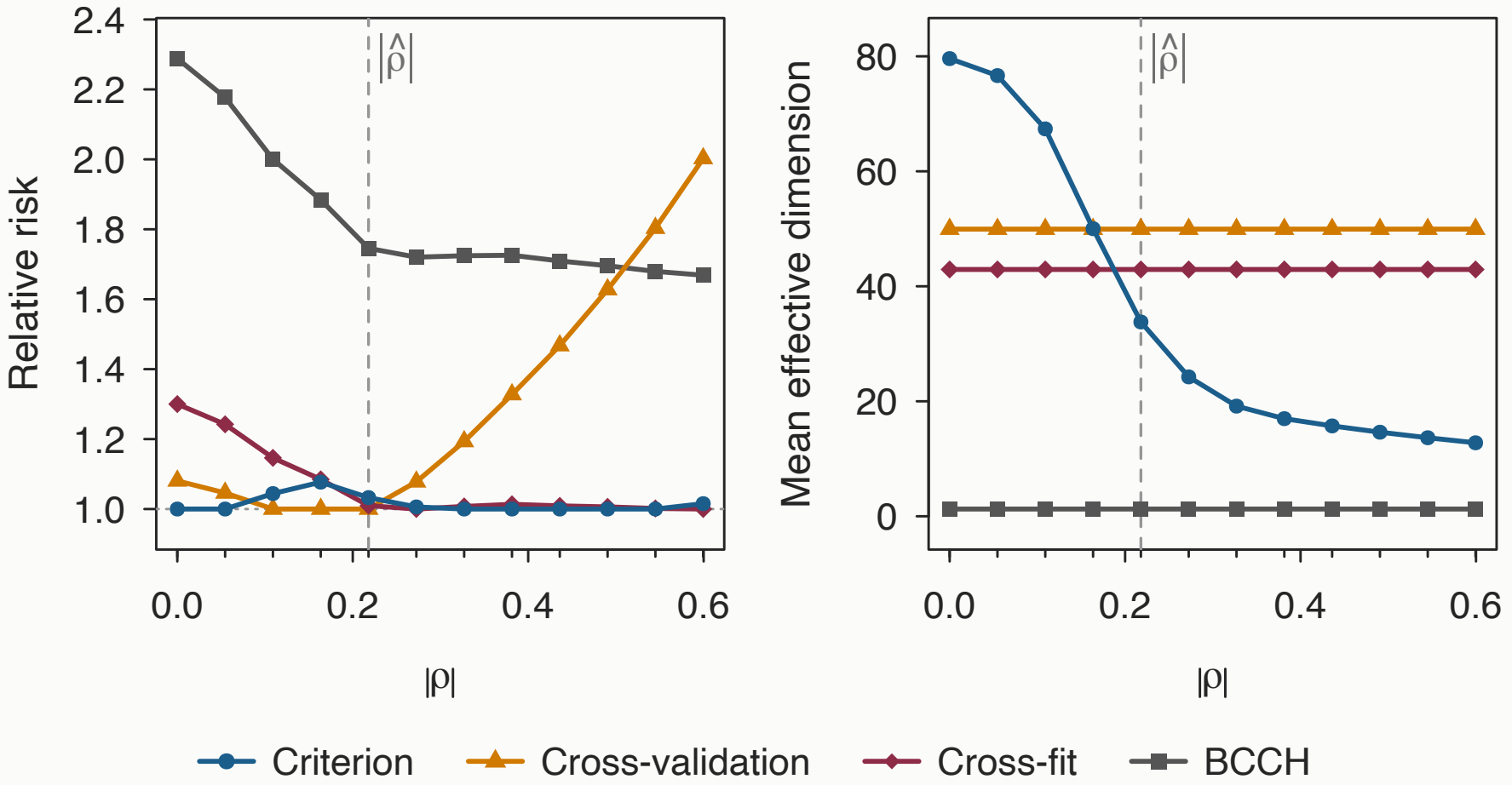
Implementation Details

Gaussian Errors: $F = 7.7$



Other Identification Strengths

Gaussian Errors: $F = 14.5$



Stronger Identification

Performance Across Identification Strengths

Average relative risk across the twelve correlations, using the main signal:

F	Criterion	CV	Cross-fit	BCCH
2.7	1.072	1.034	1.589	14.529
4.4	1.051	1.074	1.240	1.813
7.7	1.034	1.126	1.124	1.824
14.5	1.015	1.302	1.068	1.835
28.0	1.036	1.301	1.038	1.908

- Criterion is within 20% of the best rule in every main-signal design;
- CV can do better with weak identification or intermediate endogeneity.

The large BCCH entry at $F = 2.7$ reflects rare empty fits and use of the pilot fallback.

Non-Gaussian Errors

Repeat the same designs with smooth-transformation and common-scale Laplace errors. The criterion's relative risk remains similar:

	Main signal		Dense signal	
Errors	Average	Maximum	Average	Maximum
Gaussian	1.042	1.199	1.035	1.175
Smooth transformation	1.045	1.188	1.041	1.221
Common-scale Laplace	1.041	1.174	1.035	1.193

The smooth-transformation design satisfies the extension's conditions on the errors, the Laplace design does not.

Error Designs | Smooth Transformation | Laplace

Conclusion

Choose the dictionary and penalty to target accuracy of the structural estimate.

- Second-order AMSE balances first-stage fit against a cost of effective dimension that increases with endogeneity;
- Feasible selection is asymptotically AMSE-optimal among the candidate implementations, under Gaussian errors and a class of non-Gaussian errors;
- In calibrated simulations, the criterion adapts to endogeneity and can substantially reduce risk relative to cross-validation and plug-in rules.

References

Agan, Amanda, Jennifer L. Doleac, and Anna Harvey (2023). "Misdemeanor Prosecution". In: *The Quarterly Journal of Economics* 138.3, pp. 1453–1505. DOI: 10.1093/qje/qjad005.

Bald, Anthony et al. (2022). "The Causal Impact of Removing Children from Abusive and Neglectful Homes". In: *Journal of Political Economy* 130.7, pp. 1919–1962. DOI: 10.1086/719856.

Bazzi, Samuel et al. (2023). "The Other Great Migration: Southern Whites and the New Right". In: *The Quarterly Journal of Economics* 138.3, pp. 1577–1647. DOI: 10.1093/qje/qjad014.

— (2026). "The Confederate Diaspora". In: *The Review of Economic Studies*. DOI: 10.1093/restud/rdag027. Forthcoming.

Bellec, Pierre C. and Cun-Hui Zhang (2021). "Second-Order Stein: SURE for SURE and Other Applications in High-Dimensional Inference". In: *The Annals of Statistics* 49.4, pp. 1864–1903. DOI: 10.1214/20-AOS2005.

Belloni, Alexandre and Victor Chernozhukov (2013). "Least Squares After Model Selection in High-Dimensional Sparse Models". In: *Bernoulli* 19.2, pp. 521–547. DOI: 10.3150/11-BEJ410.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on Treatment Effects after Selection among High-Dimensional Controls". In: *The Review of Economic Studies* 81.2, pp. 608–650. DOI: 10.1093/restud/rdt044.

References

Belloni, Alexandre et al. (2012). "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". In: *Econometrica* 80.6, pp. 2369–2429. DOI: 10.3982/ECTA9626.

Beraja, Martin et al. (2023). "AI-tocracy". In: *The Quarterly Journal of Economics* 138.3, pp. 1349–1402. DOI: 10.1093/qje/qjad012.

Carrasco, Marine (2012). "A Regularization Approach to the Many Instruments Problem". In: *Journal of Econometrics* 170.2, pp. 383–398. DOI: 10.1016/j.jeconom.2012.05.012.

Chernozhukov, Victor et al. (2018). "Double/Debiased Machine Learning for Treatment and Structural Parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. DOI: 10.1111/ectj.12097.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov (2020). *On Cross-Validated Lasso in High Dimensions*. arXiv:1605.02214v6, extended version. URL: https://arxiv.org/abs/1605.02214v6.

— (2021). "On Cross-Validated Lasso in High Dimensions". In: *The Annals of Statistics* 49.3, pp. 1300–1317. DOI: 10.1214/20-AOS2000.

Chetverikov, Denis and Jesper Riis-Vestergaard Sørensen (2025). "Selecting Penalty Parameters of High-Dimensional M-Estimators Using Bootstrapping after Cross Validation". In: *Journal of Political Economy* 133.10, pp. 3208–3248. DOI: 10.1086/736770.

References

Cuddy, Emily and Janet Currie (2026). "Rules versus Discretion: Treatment of Mental Illness in US Adolescents". In: *Journal of Political Economy* 134.1, pp. 478–522. DOI: 10.1086/738249.

Derenoncourt, Ellora (2022). "Can You Move to Opportunity? Evidence from the Great Migration". In: *American Economic Review* 112.2, pp. 369–408. DOI: 10.1257/aer.20200002.

Do, Quy-Toan and Hanan G. Jacoby (2025). "Optimal Pricing of a New Utility Service: The Case of Piped Water in Vietnam". In: *The Review of Economic Studies* 92.3, pp. 1773–1800. DOI: 10.1093/restud/rdae074.

Donald, Stephen G. and Whitney K. Newey (2001). "Choosing the Number of Instruments". In: *Econometrica* 69.5, pp. 1161–1191. DOI: 10.1111/1468-0262.00238.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021). "Deep Neural Networks for Estimation and Inference". In: *Econometrica* 89.1, pp. 181–213. DOI: 10.3982/ECTA16901.

Gilchrist, Duncan Sheppard and Emily Glassberg Sands (2016). "Something to Talk About: Social Spillovers in Movie Consumption". In: *Journal of Political Economy* 124.5, pp. 1339–1382. DOI: 10.1086/688177.

Hahn, Jinyong, Jerry Hausman, and Guido Kuersteiner (2004). "Estimation with Weak Instruments: Accuracy of Higher-Order Bias and MSE Approximations". In: *The Econometrics Journal* 7.1, pp. 272–306. DOI: 10.1111/j.1368-423X.2004.00131.x.

References

Kuersteiner, Guido and Ryo Okui (2010). "Constructing Optimal Instruments by First-Stage Prediction Averaging". In: *Econometrica* 78.2, pp. 697–718. DOI: 10.3982/ECTA7444.

Nagar, Anirudh L. (1959). "The Bias and Moment Matrix of the General *k*-Class Estimators of the Parameters in Simultaneous Equations". In: *Econometrica* 27.4, pp. 575–595. DOI: 10.2307/1909352.

Newey, Whitney K. and Richard J. Smith (2004). "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators". In: *Econometrica* 72.1, pp. 219–255. DOI: 10.1111/j.1468-0262.2004.00482.x.

Okui, Ryo (2011). "Instrumental Variable Estimation in the Presence of Many Moment Conditions". In: *Journal of Econometrics* 165.1, pp. 70–86. DOI: 10.1016/j.jeconom.2011.05.007.

Paravisini, Daniel et al. (2015). "Dissecting the Effect of Credit Supply on Trade: Evidence from Matched Credit-Export Data". In: *The Review of Economic Studies* 82.1, pp. 333–359. DOI: 10.1093/restud/rdu028.

Reimers, Imke and Joel Waldfogel (2021). "Digitization and Pre-purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings". In: *American Economic Review* 111.6, pp. 1944–1971. DOI: 10.1257/aer.20200153.

References

Rothenberg, Thomas J. (1984). "Approximating the Distributions of Econometric Estimators and Test Statistics". In: *Handbook of Econometrics*. Ed. by Zvi Griliches and Michael D. Intriligator. Vol. 2. Elsevier. Chap. 15, pp. 881–935. DOI: 10.1016/S1573-4412(84)02007-9.

Ryan, Nicholas and Anant Sudarshan (2022). "Rationing the Commons". In: *Journal of Political Economy* 130.1, pp. 210–257. DOI: 10.1086/717045.

Sargan, J. Denis (1976). "Econometric Estimators and the Edgeworth Approximation". In: *Econometrica* 44.3, pp. 421–448. DOI: 10.2307/1913972.

Tibshirani, Ryan J. and Jonathan Taylor (2012). "Degrees of Freedom in Lasso Problems". In: *The Annals of Statistics* 40.2, pp. 1198–1232. DOI: 10.1214/12-AOS1003.

Velez, Amilcar (2026). "Debiased Machine Learning with Many Cross-Fitting Folds". Working Paper, Cornell University, arXiv:2411.01864v2.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242. DOI: 10.1080/01621459.2017.1319839.

Wainwright, Martin J. (2009). "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE Transactions on Information Theory* 55.5, pp. 2183–2202. DOI: 10.1109/TIT.2009.2016018.

References

Zhao, Peng and Bin Yu (2006). "On Model Selection Consistency of Lasso". In: *Journal of Machine Learning Research* 7, pp. 2541–2563.

MSE Nonexistence

Consider the infeasible estimator that uses the true first-stage

$$\widehat{\beta}^{\circ} := \frac{\mathbb{E}_n[y_i \Pi_i]}{\mathbb{E}_n[x_i \Pi_i]}$$

Under our assumptions, this is a ratio of two normal distributions and thus does not have any finite moments.

Back

Proof Idea: Residual Projection

Let $\widehat{u}_{c,i} = x_i - \widehat{\Pi}_{c,i}$. The LASSO KKT conditions imply

$$\left\|\mathbb{E}_n[z_{c,i}\widehat{u}_{c,i}]\right\|_\infty \le \lambda_c, \qquad \widehat{\pi}_c'\mathbb{E}_n[z_{c,i}\widehat{u}_{c,i}] = \lambda_c\|\widehat{\pi}_c\|_1.$$

For any $q \in \mathcal{K}_c$, Hölder gives $\widehat{\pi}_c'\mathbb{E}_n[z_{c,i}q_i] \le \lambda_c\|\widehat{\pi}_c\|_1$. Hence

$$\mathbb{E}_n[\widehat{\Pi}_{c,i}(q_i - \widehat{u}_{c,i})] = \widehat{\pi}_c'\mathbb{E}_n[z_{c,i}q_i] - \lambda_c\|\widehat{\pi}_c\|_1 \le 0.$$

Expanding the squared distance, we obtain

$$\begin{aligned}&\mathbb{E}_n[(x_i - q_i)^2] - \mathbb{E}_n[(x_i - \widehat{u}_{c,i})^2] \\ &\quad = \mathbb{E}_n[(q_i - \widehat{u}_{c,i})^2] - 2\mathbb{E}_n[\widehat{\Pi}_{c,i}(q_i - \widehat{u}_{c,i})] \ge 0.\end{aligned}$$

Since $\widehat{u}_c \in \mathcal{K}_c$, the residual is the closest feasible point:

$$\widehat{u}_c = \arg\min_{q\in\mathcal{K}_c} \mathbb{E}_n[(x_i - q_i)^2] = P_{\mathcal{K}_c}(x).$$

Back

Proof Idea: The Stein Cross Term

Write $\zeta_c = v'\widehat{\Pi}_c - \sigma_v^2 d_c$, $P_c = D\widehat{\Pi}_c$, and $g(v) = v'\widehat{\Pi}_c$. Since $P_c' = P_c$ and $P_c\widehat{\Pi}_c = \widehat{\Pi}_c$,

$$\widehat{\Pi}_c'\nabla g(v) = \widehat{\Pi}_c'(\widehat{\Pi}_c + P_c v) = \|\widehat{\Pi}_c\|_2^2 + v'\widehat{\Pi}_c.$$

First-order Stein applied to $g(v)\widehat{\Pi}_c$ gives

$$\begin{aligned}\mathbb{E}[\zeta_c g(v)] &= \sigma_v^2 \mathbb{E}[\widehat{\Pi}_c'\nabla g(v)] \\ &= \sigma_v^2 \mathbb{E}[\|\widehat{\Pi}_c\|_2^2] + \sigma_v^4 \mathbb{E}[d_c] = \mathbb{E}[\zeta_c^2].\end{aligned}$$

Thus the two Stein identities give the same answer. Subtracting,

$$\sigma_v^2 \mathbb{E}[d_c \zeta_c] = \mathbb{E}[\zeta_c\{g(v) - \zeta_c\}] = 0.$$

Back

Connection to Donald and Newey

Interestingly, our criterion has the same structure as that of Donald and Newey (2001) for an OLS first stage.

Let P_K project onto K prespecified, linearly independent instruments. The OLS fit is $\widehat{\Pi}_K = P_K x$, with $d_K = K$.

Write $B_K = \|(I - P_K)\Pi\|_2^2/n$, so $h_K = \mathbb{E}_n[\Pi_i^2] - B_K$. The same moment calculation gives

$$\frac{\mathbb{E}[(\varepsilon'\widehat{\Pi}_K)^2]}{nh_K^2} = \frac{\sigma_\varepsilon^2}{\mathbb{E}_n[\Pi_i^2]} + \frac{\sigma_\varepsilon^2 B_K}{\mathbb{E}_n[\Pi_i^2]h_K} + \frac{\sigma_\varepsilon^2\sigma_v^2 K + \sigma_{\varepsilon v}^2 K(K+1)}{nh_K^2}.$$

Under their growing-K conditions, $h_K - \mathbb{E}_n[\Pi_i^2] \to 0$ and the K/n terms are lower order. We recover their criterion:

$$\frac{\mathbb{E}[(\varepsilon'\widehat{\Pi}_K)^2]}{nh_K^2} = \frac{\sigma_\varepsilon^2}{\mathbb{E}_n[\Pi_i^2]} + \underbrace{\frac{\sigma_\varepsilon^2 B_K}{\{\mathbb{E}_n[\Pi_i^2]\}^2} + \frac{\sigma_{\varepsilon v}^2 K^2}{n\{\mathbb{E}_n[\Pi_i^2]\}^2}}_{S_K^{\mathrm{DN}}} + o(S_K^{\mathrm{DN}}).$$

Back

Estimating the Error Moments

Given a root-n consistent pilot $\widehat{\beta}$, form residuals

$$\widehat{\varepsilon}_i = y_i - x_i\widehat{\beta}.$$

Estimate the error variance and covariance by

$$\widehat{\sigma}_\varepsilon^2 = \mathbb{E}_n[\widehat{\varepsilon}_i^2], \qquad \widehat{\sigma}_{\varepsilon v} = \mathbb{E}_n[\widehat{\varepsilon}_i x_i].$$

The second formula uses $\mathbb{E}[\varepsilon_i x_i] = \mathbb{E}[\varepsilon_i(\Pi_i + v_i)] = \sigma_{\varepsilon v}$.

Use these same moment estimates for every candidate, squaring $\widehat{\sigma}_{\varepsilon v}$ in $\widehat{S}_c$.

Back

Why the Cross-Term Is Negligible

In the squared-error expansion, the denominator cross-term satisfies

$$\left| \mathbb{E}\left[\left(\frac{\varepsilon' \widehat{\Pi}_c}{\sqrt{n}\, h_c} \right)^2 \frac{\widehat{h}_c - h_c}{h_c} \right] \right| = o(S_c).$$

Intuition: with the true first stage, Gaussian symmetry makes this expectation zero. Stein bounds show that estimating the first stage adds only a smaller-order term.

Back

Non-Gaussian Sampling Conditions

Conditional on the instruments, pairs (ε_i, v_i) are i.i.d. with a fixed law, and:

(i) $\varepsilon_i = a v_i + u_i$ with $a \neq 0$, $\mathbb{E}[u_i \mid v_i] = 0$, $\mathbb{E}[u_i^2 \mid v_i] = \sigma_u^2 > 0$, and $\mathbb{E}[|u_i|^4] \leq C_u$;

(ii) $v_i = \psi(e_i)$ for $e_i \sim N(0,1)$, with $\mathbb{E}[\psi(e_i)] = 0$ and

$$\psi \in C^2(\mathbb{R}), \qquad 0 < c_\psi \leq \psi'(t) \leq C_\psi < \infty;$$

(iii) Strong identification, $\underline{H} \leq \mathbb{E}_n[\Pi_i^2] \leq \overline{H}$, and $\max_i |\Pi_i| \leq K_\Pi$.

Then $\sigma_{\varepsilon v} = a\sigma_v^2$ and $\sigma_\varepsilon^2 = a^2\sigma_v^2 + \sigma_u^2$. The Gaussian case is $\psi(t) = \sigma_v t$.

Back

Why the Stein Kernel Appears

The Stein kernel is chosen so that $(\tau p_v)'(t) = -t p_v(t)$. For a smooth, compactly supported function f, integration by parts gives

$$\begin{aligned}\mathbb{E}[v_i f(v_i)] &= -\int f(t)(\tau p_v)'(t)\,dt \\ &= \int f'(t)\tau(t)p_v(t)\,dt = \mathbb{E}[\tau(v_i)f'(v_i)].\end{aligned}$$

Apply this identity coordinate by coordinate to obtain the weighted first-order identity. A second integration by parts gives the second-order identity; terms involving derivatives of the kernel cancel.

Smooth approximation extends the identities to Lipschitz fits such as LASSO.

Back

Bounding the W-Terms

Let $\eta := s\sqrt{\log(pn)/n}$. With high probability, uniformly over spans B of $\lesssim s$ columns,

$$\|P_B W P_B\|_{\mathrm{op}} \lesssim \eta, \qquad \|P_B W \Pi\|_2 \lesssim \sqrt{s \log(pn)}.$$

Write $\widehat{\Pi} = \kappa\Pi + m$, with $\kappa\Pi$ the best multiple of the signal, so $\mathbb{E}[\|m\|_2^2] = n\mathcal{A}$.

1. $\mathbb{E}[\Pi' W \Pi] = 0$, and the rest of $\widehat{\Pi}' W \widehat{\Pi}$ is $\lesssim \|P_B W \Pi\|_2 \|m\|_2 + \eta \|m\|_2^2$;
2. $|\widehat{\Pi}' W P v| \lesssim (\|PW\Pi\|_2 + \|m\|_2)\|Pv\|_2$, with $\mathbb{E}[\|Pv\|_2^2] \lesssim s \log(pn)$;
3. $|\operatorname{tr}(PWP)| \le \eta d$ and $\|PWP\|_F^2 \le \eta^2 d$ since $\operatorname{rank}(P) = \operatorname{trace}(P) = d$.

Since $d \asymp s$, $\eta \to 0$, $\log(pn)/s \to 0$, each is $o(n\mathcal{A} + s^2)$.

Back

The Paravisini et al. Setting

Paravisini et al. (2015) study Peruvian exporters in the twelve months before and after July 2008.

- **Outcome:** Change in log export volume (kilograms) for each firm–product–destination flow;
- **Endogenous variable:** Change in log total bank credit to the firm;
- **Instruments:** Firms' 2006 exposure to foreign-funded banks. The main measure is the share of credit from banks with foreign funding above 10% of assets; the dictionary also includes threshold and polynomial exposure measures.

The 2008 capital-flow reversal reduced lending by these banks. Product–destination fixed effects compare exporters facing the same export-market shocks.

Back

Simulation Calibration

The PRSW sample contains 14,208 export flows from 2,091 firms. Credit growth and instruments are demeaned within product–destination groups.

- **Main signal:** OLS refit on 17 instruments selected by CV-LASSO from the ten original instruments and their 45 pairwise products;
- **Dense signal:** OLS on all 55 instruments, rescaled to the same $H = \mathbb{E}_n[\Pi_i^2]$.

Generate data with $\beta = 1$ and

$$v_i, e_i \overset{iid}{\sim} N(0,1), \qquad \varepsilon_i = \rho v_i + \sqrt{1-\rho^2}\, e_i.$$

Rescale Π to vary the i.i.d.-equivalent expected statistic $F = 1 + nH/51$. The empirical signal strength corresponds to $F = 14.5$. These values differ from the firm-clustered statistic in the original data.

Back

Simulation Implementation

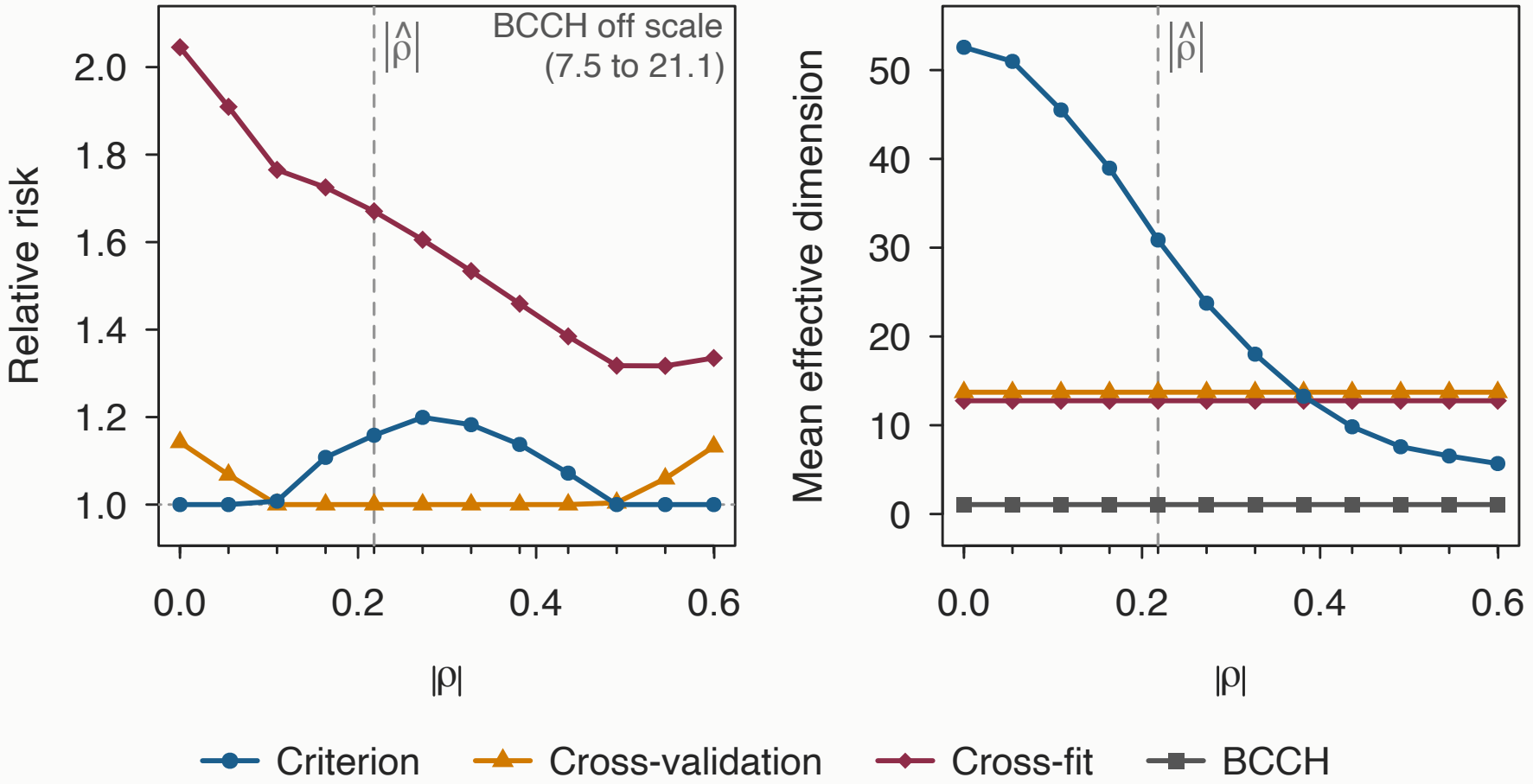
The penalty grid contains 37 multiples of the BCCH plug-in penalty, $2^{-15}, 2^{-14.5}, \ldots, 2^{3}$, for each dictionary.

- CV and cross-fitting keep observations from each firm in the same fold. Cross-fit uses ten outer folds and tunes by ten-fold CV within each outer training sample;
- The criterion estimates error moments using a cross-fitted IV pilot with BCCH fits on the Original dictionary;
- When full-sample BCCH selects no instruments, its IV estimate is replaced by the pilot.

Plug-in penalties use the known first-stage noise variance. Cross-fit's dimension averages fold-specific dimensions, weighted by held-out fold size.

Back

Gaussian Errors: $F = 2.7$

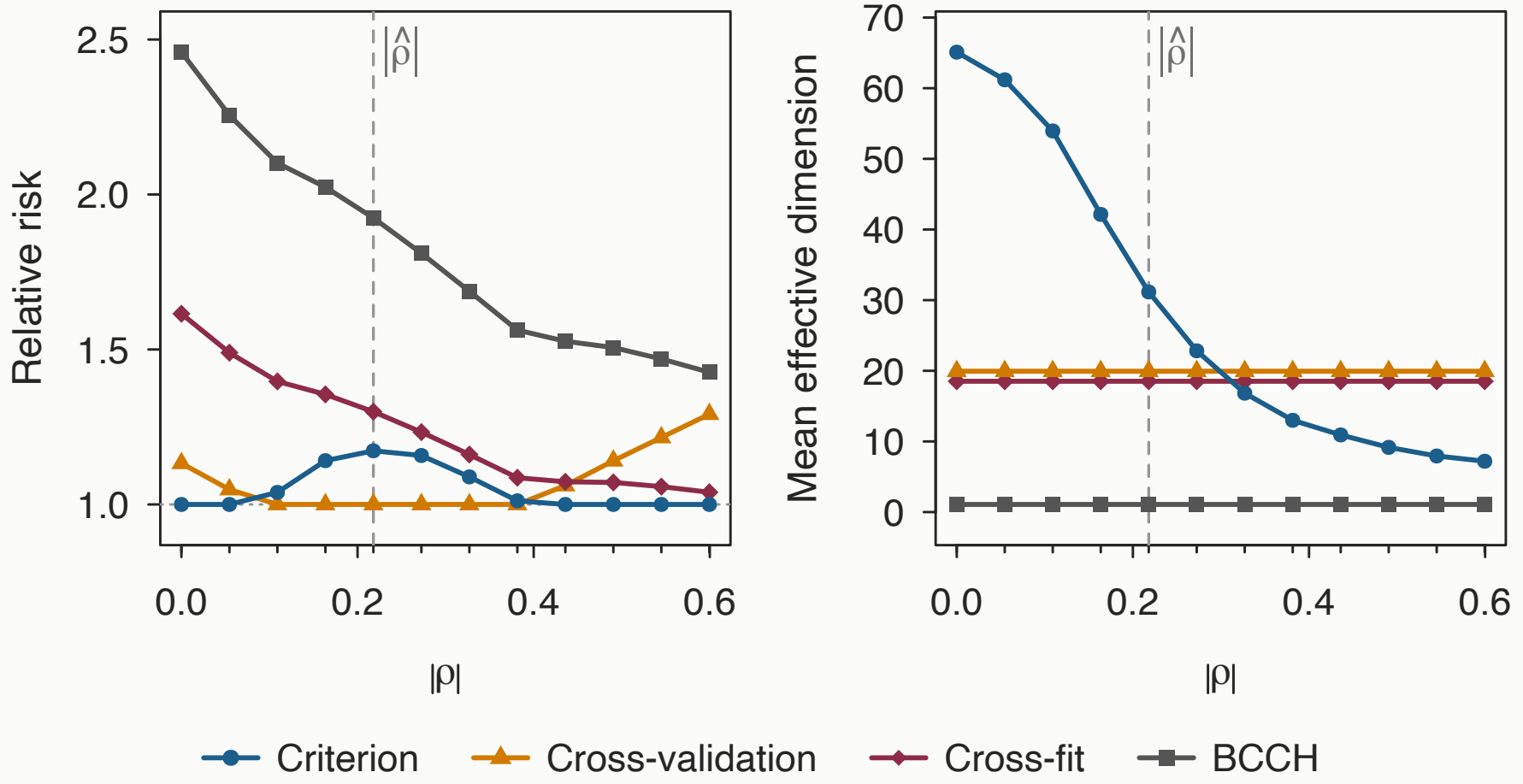


Off-scale BCCH risks include rare pilot-fallback replications.

Back

$F = 4.4$ $F = 28.0$

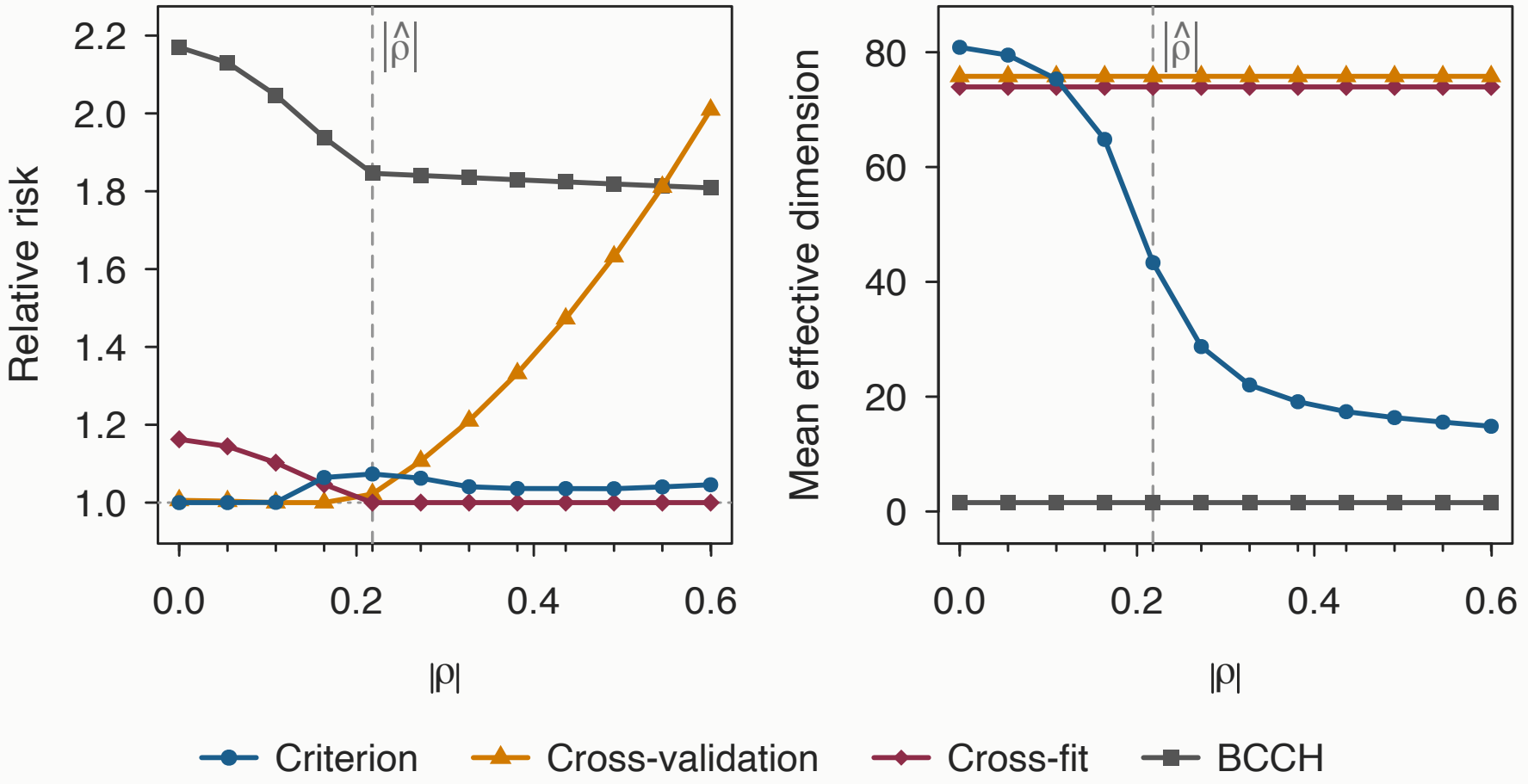
Gaussian Errors: $F = 4.4$



Back

$F = 2.7$ $F = 28.0$

Gaussian Errors: $F = 28.0$



Back

$F = 2.7$ $F = 4.4$

Non-Gaussian Simulation Designs

For independent standard Gaussian G_i, B_i, the smooth-transformation design uses

$$\psi(t) = \frac{t + t^3}{\nu(1 + 2t^2)}, \qquad v_i = \psi(G_i), \qquad \varepsilon_i = \rho v_i + \sqrt{1 - \rho^2}\, \psi(B_i),$$

where ν normalizes the variance to one. It has $\mathbb{E}[\varepsilon_i \mid v_i] = \rho v_i$ and $\mathrm{Var}(\varepsilon_i \mid v_i) = 1 - \rho^2$.

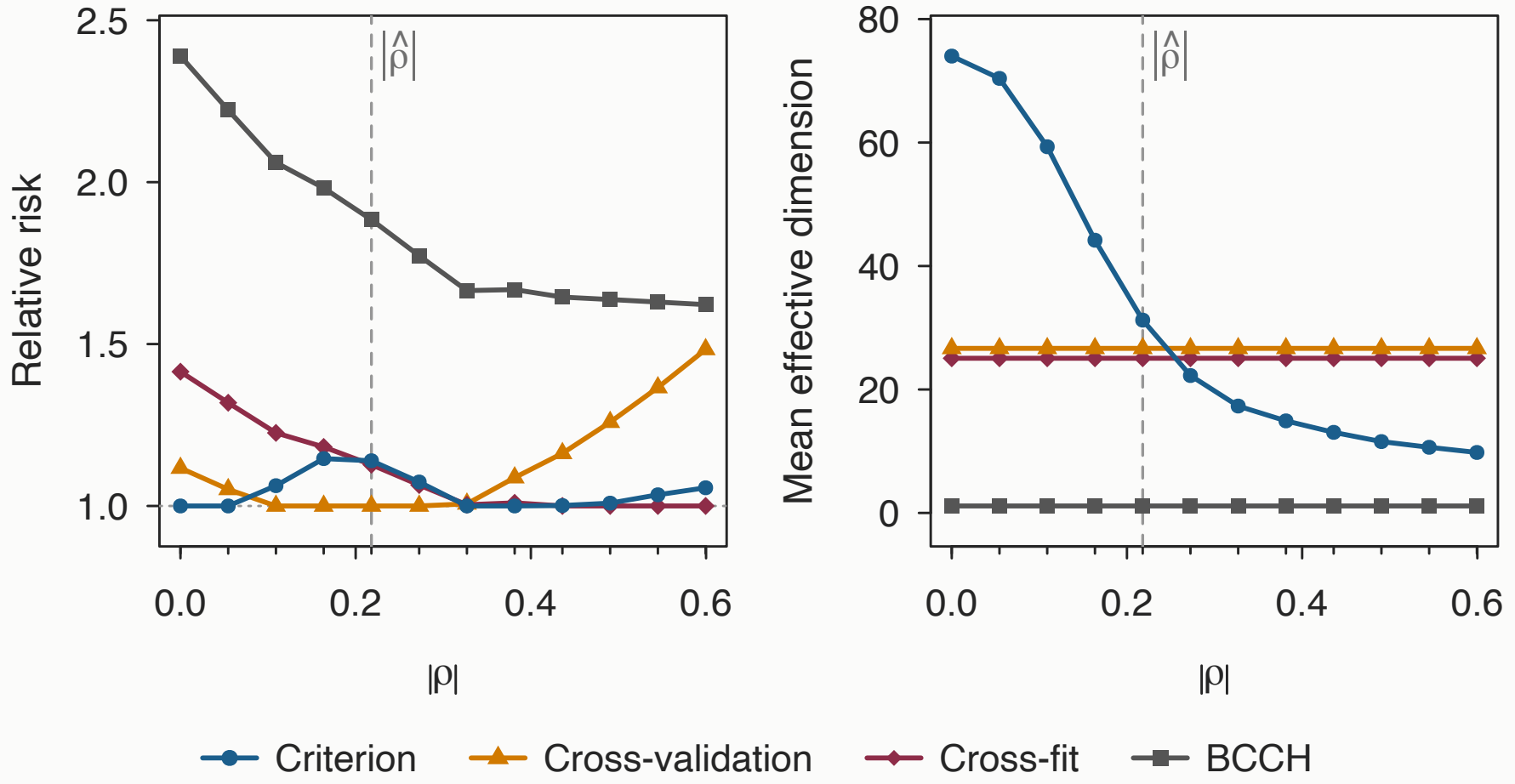
The common-scale Laplace design instead uses an independent $W_i \sim \mathrm{Exp}(1)$:

$$v_i = \sqrt{W_i}\, G_i, \qquad \varepsilon_i = \sqrt{W_i}\{\rho G_i + \sqrt{1 - \rho^2}\, B_i\}.$$

Its conditional variance is $(1 - \rho^2)\mathbb{E}[W_i \mid v_i]$, so this design goes beyond the theorem's error restrictions.

Back

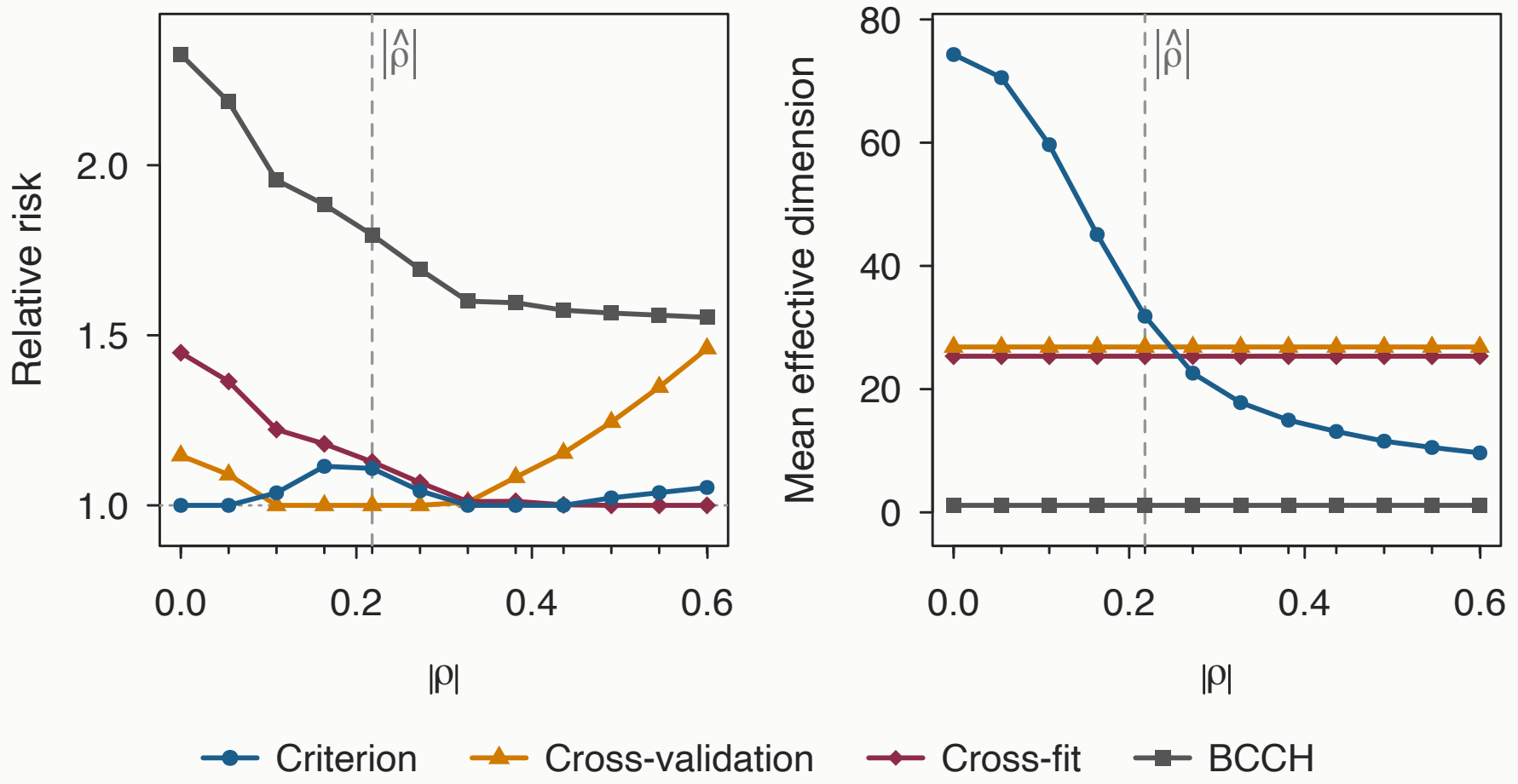
Smooth-Transformation Errors: $F = 7.7$



Back

Error Designs

Common-Scale Laplace Errors: $F = 7.7$



Back

Error Designs