The Fragility of Sparsity

Michal Kolesár, Ulrich K. Müller and Sebastian Roelsgaard
Princeton University

September 2024

Sparsity Based Estimators (SBEs)

- Belloni, Chernuzhukov, Hansen (2014) (BCH): Impose (approximate) sparsity on control coefficients γ and δ in

$$\begin{aligned} Y_i &= D_i\beta + W_i'\gamma + U_i \qquad E[U_i|D_i, W_i] = 0 \\ D_i &= W_i'\delta + \tilde{D}_i \qquad\qquad\quad E[\tilde{D}_i|W_i] = 0, \end{aligned}$$

  that is, number of non-zero coefficients is small (up to negligible approximation terms). Also see Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014).

- Whenever number of control coefficients p is larger than n, cannot run OLS

  $\Rightarrow$ Theory allows $p \gg n$, but in economic applications, p almost always smaller than n

- Efficiency gains if $p \asymp n$, potential appeal of robust "fully automatic data-driven control selection"

Fragility of Sparsity Based Estimators

1. Not invariant to linear reparameterizations of controls

   $\Rightarrow$ SBEs move up to 3 standard errors by seemingly innocuous reparameterizations in three applications

2. Sparse representations are rare

   $\Rightarrow$ Probability of a "random chosen" parameterization to be (approximately) sparse is small

   $\Rightarrow$ A thoughtless parameterization choice is unlikely to induce sparsity

3. Sparsity might not hold (potentially even for all parameterizations in large class)

   $\Rightarrow$ We develop two tests of null hypothesis of sparsity and find many rejections in the three applications

Upshot

- OLS is feasible when $p < n$
- Unless p is close to n, efficiency gains of imposing sparsity over OLS are modest
- Can avoid fragility concerns by running OLS

  $\Rightarrow$ Standard errors if $p \asymp n$: Cattaneo et al. (2018), Jochmans (2022), Kline et al. (2020), etc.
- If SBEs are employed, must think carefully about the joint issues of
  - why sparsity is defensible assumption
  - specification of controls

Literature

- Lasso variants that are invariant to choice of baseline category: Bondell and Reich (2009), Gertheiss and Tutz (2010), Stokell, Shah, Tibhsirani (2021), etc.

- Empirical evidence against sparsity: Giannone, Lenza and Primiceri (2021)

- Poor small sample performance: Wüthrich and Zhu (2023)

- Sensitivity to tuning parameter choices: Angrist and Frandsen (2022)

Outline

1. Empirical sensitivity to linear reparameterizations
2. Sparse representations are rare
3. Potential efficiency gains of imposing sparsity
4. Testing sparsity
5. Conclusion

Three Empirical Illustrations

- Three papers that leverage SBEs in their main specification

  1. BCH: Effect of abortion on crime
  2. Frerrara (2022): Effect of WW2 casualty rates of semiskilled white soldiers on post-WW2 black employment
  3. Enke (2020): Relationship between moral values and voting

- Applications make (arguably arbitrary) choices in defining control regressor matrix W. How do results change under other reasonable choices?

  1. Different ways of resolving multicollinearity
  2. When W includes powers and interactions, explore different normalizations of baseline variables (keep original, demean, subtract median, set range to $[-1, 1]$ or $[0, 1]$)

BCH

- Donohue and Levitt (2001) use 8 baseline controls and state and time fixed effects. BCH estimate first difference version on years 1986-1997 where
  - Y_{it} is change in log crime rate in state i between years t and $t-1$
  - D_{it} is change in effective abortion rate (affected by 1973 Roe decision)
- BCH estimate add
  - Interaction of baseline controls with linear and quadratic time trends
  - Lags and squared lags of baseline controls, also interacted with linear and quadratic trends
  - 49 time-invariant state-level controls (initial values, average values of various transformations of baseline controls), also interacted with trends

Multicollinearity in W

- Only 48 states considered, so 49 time-invariant controls span same column space as state fixed effects

- Time effects are included, so 2 of the time invariant variables are redundant, as are 2 interactions with linear trend and 2 interactions with quadratic trend

- A dummy baseline control is non-zero only 21 times, but is interacted with 24 variables, so 3 interactions are redundant

$\Rightarrow$ BCH drop 9 perfectly collinear columns of the original 303 columns of W to obtain $p = 294$ controls

$\Rightarrow$ There are $\binom{24}{3}\binom{49}{2}^3 \approx 3 \times 10^{12}$ equally plausible alternative ways of doing this

BCH: Effect of Resolving Multicollinearity

Outcome	OLS	Post Double Lasso		
		repl.	min	max
violent crime	0.006	-0.160	-0.216	-0.109
	(0.755)	(0.112)	(0.118)	(0.093)
property crime	-0.154	-0.110	-0.137	-0.054
	(0.223)	(0.045)	(0.045)	(0.047)
murder	2.240	-0.131	-0.225	-0.061
	(2.819)	(0.146)	(0.140)	(0.149)

$n = 576$, $p = 294$

$\Rightarrow$ Changes of 1.2 to 1.9 standard errors

Ferrara (2022)

- Decennial 1920-1960 unbalanced panel of county level observations in 16 Southern U.S. states, two-way fixed effects estimation with
  - Y_{it} : share of semi-skilled black workers
  - D_{it} : white casualty rate in WW2 interacted with post-WW2 dummy
- Additional controls:
  - Interactions between state and time dummies
  - 24 baseline controls, their squares, interactions, and interactions with state and time
  - Two baseline controls included in triple interactions with other baseline controls, state and time

$\Rightarrow$ Drop one reference state and reference decade in each interaction; Delaware only has 15 observations, but 33 state specific controls

Ferrara (2022): Effect of Resolving Multicollinearity

Outcome	OLS	Post Double-t Selection		
		repl.	min	max
% Semiskilled Black workers	0.118	0.548	0.242	0.657
	(0.126)	(0.167)	(0.126)	(0.153)

$n = 4,903$, $p = 2,252$

$\Rightarrow$ Change of over 3 standard errors

Enke (2020)

- Uses survey data to construct index of importance of universalist moral values (individual rights, justice, fairness) vs communal values (loyalty, respect)
  - Y_{it} : voting behavior
  - D_{it} : value index
  - W_{it} : 10 continuous or binary controls, plus 5 sets of categorical variables
- Requires choice of reference category for each of the 5 sets of categorical variables

Enke (2020): Effect of Resolving Multicollinearity

Outcome	OLS	Post Double Lasso		
		repl.	min	max
Trump—avg. GOP	-3.68	-1.92	-2.12	-1.84
	(1.42)	(0.94)	(0.95)	(0.93)
Trump 2016	-12.40	-12.34	-12.36	-11.96
	(1.36)	(1.05)	(1.05)	(1.03)
Trump primary	-5.32	-7.78	-8.62	-7.72
	(2.67)	(1.54)	(1.53)	(1.54)

$n = 4,903$, $p = 2,252$

$\Rightarrow$ Changes of 0.3 to 0.6 standard errors

Variable Normalization Before Interactions and Taking Powers

- BCH and Ferrara (2022) did not normalize control variables

- We consider:
  - Demeaning
  - Centering at median
  - Setting range to $[-1, 1]$
  - Setting range to $[0, 1]$

BCH: Effect of Variable Normalizations

Outcome	OLS	Post Double Lasso		
		repl.	min	max
violent crime	0.006	-0.160	-0.160	-0.122
	(0.755)	(0.112)	(0.112)	(0.097)
property crime	-0.154	-0.110	-0.127	-0.078
	(0.223)	(0.045)	(0.038)	(0.041)
murder	2.240	-0.131	-0.149	-0.066
	(2.819)	(0.146)	(0.151)	(0.167)

$n = 576$, $p = 294$

$\Rightarrow$ Changes of 0.3 to 1.3 standard errors

Ferrara (2022): Effect of Variable Normalizations

Outcome	OLS	Post Double-t Selection		
		repl.	min	max
% Semiskilled Black workers	0.118	0.548	0.482	0.548
	(0.126)	(0.167)	(0.137)	(0.167)

$n = 4,903$, $p = 2,252$

$\Rightarrow$ Change of 0.5 standard errors

Sparse Representations Are Rare

- Many ways of expressing same column space. If we pick one plausible representation are random, how likely do we get an approximately sparse one?

- Three idealized settings:

  1. All rotations of W plausible (extreme case)

  2. W consists of FE, and any representation involving sums of FEs is plausible

  3. W obtained by taking Hermite polynomials of scalar base variable after offset λ, but not sure what λ is appropriate

Approximate Sparsity

- Consider outcome regression $Y_i = D_i\beta + W_i'\gamma + U_i$

- Assume $p \asymp n$ throughout. Then representation $\tilde{W}_i = AW_i$ is *approximately sparse* if for sparsity index

$$s = o(\sqrt{p}/\log p)$$

the mean squared error approximation of $W'\gamma$ satisfies

$$\min_{||v||_0 \leq s} E[(W_i'\gamma - \tilde{W}_i'v)^2] = O(s/p).$$

Full Rotation

- Obviously extreme: For any $W_i'\gamma$, there exists rotation R so that with $\tilde{W}_i = RW_i$, $W_i'\gamma = \tilde{W}_i'\tilde{\gamma}$ with $||\tilde{\gamma}||_0 = 1$, and there exists another rotation such that $||\tilde{\gamma}||_0 = p$ and $\tilde{\gamma}$ has identical entries

- **Theorem**: Let $\tilde{W}_i = \mathcal{R}W_i$, where $\mathcal{R}$ is random with Haar measure on rotation matrices. Assume eigenvalues of $E[W_iW_i']$ are bounded away from zero and infinity. Then log of probability of obtaining approximately sparse representation is $O(-\frac{p}{4}\log p)$.

  $\Rightarrow$ For $p \geq 50$, $p^{-p/4} < 10^{-21}$

- Proof leverages that

$$\mathcal{R}\gamma \sim \frac{||\gamma||_2}{||\mathcal{Z}||_2}\mathcal{Z} \qquad \mathcal{Z} \sim \mathcal{N}(0, I_p).$$

  Tails of normal are thin, so very rare to obtain vector that is dominated by few elements.

Fixed Effects

- Consider turning age into categories. Then maybe step function could yield sparse representation of fixed effects (young vs old), or maybe three distinct coefficients (young, middle aged, old), or...

- General specification: Starting from p fixed effects Z_i, let $\tilde{W}_i = AZ_i$, where $A_{ij} \in \{0, 1\}$ and A is full rank.

  $\Rightarrow$ Generate $\mathcal{A}$ by drawing elements i.i.d. Bernoulli(q), $0 < q \leq 1/2$ until we obtain full rank matrix

- **Theorem**: Suppose for some A_0, a single coefficient on $W_i = A_0 Z_i$ is non-zero, and that the number of zeros K in the corresponding row of A_0 satisfies $0 < \lim_{n\to\infty} K/p < 1$. Further assume all baseline categories have population fractions of the same order. Then the probability of $\tilde{W}_i = \mathcal{A} Z_i$ to be approximately sparse is no larger than $(1 - q - \varepsilon)^K$ for all $\varepsilon > 0$ and large enough n.

  $\Rightarrow$ Proof leverages results in Tikhomirov (2020)

Offset in Hermite Polynomial Expansion

- Construct p scaled Hermite polynomials from scalar baseline variable $z_i \sim iid\mathcal{N}(0,1)$,
$$\tilde{W}_{ij} = H_j(z_i), \qquad j = 1, \ldots p$$
where scaling is such that $E[\tilde{W}_{ij}^2] = 1$.

- Suppose $Y_i = H_p(Z_i + \lambda) + U_i$, so for regression to be sparse, would need to use offset λ, but researcher uses zero offset.

- **Theorem:** (a) Suppose $\lambda = L/\log p$, $L > 0$. If L is fixed, then for $1 \leq j \leq L\sqrt{p}/\log p$ and all large enough p, $\tilde{\gamma}_{p-j}^2 \geq Ce^{j/2}$, where C is an absolute constant, and approximate sparsity does not hold.
(b) If $L \rightarrow 0$, then approximate sparsity holds.

  $\Rightarrow$ If λ is drawn at random from $[0,1]$, probability of approximate sparsity is of order $O(1/\log p)$

Potential Efficiency Gains of SBEs

- Recall model

$$\begin{aligned} Y_i &= D_i\beta + W_i'\gamma + U_i \quad & E[U_i|D_i, W_i] = 0 \\ D_i &= W_i'\delta + \tilde{D}_i \quad & E[\tilde{D}_i|W_i] = 0 \end{aligned}$$

- If $p < n$ can run OLS without any assumptions on γ or δ

- If $p \asymp n$ OLS is not semiparametrically efficient, but SBEs are

$\Rightarrow$ How large are the potential gains?

Potential Efficiency Gains of SBEs

- **Assumption OLS**: (i) $\{U_i, \tilde{D}_i\}$ are i.i.d. conditional on W

  (ii) $\lim_{n\to\infty} p/n = c < 1$

  (iii) For $\eta, K > 0$: $E[|U_i|^{2+\eta}|D,W] + E[|\tilde{D}_i]^4|W] \leq K$, $1/E[\tilde{D}_i^2|W] + 1/E[U_i^2|D,W] \leq K$

- **Lemma**: Let $\hat{D}_i$ be the OLS residuals from regressing D_i on W_i. Under Assumption OLS

$$\frac{\hat{\beta}_{OLS} - \beta}{s_{OLS}} \sim \mathcal{N}(0,1) \qquad s_{OLS}^2 = \left(\sum_{i=1}^{n} \hat{D}_i^2\right)^{-2} \left(\sum_{i=1}^{n} \hat{D}_i^2 U_i^2\right)$$

- Semiparametric Efficiency Bound under homoskedasticity is limit of

$$s_*^2 = \left(\sum_{i=1}^{n} \tilde{D}_i^2\right)^{-2} \left(\sum_{i=1}^{n} \tilde{D}_i^2 U_i^2\right)$$

  $\Rightarrow$ SBEs achieve bound, so potential gain of s_*^2/s_{OLS}^2 (but OLS inference small sample optimal under Gaussianity)

Potential Efficiency Gains of SBEs

- If U_i is homoskedastic and Assumption OLS holds

$$\frac{s_*^2}{s_{OLS}^2} = (1 - p/n)\kappa(1 + o_p(1)) \qquad \kappa = \frac{E[(n-p)^{-1}\sum_{i=1}^n \hat{D}_i^2]}{E[\tilde{D}_i^2]}$$

$\Rightarrow$ When $\tilde{D}_i$ is homoskedastic, $\kappa = 1$

- When $p/n = 0.2$ and $\kappa = 1$, $s_*/s_{OLS} \approx 0.9$ (and 0.7 for $p/n = 0.5$)

$\Rightarrow$ $\kappa \ll 1$ only under large positive correlation of leverages P_{ii} and $E[\tilde{D}_i^2|W]$

- Under heteroskedasticity

$$\frac{s_*^2}{s_{OLS}^2} = \frac{s_*^2/s_{*,\text{hom}}^2}{s_{OLS}^2/s_{OLS,\text{hom}}^2}(1 - p/n)\kappa(1 + o_p(1))$$

$\Rightarrow$ Large *differential* impact of heteroskedasticity corrections $s_*^2/s_{*,\text{hom}}^2$ and $s_{OLS}^2/s_{OLS,\text{hom}}^2$ needed to obtain very different conclusions

Testing Sparsity

1. Hausman test: Compare $\hat{\beta}_{OLS}$ with $\hat{\beta}_{SBE}$:

   - If sparsity holds, $\hat{\beta}_{SBE}$ is efficient and asymptotically normal

   - $\hat{\beta}_{OLS}$ is asymptotically normal regardless

   $\Rightarrow$ Large differences between $\hat{\beta}_{OLS}$ and $\hat{\beta}_{SBE}$ indicate sparsity does not hold

2. F-test: Check whether non-selected regressors explain too much of residual variation

   - Under approximate sparsity, treating Lasso selection as truth is good enough approximation

   - In high dimensions, asymptotic variance often hard to estimate; avoid by estimating variances under the null of sparsity

Hausman Test

- **Lemma**: If Assumption OLS holds, then for any asymptotically linear estimator $\hat{\beta}_*$ that achieves the semiparametric efficiency bound,

$$\frac{\hat{\beta}_{OLS} - \hat{\beta}_*}{s_H} \sim \mathcal{N}(0,1) \qquad s_H^2 = \sum_{i=1}^{n} \omega_i^2 U_i^2 \qquad \omega_i = \frac{\hat{D}_i}{\hat{D}'\hat{D}} - \frac{\tilde{D}_i}{\tilde{D}'\tilde{D}}$$

- **Theorem**: Under additional regularity conditions, same conclusion holds when U_i and $\tilde{D}_i$ are replaced by Lasso residuals

- When U is homoskedastic, $s_H^2 \approx s_{OLS}^2 - s_*^2$, so when efficiency gain is small, $\hat{\beta}_{OLS}$ and $\hat{\beta}_*$ need to be close

  $\Rightarrow$ Some authors compute SBEs as a "robustness check" of OLS. In fact, it's the opposite!

F-test

- Consider regression $\mathsf{Y}_i = X_i'\alpha + \varepsilon_i$, $E[\varepsilon_i|X_i] = 0$ (can be outcome or propensity score regression)

- Suppose we knew set $\mathcal{S}^*$ of non-zero regressors under null hypothesis of sparsity. Natural test compares restricted and unrestricted sum of squared residuals

$$\mathcal{F} = \mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y} - \mathsf{Y}'(I - P)\mathsf{Y} = \sum_{i=1}^{n} (\hat{\varepsilon}_i^*)^2 - \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

  where $P_{\mathcal{S}} = X_{\mathcal{S}}(X_{\mathcal{S}}'X_{\mathcal{S}})^{-1}X_{\mathcal{S}}'$ and $P = X(X'X)^{-1}X'$

- Don't know $\mathcal{S}^*$, but under standard lasso assumptions can construct SBE $\tilde{\alpha}$ such that $\tilde{\varepsilon}_i = \mathsf{Y}_i - X\tilde{\alpha}$ is good enough approximation to $\hat{\varepsilon}_i^* = \mathsf{Y}_i - X\hat{\alpha}_{\mathcal{S}^*}$

  $\Rightarrow$ Holds even under approximate sparsity, suggests impossibility of testing "approximate sparsity" vs "exact sparsity"

Limiting Distribution of F-test

- **Theorem**: Under suitable assumptions, and under null hypothesis of approximate sparsity

$$\frac{\mathcal{F} - \sum_{i=1}^{n} \varepsilon_i^2 P_{ii}}{\sqrt{2 \sum_{i \neq j} \varepsilon_i^2 \varepsilon_j^2 P_{ij}^2}} \Rightarrow \mathcal{N}(0, 1)$$

  and with $\tilde{\varepsilon}_i = \mathsf{Y}_i - X\tilde{\alpha}$ and $\hat{\varepsilon}_i$ the OLS residuals

$$\frac{\sum_{i=1}^{n} \tilde{\varepsilon}_i^2 - \sum_{i=1}^{n} \hat{\varepsilon}_i^2 - \sum_{i=1}^{n} \tilde{\varepsilon}_i P_{ii}}{\sqrt{2 \sum_{i \neq j} \tilde{\varepsilon}_i^2 \tilde{\varepsilon}_j^2 P_{ij}^2}} \Rightarrow \mathcal{N}(0, 1).$$

- Amounts to checking whether lasso or post-lasso residuals are too large compared to OLS residuals, allowing for heteroskedasticity

Testing Sparsity in BCH

Outcome			Collinearity		Normalization	
	Test	repl.	min	max	min	max
violent crime	H	81.7	76.0	87.4	81.7	85.8
	FO	9.5	9.5	9.5	9.5	9.5
	FP	0.3	0.0	0.9	0.0	0.6
property crime	H	82.6	61.8	93.4	70.9	89.7
	FO	12.0	12.0	12.0	12.0	12.0
	FP	28.8	12.6	35.0	0.0	29.9
murder	H	21.0	19.7	22.5	20.4	22.6
	FO	43.3	43.3	43.3	43.3	43.3
	FP	0.4	0.2	1.1	0.4	1.2

p-values in percent of Hausman test (H), F-test for outcome (FO) and F-test for propensity score (FP)

Testing Sparsity in Ferrara (2022)

Outcome	Test	repl.	Collinearity min	Collinearity max	Normalization min	Normalization max
% Semiskilled Black workers	H	0.0	0.0	5.5	0.0	0.0
	FO	0.0	0.0	0.0	0.0	0.0
	FP	34.5	19.6	49.7	34.5	53.3

p-values in percent of Hausman test (H), F-test for outcome (FO) and F-test for propensity score (FP)

Testing Sparsity in Enke (2020)

Outcome	Test	repl.	Collinearity min	Collinearity max
Trump−avg. GOP	H	6.1	4.9	9.5
	FO	13.0	5.1	13.0
	FP	94.6	63.1	96.4
Trump 2016	H	94.6	63.1	96.4
	FO	0.0	0.0	0.5
	FP	0.1	0.1	0.1
Trump primary	H	15.5	5.9	16.6
	FO	9.6	2.8	9.6
	FP	0.0	0.0	0.0

p-values in percent of Hausman test (H), F-test for outcome (FO) and F-test for propensity score (FP)

Conclusion

- If SBEs are used, then one needs to provide substantive arguments why
  - sparsity holds
  - in a particular representation of column space
- Issues not specific to SBEs: Most machine learning methods lack invariance to linear reparameterizations
  - Less of a concern when used repeatedly to produce many forecasts
  - But in economics, typically care about one particular estimate

Thank you!