Estimation and inference for high-dimensional nonparametric additive instrumental-variables regression

Ziang Niu

University of Pennsylvania

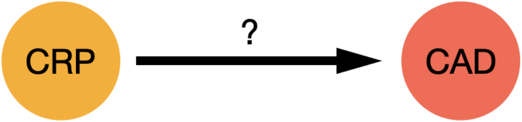
July 10th, 2022

Joint work with Wei Li and Yuwen Gu

Outline

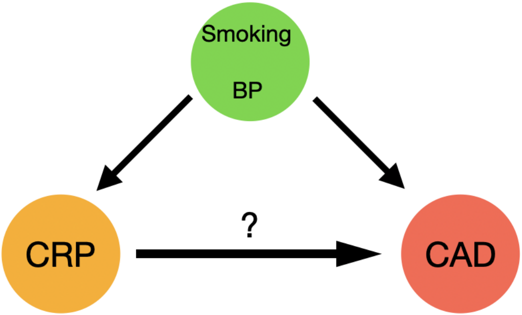
1. Introduction
2. The sparse additive instrumental-variables model
3. Non-asymptotic analysis
4. Inference
5. Simulation
6. Application
7. Conclusion

An illustrative example



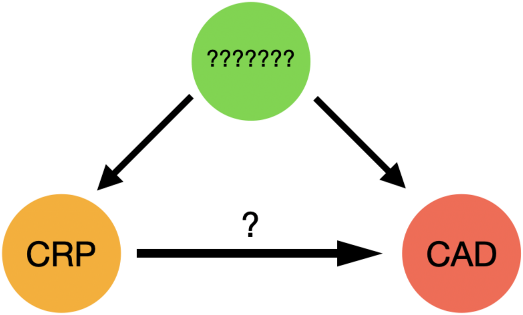
- CRP: C-reactive protein, a kind of blood proteins.
- CAD: coronary artery disease, a kind of heart disease.

An illustrative example



- Confounders: smoking, blood pressure, etc.

An illustrative example



- Confounders are not always observed!
- Solution: use instrumental variable.

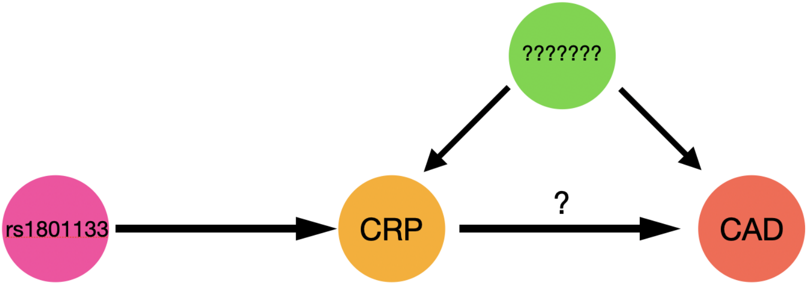
What is IV?

- Three requirements:

  (i) independence with unobserved confounders

  (ii) independence with outcome given treatments and confounders

  (iii) correlation with treatments



- CRP: gene expression
- Single nucleotide polymorphisms (SNP) as IV

Two-stage least squares (2sls)

- Consider the following linear model with an endogenous treatment X:

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

where $E(\epsilon \mid X) \neq 0$.

- Directly applying OLS will yield inconsistent estimators; suppose we have an IV Z satisfying $E(\epsilon \mid Z) = 0$.
- 2sls: first regress X on Z and obtain predictions $\widehat{X}$; second regress Y on $\widehat{X}$ to obtain an estimator $\widehat{\beta}$.
- $\widehat{\beta}$ is consistent no matter which model is used in the first stage, because $E\{f(Z)(Y - \beta X)\} = 0$. We often use linear model as a working model.

Two-stage least squares (2sls)

- See the following toy example:

$$Y = 1 + X + U + \epsilon, \quad X = 1 + Z + Z^2 + \sin(Z) + Z^3 + U + \eta.$$

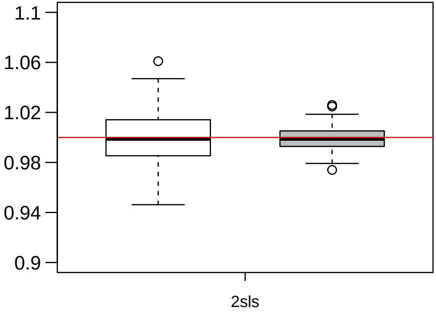
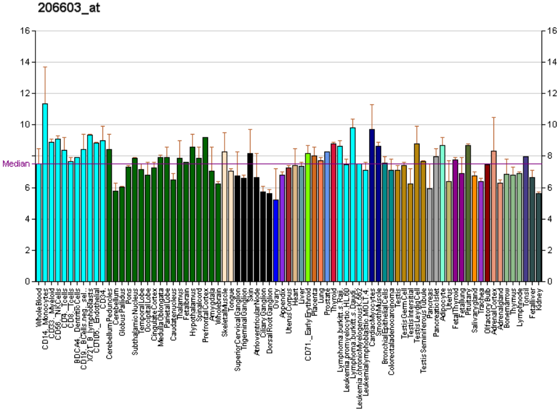


Figure: 2sls estimators under sample sizes $n = 200$ and $n = 1000$.

Challenges for 2sls

- Both SNPs and gene expression are potentially high dimensional variables.



2sls under high-dimensionality

- Consider the following example:

$$Y = 1 + X + U + \epsilon, \quad X = 1 + Z_1^2 + Z_2 + \sin(\pi Z_3) + Z_4^2 + Z_5^2 + U + \eta,$$

where $Z \in \mathbb{R}^q$ and $q = 1000$.

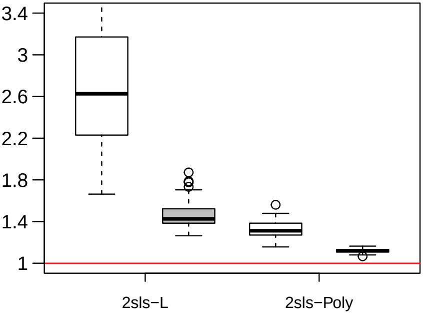


Figure: 2sls estimators under sample sizes $n = 200$ and $n = 1000$.

Most relevant work

- When instruments and treatments are both high-dimensional, linear models have been proposed (Lin et al., 2015; Zhu, 2018; Gold et al., 2020).
- Nonlinear effects of the SNPs on gene expressions are likely to exist (Wang et al., 2015; Zhang & Ghosh, 2017; Zhao et al., 2019). These methods employ kernel-based procedures to capture nonlinear relationships, and hence are not very effective when applied to the high-dimensional regime.
- Zhu (2018) considers the high-dimensional linear instrumental-variables regression for peer effect estimation in econometrics, e.g., analyzing the effects of peers' output on a firm's production output using panel data. The Research and Development expenditures of peer firms from a previous period are treated as potential instrumental variables.

The sparse additive instrumental-variables model

- $X \in \mathbb{R}^{n \times p} \Rightarrow$ Treatments $(p > n)$
- $Y \in \mathbb{R}^{n} \Rightarrow$ Outcome
- $Z \in \mathbb{R}^{n \times q} \Rightarrow$ Instrumental Variables $(q > n)$
- Model setup

$$X_\ell = \sum_{j=1}^{q} f_{j\ell}(Z_j) + \eta_\ell, \ \ell = 1, \ldots, p$$

$$Y = X\beta + \epsilon, \ E(\epsilon \mid X) \neq 0$$

Two-stage estimation procedure

- $f_{j\ell}(z_j) \approx \sum_{k=1}^{m} \overline{\gamma}_{kj\ell}\phi_k(z_j)$, where $\{\phi_k(\cdot),\ k = 1, \ldots, m\}$ are some approximation basis functions. Solve the group lasso problem:

$$\widehat{\gamma}_\ell := \arg\min_{\boldsymbol{\gamma}_\ell} \frac{1}{2n}\|X_\ell - U\boldsymbol{\gamma}_\ell\|_2^2 + \lambda_\ell \sum_{j=1}^{q} \|\gamma_{j\ell}\|_2,$$

  where $U = \{\phi_k(Z_{ij})\}$ is the spline matrix.

- Solve the following lasso problem

$$\widehat{\beta} := \arg\min_{\beta} \frac{1}{2n}\left\|Y - \widehat{X}\beta\right\|_2^2 + \mu\|\beta\|_1,$$

  where $\widehat{X} = (\widehat{X}_1, \ldots, \widehat{X}_p)$ with $\widehat{X}_\ell = U\widehat{\gamma}_\ell$.

- Sparsity: r for the additive model in the first stage; s for the second stage.
- IV observations in an additive function: $F_{j\ell} = \{f_{j\ell}(z_{1j}), \ldots, f_{j\ell}(z_{nj})\}^{\mathrm{T}} \in \mathbb{R}^n$.

First stage non-asymptotic analysis

Theorem 1

There exist positive constants c_1, c_2, and c_3 such that if

$$\lambda_{\max} = \max_{\ell} \lambda_\ell = \max\left[c_1 \sigma_{\max}\left\{\frac{\log(pqm)}{n}\right\}^{1/2}, c_2 r m^{-(2d+1)/2} + c_3 r \left\{\frac{\log(pqm)}{mn}\right\}^{1/2}\right],$$

then for sufficiently large n, with probability at least $1 - 20(pqm)^{-1}$, the regularized estimator $\widehat{\gamma}_\ell$ satisfies

$$\max_{\ell} \left\| \sum_{j=1}^{q} F_{j\ell} - U\widehat{\gamma}_\ell \right\|_2^2 \leq \frac{50 rmn\lambda_{\max}^2}{\rho}, \quad \max_{\ell} \sum_{j=1}^{q} \|\widehat{\gamma}_{j\ell} - \bar{\gamma}_{j\ell}\|_2 \leq \frac{32 rm\lambda_{\max}}{\rho},$$

where $\sigma_{\max} = \max_\ell \sigma_\ell$, $m = \Theta\{n^{1/(2d+1)}\}$, and $r^2 = o[n/\{m^4 \log(pqm)\}]$.

- Average in-sample prediction consistency: $r^3 = o\{n^{2d/(2d+1)}/\log(pqm)\}$
- Coefficients estimation consistency: $r^4 = o\{n^{(2d-1)/(2d+1)}/\log(pqm)\}$

Second-stage nonasymptotic analysis

Theorem 2

Let the regularization parameter $\lambda_{\max}$ be chosen as in Theorem 1. Further assume $\lambda_{\max}$ satisfies $560 C_0 \lambda_{\max}(2rm/\rho)^{1/2} \leq \kappa^2/(4rs)$. If we choose the second-stage regularization parameter as

$$\mu = 2r\lambda_{\max}(7\sigma_0 + 8\sqrt{5}B\sigma_{\max} + 30B)(2m/\rho)^{1/2},$$

then with probability at least $1 - 234(pqm)^{-1}$, the estimator $\widehat{\beta}$ satisfies

$$\|\widehat{\beta} - \beta\|_1 \leq \frac{64}{\kappa^2} s\mu, \quad \|\widehat{X}(\widehat{\beta} - \beta)\|_2^2 \leq \frac{64}{\kappa^2} ns\mu^2.$$

- Consistency is guaranteed if we take $\mu^2 = O\{r^4 \log(pqm)/n^{2d/(2d+1)}\}$ and $s^2 r^5 = o\{n^{2d/(2d+1)}/\log(pqm)\}$.
- When r is fixed, we have $s^2 = o[n/\{m \log(pqm)\}]$. This almost recovers the sparsity in the classical lasso setting when d is large enough.

Inference: one-step Newton-Raphson iteration

- Moment condition:

$$Z \perp\!\!\!\perp \epsilon \quad \Rightarrow \quad \mathbb{E}\left\{F(Z)^\top \frac{(Y - X\beta)}{n}\right\} = 0,$$

where $F(Z) = \sum_{j=1}^{q} F_j$ with $F_j = (F_{j1}, \ldots, F_{jp}) \in \mathbb{R}^{n\times p}$.

- Take derivative and obtain the following matrix:

$$\mathbb{E}\{-F(Z)^\top X\}/n = \mathbb{E}\{-F(Z)^\top F(Z)/n\}$$

- One step update: given the lasso estimate $\widehat{\beta}$,

$$\widetilde{\beta} = \widehat{\beta} + \left\{\mathbb{E}\left(\frac{F(Z)^\top F(Z)}{n}\right)\right\}^{-1} \frac{F(Z)^\top (Y - X\beta)}{n}$$

Inference: one-step Newton-Raphson iteration

- $F(Z)$ is unknown: estimate $F(Z)$ with $U\widehat{\Gamma}$.

$$\frac{F(Z)^\top(Y - X\beta)}{n} \approx \frac{(U\widehat{\Gamma})^\top(Y - X\beta)}{n} \approx \frac{(U\widehat{\Gamma})^\top(Y - X\widehat{\beta})}{n}$$

- Estimate precision matrix $\Omega := [\mathbb{E}\{F(Z)^\top F(Z)/n\}]^{-1}$.
- The rows $\widehat{\theta}_\ell$ of the estimator $\widehat{\Omega}$ are obtained by solving the following constrained L_1-minimization program:

$$\min_{\theta_\ell \in \mathbb{R}^p} \|\theta_\ell\|_1, \text{ subject to } \|\widehat{\Sigma}_F \theta_\ell - e_\ell\|_\infty \leq \upsilon \quad (\ell = 1, \ldots, p),$$

where $\widehat{\Sigma}_F = \frac{F(Z)^\top F(Z)}{n} \approx \frac{\widehat{\Gamma}^\top U^\top U\widehat{\Gamma}}{n}$. Then we can obtain the estimate

$$\widetilde{\beta} = \widehat{\beta} + \widehat{\Omega}\frac{(U\widehat{\Gamma})^\top(Y - X\widehat{\beta})}{n}$$

where $\widehat{\Omega} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p) \in \mathbb{R}^{p \times p}$.

Asymptotically normal

Theorem 3

With technical assumptions, we have

$$\sqrt{n}(\widetilde{\beta}_\ell - \beta_\ell) \rightsquigarrow \mathcal{N}(0, \omega_\ell^2),\ \ell = 1, \ldots, p$$

where $\omega_\ell = \mathrm{Var}(\epsilon)\theta_{\ell\ell}$. *Define*

$$\widehat{\omega}_\ell = \widehat{\sigma}_0 \big(\widehat{\theta}_\ell^{\mathrm{T}} \widehat{\Gamma}^{\mathrm{T}} U^{\mathrm{T}} U \widehat{\Gamma} \widehat{\theta}_\ell / n\big)^{1/2},\ \ \widehat{\sigma}_0 = n^{-1/2} \| Y - X\widehat{\beta} \|_2.$$

Then $\widehat{\omega}_\ell$ *is a consistent estimator of* ω_ℓ *for each* $\ell \in \{1, \ldots, p\}$.

Simulation setup

(1) Estimation:

- $p = q = 600$, and vary n from 100 to 2100.
- both linear and nonlinear treatment models are considered.

(2) Inference:

- varying $p = q$ and n.
- consider a more challenging nonlinear treatment model.

Estimation results

Table: L_1 errors of our method, the two-stage regularized least squares (2SR), and the one-stage lasso penalized least squares (PLS), averaged over one hundred replications when $p = 600$. Standard deviations are shown in the parentheses.

Sample	Linear			Nonlinear		
size	Our method	2SR	PLS	Our method	2SR	PLS
100	1.26 (0.53)	2.52 (1.19)	0.86 (0.22)	2.96 (1.41)	-	1.25 (0.38)
300	0.50 (0.23)	0.59 (0.30)	0.51 (0.16)	0.74 (0.28)	-	0.79 (0.27)
600	0.34 (0.15)	0.27 (0.11)	0.46 (0.17)	0.43 (0.17)	-	0.89 (0.27)
900	0.25 (0.10)	0.23 (0.08)	0.52 (0.14)	0.32 (0.13)	-	1.10 (0.23)
1200	0.19 (0.09)	0.19 (0.08)	0.61 (0.18)	0.27 (0.13)	-	1.18 (0.18)
1500	0.17 (0.08)	0.18 (0.08)	0.70 (0.25)	0.24 (0.10)	-	1.27 (0.16)
1800	0.16 (0.07)	0.17 (0.08)	0.81 (0.29)	0.21 (0.09)	-	1.34 (0.17)
2100	0.14 (0.05)	0.16 (0.07)	1.09 (0.42)	0.21 (0.11)	-	1.43 (0.16)

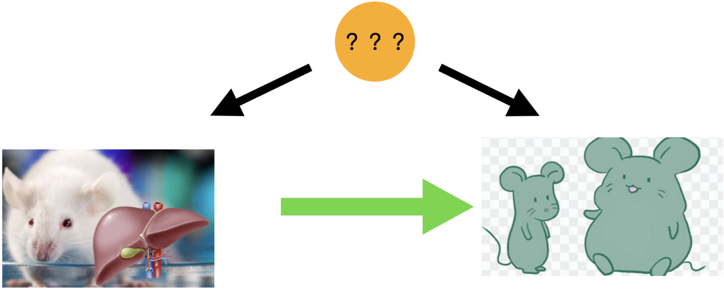
Inference results

Table: Coverage probabilities and lengths of the 95% confidence intervals by our method and the method by Gold et al. (2020). Numbers shown are multiplied by one hundred.

Dimension	Sample	Our method		Gold et al. (2020)	
	size	Coverage	Length	Coverage	Length
250	200	92.0	0.396	87.3	2.663
400	300	93.5	0.264	89.0	1.585
500	400	94.0	0.245	87.1	2.102
600	500	93.7	0.140	87.8	2.614

Application: mouse obesity data

- Gene expression: liver tissue in obese mice data.



Mouse obesity data

- $n = 287$ mice fed with high-fat western diet (144 female and 143 male).

Mouse obesity data

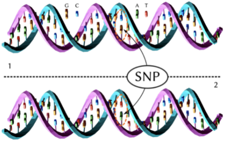
- $n = 287$ mice fed with high-fat western diet (144 female and 143 male).
- $p = 2816$ proteins in liver tissues are measured.

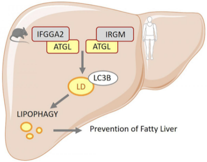
Mouse obesity data

- $n = 287$ mice fed with high-fat western diet (144 female and 143 male).
- $p = 2816$ proteins in liver tissues are measured.
- The mice were genotyped with $q = 1250$ SNPs.

Mouse obesity data

- $n = 287$ mice fed with high-fat western diet (144 female and 143 male).
- $p = 2816$ proteins in liver tissues are measured.
- The mice were genotyped with $q = 1250$ SNPs.





Application: analysis

- Adjust the body weight (Y) with sex of mice and subtract the effect of sex on weight
- Apply the proposed estimation method to obtain 28 genes.
- Compute the selection probability of each gene over 100 subsamples with size $\lfloor n/2 \rfloor$ for a sequence of tuning parameter.
- Set the threshold probability to 0.5

Stability selection results

Table: Stability selection: the selected genes with superscript “*” denote the ones overlapping with 2SR (Table 3, Lin et al., 2015).

Gene Name	Selection Probability	Gene Name	Selection Probability
Vwf*	0.77	Krtap19-2	0.59
Akap12	0.63	Tmem184c	0.74
2010002N04Rik*	0.84	Igfbp2*	0.51
Slc43a1	0.76	Gstm2*	0.91
Ccnl2*	0.54	D14Abb1e	0.52
B4galnt4	0.71		

- **Igfbp2**, **Ccnl2**, **Vwf**, **Gstm2**, and **2010002N04Rik** are also selected in Lin et al. (2015).
- Insulin-like growth factor binding protein 2 (**Igfbp2**) has been shown to protect against the development of obesity (Wheatcroft et al. 2007).

Stability selection results

Table: Stability selection: the selected genes with superscript "*" denote the ones overlapping with 2SR (Table 3, Lin et al., 2015).

Gene Name	Selection Probability	Gene Name	Selection Probability
Vwf*	0.77	Krtap19-2	0.59
Akap12	0.63	Tmem184c	0.74
2010002N04Rik*	0.84	Igfbp2*	0.51
Slc43a1	0.76	Gstm2*	0.91
Ccnl2*	0.54	D14Abb1e	0.52
B4galnt4	0.71		

- **Slc43a1**, **B4galnt4**, **Tmem184c**: potential factors leading to obesity.
- Solute Carrier Family 43 Member 1 (**Slc43a1**) is a protein coding gene; Gill et al. (2010) found that the expression of Slc43a1 in the fat mice group is quite different from that in the lean mice group.

Inference results

Table: 95% confidence intervals for the causal effects of the genes on the body weights of the mice. Shown are only the genes whose corresponding intervals do not contain zero.

Gene Name	Confidence Interval	Gene Name	Confidence Interval
Anxa5	(0.010, 7.269)	Kif22	(0.615, 7.930)
Vwf	(0.500, 7.841)	Gstm2	(0.537, 8.231)
Aqp8	(0.066, 6.855)	Gpld1	(−7.448, −0.447)
Lamc1	(0.094, 5.877)	Slc43a1	(−6.641, −1.412)
Acot9	(0.056, 8.298)	Abca8a	(−7.152, −0.072)
Anxa2	(1.086, 9.331)	Cyp4f15	(−7.468, −0.250)
2010002N04Rik	(1.343, 8.240)	Igfbp2	(−6.451, −0.666)
Msr1	(0.004, 6.783)		

- **Igfbp2**, **Ccnl2**, **Vwf**, **Gstm2**, and **2010002N04Rik** are also shown to have high selection probability.
- Annexin A2 **Anxa2** plays a role in the regulation of cellular growth and in signal transduction pathways (Wang et al., 2019).

Inference results

Table: 95% confidence intervals for the causal effects of the genes on the body weights of the mice. Shown are only the genes whose corresponding intervals do not contain zero.

Gene Name	Confidence Interval	Gene Name	Confidence Interval
Anxa5	(0.010, 7.269)	Kif22	(0.615, 7.930)
Vwf	(0.500, 7.841)	Gstm2	(0.537, 8.231)
Aqp8	(0.066, 6.855)	Gpld1	(−7.448, −0.447)
Lamc1	(0.094, 5.877)	Slc43a1	(−6.641, −1.412)
Acot9	(0.056, 8.298)	Abca8a	(−7.152, −0.072)
Anxa2	(1.086, 9.331)	Cyp4f15	(−7.468, −0.250)
2010002N04Rik	(1.343, 8.240)	Igfbp2	(−6.451, −0.666)
Msr1	(0.004, 6.783)		

- Cytochrome P450, family 4, subfamily f, polypeptide 15 (**Cyp4f15**) controls the omega-hydroxylated fatty acids in the liver tissue, which can be used for energy production (Hardwick et al., 2009).

Summary and discussion

- We propose an estimation and inference approach for high-dimensional nonparametric additive instrumental-variables model.

Summary and discussion

- We propose an estimation and inference approach for high-dimensional nonparametric additive instrumental-variables model.
- We apply our method to the mouse obesity data and find some genes that are not previously detected.

Summary and discussion

- We propose an estimation and inference approach for high-dimensional nonparametric additive instrumental-variables model.
- We apply our method to the mouse obesity data and find some genes that are not previously detected.
- Fully nonparametric method with Neural Network is possible: DeepIV (Hartford et al. ICML 2017) $\Rightarrow$ No statistical guarantee!

Summary and discussion

- We propose an estimation and inference approach for high-dimensional nonparametric additive instrumental-variables model.
- We apply our method to the mouse obesity data and find some genes that are not previously detected.
- Fully nonparametric method with Neural Network is possible: DeepIV (Hartford et al. ICML 2017) ⇒ No statistical guarantee!
- It is also interesting to consider other types of outcome in the presence of high-dimensional endogeneity issues.

Thank You!