High-Dimensional Methods: Examples for Inference on Structural Effects

V. Chernozhukov, M. Gentzkow, C. Hansen, J. Shapiro, M. Taddy

July 16, 2013

Example 1: 401(k)'s and Assets

Estimate the effect of 401(k) eligibility on measure of accumulated assets, (e.g. PVW 1994, 1995, 1996)

$$y_i = d_i\alpha_0 + x_i'\beta + \zeta_i$$

- y_i = net financial assets or total wealth,
- d_i = eligible for 401(k),
- x_i = controls for individual characteristics. PVW argue important to control for income
  - income (<10k, 10k-20k, 20k-30k, 30k-40k, 40k-50k, 50k-75k, 75k+), age, age^2, family size, education (high school, some college, college), married, two-earner, defined benefit, ira, home-owner
- $n = 9915$

Baseline estimates: Net-TFA: 9216.5 (1340.6); TW: 6612.0 (2110.1)

Did we control sufficiently for income?

Previous model provides nice baseline, but we might wonder

- Are 7 dummies for income categories sufficient to control for income?
  - More complex nonlinearity?
  - Interactions?
- Could we improve efficiency?
  - Even if eligibilty were randomly assigned, might want to introduce controls to improve efficiency
  - "Over-controlling"?
- Other important variables?

Regularization

All treatment effects estimation problems involve regularization/variable selection.

The only valid inference without dimension reduction is to conclude that one cannot learn from the data.

Regularization

Dimension reduction strategies:

- Randomized treatment assignment: Reduces dimension of control vector to 0.
  - May still want to do variable selection to improve efficiency ("single-selection" procedures)
  - Stratified sampling
- Intuition
- Formal Model Selection
  - Will fail without good intuition (Needle in a haystack: Want a big needle or a small haystack)

Intuitive Dimension Reduction

401(k) Example Baseline:

- Data Selection
  - Wave 4 of 1990 SIPP
  - Ignore panel aspect
  - Impose criteria (e.g. age selection) to limit sample to "interesting" population (control)
- Variable Selection
  - 655 raw variables (obviously many administrative/technical, highly unlikely to be related to problem of interest)
  - Select 9 as controls: income, age, family size, education, married, two-earner, defined benefit, ira, home-owner
- Functional form
  - no interactions, dummies for income categories, quadratic in age, dummies for schooling levels

"Formal" Dimension Reduction: Model

Model:

$$
\begin{aligned}
y_i &= d_i\alpha_0 + g(x_i) + \zeta_i \approx d_i\alpha_0 + z_i'\beta_g + \zeta_i \\
d_i &= m(x_i) + u_i \approx z_i'\beta_m + u_i \\
\mathrm{E}[\zeta_i|x_i, d_i] &= \mathrm{E}[u_i|x_i] = 0
\end{aligned}
$$

where z_i is a function of x_i.

"Reduced forms":

$$
\begin{aligned}
y_i &= z_i'\widetilde{\beta} + \widetilde{\zeta}_i \\
d_i &= z_i'\beta_m + u_i \\
\widetilde{\beta} &= \beta_g + \alpha_0\beta_m \\
\widetilde{\zeta}_i &= \zeta_i + \alpha_0 u_i
\end{aligned}
$$

"Formal" Dimension Reduction: Intuition

Intuition for what we get from each reduced form:

- Reduced form for treatment:
  - Find variables that are strongly related to treatment
  - Ignoring these potentially leads to omitted-variables-bias (OVB)
- Reduced form for outcome:
  - Find variables that are strongly related outcome
  - Improve efficiency
  - Reduce OVB which would result if associated coefficient in treatment equation is small but non-zero

Reduced forms are parts of problem data are informative about (predictive relationships)

Good methods for estimating these parts of the model

"Formal" Dimension Reduction: Choices

- Choice 1 (Data Set): Same as above
- Choice 2 (Baseline Variables): Same as above
  - Could consider adding additional variables
  - Each variable makes the haystack bigger
  - Think carefully. Do we want more variables or more flexibility in other dimensions?
- Choice 3 (Functional Form): Want to try to be very flexible in income (Details to follow)
- Choice 4 (Selection Method/Auxiliary Parameters): LASSO

Review: Feasible LASSO Allowing Heteroskedasticity

LASSO (heteroskedastic):

$$\widehat{\beta}_L = \arg\min \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \widehat{\lambda} \sum_{j=1}^{p} |\widehat{\gamma}_j \beta_j|$$

for a generic outcome (y) and set of regressors (x)

Need to fill in values for $\widehat{\lambda}$ and $\widehat{\gamma}_j$ for $j = 1, ..., p$.

Review: Feasible LASSO Allowing Heteroskedasticity

$\widehat{\gamma}_j$:

- Want $\widehat{\gamma}_j \approx \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \varepsilon_i^2$
- Estimate iteratively for a given value of $\widehat{\lambda}$.
  - Start with initial guess for β (β^*) $\rightarrow e_i = y_i - x_i'\beta^*$ (initial guess for $\{\varepsilon_i\}_{i=1}^n$)
  - Estimate new value of β by solving above problem for given value of $\widehat{\lambda}$ with $\widehat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 e_i^2$
  - Take new value of β to form new residuals and new set of $\widehat{\gamma}_j$ estimates
  - Iterate to convergence (or max number of iterations)

Review: Feasible LASSO Allowing Heteroskedasticity

$\widehat{\lambda}$:

1. Cross-validation
   - 10-fold CV in all examples
   - Use 1 s.e. rule as discussed by M. Taddy previously
   - Iterative estimation of penalty loadings inside of CV loop
     - Less stable than vanilla LASSO CV
     - 1 s.e. rule seems to add some stability
2. Theoretical value:
   - Simple bound 1: $2.2\Phi^{-1}(1-\frac{q}{2p})$
   - Simple bound 2: $2.2\sqrt{2n\log(2p/q)}$
   - Theoretically need $q \to 0$, q size of test of hypothesis that "biggest" coefficient equals 0 when all coeffients equal 0. We use $q = .05$ or $q = .1/\max\{p, n\}$ in our examples. (Bounds above are bounds on this critical value.)

"Formal" Dimension Reduction: Variables

- Dummies: married, two-earner, defined benefit, ira, home-owner
- Schooling: 5^{th} degree orthogonal polynomial (generated by Stata *orthpoly*)
- Family Size: 3^{rd} degree orthogonal polynomial (generated by Stata *orthpoly*)
- Age: Cubic spline with 10 equally spaced knots (30 terms)
- Income: Cubic spline with 15 equally spaced knots (45 terms)
- Interactions:
  - 1. Dummies with Schooling, Family Size, and Age terms (190 interactions)
  - 2. Income terms interacted with A. Dummies, Schooling, Family Size, and Age terms and B. interactions in 1. (10,485 interactions)
- 10,763 total variables

Dimension Reduction

So, n = 9,915; p = 10,763

Cannot identify effect of 401(k) if we think we actually need all of these terms to adequately control for income.

Have already intuitively reduced dimensionality by focusing on 9 controls that seem plausibly related to income and assets

Baseline results resolve this identification problem by assuming the functional form is known.

Rather than assume functional form, try to learn it from data using variable selection (from a flexible but still very parsimonious specification).

Results:

Note: Before constructing splines, income and age were put on [0,1] interval.

401(k) Eligibility:

- Selected variables (CV and Plug-in): *income*, $\max(0, \textit{income} - .33)$, *two–earner*, *defined–benefit*, *home–owner*, $\textit{education}^3$, $\max(0, \textit{age}^3 - .4)$, $\max(0, \textit{age}^3 - .5)$, $\textit{home–owner} * \textit{income}$
- Estimate 401(k) effects obtained by combining these variable with those selected for relevant outcome and running OLS

Results:

Net Financial Assets:

- Selected variables (CV and Plug-in): *income*, *two–earner*, *age*, *two–earner* $*$ *family–size*, *ira* $*$ *age*, *home–owner* $*$ *age*^{2}, *ira* $*$ *income*, *home–owner* $*$ *income*, *home–owner* $*$ *income*^{2}
- Estimated Effect: 8687.43 (1274.76) [Recall baseline estimates: 9216.5 (1340.6)]

Total Wealth:

- Selected variables (CV and Plug-in): *income*, *two–earner*, *ira* $*$ *age*, *home–owner* $*$ *age*, *home–owner* $*$ *age*^{2}, *ira* $*$ *income*, *home–owner* $*$ *income*, *age* $*$ *income*
- Estimated Effect: 5374.79 (1990.94) [Recall baseline estimates: 6612.0 (2110.1)]

ATE with Heterogeneous Treatment Effects:

We could also estimate the ATE allowing for full heterogeneity of treatment effects.

Model:

$$y_i = d_i g_1(x_i) + (1 - d_i) g_0(x_i) + \zeta_i$$
$$d_i = m(x_i) + u_i \quad \text{(as before)}$$

Following Hahn (1998), can estimate ATE (α) as

$$\widehat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{d_i(y_i - \widehat{g}_1(x_i))}{\widehat{m}(x_i)} - \frac{(1 - d_i)(y_i - \widehat{g}_0(x_i))}{1 - \widehat{m}(x_i)} + \widehat{g}_1(x_i) - \widehat{g}_0(x_i) \right)$$

where we obtain estimates of the functions $g_0(\cdot)$, $g_1(\cdot)$, and $m(\cdot)$ as above using variable selection methods.

ATE with Heterogeneous Treatment Effects:

Estimates of $m(\cdot)$ were obtained previously.

Net Financial Assets and Total Wealth:

- Use the same set of covariates broken out by values of 401(k) eligibilty dummy
- Use the CV penalty parameter from before scaled for the number of observations in each category (3682 are eligible, 6233 not-eligible)
  - I.e. multiply CV penalty parameter by $\sqrt{3682/9915}$ in eligible models and by $\sqrt{6233/9915}$ in not-eligible models

ATE Results:

- Financial Assets:
  - Selected Variables (eligible): $income$, IRA, age, $two\text{–}earner * family\text{–}size$, $IRA * age$, $home\text{–}owner * age$, $IRA * income$, $home\text{–}owner * income$
  - Selected Variables (non-eligible): IRA, $home\text{–}owner$, age, $IRA * age$, $home\text{–}owner * age$
  - Estimates: 8032.54 (1136.6)
- Total Wealth:
  - Selected Variables (eligible): $income^2$, $two\text{–}earner$, age, $IRA * age$, $home\text{–}owner * age$, $IRA * income$, $home\text{–}owner * income$, $age * income$
  - Selected Variables (non-eligible): $income$, IRA, $IRA * age$, $home\text{–}owner * age$, $home\text{–}owner * age^2$, $IRA * income$, $home\text{–}owner * income$
  - Estimates: 6180.29 (1828.5)

Example 2: Effect of Abortion on Crime

Goal: Understand causal effect of d_{it} (abortion) on y_{it} (crime). (Donohue and Levitt 2001)

Problem: Abortion rates are not randomly assigned

Key concern:

- states are different for lots of reasons
- crime rates in states evolve differently for lots of reasons
- factors that are associated to differences in states, state evolutions, etc. may also be related to differences in abortion rates, abortion rate evolution, etc.

Most of the favorite stories for confounding obviously fit here.

Example 2: Baseline Model

Donohue and Levitt (2001) baseline model

$$y_{it} = d_{it}\alpha_0 + x_{it}'\beta_g + \gamma_t + \delta_i + \varepsilon_{it}$$

- y_{it} = crime-rate (violent, property, or murder per 1000)
- d_{it} = "effective" abortion rate
- x_{it} = eight controls: log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, AFDC generosity at time $t - 15$, a dummy for concealed weapons law, and beer consumption per capita
- γ_t time effects
- δ_i state effects

Baseline Results:

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
DL Table 4	-.129	.024	-.091	.018	-.121	.047
First-Diff	-.152	.034	-.108	.022	-.204	.068

Use first-differences from now on.

Assumes confounds are time invariant, state invariant, or captured by small set of variables in x_{it}

State-specific characteristics related to features of abortion only allowed to be related to level of crime rate

I.e. evolution of abortion rates and crime rates unrelated after subtracting the mean

Flexible Trends

"Model" with flexible trends:

$$y_{it} = d_{it}\alpha_0 + g(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) + \zeta_{it}$$
$$d_{it} = m(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) + u_{it}$$

Want to allow flexible state-specific trends BUT clearly can't learn about about effect if trends are allowed to do anything.

I.e. if $m(\cdot)$ and $g(\cdot)$ can vary arbitrarily across i and t clearly can't identify α_0

Need to reduce the dimension.

Intuitive Dimension Reduction

Abortion Baseline:

- Functional form
  - $g(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) = x_{it}'\beta_g + \gamma_t + \delta_i$
  - additively separable
  - (Correlated) Evolution of state crime and abortion rates captured by macro-economy (γ_t), constant state-specific level shifts from aggregate (δ_i), and small number of time varying variables
- Variable Selection
  - Select 8 time varying state-level control variables (of the many state-level macro series available)

What if there are (correlated) differences in abortion and crime evolution not captured by aggregate evolution?

Formal Dimension Reduction

Model (First-Differences):

$$\Delta y_{it} = \Delta d_{it}\alpha_0 + \gamma_t + w_{it}'\pi_1 + \varepsilon_{it}$$
$$\Delta d_{it} = \kappa_t + w_{it}'\pi_2 + u_{it}$$

- Use first-difference to remove state effects
- time effects (not-selected over), included in both reduced form models
- With $w_{it} = \Delta x_{it}$ first-difference version of Donohue and Levitt (2001) model. (Results presented above.)

Formal Dimension Reduction: Variables

- w_{it}:
  1. differences in eight original controls
  2. initial conditions of controls, abortion rate, crime rate
  3. within state averages of controls, abortion rate
  4. t, t^2
  5. interactions of 1-3 with 4
  - corresponds to a model for crime and abortion rates with a cubic trend that may depend on baseline state characteristics
- $p = 284$
- $n = 576$

Variables in 2-5 motivated by a desire to have a flexible, sensible model of evolution

Kitchen Sink:

Why not use "everything"?

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
All Controls	.006	.755	-.154	.224	2.240	2.804

A flexible cubic trend arguably isn't going crazy, but everything is rendered very imprecise.

Probably a lot of things added aren't really important

Variable Selection:

Variables in each equation selected using LASSO.

- 10-Fold CV
- Plug-in penalty parameter: $2.2\sqrt{2n(\log(2p/.05))}$
- Penalty loadings estimated using iterative Lasso with 100 max iterations

Variable Selection:

Violent crime selection:

- Abortion equation (CV and Plug-in): (11) lagged prisoners, lagged police, lagged unemployment, initial income, initial income difference $\times$ t, initial beer consumption difference $\times$ t, initial income $\times$ t, initial prisoners squared $\times$ t^2, average income, average income $\times$ t, initial abortion rate
- Crime equation (CV): No variables
- Crime equation (Plug-in): Initial difference in abortion rate $\times$ t, Initial abortion rate $\times$ t

Variable Selection:

Property crime selection:

- Abortion equation (CV): (7) lagged prisoners, lagged income, initial income, initial income difference, initial income difference $\times$ t, average income, initial abortion rate
- Abortion equation (Plug-in): (12) lagged prisoners, lagged police, lagged income, Initial income difference, initial income, initial income difference $\times$ t, initial beer difference $\times$ t, initial prisoners squared $\times$ t, initial prisoners squared $\times$ t^2, initial beer squared $\times$ t^2, average income, initial abortion rate
- Crime equation (CV): No variables
- Crime equation (Plug-in): (3) Initial income squared $\times$ t, Initial income squared $\times$ t^2, average AFDC squared

Variable Selection:

Murder rate selection:

- Abortion equation (CV): (5) lagged prisoners, lagged unemployment, initial income $\times$ t, average income $\times$ t, initial abortion rate
- Abortion equation (Plug-in): (9) lagged prisoners, lagged unemployment, initial unemployment difference squared, initial prisoners $\times$ t, initial income *times* t, initial beer difference $\times$ t^2, average income $\times$ t, initial abortion rate, initial abortion rate $\times$ t
- Crime equation (CV and Plug-in): No variables selected

Estimated Effects of Abortion on Crime

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
DL Table 4	-.129	.024	-.091	.018	-.121	.047
First-Diff	-.152	.034	-.108	.022	-.204	.068
All Controls	.006	.755	-.154	.224	2.240	2.804
Post-DS (CV)	-.119	.120	-.042	.059	-.122	.131
Post-DS (Plug-in)	-.174	.120	-.052	.070	-.123	.148

- "Post-DS" Results in-line with critique raised by Foote and Goetz (2008).

Example 3: Institutions and Growth (AJR 2001)

Equation of interest:

$$\log(GDP\ per\ capita_i) = \alpha(Protection\ from\ Expropriation_i) + x_i'\beta + \varepsilon_i$$

Endogeneity/Simultaneity:

- better institutions may lead to higher incomes
- higher incomes may lead to the development of better institutions

Instrument

Instrument = European Settler Mortality:

- First Stage: settlers set up better institutions in places they might stick around in (i.e. where they were less likely to die) and institutions are highly persistent
- Exclusion: GDP, while persistent, is unlikely to be strongly influenced by the factors that determined the exact development of institutions 100+ years ago except through the institutions established

Controls

- There may be other factors that are highly persistent and related to development of institutions and GDP
- leading candidate - geography (Geographic Determinism)
- Want to control for geography and use variation in mortality not captured by geography
  - Baseline AJR results control linearly for latitude
  - AJR consider continent dummies, split by continent, first-stage gets weak with some of these

Baseline estimates find strong positive effect of institutions:

- First-stage: -0.5372 (0.1545)
- α: 0.9692 (0.2128)

Setting

Want to control flexibly for geography but may lose power to identify effect of institutions.

IV with one instrument and unknown controls:

$$
\begin{aligned}
y_i &= \alpha d_i + x_i'\beta + \varepsilon_i \\
d_i &= \pi_1 z_i + x_i'\Pi_2 + v_i \\
z_i &= x_i'\gamma + u_i
\end{aligned}
$$

- Believe z_i is a valid instrument after controlling for x_i.

Setting

Same as estimating regression coefficients after conditioning but now have three reduced form/prediction equations:

$$
\begin{aligned}
y_i &= x_i'\widetilde{\beta} + \widetilde{\varepsilon}_i \\
d_i &= x_i'\widetilde{\Pi}_2 + \widetilde{v}_i \\
z_i &= x_i'\gamma + u_i
\end{aligned}
$$

- Do variable selection on the three equations and use union of selected variables as controls.

Variable Selection

Select over a flexible function of geography:

- Africa, Asia, North America, South America (dummies)
- latitude, latitude^2, latitude^3, (latitude-.08)$_+$, (latitude-.16)$_+$, (latitude-.24)$_+$, ((latitude-.08)$_+$)^2, ((latitude-.16)$_+$)^2, ((latitude-.24)$_+$)^2, ((latitude-.08)$_+$)^3, ((latitude-.16)$_+$)^3, ((latitude-.24)$_+$)^3

Using all these variables results in a very weak first-stage:

- First-stage: -0.2164 (0.2191)
- α: 0.9480 (0.7384) (and unreliable)

Variable Selection:

Variables for each equation selected with Lasso:

- Penalty loading calculated with iterative method
- 10-Fold CV and Plug-in $[2.2\sqrt{2n(\log(2p/\gamma))}$ with $\gamma = .1/\log(n)]$ give same results
- GDP - Africa
- Expropriation - Africa
- Mortality - Africa

Results:

	Latitude	All Controls	Selection
First Stage	-0.5372	-0.2164	-0.5429
	(0.1545)	(0.2191)	(0.1719)
Second Stage	0.9692	0.9480	0.7710
	(0.2128)	(0.7384)	(0.1971)

- First Stage - Coefficient on Settler Mortality
- Second Stage - Coefficient on Protection from Expropriation

Example 4: Eminent Domain

Estimate economic consequences of the law of takings or eminent domain (following Chen and Yeh (2010))

Consider effect on Case-Shiller Price Index.

Eminent domain (or law of takings): when a government actor physically acquires the property rights of one or more individuals

Laws/judicial decisions may not be exogenous.

Instruments

Random assignment of judges to three judge federal appellate panels

$$\Rightarrow$$

Panel demographics randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year

Why variable selection?

Many characteristics of three judge panels

Any set of characteristics of the three judge panel unrelated to structural unobservable

Some instruments may be more valuable than others

Could attempt to solve through intuition. Number of judges who are democrats:

- Judges' political affiliation known to predict decisions for many outcomes
- First Stage: 0.0664 (0.0713)
- Second Stage: -0.2583 (0.5251) (Unreliable)

Model

Econometric model:

$$y_{ct} = a_c + b_t + g_c t + \theta \ \mathit{Takings\ Law}_{ct} + W'_{ct} d + \epsilon_{ct}$$
$$\mathit{Takings\ Law}_{ct} = \alpha_c + \beta_t + \gamma_c t + W'_{ct} \delta + z'_{ct} \Pi + v_{ct}$$

- c circuit; t time
- y_{ct}: log(house price index) or log(GDP)
- $\mathit{Takings\ Law}_{ct}$: number of pro-plaintiff (overturn gov't taking) apellate takings decisions
- (a_c, α_c), (b_t, β_t), and $(g_c t, \gamma_c t)$: circuit-specific effects, time-specific effects, and circuit-specific time trends.
- (d, δ) coefficients on exogenous variables
- z_{ct} instruments with coefficients Π

Model

Controls (W_{ct}):

- $\approx$ 30 probability controls for panel demographics
- a dummy for no cases in that circuit-year
- number of takings appellate decisions

θ: effect of an additional decision upholding individual property rights on an economic outcome

Any set of characteristics of three judge panels is potentially an instrument.

Instruments

Do *ex ante* dimension reduction by intuively selecting characteristics thought to have strong signal about judge preferences over government vs. individual property rights:

- 42 baseline variables
    - gender, race, religion (jewish, catholic, protestant, evangelical, none) , political affiliation, bachelor obtained in-state, bachelor from public university, JD from a public university, has an LLM or SJD, elevated from a district court
    - number of panels with 1, 2, or 3 members with each characteristic
- $+$
    - cubic in number panels democrat, number with JD from public university, number elevated from district
    - first order interactions between all variables

Additional pre-processing

Remove instruments likely to be irrelevant based on features of instrument vector alone:

- remove any instrument with mean $< .05$, standard deviation after partialling out controls $< .000001$
- remove one from each pair of any pair with bivariate correlation $> .99$ in absolute value

Note: Selection based on characteristics of z cannot introduce bias under exclusion restriction.

Leaves 147 instruments (n $= 183$)

Obtaining the Estimates

Method: Post-LASSO

Penalty parameters and loadings:

- Penalty loading calculated through iterative scheme
- 10-fold Cross-validation to obtain penalty
- Plug-in penalty: $2.2\sqrt{n}\Phi^{-1}(1-\gamma/(2p))$, $\gamma = .1/\log(n)$ (Gives same results as 10-fold CV)

Results: Case-Shiller Prices

LASSO Selection Results:

- 1+ JD public squared

- First Stage: 0.4495 (0.0511) [Using 1+ democrat: 0.0664 (0.0713)]
- Second Stage: 0.0648 (0.0240) [Using 1+ democrat: -0.2583 (0.5251) (Unreliable)]