The Factor-Lasso and K-Step Bootstrap Approach for Inference in High-Dimensional Economic Applications

Christian Hansen   Yuan Liao

May 2017
Montreal

- Observe many control variables
- Two popular (formal) dimension blueuction techniques:
  Variable/model selection - e.g. lasso
  Factor models

Variable Selection Review

(α parameter of interest):

$$y_i = \alpha d_i + x_i'\beta + \varepsilon_i$$
$$d_i = x_i'\gamma + u_i$$

1. Allow MANY control variables
2. Impose SPARSITY on β, γ

- Literature: Belloni, Chernozhukov and Hansen (12 *REStud.*), etc.
- weak dependence among x
- just a few x have impact on y, d

Large Factor Model Review

(α parameter of interest):

$$
\begin{aligned}
y_i &= \alpha d_i + f_i' \beta + \varepsilon_i \\
d_i &= f_i' \gamma + v_i \\
x_i &= \Lambda f_i + U_i
\end{aligned}
$$

1. Most of x have impact on y, d.
2. dimension of f_i is small

- Literature: Factor augmented regressions, diffusion index forecast (e.g. Bai and Ng (03), Stock and Watson (02))
- Generally results in strong dependence among x
- Regression directly on x will generally NOT produce sparse coefficients
- Do not worry about the "remaining information" in U_i

What we aim to do

nests large factor models and variable selection.

$$\begin{aligned} y_i &= \alpha d_i + f_i'\beta + U_i'\theta^y + \varepsilon_i \\ d_i &= f_i'\gamma + U_i'\theta^d + v_i \\ x_i &= \Lambda f_i + U_i \end{aligned}$$

1. U_i represent variation in observables not captured by factors
2. estimation method: lasso on U_i.
3. Justifications of key assumptions for lasso:

- Weak dependence among regressors:

  Most variations in x are driven by factors.

- Sparsity of θ:

  only a few x have “useful remaining information” after factors are controlled.

Some "why not" questions we had...

1. control for (f_i, x_i) instead of (f_i, U_i):

$$y_i = \alpha d_i + f_i'\beta + x_i'\theta^y + \varepsilon_i$$
$$d_i = f_i'\gamma + x_i'\theta^d + v_i$$
$$x_i = \Lambda f_i + U_i$$

- within x_i: strongly correlated.
- between x_i and f_i: strongly correlated.

2. Use lots of factors

$$y_i = \alpha d_i + f_i'\beta + \varepsilon_i$$
$$d_i = f_i'\gamma + v_i$$
$$x_i = \Lambda f_i + U_i$$

- Allow $\dim(f_i)$ to increase fast with $p = \dim(x_i)$
- Assume (β, γ) sparse, then "lasso" them.
- No sufficient amount "cross-sectional" information for factors
- Estimating factors is either inconsistent or with slow rate, impacting inference on α

Some "why not" questions we had...

3. Sparse PCA

$$x_{i,l} = \lambda_l' f_i + U_i, \quad l = 1, ..., p, \quad i = 1, ..., n$$

- Most of $(\lambda_1, ..., \lambda_p)$ are zero.
- Most of x do not depend on factors. Become a sparse model:

$$y_i = \alpha d_i + x_i'\beta + \varepsilon_i$$
$$d_i = x_i'\gamma + u_i$$

What we do

$$y_i = \alpha d_i + f_i'\beta + U_i'\theta^y + \varepsilon_i$$
$$d_i = f_i'\gamma + U_i'\theta^d + v_i$$
$$x_i = \Lambda f_i + U_i, \quad i = 1, ..., n$$

- Do not directly observe (f, U) ; (θ^y, θ^d) are sparse
- $\dim(f_i)$, $\dim(\alpha)$ are small.

1. Estimate (f, U) from the third equation
2. Lasso on
$$y_i - \widehat{E(y_i|f_i)} = \widehat{U}_i'\theta^{new} + \varepsilon_i^{new}, \quad \varepsilon_i^{new} = \alpha v_i + \varepsilon_i$$
$$d_i - \widehat{E(d_i|f_i)} = \widehat{U}_i'\theta^d + v_i$$
3. OLS on
$$\widehat{\varepsilon_i^{new}} = \alpha \widehat{v}_i + \varepsilon_i$$

Extensions: I, II

I: endogenous treatment

$$
\begin{aligned}
y_i &= \alpha d_i + f_i'\beta + U_i'\theta^y + \varepsilon_i \\
d_i &= \pi z_i + f_i'\gamma + U_i'\theta^d + v_i \\
z_i &= f_i'\psi + U_i'\theta^z + u_i \\
x_i &= \Lambda f_i + U_i, \quad i = 1, ..., n
\end{aligned}
$$

II: diffusion index forecast

$$
\begin{aligned}
y_{t+h} &= \alpha y_t + f_t'\beta + U_t'\theta + \varepsilon_{t+h} \\
x_t &= \Lambda f_t + U_t, \quad t = 1, ..., T.
\end{aligned}
$$

Include U_t to capture idiosyncratic information in x_t.

Extensions: III Panel data

What we focused on in this paper:

$$y_{it} = \alpha d_{it} + (\lambda_t^y)' f_i + U_{it}'\theta^y + \mu_i^y + \delta_t^y + \epsilon_{it}$$
$$d_{it} = (\lambda_t^d)' f_i + U_{it}'\theta^d + \mu_i^d + \delta_t^d + \eta_{it}$$
$$X_{it} = \Lambda_t f_i + \mu_i^X + \delta_t^X + U_{it}, \quad i \leq n, t \leq T, \dim(X_{it}) = p$$

- μ_i and δ_t are unrestricted individual and time effects
- $p \to \infty$, $n \to \infty$,
- T is either fixed or growing but satisfy $T = o(n)$, because: need accurate estimation of U_{it}, relying on estimating Λ_t
- $n = o(p^2)$ because need accurate estimation of f_i.

Define

$$\sigma_{\eta\epsilon} = \mathrm{Var}\left(\frac{1}{\sqrt{nT}}\sum_{i,t}(\eta_{it} - \bar{\eta}_i)(\epsilon_{it} - \bar{\epsilon}_i)\right) \qquad \widehat{\sigma}_{\eta\epsilon} = \frac{1}{nT}\sum_i\left(\sum_t \widehat{\eta}_{it}\widehat{\epsilon}_{it}\right)^2$$

$$\sigma_\eta^2 = \mathrm{E}\left(\frac{1}{nT}\sum_{i,t}(\eta_{it} - \bar{\eta}_i)^2\right) \qquad \widehat{\sigma}_\eta^2 = \frac{1}{nT}\sum_{i,t}\widehat{\eta}_{it}^2$$

$$\sigma_\eta^2 \sigma_{\eta\epsilon}^{-1/2}\sqrt{nT}(\widehat{\alpha} - \alpha) \xrightarrow{d} N(0,1)$$

$$\widehat{\sigma}_\eta^2 \widehat{\sigma}_{\eta\epsilon}^{-1/2}\sqrt{nT}(\widehat{\alpha} - \alpha) \xrightarrow{d} N(0,1)$$

Additional comments:

- Not clear that you could get these results even if $\lambda_t^y = 0$ were known due to strong dependence in X resulting from presence of factors
- First taking care of factor structure in X seems potentially important

Extensions of Inference I: K-Step Bootstrap

Alternative to inference from plug-in asymptotic distribution is bootstrap inference

Full bootstrap lasso:

- Generate bootstrap data $(X_i,^*, Y_i^*)$
- $$\widehat{\beta}^* = \arg\min \frac{1}{n}\sum_{i=1}^{n}(Y_i^* - X_i^{*T}\beta)^2 + \lambda\|\beta\|_1$$
- Repeat B times.

Full bootstrap lasso is potentially burdensome.

K-Step Bootstrap

Consider a K-Step bootstrap in Andrews (2002):

- Start lasso at full sample solution ($\widehat{\beta}_{lasso}$)
- For each bootstrap data, initialize at $\widehat{\beta}_0^* = \widehat{\beta}_{lasso}$
- Employ iterative algorithms: Obtain

$$\widehat{\beta}_{lasso} = \widehat{\beta}_0^* \Rightarrow \widehat{\beta}_1^* \Rightarrow \ldots \Rightarrow \widehat{\beta}_k^*$$

- Similar to Andrews 02, each step is in closed form - fast even in large problems
- Different from Andrews 02, each step is still an l_1-penalized problem

Coordinate descent (Fu 1998)

- Update one component at a time, fixing the remaining components:

$$\min_{\beta_j} \frac{1}{n}\sum_i (Y_i^* - \underbrace{X_{i,-j}^{*\prime}\widehat{\beta}_{\ell,-j}^*}_{\text{others, known}} - X_{ij}\beta_j)^2 + \lambda|\psi_j\beta_j| = \min_{\beta_j} L_\ell(\beta_j) + \lambda|\psi_j\beta_j|$$

$$\widehat{\beta}_{\ell+1,j}^* = \arg\min_{\beta_j} L_\ell(\beta_j) + \lambda|\psi_j\beta_j|$$

for $j = 1, ..., p$.

- Each $\widehat{\beta}_{\ell+1,j}^*$ is closed form = soft-thresholding.

$$\arg\min_{\beta\in\mathbb{R}} \frac{1}{2}(z-\beta)^2 + \lambda|\beta|$$

$$= sgn(z)\max(|z|-\lambda, 0)$$

- "Composite Gradient descent" (Nesterov 07, Agarwal et al. 12 *Ann. Statist.*)
  update the entire vector at once

  $$\text{originally:} \quad \widehat{\beta}^*_{l+1} = \arg\min_{\beta}(\beta - \widehat{\beta}^*_l)' V(\beta - \widehat{\beta}^*_l) + b'(\beta - \widehat{\beta}^*_l) + \lambda\|\psi\beta\|_1$$

  Replace V by $\frac{h}{2}\times$ identity

  $$\Rightarrow \text{ the entire vector is in closed form= soft thresholding}$$

- choose h:
  if dimension is small, use $h = 2\lambda_{\max}(V)$ to "majorize" V
  If dimension is large, $2\lambda_{\max}(V)$ is unbounded (Johnstone 01)

General Conditions for Iterative Algorithms

$$Q(\beta) = \frac{1}{n}\|Y^* - X^*\beta\|_2^2 + \lambda\|\Psi\beta\|_1$$

Suppose $\widehat{\beta}_k^*$ satisfies:

1. minimization error is smaller than statistical error.

$$Q(\widehat{\beta}_k) \leq \min_{\beta} Q(\beta) + o_{P^*}(|\widehat{\beta} - \beta_0|)$$

2. sparsity:

$$|\widehat{\beta}_k|_0 = O_{P^*}(|J|_0).$$

Can be directly verified using the KKT condition

We verified both conditions for the Coordinate descent (Fu 98)

Bootstrap Confidence Interval

Let $q^*_{\tau/2}$ be the $\tau/2^{\text{th}}$ upper quantile of $\{\sqrt{nT}|\widehat{\alpha}^b - \widehat{\alpha}| : b = 1, ..., B\}$

k-step bootstrap does not affect first-order asymptotics. (proved for linear model)

- $P\left(\alpha \in \widehat{\alpha} \pm q^*_{\tau/2}/\sqrt{nT}\right) \to 1 - \tau.$
- extendable to nonlinear models with orthogonality conditions

Technical remarks

- We spent most of the time proving:
  The effect of estimating (f, U) is first-order negligible under weakest possible conditions on (n, T, p)
- Require weighted errors of the form:

$$\max_{d\leq p}|\frac{1}{n}\sum_i(\widehat{f}_i - f_i)w_{id}|, \quad \max_{d\leq p}|\frac{1}{nT}\sum_{it}(\widehat{f}_i - f_i)z_{it,d}|$$

- Easy to bound using Cauchy-Schwarz and $\frac{1}{n}\sum_i\|\widehat{f}_i - f_i\|^2$
  But very crude, leading to stronger than necessary conditions
- Need to use the expansion of $\widehat{f}_i - f_i$ ($\widehat{f}_i =$ PCA estimator)
- If $\widehat{f}_i$ has no closed form (e.g., MLE), need its Bahadur expansion

II: factor augmented regression:

$$\begin{aligned} y_t &= \alpha d_t + f_t'\beta + U_t'\theta^y + \varepsilon_t \\ d_t &= f_t'\gamma + U_t\theta^d + v_t \\ x_t &= \Lambda f_t + U_t, \quad t = 1, ..., T \end{aligned}$$

- $\alpha \perp E(y_t|f_t, U_t), E(d_t|f_t, U_t)$, Lasso does NOT affect first-order asymptotics (Robinson 88, Andrews 94, Chernozhukov et al 16)
- Apply HAC (Newey-West)

III: Out-of- sample forecast interval

$$\begin{aligned} y_{t+h} &= \underbrace{\alpha y_t + f_t'\beta + U_t'\theta}_{y_{t+h|t}} + \varepsilon_{t+h} \\ x_t &= \Lambda f_t + U_t, \quad t = 1, ..., T. \end{aligned}$$

$y_{T+h|T} \not\perp U_t'\theta$, Lasso estimation of $U_t'\theta$ DOES affect confidence interval for $y_{T+h|T}$

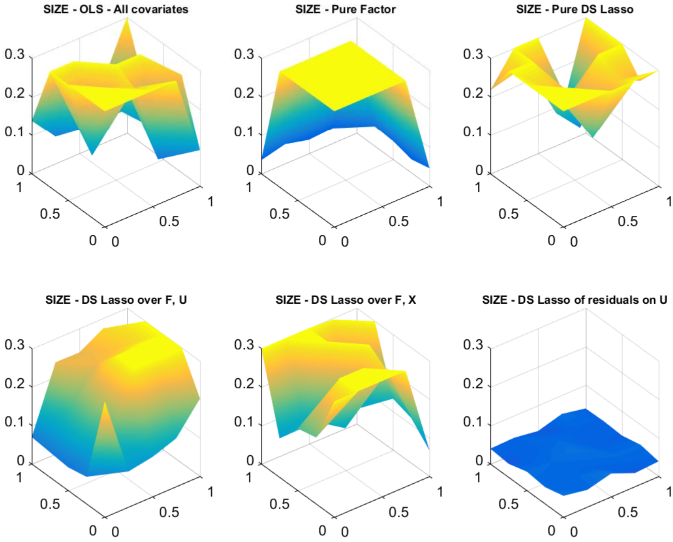
Panel Linear Model Simulations

Linear Panel Model Simulation:

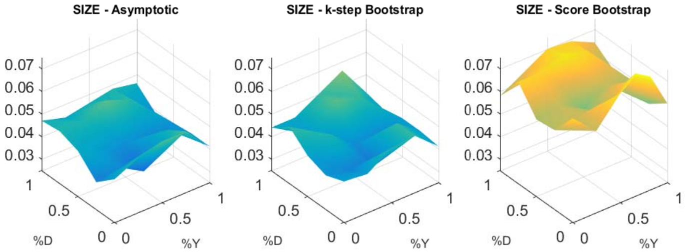
- n = 100, T = 10, p = 100 (number of covariates), r = 3 (number of factors)
- For X: Factors (on average) contribute 50% of variation; U contributing remaining 50%
- For Y and D: F and U contribute 70% of variation .

  Individual contributions of F vary. (given on horizontal axes on figures)
- $\theta_j^y = c_y 1/j^2$, $\theta_j^d = c_d 1/j^2$

Panel Linear Model Simulations: (Truncated) Size of 5% Test



Panel Linear Model Simulations: Bootstrap Size of 5% Test



score bootstrap: Kline and Santos (2012):

$$\widehat{\sigma}_{\eta}^{-2}\frac{1}{\sqrt{nT}}\sum_{it}\widehat{\eta}_{it}\widehat{\epsilon}_{it}w_{it}^{*}, \quad Ew_{it}^{*}=0, \quad Ew_{it}^{*2}=1.$$

Institutions and Growth (AJR 2001)

Equation of interest:

$$\log(\textit{GDP per capita}_i) = \alpha(\textit{Protection from Expropriation}_i) + U_i'\beta + \lambda' f_i + \varepsilon_i$$

$$(\textit{Protection from Expropriation}_i) = \pi(\textit{Early Settler Mortality}_i) + U_i'\tilde{\beta} + \tilde{\lambda}' f_i + \varepsilon_i$$

- "Protection from Expropriation" is a measure of the strength of individual property rights that is used as a proxy for the strength of institutions
- Acemoglu et al. (2001, *AER*) instrument: Early settler mortality
- Controls: Need to control for other factors that are highly persistent and related to development of institutions and GDP
  - Leading candidate: Geography (geographic determinism)

Potential Control Variables

Potential geographic controls:

1. Africa, Asia, North America, South America (dummies)
2. longitude, renewable water, land boundary, land area, coastline, territorial sea, arable land, average temperature, average high temp, average low temp, average precipitation, highest point, lowest point, low-lying area
3. 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
4. dist, dist^2, dist^3, (dist-.25)$_+$, (dist-.375)$_+$, (dist-.5)$_+$, ((dist-.25)$_+$)^2, ((dist-.375)$_+$)^2, ((dist-.5)$_+$)^2, ((dist-.25)$_+$)^3, ((dist-.375)$_+$)^3, ((dist-.5)$_+$)^3 (dist = distance from London)

Results:

	Latitude	All	Lasso	Factor	Factor-Lasso
First Stage	-0.55	-0.04	-0.33	-0.34	-0.21
s.e.	(0.17)	(0.41)	(0.19)	(0.18)	(0.20)
Second Stage	0.93	3.07	0.71	1.26	1.40
s.e.	(0.21)	(32.82)	(0.40)	(0.53)	(1.17)

- First Stage - Coefficient on Settler Mortality
- Second Stage - Coefficient on Protection from Expropriation
- When only “Latitude” is controlled, the instrument is strong
- But the instrument looks pretty weak with more controls. Thus the result is different from Acemoglu et al. (2001)’s.

Summary of empirical findings

- draw substantively different conclusions about the strength of identification than Acemoglu et al. (2001),
  due to the ability to control more.
- Overall, usefully complement the sensitivity analyses performed in empirical studies and also have the potential to strengthen the plausibility of any conclusions drawn.