Machine Learning in Economic Analysis

Serena Ng
Columbia University and NBER

September 2016
Machine Learning: What's in it for Economics
Becker Friedman Institute
University of Chicago

Two creative papers

- Structural estimation of discrete choice models using random projections to reduce data dimension.
  - 3000+ combinations of soft drinks/store.
- Analyze connectedness using a regularized SVAR.
  - connectedness has a spatial and a cyclical component.
- Structural analysis $(\widehat{\beta})$ or prediction $(\widehat{y})$?
- Explore global banks data using ML methods.

Matrix Sketching: $\widetilde{A} = \underbrace{A}_{m\times n}\underbrace{\mathcal{S}}_{n\times k}$

- Goal: given A of high dimension, map it to lower dimension while preserving the structural features of A.

- PCA: choose a small number of directions in which the original data have high variance.

  - preserves average pairwise distance, but a few distance can be drastically violated.
  - relation to factor models.
  - statistical properties can be analyzed.
  - but even partial SVD can be computationally expensive.

Random projections (RP)

RP: preserves all $\binom{n}{2}$ pairwise distance of data points.

- may sacrifice overall variance.
- worst case error bounds
- optimal from algorithmic perspective.

Random Projections

- Linear algebra: a projection is a linear transformation P from a vector space to itself such that $P = P^2$.
    - e.g $A = U\Sigma V^T = QR$, then $P = UU^T = QQ^T$.
    - P has eigenvalues 0 or 1, and P is idempotent.
- the 'projection' in RP is somewhat different
    - if $[P]_{ij}$ is iid Gaussian, the range of $P^T P$ is a uniformly distributed subspace but eigenvalues $\notin \{0, 1\}$.
    - If $[P]_{ij}$ is $\{\pm 1\}$, P is *approximately* unit length and *approximately* orthogonal.

Informal arguments of RP

- We want to put m points in $\mathbb{R}^n$ and put them in $\mathbb{R}^k$.
- Naive approach: choose k columns uniformly at random.
    - if features are *spread out* (uniformity): works well.
    - if some columns contribute more and we do not find them, the approximation will be poor.
- Idea of random projections: randomly rotate the original data to get a new random basis. In that basis, the vectors are roughly uniformly spread out.

RP Implementations

- Choice of $\mathcal{S}$
    - Dense: $\mathcal{S}_{ij} = N(0,1)$
    - Sparse: $\mathcal{S}_{ij} = \pm 1$ with prob $\frac{1}{2s}$ and 0 with prob $1 - \frac{1}{s}$.
    - SRHT, count sketch, many alternatives.
- Sketching error and k:
    - With probability at least $1/2$, all pairwise distance will be preserved if, for $\epsilon \in (0, 1/2)$, $k \propto \frac{\log(m)}{\epsilon^2 - \epsilon^3}$.
    - k is logarithmic in m but does not depend on n.
    - Worst-case approximation error depends on ϵ.

Remarks

1 Projected data have no interpretation. Do we care?

2 Favorable worst-case errors $\Rightarrow$ favorable MSE$(\widehat{\beta})$?

- Linear Regression: $y = X\beta_0 + e$, $\min_\beta \|\mathcal{S}y - \mathcal{S}X\beta\|$.
- $\mathcal{S}$, random sampling/rescaling matrix. Let $W = \mathcal{S}'\mathcal{S}$.

$$\widehat{\beta}_W = (X^T W^{-1} X)^{-1} X^T W y.$$

$\widehat{\beta}_W$ depends on random weights. Some issues:

Remarks

1. Projected data have no interpretation. Do we care?

2. Favorable worst-case errors $\Rightarrow$ favorable MSE($\widehat{\beta}$)?

   - Linear Regression: $y = X\beta_0 + e$, $\min_\beta \|\mathcal{S}y - \mathcal{S}X\beta\|$.
   - $\mathcal{S}$, random sampling/rescaling matrix. Let $W = \mathcal{S}'\mathcal{S}$.

$$\widehat{\beta}_W = (X^T W^{-1} X)^{-1} X^T W y.$$

$\widehat{\beta}_W$ depends on random weights. Some issues:

- Like GLS. Finite sample properties not known.
- GLS improves efficiency, here weighting adds noise.
- Is strict exogeneity satisfied? $E[e_i^*|X_1^*, \ldots, X_n^*]=0$?
- Do we care about mse($\widehat{\beta}$)? or mse($\widehat{y}$))
- Know little even in point identified models.

An RP alternative: Random Sampling of $A = U\Sigma V^T$

$A_{m\times n}$, $m >> n$. Choose k rows.

- Select representative rows to capture the structure of U.
- Statistical leverage scores: $\ell_i = \| U_i \|_2^2 \,, i = 1, \ldots, m$.
- Importance sampling distribution: $p_i = \frac{\ell_i}{n}$
- $\ell_i = H_{ii} = [A(A'A)^{-1}A']_{ii}$. Hat matrix. Choose rows with large influence to account for non-uniformity.
- Error bound: $\left\| U^T U - \widetilde{U}^T \widetilde{U} \right\|_2 < \epsilon$.

Regression: (y, X), $n >> p$. Choose r rows.

(Drineas at al, 2011): If $r = O(f(p, \epsilon, \delta))$ with prob $> 1 - \delta$,

$$||\widehat{\beta} - \widehat{\beta}_W||_2 \leq \frac{\epsilon}{\sigma_{\min}(X)} ||Y - X\widehat{\beta}||_2.$$

Regression: (y, X), $n >> p$. Choose r rows.

(Drineas at al, 2011): If $r = O(f(p, \epsilon, \delta))$ with prob $> 1 - \delta$,

$$||\widehat{\beta} - \widehat{\beta}_W||_2 \leq \frac{\epsilon}{\sigma_{\min}(X)}||Y - X\widehat{\beta}||_2.$$

- Sampling with replacement. $w_i \sim$ scaled multinomial with $E[w_i] = 1$. $\widehat{\beta}_{OLS} = \widehat{\beta}_W(1)$. TSE of $\widehat{\beta}_W$ around $w_0 = 1$:

$$\mathbb{E}_W[\widehat{\beta}_W|y] = \widehat{\beta}_{OLS} + E_W[R_W].$$

Regression: (y, X), $n >> p$. Choose r rows.

(Drineas at al, 2011): If $r = O(f(p, \epsilon, \delta))$ with prob $> 1 - \delta$,

$$||\widehat{\beta} - \widehat{\beta}_W||_2 \leq \frac{\epsilon}{\sigma_{\min}(X)} ||Y - X\widehat{\beta}||_2.$$

- Sampling with replacement. $w_i \sim$ scaled multinomial with $E[w_i] = 1$. $\widehat{\beta}_{OLS} = \widehat{\beta}_W(1)$. TSE of $\widehat{\beta}_W$ around $w_0 = 1$:

$$\mathbb{E}_W[\widehat{\beta}_W | y] = \widehat{\beta}_{OLS} + E_W[R_W].$$

- R_W depends on sampling process and $\text{var}_W(\widehat{\beta}_W | y)$ decreases with rows selected.

- Favorable algorithmic properties (worse case error bounds) may not translate into good (statistical properties) MSE. (Ma, Mahoney, Yu 2015).

Summary of subspace sampling methods

- Random projections:
  - $\mathcal{S}$ is data oblivious. Uniformize data, then sample.
  - Projected data are linear combin. of the original data.
- Leverage score sampling:
  - $\mathcal{S}$ depends on data through leverage scores.
  - The columns of submatrix are columns of A.

Summary of subspace sampling methods

- Random projections:
  - $\mathcal{S}$ is data oblivious. Uniformize data, then sample.
  - Projected data are linear combin. of the original data.
- Leverage score sampling:
  - $\mathcal{S}$ depends on data through leverage scores.
  - The columns of submatrix are columns of A.
- Suggestions and Questions:
- If columns contribute uniformly, can just sample u.a.r. Document properties of the data?
- β is homogeneous. How much data to use? Aggregate?
- Understand a well identified example first?

Using these methods for summarizing data

- Lots are still unknown about statistical implications of subspace sampling methods for $\widehat{\beta}$.
- How useful are they in describing data, $(\widehat{y})$?
- Global banking data (96 banks, 2675 days).
- Four observations
  - clusters
  - row and column leverage scores
  - common factors or network spillovers?
  - connectedness: top-down or bottom up.

Frank Diebold kindly provided the data.

1. Kmeans

Group 1

jpm	bac	c	wfc	ms	bk.us
pnc.us	cof	stt.us	fitb.us	rf.us	sti.us
gs	usb	axp	bbt	mqg.au	
na.t	td.t	ry.t	bns.t	bmo.t	cm.t

1. Kmeans

Group 1

jpm	bac	c	wfc	ms	bk.us
pnc.us	cof	stt.us	fitb.us	rf.us	sti.us
gs	usb	axp	bbt	mqg.au	
na.t	td.t	ry.t	bns.t	bmo.t	cm.t

Group 1: Canada/US (23)

Group 2

hsba.ln	bnp.fr	dbk.xe	barc.ln
aca.fr	gle.fr	rbs.ln	san.mc
inga.ae	lloy.ln	ucg.mi	ubsn.vx
csgn.vx	ndasek.sk	isp.mi	bbva.mc
cbk.xe	stan.ln	danske.ko	dnb.os
shba.sk	seba.sk	kbc.bt	sweda.sk
ebs.vi	bmps.mi	sab.mc	pop.mc
bir.db	bp.mi	aib.db	ete.at
poh1s.he	uni.mi	bcp.lb	bes.lb
mb.mi			

.ln=uk, .fr=france, .ae=netherlands, .db=ireland, .vx=switzerland,.lb=portugal, .xe=germany, .vi=austria,.ko=denmark,.mi=italy.

Group 2

hsba.ln	bnp.fr	dbk.xe	barc.ln
aca.fr	gle.fr	rbs.ln	san.mc
inga.ae	lloy.ln	ucg.mi	ubsn.vx
csgn.vx	ndasek.sk	isp.mi	bbva.mc
cbk.xe	stan.ln	danske.ko	dnb.os
shba.sk	seba.sk	kbc.bt	sweda.sk
ebs.vi	bmps.mi	sab.mc	pop.mc
bir.db	bp.mi	aib.db	ete.at
poh1s.he	uni.mi	bcp.lb	bes.lb
mb.mi			

.ln=uk, .fr=france, .ae=netherlands, .db=ireland, .vx=switzerland,.lb=portugal, .xe=germany, .vi=austria,.ko=denmark,.mi=italy.

Group 2: Europe (37)

Group 3

x8306.to	x8411.to	x8316.to	nab.au
cba.au	x600036.sh	anz.au	wbc.au
x600000.sh	sber.mz	x600016.sh	itub4.br
x8308.to	x8604.to	x8309.to	sbin.in
bbdc4.br	d05.sg	x000001.sz	x053000.se
dexb.bt	x055550.se	x600015.sh	u11.sg
x024110.se	maybank.ku	sbk.jo	x8354.to
x8332.to	x8331.to	cimb.ku	bankbaroda.in
isctr.is	x8377.to	x8355.to	x8418.to

.to=Japan .sh=China

.se=korea .ku=malaysia .mz=russia

Group 3

x8306.to	x8411.to	x8316.to	nab.au
cba.au	x600036.sh	anz.au	wbc.au
x600000.sh	sber.mz	x600016.sh	itub4.br
x8308.to	x8604.to	x8309.to	sbin.in
bbdc4.br	d05.sg	x000001.sz	x053000.se
dexb.bt	x055550.se	x600015.sh	u11.sg
x024110.se	maybank.ku	sbk.jo	x8354.to
x8332.to	x8331.to	cimb.ku	bankbaroda.in
isctr.is	x8377.to	x8355.to	x8418.to

.to=Japan .sh=China

.se=korea .ku=malaysia .mz=russia

Group 3: Asia, (36)

Group 3

x8306.to	x8411.to	x8316.to	nab.au
cba.au	x600036.sh	anz.au	wbc.au
x600000.sh	sber.mz	x600016.sh	itub4.br
x8308.to	x8604.to	x8309.to	sbin.in
bbdc4.br	d05.sg	x000001.sz	x053000.se
dexb.bt	x055550.se	x600015.sh	u11.sg
x024110.se	maybank.ku	sbk.jo	x8354.to
x8332.to	x8331.to	cimb.ku	bankbaroda.in
isctr.is	x8377.to	x8355.to	x8418.to

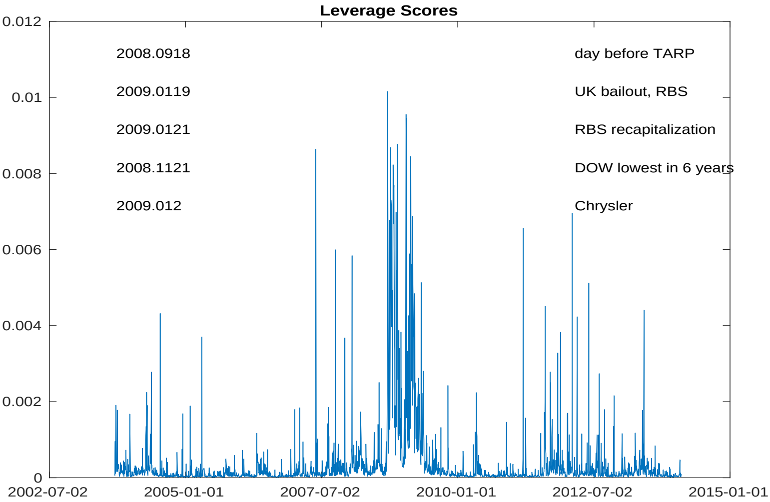
.to=Japan .sh=China

.se=korea .ku=malaysia .mz=russia

Group 3: Asia, (36)

A clear geographical component in the data

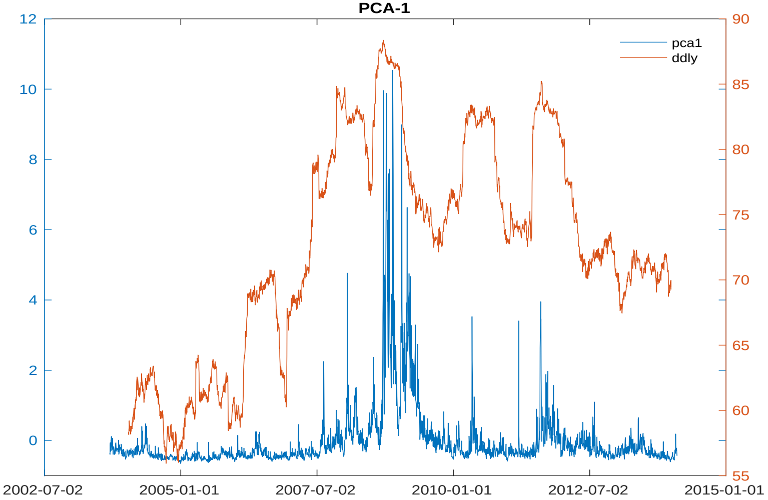
2. Row Leverage Scores



Influential Columns

	bank		Assets
1	bac	Bank of America (USA)	2416
2	gle.fr	Societe Generale (FR)	1703
3	stan.ln	Standard Chartered (UK)	674
4	wbc.au	Westpac Banking (Australia)	650
5	sber.mz	Sherbank Rossii (Russia)	552
6	X600016.sh	China Minsheng (China)	533
7	bmo.t	Bank of Montreal (Canada)	515
8	itub4.br	Itau Unibano (Brazil)	435
9	X8308.to	Resona Holdings (Japan)	434
10	shba.sk	Svensaka (Sweden)	388

3. Common factors, network effects, or both?



4. To/From/Connectedness: Really cool idea

Sparse reduced-form VAR in n variables

$$y_{it} = \sum_{k=1}^{p} \phi_{ik} y_{it-p} + \sum_{j\neq i}^{n} \sum_{k=1}^{p} \phi_{jk} y_{j,t-k} + u_{it} = \underbrace{\Phi(L)}_{sparse} \underbrace{y_{t-1}}_{n^2\times p} + u_{it}$$

SVAR: n variables, n structural shocks e

$$\underbrace{y_t}_{n\times 1} = \Psi(L)u_t = \Psi(L) \underbrace{H}_{n\times n} e_t$$

- Remark: name of shock= name of variable?
- Need $n(n-1)/2$ restrictions on H.

Regularization or Aggregation? $y_{it}^{-} = \sum_{j \neq i}^{n} y_{jt}$

Suggest n Reduced-form bivariate VAR:

$$\begin{aligned} y_{it} &= \sum_{k=1}^{p} \gamma_{ik} y_{it-k} + \sum_{k=1}^{p} \gamma_{ik} y_{it-k}^{-} + \epsilon_{it} \\ &= \sum_{k=1}^{p} \gamma_{ik} y_{it-k} + \underbrace{\Gamma(L)}_{dense} \underbrace{y_{it}^{-}}_{\text{scalar}} + \epsilon_{it} \end{aligned}$$

$$\underbrace{y_{it}}_{2\times 1} = \Psi_i(L) u_{it} = \Psi_i(L) H_i e_{it}$$

- Easier to justify $n(n-1)/2$ restrictions on H when $n = 2$.
- Two variables, 'to' and 'from' naturally defined.

Connectedness: To

horizon in days

1	10	20	banks	
0.13	0.28	0.34	rbs.ln	royal bank, scotland
0.11	0.26	0.32	c	citibank
0.10	0.25	0.31	bac	bank of america
0.14	0.26	0.30	barc.ln	barclays
0.08	0.22	0.29	sti.us	sun trust
0.09	0.23	0.28	ms	morgan stanley
0.15	0.24	0.28	gle.fr	societe generale
0.12	0.23	0.28	ubsn.vx	ubs (switzerland)
0.15	0.24	0.27	inga.ae	ing (netherlands)
0.08	0.21	0.27	wfc	wells fargo

Connectedness: From

horizon in days				
1	10	20	banks	
0.01	0.15	0.24	hsba.ln	hsbc (uk)
0.00	0.15	0.23	mqg.au	macquarie (aus)
0.01	0.16	0.23	dbk.xe	deutsche bank
0.01	0.15	0.23	ebs.vi	erste (austria)
0.01	0.15	0.23	csgn.vx	credit suisse
0.01	0.13	0.21	aca.fr	credit agricole
0.01	0.14	0.21	cbk.xe	commerzbank
0.02	0.13	0.21	dnb.os	dnb (norway)
0.00	0.12	0.20	ndasek.sk	nordea (sweden)
0.01	0.13	0.20	danske.ko	danske (denmark)

- Can we go beyond connectedness (descriptive) to structural modeling?