Big Data: Big N

V.C. 14.387 Note

December 2, 2014

Examples of Very Big Data

- Congressional record text, in 100 GBs
- Nielsen's scanner data, 5TBs
- Medicare claims data are in 100 TBs
- Facebook 200,000 TBs
- See **"Nuts and Bolts of Big Data", NBER lecture, by Gentzkow and Shapiro.** The non-econometric portion of our slides draws on theirs.

Map Reduce & Hadoop

The basic idea is that you need to divide work among the cluster of computers since you can't store and analyze the data on a single computer.
Simple but powerful algorithm framework. Released by Google around 2004; Hadoop is an open-source version.
Map-Reduce algorithm has the following steps:

1. Map: processes "chunks" of data to produce "summaries"
2. Reduce: combines "summaries" from different chunks to produce a single output file

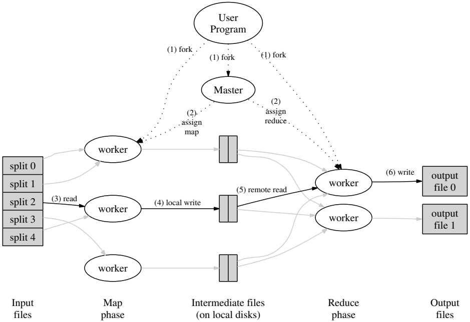
Examples

- Count words in docs i. Map: $i \mapsto$ set of $(word, count)$ pairs, C_i Reduce: Collapse $\{C_i\}$ by summing over $count$ within $word$.
- Hospital i. Map: $i \mapsto$ records H_i for patients who are 65+. Reduce: Append elements of $\{H_i\}$.

Map-Reduce Functionality

- Partitions data across machines
- Schedules execution across nodes
- Manages communication across machines
- Handles errors, machine failure

MapReduce: Model



Courtesy of Jeffrey Dean and Sanjay Ghemawat. Used with permission.

Amazon Web Services

- Data centers owned and run by Amazon. You can rent "virtual computers" minute-by-minute basis
- more than 80% of the cloud computing market
- nearly 3,000 employees
- cost per machine: 0.01 to 4.00 /hour
- Several services in AWS
- S3 (Storage)
- EC2 (Individual Machines)
- Elastic Map Reduce
- distribute the data for Hadoop clusters

Distributed and Recursive Computing of Estimators

We want to compute the least squares estimator

$$\hat{\beta} \in \arg\min_b n^{-1} \sum_{i=1}^{n} (y_i - x_i' b)^2.$$

The sample size n is very large and can't load the data into a single machine. What could we do if we have a single machine or many machines?

> Use the classical sufficiency ideas to distribute jobs across machines, spatially or in time.

The OLS Example

- We know that
$$\hat{\beta} = (X'X)^{-1}(X'Y).$$
- Hence we can do everything we want with just:
$$X'X, \quad X'Y, \quad n, \quad S_0,$$
where S_0 is a "small" random sample $(Y_i, X_i)_{i \in I_0}$ with sample size n_0, where n_0 is large, but small enough that the data can be loaded in the machine.
- We need $X'X$ and $X'Y$ to compute the estimators to compute the estimator.
- We need S_0 to compute robust standard errors and we need to know n to scale these standard errors appropriately.

The OLS Example Continued

- The terms like $X'X$ and $X'Y$ are sums that can be computed by distribution of jobs over many machines:

1. Suppose machine j stores sample $S_j = (X_i, Y_i)_{i \in I_j}$ of size n_j.
2. Then we can map S_j to the sufficient statistics

$$T_j = \Big(\sum_{i \in I_j} X_i X_i', \sum_{i \in I_j} X_i Y_i, n_j \Big)$$

   for each j.
3. We then collect $(T_j)_{j=1}^{M}$ and reduce them further to

$$T = \sum_{j=1}^{M} T_j = (X'X, X'Y, n).$$

The LASSO Example

The Lasso estimator minimizes

$$(Y - X\beta)'(Y - X\beta) + \lambda\|\Psi\beta\|_1, \quad \Psi = \mathrm{diag}(X'X)$$

or equivalently

$$Y'Y - 2\beta'X'Y + \beta'X'X\beta + \lambda\|\Psi\beta\|_1.$$

Hence in order to compute Lasso and estimate noise level to tune λ we only need to know

$$Y'X, \quad X'X, \quad n, \quad S_0.$$

Computation of sums could be distributed across machines.

The Two Stage Least Squares

The estimator takes the form

$$(X'P_ZX)^{-1}X'P_ZY = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y.$$

Thus we only need to know

$$Z'Z, \quad X'Z, \quad Z'Y, \quad n, \quad S_0.$$

Computation of sums could be distributed across machines.

Digression: Ideas of Sufficiency are Extremely Useful in Other Contexts

- Motivated by J. Angrist, *Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records*, AER, 1990.
- We have a small sample $S_0 = (Z_i, Y_i)_{i \in I_0}$, where Z_i are instruments (that also include exogenous covariates) and Y_i are earnings. In ML speak, this is called "labelled data" (they call Y_i labels, how uncool)
- We also have huge ($n \gg n_0$) samples of unlabeled data (no Y_i recorded) from which we can obtain $Z'X$, $X'X$, $Z'Z$ via distributed computing (if needed).
- We can compute the final 2SLS-like estimator as

$$\frac{n}{n_0} \cdot (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}\sum_{i=1}^{n_0} Z_i Y_i$$

Can compute standard errors using S_0.

Exponential Families and Non-Linear Examples

Consider estimation using MLE based upon exponential families. Here assume data $W_i \sim f_\theta$, where

$$f_\theta(w) = \exp(T(w)'\theta + \varphi(\theta)).$$

Then the MLE maximizes

$$\sum_{i=1}^{n} log f_\theta(W_i) = \sum_{i=1}^{n} T(W_i)'\theta + \varphi(\theta) =: T'\theta + n\varphi(\theta).$$

The sufficient statistic T can be obtained via distributed computing. We also need an S_0 to obtain standard errors. Going beyond such quasi-linear examples could be difficult, but possible.

M- and GMM - Estimation

The ideas could be pushed forward using 1-step or approximate minimization principles. Here is a very crude form of one possible approach.

Suppose that $\hat{\theta}$ minimizes

$$\sum_{i=1}^{n} m(W_i, \theta).$$

Then given an initial estimator $\hat{\theta}_0$ computed on S_0 we could do Newton iterations to approximate $\hat{\theta}$:

$$\hat{\theta}_{j+1} = \hat{\theta}_j - \left(\sum_{i=1}^{n} \nabla_\theta^2 m(W_i, \hat{\theta}_j) \right)^{-1} \sum_{i=1}^{n} \nabla_\theta m(W_i, \hat{\theta}_j).$$

Each iteration involves sufficient statistics

$$\sum_{i=1}^{n} \nabla_\theta^2 m(W_i, \hat{\theta}_j), \quad \sum_{i=1}^{n} \nabla_\theta m(W_i, \hat{\theta}_j)$$

which can obtained via distributed computing.

Conclusions

- We discussed the large p case, which is difficult. Approximate sparsity was used as a generalization of the usual parsimonius approach used in empirical work.
- A *sea of opportunities* for exciting empirical and theoretical work.
- We discussed the large n case, which is less difficult. Here the key is the distributed computing. Also big n samples often come in "unlabeled" form, so you need to be creative in order to make good use of them.
- This is an *ocean of opportunities*.

14.387 Applied Econometrics: Mostly Harmless Big Data
Fall 2014