Models for Millions

Bob Stine
Department of Statistics
The Wharton School, University of Pennsylvania
stat.wharton.upenn.edu/~stine

34th NJ ASA Spring Symposium
June 7, 2013

Introduction

Statistics in the News

- Hot topics
  - Big Data
  - Business Analytics
  - Data Science

CIO Journal.

CIO Report | Consumerization | Big Data | Cloud | Talent & Management | Security

April 10, 2013, 2:59 PM ET

Like It or Not, You're in the Data Business

- Are the authors talking about statistics?
  - Or about …
    - information systems?
    - database technology?
    - visualization, eye candy?

JOURNAL REPORTS | Updated March 8, 2013, 12:49 p.m. ET

Help Wanted!

Data, data everywhere—and not enough people to decipher it

Data Science: The Numbers of Our Lives

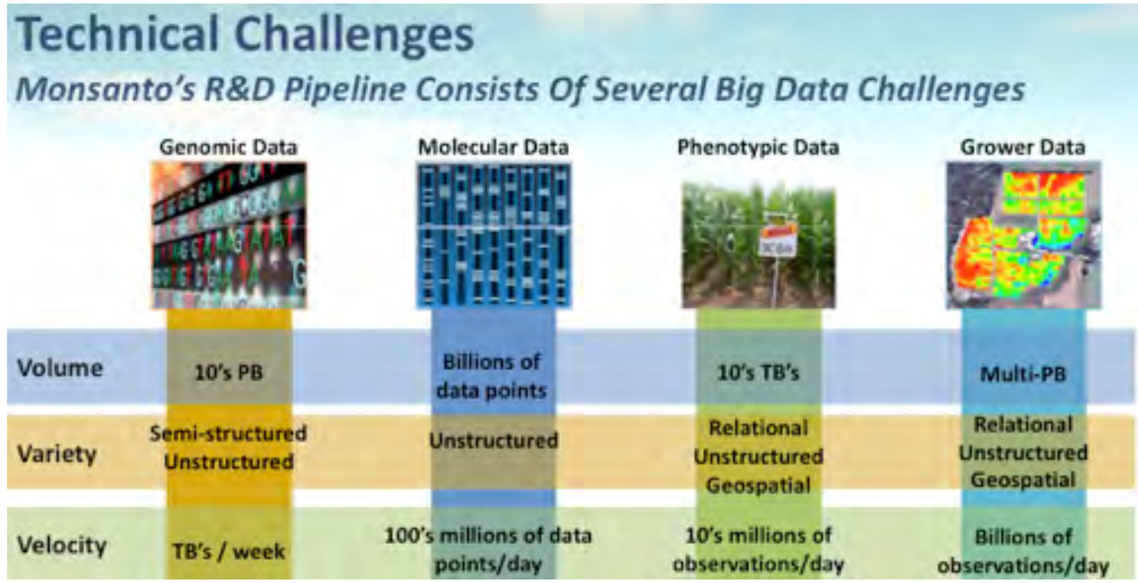
By CLAIRE CAIN MILLER
Published: April 11, 2013

HARVARD BUSINESS REVIEW calls data science "the sexiest job in the 21st century," and by most accounts this hot new field promises to revolutionize industries from business to government, health care to academia.

Even Farming…

How B.I. and Data Make a More Efficient Farm

by David Strom | September 17, 2012



Technical Challenges

Monsanto's R&D Pipeline Consists Of Several Big Data Challenges

	Genomic Data	Molecular Data	Phenotypic Data	Grower Data
Volume	10's PB	Billions of data points	10's TB's	Multi-PB
Variety	Semi-structured Unstructured	Unstructured	Relational Unstructured Geospatial	Relational Unstructured Geospatial
Velocity	TB's / week	100's millions of data points/day	10's millions of observations/day	Billions of observations/day

ALPRO™ – Milking

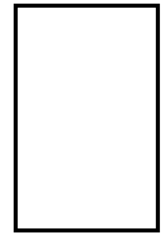
From milking point control and monitoring milking performance, to checking that your milking protocols are adhered to, ALPRO delivers the information you need to fine-tune your milk production.

Business intelligence: it's not just for big-city businesses anymore.

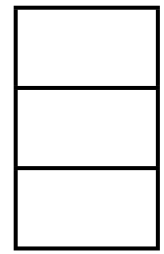
Big Data

Notation
n = # rows of X
p = #columns of X

- Recent modeling projects

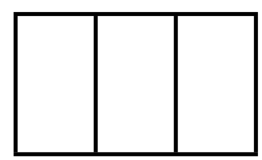


- Credit scoring
  - 75,000 cases
  - 15,000+ possible explanatory variables
- Spatial time series
  - 3,000 locations
  - 100 time points
  - 20+ features at each location and time



- Text
  - Real estate listings
    - 6,000 prices, millions of possible descriptions

  - Tagging
    - 1.2 million words, 60,000+ 'explanatory variables'

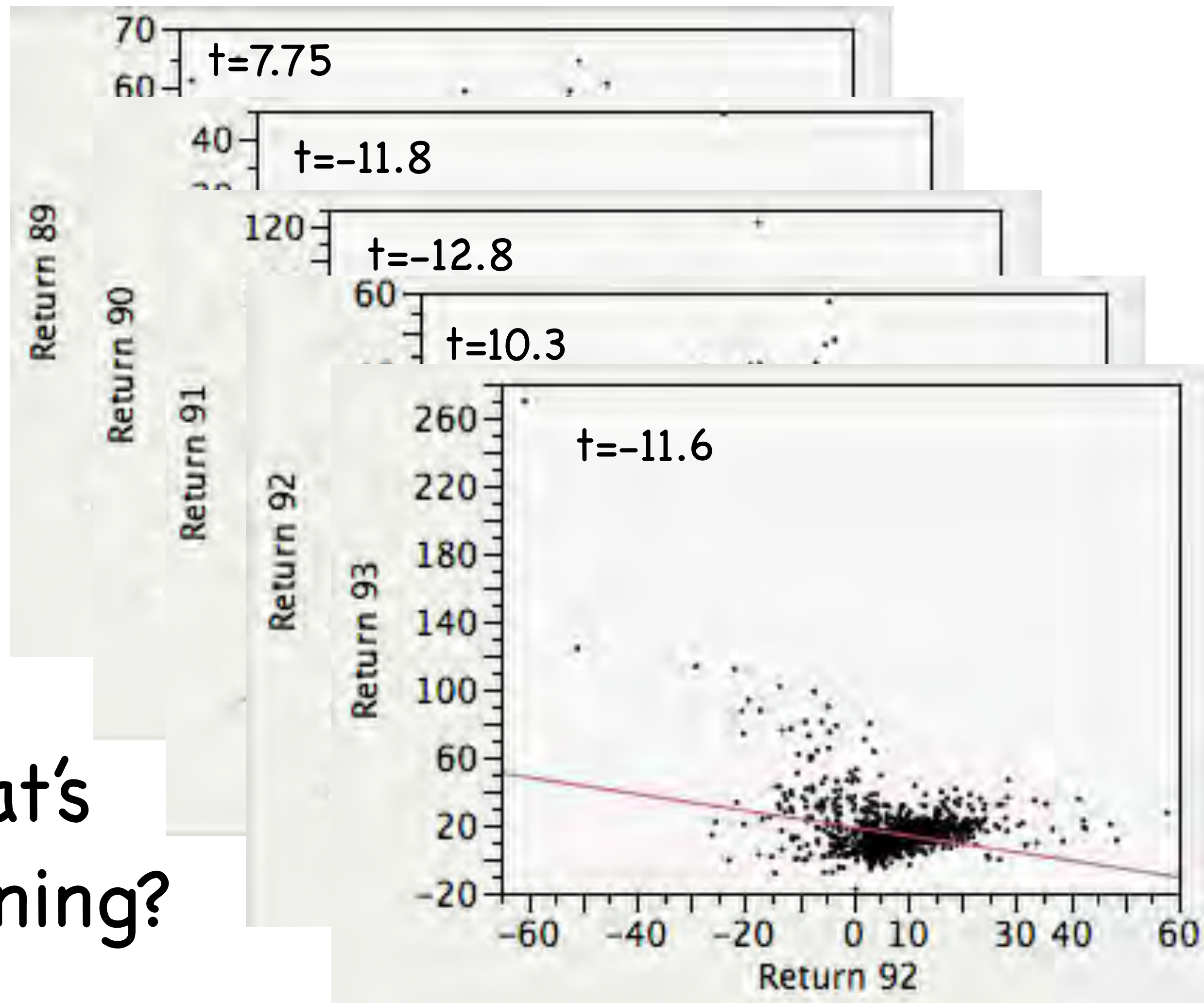


Is Big Data Really So Big?

- Not always so large as they may seem
  - Repeated measurement ≠ more degrees of freedom
  - What is the relevant source of variation?
- Transfer learning problem
  - Machine learning
  - Build model for structure of text on corpus such as the New York Times
  - What transfers from that model to
    Washington Post?
    Richmond Times-Dispatch?
- Implications for estimates of standard error

Example of Dependence

- Predict returns on mutual funds
  - Do funds that do well in one year anticipate doing well (or poorly) the next year?



What's happening?

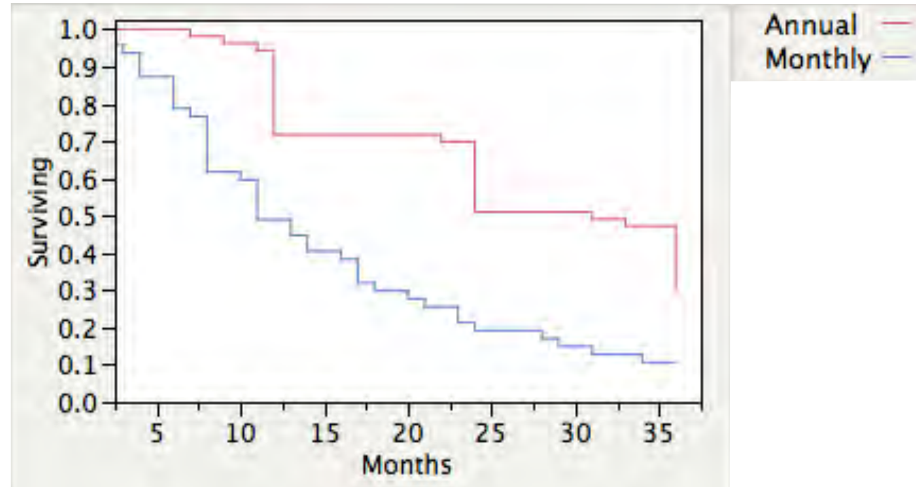
Does Big Data Imply Big Models?

- Perhaps all one needs is a very simple analysis
  - Google
  - Massive hardware
  - Extensive data
- Text modeling
  - Hard problem: predict next word in sentence

    I took a walk _____
  - Tabulation of all 5-grams (5 word sequence)
  - Replace modeling with frequency table
- Web page design
  - Continuous experimentation
  - Randomized, two-sample t-test

Simple Models Can Be Better

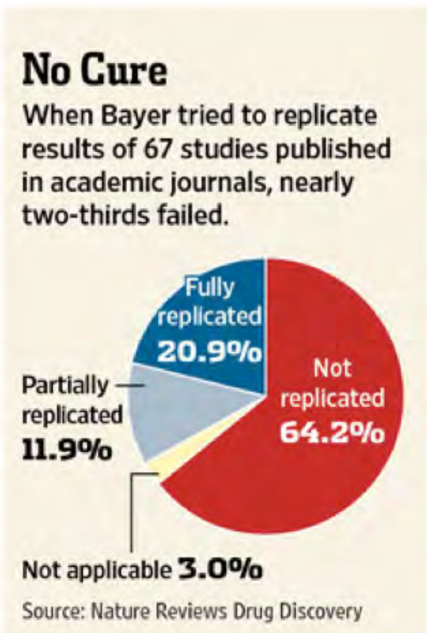
- Association rules
  - Low tech...
    Build tables
  - Identify association
  - Low-tech ≠ low impact...
    grab low-hanging fruit

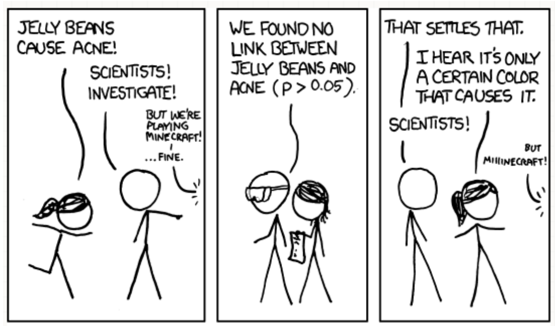


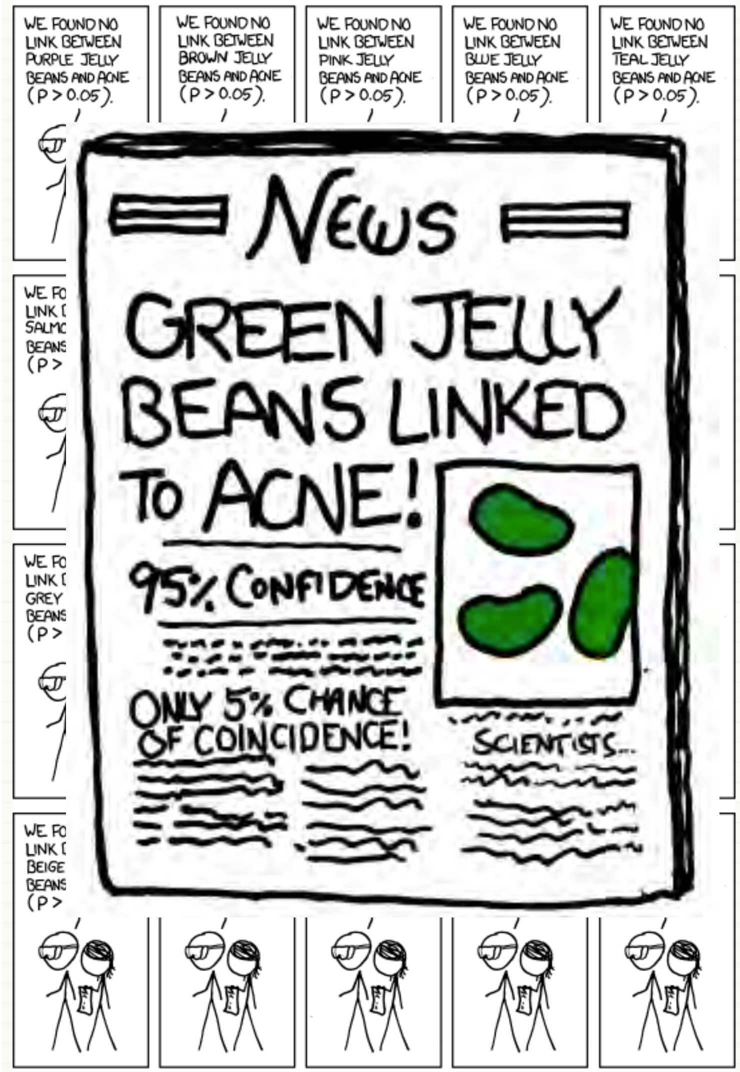
- Predictive modeling via support vector machine
  - High tech...
    Locate separating hyperplanes in kernel space
  - Identify predictive features
  - High-tech ≠ high impact...
    Complexity vs communication

Simple might be right!

- Recent WSJ story on reproducibility and proliferation of research...







XKCD

Attractive Misconceptions*

- Thinking the true predictor is in my data rather than running an experiment
  - Reject inference and white cars
  - Training: we give students the data
- Outliers don't matter with millions of cases
  - Central limit theorem
  - Corollary: estimators are normally distributed.
- Methods are black boxes
  - Lasso is popular, so it's best for my application.
- Cross-validation keeps me out of trouble
  - As long as the model validates well out-of-sample, the predictions are reliable.

*ie, Lessons I have learned the hard way.

Plan

- Familiar context
  - Fit LS regression of continuous Y to large collection of possible explanatory variables
- Two themes
  - Reducing dimensions
    - Columns: Random projections
    - Row: Subsampling
  - Streaming
    - Sequential from rows
    - Sequential from columns
  - Mixtures of the two (VIF regression)
- Comments
  - Regularization (shrinkage) can be added
  - Where are the Bayesian models?

Dimension Reduction

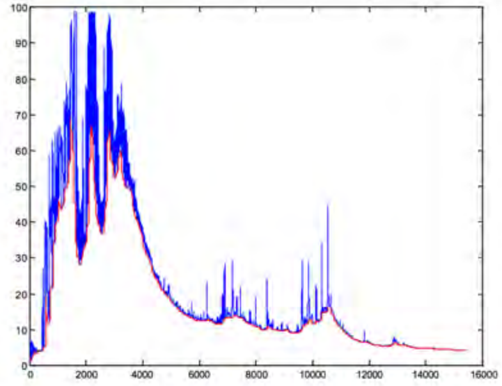
Reducing Columns

- Context
  - PCA, common column scales
  - Huge $p >> n$
- Random projection
  - Methods based on random projection have revived interest in PCA
- Idea
  - Use random projections to reduce the data matrix to a size amenable to calculation.
  - Explanatory variables in $n \times p$ matrix X
  - Pick $d << p$
  - Multiply X by a $p \times d$ matrix of random numbers Ω so that resulting dimension is $n \times d$.

Arcene Example

- Automation
  - Automated data collection produces extensive measurements, here p=10,000 features
  - Only n=200 cases
- Arcene example from UCI
  - Mass spectrometer measurements
  - Origin: Separate normal cells from cancerous cells
  - Make into a regression problem
    - Use continuous response, not the 0/1 indicator in respository

- Complications galore…
  - Collinear: sampling smooth function
  - Too many 'perfect' solutions
  - Hard to test out-of-sample because so few cases



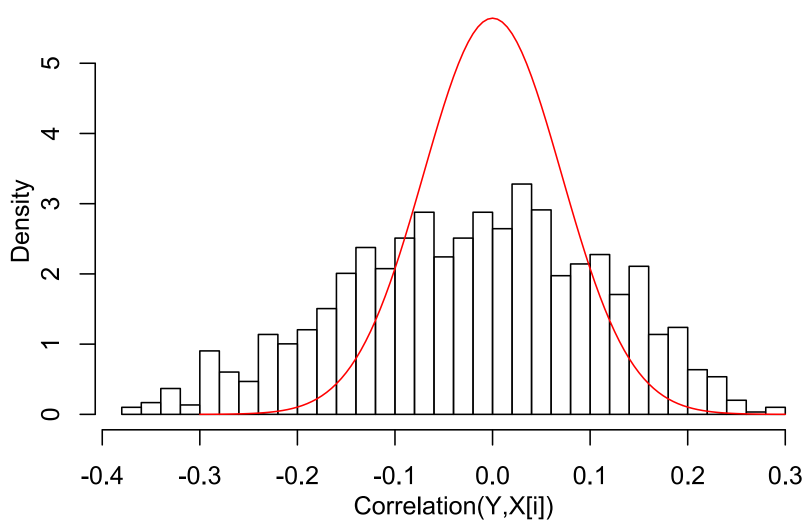
Wharton
Department of Statistics

UCI = Univ of Ca Irving ML databases, http://archive.ics.uci.edu

Marginal Analysis

- Marginal correlations (X_i, Y) show signal
  - Deviate from distribution of random noise (red)
- But: weakly spread over many coordinates
  - Multiple regression finds weak effects
  - $R^2 = 0.19$ is larger than might expect

Null: Expect p/n
$R^2 = 10/200 = 0.05$



	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.059631	1.301089	0.814	0.41643	
x1	-0.004818	0.005775	-0.834	0.40512	
x2	-0.007273	0.003887	-1.871	0.06288	.
x3	0.004149	0.003295	1.259	0.20954	
x4	-0.003342	0.001279	-2.614	0.00967	**
x5	-0.007191	0.006658	-1.080	0.28153	
x6	0.002474	0.002162	1.144	0.25401	
x7	-0.001173	0.001457	-0.805	0.42188	
x8	0.001113	0.009964	0.112	0.91116	
x9	-0.008695	0.004328	-2.009	0.04599	*
x10	0.000841	0.002593	0.324	0.74604	

$R^2 = 0.19$

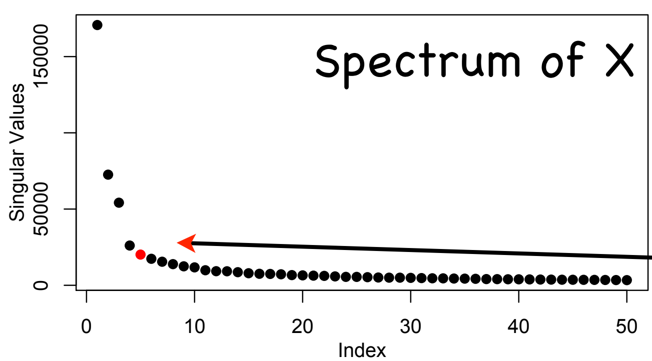
Wharton
Department of Statistics

PCA Analysis

- Compute singular value decomposition

$$X = U\,D\,V'$$

  - Columns of U, V are orthonormal
  - D is a diagonal matrix of singular values (spectrum of X)
- Doable in R if X is 200×10,000 matrix
  - Regression finds clear, strong effect in U_5



	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.4084	2.2091	-0.638	0.5245	
U1	-20.4379	31.1613	-0.656	0.5127	
U2	-6.4944	2.5894	-2.508	0.0130	*
U3	0.2883	1.9062	0.151	0.8799	
U4	-1.8998	2.3440	-0.810	0.4186	
U5	14.9618	1.9141	7.817	3.36e-13	***

R^2=0.27

Random Projection

- Project down to smaller size
  - Example with d=100
  - Compare random projections to exact from R
- Procedure
  - $P_0 = X\,\Omega$, Ω is $10{,}000 \times d$ random matrix
  - $P_1 = XX'\,P_0$ is one step of power method
  - Take first few columns of U from SVD of P_j
- Compare to fit with exact SVD

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0547	0.1345	0.407	0.68474	
U1.1	-5.2929	1.9025	-2.782	0.00593	**
U1.2	-0.3964	1.9025	-0.208	0.83516	
U1.3	0.2356	1.9025	0.124	0.90157	
U1.4	-15.1852	1.9025	-7.982	1.23e-13	***
U1.5	0.1092	1.9025	0.057	0.95428	

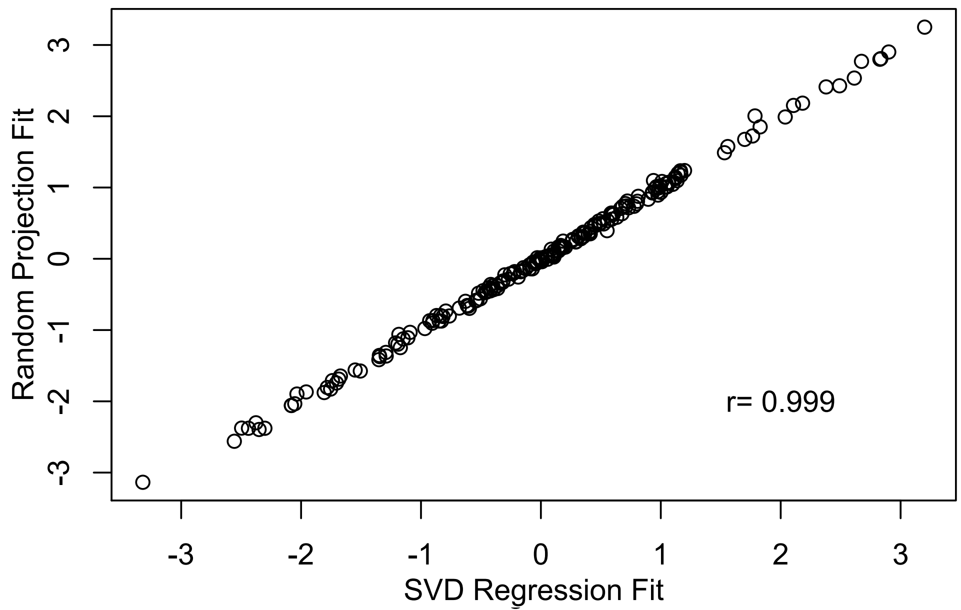
Random Projection
one iteration

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.4084	2.2091	-0.638	0.5245	
U1	-20.4379	31.1613	-0.656	0.5127	
U2	-6.4944	2.5894	-2.508	0.0130	*
U3	0.2883	1.9062	0.151	0.8799	
U4	-1.8998	2.3440	-0.810	0.4186	
U5	14.9618	1.9141	7.817	3.36e-13	***

Exact

Comparison of Fits

- Reconstruction
  - Random projection preserves subspace holding range of matrix, but not necessarily in the same coordinates.
  - Eg: different components appear in regression
- Comparison of fits shows same subspace



After one iteration

A really big X matrix?

- Arcene example is 'small': we can do do the exact SVD quickly in R.
- Suppose X had more columns, say

$$10{,}000^2 = 100{,}000{,}000$$

such as from the interaction space of X.

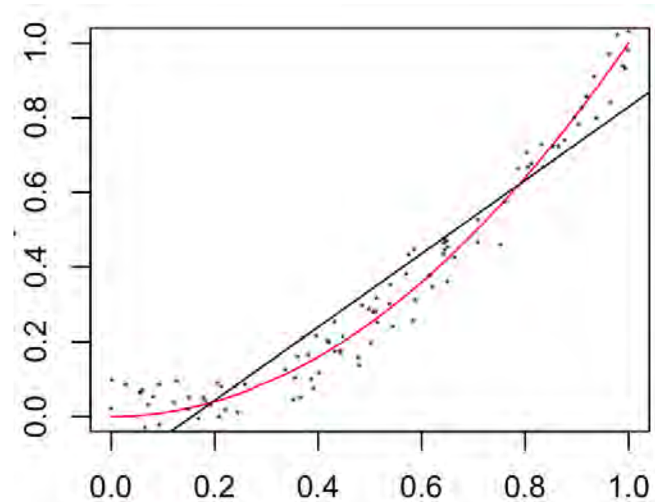
Okay, half that

- Linear models often approximate non-linear structure...

first 10 PCs of X

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.3808	2.2385	-1.064	0.288890	
U.1	-23.0907	31.5477	-0.732	0.465118	
U.2	-4.8377	2.2556	-2.145	0.033250	*
U.3	5.0956	1.3908	3.664	0.000322	***
U.4	0.1807	1.9608	0.092	0.926659	
U.5	-2.2772	1.4018	-1.625	0.105935	
U.6	0.3524	1.4127	0.249	0.803301	
U.7	5.3127	1.3850	3.836	0.000170	***
U.8	4.8648	1.6621	2.927	0.003844	**
U.9	-1.7501	1.4451	-1.211	0.227389	
U.10	-0.8872	1.6176	-0.548	0.584012	

R^2=0.23



Random Projection

- Random projection with 50,000,000 explanatory variables (X_j X_k)
  - Cannot compare to the exact solution for this one
  - Runs 'quickly': about 5 minutes on laptop!
- Fitted model on 5 elements of the random projection of the quadratic X's

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.14661	0.12277	1.194	0.2339	
Qq1.1	20.64135	1.50690	13.698	<2e-16	***
Qq1.2	0.47731	1.25891	0.379	0.7050	
Qq1.3	-1.90151	1.08258	-1.756	0.0806	.
Qq1.4	0.71912	1.03463	0.695	0.4879	
Qq1.5	-0.07981	1.03956	-0.077	0.9389	

One power iteration

R^2=0.23 → R^2=0.46 → R^2=0.57

Wharton
Department of Statistics

Postscript...

- What's the response in that regression?
  - What Y variable lives in the quadratic space?
- Short answer: Kernel trick
  - Compute the quadratic kernel of the data
  - Find the SVD
  - Let Y be one of the singular vectors
- Story for another day ...

Reducing Rows

- Context
  - Very large n >> moderate p
  - Again, less interested in selecting specific Xs
- Common sense
  - Don't need to fit a model more precisely than needed for statistical precision/selection.
  - However…
    More data reveals a more interesting model, one with subtle effects
- Speed of OLS
  - $b = (X'X)^{-1}X'Y$
  - Slow part if n >> p is computing $X'X$ $O(np^2)$

Case Sampling

Not sampling on the response!

- Exploit familiar property of regression
  - Precision of slope is maximized by finding cases with large variation in Xs
  - Task becomes finding cases with high leverage
- Machine learning has developed methods to seek high-leverage points
  - Hard to find sequentially
- Simple improvement
  - Sample m << n cases to estimate X'X
  - Use all n cases to estimate X'Y

$b=(X'X)^{-1}X'Y$

- Leverage points however may not be your friends in modeling large data sets...

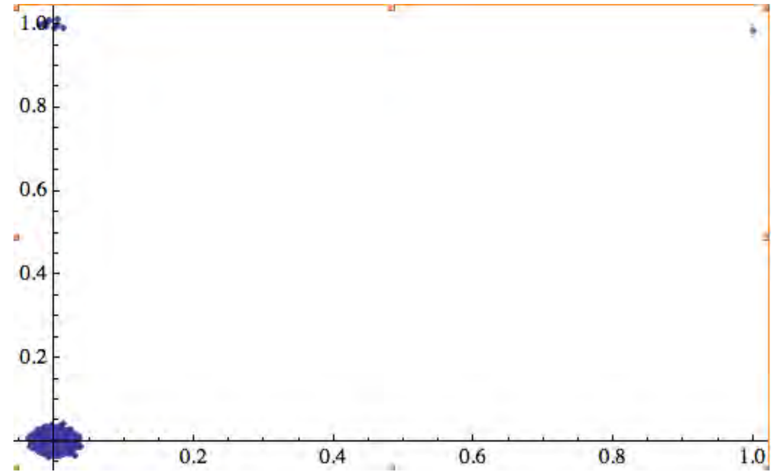
Misconceptions

Outliers in Big Data

- Sparse data
  - n=10,000
  - X ≈ 0

    Y=0 for 9,990, Y=1 for 10
  - X ≈ 1

    Y=1 for one case



- What's the appropriate p-value?
- Classical OLS
  - Use residual after fit slope, as if right model
  - t ≈ 10, pick your level of significance!
- Common sense
  - p = 1/1000 more sensible p-value

Streaming Methods

Cases

Variables

Combined

Streaming Cases

- Context
  - Huge number of cases, more than memory holds
- Idea
  - Compute estimates as read in data so do not have to store all data
  - Calculations can be split over network
- Different take on OLS
  - OLS estimate for n-1 cases

$$b_{n-1} = (X'X)^{-1}X'Y$$

  - The estimate for n cases is

$$\begin{aligned} b_n &= b_{n-1} + (X'X)^{-1}x_n(y_n - x_n'b_{n-1})/(1+h_n) \\ &= b_{n-1} + [(1+h_n)(X'X)]^{-1}\, x_n\, e \end{aligned}$$

where the leverage $h_n = x_n'(X'X)^{-1}x_n$. slow step

Wharton
Department of Statistics

Stochastic Gradient

- Build up normal equations and solutions by randomly sampling cases
- Stochastic gradient
  - Robbins & Monro
  - To minimize $(y_i - x_i'b)^2$ w.r.t. b, step in the direction of the negative gradient,

$$x_i(y_i - x_i'b) = x_i \; e_i$$

- Full least squares solution uses X'X

$$b_n = b_{n-1} + [(1+h_n)(X'X)]^{-1} \; x_n \; e$$

- Pretend X'X is diagonal, and life moves faster

$$b^*_n \;\; = b^*_{n-1} + \delta_n \; D^{-1} \; x_n \; e^*$$

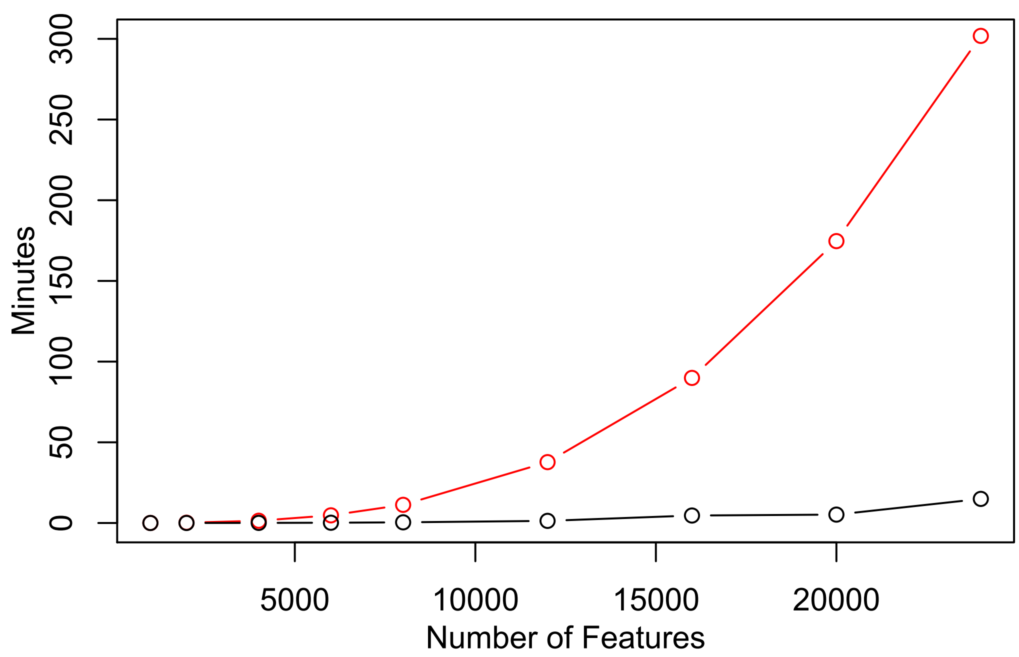
with D = diagonal (X'X) and δ_n is a learning rate.

How fast is it?

- Goal in stochastic gradient is to run as fast as you can read data!

n	p	OLS	SG
2,500	500	<0.1	<0.1
5,000	1,000	0.7	0.2
10,000	2,000	9.5	0.9
20,000	4,000	84	3.7
40,000	8,000	675	25
80,000	16,000	5394	276
100,000	20,000	10480	312

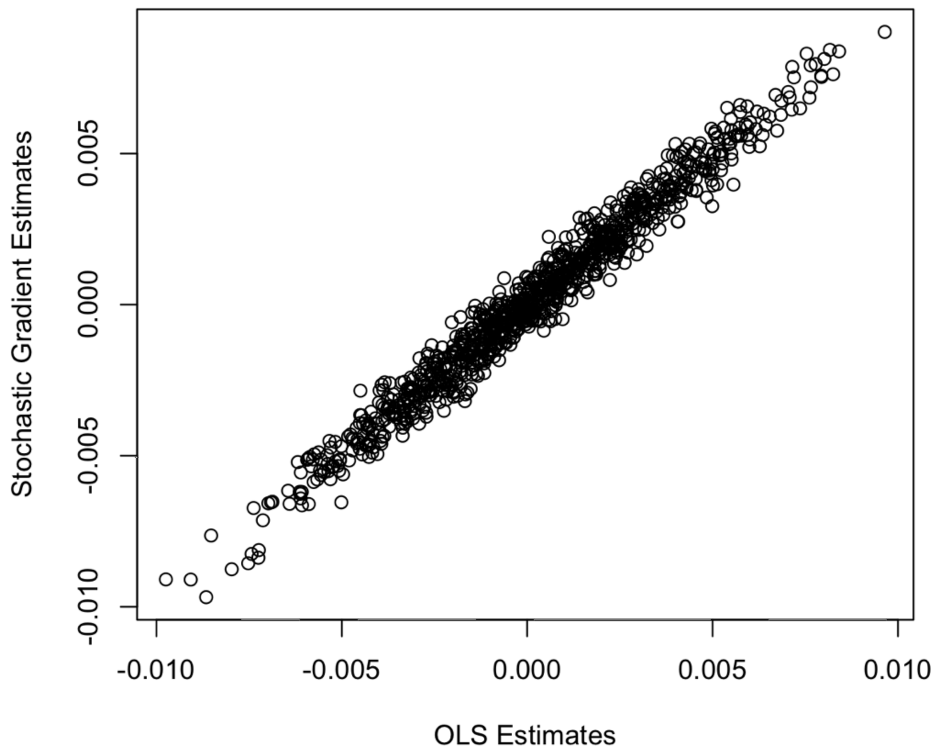
n=5p seconds



How good are estimates?

- Graph plots estimated coefficients from one-pass of stochastic gradient versus exact OLS
- Deviation from OLS below standard error
  - Small error relative to variation in estimates

p=1000



At least when there is not much collinearity!

Statistical Significance?

- Don't have X'X so don't have usual SE
  - How to evaluate modeling?
- Cross-validation
  - Less sensitive to modeling assumptions
  - Split data

    Training data: Fit model on part of the data

    Test data: Reserved data
  - Compare fit in two datasets
- Three way split becoming necessary
  - Training data
  - Tuning data…
    - Set tuning parameters, such as level of shrinkage
  - Testing data

Misconceptions

Population Drift

- Cross-validation is an optimistic assessment
  - One of few places when have random sample
- Credit scoring
  - Predict performance of applicants
  - Cross-validation shows model spot on
- Data collection is a long process
  - Gather data over 1-2 years
  - Takes 1-2 more years to find the response
- The world changed!
  - Booming economy during data collection
  - Collapsing recession when implemented
  - No way CV could see this problem

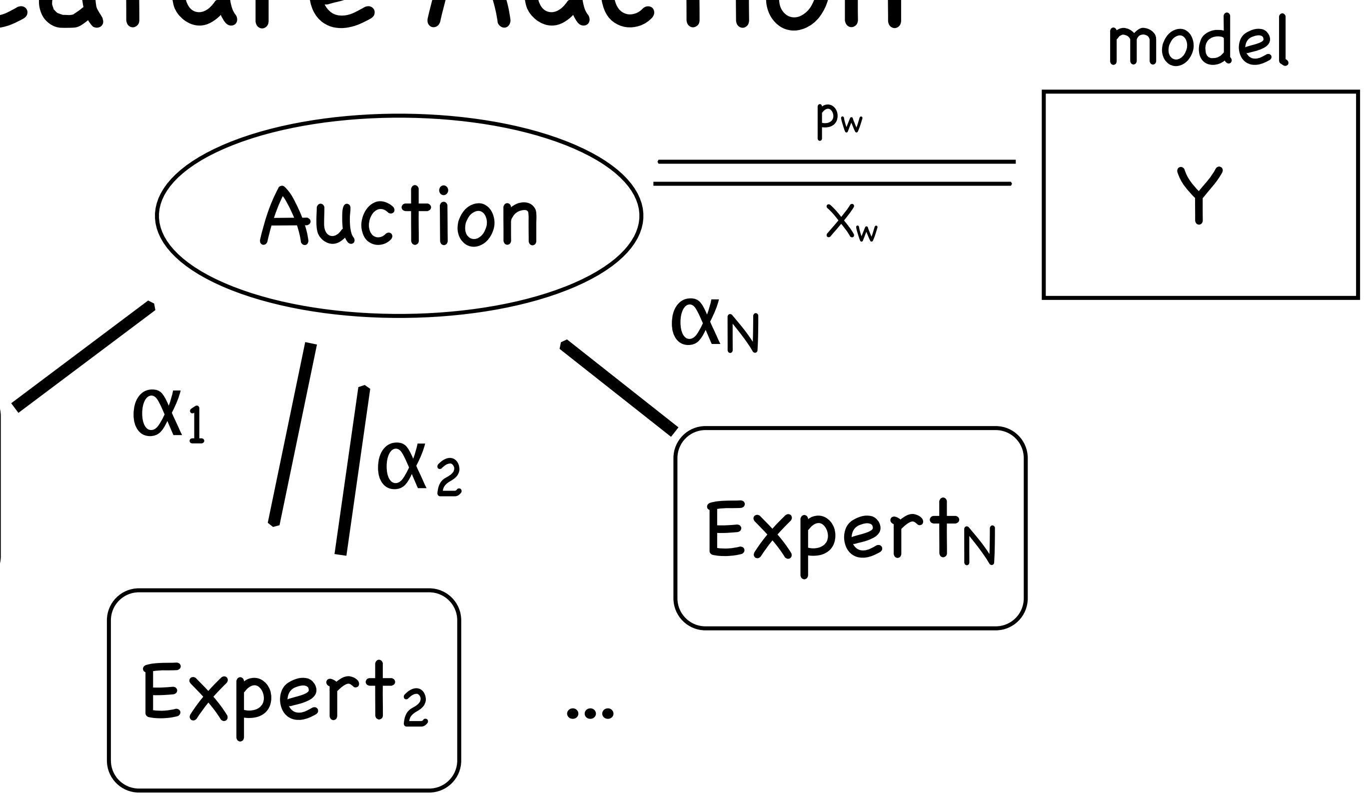
More issues …
Variation?
How to allocate?

Streaming Variables

- Context
  - Huge number of variables
  - Want to preserve scales
- Idea
  - Stepwise search pays a large cost for searching
    Bonferroni p-value threshold 0.05/millions
  - Streaming: Examine features one at a time
    - Resembles forward stepwise, but
      without sorting/ordering based on p-values
- Exploit context
  - "Scientist" orders variables, defines search strategy
  - Adaptive:Build interactions as features added

Feature Auction

Collection of experts bid for the opportunity to recommend feature



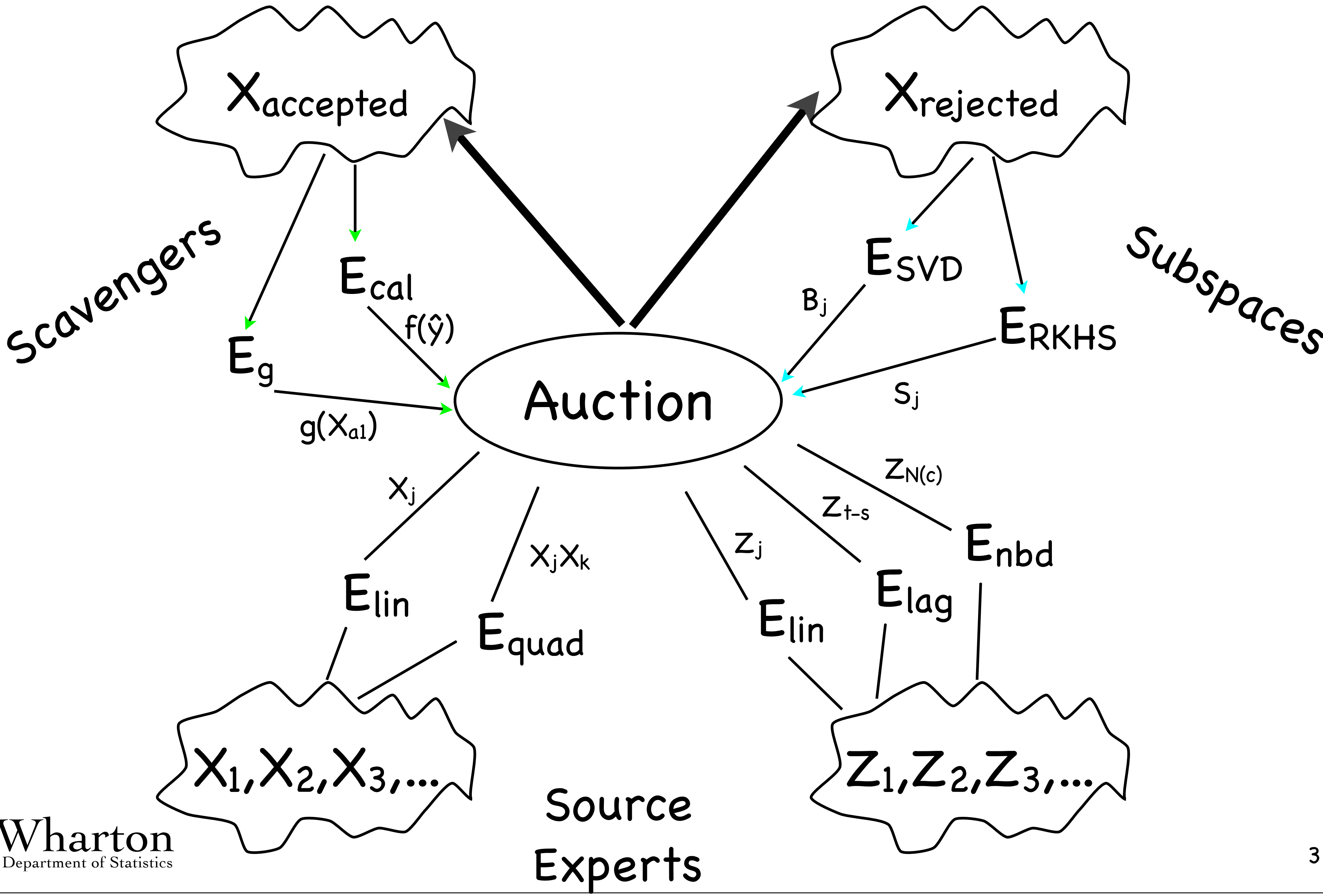
Auction collects winning bid α_2

Expert supplies values of recommended feature X_w

Expert receives payoff ω if $p_w \le \alpha_2$

Experts only learn if the bid was accepted, not the value of b or the p-value.

Experts


$X_{accepted}$
$X_{rejected}$
Scavengers
Subspaces
E_{cal}
E_g
$f(\hat{y})$
$g(X_{a1})$
E_{SVD}
E_{RKHS}
B_j
S_j
Auction
X_j
$X_j X_k$
E_{lin}
E_{quad}
Z_j
Z_{t-s}
$Z_{N(c)}$
E_{lin}
E_{lag}
E_{nbd}
$X_1, X_2, X_3, \ldots$
$Z_1, Z_2, Z_3, \ldots$
Source
Experts

Experts

- Expert
  Strategy for creating list of features. Experts embody domain knowledge, science of application.
- Source experts
  - A collection of measurements (eg, synonyms, clusters)
  - Components of a subspace basis (PCA, RKHS)
  - Lags of a time series
- Scavenger experts
  - Interactions
    - among features accepted into model
    - among features rejected by model
    - between those accepted with those rejected
  - Transformations
    - segmenting, as in scatterplot smoothing
    - polynomial transformations

Winning Experts

- Expert is rewarded if correct
  - Experts have alpha-wealth
  - If recommended feature is accepted in the model, expert earns ω additional wealth
  - If recommended feature is refused, expert loses bid
- As auction proceeds, it...
  - Rewards experts that offer useful features.
  - Eliminates experts whose features are not accepted.
  - Taxes fund scavenger experts
    - Ensure that continue to control overall FDR
- Critical
  - Adjust for multiplicity
  - p-values determine useful features

Robust Standard Errors

- p-values are critical, but…
  - Error structure often heteroscedastic
  - Observations frequently dependent
- Dependence
  - "Observations"
    - Spatial time series at multiple locations
    - Documents from various news feeds
  - Transfer learning problem
- Examples
  - Use sandwich-type estimate of standard error

heteroscedasticity

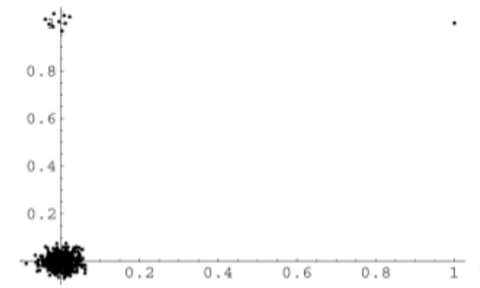
$$\text{var}(b) = (X'X)^{-1}X'E(ee')X(X'X)^{-1}$$
$$= (X'X)^{-1}\, X'D^2X\, (X'X)^{-1}$$

dependence

$$\text{var}(b) = (X'X)^{-1}X'E(ee')X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}\, X'BX\, (X'X)^{-1}$$

Flashback…

- Heteroscedastic error
  - Estimate standard error with outlier
  - Sandwich estimator allowing heteroscedastic error variances gives a t-stat ≈ 1, not 10.



- Dependent error
  - Even more important need for accurate SE
  - Netflix example
    Bonferroni (or hard thresholding) overfits due to dependence in responses.
  - Spatial modeling
    Everything seems significant unless incorporate dependence into the calculation of the SE

Control for Over Fitting

- Alpha investing
  - Test possibly infinite sequence of m hypotheses

    $H_1, H_2, H_3, \ldots H_m \ldots$

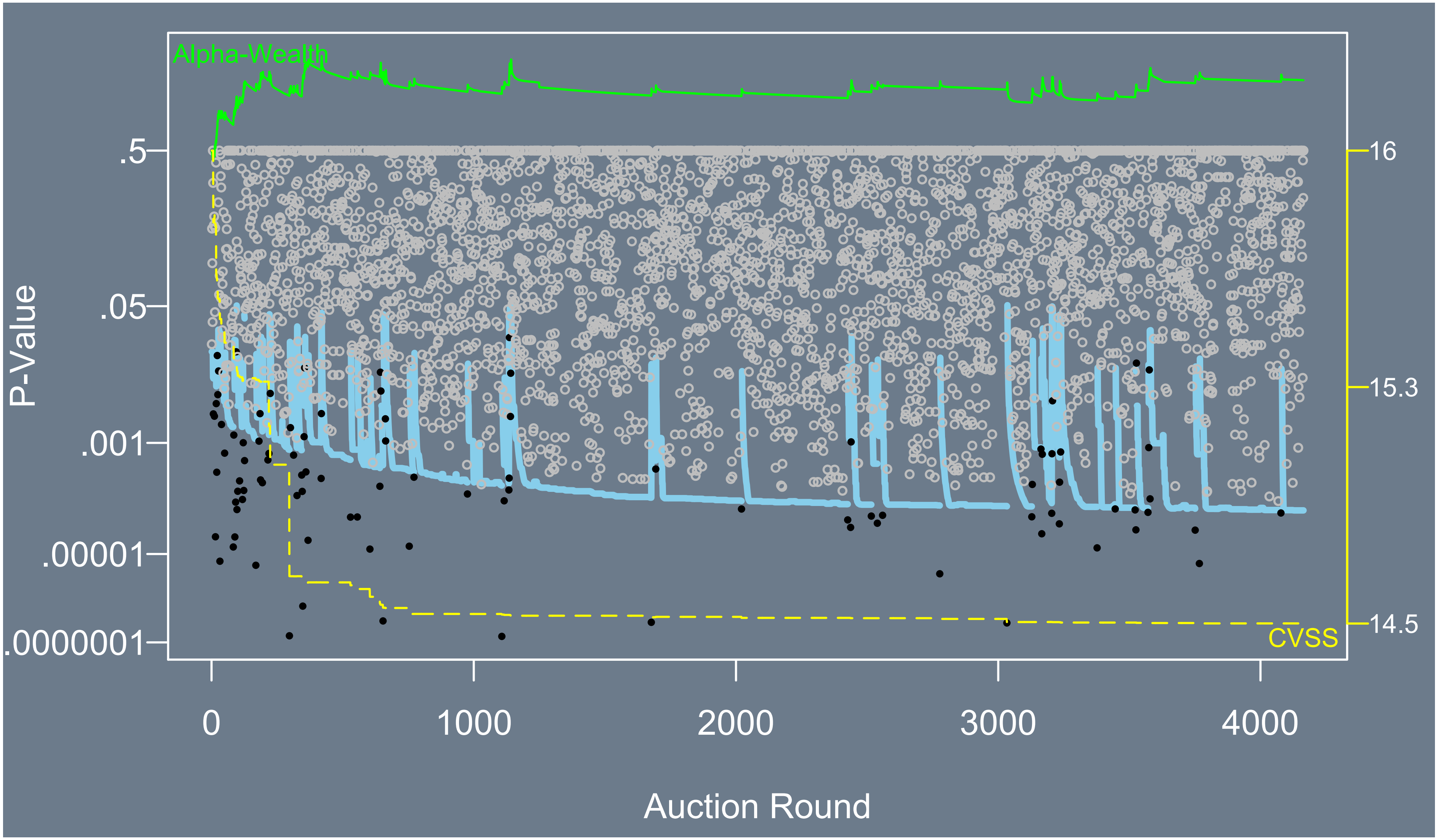
    obtaining the p-values $p_1, p_2, \ldots$
- Procedure
  - Start with an initial alpha wealth W_0
  - Invest wealth $0 \le \alpha_j \le W_j$ in the test of Hj
  - Change in wealth depends on test outcome

    If reject, wealth goes up by payout $\omega - \alpha_j$

    If don't reject, wealth goes down by α_j
- Properties
  - Controls expected false discovery rate
  - Can reproduce Bonferroni or FDR methods

Auction Run

First 4,000 rounds of auction modeling.



Wharton
Department of Statistics

Streaming Cases & Variables

- Background
  - A variance inflation factor (VIF) is a diagnostic for collinearity in regression
- VIF compares variances of slope estimates
  - Variance of b_k were it uncorrelated with others

$$var(b_x) = s^2/(x_k'x_k)$$

  - Actual variance is larger due to collinearity

$$var(b_k) \approx VIF_k\ s^2/(x_k'x_k)$$

  where $1 \le VIF_k = 1/(1-R^2_{k|rest})$

- Handy interpretation
  - Is x_k not significant because

    It is not useful?

    Redundant?

Wharton
Department of Statistics

VIF Regression

- Idea
  - Speed up the slow step in forward stepwise
- Usual selection
  - Has variables X and residual

    $$e = (I - X(X'X)^{-1}X')\,y = (I - H)\,y$$

  - Partial t-statistic for testing another variable z with partial regression $z^* = (I-H)z$ $O(np^2)$ given $(X'X)^{-1}$

    $$t^2 = (z^{*\prime}e)^2/(s^2\, z^{*\prime}z^*)$$

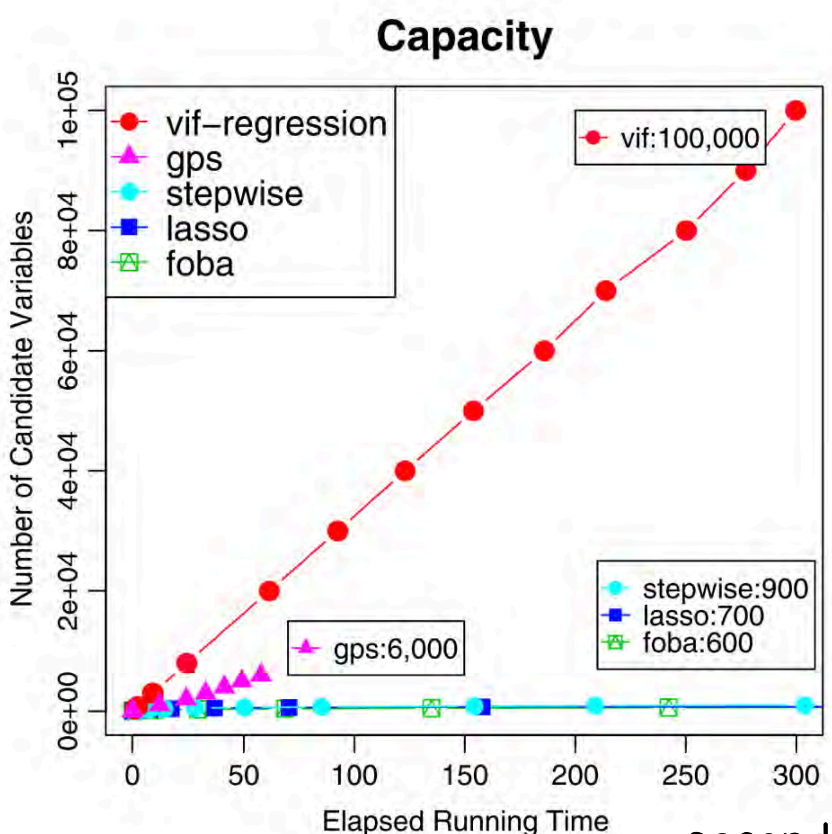
- Re-express t-statistic using VIF

  $$t^2 = (z'e)^2/(s^2\, z'z\, VIF_k)$$

- Conservatively estimate VIF_k from subsample

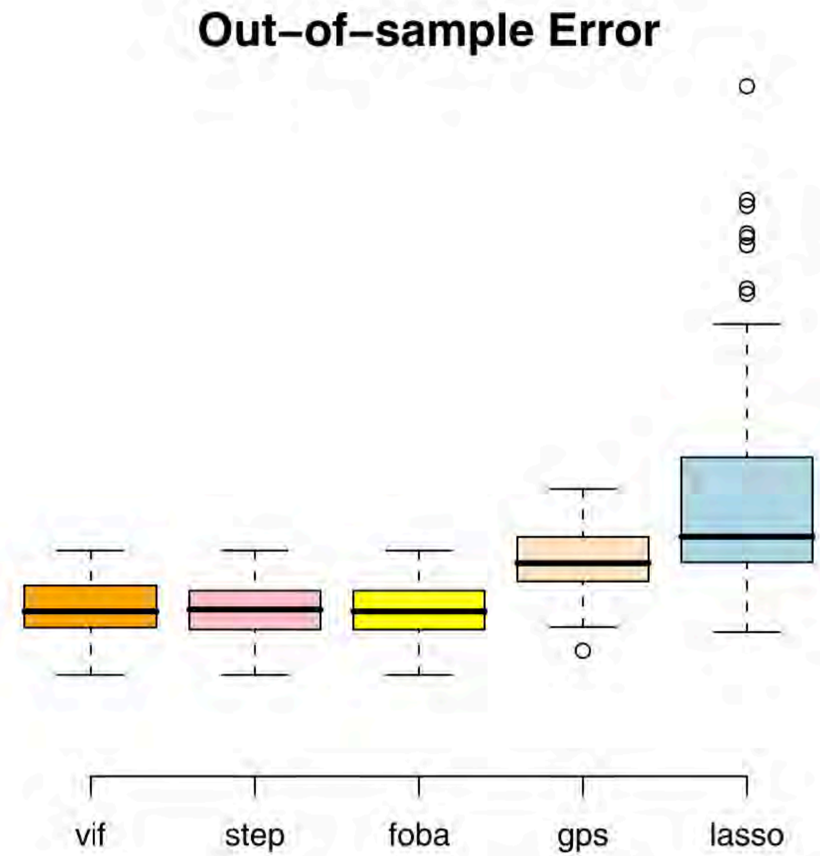
Performance

- Faster than rivals
  - Plus smaller out-of-sample error



seconds

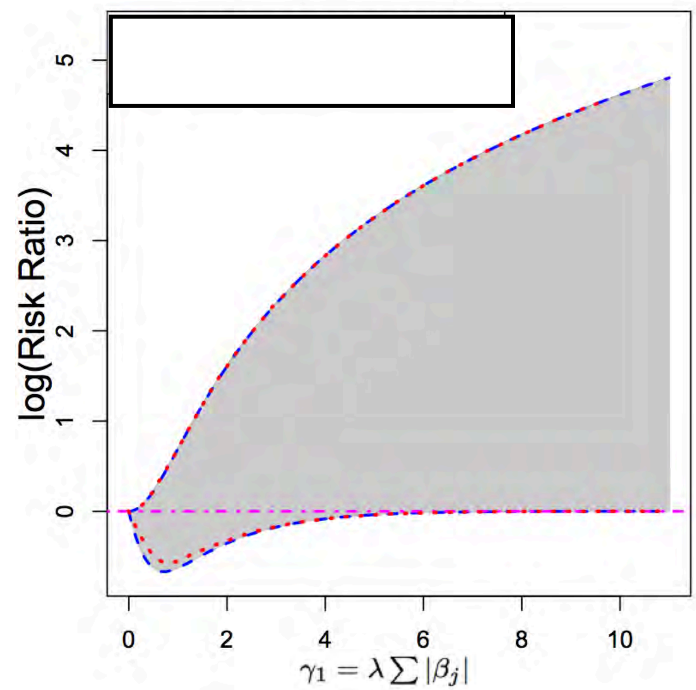
n=1000



n=1000, p=500

Comment on L_1

- Success of lasso depends on nature of underlying model
- Risk comparison
  - Compare the risk of the model identified by subset selection to the model identified by lasso (L_1).
  - Grey region in plot represent possible model datasets
- Take-away
  - In models for which lasso identifies high penalty, L_0 has better performance.
  - Why? It shrinks them all.



Wrap-Up

- Dimension reduction
  - Random projection
  - Subsampling
- Streaming
  - VIF regression
  - Alpha investing, auction models
- Issues
  - Importance of substantive insight
  - Prediction/association vs causation
  - Dependence, population drift

References

- Stochastic Gradient
  - Papers of John Langford, Microsoft Research
- Random projection
  - Halko, Martinsson, and Tropp, SIAM Review, 2011
- VIF Regression
  - "VIF Regression: A Fast Regression Algorithm for Large Data", JASA, 2011, Lin, Foster and Ungar
- Alpha investing
  - "α-investing: a procedure for sequential control of expected false discoveries", JRSSB, 2006
- Improved stepwise regression
  - "Variable selection in data mining: Building a predictive model for bankruptcy", JASA, 2004
- Streaming feature selection
  - "Streamwise feature selection", JMLR, 2006, with Foster, Ungar, and Zhou.

Thanks!