Streaming Feature Selection

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

Plan

- Motivating applications
  - Predictive models
- Sequential testing
  - Alpha investing
- Robust standard errors
  - Sandwich estimator
- Auction framework
- Collaborators
  - Dean Foster
  - Dongyu Lin

Applications

Modeling Challenges

Rare Events

bankruptcy

random forest

Function Estimation

smoothing

wavelets/Dantzig

Linguistics, Text Mining

cloze

TF-IDF

Spatiotemporal Models

disease

MRF + MCMC

TF-IDF: term frequency-inverse document frequency
frequency in document relative to frequency in corpos

MRF: Markov random fields

Text Mining

- Variety of applications…
  - Word disambiguation
    Does "Georgia" refer to a person, US state, or perhaps to a Nation?
  - Tagging parts of speech
    Identifying noun, verb, adjective…
  - Cloze (predicting the next word)
    "…in the midst of modern life the greatest, ___"
- Huge corpus of data from various sources
  - x,000,000 cases
  - novels, news feeds, web pages
  - downloaded the entire text of Wikipedia for testing disambiguation methods

Challenges in Text

- Cloze
  - Is the next word "the" or "her"?
  - "...in the midst of modern life the greatest, ___"
  - Balanced training data with 50/50 rate
- Possible predictors
  - Word frequencies (bag of words)
  - Neighboring sentences/words
  - Parts of speech, tree banks, stem words, synonyms
- Over-fitting?
  - Transfer learning
  - Do predictors in the context of one source (Washington Post) carry over to models for another (NY Times)?

Spatial Temporal Models

- Questions
  - Predict default rates in mortgages, credit cards
- Spatial time series
  - 3,000 counties in US, quarterly since 1997
  - vec(data) gives n = 210,000 (next individuals!)
- Possible predictors
  - Macroeconomic factors, at some geographic unit
  - Personal payment history
  - Local trends
- Modeling issues
  - All sorts of dependence (spatial, temporal)
  - Heterogeneity among observations (counties)
  - Population drift

Goals

- "Turnkey" predictive model that is
  - Competitive with best in each domain
  - Fast
- Stepwise regression (gradient descent)
  - Question is which features (direction)
  - Leverage extensive domain knowledge
  - Regression benefits: well-understood, diagnostics, etc
- Tolerate complex error structure
  - Variety of sources of dependence
  - Heterogeneity of variation
- Avoid over-fitting, "expensive" cross-validation.

Modeling Challenges

<u>Rare Events</u>
bankruptcy
random forest

<u>Function Estimation</u>
smoothing
wavelets/Dantzig

<u>Linguistics, Text Mining</u>
cloze
TF-IDF

<u>Spatiotemporal Models</u>
disease
MRF + MCMC

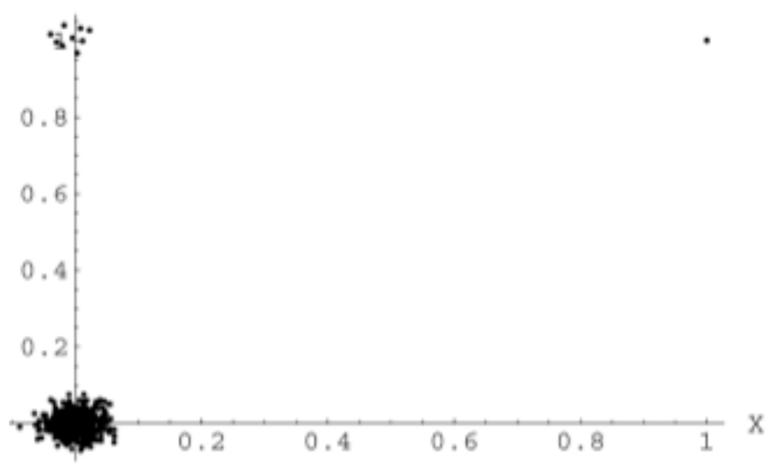
TF-IDF: term frequency-inverse document frequency
frequency in document relative to frequency in corpos

MRF: Markov random fields

Methods

Lessons from Prior Modeling

- Bankruptcy: $n=500{,}000$, $p=60{,}000+$, 450 events
- "Breadth-first" search causes problems
  - Slow, memory hog
  - Severe penalty on largest z-score, sqrt(2 log p)
- If tested features are mostly interactions, then selected features are mostly interactions
  - Example
    $\mu \gg 0$ and $\beta_1, \beta_2 \neq 0$, then $X_1 * X_2 \Rightarrow c + \beta_1 X_1 + \beta_2 X_2$
- Outliers cause problems even with large n



Real p-value ≈ 1/1000,
but
usual t-statistic ≈ 10

Wharton
Department of Statistics

Reaction to Lessons

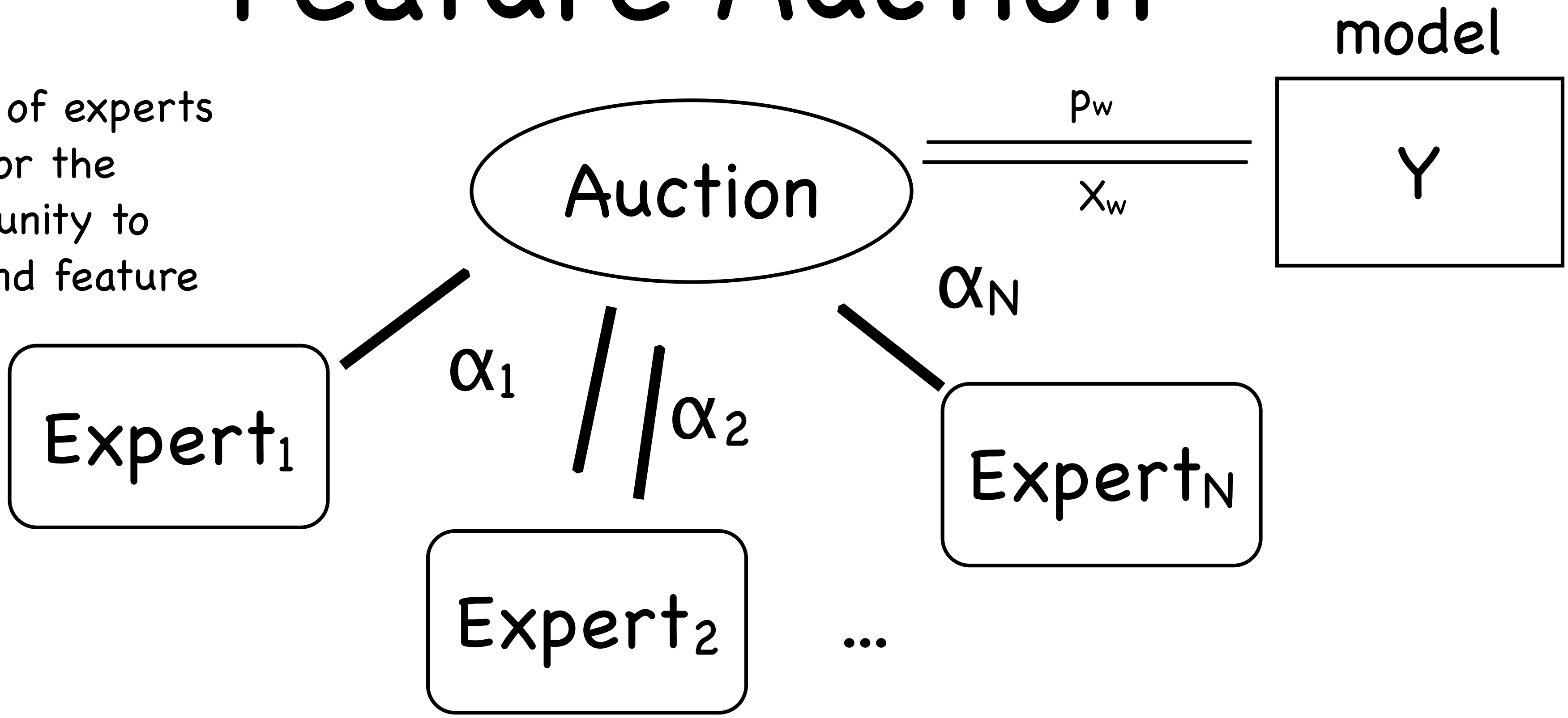
- Breadth-first becomes <u>streaming selection</u>
  - Test a sequence of possible features
  - Examining each is very fast
  - Over-fitting? Multiplicity adjustments?
- Equal significance levels replaced by levels that vary with the type of the variable
  - Simple Bonferroni procedure
  - Divide α level equally between linear & interactions
    - p linear: test each at level $\alpha/(2p)$
    - p^2 interactions: test at level $\alpha/(2p^2)$
- Rather than trust model to obtain standard errors, use a more robust estimate.

Methods Summary

- Supercharged stepwise regression
- Auction
  Explore more expansive feature space
- Robust standard errors (ultimately p-values)
  Allow for dependence and heterogeneity
- Alpha investing
  Control over-fitting adaptively

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

- 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
- Parasitic 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. These then can win later bids and recommend more X's
  - Eliminates experts whose features are not accepted.
- Taxes fund parasitic experts
  - Ensure that continue to control overall FDR
- Critical
  - control multiplicity in a sequence of hypotheses
  - 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
    When train on observations from selected regions or document sources, what can you infer to others?
- What are the right degrees of freedom?
  - Tukey story

Sandwich Estimator

- Usual OLS estimate of variance
  - Assume your model is true

$$\begin{aligned} var(b) &= (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'X)\ (X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

- Sandwich estimators
  - Robust to deviations from assumptions

heteroscedasticity

$$\begin{aligned} var(b) &= (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= (X'X)^{-1}\ X'D^2X\ (X'X)^{-1} \end{aligned}$$

diagonal

dependence

$$\begin{aligned} var(b) &= (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\ X'BX\ (X'X)^{-1} \end{aligned}$$

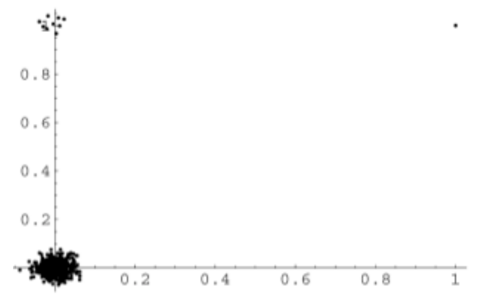
block diagonal

Essentially the "Tukey" method

Wharton
Department of Statistics

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.
  - Credit modeling
    Everything seems significant unless incorporate dependence into the calculation of the SE

Alpha Investing

- Situation
  - Test possibly infinite sequence of m hypotheses

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

    obtaining the p-values $p_1, p_2, \ldots$
  - Order of tests may depend prior outcomes
- Procedure
  - Start with an initial alpha wealth $W_0 = \alpha$
  - Invest wealth $0 \le \alpha j \le Wj$ in the test of Hj
  - Change in wealth depends on test outcome
  - $\omega \le \alpha$ denotes the payout earned by rejecting

$$
W_j - W_{j-1} = \begin{cases} \omega \text{ if } p_j \le \alpha_j \\ -\alpha_j/(1-\alpha_j) \text{ if } p_j \le \alpha_j \end{cases}
$$

Properties of Alpha Investing

- Provides uniform control of the expected false discovery rate. At any stopping time during testing, martigale argument shows

$$\sup_{\theta} \frac{E(\#\text{false rejects})}{E(\#\text{rejects})+1} \leq \alpha$$

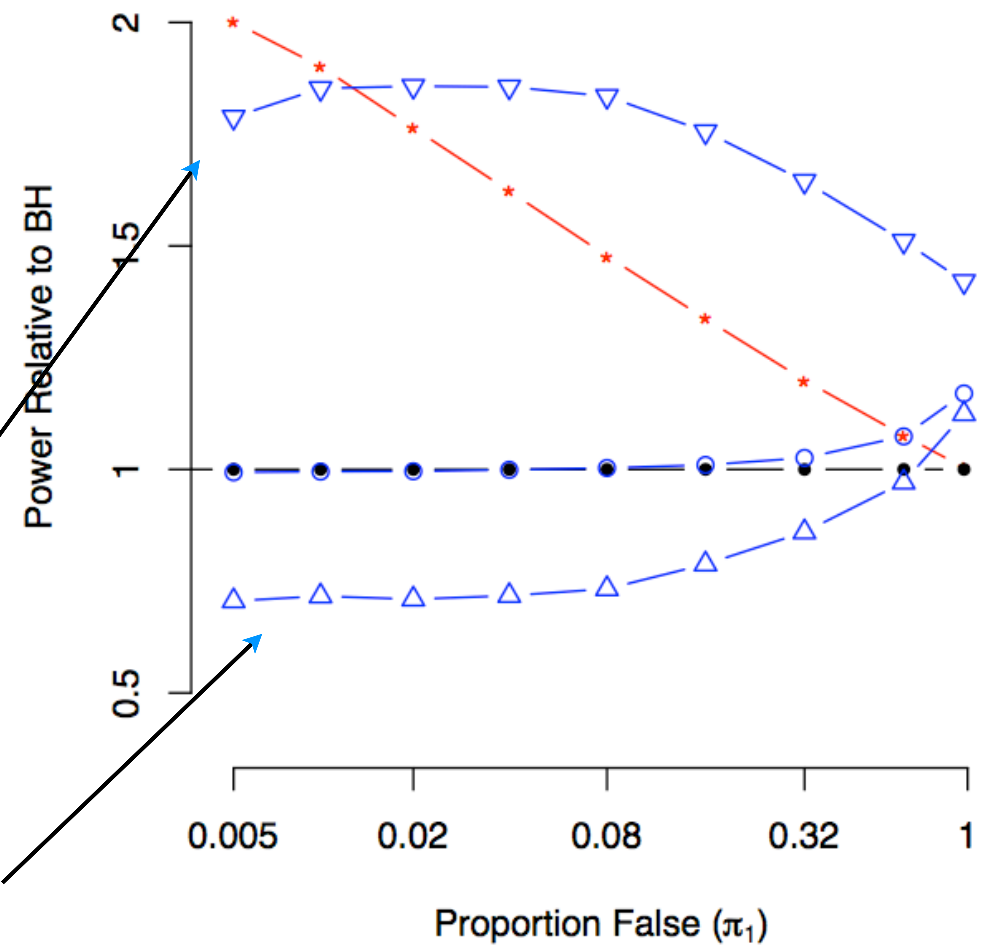
- Flexibility in choice of how to invest alpha-wealth in test of each hypothesis
  - Example. Invest more when just reject if suspect that significant results cluster.
  - Universal strategies
- Avoids need to compute p-values in advance

Connections

- Bonferroni test of $H_1, \dots, H_m$
  - Set $W_0 = \alpha$ and reward $\omega = 0$
  - Bid $\alpha_j = \alpha/m$
- Step-down test of Benjamini & Hochberg
  - Set $W_0 = \alpha$ and reward $\omega = \alpha$
  - Test all m at level α/m
  - If none are significant, done
  - If one is significant, earn α back
    - Test remaining m-1 conditional on $p_j > \alpha/m$

Benefits of Knowledge

- Test m = 200 hypotheses
- Compare power to Benjami-Hochberg
- Signal from spike and slab prior



Oracle BH

correct order

random order

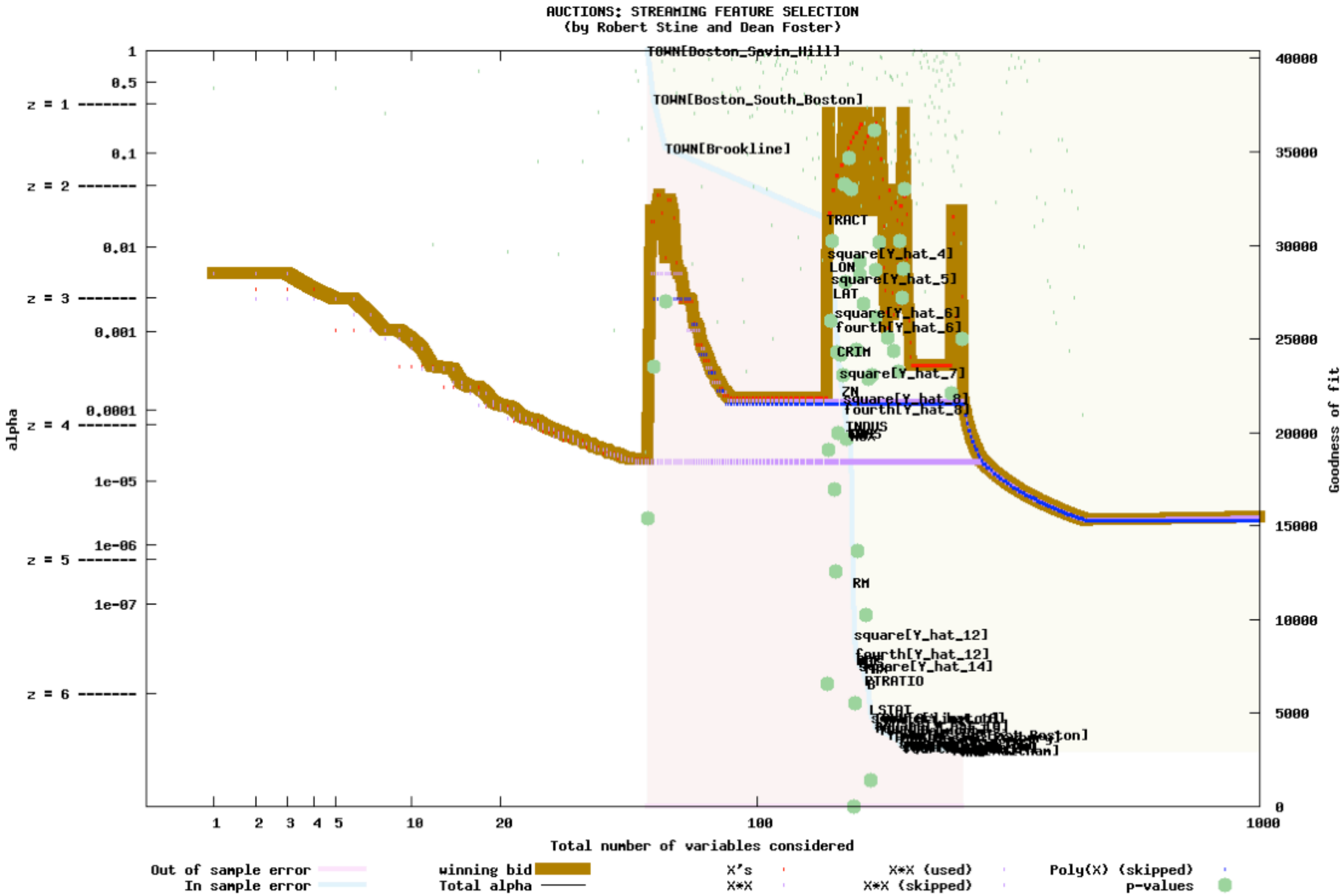
Alpha investing

Example Results

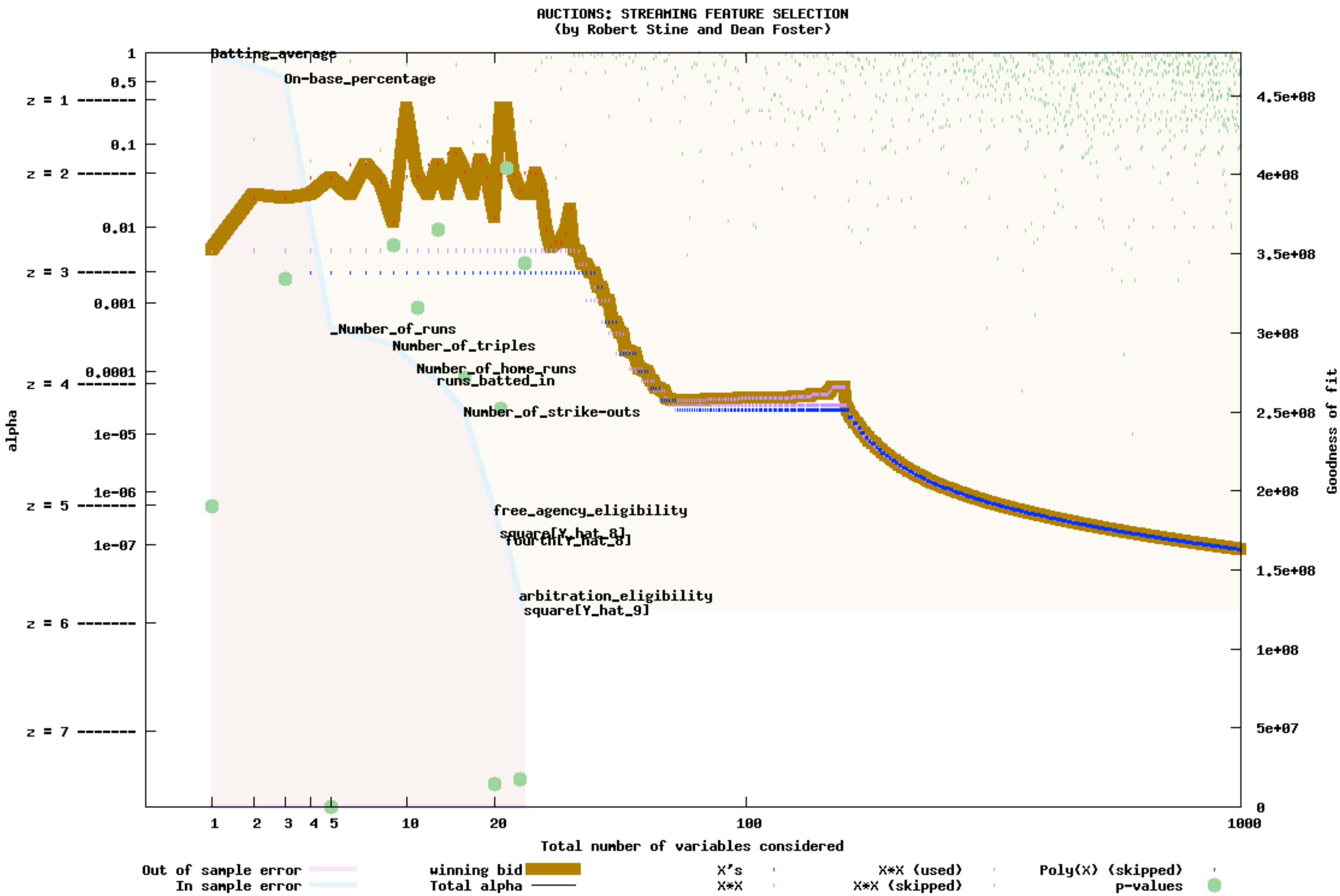
Examples

- On-line prototype used in classroom
  - Limited experts
  - www-stat.wharton.upenn.edu/~foster
- Data
  - Supply a csv file or use one provided
- Graphical summary
  - all expert bids and winning bid
  - p-value of result
  - accepted variable names
  - CVSS

Boston Housing



Baseball



Wharton
Department of Statistics

Next Steps

- Very much a work in progress
- Improved experts
  - Identify common expert classes that appear in various situations (eg, cluster detection)
  - Neighborhood structure
    - geographical
    - temporal
- Better software
  - Front end
  - Back end

Discussion

- Expert bidding
  - Aggressive vs Passive
  - "Stacking the deck"
- Anonymous vs attributed variables
  - Stat traditionally models X_1, X_2, ...
  - Right emphasis?
- Standard errors are only part of the path to a good p-value
  - Other bounds often useful (Bennett type)

References

- Feature auction
  - www-stat.wharton.upenn.edu/~stine
- Alpha investing
  - "α-investing: a procedure for sequential control of expected false discoveries", JRSSB, 2006
- Early improved stepwise regression
  - "Variable selection in data mining: Building a predictive model for bankruptcy", JASA, 2004
- Robust standard errors
  - "Variable selection in models with blockwise dependence", Lin and Foster.

Thanks!

Wharton
Department of Statistics