

Variable Selection in Wide Data Sets

Bob Stine

Department of Statistics

The Wharton School of the University of Pennsylvania

www-stat.wharton.upenn.edu/stine

Overview

- Problem
 - Picking the features to use in a model
 - Wide data sets imply many choices
- Examples
 - Simulated idealized problems
 - Predicting credit risk
- Themes
 - Modifying familiar tools for data mining
 - Dealing with the problem of multiplicity

Collaboration with Dean Foster

Questions

Credit scoring

Who will repay the loan? Who will declare bankruptcy?

Identifying faces

Is this a picture of a face?

Genomics

Does a pattern of genes predict higher risk of a disease?

Text processing

Which references were left out of this paper?

Questions

Credit scoring

Who will repay the loan? Who will declare bankruptcy?

Identifying faces

Is this a picture of a face?

Genomics

Does a pattern of genes predict higher risk of a disease?

Text processing

Which references were left out of this paper?

These are great statistics problems, so...

Why not use our workhorse, regression?

Its familiar with diagnostics available.

Regression Models

Simple structure

- n **independent** observations of m features
- q predictors in model with error variance σ^2 :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \epsilon$$

Calibrated predictions allow various link functions ...

- Linear regression has identity, logistic has logit link.

$$E(Y|X) = h(\beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q)$$

Rich feature space is key to the model

- Usual mix: continuous, categorical, dummy vars, ...
- Missing data indicators
- Combinations (principal components) and clusters
- Nonlinear terms, transformations (quadratics)

Hard Question

If we allow for the possibility of **interactions** among the predictors, then the feature space becomes larger still.

Which features generate the best predictions?

Also hope to do this without cross-validation.

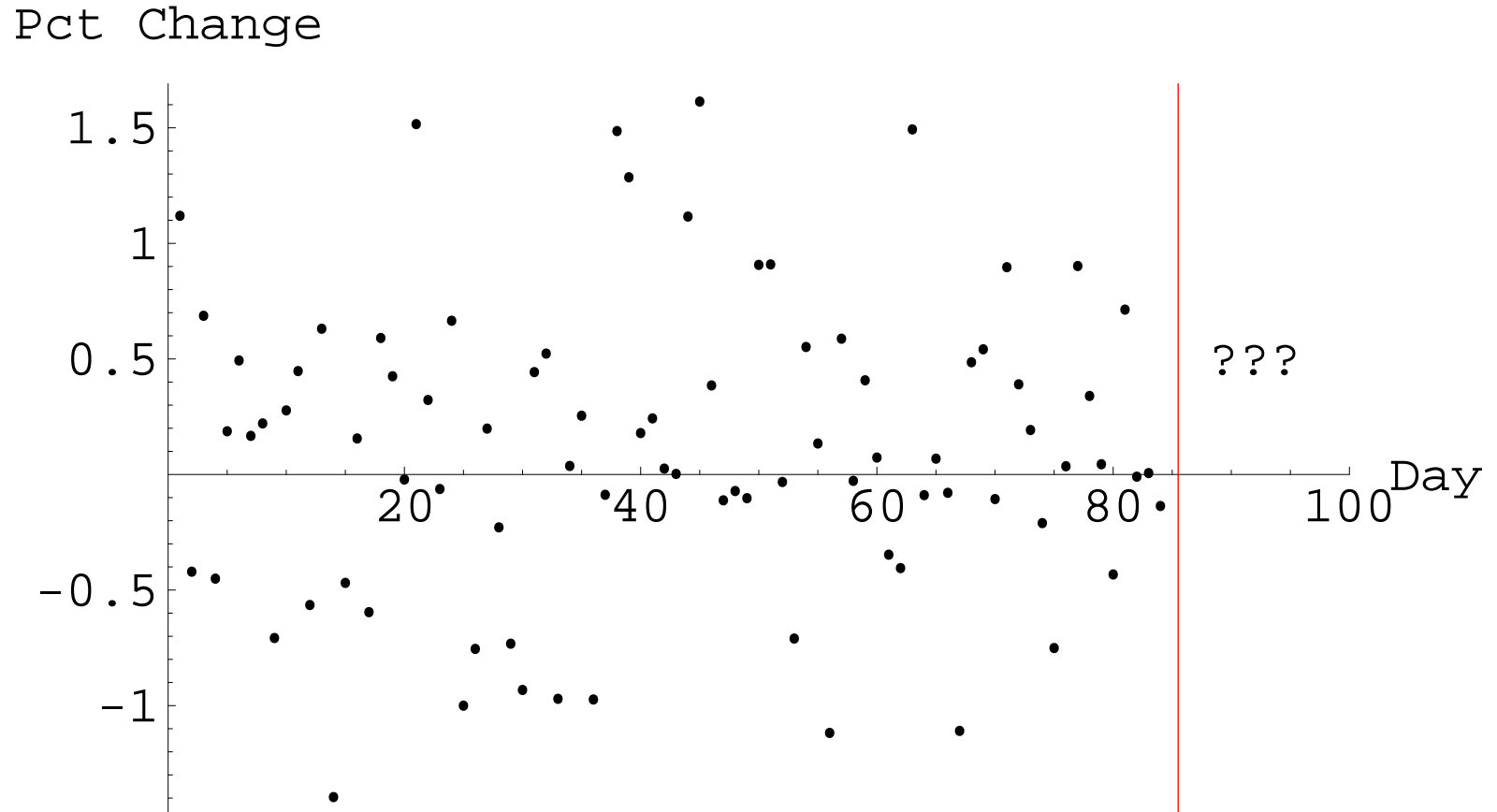
A favorite example: Stock Market

Where's the stock market headed in 2005?

A favorite example: Stock Market

Where's the stock market headed in 2005?

Daily percentage changes in the closing price of the S&P 500 during the last 3 months of 2004 (85 trading days) ...



Predicting the Market

Problem

Predict returns on S&P 500 in 2005 using features built from 12 *exogenous* factors (*a.k.a.*, technical trading rules).

Regression model

$R^2 = 0.85$ using $q = 28$ predictors.

With $n = 85$, $F = 12$ with p-value < 0.00001 .

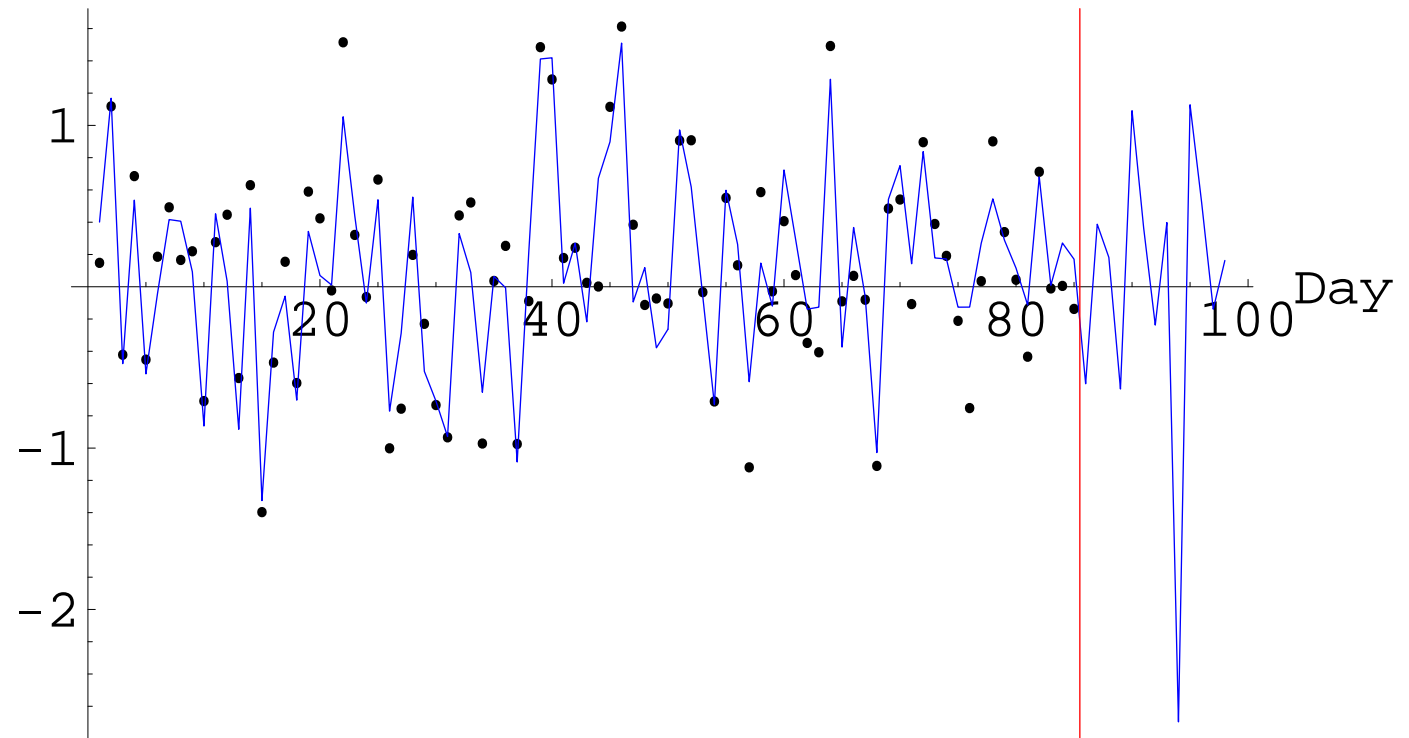
Coefficients are impressive...

Term	Estimate	Std Error	t-Ratio	p-value
Intercept	0.323	0.078	4.14	0.0001
X4	0.172	0.040	4.34	0.0000
(X1)*(X1)	-0.202	0.039	-5.16	0.0000
(X1)*(X5)	0.256	0.048	5.34	0.0000
(X2)*(X6)	0.289	0.044	6.59	0.0000
(X7)*(X9)	0.249	0.046	5.37	0.0000

Model Fit and Predictions

Predictions track the historical returns closely, even matching turning points.

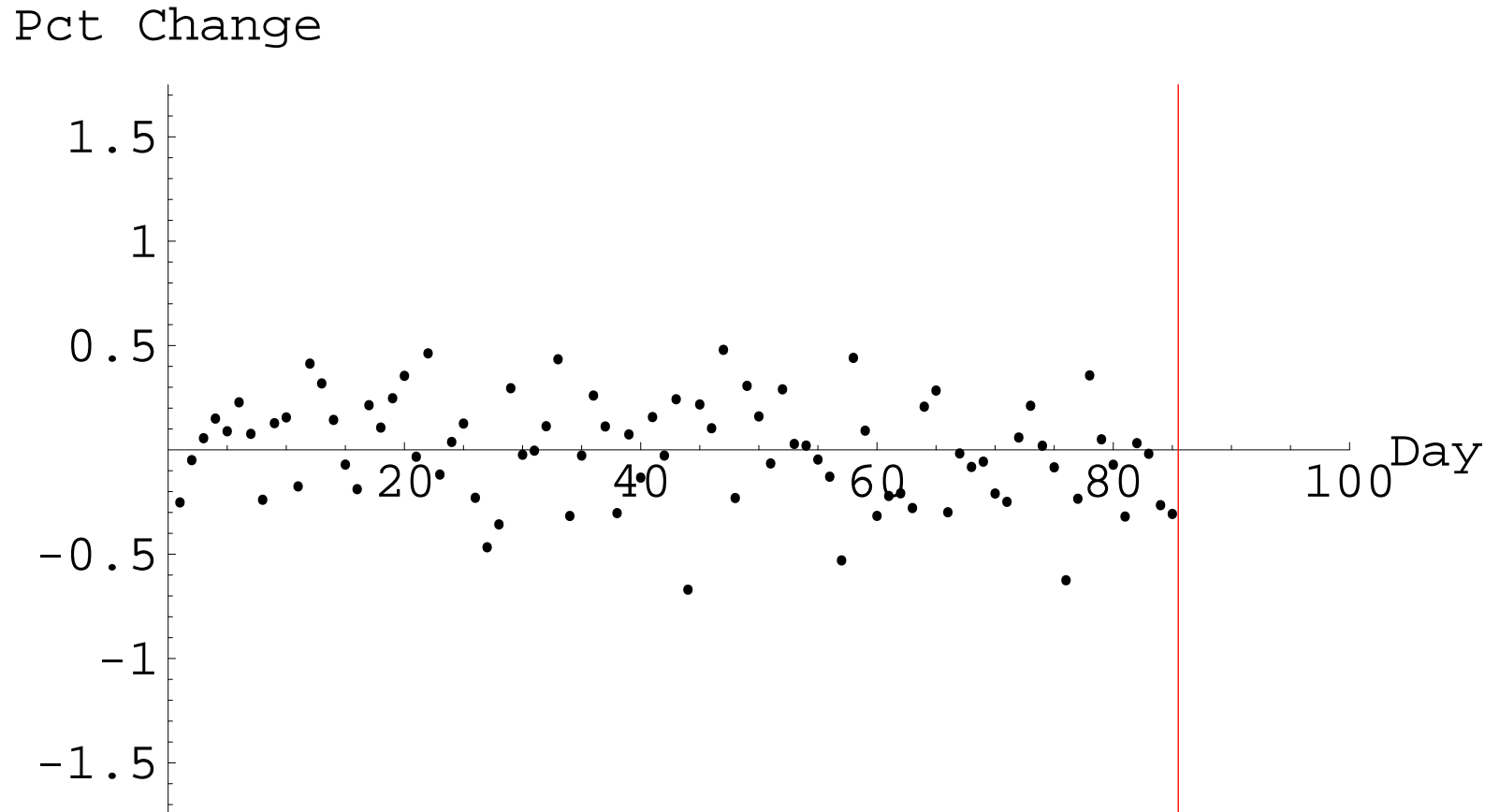
Daily Pct Change



Better short the market that day... No guts, no glory!

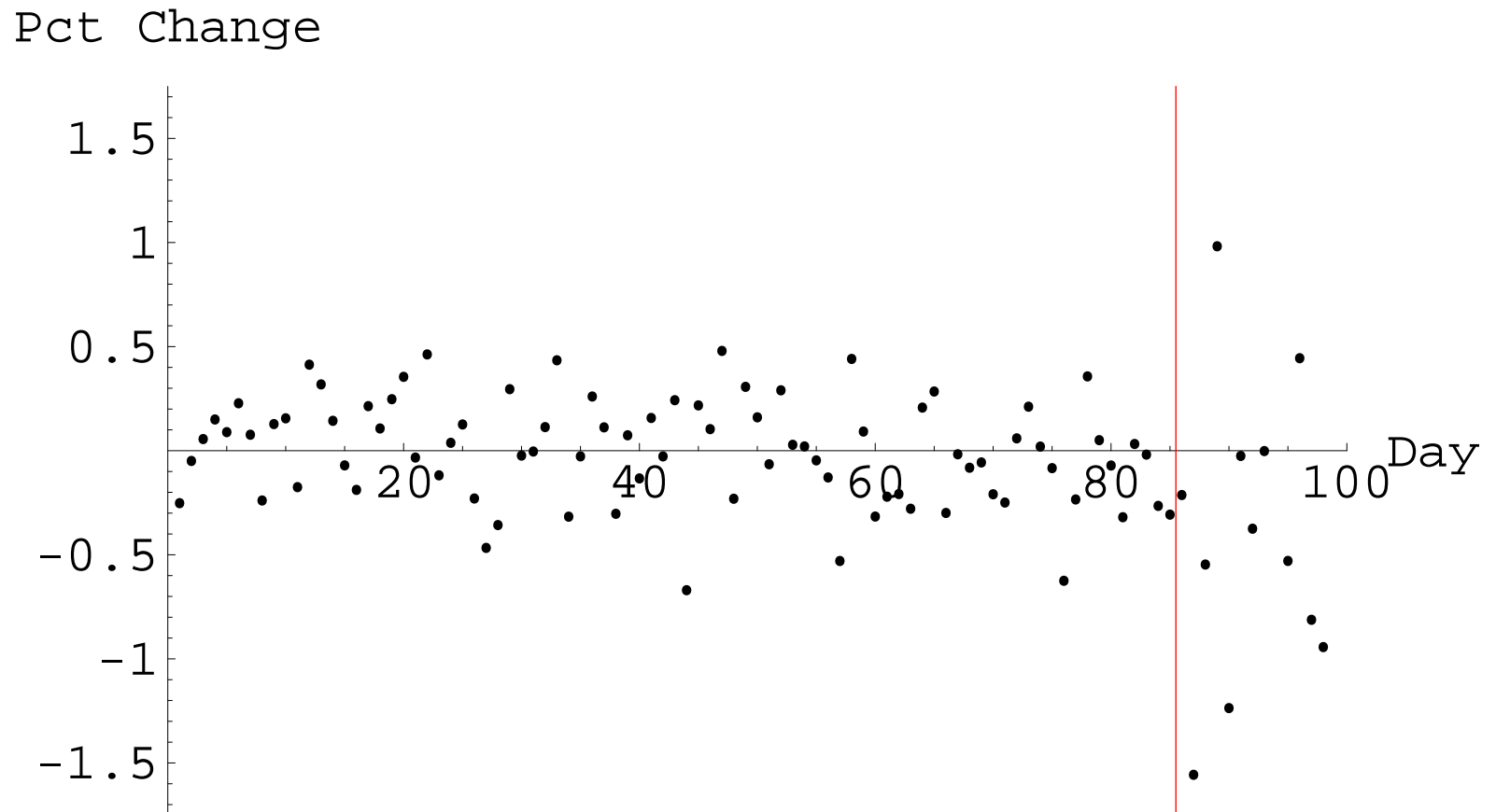
Prediction Errors

In-sample prediction errors are small...



Prediction Errors

How'd you lose the house? Even Bonferroni lost his house!



What was that model?

Exogenous base variables

12 columns of Gaussian noise.

Null model is the right model.

Selecting predictors

Turn stepwise regression loose on $m = 90$ features

12 linear + 12 Squares + 66 Interactions

with “promiscuous” settings for adding variables,

prob-to-enter = 0.16 (AIC)

Then run stepwise backward to remove extraneous effects.

Why does this fool Bonferroni?

s^2 becomes biased, so remaining predictors appear more significant. Often “cascades” into a perfect fit.

Cannot fit saturated model to get unbiased estimate of σ^2 because $m > n$.

Big Picture, 1

Moral

Stepwise regression can make a silk purse from sow's ear.

Must control the fitting process

Use Bonferroni from the start, not after the fact. No predictor joins model — the right answer.

So, how does one control the fitting process?

Particularly when there are many possible choices.

Second Example: Predicting Bankruptcy

Predict onset of personal bankruptcy

Estimate probability customer declares bankruptcy.

Challenge

Can **stepwise regression** predict as well as commercial “data-mining” tools or substantive models?

Many features About 350 “basic” variables

- Short time series for each account
- Spending, utilization, payments, background
- Missing data and indicators
- Interactions are important (cash advance in Vegas)

$m > 67,000$ predictors!

- **Transaction** history would vastly expand the problem.

Complication: Sparse Response

How much data do you really have?

3 million months of activity, each with 67,000 predictors.

Needle in many haystacks

Observe only 2,244 bankruptcies.

Further complication

What loss function should be minimized?

Profitable customers look risky. Want to lose them?

Who is the ideal customer?

“Borrow lots of money and pay it back slowly.”

Cross-validation

Less appealing with so few events.

Approach

Forward stepwise search

Use p-values to determine whether to add predictors.

Two questions

1. How to calculate a p-value?
2. Where to put the threshold for the p-value?

Risk inflation criterion (RIC) ($m = \#$ possible predictors)

Hard thresholding, Bonferroni, Fisher's method

$$\text{Add } X_j \iff t_j^2 > 2 \log m \iff p_j < \frac{1}{m}$$

Obtains RIC bound (Foster & George 1994)

$$\min_{\hat{\beta}} \max_{\beta} E \frac{\|Y - X\hat{\beta}\|^2}{|\beta| \sigma^2} \leq 2 \log m$$

where $|\beta| = \#\{\beta_j \neq 0\}$.

Adaptive Variable Selection

Sparse models

RIC best when “truth” is sparse, but lacks power if much signal. **False discovery rate** motivates an alternative.

Adaptive threshold If you’ve added q predictors,

$$\text{Add } X_j \iff p_j < \frac{q+1}{m}$$

Further motivation

- Half-normal plot (Cuthbert Daniel?)
- Generalized degrees of freedom (Ye 1998, 2002)
- Empirical Bayes (George & Foster 2000)
- Information theory (Foster, Stine & Wyner 2002)

Universal prior Which predictors minimize this ratio?

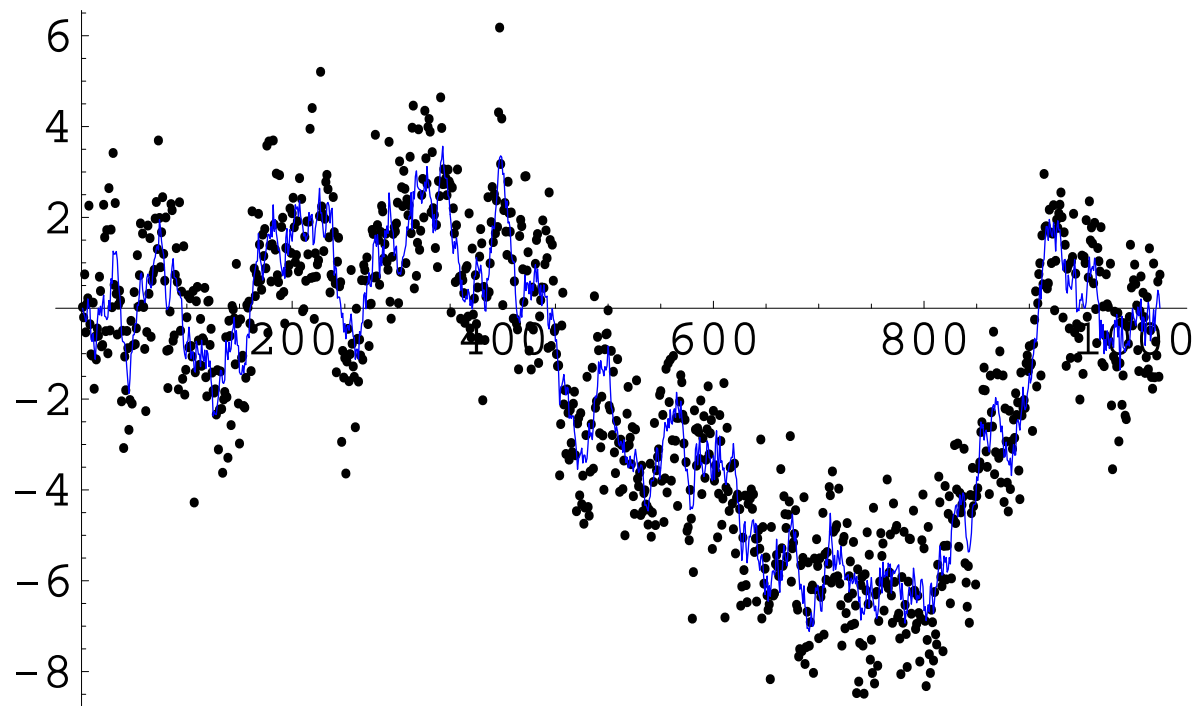
$$\min_{\hat{q}} \max_{\pi} \frac{E \|Y - \hat{Y}(\hat{q})\|^2}{E \|Y - \hat{Y}(\pi)\|^2} \text{ for } \beta \sim \pi$$

Example: Finding Subtle Signal

Signal is a **Brownian bridge**

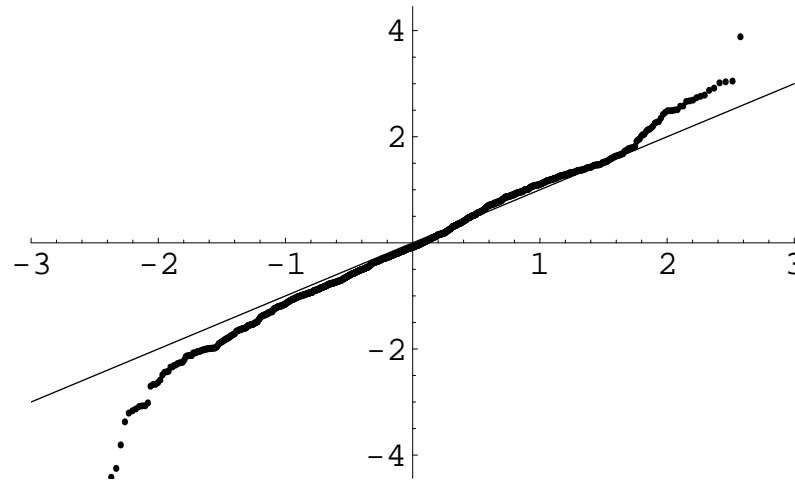
Stylized version of financial volatility.

$$Y_t = BB_t + \sigma \epsilon_t$$

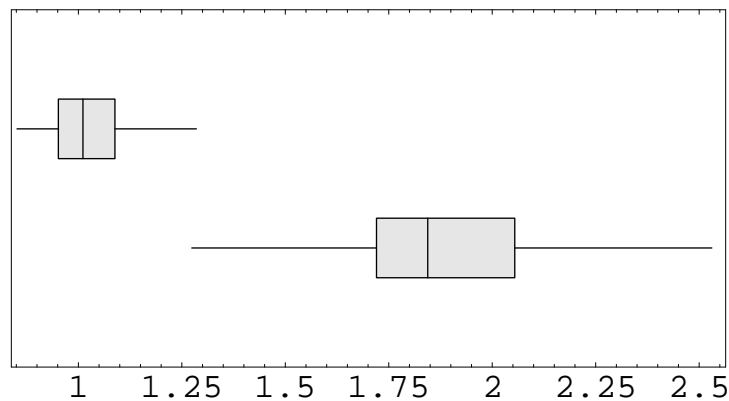


Example: Finding Subtle Signal

Wavelet transform has many coefficients but none is very large relative to the others.



MSE with adaptive (top) vs. hard (bottom)



But it fails for bankruptcy!

Thresholding classical p-values does not work

Fit on 600,000, predict 2,400,000.

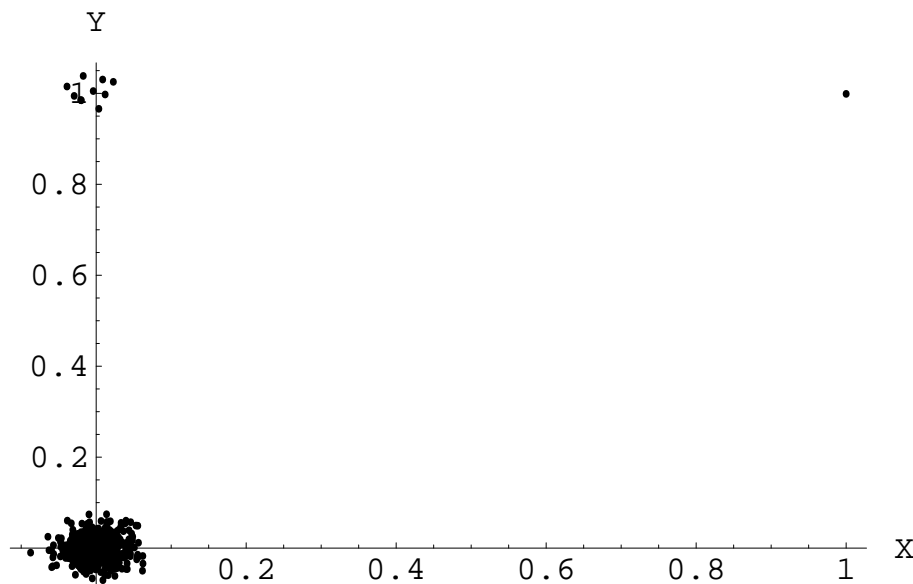
As fitting proceeds, out-of-sample error shows sudden spike.

Diagnostic **plots** revealed the problem.

Stylized problem $n = 10,000$, dithered

$X_1 = 1, X_2, \dots, X_{10,000} \sim N(0, 0.025)$

$P(Y = 1) = 1/1000$, independent of X .



$t = 14$, but common sense suggests $p \approx 1/1000$.

What went wrong?

Theory implies p-values are known

- Normal distribution on error (thin tailed).
- Know “true” error variance σ^2 , plus its constant.
- Predictors are orthogonal.

$$\Rightarrow \hat{\beta}_j \stackrel{\text{iid}}{\sim} N(\beta_j, \text{SE}_j)$$

What went wrong?

Theory implies p-values are known

- Normal distribution on error (thin tailed).
- Know “true” error variance σ^2 , plus its constant.
- Predictors are orthogonal.

$$\Rightarrow \hat{\beta}_j \stackrel{\text{iid}}{\sim} N(\beta_j, \text{SE}_j)$$

In practice...

- Is any data ever normally distributed?
 - Don't believe there's a true model, much less fixed error variance.
 - Collinear features, frequently with $m > n$.
- \Rightarrow At best, can only guess p-values.

Honest p-values

Concern from stock example

Must avoid “false positives” that lead to inaccurate predictions, cascade of mistakes.

Robustness of validity

Rank regression would also protect you to some extent, but has problems dealing with extreme heteroscedasticity.

Alternatives

Method	Requires
Robust estimator (e.g. ranks)	Homoscedastic
White estimator	Symmetry
Bennett bounds	Bounded Y

Bounded influence estimators

Another possibility, but can it be computed fast enough?

White Estimator

$$Y = \hat{\beta}_0 + \hat{\beta}_{q,1}X_{q,1} + \cdots + \hat{\beta}_{q,q}X_{q,q} + \epsilon$$

Sandwich formula (H. White, 1980, *Econometrica*)

$$\text{Var}(\hat{\beta}_q) = (X'_q X_q)^{-1} X'_q \underbrace{\text{Var}(\epsilon)} X_q (X'_q X_q)^{-1}$$

Estimate variance using the residuals from **prior** step:

$$\text{Var}(\hat{\beta}_q) = (X'_q X_q)^{-1} X'_q \underbrace{\text{Diag}(e^2_{q-1})} X_q (X'_q X_q)^{-1}$$

Result in stylized problem

Estimated SE is 10 times larger, more appropriate p-value.

Moral: Watch your null hypothesis!

Only test $H_0 : \beta_{q,q} = 0$ rather than

$$H_0 : \beta_{q,q} = 0 \quad \& \quad \text{Var}(\epsilon_i) = \sigma^2$$

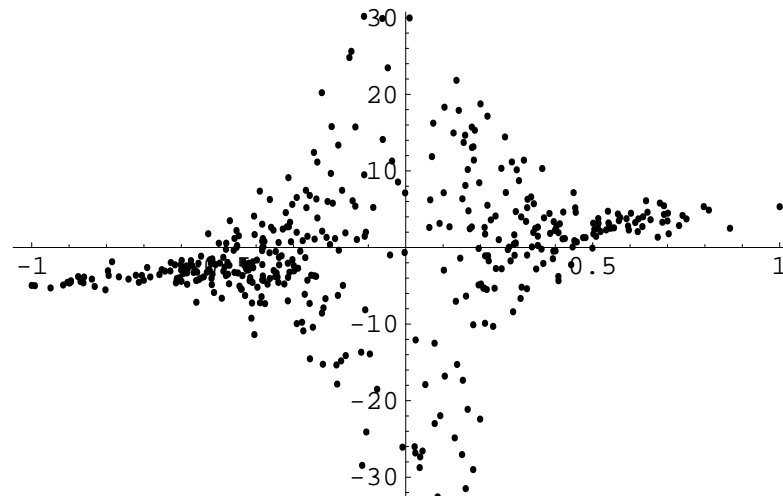
Example: Finding Missed Signal

Question

Does White estimator ever find an effect that OLS misses?
Usually OLS underestimates SE.

Heteroscedastic data

High variance at 0 obscures the differences at extremes.



Least squares

Standard OLS reports $SE = 3.6$ giving $t = 1.2$.
White $SE \approx 0.9$ and finds the underlying effect.

Honest p-values for Testing $H_0 : \beta = 0$

Three methods

Each method makes some assumption about the data in order to produce reliable p-value.

Method	Requires
Robust estimator (e.g. ranks)	Homoscedastic
White estimator	Symmetry
Bennett bounds	Bounded Y

Bennett Inequality

Think differently

Sampling distribution of $\hat{\beta}$ is not normal because huge leverage contributes “Poisson-like” variation.

Bennett inequality (Bennett, 1962, *JASA*)

Bounded independent r.v. U_1, \dots, U_n with $\max |U_i| < 1$,
 $E U_i = 0$, and $\sum_i E U_i^2 = 1$,

$$P\left(\sum_i U_i \geq \tau\right) \leq \exp\left(\frac{\tau}{M} - \left(\frac{\tau}{M} + \frac{1}{M^2}\right) \log(1 + M\tau)\right)$$

If $M\tau$ is small, $\log(1 + M\tau) \approx M\tau - M^2\tau^2/2$

Allows heteroscedastic data

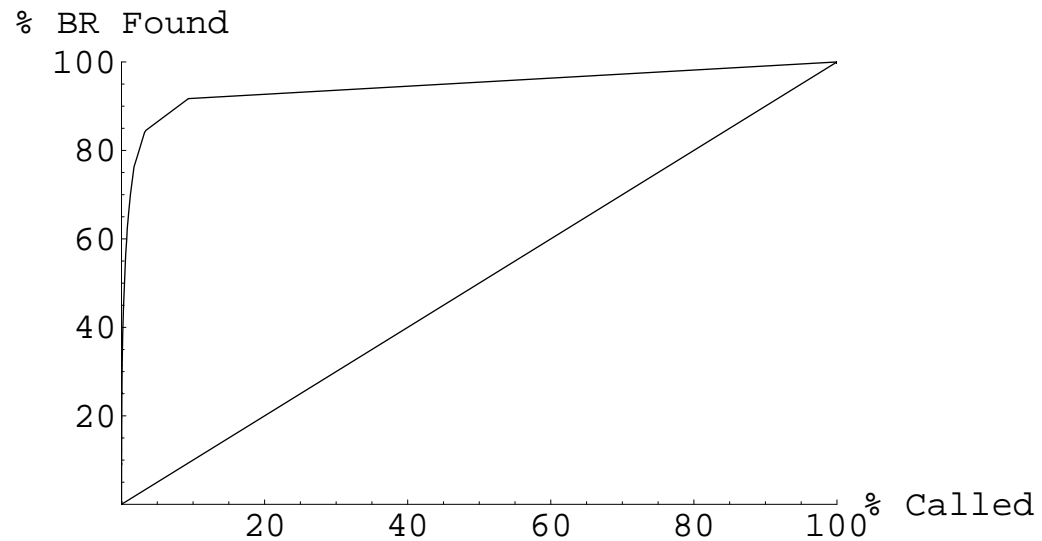
Free to divy up the variances as you choose, albeit only for bounded random variables.

In stylized example assigns p-value $\approx 1/100$.

Classification Results

Success!

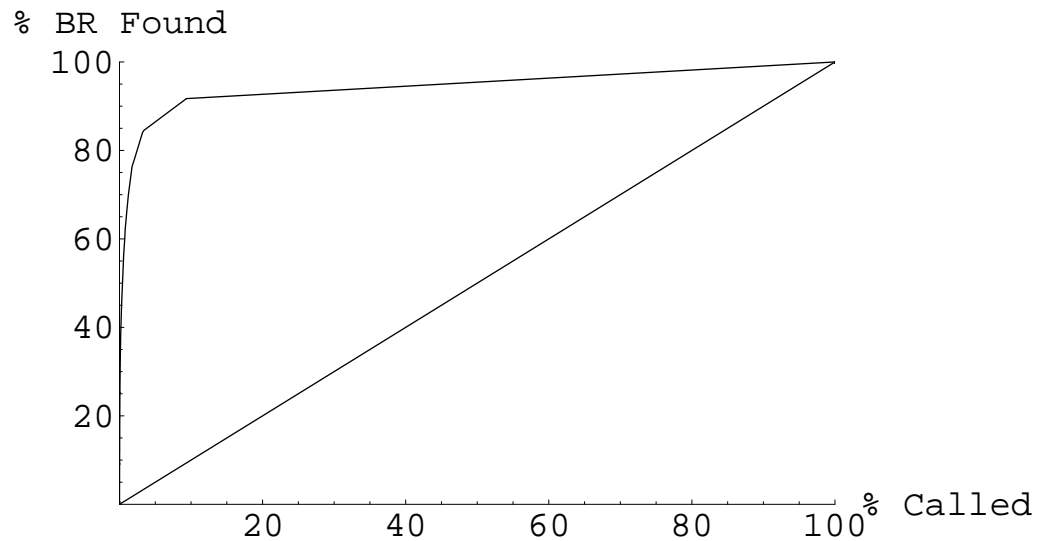
Not only does the model obtain smaller costs than C4.5 (w/wo boosting), it has huge lift:



Classification Results

Success!

Not only does the model obtain smaller costs than C4.5 (w/wo boosting), it has huge lift:



But...

- Most predictors were interactions.
- Know that you missed some things.
- Slloooowwwwwwwwwww.

Big Picture, 2

Moral

Better get the right p-values

Successful

Carefully computed standard errors and p-values

+

Adaptive thresholding rules

= Competitive procedure

Big Picture, 2

Moral

Better get the right p-values

Successful

Carefully computed standard errors and p-values

+

Adaptive thresholding rules

= Competitive procedure

Downside

- Know that we're not leveraging domain knowledge.
- Model is more complicated than need be.
- Took forever to run.

Obesity

The issues arising the bankruptcy application only get worse as the data set gets wider...

Obesity

The issues arising the bankruptcy application only get worse as the data set gets wider...

and data sets seem to be getting wider and wider.

<i>Application</i>	<i>Number of Cases</i>	<i>Number Raw Features</i>
Bankruptcy	3,000,000	350
Faces	10,000	1,400
Genetics	1,000	10,000
CiteSeer	500	10,000,000

Sequential Selection

Simple idea

Search features in *some* order rather than all at once.

Opportunities

- Incorporate substantive knowledge into search order
- Sequential selection of features based on current status
- Open vs. closed view of space of predictors
- Run faster

Heuristics

- Theory from adaptive selection suggests that if you can order predictors, then little to be gained from knowing β .
- Alpha spending rules in clinical trials.
- Depth-first rather than breath-first search.

Multiple Hypothesis Testing

Test m null hypotheses $\{H_1, \dots, H_m\}$, $H_j : \theta_j = 0$.

		Accept H_0	Reject H_0	
True	H_0	$U^\theta(m)$	$V^\theta(m)$	m_0
State	H_0^c	$T^\theta(m)$	$S^\theta(m)$	$m - m_0$
		$m - R(m)$	$R(m)$	m

$$V^\theta(m) = \sum_{j=1}^m V_j^\theta \text{ indicators}$$

Multiple Hypothesis Testing

Test m null hypotheses $\{H_1, \dots, H_m\}$, $H_j : \theta_j = 0$.

		Accept H_0	Reject H_0	
True	H_0	$U^\theta(m)$	$V^\theta(m)$	m_0
State	H_0^c	$T^\theta(m)$	$S^\theta(m)$	$m - m_0$
		$m - R(m)$	$R(m)$	m

$$V^\theta(m) = \sum_{j=1}^m V_j^\theta \text{ indicators}$$

Classical criterion Family wide error rate

$$\text{FWER}(m) = P_0(V^\theta(m) \geq 1)$$

Multiple Hypothesis Testing

Test m null hypotheses $\{H_1, \dots, H_m\}$, $H_j : \theta_j = 0$.

		Accept H_0	Reject H_0	
True	H_0	$U^\theta(m)$	$V^\theta(m)$	m_0
State	H_0^c	$T^\theta(m)$	$S^\theta(m)$	$m - m_0$
		$m - R(m)$	$R(m)$	m

$$V^\theta(m) = \sum_{j=1}^m V_j^\theta \text{ indicators}$$

Classical criterion Family wide error rate

$$\text{FWER}(m) = P_0(V^\theta(m) \geq 1)$$

Classical procedure Bonferroni

Test each H_j at level α_j so that

$$P_0(V_j^\theta = 1) \leq \alpha_j, \quad \sum_j \alpha_j \leq \alpha$$

False Discovery Rate

Approach Control testing procedure once it rejects.

False Discovery Rate

Approach Control testing procedure once it rejects.

FDR criterion (Benjamini & Hochberg 1995, *JRSSB*)

Proportion of false positives among rejects

$$\text{FDR}(m) = E_{\theta} \left(\frac{V(m)}{R(m)} \mid R(m) > 0 \right) P(R(m) > 0) .$$

False Discovery Rate

Approach Control testing procedure once it rejects.

FDR criterion (Benjamini & Hochberg 1995, *JRSSB*)

Proportion of false positives among rejects

$$\text{FDR}(m) = E_{\theta} \left(\frac{V(m)}{R(m)} \mid R(m) > 0 \right) P(R(m) > 0) .$$

Step-down testing (procedure)

Order p-values of m **independent** tests of hypotheses

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

Starts with Bonferroni comparison $p_{(1)} \leq \alpha/m$, then gradually raises threshold as more hypotheses are rejected.

Reject $H_{(j)}$ if $p_{(j)} \leq \alpha j/m$.

Controls FWER and FDR with more power.

An Alternative to FDR

Count the number of *correct* rejections $S^\theta(m)$ in excess of a fraction of the total number rejected.

$$\text{EDC}_{\alpha,\gamma}(m) = E_\theta[S^\theta(m) - \gamma R(m)] + \alpha, \quad 0 < \alpha, \gamma < 1.$$

Heuristically, α controls FWER and γ controls FDR.

Testing procedure “controls EDC” $\iff \text{EDC} \geq 0$

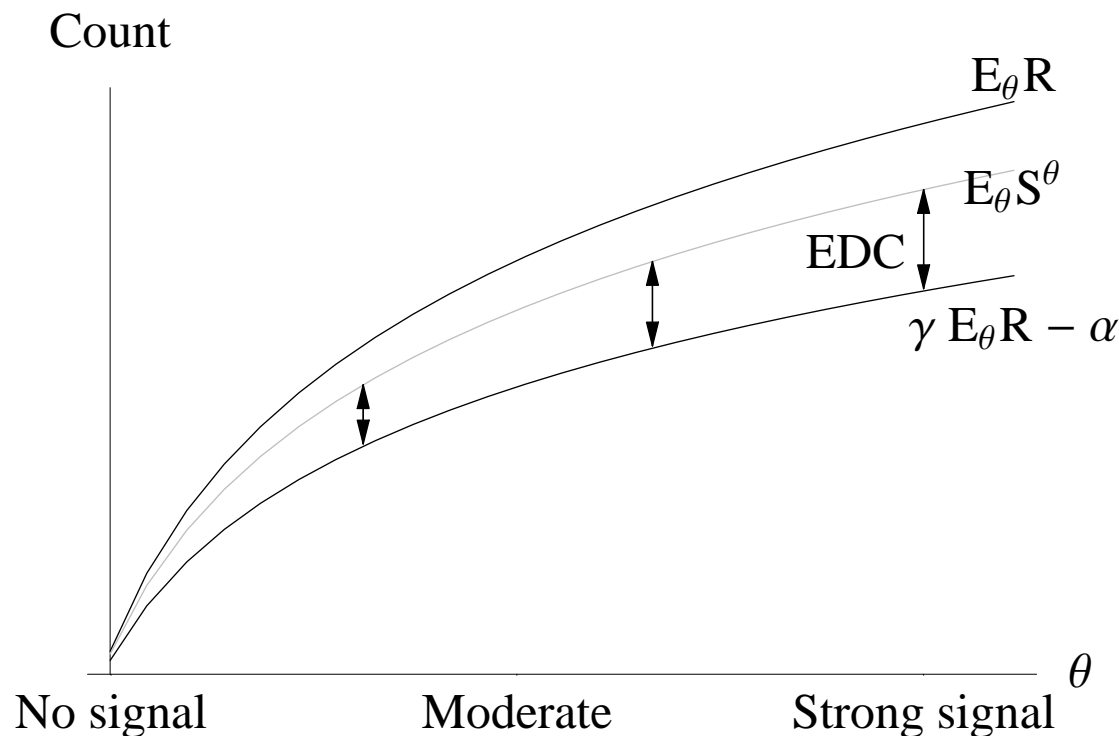
An Alternative to FDR

Count the number of *correct* rejections $S^\theta(m)$ in excess of a fraction of the total number rejected.

$$\text{EDC}_{\alpha,\gamma}(m) = E_\theta[S^\theta(m) - \gamma R(m)] + \alpha, \quad 0 < \alpha, \gamma < 1.$$

Heuristically, α controls FWER and γ controls FDR.

Testing procedure “controls EDC” $\iff \text{EDC} \geq 0$



Comparison to FDR

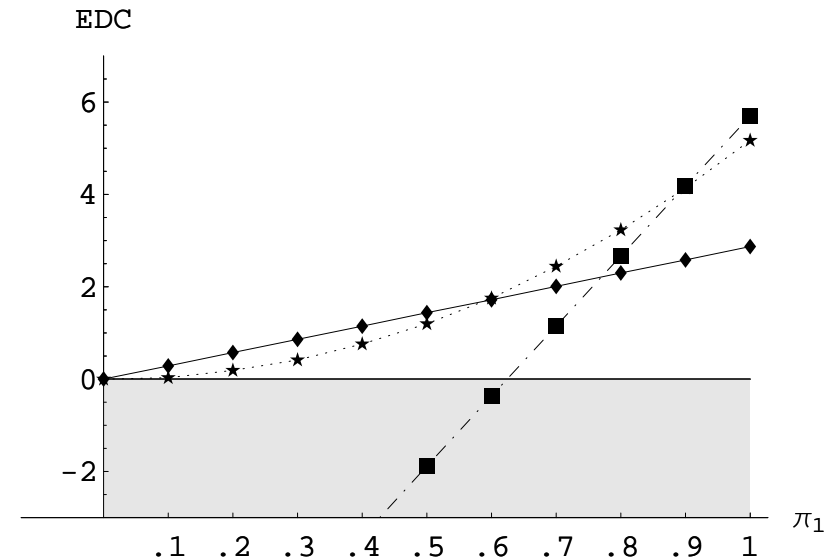
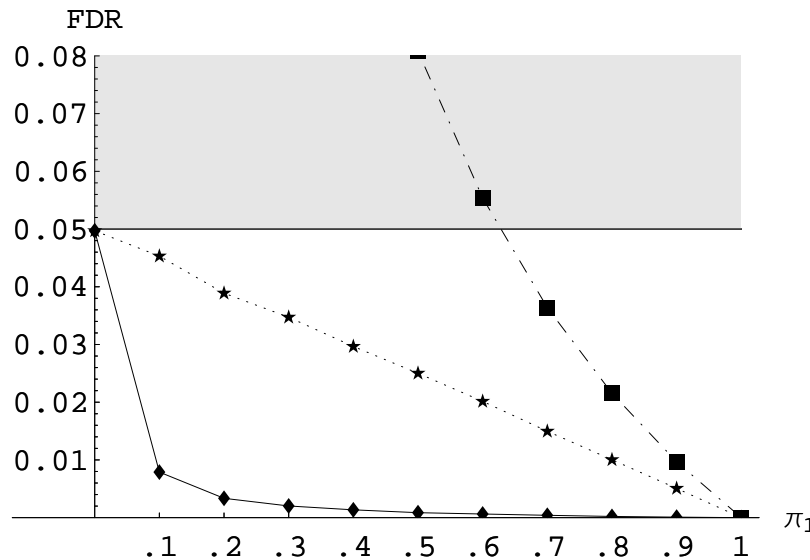
Simulate tests of $m = 200$ hypotheses with

$$\mu_j \sim \begin{cases} 0 & w.p. \quad 1 - \pi_1 \\ N(0, \sigma^2) & w.p. \quad \pi_1 \end{cases}$$

Comparison to FDR

Simulate tests of $m = 200$ hypotheses with

$$\mu_j \sim \begin{cases} 0 & w.p. 1 - \pi_1 \\ N(0, \sigma^2) & w.p. \pi_1 \end{cases}$$



Three methods: naive fixed α level, step-down, Bonferroni.

Alpha-Investing Rules

Sequentially tests hypotheses in style of alpha-spending rules.

Alpha-Investing Rules

Sequentially tests hypotheses in style of alpha-spending rules.

Alpha-wealth

Test next hypothesis H_j with level α_j up to its current alpha-wealth,

$$0 < \alpha_j \leq W(j - 1)$$

Alpha-Investing Rules

Sequentially tests hypotheses in style of alpha-spending rules.

Alpha-wealth

Test next hypothesis H_j with level α_j up to its current alpha-wealth,

$$0 < \alpha_j \leq W(j - 1)$$

P-value

Determines how alpha-wealth changes:

$$W(j) \approx W(j - 1) + \begin{cases} \omega - p_j & \text{if } p_j \leq \alpha_j, \\ -\alpha_j & \text{if } p_j > \alpha_j. \end{cases}$$

Payout

If the test rejects, the rule “earns” payout ω toward future tests. Otherwise, its wealth decreases by α_j .

Alpha-investing Controls EDC

Benefit of EDC

Because it uses differences in counts rather than expected ratio, we can prove alpha investing satisfies EDC criterion.

Theorem

An alpha-investing rule with initial wealth $W(0) \leq \alpha$ and payoff $\omega \leq 1 - \gamma$ controls EDC,

$$\inf_M \inf_{\theta} E_{\theta} \text{EDC}_{\alpha, \gamma}(M) \geq 0$$

for any stopping time M .

Generality

The tests need not be independent, just “honest” in the sense that

$$E(V_j | R_1, \dots, R_{j-1}) \leq \alpha_j \quad \text{under } H_j$$

Comparison to Step-Down Testing

Simulation

Same setup.

200 hypotheses, varying proportion of signal π_1 .

Testing procedures

- Step-down testing
- Alpha-investing: “conservative” and “aggressive”

Two scenarios: What is known?

Order of testing differs in the two cases:

Nothing: random order.

Lots: order of $|\mu_j|$ (not \bar{Y}_j)

Approaches to Alpha-Investing

Conservative

Micro-investing matches performance of FDR.

Suppose FDR rejects k hypotheses, and $\alpha = \omega = W(0)$.

- Test all m hypotheses at Bonferroni level α/m .
Rejects $H_{(1)}$: pay $\alpha = m \times \alpha/m$ to earn α .
- Test other $m - 1$ hypotheses given $p_j > \alpha/m$.
Rejects $H_{(2)}$: pays α and earns α .
- Continue...

Rejects at least those FDR rejects, and perhaps more.

Approaches to Alpha-Investing

Conservative

Micro-investing matches performance of FDR.

Suppose FDR rejects k hypotheses, and $\alpha = \omega = W(0)$.

- Test all m hypotheses at Bonferroni level α/m .
Rejects $H_{(1)}$: pay $\alpha = m \times \alpha/m$ to earn α .
- Test other $m - 1$ hypotheses given $p_j > \alpha/m$.
Rejects $H_{(2)}$: pays α and earns α .
- Continue...

Rejects at least those FDR rejects, and perhaps more.

Aggressive

Spend more α testing leading hypotheses since these should have the most signal — if your science is right.

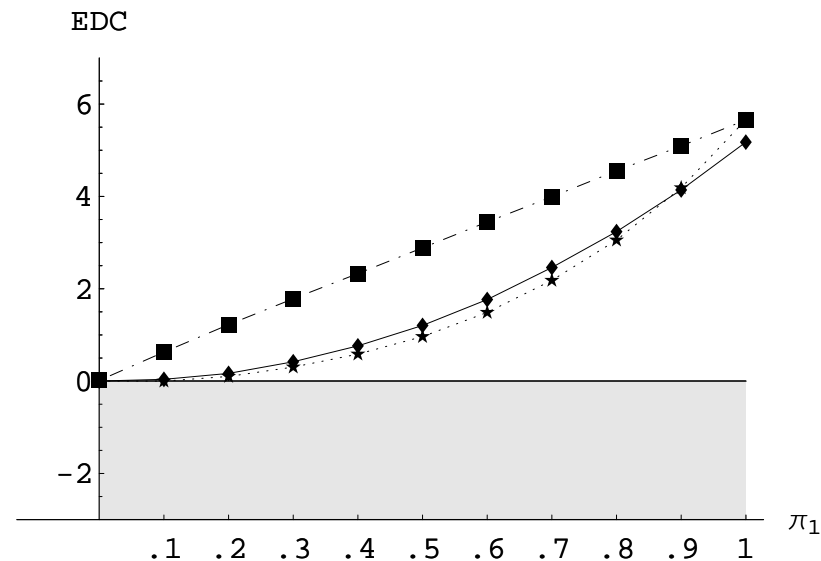
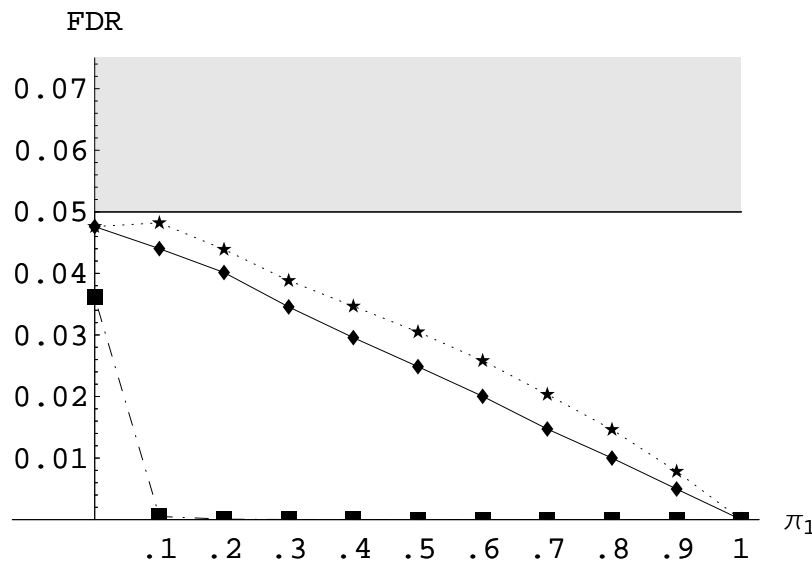
e.g., Set level as

$$\alpha_j = \frac{cW(0)}{j^2}, \quad j = 1, 2, \dots$$

Results of Simulation

Simulate tests of $m = 200$ hypotheses with

$$\mu_j \sim \begin{cases} 0 & w.p. \ 1 - \pi_1 \\ N(0, \sigma^2) & w.p. \ \pi_1 \end{cases}$$



Step-down testing . . .

Alpha-investing (conservative —, aggressive — · —)

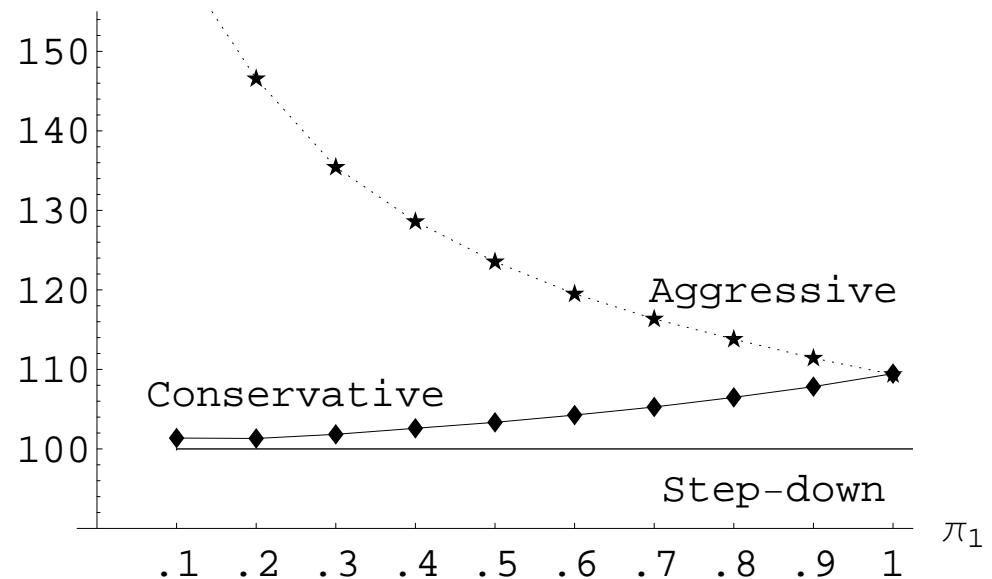
Knowledge = Power

Percentage rejection relative to step-down testing.

$$100 \frac{S^\theta(m, \text{aggressive investing})}{S^\theta(m, \text{step-down})} > 150\%$$

Aggressive alpha investing works well when the information it uses is accurate.

% Rejected vs Step-Down



Going Further

In for a penny, in for a pound

EDC and alpha-investing provide framework for testing a *stream* of hypotheses using *several* investing rules.

Multiple predictor streams

- Start with raw variables as basic stream.
- Once select X_1 and X_2 , try $X_1 * X_2$.

Auction = Multiple alpha-investing rules

- Two rules, one for X 's and second for interactions.
- Each starts with wealth $\alpha/2$.
- Easy to wager more for m X 's than m^2 interactions.
- Best strategy accumulates wealth, makes most choices.

It works!

- Have run auctions through 1,000,000 predictors.
- Finds linear effects *and* high-order interactions.

Discussion

Variable selection

- Importance of p-values (which not all methods have).
- Role for cross-validation.
- Provably as good as class of Bayes estimators.

Discussion

Variable selection

- Importance of p-values (which not all methods have).
- Role for cross-validation.
- Provably as good as class of Bayes estimators.

Strategies for feature creation

- Room for serious improvement
- Substantive expert can order features (chemist)
- Parasitic rules that exploit substantive rules

Discussion

Variable selection

- Importance of p-values (which not all methods have).
- Role for cross-validation.
- Provably as good as class of Bayes estimators.

Strategies for feature creation

- Room for serious improvement
- Substantive expert can order features (chemist)
- Parasitic rules that exploit substantive rules

Multiple testing using EDC

- Adaptive sequential trials
- Universal wagering