Variable Selection in Very Large Models with Rare Events

Bob Stine & Dean Foster

Department of Statistics, The Wharton School

University of Pennsylvania, Philadelphia PA

www-stat.wharton.upenn.edu/~bob

November 27, 2000

- Prediction problems
- Methods for automatic selection of predictors
  - What they are and where they come from
  - How and when they work
- Application in predicting bankruptcy
  - Over-sampling
  - Sparse data with rare events
  - Heteroscedasticity

Some Modern Prediction Problems

Credit modeling, scoring

Can you predict who will declare bankruptcy?

Risk factors for a disease

Which factors indicate risk for osteoporosis?

Direct mail advertising

Who should receive a solicitation for a donation?

Internet/e-commerce

If you bought this CD, which others might you buy?

Financial forecasting

Which factors predict movement in stock returns?

Aren't these great statistics problems?

- Contemporary (e-commerce after all)
- Big dollars at stake
- Role in business strategy
- Lots of computing with rich data structures

Predicting Osteoporosis

Osteoporosis

Degenerative loss of calcium as we age leads to weakened bones that easily fracture in a fall, leading to hospitalization and further complications.

Goal

Estimate likelihood of osteoporosis in women without requiring an *expensive* diagnostic x-ray.

Preference for self-reported measurements.

Data

- Training data (1000 women from clinics)
- *Extensive feature set*
  - Self-reported forms
  - Clinical data obtained by MD
  - Lab measurements (blood sample analysis)
- Validation data (sample of 250, only self-reported)

Trade-off

Sensitivity vs specificity:

Find those at high risk *without*
sending too many for expensive diagnosis.

Predicting Bankruptcy

Goal

Predictive model for personal bankruptcy...

Based on the recent history of an *individual* credit-card holder, estimate the probability that the card holder will declare bankruptcy during the next credit cycle.

Data

- Large data set: 250,000 bank card accounts
- About 350 "basic" predictors (aka, features)
  - Short monthly time series for each account
  - Credit limits, spend, payments, bureau info
  - Demographic background
  - Interactions are important (AC and cash adv.)

67,000 predictors???

Bankruptcy is rare

2,244 bankruptcies in

$$12 \times 250,000 = 3 \text{ million account-months}$$

Trade-off

Profitable customers look risky. Want to lose them?

"Borrow lots of money and pay it back slowly."

Modeling Questions

Structure — What type of model?

- Linear regression model
- Modern forms of regression
  - Additive models (smoothing), interactions (MARS)
  - Neural nets, projection pursuit
  - Regression trees (CART), piecewise fits

Identification — Which predictors to use?

Without a "true model" and with many predictors, emphasis shifts from hypothesis testing and consistency to finding features that are predictive.

Big features ($|t| \approx 10$) are easy to find, so gains are made in locating more subtle features ($|t| \approx 2$):

- Informed combinations e.g. *limit – curr balance*
- Time lags, other transformations e.g. logs, ratios
- Interactions "less-informed" combinations $\Leftarrow$

Search — How do you find them?

- Cannot try all possible solutions.
- Greedy methods
- We still rely on brute force. Getting cheaper!

Traditional Variable Selection Criteria

Regression context

- p potential predictors, n observations
- q non-zero predictors give error variance σ^2:

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

Akaike Information Criterion – AIC

- Unbiased estimate of the prediction MSE
- Identify the order of an autoregression (sequential)
- Selection criterion (orthogonal regression)

$$\text{Pick } X_j \quad \Leftrightarrow \quad |t_j| > \sqrt{2}$$

- Picks many predictors: 16% when no signal

C_p, leave-one-out cross-validation

Equivalent to AIC for large sample sizes.

Bayesian Information Criterion – BIC

- Estimates the Bayes factor, ratio of posterior $\Pr\{H_k\}$
- Selection criterion (orthogonal regression)

$$\text{Pick } X_j \quad \Leftrightarrow \quad |t_j| > \sqrt{\log n}$$

- Parsimonious if $n \gg p$, promiscuous if $n \ll p$.

Hard Thresholding and Bonferroni

Minimax variable selection

- Which predictors minimize maximum prediction MSE?

$$\min_{\hat{q}} \max_{\beta} E \left\| Y - \hat{Y}(\hat{q}) \right\|^2$$

- Answer: (disappointing) Constant risk — pick them all!

Competitive analysis

- Which predictors minimize *ratio* of prediction MSEs?

$$\min_{\hat{q}} \max_{\beta} \frac{E \left\| Y - \hat{Y}(\hat{q}) \right\|^2}{q\sigma^2}$$

- Answer: (Donoho&Johnstone, Foster&George 1994)

$$\text{Pick } X_j \quad \Leftrightarrow |t_j| > \sqrt{2 \log p}$$

Heuristic for hard thresholding

- It's almost Bonferroni! ($\sqrt{2 \log p}$ is a bit less strict)
- Fisher's (1927) test for the max of periodogram.
- If have a sample of $X_1, \ldots, X_p \sim N(0,1)$ then

$$\Pr\left\{\max(|X_1|, \ldots, |X_p|) > \sqrt{2 \log p}\right\} \to 0.$$

Data Compression

File compression

Disk compression utilities: WinZip, Stacker, Stuffit.

How do they work?

How to compress a file of characters into a sequence of bits (0's and 1's) without losing information (lossless)?

Sample problem

Alphabet of 4 characters: $a,\ b,\ c,\ d$.

What would you need to know in order to compress a file of these characters?

Question rephrased

View file as a sequence $Y_1,\ Y_2,\ \ldots,\ Y_n$ of *iid* discrete r.v.'s,

$$Y_1,\ Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} P(y)\ .$$

Let $\ell(y)$ denote code length for y. What is the smallest compressed file length (on average),

$$\min_{\ell} E \sum_{i=1}^{n} \ell(Y_i) = n \ \min_{\ell} E\, \ell(Y_1)\ ?$$

What code achieves this limit?

Examples of Codes

Canonical question

How to compress a file to as small a size as possible?

Answer: Use short symbols for letters of high probability.

Two codes

- Code I: fixed-length, like ASCII but 2 bits each
- Code II: variable-length, matched to probability

Symbol y	Code I	Code II	$p(y)$
a	00	0	$1/2 = 1/2^1$
b	01	10	$1/4 = 1/2^2$
c	10	110	$1/8 = 1/2^3$
d	11	111	$1/8 = 1/2^3$

Example

String	Code I	Code II	P(String)
baa	010000	1000	$\frac{1}{4}\frac{1}{2}^2 = \frac{1}{2^4}$
dad	110011	1110111	$\frac{1}{8}\frac{1}{2}\frac{1}{8} = \frac{1}{2^7}$

Both are prefix codes

No delimiters, even though Code II varies in length.

Codes = Probability

Optimal codes

For a (discrete) random variable

$$Y \sim P(Y) \ ,$$

the avg length of any code must be at least the *entropy*

$$E\,\ell(Y) \geq E\,\log_2 \frac{1}{P(Y)}$$

Connection to statistics

Every code implies a probability model:

$$\text{code for } y \text{ has } \ell(y) \text{ bits } \iff P(y) = 2^{-\ell(y)} \ .$$

Proper distributions

The Kraft inequality implies that

$$\sum_y 2^{-\ell(y)} \leq 1$$

so that the implied probabilities sum to less than 1,

$$\sum_y 2^{-\ell(y)} = \sum_y P(y) \leq 1.$$

Example: Bernoulli Coding

Beroulli data

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} B(\theta), \quad 0 \le \theta \le 1 \, .$$

How well can these be compressed?

Choose the code that maximizes the probability, or minimizes the negative log-likelihood:

$$\begin{aligned} L_\theta(Y) &= -\log_2 P_\theta(Y_1, \ldots, Y_n) \\ &= S \log_2 \frac{1}{\theta} + (n - S) \log_2 \frac{1}{1-\theta} \, , \quad S = \sum_i Y_i \end{aligned}$$

Examples

The number of bits to encode $Y_1, \ldots, Y_n$ for two cases

"random", half successes	$S = n/2$	$L_{1/2}(Y) = n$
rare, one success	$S = 1/n$	$L_{1/n}(Y) = \log_2 n$

Boolean entropy function

What's the parameter?

Cannot decode the message

Prior results *only* apply when both sender and receiver know the encoding parameter value!

$L_\theta(Y)$ is the length for *only* the data, not θ.

Two-part code (Davisson, 1973)

Add a prefix that encodes the parameter.

How to encode the parameter?

Encode S directly requires a total length of

$$L(Y) = \log_2 n + L_{S/n}(Y)$$

Too long. We can do better!

Round the parameter (Rissanen, 1978)

Round S to grid with standard error spacing, say $\tilde{S}$, then total code length is

$$L(Y) = \tfrac{1}{2} \log_2 n + L_{\tilde{S}/n}(Y) + \text{divergence}$$

but divergence is less than one or two bits.

Bayesian tie Code for parameter suggests priors...

$$L(Y) = \text{ Length parameters } + \text{ Length data}$$

$$\Rightarrow \quad P(Y) = P(\theta)\ P(Y|\theta)$$

Coding and Model Selection

Idea

"Good" models yield codes that compress the observed data into a shorter **message** than bad models.

Maximum likelihood

Given a parametric model $P_\theta(Y)$ for the response data Y,

$$\begin{aligned} \min_\theta(\text{code length for data}) &= \min_\theta \log 1/P_\theta(Y) \\ &= \max_\theta P_\theta(Y) \end{aligned}$$

Rissanen's MDL

Pick model obtaining shortest encoding of **itself** and **data**. The message must encode *both* the slopes and the residuals (everyone knows all of the X's):

$$\begin{aligned} \text{Message length} &= L(\text{parameters}) + L(\text{data}) \\ &= L(\text{parameters}) + c \text{ log-likelihood} \end{aligned}$$

$\Rightarrow$ penalized likelihood.

Regression Model

Usual model

From among a collection of

$$p \text{ possible predictors } X_1, \ldots, X_p,$$

fit a model with q predictors of the form

$$Y = \beta_{j_1} X_1 + \cdots + \beta_{j_q} X_q + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

Subset/selection indicator

Let

$$\gamma = (\gamma_1, \ldots, \gamma_p)$$

denote a sequence of p 0's and 1s. Denote a subset of β by

$$\beta_\gamma \quad \text{defined by} \quad \beta_j \neq 0 \iff \gamma_j = 1$$

Simplifying assumptions

- $p \leq n$ **orthonormal** predictors with $X_j' X_j = 1$.
- $\sigma^2 = 1$ is known, implying $\hat{\beta}_j$ is a z-score.
- n and X are known and fixed, so are not coded.

Two-Part Codes for Regression

Model prefix encodes $\hat{\gamma}$ and $\hat{\beta}$

1. Encode for the selection indicator

$$\gamma = (1, 0, 0, \ldots, \gamma_j, \ldots, 1)$$

2. Encode for rounded $\hat{\beta}_\gamma$ estimates

Overall code

$$L(Y) = L(\hat{\gamma}) + L(\hat{\beta}_{\hat{\gamma}}) + \underbrace{L_{\hat{\beta}_\gamma}(Y)}_{\text{residual SS}}$$

Goal

Find the shortest message (ie, highest probability)

Protection

Automatic penalty for over-fitting:
the more predictors included in the model, the longer the prefix since more estimates must be added.

$$\text{Add } \hat{\beta}_j \quad \Longleftrightarrow \quad (\downarrow \text{ Residual SS}) > (\uparrow L(\text{parameters}))$$

Two-Part Codes for Regression, Examples

Context

1. Select from a collection of p orthonormal predictors.
2. Pretend error variance $\sigma^2 = 1$ is known.

$\rightarrow$ know $SE(\hat{\beta}_j) = 1$.

AIC code

Send $\hat{\gamma}$ as p bits, then send $\hat{\beta}_j$. When is a predictor added?

$$\begin{array}{rcl} \text{improvement in likelihood} & > & \text{length for parameter} \\ \dfrac{z_j^2}{2 \log 2} & > & 2 \log_2 z_j \end{array}$$

Add $\hat{\beta}_j$ if $|z_j| > 2.xx$

RIC code

Send $\hat{\beta}_j$ to the receiver as a pair

$$(j, \hat{\beta}_j) \quad \Rightarrow \quad (\log_2 p, 1 + 2 \log_2 z_j) \text{bits}$$

When is a predictor added?

$$\begin{array}{rcl} \dfrac{z_j^2}{2 \log 2} & > & 1 + \log_2 p + 2 \log_2 z_j \end{array}$$

Add $\hat{\beta}_j$ if $|z_j| > \sqrt{2 \log p}$.

Adaptive Criteria

Why assume a fixed number of predictors q?

Method	Code for $\hat{\gamma}$	Expecting in model
AIC	p bits	half of X_j
RIC	index=$\log p$ bits	one X_j

Simple adaptive solution (Foster and Stine, 1997)

Compress γ using any "universal" method, coding $\hat{\gamma}$ in about

$$\frac{1}{2}\log p + L_{\tilde{q}/p} \text{ bits}$$

Resulting criterion adds $\hat{\beta}_j$ if

$$|z_j| > \sqrt{2\log p/q}$$

Other paths to similar adaptive selection

- Empirical Bayes: estimate the distribution of γ.
- Half-normal plots
- Simes method, step-up testing

Discussion

Sources of prediction error

- Include an extraneous predictor
- Omit a useful predictor
- Random estimation error

Weakness of "Bonferroni"

Includes too few predictors: prediction error dominated by omitting useful predictors.

Adaptive variable selection (a.k.a. multiple testing)

- Bonferroni unpopular because of low power.
- Simes method – step-up/step-down tests:

$$|t_{(1)}| \geq |t_{(2)}| \geq \cdots \geq |t_{(p)}|$$

1. Compare $t_{(1)}$ to $\sqrt{2 \log p}$
2. Compare $t_{(2)}$ to $\sqrt{2 \log p/2}$
3. ... compare $t_{(q)}$ to $\sqrt{2 \log p/q}$

$\Rightarrow$ Once you find one variable, easier to add more.

- Related to empirical Bayes and info theory

Prediction error of adaptive model is within a factor of the prediction error of "expert" model that knows the true β_j, but not the coordinates!

Thresholding Summary (Orthogonal)

Criterion	Threshold on z Scale	Origin
Least squares	0	Gauss, minimax
AIC	$\sqrt{2}$	Unbiased diverg (Akaike 73)
C_p	$\sqrt{2}$	Pred error (Mallows 73)
BIC	$\sqrt{\log n}$	Bayes (Schwarz 78)
MDL	$\sqrt{\log n}$	Coding (Rissanen 83)
RIC	$\sqrt{2 \log p}$	Relative risk (F & G 94)
hard threshold	$\sqrt{2 \log p}$	Minimax (D & J 94)
Adaptive	$\sqrt{2 \log q/p}$	E Bayes (Foster&George 96) Mult hypoth (Benjamini 97) Coding (Foster&Stine 97)

Predicting Personal Bankruptcy

Goal

Identify customers at "high" risk of declaring bankruptcy.

"High" might mean Pr{Bankrupt next month} = 0.10.

Data

- Records for $n = 250,000$ card holders
- Demographic data (e.g., location, home ownership)
- Year of longitudinal data
  - Some gathered monthly, others quarterly and annual
- Derived data
  - Interactions (regional differences, nonlinear)
  - Missing data

Linear model

Consider selecting from

$$p = \mathbf{67,000} \quad \text{candidate predictors}$$

including interactions and missing indicators.

Needle in the haystack

Bankruptcy is a rare event in our data:

2,244 events in 3,000,000 months of data

Variable Selection in Bankruptcy Model

Cross-validation Split sample into two parts

- 20% for estimation ($n = 600,000$)
  - 450 bankruptcy events
- 80% for validation ($n = 2,400,000$)
  - 1,794 remaining bankruptcy events

Estimation

- Over-sample
  Use all 450 bankrupt cases, but only 2.5% of the rest, for an estimation sample of about 15,000.
- Adaptive selection
  Comparing $|t|$ to $\sqrt{2 \log p/q}$.

Objective

Predict the validation sample.

Hope

Validation should not be necessary.

Would be able to find a better model with *all* 2,244 bankrupt cases used for estimation rather than reserved for validation.

Effects of Over-sampling

Over-sample bankrupt events

- Use *all* bankrupt events, 2.5% of rest.
- Unlike logistic regression, linear regression slopes are *biased unless we adjust* for sampling wts w_i.

Weighted least squares estimator

$$\hat{\beta}_w = (X'WX)^{-1}X'WY, \quad W = \text{diag}(w_i)$$

Standard error

$$\begin{aligned} \text{Var}(\hat{\beta}_w) &= (X'WX)^{-1}\,(X'W\ \text{Var}(Y)WX)\,(X'WX)^{-1} \\ &= \sigma^2(X'WX)^{-1} \end{aligned}$$

IF

$$\text{Var}(Y) = \sigma^2 W^{-1}$$

which is not likely since w_i are sampling weights.

Homoscedastic case

Assuming constant variance σ^2, left with

$$\text{Var}(\hat{\beta}_w) = \sigma^2(X'WX)^{-1}\,(X'W^2X)\,(X'WX)^{-1}$$

which greatly complicates the *search* for predictors.

Dare We Assume Constant Variance?

Discrete data

- Response Y is 0/1 indicator with most $Y = 0$.
- Many predictors are also 0/1:
  Indicators, missing data, interactions

Stylized testing problem (assume $n_0 \gg n_1$)

$n_0 : Y_{0i} = 0$ at $X = 0$ $\quad$ $n_1 : Y_{1i} \sim N(0, 1)$ at $X = 1$

Correct test for mean shift One-sample t

$$t_1 = \frac{\sqrt{n_1}\,\overline{Y_1}}{s_1}$$

Two-sample t test *assuming* homoscedastic

$$t_2 = \frac{\overline{Y_1} - \overline{Y_0}}{s\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} = t_1 \times \frac{\sqrt{n_0}}{\sqrt{n_1}}$$

and test statistic is inflated since $n_0 \gg n_1$

'Robust' Variance Estimate

White's estimator

$$\mathrm{Var}(\hat{\beta}_w) = (X'WX)^{-1}\,(X'W\ \underbrace{\mathrm{Var}(Y)}\ WX)\,(X'WX)^{-1}$$

Estimate $\mathrm{Var}(Y)$ using the squared residuals from fitted model

$$E = \mathrm{diag}(e_i), \quad e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \cdots - \hat{\beta}_q X_q$$

Obtain SE from

$$\mathrm{var}(\hat{\beta}_w) = (X'WX)^{-1}\,(X'W\ E^2\ WX)\,(X'WX)^{-1}$$

(Use the binomial variance form in this case?)

Further complication Intermediate output...

Term	Estimate	Homo SE	t	Hetero SE	t
$X_{20} * X_{317}$	0.9992	0.01970	51	0.00003	29857
$X_{25} * X_{348}$	0.9992	0.01970	51	0.00003	29857

What happened?

$Y = 1$	300	2
$Y = 1$	15000	0
	$X = 0$	$X = 1$

Rate 1/30

Stylized Testing Problem

Stylized problem $n_0 \gg n_1$

$$n_0 : Y_{0i} = 0 \text{ at } X = 0 \qquad n_1 : Y_{1i} \sim N(0,1) \text{ at } X = 1$$

Standard two-sample test

Inflated t-statistic since data lacks constant variance and observe many more in group at $X = 0$.

One-sample test

Fails for *sparse data* when only one observation in group at $X = 1$.

Conservative one-sample test

Estimate assuming no signal using conservative estimate of variance

$$t_1' = \frac{\sqrt{n_1}\,\overline{Y_1}}{s_1'}, \qquad s_1'^2 = \frac{\sum_i Y_i^2}{n_1}$$

Current Procedure

Estimator

Use sampling weights, but not heteroscedastic weights:

- WLS down-weights observations with large variance. which tends to occur for "large" $\hat{Y} > 0.25$, say
- WLS down-weights observations of most interest.

Standard error

Recognize the heteroscedasticity when estimating SE.

Search procedure

Stepwise search, with adjustment for survey weight to get correct forward selection...

- Sort omitted predictors by change in residual SS and note the associated t ratio.
- Add variable with most explanatory power **if** "look ahead" SE is significant using adaptive threshold.
- Estimate SE using current, not updated residuals. i.e., to evaluate at step k, estimate SE using residuals from step $k-1$.

$$\text{Var}(\hat{\beta}_w) = (X_k' W X_k)^{-1} \, (X_k' W \, E_{k-1}^2 \, W X_k) \, (X_k' W X_k)^{-1}$$

- Bonus: easier to compute look-ahead selection step.

Validation Results

Sums of squares for in-sample and validation

Validation Results

"Lift" chart

Validation Results

"Lift" table

Lower	Cum Count	Cum % BR
-0.10	103	0.0006
-0.05	1199	0.0022
0.00	1864748	0.0767
0.05	2325087	0.5258
0.10	2330767	0.7111
0.15	2331750	0.8057
0.20	2332064	0.8768
0.25	2332208	0.9199
0.30	2332280	0.9451
0.35	2332325	0.9574
0.40	2332345	0.9675
0.50	2332372	0.9838
0.60	2332379	0.9899
0.70	2332383	0.9961
0.80	2332385	0.9983
1.10	2332385	0.9989

Validation Results

Calibration chart

Interpreting Bankruptcy Model

Coefficients from fitted model

Arranged by order of variable indices reveals same "base terms" appearing in selected terms:

Term	Estimate	Homo SE	t	Hetero SE	t
$X_{59} * X_{263}$	0.002063	0.00022	9.20	0.00036	5.66
$X_{59} * X_{292}$	0.000460	0.00004	10.99	0.00012	3.89
$X_{59} * X_{284}$	0.000275	0.00003	10.07	0.00006	4.80
$X_{215} * X_{292}$	0.000021	0.00000	13.41	0.00000	6.15
$X_{284} * X_{292}$	0.002664	0.00011	25.30	0.00058	4.62
$X_{292} * X_{298}$	-0.007760	0.00066	-11.77	0.00291	-2.67

"Interpretation" of model

- Combination of terms suggests *multiplicative* model.
- Logistic regression requires only base terms – the interactions are no longer needed!

Discussion

Adaptive variable selection

- Powerful technique, strong theoretical basis
- Crucial role of standard error estimates
- Adaptive cut-off finds structure Bonferroni misses
- Significant terms also help in validation

Implications for practice

- Automated search with good validation properties
  - Use more to estimate
- Supplement to "manual" analysis

Next steps Better searching ...

- Grow the model using all bankruptcy data
- Improved backward elimination
- Searching for other interactions (detecting multiplicative structure with better selection method)
- Logistic regression as base model.