

**Yale University**

---

**From the Selected Works of Shuangge Ma**

---

June 6, 2009

# A Tale of Two Streets: Incorporating grouping structure in high dimensional data mining

Shuangge Ma, *Yale University*



Available at: <https://works.bepress.com/shuangge/8/>

# A Tale of Two Streets

Incorporating grouping structure in  
high dimensional data mining

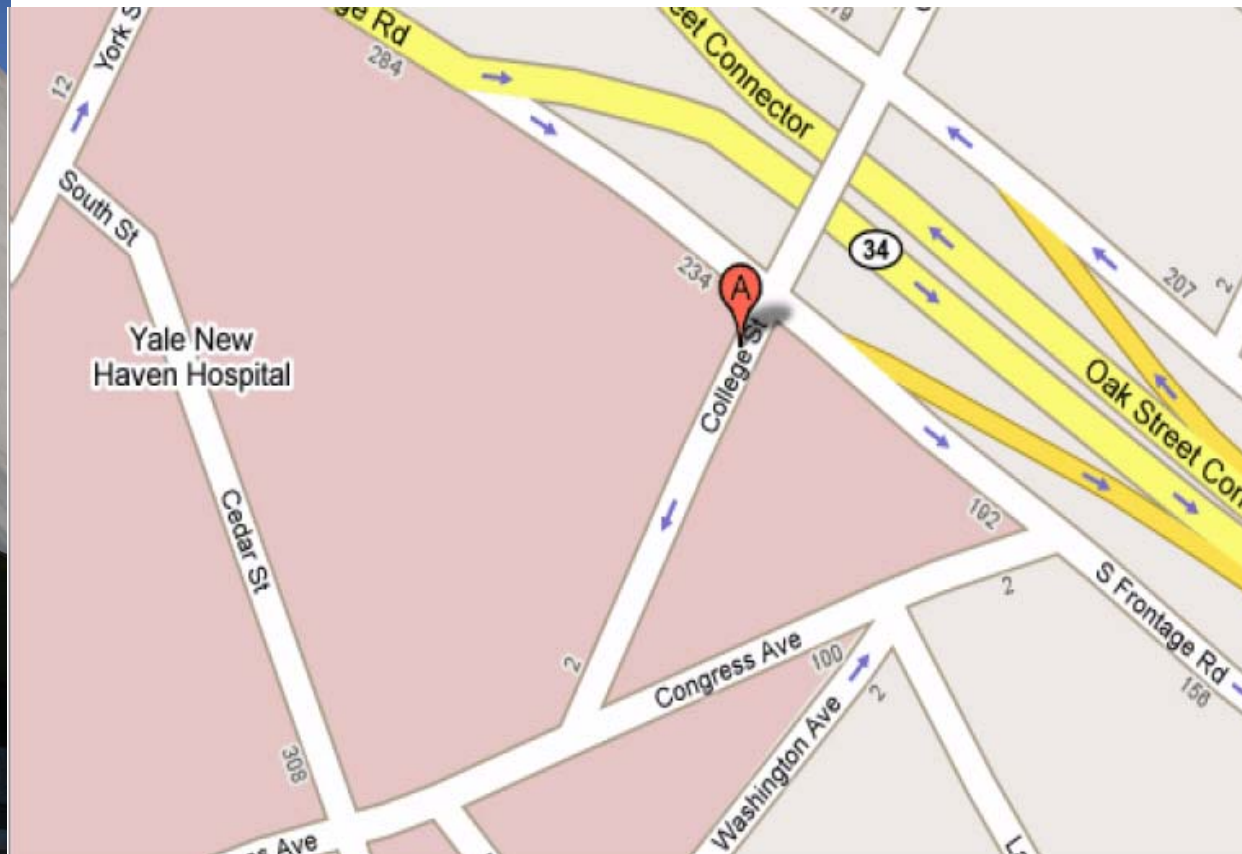
马双鸽

Assistant Professor

School of Public Health, Yale University

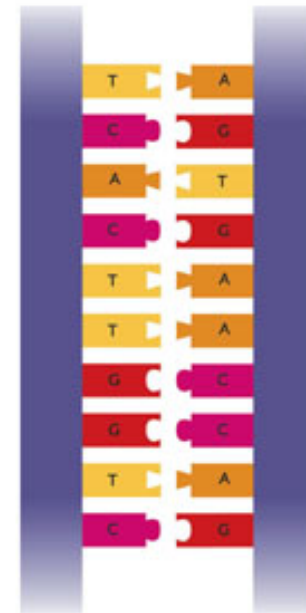
# College Street (New Haven, CT)

## School of Public Health, Yale University



# Cancer Informatics

- Genetic mutations and defects are responsible for cancer occurrence and progression
- Our research: Predict occurrence of cancer using genetic/genomic measurements obtained through genome profiling



# Cancer Informatics

- Occurrence of cancer: Yes/No

## Human Genome:

- 23 chromosomes
- 20,000+ genes
- 250,000+ SNPs (single nucleotide polymorphisms)
- 6,000,000,000 basepairs

# Cancer Informatics:

## A high dimensional data mining problem

- Consider the 20,000+ genes
- Our problem: Binary classification/prediction
- Data characteristics:
  - High dimensionality
  - Most genes are just noises
  - Some genes work together

CITY  
EDITION

London  Herald

LATE  
PRICES

No. 36457

FRIDAY 25th OCTOBER 1929

14

# WALL STREET CRASH!

## Black Thursday in America Stocks Plunge and Eleven Commit Suicide

Panic selling in the New York Stock Market caused an extraordinary atmosphere of chaos and panic throughout the market with orders from their brokers to sell at any price.

Market sales in the early morning continued at an extraordinary atmosphere of chaos and panic throughout the market with orders from their brokers to sell at any price.

In the situation created while back, the market broke down entirely as the word broke down completely for the market. Orders were issued for any price and the value of some companies failed during the course of the morning.

The huge early morning sales in the market were a result of orders to sell from the London market and the London market broke down completely as the word broke down completely for the market. Orders were issued for any price and the value of some companies failed during the course of the morning.

### Cash Selling

A crisis morning in New York's trading markets was followed by the collapse of 100 Morgan & Co and prices plummeted rapidly in the afternoon as the market broke down completely. The London market broke down completely as the word broke down completely for the market. Orders were issued for any price and the value of some companies failed during the course of the morning.

Washington also issued a statement to the effect that "significant business was still fundamentally sound and that the market was affected only by a temporary panic." The following morning the market was still in a state of confusion and the value of some companies failed during the course of the morning.

### Bank

Despite the selling mood of the great and good, the day's action was dominated by the actions of the market and the value of some companies failed during the course of the morning.

Parking lot filled with cars around the state of George Washington in New York.

## What Went Wrong?

The initial reaction was of a panic selling which, it is believed, led to a further decline in the market.

### 4pm

The day's action was dominated by the actions of the market and the value of some companies failed during the course of the morning.

market, values of stocks fell sharply and the market broke down completely.

Despite the selling mood of the great and good, the day's action was dominated by the actions of the market and the value of some companies failed during the course of the morning.

City of New York. The market broke down completely as the word broke down completely for the market. Orders were issued for any price and the value of some companies failed during the course of the morning.

### Speculators

Speculators were also affected by the market and the value of some companies failed during the course of the morning.

100s, are speculators have finished and the market broke down completely. The London market broke down completely as the word broke down completely for the market. Orders were issued for any price and the value of some companies failed during the course of the morning.

### OTHER NEWS

London News to be held in the Government tomorrow due to the fact that the market broke down completely.

News to be held in the Government tomorrow due to the fact that the market broke down completely.

News to be held in the Government tomorrow due to the fact that the market broke down completely.

News to be held in the Government tomorrow due to the fact that the market broke down completely.

### COLEMAN'S "WINGED" H



# Are you going to default?

[illegible][illegible][illegible][illegible][illegible]



# Mortgage (as well as many other business sectors)

- Default: Yes/No

Applicant's information:

1. Financial situation

2. Education

3. Family members' information

..... A long list of variables  
from questionnaires

# Cancer/Mortgage

## A high dimensional data mining problem

- Our problem: Binary classification/prediction
- Data characteristics:
  - High dimensionality
  - Most measurements are just noises
  - Some measurements work together

# Cancer/Mortgage

## A high dimensional data mining problem

- Our problem: Binary classification/prediction  
Construct binary classification models
- Data characteristics:
  - High dimensionality
  - Most measurements are noises  
Variable selection to reduce dimension & remove noises
  - Some measurements work together  
Account for the grouping structure!!!

# Penalized Regularization in Data Mining

- Binary outcome  $Y = 0, 1$
- Covariate  $X = (X^1, X^2 \dots X^K)$
- $X^i$  is the  $m_i$  dimensional measurements of the  $i^{th}$  group (of variables)
- For example:  $X^i = (\text{year of education; highest degree; area of education...})$

# Penalized Regularization in Data Mining

- Assume the logistic regression model  $Pr(Y = 1|X) = \text{logit}(\sum_{i=1}^K \beta^i X^i)$ .

Denote  $\beta = (\beta^1 \dots \beta^K)$

- Assume  $n$  iid observations

- Log-likelihood  $R_n(\beta) = \sum_i Y_i \log \left( \frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)} \right) + (1 - Y_i) \log \left( \frac{1}{1 + \exp(\beta X_i)} \right)$

# Penalized Regularization in Data Mining

- Penalized regularization  $\hat{\beta} = \operatorname{argmax} R_n - \lambda_n \operatorname{Pen}(\beta)$
- $\operatorname{Pen}(\beta) = \sum_i \|\beta^i\|_2^\gamma$  with  $0 < \gamma < 1$

So what we do is:

- We maximize the likelihood function
- But, we add a penalty, which gives us a **penalized** likelihood function
- Most recently, penalized data mining is one of the hottest areas

So, what's the deal with  $Pen(\beta) = \sum_i ||\beta^i||_2^2$

- It is called the “group bridge” penalty
- When sample size < number of covariates, it can lead to a unique estimate → **Regularization**
- Many of the estimates are zero → **Selection**
- The penalty is on the “group norm”. Thus, it can account for the **grouping structure**



# We bother with group bridge because ...

- It is estimation consistent
- It is variable selection consistent, meaning: we can separate "important variables" from "noisy ones"
- So in the future, instead of asking for a huge number of questions, we can ask only a few
- It saves time and cost

# We bother with group bridge because ...

- Variable selection consistency + estimation consistency. **The Oracle is with us!**



# We bother with group bridge because ...

- Many other data mining methods
  - cannot account for the grouping structure [you lose efficiency]
  - Although they may work well with a few datasets, without statistical backup, it is hard to say how well they work in general

## Some non-trivial tasks

- Computing [the objective function is not concave]
  - An iterative approach has been developed
- Tuning parameter selection
  - Cross validation
- Evaluation of significance and reproducibility
  - Occurrence index

## All right. Now the real question:

- How much do we really gain in practice?
- No one likes a toy that works only “in theory”
- In extensive numerical studies, the group bridge can beat alternatives:
  - A few percent in classification error
  - A lot in variable selection

# What's next

- On College street: let's cure cancer!
  - Diagnosis and prognosis gene signatures are being constructed for lymphoma, breast cancer, leukemia ...
- On Wall street: let's make money! Or at least lose less ☹
  - Efficient penalized data mining tools are being introduced

# Acknowledgements

- Conference organizers and RenMin University
- Funding support from NIH/NSF, USA
- Main collaborators: Dr. Jian Huang, University of Iowa



Thanks for your  
attentions!