Data Mining for Outliers

Ruben Zamar

Department of Statistics

University of British Columbia

Vancouver, Canada

William J. Welch
Fei Yuan
Yi Lin
Hui Shen
Guohua Yan
Mohua Podder

OUTLINE

- Robust Data Mining?
- Finding Homologous Proteins
- Finding the Needle Outside the Haystack

TYPICAL STEPS IN DATA MINING

TYPICAL STEPS IN DATA MINING

➤ DEFINING THE MINING GOAL

TYPICAL STEPS IN DATA MINING

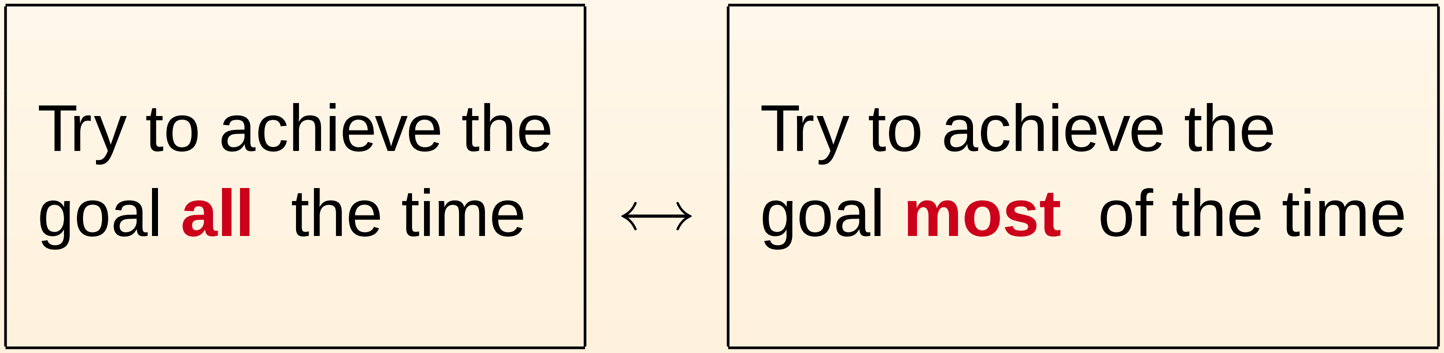
- ➤ DEFINING THE MINING GOAL

- ➤ CHOOSING A SCORING SCHEME

TYPICAL STEPS IN DATA MINING

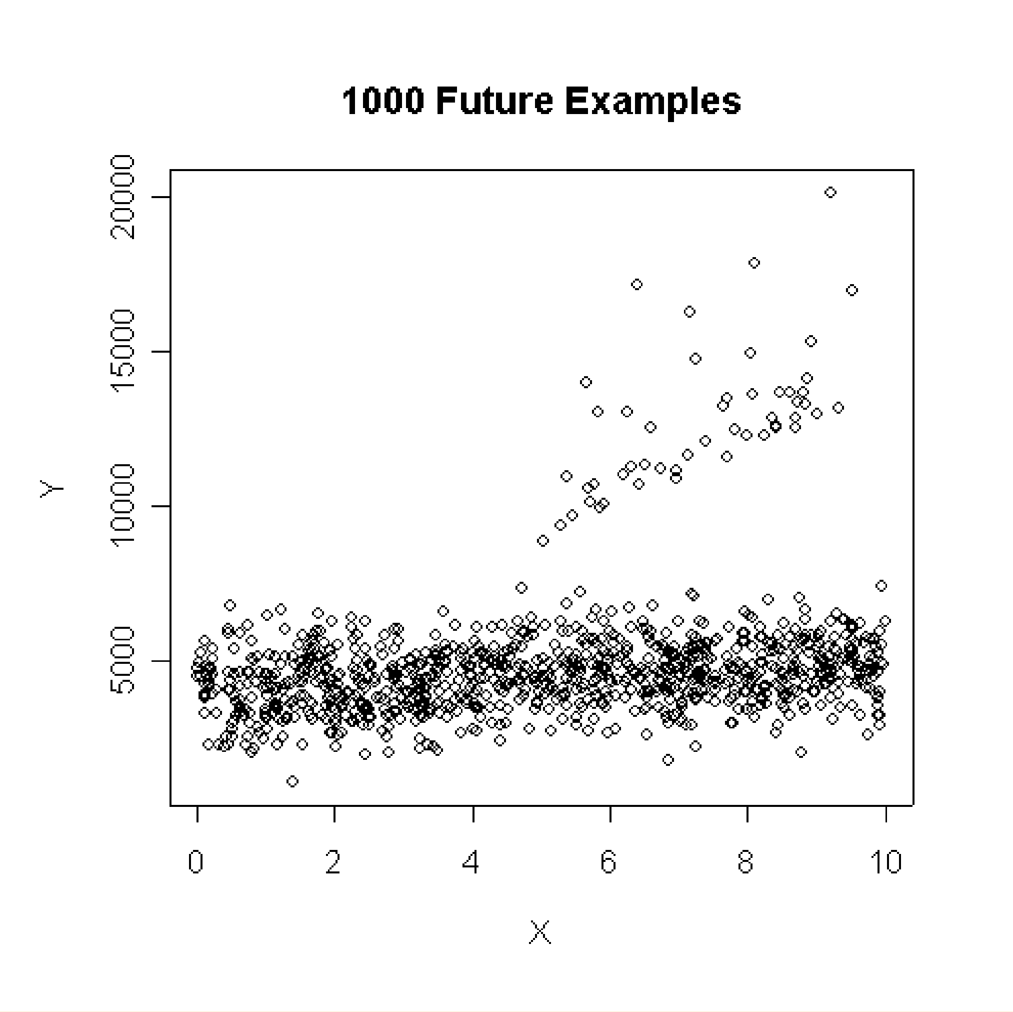
- ➤ DEFINING THE MINING GOAL
- ➤ CHOOSING A SCORING SCHEME
- ➤ NUMERICAL IMPLEMENTATION

A ROBUSTNESS ISSUE

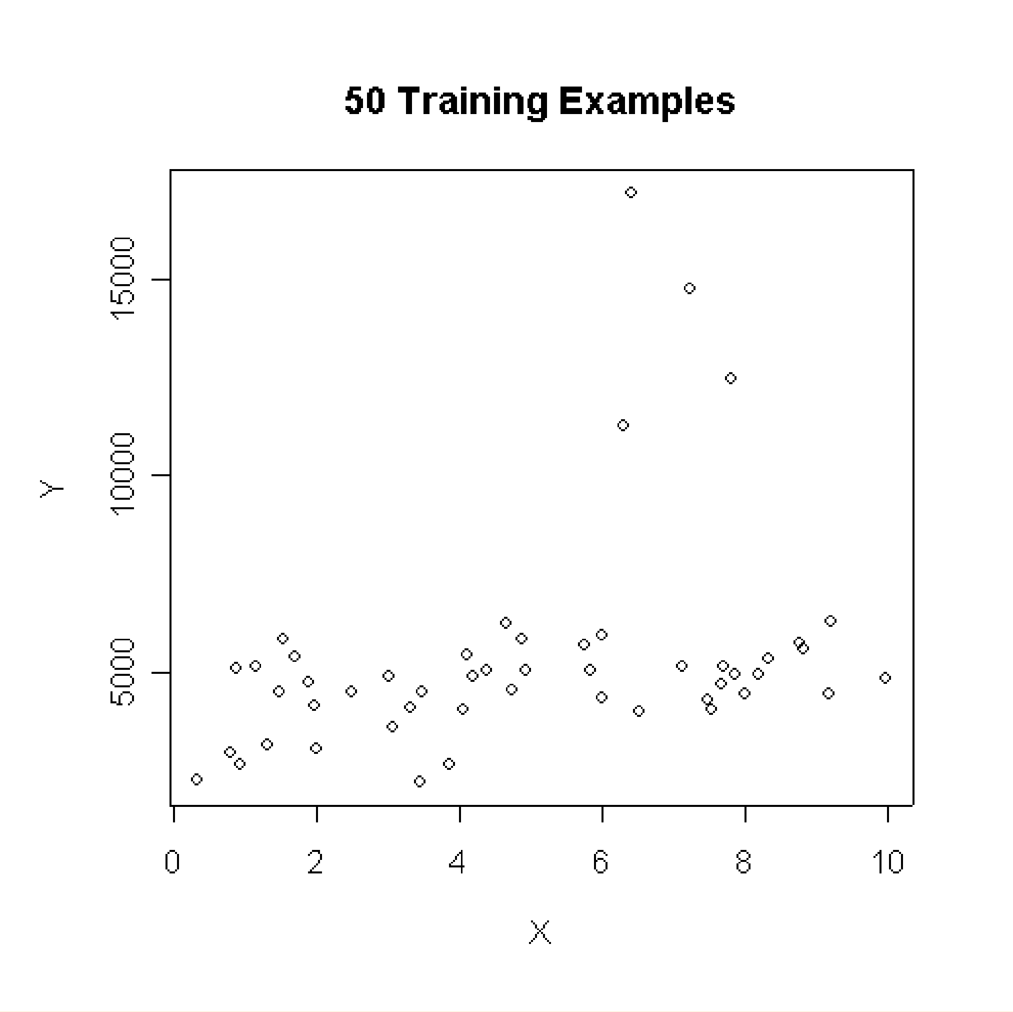


Try to achieve the goal **all** the time ⟷ Try to achieve the goal **most** of the time

TARGET POPULATION



TRAINING SAMPLE



LINEAR PREDICTION

LINEAR PREDICTION

➤ Prediction of Y using X

LINEAR PREDICTION

➤ Prediction of Y using X

➤ Try to perform well on **all future predictions**

LINEAR PREDICTION

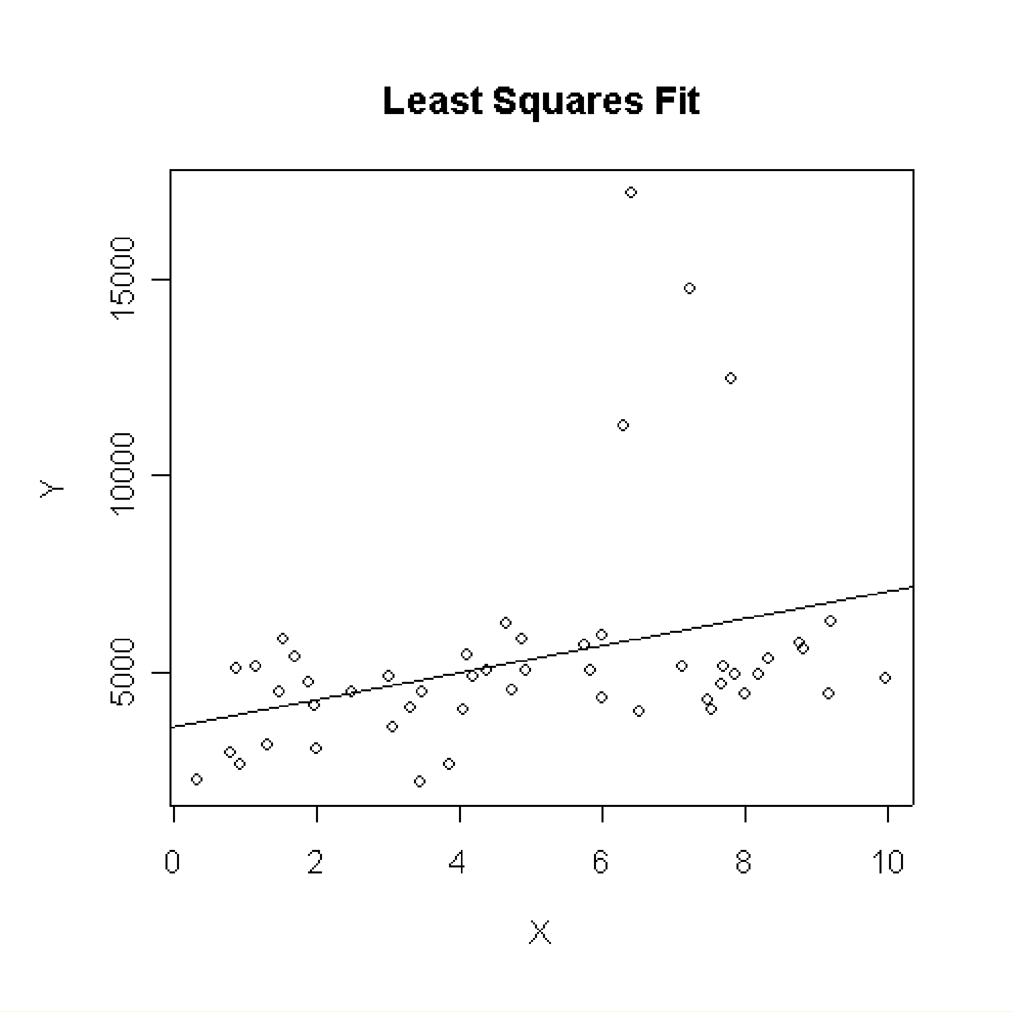
➤ Prediction of Y using X

➤ Try to perform well on **all future predictions**

➤ Minimize

$$\sum_{i=1}^{50} (y_i - a - bx_i)^2$$

LS PREDICTION EQUATION



A ROBUST APPROACH

A ROBUST APPROACH

- Construct an equation that works well on **the majority of the future predictions**

A ROBUST APPROACH

- Construct an equation that works well on **the majority of the future predictions**

- Minimize **trimmed squared-prediction error**

A ROBUST APPROACH

- Construct an equation that works well on **the majority of the future predictions**

- Minimize **trimmed squared-prediction error**

$$\begin{aligned} r_i &= (y_i - a - bx_i)^2 \\ r_{(1)} &\leq r_{(2)} \leq \cdots \leq r_{(50)} \end{aligned}$$

A ROBUST APPROACH

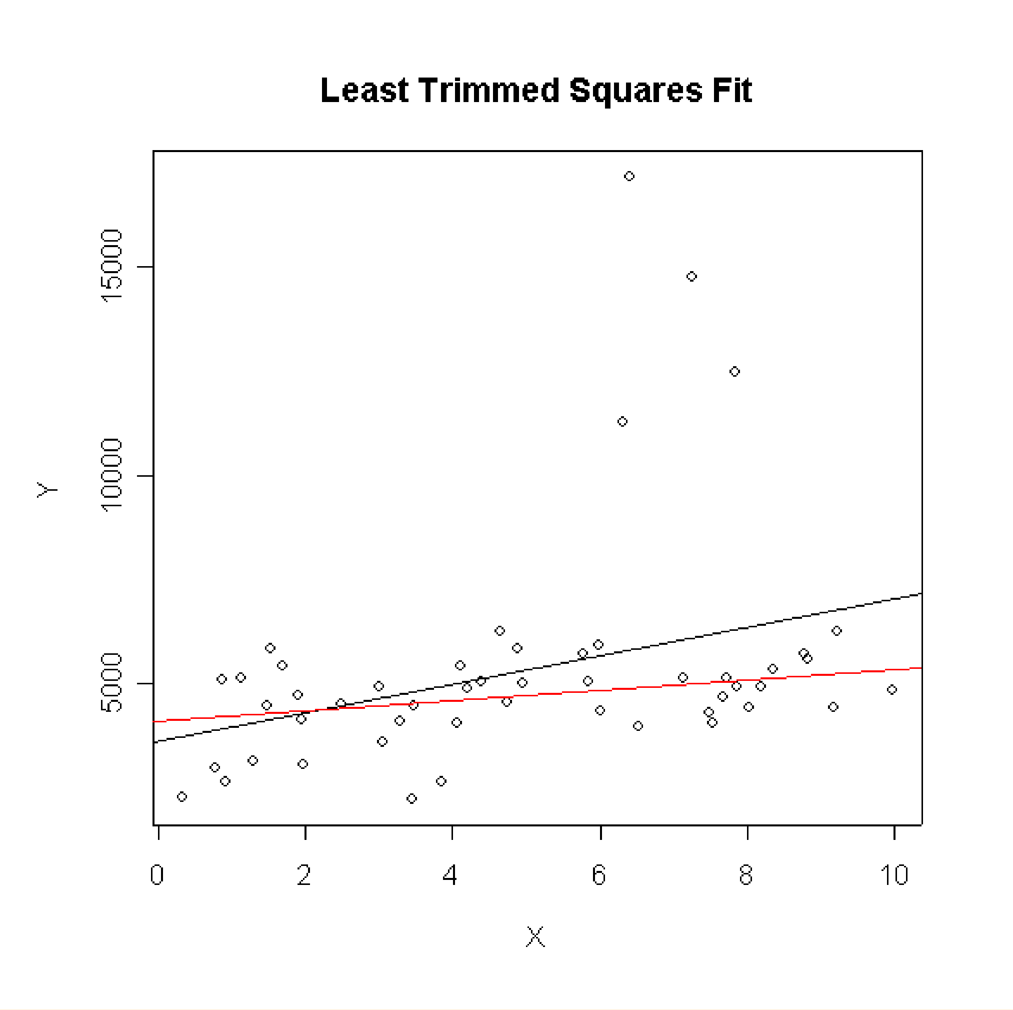
- ➤ Construct an equation that works well on **the majority of the future predictions**

- ➤ Minimize **trimmed squared-prediction error**

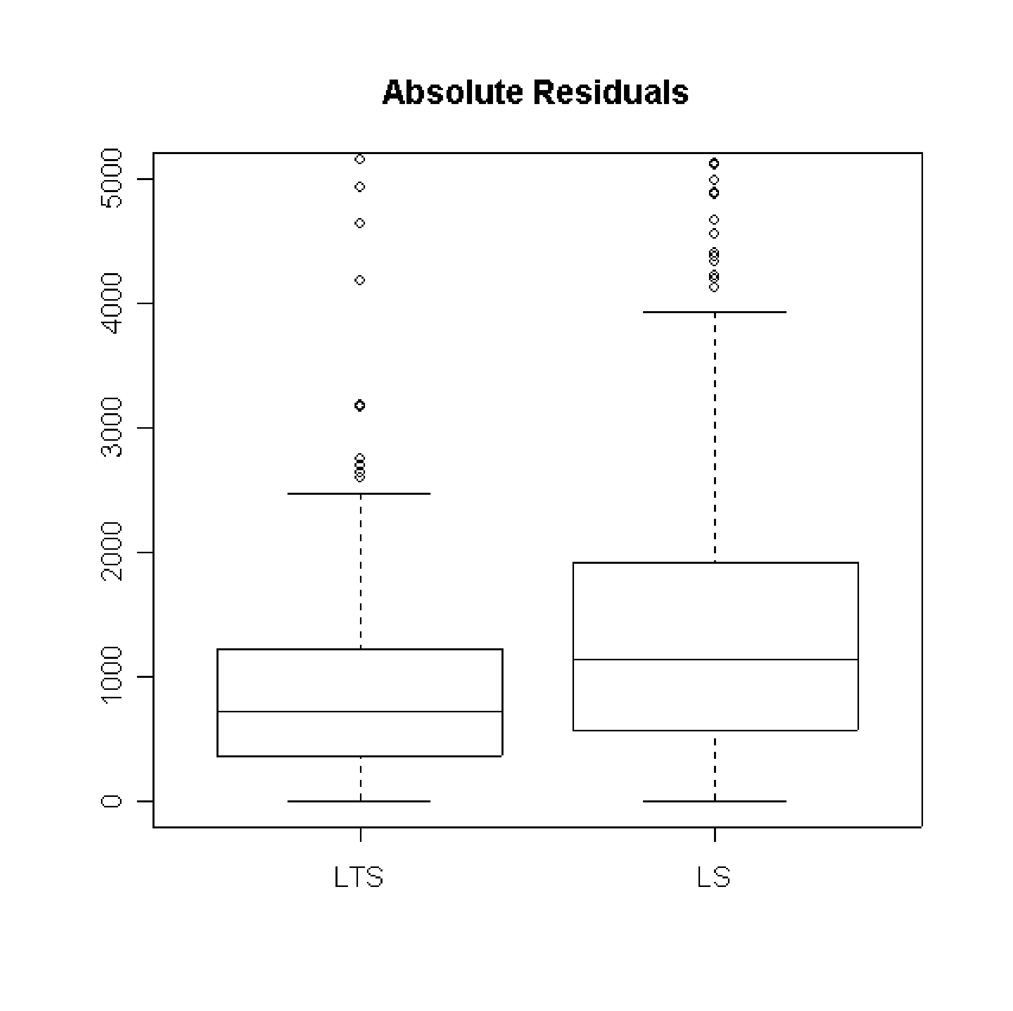
$$
\begin{aligned}
r_i &= (y_i - a - bx_i)^2 \\
r_{(1)} &\leq r_{(2)} \leq \cdots \leq r_{(50)}
\end{aligned}
$$

$$R(a,b) = \min_{a,b} \sum_{i=1}^{30} r_{(i)}(a,b)$$

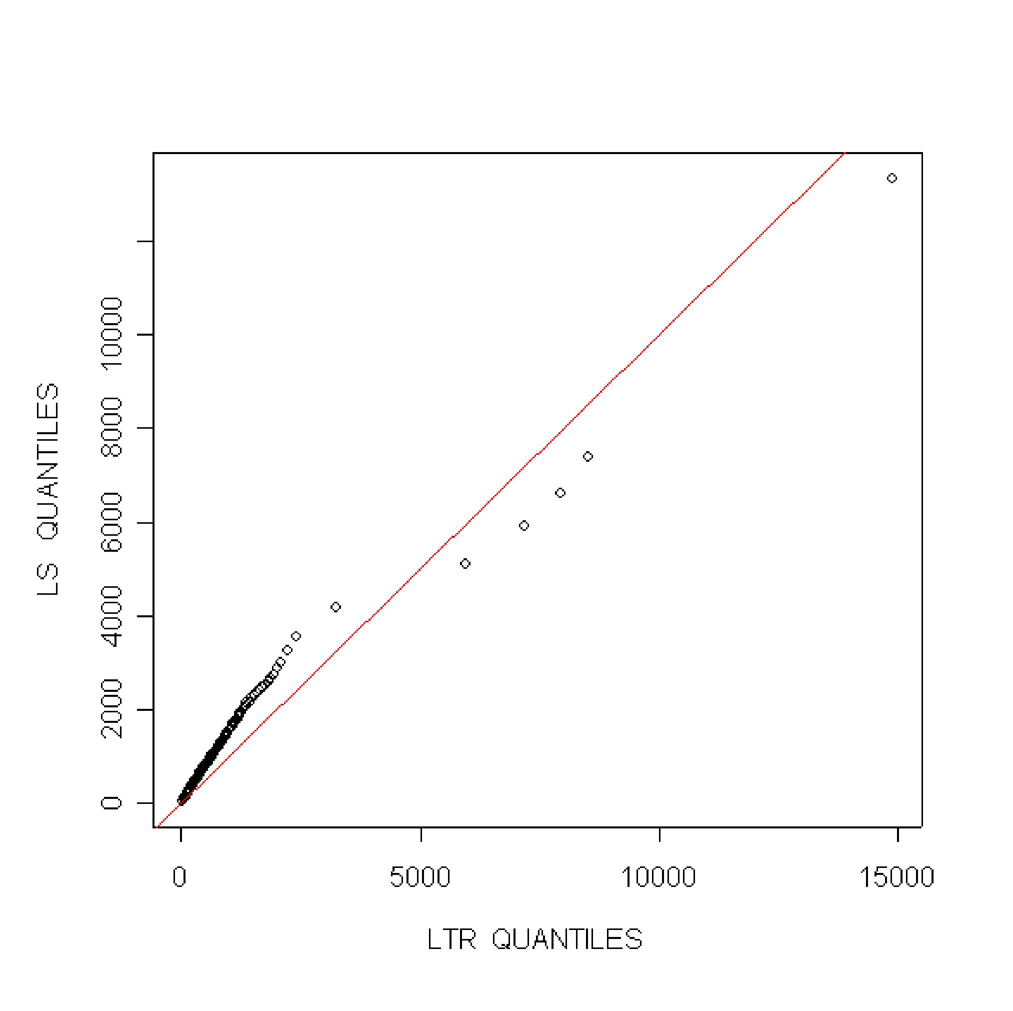
LTR FIT



ABSOLUTE PREDICTION ERROR



Q-Q PLOT


LS QUANTILES
0
2000
4000
6000
8000
10000
0
5000
10000
15000
LTR QUANTILES

CONCLUSION

CONCLUSION

AN ARGUABLY BETTER PREDICTION STRATEGY RESULTED FROM:

CONCLUSION

AN ARGUABLY BETTER PREDICTION STRATEGY RESULTED FROM:

1) A MORE MODEST PREDICTION GOAL

CONCLUSION

AN ARGUABLY BETTER PREDICTION STRATEGY RESULTED FROM:

1) A MORE MODEST PREDICTION GOAL

2) A MORE ROBUST SCORING PROCEDURE

SEARCHING FOR HOMOLOGOUS PROTEINS (SUPERVISED LEARNING)

SEARCHING FOR HOMOLOGOUS PROTEINS (SUPERVISED LEARNING)

➤ DATA (from the KDD Data Cup 2004)

SEARCHING FOR HOMOLOGOUS PROTEINS (SUPERVISED LEARNING)

➤ DATA (from the KDD Data Cup 2004)

- *74 features (variables) measured on 145,751 proteins (cases)*

SEARCHING FOR HOMOLOGOUS PROTEINS (SUPERVISED LEARNING)

➤ DATA (from the KDD Data Cup 2004)

- *74 features (variables) measured on 145,751 proteins (cases)*
- *Proteins are grouped into 153 blocks corresponding to 153 different native sequences*

SEARCHING FOR HOMOLOGOUS PROTEINS

➤ FEATURES

- *Length of alignment*
- *Percentage of sequence identity*
- *Z score for global sequence alignment*
- *Several scores of local sequence alignment*
- *...*
- *http://kodiak.cs.cornell.edu/kddcup/protein_description.pdf*

SEARCHING FOR HOMOLOGOUS PROTEINS

➤ Block Size (Number of Candidate Proteins per Block)

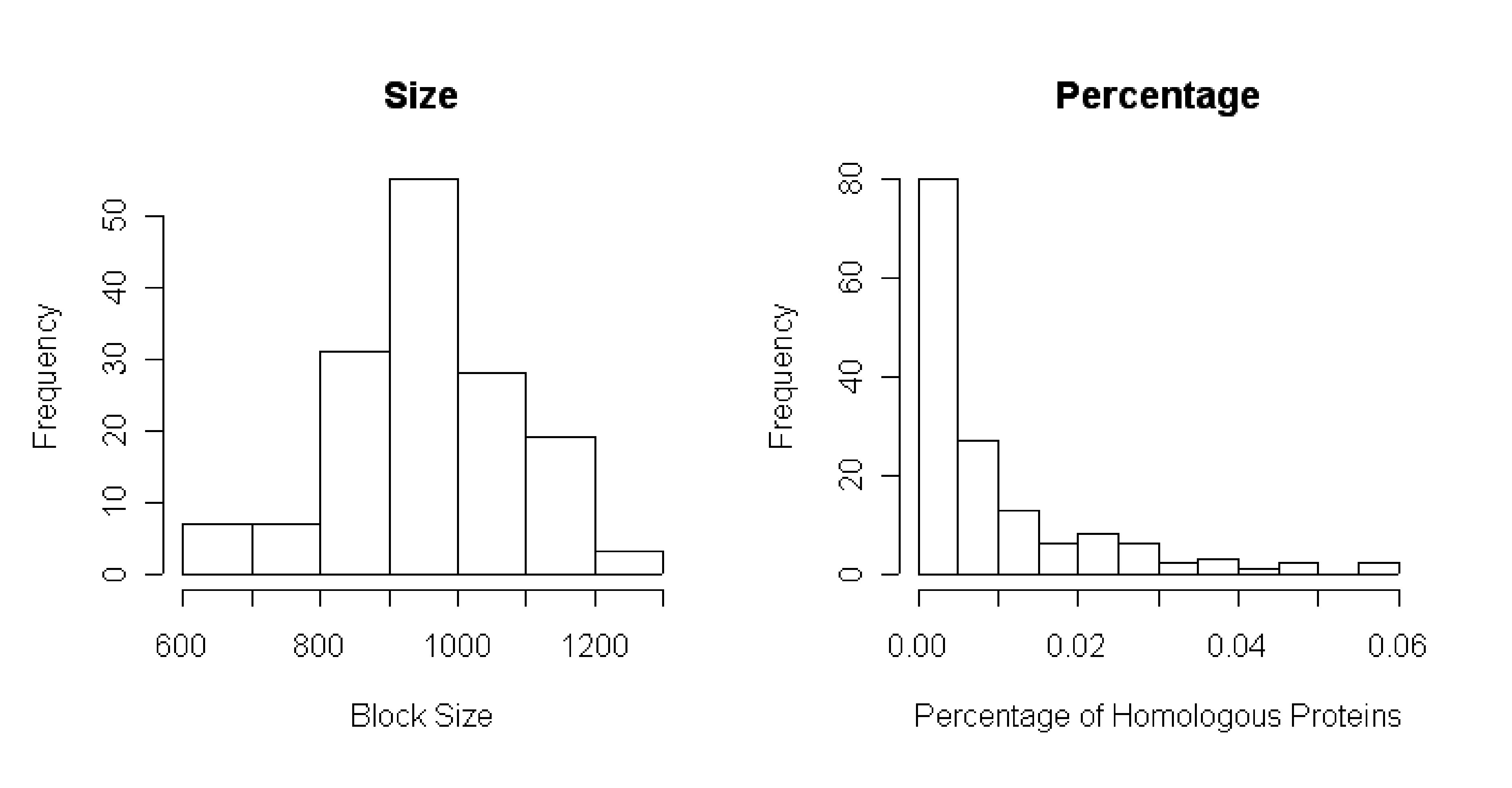
- *Smallest Block Size = 620,*
- *Largest Block Size = 1244,*
- *Median Block Size = 962*

SEARCHING FOR HOMOLOGOUS PROTEINS

➤ Percentage of Homologous Proteins per Block (hits)

- *Smallest Percentage = 0.08%*
- *Largest Percentage = 5.8%*
- *Median Percentage = 0.04%*
- *70% of the blocks have less than 1% homologous proteins*

BLOCKS SIZE AND PERCENTAGE OF TARGET PROTEINS



SEARCHING FOR HOMOLOGOUS PROTEINS

➤ GOAL: to predict which proteins are homologous to each of the 153 “target” native sequences.

➤ TASK: prioritize the candidate proteins in each block from top to bottom

SEARCHING FOR HOMOLOGOUS PROTEINS

➤ GOAL: to predict which proteins are homologous to each of the 153 “target” native sequences.

➤ TASK: prioritize the candidate proteins in each block from top to bottom

- *Proteins in each block must be assigned probabilities of being homologous*
- *Proteins in each block are then ranked from first to last according to these probabilities*

PERFORMANCE MEASURES

PERFORMANCE MEASURES

$$t_j = \begin{cases} 1 & \text{If the } j^{th}\text{-ranked protein in the block} \\ & \text{is homologous (a hit)} \\ & \\ 0 & \text{If the } j^{th}\text{-ranked protein in the block} \\ & \text{is not homologous (a miss)} \end{cases}$$

TOPK

$$TOP_k \;=\; \max\{t_j : j = 1, 2, ..., k\}$$

TOPK

$$TOP_k = \max\{t_j : j = 1, 2, ..., k\}$$

For example

$$TOP_1 = 1, \quad \text{IF TOP RANKED IS A HIT}$$

TOPK

$$TOP_k = \max\{t_j : j = 1, 2, ..., k\}$$

For example

$$TOP_1 = 1, \text{ IF TOP RANKED IS A HIT}$$

Average TOP_1 (over blocks) is a robust performance measure.

RANK OF THE LAST POSITIVE

$$RKL \;=\; \max\{j : t_j = 1\}$$

RANK OF THE LAST POSITIVE

$$RKL \;=\; \max\{j : t_j = 1\}$$

Average RKL (over blocks) is a non-robust performance measure.

MEAN SQUARED ERROR

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(\pi_j - t_j)^2$$

AVERAGE PRECISION

$$AP = \frac{\sum_{j \in J} \left(\frac{1}{j} \sum_{k=1}^{j} t_k \right)}{\sum_{j=1}^{n} t_j}$$

$$J = \{ j : t_j = 1 \}$$

OUR ANALYSIS

➤ One, two and three-dimensional data exploration showed that

- *Some features are highly correlated*
- *Some variables seemed promising and others seemed random noise*
- *No obvious pattern differentiates the blocks*

OUR ANALYSIS

➤ One, two and three-dimensional data exploration showed that

- *Some features are highly correlated*
- *Some variables seemed promising and others seemed random noise*
- *No obvious pattern differentiates the blocks*

➤ Tried different classification strategies including

- *Bayesian factor based on one-dimensional kernel density estimates*
- *Linear and quadratic discriminant analysis*
- *Recursive partitioning*
- *Nearest neighbor*
- *Logistic regression*
- *etc.*

OUR RESULTS

OUR RESULTS

- Selection of variables appeared to be much more important than the selection of classification tools.

OUR RESULTS

- ➤ Selection of variables appeared to be much more important than the selection of classification tools.
- ➤ Restricted attention to logistic regression and TOP1, which is at the same time the most challenging and robust measure

OUR RESULTS

- Selection of variables appeared to be much more important than the selection of classification tools.
- Restricted attention to logistic regression and TOP1, which is at the same time the most challenging and robust measure
- Used two fold cross-validation and stepwise forward selection to choose variables

OUR RESULTS

- ➤ Selection of variables appeared to be much more important than the selection of classification tools.
- ➤ Restricted attention to logistic regression and TOP1, which is at the same time the most challenging and robust measure
- ➤ Used two fold cross-validation and stepwise forward selection to choose variables
- ➤ Performance improved as variables entered the model up to a certain point and then begun to deteriorate

OUR RESULTS

- Selection of variables appeared to be much more important than the selection of classification tools.
- Restricted attention to logistic regression and TOP1, which is at the same time the most challenging and robust measure
- Used two fold cross-validation and stepwise forward selection to choose variables
- Performance improved as variables entered the model up to a certain point and then begun to deteriorate
- Variables: $X_{53}, X_{63}, X_{38}, X_{58}, X_{63}, X_{35}, X_{15}, X_{8}, X_{12}, X_{26}, X_{36}$

OUR RESULTS

PERFORMANCE	OUR	RANK	THE BEST
TOP1	0.8867	8	0.9200
RMS	0.0383	6	0.0350
RKL	52.8466	4	45.6200
APR	0.8206	6	0.8412

FINDING THE NEEDLE OUTSIDE THE HAYSTACK

FINDING THE NEEDLE OUTSIDE THE HAYSTACK

- Now we consider a different problem:

FINDING THE NEEDLE OUTSIDE THE HAYSTACK

- Now we consider a different problem:

FINDING HOMOLOGOUS PROTEINS

WITHOUT A TRAINING SAMPLE

LOOKING OUTSIDE THE HAYSTACK

LOOKING OUTSIDE THE HAYSTACK

➤ Homologous proteins are a small minority in a see of candidate proteins.

LOOKING OUTSIDE THE HAYSTACK

- Homologous proteins are a small minority in a see of candidate proteins.
- Their features may then appear as **“outliers”** in several low dimensional spaces.

LOOKING OUTSIDE THE HAYSTACK

➤ Homologous proteins are a small minority in a see of candidate proteins.

➤ Their features may then appear as **"outliers"** in several low dimensional spaces.

➤ STRATEGY: for each pair of variables, calculate Mahalanobis distances using a fast and robust bivariate covariance matrix.

LOOKING OUTSIDE THE HAYSTACK

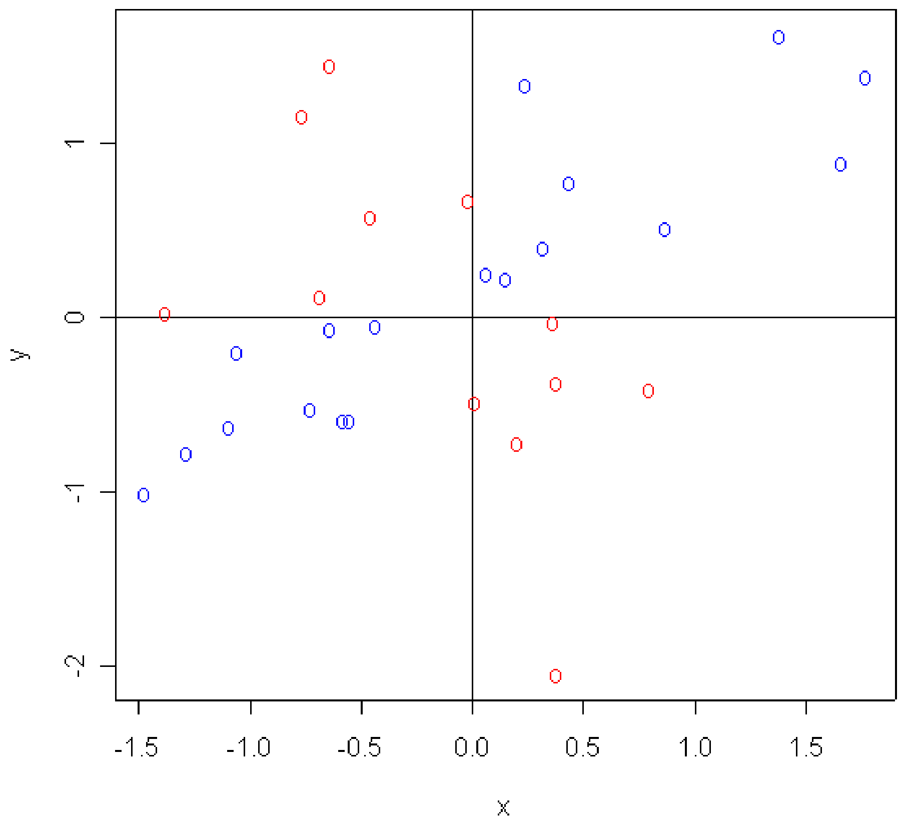
➤ Homologous proteins are a small minority in a see of candidate proteins.

➤ Their features may then appear as **“outliers”** in several low dimensional spaces.

➤ STRATEGY: for each pair of variables, calculate Mahalanobis distances using a fast and robust bivariate covariance matrix.

➤ We used **coordinate-wise medians** the **quadrant correlation**.

QUADRANT CORRELATION



LOOKING OUTSIDE THE HAYSTACK

LOOKING OUTSIDE THE HAYSTACK

- CALCULATE THE MAHALANOBIS DISTANCE RANK OF EACH PROTEIN FOR EACH PAIR OF VARIABLES

LOOKING OUTSIDE THE HAYSTACK

➤ CALCULATE THE MAHALANOBIS DISTANCE RANK OF EACH PROTEIN FOR EACH PAIR OF VARIABLES

➤ CALCULATE THE AVERAGE RANK FOR EACH PROTEIN (AVERAGE OVER ALL PAIRS OF VARIABLES)

LOOKING OUTSIDE THE HAYSTACK

➤ CALCULATE THE MAHALANOBIS DISTANCE RANK OF EACH PROTEIN FOR EACH PAIR OF VARIABLES

➤ CALCULATE THE AVERAGE RANK FOR EACH PROTEIN (AVERAGE OVER ALL PAIRS OF VARIABLES)

➤ PRIORITIZE THE PROTEINS ACCORDING TO THEIR AVERAGE RANKS

RESULTS

PERFORMANCE	RESULT
TOP1	0.74
TOP2	0.79
TOP3	0.80
TOP4	0.83

THANKS

FOR

YOUR ATTENTION