Which measure is best?

A case study clustering stocks

Michael William Boldt

CSci 8363, Spring 2005

Introduction

- Much work goes into data mining methods
- Choosing a method for a given data set is also important
- One application: clustering stock price data into industries

What's been done

- Back & Weigend
  - Applied Independent Component Analysis (ICA) to Japanese stock data
  - Conclude ICA provides insight Principal Component Analysis does not
- Gavrilov et. al
  - Evaluated different methods of clustering stock data
    - Data representation
    - Normalization
    - Dimension reduction

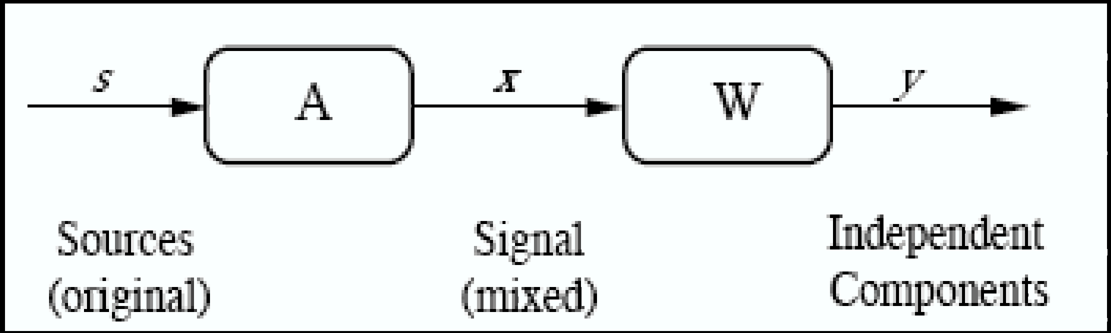
What I'll do

- Modify method comparison experiment
  - Add ICA as dimension reduction method
  - Use recent data
- Goals
  - Evaluate ICA as dimension reduction technique for this data set
  - Validate original results with recent data
- Hypothesis
  - ICA will yield most accurate clustering
  - Original results will hold with recent data

Outline

- ICA
  - Problem
  - Applications
  - Brief algorithm overview
- Experiment
  - Data
  - Methods
  - Results
- Summary

ICA problem (1/2)



- Known as “Blind Source Separation”
- Assume data is linear combination of statistically independent sources
- Know nothing about original sources or how they’re combined
- Extract statistically independent components to estimate original sources

ICA problem (2/2)

- Let
  - X: rows are observations
  - S: rows are unknown statistically independent source signals
  - A: unknown mixing matrix
  - $X = AS$
- We want to separate data into sources
  - $Y = WX \approx WAS$
  - Y: computed independent component
  - W: demixing matrix

ICA applications

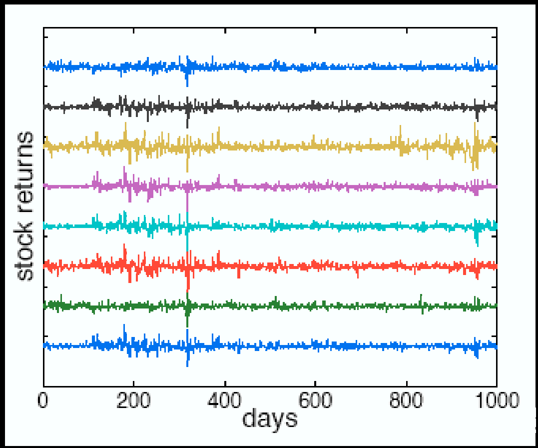
- Electrophysiology
- MRI analysis
- Face recognition
- Lip-reading

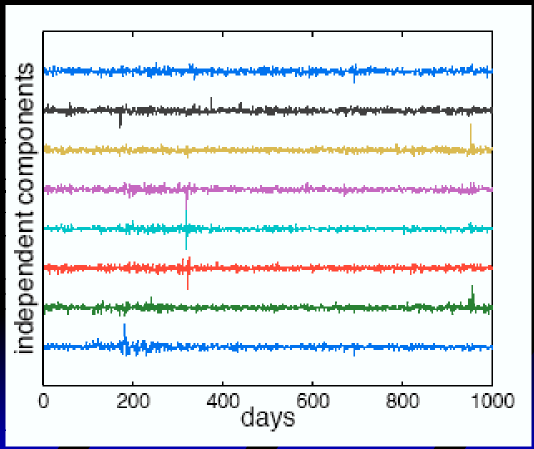
ICA basic algorithm

- Preprocessing
  - Center data (subtract mean)
  - Decorrelate/whiten/sphere data (make covariance matrix identity)
  - Results in zero-mean, unit variance, zero correlation
- Minimize gaussianity of data
  - Equivalent to maximizing independence

ICA vs. PCA on stock data (1/2)

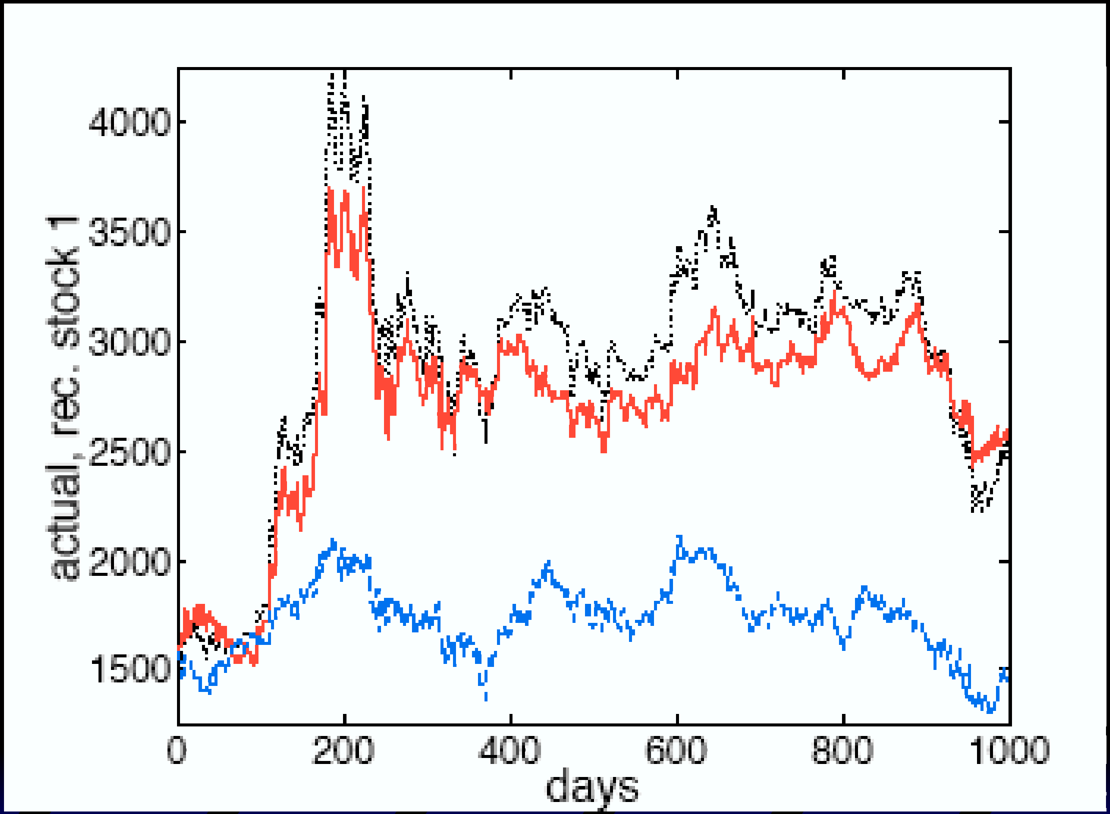
- Price shocks identified better by ICA





ICA vs. PCA on stock data (1/2)

- PCA gives best fit, but ICA offers more structural insight



Experiment data

- 1 year daily S&P 500 prices
- Some stocks not complete year
  - Members of index can change
  - Original study set missing days to 0 when necessary

Experiment data representation

- Daily opening prices
- “First derivative”
  - $p_i = p_{i+1} - p_i$

Experiment normalization

- None
- Global
  - Center
  - Divide by 2-norm
- Piecewise
  - Split sequence into windows
  - Apply global normalization to each window

Experiment dimension reduction

- None
- PCA
- Aggregation
  - Split sequence into windows
  - Replace window by mean
- I will use ICA

Experiment clustering method

- Hierarchical Agglomerative Clustering (HAC)
- Series of binary merges
- Best results: smallest maximum distance b/w inter-cluster elements

Evaluating and comparing results

- Ground-truth: stock industries
- Given clusterings $C = C_1 \ldots C_k$, $C' = C'_1 \ldots C'_k$
  - $S(C_i, C'_j) = 2\frac{|C_i \cap C'_j|}{|C_i| + |C'_j|}$
  - $S(C, C') = (\sum_i \max_j S(C_i, C'_j))/k$

Previous results (1/3)

- {raw, first derivative} $\times$ {global, none} $\times$ {dimensions}

FD	Norm	Dims	Sim(S&P,HAC)	Sim(HAC,S&P)
N	N	all	0.183	0.210
N	N	5	0.197	0.210
N	Y	all	0.222	0.213
N	Y	10	0.211	0.212
Y	N	all	0.154	0.198
Y	N	50	0.172	0.207
Y	Y	all	0.290	0.298
Y	Y	100	0.310	0.310

Table 1: The clustering results, with PCA dimensionality reduction

Previous results (2/3)

- {raw, first derivative} $\times$ {global, none} $\times$ {window size}

FD	Norm	AggWin	Sim(S&P,HAC)	Sim(HAC,S&P)
N	N	none	0.183	0.210
N	N	5	0.192	0.217
N	N	10	0.193	0.215
N	N	20	0.192	0.213
N	Y	none	0.228	0.217
N	Y	5	0.217	0.212
N	Y	10	0.221	0.216
N	Y	20	0.215	0.220
Y	N	none	0.152	0.197
Y	N	5	0.190	0.211
Y	N	10	0.195	0.217
Y	N	20	0.178	0.208
Y	Y	none	0.288	0.294
Y	Y	5	0.225	0.217
Y	Y	10	0.230	0.231
Y	Y	20	0.211	0.211

Table 2: The clustering results, with dimensionality

Previous results (3/3)

- {raw, first derivative} $\times$ {piecewise} $\times$ {window size}

Window	FD	Sim(S&P,HAC)	Sim(HAC,S&P)
10	N	0.322	0.326
15	N	0.307	0.314
30	N	0.270	0.273
45	N	0.266	0.281
60	N	0.246	0.241
75	N	0.255	0.257
10	Y	0.338	0.334
15	Y	0.346	0.339
30	Y	0.330	0.329
45	Y	0.346	0.333
60	Y	0.316	0.310
75	Y	0.310	0.297

Table 4: The clustering results, with piecewise normalization

Summary

- Methods are important, but so is matching methods to data
- ICA gives insight into stock market data beyond PCA
- Some methods claimed better at clustering stocks
  - First derivative
  - Piecewise normalization
- My project will combine these concepts