Introduction and overview

Stats 503

Prof. Liza Levina

Learning from Data

- Fact: The amount of data and information collected and stored is constantly increasing, due to advances in data collection, computerization of many aspects of life and breakthroughs in storage technology.
- Consequence: Statistical problems have increased both in size and complexity.
- The data analyst's job: make sense of all these data! Identify patterns and trends, uncover ”interesting” relationships among the variables and/or the observations, predict future behavior.

- Technology helps
  - Faster computers $\Rightarrow$ more flexible and thus more powerful techniques $\Rightarrow$ fewer modeling assumptions
  - New graphic capabilities (a picture is worth a thousand words...)
- But not always: Faster computers do not solve all problems
  - Some problems are inherently computationally intractable
  - "Easy" black-box data analysis can lead to a lot of misuse and misunderstanding
  - Flexible models can overfit (too much of a good thing)
  - Understanding underlying assumptions and interpreting conclusions correctly remains as important as ever

What is “multivariate analysis”?

- The name historically refers to a particular set of techniques
- Multivariate data: $X = \{X_1, \ldots, X_p\}$, the variables $X_1, \ldots, X_p$ can be quantitative, ordinal, categorical, or a mix of all of the above.
- This is in contrast to univariate data, where there is only one variable X
- Response: an additional variable Y (scalar- or vector-valued) that depends on X.
- When a response is present, it is usually of interest to understand the relationship between Y and X and/or predict Y from X.

Supervised vs unsupervised learning

Unsupervised learning: only X is observed

- Goal: understand/summarize/visualize the relationships between the variables in X
- Examples: principal components analysis, clustering

Supervised learning: X and Y are observed

- Goal: understand/summarize/visualize the relationships between X and Y, learn to predict Y from X
- Examples: regression (continuous Y), classification (categorical Y), ANOVA (categorical X, continuous Y)

This course covers

- Unsupervised techniques
  - Principial components analysis
  - Dimension reduction
  - Clustering
- Supervised techniques
  - Model-based classification (discriminant analysis, logistic regression)
  - Model-free classification (trees, support vector machines, ensemble methods)
- Categorical data analysis (briefly)
- Visualization as appropriate

Some important issues we'll talk about

- Underlying probability models and statistical inference – where possible
- The role of the multivariate normal distribution
- Computational inference: bootstrap, permutation tests
- Algorithmic considerations, where possible: do the methods scale to "Big Data"?
- Interpretation: what the analysis does and does not tell us

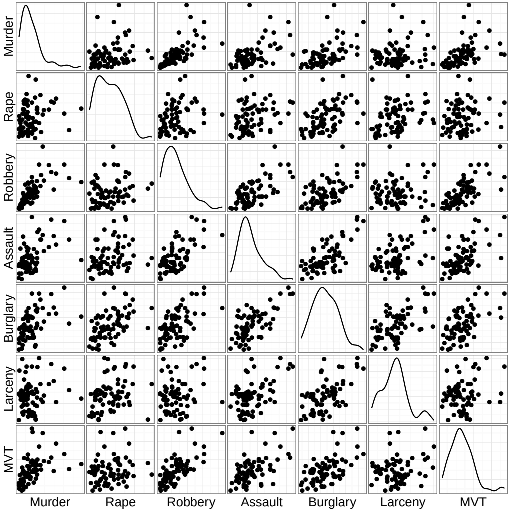
Example: U.S. cities crime data

The data give crime rates per 100,000 people for 73 large U.S. cities. The variables are:

1. Murder
2. Rape
3. Robbery
4. Assault
5. Burglary
6. Larceny
7. Motor Vehicle Thefts (MVT)

Goal: summarize, visualize – unsupervised analysis

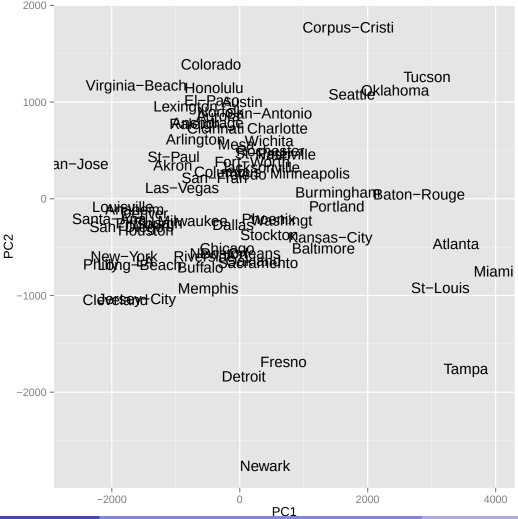
Scatterplot matrix of U.S. cities crime data



Scatterplots of many variables can be hard to read.

A 2-d representation of U.S. cities crime data

Can combine the variables and produce a safety "index": a principal components analysis plot



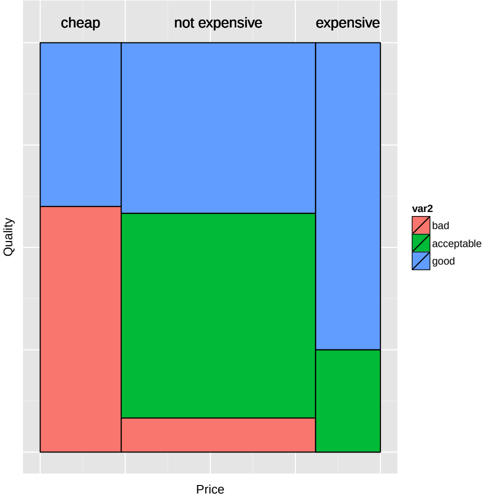
Example: sleeping bags (categorical data)

- The variables are price, fiber and quality for 21 sleeping bags
- All variables are categorical; cannot do a scatterplot.
- Goal: understand something about the relationship between price and quality of available sleeping bags – unsupervised analysis

	cheap	not expensive	expensive	down fibers	synthetic fibers	good	acceptable	bad
Brand	**Price**			**Fiber**		**Quality**		
One Kilo Bag	1	0	0	0	1	1	0	0
Sund	1	0	0	0	1	0	0	1
Kompakt Basic	1	0	0	0	1	1	0	0
Finmark Tour	1	0	0	0	1	0	0	1
Interlight Lyx	1	0	0	0	1	0	0	1
Kompakt	0	1	0	0	1	0	1	0
Touch the Cloud	0	1	0	0	1	0	1	0
Cat's Meow	0	1	0	0	1	1	0	0
Igloo Super	0	1	0	0	1	0	0	1
Donna	0	1	0	0	1	0	1	0
Tyin	0	1	0	0	1	0	1	0
Travellers Dream	0	1	0	1	0	1	0	0
Yeti Light	0	1	0	1	0	1	0	0
Climber	0	1	0	1	0	0	1	0
Viking	0	1	0	1	0	1	0	0
Eiger	0	0	1	1	0	0	1	0
Climber light	0	1	0	1	0	1	0	0
Cobra	0	0	1	1	0	1	0	0
Cobra Comfort	0	1	0	1	0	0	1	0
Foxfire	0	0	1	1	0	1	0	0
Mont Blanc	0	0	1	1	0	1	0	0

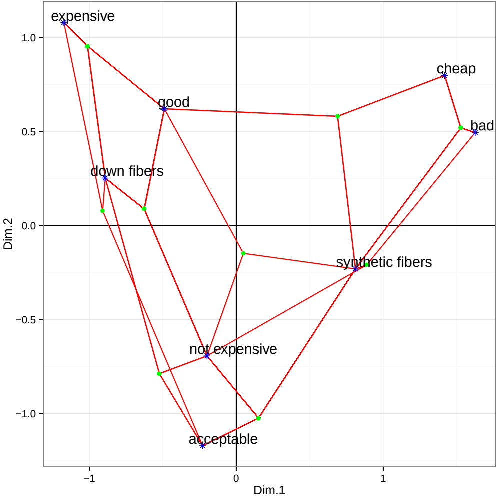
How do we visualize the sleeping bag data?

A panel plot for price and quality variables



How do we visualize the sleeping bag data?

A plot from multiple correspondence analysis

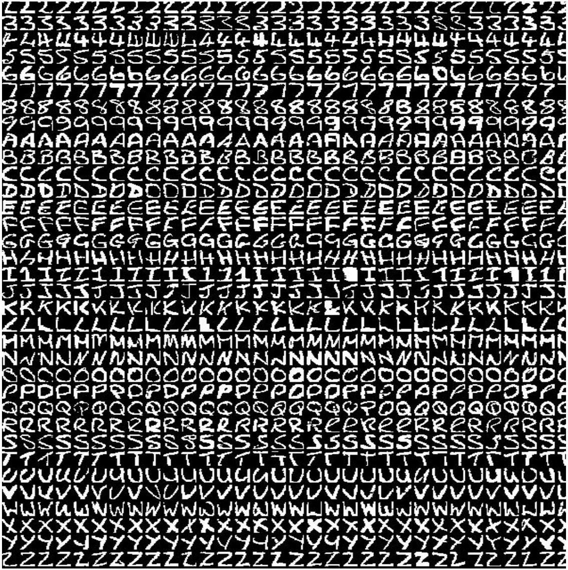


green points represent the sleeping bags

Some findings off the sleeping bags picture

- there are good, expensive, down-filled sleeping bags
- there are bad, cheap, synthetic-filled sleeping bags
- there are some expensive ones of acceptable quality and some cheap ones of good quality
- there are no bad expensive sleeping bags
- all expensive bags are filled with down

Example: optical character recognition



Example: handwritten letters and digits dataset

- Data: images of single handwritten letters and digits
- Each image is 20×16 pixels, with pixel intensities from 0 to 255. This vector of 320 quantitative variables is X (features).
- Response/outcome: the identity of each image $\{A, B, ..., Z, 0, 1, ..., 9\}$. This categorical variable with 36 levels is Y.
- Goal: build an algorithm (classifier, learner) to predict the identity Y from pixel values X using a training dataset of labelled images – supervised analysis
- A good algorithm should predict well not only on training data, but also on test data (pairs of X and Y that have not been used to build/train the algorithm).

Example: DNA expression data

- DNA is the basic material that makes up human chromosomes.
- DNA microarrays and other gene chips are new technologies measuring quantitative expression of thousands of genes simultaneously from a single sample of cells.
- Here is a tiny sample of DNA expression data: 3 genes (variables) and 4 samples (observations).

21652	3.2025	1.6547	3.2779	1.0060
25725	0.0681	0.0710	0.1160	0.1906
22260	0.1243	0.0520	0.1014	0.1035

- The full dataset has approximately 7000 genes (rows) and around 100 samples (columns), where the samples correspond to different cancer tumors.

What can one learn from expression data?

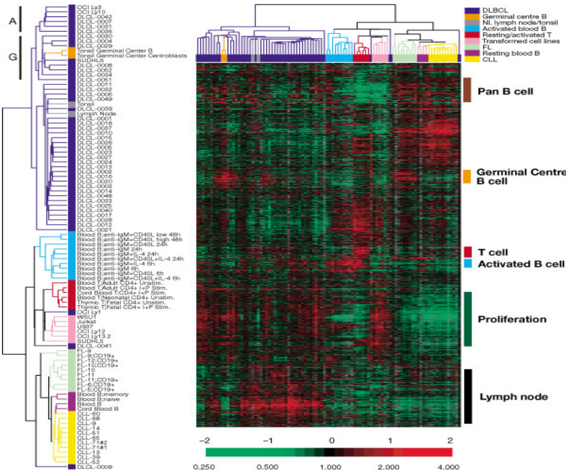
Typical unsupervised questions (Hastie et al., 2001):

- Which samples are most similar to each other, in terms of their expression profiles across genes? (clustering)
- Which genes are most similar to each other, in terms of their expression profiles across samples? (clustering)
- Do ”interesting” patterns exist between subsets of genes and samples (e.g. very high/low expression levels)?

Typical supervised questions:

- Can type of tumor be predicted from gene expression levels?
- Which genes are most predictive for which tumors?

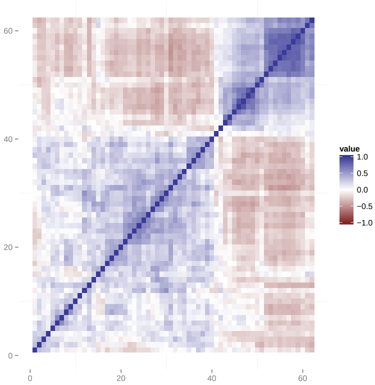
Heat map of DNA microarray data after clustering



Picture taken from Alizadeh et. al (2000), *Nature*

Another visualization example: the correlation matrix

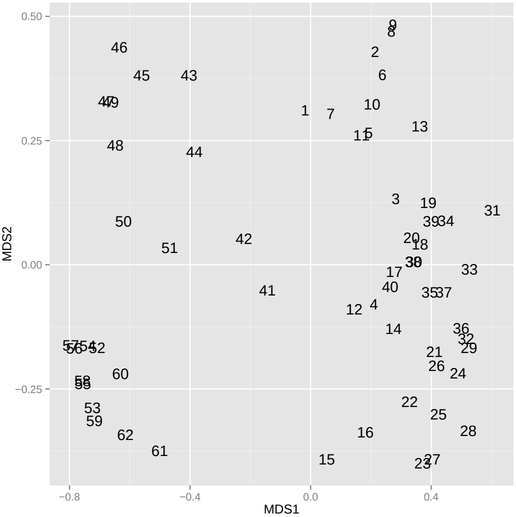
- 62×62 correlation matrix of 62 lymphona samples, computed from gene expression measurements of 4000+ genes, from the previous example of Alizadeh et. al (2000).
- How are these lymphona samples related to each other?
- Too many numbers to examine – visualize this matrix via a heatmap:



Distance-based representation

- How do we see groups in the tumors more clearly?
- Another look: plot samples as points in the plane, keeping their distances as close as possible to those implied by correlations (small distance = high correlation)

A correlation distance-based map of the tumors



Good quotes to keep in mind

Essentially, all models are wrong, but some are useful.
– George Box (Box and Draper, 1987).

There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea.
– Andreas Buja (quote taken from Hastie et al., 2001))

Practice

- Join up with one or two neighbors
- Brainstorm as a group and come up with an example of multivariate data that you'd be interested in analyzing
- Formulate one specific question about your example and decide whether it is a supervised or an unsupervised question