Exploratory Data Analysis and Cancer Genomics

BD2K

Andrew Nobel

February, 2020
Overview

- The Scientific Method: Then and Now
- Reproducible Research
- Exploratory Data Analysis
- Principal Component Analysis
- Clustering
- Correlation Mining
- Community Detection
The Scientific Method
The Scientific Method: Flow Chart (from science buddies.org)

1. Ask a Question
2. Do Background Research
3. Construct a Hypothesis
4. Test with an Experiment
5. Procedure Working?
   - No: Troubleshoot procedure. Carefully check all steps and set-up.
   - Yes: Analyze Data and Draw Conclusions
     - Results Align with Hypothesis
     - Results Align Partially or Not at All with Hypothesis
       - Communicate Results

   Experimental data becomes background research for new/future project. Ask new question, form new hypothesis, experiment again!
Paradigm Shift

**Traditional Scientific Method:** Hypothesis Driven

- Formulate a hypothesis
- Collect data to confirm/refute hypothesis

**Modern Scientific Method:** Data Driven

- Acquire data from high-throughput measurement technologies
- Mine the data for possible hypotheses
- Use the data again to test selected hypotheses
**General Principle:** If you have enough data, and you ask enough questions, you are bound to find something interesting, *just by chance.*

**Bob:** I found a needle in a haystack!

**Amy:** That’s surprising! How many haystacks did you look in?

**Bob:** A thousand.

**Amy:** Oh, maybe that’s not so surprising.
Two Facets of Reproducible Research

I. Reproducibility of analysis: Can we replicate the analysis?
   ▶ Public access to raw data and preprocessing steps
   ▶ Public access to general and special purpose software
   ▶ Careful step-by-step documentation of data analysis

II. Reproducibility of conclusions: Are the conclusions true?
   ▶ Are data, methods, and assumptions of initial study sound?
   ▶ Are results of initial study robust?
   ▶ Do similar experiments with different data yields the same conclusion?
Reproducibility Crisis

**2015:** Re-examination of 100 psychology studies
- Only 33 studies were reproducible

**2012:** Re-examination of 53 landmark studies in oncology and hematology.
- Only 6 studies were reproducible

**2009:** Re-examination of 18 gene expression studies
- Only 2 studies were reproducible
Lack of Reproducibility: Some Causes

**Experimental Process**

- Discount contradictory data; collect more data until results look good
- Change initial hypothesis after seeing the data
- Try out many hypotheses until you find one supported by data
- Fabrication of data and/or mis-use of data analysis (infrequent)

**Publication Process**

- Submission bias (of researcher): Only submit positive results
- Publication bias (of journal): Only publish positive results
50% selectively reported only studies that were successful.

58% looked at initial results, and then decided if they should collect more data.

43% threw out “bad” data.

35% reported unexpected findings as predicted from the outset.
Exploratory Data Analysis
Exploratory Data Analysis

First look at a data set, typically in the form of a matrix of numbers.

▶ Visualization
▶ Identifying patterns or regularities of interest
▶ Hypothesis generation

Preliminaries

▶ Identifying and addressing outliers and extreme values
▶ Imputing missing values
▶ Normalization: removing systematic differences between samples
▶ Transforming data values using logarithm or other functions
▶ Checking distributional/model assumptions
Finding Patterns: The Narrative Impulse

More Than Coincidence?

A CUMBER SOME APPARATUS

SOME CUCUMBERS AND ASPARAGUS

Drawing by B. Kliban
DNA contains the original codes for making the proteins that living cells need. mRNA is a copy of a gene located on the DNA molecule. mRNA will leave the nucleus of the cell and the ribosome will read its coding sequences and put the appropriate amino acids together.
Gene Expression Data

Arguably, the most important and common type of high-throughput data in genomics and biomedicine

**Data set:** Expression level of $p$ genes in $n$ samples

- $p = \#$ genes (usually tens or hundreds of thousands)
- $n = \#$ samples (usually tens or hundreds)

**Data matrix:** Array $X = \{x_{ij}\}$ with $p$ rows and $n$ columns

- each column contains expression measurements for a single sample
- each row contains expression measurements for a single gene
- $x_{ij} = \text{expression level of gene } i \text{ in sample } j$

*Other data types:* copy number, methylation, genotype, micro-RNA
Example: Gene Expression Data from Breast Cancer

Heat map of gene expression data from The Cancer Genome Atlas (TCGA)

- **Samples:** $n = 200$ tumors from two breast cancer subtypes
  - 100 Luminal A tumors
  - 100 Basal tumors
- **Variables:** $p = 11,000$ genes (post-filtering)
Univariate Sample $x = x_1, \ldots, x_n$

Statistics

- Sample mean $m(x) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

- Sample variance $s^2(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ and sample SD $s(x)$

- Standardized sample $\tilde{x}$ with $\tilde{x}_i = (x_i - \bar{x})/s(x)$

- Quantiles, percentiles, and order statistics

Visualization

- Histogram/density plots

- Bar and whisker plots, QQ plots
Bivariate Sample \((x, y) = (x_1, y_1), \ldots, (x_n, y_n)\)

**Statistics**

- **Sample covariance of** \(x\) **and** \(y\)
  \[
s(x, y) = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = n^{-1} \sum_{i=1}^{n} x_i y_i - \bar{x} \bar{y}
  \]

- **Sample correlation of** \(x\) **and** \(y\)
  \[
r(x, y) = \frac{s(x, y)}{s(x) s(y)} \in [-1, 1]
  \]

**Visualization**

- **Scatter-plot** \(\{(x_i, y_i) : 1 \leq i \leq n\} \subseteq \mathbb{R}^2\)
Exploratory Analysis of Genomic Data

**Step 1a:** Univariate analysis of columns and rows of data matrix $X$
- sample/variable means and standard deviations
- histograms of these

**Step 1b:** Bivariate analysis of columns and rows of data matrix $X$
- scatter plots
- $n \times n$ matrix of sample $x$ sample correlations
- $p \times p$ matrix of gene $x$ gene correlations

**Next:** Differential expression, Principal component analysis (PCA), Clustering
Two Uncorrelated Genes
Two Positively Correlated Genes
Two Negatively Correlated Genes
Aside: Regression Line and R-squared

**Def’n:** Sample regression line of $y$ on $x$ is the line $\ell^*(x)$ minimizing
\[
\text{MSE}(\ell) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \ell(x_i))^2
\]
over all linear functions $\ell(x) = ax + b$.

**Fact:** Sample regression line $\ell^*$ of $y$ on $x$ is given by
\[
\ell^*(x) = m(y) + \frac{s(x, y)}{s^2(x)} [x - m(x)]
\]
and satisfies $\text{MSE}(\ell^*) = s^2(y)[1 - r^2(x, y)]$.

**Note:** $s^2(y) = \text{MSE}$ of constant line $l(x) = m(y)$. 
Scatterplot: Mean and SD of Expression
Scatterplot of SD (expression) for Two Subtypes
Heatmap: Correlation Matrix of Samples \((n \times n)\)
Heatmap: Correlation Matrix of Genes \( (p \times p) \)
Differential Expression Analysis
Example of Differential Gene Expression

Expression Levels of Gene GABRP.2568 in Basal vs Luminal A Subgroups

- Blue line: Basal
- Red line: Luminal A

Density axis ranges from 0.00 to 0.25.
Gene Expression axis ranges from -4 to 8.
Differential Expression Analysis

**Step 1:** For each gene

- Compute a two-sample t-statistic to assess whether it is differentially expressed in two cancer subtypes
- Use the t-statistic to assign a p-value assessing the evidence for differential expression

**Step 2:** Use a multiple testing procedure to identify genes that are differentially expressed. Procedures include

- Bonferroni correction (controls family-wise error rate)
- Benjamini-Hochberg (controls false discover rate)
$t$-Statistics for Differential Expression (11K genes)

$|t|$-cut-offs ($\alpha = 0.05$): 1.97 (no correction), 2.976 (FDR), 4.723 (FWER)
P-values for Differential Expression

\( p \)-cut-offs (\( \alpha = .05 \)): .05 (no correction), .00342 (FDR), \( 4.38e^{-6} \) (FWER)
Principal Component Analysis
Principal Component Analysis (PCA)

**Given:** High dimensional samples $x_1, \ldots, x_n \in \mathbb{R}^p$ with $\sum_i x_i = 0$

**Goal:** Find a linear subspace $V$ of $\mathbb{R}^p$ meeting two criteria

- *Dimension reduction:* Dimension of $V$ is small
- *Approximation:* Each sample $x_j$ close to its projection onto $V$

**Idea:** Find good low dimensional approximation of the data.
Simplest case: Approximating subspace $V$ is one-dimensional, a line in $\mathbb{R}^p$ determined by a unit vector $v$.

Turns out

- Finding a good direction is equivalent to maximizing the variance of the orthogonal projections of the samples $x_1, \ldots, x_n$ onto $v$.

- The best direction corresponds to leading eigenvector $v_1$ of the $p \times p$ sample covariance matrix $S = n^{-1}XX^T$, with $X = [x_1, \ldots, x_n]$.

- Other directions $v_2, v_3, \ldots$ can be obtained from other eigenvectors of $S$. 
Figure: Projections of Sample data onto the first four principal components of the TCGA dataset. Colors represent subtype of cancer: Luminal A and Basal.
Image Data

- **Data**: $X = 458 \times 685$ matrix of pixel intensities

- **Idea**: Project columns of the image onto $d$ leading eigenvectors of their sample covariance matrix. Consider quality of reconstruction.
Proportion of Variation Explained
Image Reconstruction

\[ d = 1, \text{ PVE} = 80.42 \]

\[ d = 3, \text{ PVE} = 88.91 \]

\[ d = 5, \text{ PVE} = 92.99 \]

\[ d = 10, \text{ PVE} = 95.79 \]

\[ d = 20, \text{ PVE} = 97.24 \]

\[ d = 40, \text{ PVE} = 98.18 \]
Clustering
General Setting

**Given:** Vectors $x_1, \ldots, x_n \in \mathbb{R}^d$

**Goal:** Identify group structure. Divide vectors into a small number of disjoint groups, called *clusters*, such that

- distances between vectors in the same cluster are small
- distances between vectors in different clusters are large

**Areas of application**

- Genomics and Biology
- Computer Science
- Psychology and Social Sciences
The k-Means Algorithm

**Given:** Observations \( x_1, \ldots, x_n \in \mathbb{R}^d \) and desired number of clusters \( k \)

**Initialize:** Cluster centers \( C_0 = \{c_0(1), \ldots, c_0(k)\} \) with \( c_0(i) \in \mathbb{R}^d \)

**Iterate:** For \( m = 1, 2, \ldots \) do:

- Let \( \pi_m \) be the nearest neighbor partition of the cluster centers \( C_{m-1} \)
- Let \( C_m = \{c_m(1), \ldots, c_m(k)\} \) be the centroids (averages) of the vectors in the cells of \( \pi_m \)

**Stop:** When \( \text{Cost}(C_m) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x_i - c_m(j)\|^2 \) stabilizes
Agglomerative Clustering

**Stage 0:** Assign each object $x_i$ to its own cluster

**Stage $k$:**
- Find the two closest clusters at stage $k - 1$
- Combine them into a single cluster

**Stop:** When all objects $x_i$ belong to a single cluster

**Output:** Dendrogram = binary tree where every node corresponds to a cluster, height of a node is distance between its children.

**Note:** Distance $d(C, C')$ between clusters $C, C'$ measured in different ways

\[
\min_{x_i \in C, x_j \in C'} d(x_i, x_j) \quad \text{or} \quad \frac{1}{|C| \cdot |C'|} \sum_{x_i \in C, x_j \in C'} d(x_i, x_j)
\]
Clustering Samples of TCGA Data

Colors: Luminal A and Basal
Clustering Rows and Columns
Row-Column Clustering (aka Co-Clustering)

Independently cluster rows and columns of the data matrix.

Result is a checkerboard partition

Note: Red, green blocks correspond to large average submatrices representing sample-variable interactions. Information about

- disease subtypes
- regulatory pathways
Extension of Co-Clustering: Biclustering

Three overlapping Biclusters.

Direct search for large average submatrices
Allows rows and columns to overlap
**Large Average Submatrix Search Procedure (Shabalin et al. 2010)**

**Input:** An $n \times n$ matrix $X$ and integer $1 \leq k \leq n$.

**Loop:** Select $k$ columns $J$ at random. Iterate until convergence.

- Let $I := k$ rows with largest sum over columns in $J$.
- Let $J := k$ columns with largest sums over rows in $I$.

**Output:** *Locally optimum* submatrix associated with $I$, $J$.

**In Practice**

- Repeat 1000 times, adaptively choosing submatrix dimensions
- Output submatrix with largest average
- Residualize and repeat
Mining Differential Correlation
Mining Differential Correlation

SAMPLES
Condition 1 | Condition 2

VARIABLES
A

Higher Correlation | Lower Correlation
Overall Goal: Adaptively identify differentially correlated variable sets $A$.

- Candidate variable set(s) $A$ not specified in advance
- Special case of differential analysis for weighted networks

Non-Assumptions: True (population) correlations may be complex

- May not be diagonal, banded, or sparse.

Note: Differential correlation distinct from differential expression, clustering
Example: TCGA

Figure: Sample correlation matrices from Her-2 and Luminal B cancer subtypes. Differentially correlated set of 165 genes (A) and 200 randomly chosen genes (B).
Application: Brain Connectome

FMRI data from Human Connectome Project (www.humanconnectome.org)

**Single subject:** 97K brain locations (37K voxels + 60K greyordinates)

- Condition 1: 316 language tasks
- Condition 2: 284 motor tasks

**DCM output:** 5 sets of brain locations
Brain Connectome: Differential Correlation

**First DC set:** 1200 locations with $\bar{r}(C_1) = .24$ and $\bar{r}(C_2) = .05$

**Visualization:** DC locations on L/R hemisphere show clear spatial structure
Brain Connectome: Differential Expression

**Visualization:** Top 1200 locations as ranked by standard t-test
Community Detection in Networks
Undirected Networks

**Simple Graph** $G = (V, E)$ where

- **Node set** $V = [n] = \{1, \ldots, n\}$
- **Edge set** $E$ with $\{u, v\} \in E$ if $u$ is linked to $v$
- No self-loops or multi-edges

**Degree Sequence** $d = \{d(1), \ldots, d(n)\}$ with

$$d(u) = \sum_{v \in V} \mathbb{I}(\{u, v\} \in E) = \text{number of edges incident on } u$$
Given $G = (V, E)$ identify sets $C_1, \ldots, C_k \subseteq V$ such that

- Edge density within sets $C_i$ is large
- Edge density between sets $C_i$ is small
- Sets $C_i$ called communities
Community Detection: Applications

Exploratory Analysis of

- Social networks
- Genetic networks
- Communication networks
Community detection and clustering share common goal of grouping objects, but differ in fundamental ways:

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Community Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Vectors</td>
<td>Nodes</td>
</tr>
<tr>
<td>Similarity (continuous)</td>
<td>Connectivity (binary)</td>
</tr>
<tr>
<td>Metric structure</td>
<td>Relational structure</td>
</tr>
</tbody>
</table>
Facebook Network of J. Wilson

- Nodes = friends of JW on FB (561)
- Edges between FB friends (8375)
- Friends divided into 8 different groups

Results of community detection

- 7 communities detected
- Match score = .87 out of 1
Conclusion
Recap

- The Scientific Method: Then and Now
- Reproducible Research
- Exploratory Data Analysis
- Principal Component Analysis
- Clustering and Biclustering
- Community Detection
- Correlation Mining
Acknowledgements

Folks

▶ Kevin O’connor
▶ Carson Mosso
▶ Mark He
▶ Kelly Bodwin
▶ Katie Hoadley
▶ Chuck Perou