High-dimensional data visualization & analysis in R

Simina Boca, Matt McCoy

Innovation Center for Biomedical Informatics, Georgetown University Medical Center

October 24, 2018

Process of data analysis

- Data analysis is a multi-stage process!
- Generally process goes from experimental design to inference.
  - Includes data cleaning, exploratory data analysis prior to modelling.
  - Sometimes the pre-inference steps are called “data janitorial services” but they are where I spend most of my time and essential to really understanding what is going on!

Leek and Peng, *Nature*, 2015

High-dimensional data

- Generally defined as scenarios where the number of features being measured (p) is larger than the number of samples (n).
- p can often be very large (hundreds, thousands, millions!)
- Makes it harder to visualize and analyze the data
- Examples?

High-dimensional data

- Generally defined as scenarios where the number of features being measured (p) is larger than the number of samples (n).
- p can often very large (hundreds, thousands, millions!)
- Makes it harder to visualize and analyze the data
- Examples? All omics datasets, imaging data, wearable technology and other time-series data etc

Is it better to have more data?

Is it better to have more data?

- It can be, but it also depends on context!
- Need to keep in mind all the ideas used for “small data” when analyzing “big data” as well as additional issues that may arise.
- The study/context in which the data were collected is key, otherwise you may answer the wrong question!

Exploratory data analysis (EDA)

- Look at dataset in several different ways before starting to model!
  - Immediately starting to model or perform t-tests etc is almost always a bad idea.
- EDA is essential for any data analysis!
- Consider a combination of:
  - Tables: Number of missing or unusual values, ranges, quantiles
  - Graphs: Boxplots, scatterplots, clustering

Exploratory data analysis (EDA)

- For high-dimensional data, typically have demographic/clinical characteristics for individuals or samples along with omics or imaging data - need to do EDA for both!
  - In our example, have disease status (Duchenne muscular dystrophy vs healthy), age, and study site, along with metabolite measurements.
- Tips at `https://github.com/SiminaB/Mentoring/blob/master/Omics_exploratory_analysis_checklist.md` and `https://leanpub.com/exdata` (book by Roger Peng - can download for free!)

Visualizing high-dimensional data

- "You have to make big data into small data to visualize it."[I forget who said this]
- Can think of your data as an n x p matrix.
- Graphs are only in 2D, though can represent additional dimensions using colors, symbols etc.
- How can we reduce dimensionality?

Visualizing high-dimensional data

- How can we reduce dimensionality?
- One commonly used approach is principal component analysis (PCA), which decomposes the matrix into the directions of highest variability that can be expressed as orthogonal linear combinations of the original features.
  - The first principal component (PC) is an n x 1 vector that captures the most variability from the original dataset (among all possible linear combinations), the 2nd PC is an n x 1 vector that captures the most variability orthogonal on the first etc.
  - Often look at the first 2 PCs - in the resulting plot, each point is a sample and the value is a combination of all the features.
  - Generally helpful to perform PCA on omics dataset, then represent clinical/demographic characteristics via different colors, shapes etc.

Why is PCA helpful?

- It gives a general idea of overall features of the dataset (can formalize with statistical tests)
- It can identify outlying samples (which could not be identified looking at thousands or tens of thousand of dimensions!)
- It can identify potential artifacts or batch effects, like samples processed in different ways

Read Leek et al, *Nature Reviews Genetics*, 2010, for more details on batch effects and using PCA or clustering to detect them.

Performing statistical analysis for high-dimensional data

- Many techniques exist, depending on whether the goal is inference or prediction.
- Will focus on inference here, specifically hypothesis testing: What metabolites are significantly associated with Duchenne muscular dystrophy, adjusting for age and study site and allowing for different trends with age in cases and controls?
- Approach is to fit a regression model for each metabolite, then adjust for multiple testing.
- In the end, get a list of metabolites that should be followed up in future studies.

Multiple testing issues

- If the null hypothesis (of no difference) holds true when performing a single statistical test, then a p-value < 0.05 will be observed 5% of the time (over repeated experiments).
- When performing m statistical tests, even if the null hypotheses are all true, we will expect $0.05 \times m$ rejections at the $p = 0.05$ level.
- Can address this in multiple ways, for example by controlling the family-wise error rate (FWER) via the Bonferroni correction or the false discovery rate (FDR) via the Benjamini-Hochberg (BH) procedure.

Reproducibility: An essential component of a data analysis

- Reproducibility refers to making the data and code of a project available so that anyone can run the same analysis and get the same results.
- When using R, one approach is to use R markdown, which combines text, code, and output.
- Reproducibility is helpful for the scientific community, but also for yourself.

Reproducibility: An essential component of a data analysis