Dimension Reduction

Katie Pollard

BMI 206

In this unit we will learn …

- Why we use dimension reduction in bioinformatics
- Techniques for visualizing high dimensional data using dimension reduction / manifolds
- Relationship between linear regression and principal components analysis (PCA)

Why perform dimension reduction?

- Lower computational demands
  - cpu
  - memory
  - storage / network (i.e., compression)
- Visualization
- De-noising
- Identifying important variables (i.e., variable selection)

How does dimension reduction work?

- Have a large n-by-p matrix for n samples and p variables (covariates, features) per sample
- How close are the n samples in l-dimensional space where $l < p$?
- Preserve relationships from p-dimensional space as well as possible by some constraint.
- Typically involves a measure of distance
  - covariance
  - other distances
- Can also do the inverse for p n-vectors.

Lower dimension representations

- One profile per cluster (after clustering p covariates)
  - medoids
  - medians / means
- Linear combinations of variables, e.g.,
  - Principal components analysis (PCA)
  - Independent components analysis (ICA)
- Non-linear combinations of variables, e.g.,
  - T distributed stochastic neighbor embedding (t-SNE)
  - Multi-dimensional scaling (MDS)

Many methods produce embeddings or manifolds

Definitions for lower dimension representations

- Embed = form without intersections in a higher dimensional space (more generally, a subspace)
- Manifold = a locally Euclidean topological space