12.3 Missing Values and Matrix Completion

Missing Values

- We talked about missing data a bit in Stat 140A.
  - We used mean imputation and multiple imputation to fill them in.
- Missing data can also be tackled by principal components based imputation.
  - This process is called *matrix completion*

Types of Missingness

- **Missing at Random** (MAR): the probability an observation is missing depends on other observed values in the dataset, but not on the value of the missing data itself.
- **Missing Completely at Random** (MCAR): the probability of an observation being missing is unrelated to both the observed and unobserved data.
- **Missing Not at Random** (MNAR) means there is some systematic reason for the missingness that is directly related to the unoserved value itself.

Approaches to Missing Data

- Mean imputation is appropriate only if the data are missing completely at random (MCAR).
- Multiple imputation is appropriate for MCAR and missing at random (MAR)
- Matrix completion also requires the data be at least MAR.
- If missing data is MNAR, more involved approaches are required.

Missing Data

- We can also use missing data methods where data is missing by necessity.
- Ex: a matrix of ratings given by n customers to the entire Netflix catalog of p movies.
  - Most values will be missing, but if we impute them accurately, we can use that to recommend movies people haven't seen yet.

Principal Components with Missing Values

Recall: The first M principal components provide the "best" approximation to the data matrix X.

- Suppose some of the observations x_{ij} are missing.
- We can both impute the missing values and solve the principal component problem simultaneously!

Principal Components with Missing Values

Consider

$$\underset{A \in R^{n \times M}, B \in R^{p \times M}}{\text{minimize}} \left\{ \sum_{(i,j) \in O} \left(x_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\}$$

where O is the set of all *observed* paired of indices (i, j), a subset of all possible $n \times p$ pairs.

Once we've solved this, we can

- estimate a missing observation using $\hat{x}_{ij} = \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm}$
- (approximately) recover the M principal component scores and loadings, as we did when the data were complete

Principal Components with Missing Values

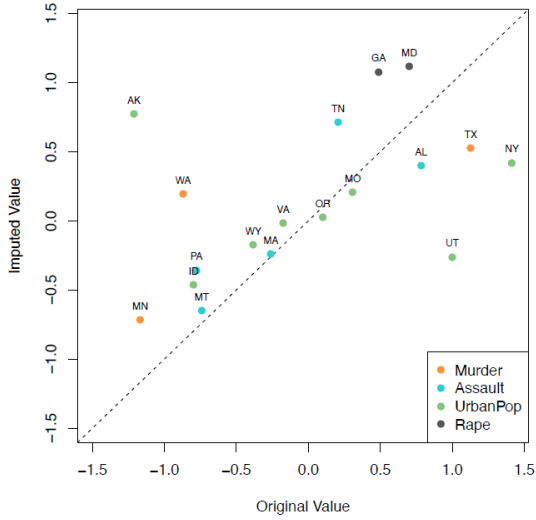
This minimization problem is difficult, but we can use an algorithmic approach:

1. Create a complete data matrix $\tilde{X}$ of dimension $n \times p$ of which the (i, j) element $\tilde{x}_{ij}$ is x_{ij} if observed and $\bar{x}_j$ otherwise.
2. Iterate until the objective fails to decrease:
   a. Solve
   $$\underset{A \in R^{n \times M}, B \in R^{p \times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\tilde{x}_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\}$$
   by computing the principal components of $\tilde{X}$
   b. For each unobserved element, set $\tilde{x} \leftarrow \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm}$
   c. Compute the objective
   $$\sum_{(i,j) \in O} \left(x_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2$$
3. Return the estimated missing entries $\tilde{x}_{ij}$

Example: USArrests

- There are $p = 4$ variables and $n = 50$ states.
- 20 states selected and, for each, one of the four variables was randomly deleted.
  - This results in 10% missing data.
- The previous algorithm is applied with $M = 1$ principal component.

Example: USArrests



- The correlation between the true and imputed values is around 0.63

Example: USArrests

Is a 0.63 correlation good?

- If we estimate those values using the complete data, we find an average correlation of 0.79
- So this imputation method is pretty good!
- This data also has only 4 variables - methods like this one work better with more variables.

Recommender Systems: Netflix Data

	Jerry Maguire	Oceans	Road to Perdition	A Fortunate Man	Catch Me If You Can	Driving Miss Daisy	The Two Popes	The Laundromat	Code 8	The Social Network	⋯
Customer 1	•	•	•	•	4	•	•	•	•	•	⋯
Customer 2	•	•	3	•	•	•	3	•	•	3	⋯
Customer 3	•	2	•	4	•	•	•	•	2	•	⋯
Customer 4	3	•	•	•	•	•	•	•	•	•	⋯
Customer 5	5	1	•	•	4	•	•	•	•	•	⋯
Customer 6	•	•	•	•	•	2	4	•	•	•	⋯
Customer 7	•	•	5	•	•	•	•	3	•	•	⋯
Customer 8	•	•	•	•	•	•	•	•	•	•	⋯
Customer 9	3	•	•	•	5	•	•	1	•	•	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Recommender Systems

- This particular example has $n = 480,189$ customers and $p = 17,770$ movies.
- On average, each customer had seen around 200 of the movies, so 99% of the matrix is missing.
- By imputing the missing review values, Netflix can attempt to recommend movies people might like.