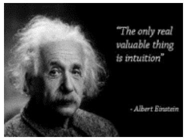
Dimension Reduction Models

XI(✗) – X’s are not independent (are correlated)



Dimension Reduction: Intuition

- Some models may need **many** important and somewhat **correlated variables**, which is particularly problematic if the **ratio of variables to observations** is large – i.e., reduced degrees of freedom
- The methods covered so far have addressed dimensionality issues by either using a **subset** of variables or by **shrinking** their coefficients
- Business models generally don't include too many variables, but other fields like **biology** often have models with thousands of variables – impractical so select a subset or use shrinkage
- **Survey data** is notorious for having large number of variables too.
- In such cases, it helps to explore the linear relationships among the variables and use the observed correlation to create new variables that are **linear combinations** (i.e., **components**) of the original variables.
- When we do this, a **few** of the new **components** may explain a large portion of the variance in the data, thus helping **reduce** the model **dimension** without losing much explanatory power.

Dimension Reduction Methods

- The basic idea is if we have **P** somewhat correlated **predictors** it is possible to transform these into **M linear combinations**, such that **P > M**, thus **reducing** the number of **variables** in a model.
- **Dimension reduction =** reduce the estimation of **P+1** coefficients **(β_0, β_1, β_2,… β_P)** to estimating **M+1** coefficients **(α_0, α_1, α_2,… α_M)**
- ***Example:*** *if we suspect that a vehicle's volume, horsepower, and weight affect the vehicle's gas mileage, but these 3 variables are highly correlated, we could combine them into a new variable called something like "size" composed of some percentage of volume, plus some of horsepower, plus some of weight, reducing the model variables from 3 to 1.*
- Naturally, we also **lose** some **interpretability**, so it is a **tradeoff**
- Two popular dimension reduction methods are **Principal Components Analysis** (PCA) and **Partial Least Squares** (PLS), both of which use the **correlation matrix** of P predictors to find M (<P) linear combinations of the P predictors
- These methods are may **increase bias** but substantially **reduce variance** of the coefficients, particularly when **P** is **large relative to N**

KOGOD SCHOOL
of
BUSINESS