Lecture #8: High Dimensionality & Principal Components Analysis (PCA)

CS 109A, STAT 121A, AC 209A: Data Science

Weiwei Pan, Pavlos Protopapas, Kevin Rader

Fall 2016

Harvard University

Announcements

- Project Milestone #2: not due on Wednesday. Look for an announcement in a few days.
- Project Assignments: look for the official assignments to come out Thursday.
- Religious conflicts: we are finalizing a policy regarding the quizzes, etc...
- Vocareum submission issues: they keep logs of button clicks. Be sure to (i) your file is not the blank notebook starter file and (ii) check under the submission status that you submitted it.
- Midterm details coming.

Quiz Time

Time for Quiz(zes).

Lecture Outline

More on Interaction Terms

High Dimensionality

Principal Components Analysis (PCA)

PCA for Regression (PCR)

Interaction Terms: a Review

Recall from lecture last time that an interaction term between predictors X_1 and X_2 can be incorporated into a regression model by including the multiplicative (aka, cross) term in the model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon$$

What is the point of an interaction term?

Interaction terms: some guidelines

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon$$

It is good practice to always include the *main effects* terms $\beta_1 X_1$ and $\beta_2 X_2$ when including the interaction term in an model. Why?

If X_1 is a binary predictor (say to indicate taxi $X_1 = 1$ vs. uber pickups to predict Y the length of ride (in minutes) at X_2 time of day of pickup), then what is the interpretation of β_3?

Under what type of model would it make sense to include the interaction term without one of the main effects?

The mass of interaction terms

Imagine we are trying to predict the systolic blood pressure (BP) of thousands of patients ($n \approx 10,000$). There could potentially dozens of patient characteristics that could be important (age, sex, weight, BMI, etc...) and potentially hundreds of drugs (various beta-blockers, pain medication, etc...) that could affect BP as well. Thus $p \approx 200$ or greater.

It may also be of interest to determine if these drugs interact with each other, and whether the drugs interact with patient characteristics.

We could model these data using a model with many, many interaction terms.

Counting up interaction terms

A *two-way interaction* term is the cross between two main effects, and a three-way interaction terms is the cross between three main effects, etc...

Say $n = 10,000, p = 200$. How many two-way interaction terms are there? How many three-way? How many total number of terms (main effects and interaction terms) are there in total?

$$\binom{p}{2} = \frac{200 \cdot 199}{2} = 19900, \qquad \binom{p}{3} \approx 1.3 \text{ million}$$

$$2^p \approx 1.6 \times 10^{60}$$

What would happen if you attempted to fit the "full" regression model with all main effects and all interaction terms (or even just all the two-way interaction terms)? What if you had enough observations?

Too many predictors ($p > n$)

This model would be called *unidentifiable* since the number of model parameters exceeds the number of observations. This means that not all of the parameters could be estimated at once.

Note: this does not even consider non-linear terms (for example, X_j^2) which would make this model even more intractable or unidentifiable.

In practice what should an analyst do?

They should either (i) only consider scientifically important interaction terms, (ii) build a model through stepwise regression (from the ground up), or (iii) take a different approach to deal with this *high dimensionality*.

Lecture Outline

More on Interaction Terms

High Dimensionality

Principal Components Analysis (PCA)

PCA for Regression (PCR)

When High Dimensionality occurs

In the previous section we saw that when the number of parameters exceeds the number of observations, then the issue of high dimensionality means that we cannot estimate everything in our model. This can occur when we are considering lots of interaction terms.

But this can occur in other settings too when the number of main effects is high. For example:

- The predictors are genomic markers in a computational biology problem.
- The predictors are someone's complete browsing history online
- The predictors are the occurrences of all words in the dictionary in a text analysis problem

Framework for Dealing with High Dimensionality

One way to reduce the dimensions (rather than shrinking effects towards zero like in Ridge and LASSO which does not improve dimensionality issues) is to redefine your set of predictors as a *linear combination* of the original X_j's.

Let $Z_1, Z_2, ..., Z_M$ represent $M < p$ linear combinations of the original p predictors. More specifically:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for fixed constants ϕ_{jm}. Then a regression model can be fit as such

$$Y = \theta_0 + \theta_1 Z_1 + ... + \theta_m Z_m + \epsilon$$

Thus instead of having to estimate $p + 1$ coefficients, there are now a reduced set of parameters to estimate ($M + 1$ of them).

Framework for Dealing with High Dimensionality (cont.)

Any method to reduce dimensionality should act in 2 steps:

1. Determine the $Z_1, ..., Z_M$ linearly transformed predictors
2. Estimate the coefficients $\theta_1, ..., \theta_M$ (aka, fit the model).

The method to determine $Z_1, ..., Z_M$ or which model to fit can vary by user. We will explore the creation of the reduced set of predictors $Z_1, ..., Z_M$ through PCA.

Lecture Outline

More on Interaction Terms

High Dimensionality

Principal Components Analysis (PCA)

PCA for Regression (PCR)

Principal Components Analysis (PCA)

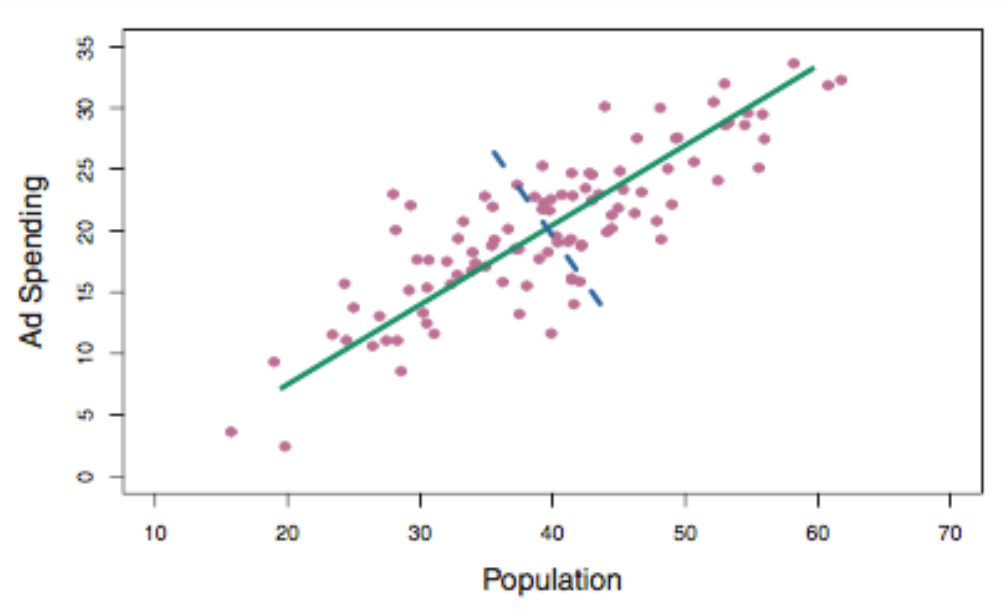
Principal Components Analysis (PCA) is an approach to automatically determine the linear combinations of predictors to consider in order to reduce the dimensionality of the problem.

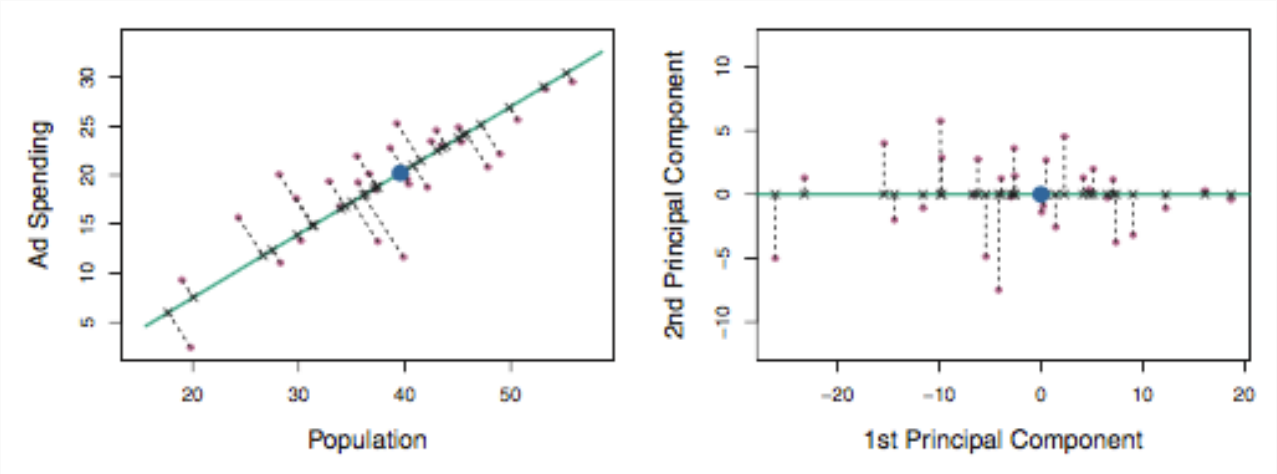
PCA is an unsupervised technique (ignores Y). In theory it creates a set of p (same size as the original predictor set) linearly transformed predictors $(Z_1, ..., Z_p)$ that:

- Require $\sum_{j=1}^{p} \phi_j^2 = 1$
- Are orthogonal to each other.
- Are in order contain more variability of the predictors as they are calculated $(Z_1 > Z_2 > ... > Z_p)$.

Z_1 can thus be thought of as the combination of predictors that represents them the best. Z_2 is the next best combination of all predictors that is also unrelated to Z_1. A picture is worth a thousand words...

Visualization of PCA





The Math behind PCA

PCA is a well-known result from linear algebra. Let $\mathbf{Z}$ be the $n \times p$ matrix consisting of columns $Z_1, ..., Z_p$ (the resulting PCA vectors), $\mathbf{X}$ be the $n \times p$ matrix of $X_1, ..., X_p$ of the original data variables (each re-centered to have mean zero, and without the intercept), and let $\mathbf{W}$ be the $p \times p$ matrix whose columns are the eigenvectors of the square matrix $\mathbf{X}^T\mathbf{X}$, then

$$\mathbf{Z}_{n\times p} = \mathbf{X}_{n\times p}\mathbf{W}_{p\times p}$$

Implementation of PCA using linear algebra

To implement PCA yourself using this linear algebra result, you can perform the following steps:

1. Subtract off the mean for each of your predictors (so they each have mean zero).
2. Calculate the eigenvectors of the $\mathbf{X}^T\mathbf{X}$ matrix and create the matrix with those columns, $\mathbf{W}$, in order from largest to smallest eigenvalue.
3. Use matrix multiplication to determine $\mathbf{Z} = \mathbf{XW}$

Note: this is not efficient from a computational perspective. This can be sped up using Cholesky decomposition.

However, PCA is easy to perform in Python using the `decomposition.PCA` function in the `sklearn` package.

Lecture Outline

More on Interaction Terms

High Dimensionality

Principal Components Analysis (PCA)

PCA for Regression (PCR)

Using PCA for Regression

PCA is easy to use in Python, so how do we then use it for regression modeling in a real-life problem?

If we use all p of the new Z_j, then we have not improved the dimensionality. Instead, we select the first M PCA variables, $Z_1, ..., Z_M$, to use as predictors in a regression model.

The choice of M is important and can vary from application to application. It depends on various things, like how collinear the predictors are, how truly related they are to the response, etc...

What would be the best way to check for a specified problem?

Train and Test!!!

Using PCA for a simplified real world problem

Let's predict income in the General Social Survey (n=1379) using 6 predictors:

- age
- educ (in years)
- sex (female = 1)
- crack (ever smoked crack = 1)
- foreignborn (born overseas = 1)
- numchildren

We first determine the principal components...

Using PCA for a simplified real world problem