

Principal Component Analysis .

Idea: instead of using as variables the actual measured quantities, look at the directions of greatest variation in your data. "principal components".

This idea is often useful for understanding your data set.

Examples:

(1) Psychology: we have data from a 100-question personality test.

For every person we have 100 measured variables. (All answers to questions.)

If we apply PCA to this data we find that almost all the variation

is explained by the first 5 principal components.

When we look at the actual questions we see that these components have natural interpretations:

Intraversion / Extraversion.

Conscientiousness

Neuroticism / Stability

etc.

That is, PCA discovers for you the theories of personality.

(2) Stock market data:

We have daily returns for ~ 4000 publicly traded US stocks.

Most of the variation in the returns is explained by the first 3 principal components.

(1) "Market Factor": How is the overall market doing? (SP 500 index, say)

(2) Growth vs Value

↑
~~BA~~ Companies which are growing profits/revenue quickly but don't necessarily making much money now.
↓
Makes money now, but business not growing.

(3) Small vs Large $52 = 1.8 \cdot 8$
(measured by Market Capitalization)

A problem with PCA:

Let's say we are trying to predict Y from variables $X_1, \dots, X_{p-1}$.

We can use PCA to isolate the

directions of greatest variation in X .

What if these directions are not useful for predicting Y ?

There is another technique, "PLS"
Partial Least Squares. (Section 11.2)

We look instead at the directions in the X -data that have maximum covariance with Y , instead of maximum variance.

This might lead to a very different result.

Issues with both: interpretability.

Often we can give a natural interpretation of the first few ~~comp~~ principal components.

No guarantee that this will be true.

Requires (usually) domain-specific knowledge.

Standard example (3).

We measure interest rates on US Treasury Bonds at different maturities:

1yr , 2yr. 3yr. 30-year

What are the principal components?

Level, Slope, Curvature.

When the economy is expanding.