Session 2: Principal Components Analysis

Ben Shepherd

Principal, Developing Trade Consultants Ltd.

ARTNeT/RIS Capacity Building Workshop on the Use of Gravity Modeling

Friday, November 11, 2011
1. Introduction

2. Principal Components Analysis
   - Overview
   - Mechanics of PCA

3. Summary
Many trade facilitation papers have noted that:

- Trade facilitation is a multi-dimensional concept covering, for example: air and maritime port infrastructure; transparency; and service sector regulation; BUT
- Indicators for each of these dimensions tend to be very strongly correlated.

This leaves us with a problem: we don’t want to leave out an important dimension, but by including everything, we lose precision in our estimates.
An obvious way to try and deal with this problem is to construct a summary indicator, or a small number of them, covering each of the dimensions we are interested in.

By doing this, we are trading off two considerations:

- Summarizing entails a loss of some variation in the data, and thus some information;
- Summarizing allows us to get a more precise estimate of the effect of this less-informative variable.
Averaging indicators is the most straightforward way of summarizing them:

- Simple average: each indicator has the same weight.
- Weighted average: different weights are applied to different indicators, usually according to professional judgment.

Wouldn’t it be nice to have a statistical methodology that allows us to derive a meaningful weighted average? It might be more acceptable to many audiences than professional judgment based on N=1...
1. Introduction

2. Principal Components Analysis
   - Overview
   - Mechanics of PCA

3. Summary
PCA is just such a methodology.

It is a statistical technique that allows us to derive one or more summary measures ("principal components") from a set of indicators.

Each PC is a weighted average of the underlying indicators. We use some math magic to choose weights such that the PC accounts for a maximum amount of the variance in the underlying indicators.
Consider the WMO trade facilitation indicators as a PCA problem.

We have four strongly correlated dimensions:

- Maritime port infrastructure
- Air transport infrastructure
- IT availability
- Corruption

We could use PCA to produce a single variable “trade facilitation” that captures as much as possible of the original variance in these indicators.
Outline

1. Introduction

2. Principal Components Analysis
   - Overview
   - Mechanics of PCA

3. Summary
PCA basically relies on some matrix algebra. Let’s get a feel for how it works.

We have $p$ variables observed across $i$ cross-sectional units, like countries. Their correlation matrix is:

$$\Sigma_{p \times p} = \begin{bmatrix}
1 & \rho_{12} & \cdots \\
\rho_{21} & \ddots & \\
\vdots & & 1
\end{bmatrix}$$
Ever heard of a spectral decomposition in matrix algebra? All it means is that we can write:

\[
\sum_{p \times p} = \mathbf{C} \mathbf{\Lambda} \mathbf{C}' = \begin{bmatrix}
    c_{11} & \cdots & c_{1p} \\
    \vdots & \ddots & \vdots \\
    c_{p1} & \cdots & c_{pp}
\end{bmatrix} \begin{bmatrix}
    \lambda_{11} & 0 & \cdots \\
    0 & \ddots & \vdots \\
    \vdots & \cdots & \lambda_{pp}
\end{bmatrix} \begin{bmatrix}
    c_{11} & \cdots & c_{1p} \\
    \vdots & \ddots & \vdots \\
    c_{p1} & \cdots & c_{pp}
\end{bmatrix}
\]
Mechanics of PCA

Now let’s say we want to turn our $p$ indicators into $r$ PCs, with $r < p$. The magic of algebra is such that we can write:

$$\sum_{pxp} \approx \sum_{pxp} = C^* \Lambda^* C^{'r} = C^* \Lambda^{1/2} C^{'r} = P P^{'}$$

P is a matrix of $r$ columns from which we can construct $r$ PCs. Each element of P is the proportion of the variance of each original indicator that a particular PC accounts for (“loadings”).

To produce the PCs, apply one of a number of standard normalizations to produce “scores”.

12
Think of PCA as a three step process:

- Analyze the correlation matrix and all possible PCs.
- Select an appropriate and meaningful number of PCs to summarize the data.
- Produce the PC scores.

Note that if you produce multiple PCs, the magic of the process ensures that they will all be uncorrelated!
Choosing the Number of PCs

- Start by running PCA and examining all the PCs.
- Two rules of thumb for the number of PCs to retain:
  - Any PCs with a corresponding eigenvalue greater than unity.
  - The number of PCs prior to the “flat” part of an eigenvalue scree plot.
- Sounds complex, but it’s often very straightforward.
PCA in Stata

- `pca var1 var2 [etc.], components(no. of PCs to retain)`
  - Execute with no options first. The output tells you about the eigenvectors, and the loadings

- `screeplot`
  - Performs a screeplot, so you can see how the eigenvectors look.

- `predict new_var, score`
  - Use the loading matrix and the chosen number of components to produce one or more PCs
Gravity modeling with strongly correlated regressors can result in imprecise estimates...

But excluding relevant variables from the model can result in bias!

Producing a smaller number of summary measures from a larger set of correlated indicators can be an appropriate middle way between these two problems.
Summary

PCA is a statistical methodology for summarizing correlated indicators into one or more PCs.

Each PC is a weighted average of the underlying indicators. Weights are chosen so as to maximize the explained proportion of the variance in the original set of indicators.

When multiple PCs are produced, each one is uncorrelated with the others.

Note: PCA is in fact one of a class of methodologies, Factor Analysis, that work in a very similar way to very similar ends.