Principal Components Analysis: Social Epi

Daniel J Carter

March 3, 2019

London School of Hygiene and Tropical Medicine
daniel.carter1@lshtm.ac.uk

PCA in Social Epidemiology

PCA in Social Epidemiology

PCA in Social Epi can be used in a variety of ways:

- Construct asset indices for measuring SEP - **today's practical** (Vyas & Kumaranayake, 2006)
- Unpack a complex construct, assuming you don't know its subcomponents a priori (Hurtado et al, 2011)
- Reduce questionnaires or Likert scales to their underlying constructs (Logie & Earnshaw, 2015)
- Discover how risk factors pattern (Navarro Silvera et al, 2011)
- Unpack a number of underlying constructs from a set of covariates (Carter et al, 2018; Oxlade & Murray, 2012)

Carter et al (2018)

With reference to the latter two papers, we are going to talk about how PCA can inform the relationship between poverty & TB.

Carter et al. (2018) looks at how achieving Sustainable Development Goal 1 can impact on TB.

- SDG 1 includes the subtargets of ‘eliminating extreme poverty’ and ‘social protection for all’

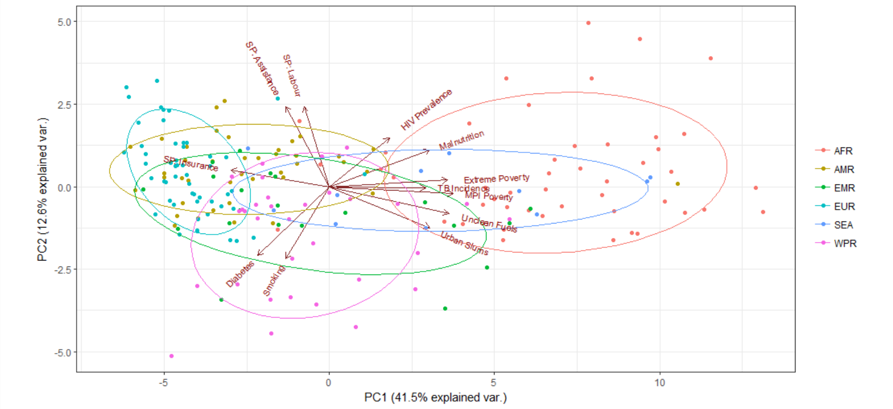
PCA was applied in the exploratory phases to determine which of the potential variables of interest were related to one another.

Specific Question of Interest:
Does an effect of social protection on TB exist at the ecological level?

SDG variables: Extreme Poverty, Multidimensional Poverty, Social Assistance, Social Insurance, Labour Market Interventions

TB covariates of interest taken from Lönnroth et al (2009) alongside their Population Attributable Fractions:

- Malnutrition: 27%
- Indoor Air Pollution: 26%
- Smoking: 23%
- HIV: 19%
- Harmful Alcohol Use: 13%
- Diabetes: 6%


PC2 (12.6% explained var.)
5.0
2.5
0.0
-2.5
-5.0
PC1 (41.5% explained var.)
-5
0
5
10
SP.Assistance
SP.Labour
HIV Prevalence
Malnutrition
SP.Insurance
Extreme Poverty
TB Incidence
MPI Poverty
Unclean Fuels
Urban Slums
Diabetes
Smoking
AFR
AMR
EMR
EUR
SEA
WPR

Questions from the biplot:

- Is TB associated with poverty?
- What is TB negatively associated with?
- Which pairs of risk factors pattern together?

Which is PC1 and which is PC2? How do you know?

- TB incidence, extreme poverty, multidimensional poverty, unclean fuels, urban slums, malnutrition, HIV prevalence, Social Insurance
- Diabetes, smoking, Social Assistance, Labour Market Interventions, HIV prevalence

So to inform the conceptual framework we concluded...

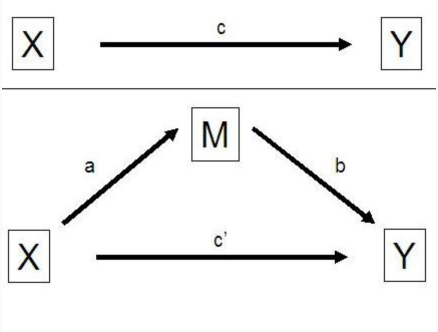
- TB is associated with poverty
- Social Insurance is negatively correlated with TB Incidence
- Diabetes & Smoking might measure one construct ('health behaviour')

Oxlade & Murray (2012)

Oxlade & Murray (2012) took self-reported TB status from the DHS survey in India as well as data about relevant covariates

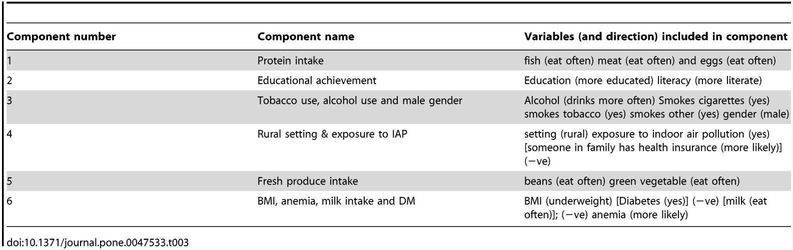
- Used PCA for variable reduction before an *assessment for mediation* to try and unpack the complex construct of poverty (X) on TB (Y).

Recall: What is the difference between mediation and confounding?



- So rather than testing every social variable (M), they tested some *linear combinations of variables* (PCs) to see whether they made the effect of c' smaller

This results in the following 6 principal components:



Component number	Component name	Variables (and direction) included in component
1	Protein intake	fish (eat often) meat (eat often) and eggs (eat often)
2	Educational achievement	Education (more educated) literacy (more literate)
3	Tobacco use, alcohol use and male gender	Alcohol (drinks more often) Smokes cigarettes (yes) smokes tobacco (yes) smokes other (yes) gender (male)
4	Rural setting & exposure to IAP	setting (rural) exposure to indoor air pollution (yes) [someone in family has health insurance (more likely)] (−ve)
5	Fresh produce intake	beans (eat often) green vegetable (eat often)
6	BMI, anemia, milk intake and DM	BMI (underweight) [Diabetes (yes)] (−ve) [milk (eat often)]; (−ve) anemia (more likely)

doi:10.1371/journal.pone.0047533.t003

- We can clearly see that the data complexity has been reduced: fish, meat, and egg intake can be summarised as one construct of protein intake
- Note though that there are three distinct nutrition factors...we might have grouped these in a conceptual framework!

- The nutritional pathways are probably more complex than we thought they were - impact on TB?
- Are the items on PC6 measuring related constructs?
- We could consider working backwards to make a conceptual framework from the results of the PCA

Note:

- Percentage of variance explained only in appendix (55.9%); HIV omitted; crowding and diabetes unaccounted for in chosen PCs

Question

PCs were retained if they displayed eigenvalues of greater than one. **Why?**

When to use PCA (and its Alternatives)

You should now...

✓ Be able to read & understand an article that uses PCA

When might you be compelled to use PCA? Simple answer:

Key Message

Whenever you want to summarise data to reduce its complexity, while still keeping the essential features of the data.

Notes and caveats:

- PCA is not the same as ‘factor analysis’ though the two are used interchangeable and very related
- Simply put, factor analysis is PCA where the components may still be correlated with one another
- PCA is built for continuous data; categorical data should be made binary, ordinal treated as continuous
- If you have a lot of categorical data consider using Multiple Correspondence Analysis or ‘Factor Analysis of Mixed Data’

You should now...

✓ Know when and why to use PCA

Any questions??

Thank you!

Intro to Practical

Constructing an Asset Index

We will be constructing an asset index as laid out in (Vyas & Kumaranayake, 2006), using data from the 2007 Tanzania AIDS indicator survey.

The data has been restricted to young women from rural areas with a recorded HIV status. We assume that the first PC represents wealth or SEP.

Specific steps we'll be undertaking:

- Selection of asset variables
- Applying PCA
- Classification into socio-economic groups

Selecting Variables

Key takeaways: choose carefully & always do your descriptive statistics first

- Look at means, frequencies, standard deviations...
- Test some correlations with variables you know to be meaningful
- Try to include assets good for PCA a priori (i.e. *minimise error*, *maximise variance*...)
- Check for 'clumping'
- No standard number of variables to include; do some sensitivity analyses

Doing the PCA

Key takeaways: choose how you will handle your data first; ensure the first PC measures wealth

- Ensure your categorical variables are coded correctly
- *A priori* specify how you will handle missing data
- Test a few different variable combinations in your PCA to see how it affects the eigenvectors (weights in Vyas paper) & eigenvalues
- Reconstruct the SEP score from the loadings on the first principal component
- Check what the second principal component measures

Using the SEP score

Key takeaways: Examine the score to check it matches your intuitions about the data

- Generates a continuous score, which may be useful for regression
- Consider categorising to get more intuitive understanding from continuous score
- Check the distribution of the score and look for skew
- Again check for clumping between different populations
- Check for internal coherence ('do my loadings make sense')

References

I Hurtado D, Kawachi I, Sudarsky J. “Social capital and self-rated health in Colombia: The good, the bad and the ugly.” *Social Science & Medicine*. 2011 Feb;72(4):58490.

II Logie CH, Earnshaw V. “Adapting and Validating a Scale to Measure Sexual Stigma among Lesbian, Bisexual and Queer Women.” *PLoS ONE*. 2015 Feb 13;10(2):e0116198.

III Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. “Drivers of tuberculosis epidemics: The role of risk factors and social determinants.” *Social Science & Medicine*. 2009 Jun ; 68(12):2240-6.

IV Navarro Silvera SA, Mayne ST, Risch HA, Gammon MD, Vaughan T, Chow W-H, et al. “Principal Component Analysis of Dietary and Lifestyle Patterns in Relation to Risk of Subtypes of Esophageal and Gastric Cancer.” *Annals of Epidemiology*. 2011 Jul;21(7):54350.

V Oxlade O, Murray M. “Tuberculosis and Poverty: Why Are the Poor at Greater Risk in India?” *PLoS ONE*. 2012 Nov 19; 7(11).

VI Vyas S, Kumaranayake L. “Constructing socio-economic status indices: how to use principal components analysis.” *Health Policy Plan*. 2006 Nov 1;21(6):45968.