NTNU | Norwegian University of Science and Technology

PRINCIPAL COMPONENT ANALYSIS WITH MISSING DATA

Torstein Nordgård-Hansen, Hanne Siri Amdahl Heglum, Anette Fagerheim Bjerke

September 23, 2024

Contents

Introduction

- Ideal world = perfect data
- Real world = messy!
  - Incomplete records
  - Not all questions answered (in questionnaires)
  - Busted sensors
  - The "trust-me bro" problem (when you could not, in fact, trust them bro)
  - etc etc
- Missing data must be handled before PCA!

Missingess mechanism [2]

- Explicit or implicit assumptions about the process that caused the data to be missing
- Missing Completely at Random (MCAR)
  - Missing values scattered across the dataset completely at random (duh)
  - No relation between missing data and observed or unobserved information.
- Missing at Random (MAR)
  - “Missing data may depend on observed data but not on unobserved data”
  - Example: Different age groups have different response rates, but age is an observed variable, so missing values are scattered randomly within age groups
- Missing Not at Random (MNAR)
  - Not MCAR or MAR.
  - Missingness depends on a variable that was not included in the data collection, or on the value of the missing score itself.

Handling missing data

Many possible solutions, including but not limited to:

- Deletion methods
  - Listwise deletion
  - Pairwise deletion
- Single imputation
  - Mean
  - Regression
  - Stochastic Regression
- Sequential methods
  - Expectation maximization
  - Multiple imputation

Listwise deletion

- A case has more than one missing value? BIN IT!
- Simple and efficient, but...
- A bit wasteful...
- Missingness must be MCAR

Pairwise deletion [4]

- Computes each individual covariance/correlation from the cases with observed values on both variables
- Instead of dropping all rows with any missing values, we only drop rows if the element of interest 'right now' is missing
- Less wasteful than listwise deletion, but...
- Also assumes MCAR missingness
- Each covariance/correlation could be based on different subsets / different number of cases
- Resulting covariance matrix may not be positive definite
- Computational issues

Single imputation methods

- Mean imputation [1]
  - A quick fix
  - Not a good solution
  - Disturbs the relations between the variables
  - Adds bias to any estimate when data is not MCAR
- Regression imputation [1]
  - Build a model based on the observed data and impute the predictions
  - Correlations are upward biased and variability is underestimated
  - Unbiased estimates of the mean when the missing data is MCAR
  - Too good to be true imputations
- Stochastic Regression imputation [6]
  - Extends the Regression Imputation
  - Adds a varying component to the predictions
  - The imputed values have the same variance as the observed values.

Expectation maximization[3]

- Use the Expectation Maximization algorithm to obtain maximum likelihood estimates of means and covariances
- Use the EM-estimated covariance matrix as basis for the PCA
- Assumes multivariate normality, but described as "robust to violations of this assumption"[3]
- Robust to MAR, but not MNAR.

Overview of multiple imputation methods

- Create multiple imputed datasets
- Perform PCA on each separately
- Recombine to one solution

Motivation

This ensemble method gives both decent results and uncertainty estimates for the loadings [2, 1].

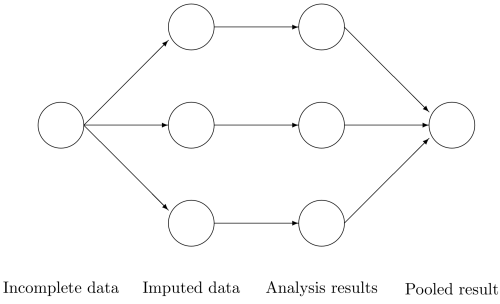

Figure by van Buuren [2].

Types of data imputation

- Several different methods can be used
- Models must contain some stochastic elements to create different imutations
- Ideally, a model should be picked based on the known structure of the data
- At the end of the day, making up data is subjective

The problem of recombining

- Averaging the results can give several problems [2]:
  - The component order may vary for similarly important components
  - Signs of componentsmay be flipped because of opposite signed parameters
  - Results will be slightly rotated
- Can be solved with for example Procrustes analysis [5, 2]
- Confidence interval for the loadings can be calculated as convex hulls, showing the uncertainty introduced by missing data

General advantages and disadvantages

Advantages

- Takes into account statistical stability [3]
- Not as prone to computational problems [2]
- Creates multiple imputated datasets as a byproduct

Disadvantages

- Requires data imputation and PCA to be performed many times, increasing computational load [3]
- May not always yield better results than simpler methods [2].

Sources I

[1] Stef van Buuren. *Flexible imputation of missing data*. eng. Boca Raton, 2018. URL: https://stefvanbuuren.name/fimd/sec-simplesolutions.html.

[2] Joost R. van Ginkel. “Handling Missing Data in Principal Component Analysis Using Multiple Imputation”. In: *Essays on Contemporary Psychometrics*. Ed. by L. Andries van der Ark, Wilco H. M. Emons, and Rob R. Meijer. Cham: Springer International Publishing, 2023, pp. 141–161. ISBN: 978-3-031-10370-4. DOI: 10.1007/978-3-031-10370-4_8.

Sources II

[3] Joost R. van Ginkel and Pieter M. Kroonenberg. “Using Generalized Procrustes Analysis for Multiple Imputation in Principal Component Analysis”. In: *Journal of classification* 31.2 (2014), pp. 242–269. ISSN: 0176-4268. DOI: 10.1007/s00357-014-9154-y.

[4] van Ginkel J. R.; Linting M.; Rippe R. C. A.; van der Voort A. “Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data”. In: *Journal of Personality Assessment* 102.3 (2019), pp. 297–308. DOI: https://doi.org/10.1080/00223891.2018.1530680.

[5] Julie Josse, Jérôme Pagès, and François Husson. “Multiple imputation in principal component analysis”. In: *Advances in data analysis and classification* 5.3 (2011), pp. 231–246. ISSN: 1862-5347. DOI: 10.1007/s11634-011-0086-7.

Sources III

[6] Michael J. Puma; Robert B. Olsen; Stephen H. Bell;Cristofer Price. *Technical Methods Report: What to Do When Data Are Missing in Group Randomized Controlled Trials*. 2009. URL: https://ies.ed.gov/ncee/pubs/20090049/section_3a.asp (visited on 09/02/2024).