Missing Data — Imputation etc

Notes from 2015-10-03

- Netflix Challenge –> becomes "fashion"
- aka "Matrix completion"

Literature

Books and Papers

- Schafer & Graham Missing Data: Our View of the State of the Art, American Psychological Association. (Schafer and Graham 2002)

- (Little and Rubin 2002) Roderick J.A. Little and Donald B. Rubin Statistical Analysis with Missing Data (Wiley; 2nd ed., 2002) Freely available (chapter wise) via ETH Library

- Joseph L. Schafer seminal monograph Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data, Chapman & Hall, London. (Schafer 1997)

Came with 4 S-PLUS "packages" called NORM, CAT, MIX, PAN} –> http://sites.stat.psu.edu/~jls/misoftwa.html

Versions of all four of these are available on CRAN:

```
sapply(c("norm", "cat", "mix", "pan"), packageDescription)
```

```
## $norm
## Package: norm
## Version: 1.0-9.5
## Date: 2013/02/27
## Title: Analysis of multivariate normal datasets with missing
##        values
## Author: Ported to R by Alvaro A. Novo <alvaro@novo-online.net>.
##        Original by Joseph L. Schafer <jls@stat.psu.edu>.
## Maintainer: John Fox <jfox@mcmaster.ca>
## Description: Analysis of multivariate normal datasets with missing
##        values
## License: file LICENSE
## URL: http://www.stat.psu.edu/~jls/misoftwa.html#aut
## Repository: CRAN
## Repository/R-Forge/Project: norm
## Repository/R-Forge/Revision: 8
## Repository/R-Forge/DateTimeStamp: 2013-02-27 16:01:38
## Date/Publication: 2013-02-28 07:11:32
## Packaged: 2013-02-27 19:16:19 UTC; rforge
## NeedsCompilation: yes
## License_restricts_use: no
## Built: R 3.2.2; x86_64-pc-linux-gnu; 2015-11-02 20:38:47 UTC; unix
##
## -- File: /home/david/R/x86_64-pc-linux-gnu-library/3.2/norm/Meta/package.rds
```

```
##
## $cat
## Package: cat
## Version: 0.0-6.5
## Date: 2012-10-30
## Title: Analysis of categorical-variable datasets with missing
##        values
## Author: Ported to R by Ted Harding and Fernando Tusell. Original
##        by Joseph L. Schafer <jls@stat.psu.edu>.
## Maintainer: Fernando Tusell <fernando.tusell@ehu.es>
## Description: Analysis of categorical-variable with missing values
## License: file LICENSE
## URL: http://www.stat.psu.edu/~jls/misoftwa.html#aut
## Packaged: 2012-10-30 16:26:21 UTC; root
## Repository: CRAN
## Date/Publication: 2012-10-30 18:21:53
## NeedsCompilation: yes
## License_restricts_use: no
## Built: R 3.2.2; x86_64-pc-linux-gnu; 2015-11-02 20:38:48 UTC; unix
##
## -- File: /home/david/R/x86_64-pc-linux-gnu-library/3.2/cat/Meta/package.rds
##
## $mix
## Package: mix
## Version: 1.0-9
## Date: 2015-06-29
## Title: Estimation/Multiple Imputation for Mixed Categorical and
##        Continuous Data
## Author: Original by Joseph L. Schafer <jls@stat.psu.edu>.
## Maintainer: Brian Ripley <ripley@stats.ox.ac.uk>
## Depends: stats
## Description: Estimation/multiple imputation programs for mixed
##        categorical and continuous data.
## License: Unlimited
## LazyData: yes
## URL: http://www.stat.psu.edu/~jls/misoftwa.html
## NeedsCompilation: yes
## Packaged: 2015-06-29 10:15:45 UTC; ripley
## Repository: CRAN
## Date/Publication: 2015-06-29 12:42:26
## Built: R 3.2.2; x86_64-pc-linux-gnu; 2015-11-02 20:38:48 UTC; unix
##
## -- File: /home/david/R/x86_64-pc-linux-gnu-library/3.2/mix/Meta/package.rds
##
## $pan
## Package: pan
## Version: 1.3
## Date: 2015-02-10
## Title: Multiple Imputation for Multivariate Panel or Clustered
##        Data
## Author: Original by Joseph L. Schafer
## Maintainer: Jing hua Zhao <jinghua.zhao@mrc-epid.cam.ac.uk>
## Suggests: mitools, lme4
## LazyData: Yes
## LazyLoad: Yes
## Description: Multiple imputation for multivariate panel or
##        clustered data.
## License: Unlimited
## URL: http://www.stat.psu.edu/~jls/misoftwa.html
## Packaged: 2015-02-10 22:15:01 UTC; jhz22
## NeedsCompilation: yes
## License_restricts_use: no
## Repository: CRAN
## Date/Publication: 2015-02-10 23:56:30
## Built: R 3.2.2; x86_64-pc-linux-gnu; 2015-11-02 20:38:49 UTC; unix
##
## -- File: /home/david/R/x86_64-pc-linux-gnu-library/3.2/pan/Meta/package.rds
```

R packages

Bioconductor:

impute : Hastie et al: knn.impute() Based on missing.pdf paper, Hastie et al. (1999).

CRAN

CRAN task view 'Multivariate' has section **Missing data** (which seems *not* comprehensive to me):

```
[mitools] provides tools for multiple imputation,
[mice] provides multivariate imputation by chained equations
[mvnmle] provides ML estimation for multivariate normal data with missing values,
[mix] provides multiple imputation for mixed categorical and continuous data.
[pan] provides multiple imputation for missing panel data.
[VIM] provides methods for the visualisation as well as imputation of missing data.
aregImpute() and transcan() from [Hmisc] provide further imputation methods.
[monomvn] deals with estimation models where the missing data pattern is monotone.
```

Schafer's ("norm", "cat", "mix", "pan") – see above

Lumley's 'mitools'

imputeR

softImpute:

```
Title: Matrix Completion via Iterative Soft-Thresholded SVD
Version: 1.4
Date: 2015-2-13
Author: Trevor Hastie and Rahul Mazumder
```

imputation : Archived in 2014 (policy violation: running on all cores)

- by Jeff Wong on Github
- also mentions the important paper by Cai, Candes, Shen et al (preprint on ArXiv), *Singular Value Thresholding Algorithm for Matrix Completion*

(We)blogs etc on R packages:

Amelia

Mad (Data) Scientist

Notes from 2015-10-15 meeting

- Focus on either continuous or mixed case
- Find data sets that are historically relevant:
  - See in the packages if there is a common package
  - Use the data sets in the appendix of (Schafer 1997), FLAS inside the `miP` package.
  - Use the synthetic bivariate set of (Schafer and Graham 2002), (normal and cluster).
  - Data sets: titanic, iris, mtcars, yeast (Lichman 2013), (noisy) time series data?

- Concentrate on the modern methods for next meeting (Troyanskaya et al. 2001) and (Hastie et al. 1999).
- Keep up with reading.
- Notation should be consistent, if possible with either (Schafer 1997) or (Little and Rubin 2002).
- Next meeting 2015-11-02 at 3pm.

Note from 2015-11-02 meeting

- vim package. See article. It allows to display the pattern of missingness.
- Start playing with mice, mix and pan packages.
- KNN: The weights in KNN are usually $w(d) = 1/d$, however nothing prevent us of having $w(d) = K(d)$ where K is any valid Kernel.
- Euclidian distance:
  - On which observations do we perform the comparison?
  - What type of standardization?
- MCAR, MAR and MNAR are an univariate properties: columns in a data matrix might have different assumptions on their missingness mechanism.
- 2014: Find paper on Weighted KNN ofr missing data.
- MICE used chained equations, AMELIA is clearly bayesian.
- Health and NHANES data sets are used widely.
- Use 20 imputation for the FLAS data sets to complete the missing variables and use it as the reference data set.
- Next meeting 2015-11-16 at 11am.

Note from 2015-11-02

- Using the FLAS data set, we should average the imputation with several packages.
- Use data frequency of the missing pattern to create artificial missingness.
- MCAR as well.
- Next meeting 2015-11-23 at 11am.

Bibliography

Hastie, Trevor, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. 1999. "Imputing Missing Data for Gene Expression Arrays."

Lichman, M. 2013. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Yeast.

Little, RJA, and DB Rubin. 2002. "Statistical Analysis with Missing Data." Wiley.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* CRC press.

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2). American Psychological Association: 147.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. "Missing Value Estimation Methods for DNA Microarrays." *Bioinformatics* 17 (6). Oxford Univ Press: 520–25.