Code for "Bayesian factor models for probabilistic cause of death assessment with verbal autopsies"

We describe Ox codes for estimtion of distributions of causes of death using the Bayesian factor model (bf.ox) and the conditionally independent model (ci.ox), and for estimation of associations in the proposed model (association.ox). Ox is a fast matrix language for statistical analysis (Doornik, 2007), and it can be downloaded from the official website[1].

1 bf.ox

bf.ox is a code for esimation of distributions of causes of death by the proposed method. It returns a MCMC sample of the cause specific mortality fraction (CSMF) accuracy and distributions of causes of death in a target site. The prior distributions are described in the manuscript.

- Line 15: set seed.
- Line 19: set sampling option.
  - nsim: number of collected MCMC samples.
  - th: number of skimming such that nsim $\times$ th = total number of MCMC iterations.
  - nburn: number of burn-in period.
  - K: number of factors.
  - R: number of the Monte Carlo simulation for evaluation of $\pi(x \mid y)$.
  - L: number of causes in the data.

[1] http://www.doornik.com/download.html

- Line 28: load data.
  - test.csv: test data with the cause of death in the first column and VA binary questions in the other columns. In a case that true causes are not known in a target site, put any cause vector in the first column.
  - training.csv: training data with the cause of death in the first column and VA binary questions in the other columns.
  - the numbering should start from zero such that $y \in \{0, \ldots, 33\}$ if data have 34 causes of death.
  - a missing value should be coded as 999.
- Line 65: MCMC sampling of parameters and latent variables.
  - sample $\mu_{\cdot j}$ (line 73), $\lambda_{cj\cdot}$ (line 89), η_i (line 104), τ_j (line 113), ϕ_j (line 118), z_{ij} (line 126).
- Line 155: estimate distributions of causes of death in a target site.
- Line 180: output result.
  - result.csv: a MCMC sample of CSMF (first columns) and the distribution of causes (other columns).

2 ci.ox

ci.ox is a code for esimation of distributions of causes of death by the conditionally independent model. It returns a MCMC sample of the CSMF accuracy and distributions of causes of death in a target site.

- Line 12: set seed.
- Line 16: set sampling option.
  - nsim: number of collected MCMC samples.
  - th: number of skimming such that nsim $\times$ th = total number of MCMC iterations.

  - nburn: number of burn-in period.
  - L: number of causes in the data.

- Line 23: load data.
  - test.csv: test data with the cause of death in the first column and VA binary questions in the other columns. In a case that true causes are not known in a target site, put any cause vector in the first column.
  - training.csv: training data with the cause of death in the first column and VA binary questions in the other columns.
  - the numbering should start from zero such that $y \in \{0, \ldots, 33\}$ if data have 34 causes of death.
  - a missing value should be coded as 999.

- Line 64: sample $\pi(x_j \mid y)$ with beta(1,1) prior from

$$\text{beta}\left(1 + \sum_{i:y_i=y} 1(x_{ij} = 0, m_{ij} = 0), 1 + \sum_{i:y_i=y} 1(x_{ij} = 1, m_{ij} = 0)\right).$$

- Line 82: estimate distributions of causes of death in a target site.

- Line 105: output result.
  - result-ci.csv: a MCMC sample of CSMF (first columns) and distributions of causes (other columns).

3 association.ox

association.ox is a code for estimation of associations of predictors with causes of death in the proposed method. It returns a MCMC sample of δ for each predictor. The prior distributions are described in the manuscript.

- Line 14: set seed.

- Line 18: set sampling option.

  - nsim: number of collected MCMC samples.
  - th: number of skimming such that nsim $\times$ th = total number of MCMC iterations.
  - nburn: number of burn-in period.
  - K: number of factors.
  - R: number of the Monte Carlo simulation for evaluation of $\pi(x \mid y)$.
  - L: number of causes in the data.

- Line 27: load data.
  - data.csv: data with the cause of death in the first column and VA binary questions in the other columns.
  - the nubering should start from zero such that $y \in \{0, \ldots, 33\}$ if data have 34 causes of death.
  - a missing value should be coded as 999.

- Line 50: MCMC sampling of parameters and latent variables.
  - sample $\pi(y)$ (line 58), sample $\mu_{\cdot j}$ (line 63), $\lambda_{cj\cdot}$ (line 79), η_i (line 94), τ_j (line 103), ϕ_j (line 108), z_{ij} (line 116).

- Line 145: compute δ.

- Line 177: output result.
  - result-association.csv: a MCMC sample of δ for each predictor.

References

Doornik, J. A. (2007). *Object-oriented matrix programming using Ox, 3rd ed.* London: Timberlake Consultants Press and Oxford: www.doornik.com.