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 proposed model (COD.ox) and the conditional independent model (CI.ox), and for estimation of conditional mutual information in the proposed model (CMI.ox). Ox is a fast matrix language for statistical analysis (Doornik, 2007), and it can be downloaded from the official website[1].

1 COD.ox

COD.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 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.

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

- Line 27: load data.
  - target.csv: data in a target site 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: data in a training site 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 66: MCMC sampling of parameters and latent variables.
  - sample $\mu_{\cdot j}$ (line 70), $\lambda_{cj\cdot}$ (line 86), η_i (line 101), τ_j (line 110), ϕ_j (line 115), z_{ij} (line 123).
- Line 152: estimate distributions of causes of death in a target site.
- Line 180: output result.
  - result.csv: a MCMC sample of CSMF (first columns) and distributions of causes (other columns).

2 CI.ox

CI.ox is a code for esimation of distributions of causes of death by the conditional 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.
  - target.csv: data in a target site 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: data in a training site 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 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 CMI.ox

COD.ox is a code for estimation of conditional mutual information of each predictor given all others in the proposed method. It returns a MCMC sample of the conditional mutual information of all predictors. 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.
  - the code uses only VA questions with the missing rate less than 5%.
- Line 57: MCMC sampling of parameters and latent variables.
  - sample $\pi(y)$ (line 61), sample $\mu_{\cdot j}$ (line 66), $\lambda_{cj\cdot}$ (line 82), η_i (line 97), τ_j (line 106), ϕ_j (line 111), z_{ij} (line 119).
- Line 148: impute missing values.
- Line 161: compute conditional mutual information.
- Line 208: output result.
  - result-cmi.csv: a MCMC sample of conditional mutual information of each predictor given all others.

References

Doornik, J. A. (2007). *Object-Oriented Matrix Programming Using Ox, 3rd ed.* London: Timberlake Consultants Press and Oxford.