Code for "Bayesian analysis of verbal autopsy data using factor models with age- and sex-dependent associations between symptoms"

We describe the Ox codes for the proposed model to estimate distributions of causes of death in a target site (BF-AS.ox) and to evaluate the relevance of predictors to causes of death (relevance.ox). In addition, we explain the code for the Bayesian factor model by Kunihama et al. (2020) (VA-BF.ox), which selects the number of factors using the cross validation method. Ox is a fast matrix language for statistical analysis (Doornik, 2007), and it can be downloaded from the official website[1].

1 BF-AS.ox

BF-AS.ox is the code for estimation of distributions of causes of death using the proposed method. It returns an MCMC sample of the cause specific mortality fraction (CSMF) in a target site. The default settings are detailed in the manuscript.

- Line 16: set seed.
- Line 20: set sampling option.
  - L: number of causes in the data.
  - K_s: upper bound of the number of factors with respect to associations shared between all age and sex groups (K) and to associations changing by age and sex (G). Assuming $K = G$, we select a best value of K from $\{1, \ldots, \text{K\_s}\}$ based on the cross validation method.
  - nsim: number of collected MCMC samples.

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

  - th: number of skimming such that nsim $\times$ th = total number of MCMC iterations.
  - nburn: number of burn-in period.
  - T: number of holds in the cross validation.
  - R: number of the Monte Carlo simulation for evaluation of $P(x \mid y, \text{age}, \text{sex})$ in (7) in the manuscript.

- Line 30: load data.
  - test.csv: test data with the cause of death in the first column, age in the second column, sex in the third column and all other VA predictors in the following columns. All predictors should be binary. 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 same structure as the test data. All causes of death should be observed.
  - the numbering should start from zero such that $y \in \{0, \ldots, 4\}$ if data have 5 causes of death.
  - a missing value should be coded as 999.

- Line 104: MCMC sampling of parameters and latent variables.
  - sample β_{yj} (line 108), $\tilde{\eta}_i$ (line 137), ϕ_B (lines 148 and 153), ϕ_Λ (line 160), ϕ_C (line 166), z_{ij} (line 174), $P(\text{age}, \text{sex} \mid y)$ (line 194) and missing values of age and sex (lines 203 and 222).

- Line 245: estimate distributions of causes of death in the target site.

- Line 180: output result.
  - BF-AS-result.csv: the MCMC sample of the CSMF in the test data.

2 relevance.ox

relevance.ox is the code for evaluation of the relevance of predictors to causes of death using the proposed model. It returns an MCMC sample of the mutual information and the

conditional mutual information for each predictor. The default settings are detailed in the manuscript.

- Line 21: set seed.
- Line 25: set sampling option.
  - L: number of causes in the data.
  - K_s: upper bound of the number of factors with respect to associations shared between all age and sex groups (K) and to associations changing by age and sex (G). With the assumption $K = G$, we select a best value of K from $\{1, \ldots, \text{K\_s}\}$ based on the cross validation method.
  - 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.
  - T: number of holds in the cross validation.
  - R: number of the Monte Carlo simulation for evaluation of $P(x \mid y, \text{age}, \text{sex})$ in (7) in the manuscript.
  - R2: number of the Monte Carlo simulation for evaluation of CMI in (8) in the manuscript.
- Line 36: load data.
  - data.csv: data with the cause of death in the first column, age in the second column, sex in the third column and all other VA predictors in the following columns. All predictors should be binary and all causes of death should be observed.
  - the numbering should start from zero such that $y \in \{0, \ldots, 4\}$ if data have 5 causes of death.
  - a missing value should be coded as 999.
- Line 110: MCMC sampling of parameters and latent variables.

  - sample β_{yj} (line 114), $\tilde{\eta}_i$ (line 143), ϕ_B (lines 154 and 159), ϕ_Λ (line 166), ϕ_C (line 172), z_{ij} (line 180), $P(\text{age}, \text{sex} \mid y)$ (line 2000), missing values of age and sex (lines 209 and 228) and $P(y)$ (line 247).

- Line 258: estimate distributions of causes of death to select the number of factors via the cross validation.

- Line 329: compute MI for each predictor.

- Line 361: compute CMI for each predictor.

- Line 177: output result.

  - result-mi.csv: the MCMC sample of MI for each predictor.
  - result-cmi.csv: the MCMC sample of CMI for each predictor.

3 VA-BF.ox

VA-BF.ox is the code for estimation of distributions of causes of death using the Bayesian factor model by Kunihama et al. (2020). In this code, the number of factors is selected via the cross validation. It returns an MCMC sample of the CSMF in a target site. The default settings are detailed in Kunihama et al. (2020).

- Line 16: set seed.

- Line 20: set sampling option.

  - L: number of causes in the data.
  - K_s: upper bound of the number of factors. We select its best value from $\{1, \ldots, \text{K\_s}\}$ based on the cross validation method.
  - 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.
  - R: number of the Monte Carlo simulation for evaluation of $\pi(x \mid y)$.

  - T: number of holds in the cross validation.

- Line 30: load data.

  - test.csv: test data with the cause of death in the first column and VA predictors in the other columns. All predictors should be binary. 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 same structure as the test data. All causes of death should be observed.
  - the numbering should start from zero such that $y \in \{0, \ldots, 4\}$ if data have 5 causes of death.
  - a missing value should be coded as 999.

- Line 206: output result.

  - BF-result.csv: the MCMC sample of the CSMF accuracy (first column) and the CSMF (other columns) in the test data.

References

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

Kunihama, T., Z. R. Li, S. J. Clark, and T. H. McCormick (2020). Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *The Annals of Applied Statistics 14*, 241–256.