A Multivariate Mixed Membership Model for malaria risk detection

Massimiliano Russo, David B. Dunson & Burton H. Singer
russo@stat.unipd.it

Department of Statistical Sciences, University of Padua, Italy

Introduction & motivation

Malaria infection risk is largely driven by human behavior and its evaluation requires considerations on biological and ecological aspect juxtaposed with **behavioral and environmental conditions**.

The study of **social surveys** allows targeting conditions that favor malaria diffusion, determining **risk profiles**, and driving malaria mitigation and prevention strategies.

We propose a novel **Multivariate Mixed Membership model** taking into account different sets of correlated risk-related variables.

In our application we consider to broad risk domains

- **Behavioural risk** (e.g. use of insecticide and presence/absence of crop on the settle).
- **Environmental risk** (e.g. quality of roof, walls and housing).

Multivariate Mixed Membership Models

The proposed model can be expressed using the following hierarchical representation

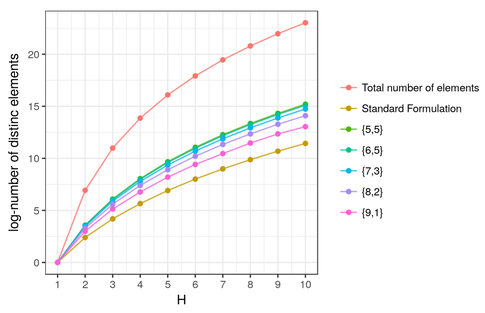
$$\begin{aligned} X_{ij} \mid Z_{ij} = h &\sim \text{Cat}(\theta^{(j)}_{h1}, \dots, \theta^{(j)}_{hd_j}) \\ Z_{ij} \mid \boldsymbol{\lambda}_i^{(\boldsymbol{g_j})} &\sim \text{Cat}(\lambda^{(g_j)}_{i1}, \dots, \lambda^{(g_j)}_{iH}) \\ (\boldsymbol{\lambda}_i^{(\mathbf{1})}, \dots, \boldsymbol{\lambda}_i^{(\boldsymbol{G})}) &\sim P \end{aligned} \tag{1}$$

- $g_j \in \{1, \dots, G\}$ for $j = 1, \dots, p$ is a group indicator for the variables.
- P is a multivariate distribution for the profiles such that each $\boldsymbol{\lambda}_i^{(g)}$ is defined on a simplex.

Integrating out the scores vector from equation we obtain the **population level model**.

$$\text{pr}(X_1 = x_1, \dots, X_p = x_p \mid \boldsymbol{\theta}) = \sum_{h_1=1}^{H} \cdots \sum_{h_p=1}^{H} \bar{a}_{h_1 \dots h_p} \prod_{j=1}^{p} \theta^{(j)}_{h_j x_j} \tag{2}$$

- $\bar{\mathcal{A}} = \{\bar{a}_{h_1 \dots h_p}; h_j = 1 \dots H; j = 1, \dots, p\}$ where $\bar{a}_{h_1 \dots h_p} = \mathbb{E}_P[\lambda^{(g_1)}_{ih_1} \cdots \lambda^{(g_p)}_{ih_p}]$.
- Equation (2) is an instance of Tucker decomposition with core tensor $\bar{\mathcal{A}}$ (**flexible and compact representation** of the probability mass function.) [e.g. 4, 1].



- $\bar{\mathcal{A}}$ has $\prod_{g=1}^{G} H^{\bar{p}_g}/p_g!$ distinct elements out of H^p (partial exchangeability assumption in (1)).
- The number of distinct elements in $\bar{\mathcal{A}}$ can be considerably bigger than $H^{\bar{p}}/p!$ (standard mixed membership with exchangeability assumption [3]).

Theorem 1. *Let π_0 be a probability tensor defined on the $d_1 \times \ldots \times d_p$ dimensional simplex, $\hat{\pi} = \{\sum_{h_1=1}^{H} \cdots \sum_{h_p=1}^{H} \bar{a}_{h_1,\dots,h_p} \prod_{j=1}^{p} u^{(j)}_{h_j i_j}, i_j = 1 : d_j; j = 1 : p\}$ the proposed model rank H approximation and $\tilde{\pi}$ be standard mixed membership model approximation of π_0 sharing the same tensor arms, then $\|\pi_0 - \hat{\pi}\|_F \leq \|\pi_0 - \tilde{\pi}\|_F$.*

Simulation Study

We consider four simulation scenarios relying on model (1) for different choices of the distribution P

Posterior mean of membership score vectors, for the proposed model (black dots) and separate model for the 2 domains using the R package `mixedMem` (red crosses). The shaded area represents the contour of the true profile distribution.

Data

We consider the case of **Machadinho Settlement Project**, located in Rondônia state, Western Brazilian Amazon (approved in 1982, occupied starting by late 1984 [2]).

- Malaria becomes a problem soon with the proliferation of Anopheles darlingi (principal diffusion vector).
- Data consists of 4 waves survey conducted in 1984, 1985, 1987 and 1995, administrated to $n = 1693$ subjects..
- A core of $p = 29$ survey items remained immutable in the considered years, and we focus on these lasts for the determination of social risk profiles.

Profile distributions

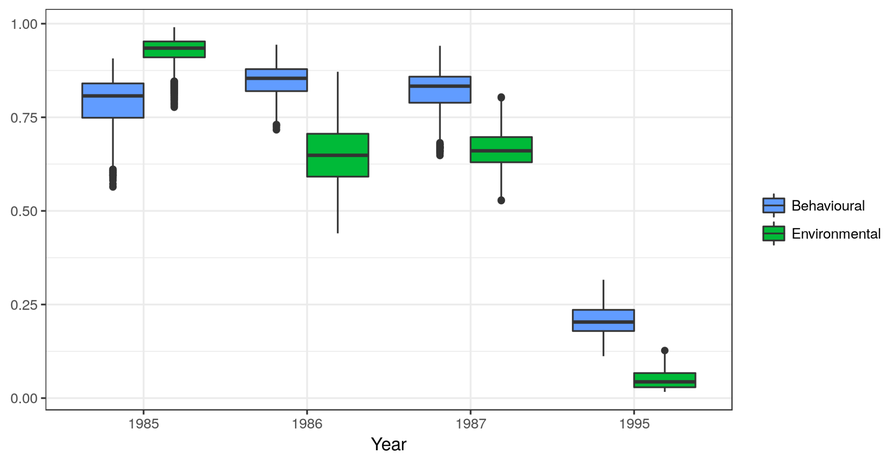
The profile distribution P is defined on the space $S = \text{conc}(S_{H_1}, \dots, S_{H_G})$, where $S_H = \{\boldsymbol{x} \in [0,1]^H : \sum_{h=1}^{H} x_h = 1\}$ be an H-dimensional simplex.
We consider a novel **Multivariate Logistic Normal distribution**
The probability density function of the given distribution is

$$f_X(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}^\star - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^\star - \boldsymbol{\mu})\right\}}{(2\pi)^{(H-1)G/2} |\boldsymbol{\Sigma}|^{1/2} \prod_{g=1}^{G} \prod_{h=1}^{H_g} x_h^{(g)}} \tag{3}$$

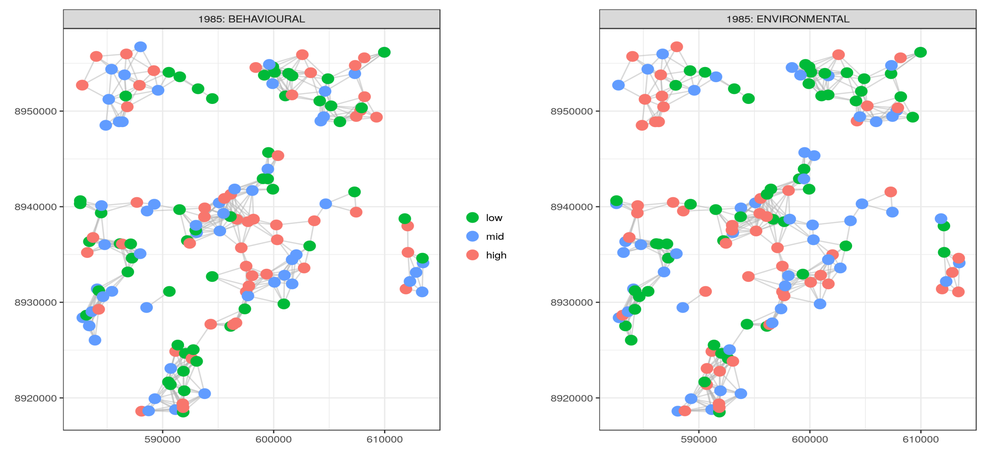
where the $\boldsymbol{x}^\star = \text{conc}\left(\left\{\log(x_h^{(g)}/x_{H_g}^{g}),\ \text{for } h = 1, \dots, H_g - 1; g = 1, \dots, G\right\}\right)$.

- $\boldsymbol{\Sigma}$ expresses within group correlation (diagonal blocks) and between group correlation (off diagonal blocks) of the log-odds for the profiles.
- Invariance respect to: change of baseline category, permutation of the labels, and merging within group categories.
- Odds ratios and log odds ratios expectation are available in closed form.
- Covariate dependence can be easily introduced in the parameter $\boldsymbol{\mu}$ leveraging a multiple regression framework.

Application to Malaria Data



Posterior distribution of expected multivariate mixed membership scores for the considered domains and years.



Spatial distribution of the multivariate membership scores for 1985. The scores have been divided in risk classes using quantiles

References

[1] Bhattacharya, A. & Dunson, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association* **107**, 362–377.

[2] de Castro, M. C., Monte-Mór, R. L., Sawyer, D. O. & Singer, B. H. (2006). Malaria risk on the amazon frontier. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 2452–2457.

[3] Erosheva, E. A., Fienberg, S. E. & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics* **1**, 346.

[4] Kolda, T. G. & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51**, 455–500.