Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXI

Bayesian inference for tensor factorization models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Bruno Scarpa

Dottorando: Massimiliano Russo

Abstract

Multivariate categorical data are routinely collected in several applications, including epidemiology, biology, and sociology, among many others. Popular models dealing with these variables include log-linear and tensor factorization models, with these lasts having the advantage of flexibly characterizing the dependence structure underlying the data. Under such framework, this Thesis aims to provide novel approaches to define compact representations of the dependence structures and to introduce new inference possibilities in tensor factorization approaches. We introduce a new class of *GROuped Tensor (GROT)* factorizations, which have superior performance in terms of data compression if compared to standard Parafac approach, using relatively few components to represent the joint probability mass function of the data. While popular Parafac factorizations rely on mixing together independent components, GROT mixes together grouped factorizations, equivalent to replacing vector arms in Parafac with low-dimensional tensor arms. We consider a Bayesian approach to inference with Dirichlet priors on the mixing weights and arm components, to obtain a combined low-rank and sparse structure, while facilitating efficient posterior computation via Markov chain Monte Carlo. Motivated by an application on malaria risk assessment, we also introduce a novel multivariate generalization of mixed membership models, which allows identification of correlated profiles related to different domains corresponding to separate groups of variables. We consider as a case study the Machadinho settlement project in Brazil, with the aim of defining survey based environmental and behavioral risk profiles and studying their interaction and evolution. To achieve this goal, we show that the use of correlated multiple membership vectors leads to interpretable inference requiring a lower number of profiles compared to standard formulations while inducing a more compact representation of the population level model. We propose a novel multivariate logistic normal distribution for the membership vectors, which allows easy introduction of auxiliary information in the membership profiles leveraging a multivariate latent logistic regression. A Bayesian approach to inference, relying on Pólya gamma data augmentation, facilitates efficient posterior computation via Markov chain Monte Carlo. The proposed approach is shown to outperform the classical mixed membership model in simulations, and the malaria diffusion application.

Sommario

I dati categoriali multivariati sono raccolti e utilizzati per una vasta gamma di applicazioni, comprendenti, tre le molte altre, epidemiologia, biologia e sociologia. Tra gli approcci più usati nell'analisi di queste variabili troviamo i modelli log-lineari e quelli per la fattorizzazione tensoriale, dove questi ultimi hanno il vantaggio di rappresentare in maniera flessibile la struttura di dipendenza sottostante i dati. Muovendosi in questo contesto, la tesi si propone di presentate nuovi approcci atti a definire in maniera compatta la struttura di dipendenza e ad introdurre nuove metodologie inferenziali nell'ambito della fattorizzazione tensoriale. Si introduce una nuova classe di fattorizzazione tensoriale raggruppata (GROT), che presenta proprietà migliori in termini di compressione rispetto agli approcci standard, quali la Parafac, utilizzando meno componenti per rappresentare la funzione di massa di probabilità dei dati. Mentre la più nota fattorizzazione di tipo Parafac associa componenti indipendenti, GROT associa fattorizzazioni raggruppate, che equivale a sostituire i vettori costituenti le componenti della Paracac, con tensori a ridotte dimensioni. Si considera un approccio bayesiano all'inferenza basandosi su priori coniugate di tipo Dirichlet sia sulle componenti di mistura, che sulle componenti della fattorizzazione, in modo da ottenere simultaneamente una rappresentazione sparsa e a rango ridotto, e facilitare la stima del modello tramite catene di Markov Monte Carlo. Partendo da un'applicazione sulla diffusione della malaria, si introduce, inoltre, una nuova generalizzazione multivariata dei modelli di tipo *Mixed Membership*, consentendo l'identificazione di profili correlati a diversi domini che corrispondono a gruppi di variabili distinti. Come caso studio si considera il progetto di insediamento di Machadinho in Brasile, con lo scopo di identificare profili di rischio ambientali e comportamentali basati su sondaggi, e di studiarne l'evoluzione e interazione. Per il conseguimento di questo obiettivo si mostra che l'uso di profili di rischio correlati da luogo a risultati interpretabili richiedendo un minor numero di profili latenti e introducendo una rappresentazione parsimoniosa del modello a livello di popolazione. Si propone una nuova distribuzione logit-normale multivariata per i vettori di appartenenza, che consente di introdurre informazione aggiuntive facendo leva su regressioni logistiche latenti. Un approccio bayesiano alla stima dei parametri, basato sulla distribuzione Pólya gamma facilità la simulazione dalla a posteriori tramite MCMC. Si mostra come il metodo proposto migliora la formulazione standard in simulazioni nell'applicazione proposta sulla diffusione della malaria.

Acknowledgements

First I want to thank my PhD advisor, Professor Bruno Scarpa since he has been always present, also beyond research. He has been a great supporter and motivator in these years of PhD life.

The second person I really want to thank is Professor David B. Dunson since he has been so generous and supportive in my PhD studies, allowing me to spend a wonderful period at Duke University, teaching me a lot about research. His passion for research has been a real stimulus for me.

A special thanks go to my (Phd) friends, in particular Sally, Charlotte, Emanulae, Alessandro, for sharing this adventure and spending this part of life together. I want also to thanks the PhD and Post docs I met while at Duke to have been so welcoming, and in particular Victor and Michael to have let me feel as at home.

I am also thankful to Tony and Daniele to have advised and helped me in many ways during these years.

My gratitude is also toward the department of Statistical Science of the University of Padova, for giving me the opportunity of being a PhD student, and to Professor Monica Chiogna and Nicola Sartori for directing the program.

Last to my family to have been always there.

Contents

List of Figures

List of Tables

Introduction

Overview

Multivariate categorical data are present in many fields of study, including psychology (e.g. Muthen and Christoffersson, 1981), social science (e.g. Santos *et al.*, 2015) and epidemiology (e.g. Landis *et al.*, 1988), among others. When dealing with this kind of variables many inferential tasks can be accomplished by modeling the *probability mass function (p.m.f.)* and its functionals.

Let $\boldsymbol{X} = (X_1, \dots, X_p)^T$, be a vector of categorical random variables, where $X_j \in \{1, \dots, d_j\}$ for $j = 1, \dots, p$. The sample space corresponding to the random variable $\boldsymbol{X}$ is the product space $\boldsymbol{\Omega} = \{1, \dots, d_1\} \times \cdots \times \{1, \dots, d_p\}$, having cardinality $|\boldsymbol{\Omega}| = \prod_{j=1}^{p} d_j$. We can immediately notice that the cardinality of such a space increases at an exponential rate with the number of variables p. Formally, the p.m.f. underlying the random variable $\boldsymbol{X}$ is defined as a function $\mathrm{p}(\omega) : \boldsymbol{\Omega} \longrightarrow [0, 1]$ for all $\omega \in \boldsymbol{\Omega}$, satisfying $\sum_{\omega \in \boldsymbol{\Omega}} \mathrm{p}(\omega) = 1$. Without any assumption we would have a parametric space $\boldsymbol{\Theta}$ having cardinality $|\boldsymbol{\Theta}| = |\boldsymbol{\Omega}| - 1$. Clearly $\mathrm{p}(\omega)$ can be parameterized by means of functions defined on a lower dimensional space. To accomplish such a goal arguably one of the most popular class of procedures is represented by log-linear models (e.g. Agresti, 2013) which directly parameterize the logarithm of the p.m.f. as a linear function encompassing the (conditional) dependence structure of the considered variables.

Log-linear models are a useful tool, but they require the specification of the dependence structure of the considered data. Learning such a structure is still feasible in small problems, but when the number of variables p increases the number of possible interactions that can be included in the model explodes. Sparsity can be explicitly incorporated, for example, Ntzoufras *et al.* (2000) and Nardi *et al.* (2012) proposed a Bayesian stochastic search and a group-lasso algorithm for model selection respectively. However, when p is moderate to large, these procedures require restrictions for computational tractability, potentially affecting flexibility. As a result, inference in log-linear models can be a cumbersome task in high-dimensional settings, where the dimension of the model space is huge.

By construction the p.m.f. $\{p(\omega); \omega \in \mathbf{\Omega}\}$ can be viewed as a *probability tensor* defined on a simplex S of dimension $|\mathbf{\Omega}| - 1$. A probability tensor is then a tensor π whose entries are non negative and sum up to 1. In view of this consideration, recent literature avoids pre-specifying the graphical dependence structure in the multivariate categorical data by means of tensor factorization.

Tensor factorization models represent a class of procedures to define low-rank approximations of multivariate tensor, not relying on highly restrictive parametric assumptions. These methods can be used to reconstruct $p(\omega)$ and its functional, relying on the low dimensional structures provided by tensor decomposition itself.

Among tensor factorization models, one of the most popular is the so called Parafac (e.g. Dunson and Xing, 2009) which defines the probability mass function for a multivariate categorical random vector as a mixture of products of multinomial distributions. This factorization provides a flexible and computationally tractable representation, but might lead to unsatisfactory characterization of the p.m.f. when the sample size is moderate to low compared to the number of variables. Tucker (Bhattacharya and Dunson, 2012) and C-Tucker (Johndrow *et al.*, 2017) factorizations mitigate such issue introducing a multivariate latent variable for the mixture weights, however, these lasts are computationally intensive to estimate, while subsequent inference is cumbersome if compared to the one provided by Parafac.

Tucker decomposition is intimately connected with *Mixed Membership (MM)* models (e.g. Airoldi *et al.*, 2014), whose main goal is to derive properties of individuals based on results of multivariate measurements, making use of subject-specific weights in a mixture model. MM or admixture models have been extensively used in many fields, and they received a particular attention in text mining for topic discovery analysis (e.g. Blei *et al.*, 2003). In this field different generalizations have been proposed, including efficient estimation methods through variational approximations (e.g. Mimno *et al.*, 2012) and different choices of the mixture weights distribution (e.g. Lafferty and Blei, 2006), however considerations on how to interpret models parameters have been for the most part ignored.

Main contributions of the thesis

In the context of Bayesian models for categorical variables this thesis aims to provide novel tools devoted to define: (i) compact representations of the dependence structure underlying the data; (ii) interpretable models exploiting new inference possibilities for tensor factorization.

In considering such goals, in the following Chapters, we introduce two new tensor factorization models, having different properties and goals. Specifically, in Chapter 1 , we introduce some basic definitions and properties of tensors and tensors algebra, mostly used in the subsequent models. In Chapter 2 we introduce a grouped tensor factorization which produces a flexible characterization of the dependence structure underlying the data, leading to a more compact representation in term of latent classes if compared to standard Parafac model. In Chapter 3, exploiting the connection between Tucker decomposition and MM models, we introduce a novel multivariate generalization of mixed membership models, which allows identification of correlated profiles related to different conceptual domains corresponding to separate groups of variables.

We show some theoretical properties of all the proposed approaches, highlighting their flexibility in simulations and real data applications.

Grouped tensor factorization

Parafac model is one of the preferred tensor factorization approach for categorical data, due to its simplicity and flexibility. The main advantage of this model is in relying on a simple functional form, being a mixture of independent multinomial distributions, which grants flexibility in many applications. Despite its generality, the simple assumption underlying Parafac models may lead to poor representation of the p.m.f., in particular when a block dependence structure is present in the data.

To overcome this issue, in Chapter 2 , we developed a new tensor factorization having m-way tensor arms (i.e. m-dimensional kernels in a mixture representation) with $m \geq 1$. We refer to this factorization as *GROuped Tensor (GROT)* reflecting the grouping of the variables in the tensor arms of the decomposition. GROT decomposition produces a flexible characterization of the dependence structure underlying the data, leading to a more compact representation in term of latent classes if compared to standard Parafac model. The main idea underlying this factorization model is to use a partition of the set of indices representing the observed variables, having blocks with low cardinality, and to represent the dependence structure in each block by means of simple models (e.g. Dirichlet-multinomial). The residual across blocks dependence is characterized by means of an univariate discrete latent variable, as in Parafac, and the induced model for the element of the p.m.f. is a mixture of multinomial distributions, but jointly for some of the variables in this case. Parafac model can be obtained as a special case where partitions' blocks have all cardinality one.

In Section 2.1 we formalize the model, proving its generality, flexibility, and compactness in representing the dependence structure with respect to Parafac factorization. The partition used to define a structure for tensor arms is usually unknown, hence in Section 2.2 we introduce a suitable prior on the space of partition based on product partition models (e.g. Hartigan, 1990) to learn such a structure directly from the data. We evaluate empirical performances of the proposed approach in challenging simulation scenarios, comparing the obtained results with standard Parafac factorization (Section 2.3). Finally, in Section 2.4 we consider an application to nucleotide sequences with the main aim being in identifying promoter sequences and contact bases for such process. We show that our proposed methodology produces interpretable results, and comparable predictive performances compared to commonly used statistical learning tools.

Multivariate Mixed Membership model

Exploiting the connection between Tucker decomposition and MM models, we introduce a novel multivariate generalization of mixed membership models, which allows identification of correlated profiles related to different conceptual domains corresponding to separate groups of variables. We show that the use of correlated multiple membership vectors leads to interpretable inference requiring a lower number of profiles compared to standard formulations, inducing a more compact representation of the population level model while allowing interpretable inference at subject-specific level.

Starting from multivariate observations, general MM models can be used to estimate subject-specific latent scores corresponding to the weights in a discrete mixture model. The key idea is that the population is composed of H pure types, or profiles, and that each subject is endowed with a vector $\boldsymbol{\lambda_i} = (\lambda_{i1}, \dots, \lambda_{iH})^T$, such that $\lambda_{ih} \in [0,1]$ and $\sum_{h=1}^{H} \lambda_{ih} = 1$, indicating the degree of similarity of subject i with each of the H profiles. The vector $\boldsymbol{\lambda_i}$ induces a soft clustering of the subjects across the multiple profiles, which is particularly appealing when an exact grouping is difficult if not impossible to obtain, as for example in identification of disease risks (e.g. Chuit *et al.*, 2001; Castro *et al.*, 2006) or political ideology (e.g. Gross and Manrique-Vallier, 2012). In many situations interpretability is one of the most important features; hence it is desirable to employ a small number of profiles, ideally $H = 2$.

To overcome these issues, in Section 3.2, we specify a class of multivariate mixed membership models (MMM) that explicitly includes the classification of blocks of variables corresponding to distinct subject matter domains and the cross-domain correlation structure. We address this goal by linking group-specific MM models through dependence in the membership scores. To effectively model multivariate membership scores, in Section 3.3 we propose a novel distribution defined on a product space composed of simplices, which allows easy introduction of auxiliary information in the membership profiles leveraging a multivariate latent regression. We shows the basic property of this distribution and consider a Bayesian approach to inference. In Section 3.5 we study the performance of our model under different simulation scenarios, and in Section 3.6 we apply the model to survey information concerning malaria diffusion in the Machadinho area.

Chapter 1

Tensors and their decomposition

In this Chapter we briefly review some mathematical notions and definitions in the field of tensors and tensor algebra. Tensors will be the one of the most used mathematical tool for the remaining of this Thesis, but their application goes far beyond what presented here. For this reason, Sections 1.1,1.2 and 1.3 are deliberately general in introducing tensors and some related algebra, and not targeted to the methodologies proposed in the next Chapters. Some models designed for categorical variables are instead discussed in Sections 1.4 and 1.5. The reader interested in a more formal and detailed introduction in tensor algebra can referrer to Kolda and Bader (2009) and references therein.

1.1 Mathematical definitions and basic properties

A tensor is a multidimensional array, or more precisely an element of a product space defined as the tensor product of p vector spaces. Formally, we can define a tensor $\mathscr{A} = \{a_{c_1 \ldots c_p}; c_j = 1, \ldots, d_j; j = 1, \ldots, p\} \in \mathbb{C}$, where $\mathbb{C} = C_1 \otimes \ldots \otimes C_p$ and d_j indicates the dimension of the space C_j. The spaces C_j composing the tensor product do not need to be homogeneous, meaning that each C_j can have its own coordinate system, and can be a general subset of the natural, real or complex numbers. To simplify the exposition, however, we consider the spaces to be homogeneous, although with possible different dimensions, which is equivalent in considering the spaces $C_j = C^{d_j}$ for $j = 1, \ldots, p$.

Some of the main characteristics of a tensor includes: the *mode* also known as dimension or order. A tensor of order one is a vector, a tensor of order 2 is a matrix, and tensors which order is greater than 2 are referred as higher order tensors. *Fibers* are defined by fixing every index but one, a matrix column is a mode-1 fiber and a matrix row is a mode-2 fiber. Third-order tensors have column, row, and tube fibers. Fibers are always assumed to be oriented as column vectors. *Slices* are two-dimensional sections of a tensor, defined by fixing all but two indices.

One of the most used norm in the tensor case is the Frobenious norm, also referred as tensor norm, and it is defined as the square root of the sum of the square of the element of a tensor. Formally:

$$||\mathscr{A}||_F = \sqrt{\sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} a^2_{c_1\dots c_p}}. \tag{1.1}$$

As in the vectors case, this norm can be used to evaluate distances between tensors defined on the same space.

An important sub-class of tensors is represented by the *cubic* tensors. Cubic tensors can be considered as a generalization of square matrices, and they are defined as tensors having all modes of the same dimensions. In our case $d_1 = \dots = d_p = d$.

Another special case is represented by *symmetric tensors*, sometime referred to as super-symmetric. These are the direct generalization of symmetric matrices in tensor algebra. Formally, given a vector of indices $\mathbf{h} = (h_1, \dots, h_p)^T$ and defining $\mathfrak{S}_{\mathbf{h}}$ to be the space of all permutation of $\mathbf{h}$, we have that $a_{\mathbf{h}} = a_{\sigma(\mathbf{h})}$ for all $\sigma \in \mathfrak{S}_{\mathbf{h}}$. This definition implies that just $H^{\bar{p}}/p!$ elements out of the H^p are distinct, where $H^{\bar{p}} = H(H+1)\cdots(H-p-1)$ is the rising factorial. It is easy to see that in 2-dimensional space the previous definition reduces to the usual symmetric matrix (i.e. equal to its transpose) and that $H^{\bar{2}}/2! = H(H+1)/2$.

An important notion is represented by the concept *rank-one* tensor, that is to say a tensor that can be written as outer product of vectors. Specifically, let $\boldsymbol{a}^{(j)} \in C^{d_j}$ be a collection of vectors for $j = 1, \dots, p$, $\mathscr{A} \in C$ is a rank-one tensor if it can be expressed as $\mathscr{A} = \boldsymbol{a}^{(1)} \otimes \dots \otimes \boldsymbol{a}^{(p)}$. This implies that each element of the tensor is the product of the corresponding vector elements $a_{c_1\dots c_p} = a^{(1)}_{c_1} \cdots a^{(p)}_{c_p}$.

Unlike matrices, however, there is not a uniquely defined notion of rank of a tensor, and consequently there are multiple definition of low-rank approximations.

1.2 Parafac rank and Parafac decomposition of a tensor

An appealing possibility is in expressing a tensor as a finite sum of low dimensional objects, such as vectors, matrices and small tensors. These representations allow compress storage of a tensor while simplifying several statistical tasks including, among others: regression, missing data imputation and dependence testing.

Such representations can be obtained, for example, by means of a low-rank approximations of the original tensor, given a certain definition of rank. A widely used definition of rank is the smallest number of rank-one tensors that can be used to rewrite the starting tensor. Formally:

$$\operatorname{rank}(\mathscr{A}) = \left\{ \min H : \mathscr{A} = \sum_{h=1}^{H} \lambda_h \bigotimes_{j=1}^{p} \boldsymbol{a}_h^{(j)} \right\}. \tag{1.2}$$

This rank is also known as Parafac rank because of the related tensor decomposition which expresses a general tensor as a finite sum of rank-one tensors

$$\mathscr{A} = \sum_{h=1}^{H} \mathscr{A}_h, \quad \mathscr{A}_h = \bigotimes_{j=1}^{p} \boldsymbol{a}_h^{(j)}.$$

This form of a tensor is mostly referred to as Parafac (parallel factor); alternative names used in the literature includes: Polyadic form of a tensor and CanDecomp or CP for canonical decomposition.

It is sometime useful to normalize to length one the rank-one tensors composing the previous sum, by introducing a weight vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_H)^T$. In this case Parafac can be expressed as a weighted sum of rank-one tensors

$$\mathscr{A} = \sum_{h=1}^{H} \lambda_h \bigotimes_{j=1}^{p} \boldsymbol{a}_h^{(j)}. \tag{1.3}$$

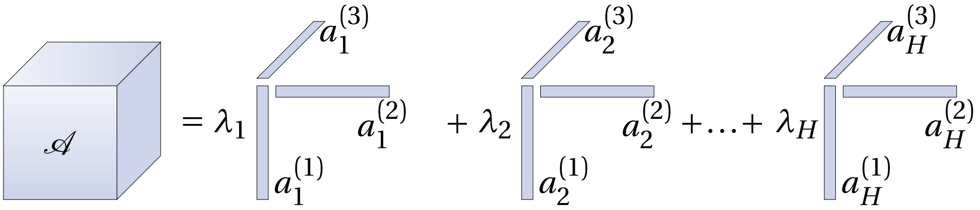


FIGURE 1.1: Parafac decomposition of a 3-dimensional cubic tensor.

Definition (1.2) implies that Parafac can be considered as an alternative way of writing a tensor $\mathscr{A}$, and open the possibility of approximating it with a smaller number of terms $\bar{H} < H$. The obtained decomposition is akin to the low-rank approximation of a matrix. In this latter case, Eckart and Young (1936) showed that the best H-rank approximation of a matrix can be obtained relying on the first H factors of the *Singular Value Decomposition (SVD)*. This formulation implies that low-rank decomposition of order $H-1$ is obtained by omitting the last component out the H-rank decomposition. Hence, all the decomposition of a matrix can be computed at once, relying on the SVD. Unfortunately, this property is not shared, in general, by higher order tensors, and the elements composing the $(\bar{H}-1)$-rank decomposition are not related to the ones composing the $\bar{H}$-rank one.

Let us assume the tensor $\mathscr{A}$ to be an observed tensor, the best H-rank approximation of this tensor can be computed solving the following minimization problem

$$\min_{\hat{\mathscr{A}}} ||\hat{\mathscr{A}} - \mathscr{A}||_F, \quad \text{such that} \quad \hat{\mathscr{A}} = \sum_{h=1}^{H} \lambda_h \boldsymbol{a}_h^{(1)} \otimes \cdots \otimes \boldsymbol{a}^{(p)}, \tag{1.4}$$

where $||\cdot||_F$ is the Frobenious norm defined in (1.1).

The most intuitive way of obtaining a numerical solution for minimization problem (1.4) is by using an alternate least square procedure. In fact, by fixing all elements of the decomposition but one, the problem is a simple least square minimization, having an analytic solution. A step by step pseudo-code for the algorithm can be found in Figure 3.3 of Kolda and Bader (2009). Many alternatives to least squares have been proposed over the years, for an overview of these methods we refer again to Kolda and Bader (2009) and related bibliography.

1.3 Tucker decomposition and compression

One of main aim of a tensor decomposition is in expressing a tensor in a compressed form. In this context, the idea of compression is related to storing an high-dimensional object by means of low dimensional ones, which occupy less physical memory. Although Parafac decomposition can be used in such a way, alternative tensor decompositions can produce a better compression.

Tucker decomposition (Tucker, 1963) is an alternative to Parafac consisting in a lower dimensional core tensor multiplied by a matrix along each mode, as exemplified in Figure 1.2.

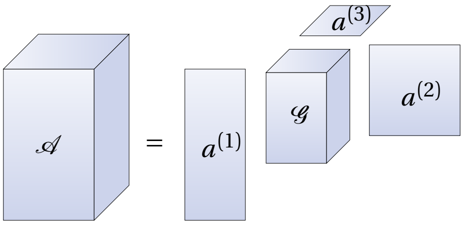


FIGURE 1.2: Tucker decomposition of a 3-dimensional cubic tensor.

Tucker decomposition can be expressed as

$$\mathscr{A} = \left\{ a_{c_1 \dots c_p} = \sum_{h_1=1}^{H_1} \cdots \sum_{h_p=1}^{H_p} g_{h_1 \dots h_p} \prod_{j=1}^{p} a^{(j)}_{h_j c_j}, c_j = 1, \dots d_j; j = 1, \dots, p \right\}. \tag{1.5}$$

Equation (1.5) is based on a different definition of rank compared to Parafac, namely the multi-rank of the tensor $\mathscr{A}$. The multi-rank of a tensor is the collection of the n-ranks respect to all its modes. The n-rank is defined as the dimension of the vector space spanned by the mode-n fiber. If one or more of the element of the multi-rank $(H_1, \dots, H_p)$ in equation (1.5) is smaller than the corresponding n-rank of $\mathscr{A}$ we have a low-rank approximation of the original tensor.

Parafac and Tucker decompositions are closely related, in fact we can see Parafac as a special case of Tucker where the core tensor $\mathscr{G}$ is super-diagonal, i.e. $g_{h_1 \dots h_p} = \lambda_h$ if $h_1 = \dots = h_p = h$ and 0 otherwise, and cubic, i.e. $H_1 = \dots = H_p = H$.

Wang and Ahuja (2005) and Kim and Choi (2007) empirically noted that Tucker achieves better data compression than Parafac decomposition. This is easier to understand by considering the way a tensor decomposed using Tucker can be reparametrize in Parafac form, expressed in Proposition 1.

Proposition 1. It is always possible to reparametrize a tensor expressed in Tucker form (1.5) in its equivalent Parafac form (1.3).

Proof. Letting $\bar{\boldsymbol{g}} = \text{vec}(\mathscr{G}) = (g_{1\dots1}, g_{1\dots2}, \dots, g_{H_1 \dots H_p})^T$ we can rewrite expression (1.5) as $\{\sum_{h=1}^{\bar{H}} \bar{g}_h \mathscr{A}_{h c_1 \dots c_p}; c_j = 1, \dots, d_j; j = 1, \dots, p\}$, where $\mathscr{A}_{h c_1 \dots c_p} = a_{h_1 c_1} \cdots a_{h_p c_p}$ for $h = 1, \dots, \bar{H} = H_1 \cdots H_p$. □

Form Proposition 1 we can notice how the Parafac rank can be considerably higher than the n-rank composing the Tucker multi-rank, that for this reason can produce a more compact approximation. As in the case of Parafac, for a sufficiently high value of the rank, Tucker is a way of re-expressing the original tensor.

Since the introduction different algorithms have been proposed to compute the elements of the Tucker decomposition, and arguably the most famous is the so called *Higher Order Singular Value Decomposition (HOSVD)*, which consists in iteratively computing the left singular value of the matrix form of the starting tensor, obtained by fixing all the modes but one. The details of the procedure are highlighted in Figure 4.3 of Kolda and Bader (2009).

1.4 Categorical variables and probabilistic tensors

Tensor data may occur in many different statistical applications including physics, chemistry and neuroscience. For example, RGB images can be seen as 3-dimensional tensors, and the previously described low-rank approximations can be naturally used to fulfill modelistic tasks such as classification and regression.

However, tensor factorizations are useful tools also in other fields of statistical research, where a notable example is represented by models for multivariate categorical variables. Let $\boldsymbol{X} = (X_1, \dots, X_p)^T$, be a vector of categorical random variables, where $X_j \in \{1, \dots, d_j\}$ for $j = 1, \dots, p$. We can notice that the sample space $\boldsymbol{\Omega} = \{1, \dots, d_1\} \times \cdots \times \{1, \dots, d_p\}$ is huge, and that it increases at an exponential rate with the number of variables p. Figure 1.3 gives an idea of the log-dimension of the space when $d_1 = \dots = d_p = d$.

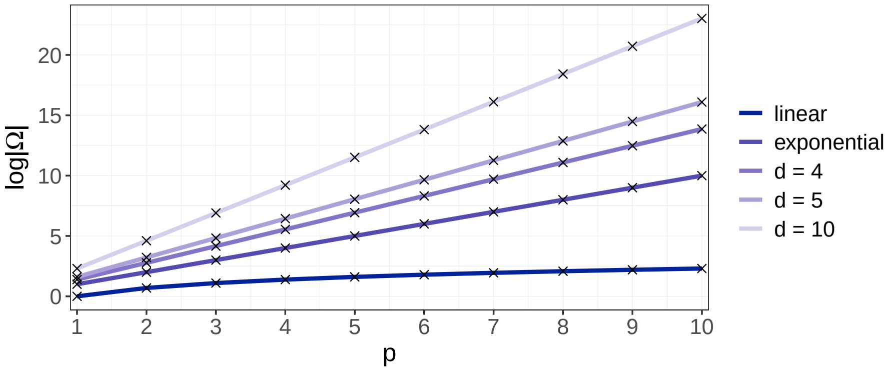


FIGURE 1.3: Log-cardinality of the sample space $|\Omega|$ for different values of p, in the case $d_1, \dots, d_p = d$. Linear and exponential rates have been added for comparison.

From a probabilistic perspective random variable $\boldsymbol{X}$ is described by means of a p.m.f. $\mathrm{p}(\omega) : \Omega \to [0,1]$ for all $\omega \in \Omega$, satisfying $\sum_{\omega \in \Omega} \mathrm{p}(\omega) = 1$. By construction the p.m.f. of a categorical random vector can be viewed as a *probability tensor* defined on the space $C = \boldsymbol{\Omega}$. A probability tensor is a tensor $\boldsymbol{\pi}$ which entries are non negative and add up to 1, or in other words $\boldsymbol{\pi}$ is defined on the $(|\boldsymbol{\Omega}| - 1)$-dimensional simplex. Clearly the p.m.f. of the random variable $\boldsymbol{X}$ can be parametrized by means of a function defined on a low dimensional space. This goal can be achieved, for example using log-linear models, which directly parametrize the logarithm of the p.m.f. as a linear function. The selected linear function directly includes the conditional dependence structure underlying the data. Tensor factorization methods however do not rely on highly restrictive assumptions, and directly reconstruct $\mathrm{p}(\omega)$ or some of its functional exploiting the low dimensional structures composing tensor decomposition itself.

It is worth noticing that in any model devoted to learn the p.m.f. of a vector of categorical variables, we necessarily have to make some assumptions since we are practically always trying to solve a dense problem by means of sparse observations, since we rarely are able to have a sample size of the order of $|\mathbf{\Omega}|$.

1.5 Low-rank approximations of probabilistic tensors

In this section we review some statistical approaches for large contingency tables which rely on the tensor factorization models presented above.

1.5.1 Latent class analysis and Parafac

In general dealing with non-ordered categorical data is a challenging problem, and the use of low-rank decompositions is an appealing alternative to log-linear models.

One of the first attempt to deal with this data was made by Lazarsfeld during the 1940s, and published for the first time in his contribution to Studies in Social Psychology in World War II (Lazarsfeld, 1950). The proposed model, mostly known as *Latent Class Analysis (LCA)*, was intended as an alternative to factor analysis to deal with categorical data. Factor analysis, in fact, postulates the existence of factors to explain the correlations among a set of continuous variables. LCA models, on the other hand, explicitly apply to the discrete variables that are created from responses to questionnaire items.

From a tensor algebra perspective, LCA is a low-rank Parafac approximation of the p.m.f. underlying the observed data. The used Parafac decomposition, is slightly different from the one described in (1.3), since it should produce a probability tensor.

To obtain a probability tensors from expression (1.3), different constraint might be considered. However, by letting the mixture weights $\boldsymbol{\lambda}$ and the tensor arms $\boldsymbol{\psi}_{\boldsymbol{h}}^{(j)}$ for $j = 1,\ldots,p$ and $h = 1,\ldots,H$, be defined on simplicies of appropriate dimension, we get a discrete mixture model representation. Such representation is the one originally proposed in defining LCA while, to the best of our knowledge, a formal connection with Parafac was made many years later in Dunson and Xing (2009).

The main assumption of this latent variable representation is that the observed data derive from a composite population having H sub-populations. Each of the observed subject i, for $i = 1,\ldots,n$, is hence endowed with an indicator variable $Z_i \in \{1,\ldots,H\}$, expressing to which sub-population the individual i belongs to. The random variable Z_i is indeed not observed and its values can be considered as missing or latent, and "predicted" using the available data.

An additional hypothesis is that, within each sub-population observed variables are independent. This hypothesis correspond to express each p.m.f. conditional on the latent indicator as a product of univariate p.m.f., hence rank-one probability tensors.

Formally, we can define $\mathrm{pr}(X_{i1} = x_1, \dots, X_{ip} = x_p \mid Z_i = h) = \psi^{(1)}_{hx_1} \cdots \psi^{(p)}_{hx_p}$. Letting $\mathrm{pr}(Z_i = h) = \lambda_h$, and marginalizing out the latent variable Z_i respect to its distribution, the global p.m.f. can be expressed as:

$$\mathrm{pr}(X_{i1} = x_p, \dots, X_{ip} = x_p \mid \boldsymbol{\Psi}, \boldsymbol{\lambda}) = \sum_{h=1}^{H} \lambda_h \prod_{j=1}^{p} \psi^{(j)}_{hx_j}. \tag{1.6}$$

Equation (1.6) is basically identical to Equation (1.3) with the only difference being in the space in which the tensor is defined and in the imposed constraints on tensor arms and weights.

The parameters $\boldsymbol{\lambda}$ in the aforementioned factorization are latent factors, and can be considered either as unknown parameters, or as random variable in a Bayesian framework. We follow this latter approach, hence all the p.m.f. involving the latent factors are defined conditional to them. This is not, however, a formal requirement and the symbol "|" can be substituted with ";".

Equation (1.6) is the contribution to the likelihood for subject i, and to estimate its parameters, at first method of moments was proposed (e.g. Lazarsfeld and Henry, 1968). A major drawback with this approach is that the subset of moments used in the procedure has to be chosen a priori. Usually $\mathbb{E}[X_j]$ and $\mathbb{E}[X_j X_k]$ for some $j, k \in \{1, \dots, p\}$ are used. This procedure was initially preferred to the maximum likelihood because of the complexity of likelihood function evaluation and numerical maximization through Newton-Raphson algorithm. This maximization was indeed too demanding for the computers back in 70s. Goodman (1974) proposed an efficient iterative proportional fitting algorithm to maximize the likelihood deriving from (1.6) solving the computational issues linked to the standard Newton-Raphson procedure. Nowadays a more efficient alternative is provided by the *Expectation Maximization (EM)* (Dempster *et al.*, 1977) algorithm which directly uses the incomplete data structure provided by the latent variable representation in order to maximize the likelihood. A recent development on this thread, concerning LCA estimation in presence of covariates is given in Durante *et al.* (2019).

However, likelihood maximization approaches, without imposing any additional constraint, tends to be unstable, mostly because of the presence of many local modes. Additionally, the uniqueness of the solution is not guaranteed in general cases.

The rank of the decomposition H is usually fixed in the maximization process, and treated as a tuning parameter selected relying on information criteria such as AIC/BIC or by means of cross validation.

Dunson and Xing (2009) overcame the rank selection problem recasting LCA in a Bayesian setting, learning the rank directly from the data through a Dirichlet process prior on the mixture weights of the decomposition. The proposed prior can be induced through the following specification:

$$\begin{aligned} \boldsymbol{\pi} &= \sum_{h=1}^{\infty} \lambda_h \boldsymbol{\psi}_h, \qquad \boldsymbol{\psi}_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \\ \boldsymbol{\psi}_h^{(p)} &\sim P_{0j}, \text{ indipendently for } j = 1, \dots, p, \\ & \qquad h = 1, \dots, \infty, \\ \boldsymbol{\lambda} &\sim Q, \end{aligned} \tag{1.7}$$

where P_{0j} are probability measure on the $(d_j - 1)$-dimensional simplex and Q is a probability measure on the countably infinite probability simplex. In the original proposal Dunson and Xing (2009) suggest the use of finite Dirichlet for P_{0j} and a Dirichlet process for Q.

They show that the prior induced by formulation (1.7) gives rise to many desirable statistical proprieties. Specifically, leveraging the duality with Parafac tensor decomposition, Corollary 1 in Dunson and Xing (2009) states that LCA can express any possible multivariate categorical data distribution. More importantly, Theorem 2 proves full support in Frobenius norm of the proposed prior which also implies full support for the posterior, since we are dealing with finitely many parameters.

The infinite dimensional prior for the mixing weights used in specification (1.7) can be approximated by means of a finite Dirichlet distribution. In particular, by choosing a symmetric Dirichlet prior with parameter H^{-1}, we obtain a finite dimensional approximation of the Dirichlet process through prior truncation (Ishwaran and Zarepour, 2008). This choice has the effect of enhancing computational flexibility, moreover, as noted in Rousseau and Mengersen (2011) the choice of small hyperparameters in finite mixture models allow deletion of redundant components. In most applications the difference in posterior inference is almost always negligible, hence this last strategy might be preferred. Additionally, provided that H is a sensible upper bound theoretical proprieties of model formulation (1.7) remain valid.

The introduction of Bayesian LCA inspired several papers dealing with flexible modeling of multivariate categorical variables in different contexts. In this branch of literature we find, among many others: a model to deal with time variant categorical data (Kunihama and Dunson, 2013); an alternative prior specification for the Parafac decomposition inducing shrinkage towards independence (Zhou *et al.*, 2015); a Bayesian non parametric model for retrospective likelihood in case/control studies (Zhou *et al.*, 2016); a variable selection approach to directly exclude from the factorization the independent variables (Papathomas and Richardson, 2016); a method to test group differences in high dimensional categorical data (Russo *et al.*, 2018).

1.5.2 Subject-specific inference and soft clustering in Tucker decomposition

Tucker decomposition has been extensively used in statistics as implicit parametrization for the p.m.f. underlying multivariate categorical variables, however, as in the case of LCA and Parafac such connection was made just in recent years. Again, as the case of Parafac, the model has been introduced leveraging a discrete mixture representation. Specifically, a generative mechanism producing a p.m.f. having a Tucker decomposition form may be obtained as product of conditionally independent mixture models, having subject-specific mixture weights. The resulting model has a nice interpretation both in term of characteristic of the underlying population, expressed by tensor arms, than, from a subject-specific perspective. Indeed, the subject-specific mixture weights can be interpreted as a proximity measure of each subject respect to the estimated population characteristics.

This conditional model representation has been used, with some differences in the underlying hypothesis, at least in four different statistical problems and their generalization, namely: (i) *Grade of Membership models (GoM)* (e.g. Woodbury *et al.*, 1978); (ii) Admixture models (Pritchard *et al.*, 2000); (iii) *Latent Dirichlet Allocation (LDA)* (e.g. Blei *et al.*, 2003); (iv) *Mixed Membership models (MM)* (e.g. Erosheva and Fienberg, 2005).

All previous models are deeply connected even if proposed independently for different purposes. In particular, (i) was the first to be used, while (iv) is a more general class including (ii) and (iii), which have been introduced to fulfill substantially different goals. GoM are based on likelihood framework while (iii)–(iv) rely on a Bayesian perspective. Despite their differences in generative mechanism, application and estimation techniques, all the proposed models share the common aim of allowing subject-specific inference in a multivariate categorical variables context. In all the aforementioned models, subject-specific mixture weights induce a soft, or fuzzy, clustering giving rise to a more compact representation of the p.m.f. if compared to standard mixture models.

Given a collection of categorical random variables $(X_{i1},\dots,X_{ip})^T$ for $i=1,\dots,n$ and $j=1,\dots,p$ such that $X_{ij}\in\{1,\dots,d_j\}$, a mixed membership model can be defined as follows:

$$\begin{aligned} X_{ij}\mid Z_{ij}=h,\boldsymbol{\psi}_h^{(j)} &\sim \text{Cat}(\psi_{h1}^{(j)},\dots,\psi_{hd_j}^{(j)}),\\ Z_{ij}\mid\boldsymbol{\lambda}_i &\sim \text{Cat}(\lambda_{i1},\dots,\lambda_{iH}),\\ \boldsymbol{\lambda}_i &\sim P, \end{aligned} \tag{1.8}$$

where P is the distribution of the membership score vectors associated with each observation i. The distribution P is defined on the $(H-1)$-dimensional simplex and popular choices for its specification include Dirichlet (Blei *et al.*, 2003) and logistic normal (Lafferty and Blei, 2006). From model (1.8) we can immediately notice that there is a population level assumption, i.e. the population is composed of H sub-populations, and an individual level assumption, for which each subject has a degree of similarity with the type h expressed by λ_{ih}. The vector $\boldsymbol{\lambda}_i$ informs how subject i varies from other individuals in the population.

Leveraging the local independence assumption in formulation (1.8) the probability distribution for the generic subject i can be expressed, integrating out the latent variable $\boldsymbol{Z}_i=(Z_{i1},\dots,Z_{1p})^T$, as

$$\begin{aligned} \text{pr}(X_{i1}=x_1,\dots,X_{ip}=x_p\mid\boldsymbol{\lambda}_i,\boldsymbol{\psi}) &= \prod_{j=1}^{p}\sum_{h=1}^{H}\lambda_{ih}\psi_{hx_j}^{(j)}\\ &= \left(\sum_{h_1=1}^{H}\lambda_{ih_1}\psi_{h_1x_1}^{(1)}\right)\cdots\left(\sum_{h_p=1}^{H}\lambda_{ih_p}\psi_{h_px_p}^{(p)}\right)\\ &= \sum_{h_1=1}^{H}\cdots\sum_{h_p=1}^{H}\prod_{j=1}^{p}\lambda_{ih_j}\psi_{h_jx_j}^{(j)}, \end{aligned} \tag{1.9}$$

which is a product of conditional independent mixture models. The population model can be retrieved integrating out the random effect $\boldsymbol{\lambda_i}$ with respect to its distribution P

$$\text{pr}(X_1=x_1,\dots,X_p=x_p\mid\boldsymbol{\psi})=\sum_{h_1=1}^{H}\cdots\sum_{h_p=1}^{H}a_{h_1\dots h_p}\prod_{j=1}^{p}\psi_{h_jx_j}^{(j)}, \tag{1.10}$$

where $a_{h_1\dots h_p}=\mathbb{E}_P[\lambda_{ih_1}\cdots\lambda_{ih_p}]$ is the expectation of the product of the score vector elements over P. Depending on the choice of P, the expectation $a_{h_1\dots h_p}$ may or may not have a closed form expression.

Equation (1.10) is an instance of Tucker tensor decomposition defined in Equation (1.5), and it is possible to show that it is a flexible representation for the probability mass function of unordered categorical random variables, since there always exists an H such that any probability mass function can be characterized as in (1.10). This is a direct consequence of the fact that any tensor can be expressed in Tucker form, for a sufficiently high rank. Additionally, representation (1.10) is more compact than a standard discrete mixture model representation for the p.m.f. (see for example Bhattacharya and Dunson, 2012).

Leveraging Proposition 1, equation (1.10) can be interpreted as a constrained discrete mixture model with H^p latent components. In fact, the core tensor $\mathscr{A} = \{a_{h_1,\dots,h_p}; h_j = 1,\dots,H; j = 1,\dots p\}$ is specified to be a cubic symmetric tensor, defined in Section 1.1.

From modelistic perspective, such constraints derive from the exchangeability assumption for the profiles probability in model (1.8) (e.g. Erosheva *et al.*, 2007). The resulting constraint Tucker is a less compact representation than an unconstrained decomposition, meaning that the dimension of the core tensor needed to reach a certain approximation of the p.m.f. is expected to be larger than in (1.5). This might not represent an issue when the number of pure types is chosen adaptively but, when a small H is fixed a priori, this representation may lead to an unsatisfactory approximation of the probability mass function.

As briefly mentioned above, members of the class of MM models can be estimated in many different frameworks, leading to alternative algorithms and techniques for model inference. It is worth noticing that model (1.10) is a generative mechanism, not a Bayesian specification, hence after choosing a specific distribution for P we can obtain maximum likelihood estimate of the parameters through numeric optimization. This approach was followed in many application, especially in the case of GoM and its extension (e.g. Berkman *et al.*, 1989; Castro *et al.*, 2006). In this context, Manton *et al.* (1987) propose to optimize the likelihood conditioning on the score distribution P. This proposal also justifies, from a generative perspective, the first algorithm used in Woodbury *et al.* (1978), who considered the scores $\boldsymbol{\lambda}_i$ directly as model parameters, deriving their values jointly with the profiles parameters using a Newton-Raphson style approach.

Both admixture and mixed membership models were directly introduced following a Bayesian specification. Model formulation are slightly different to accomodate the different application, however, in both Pritchard *et al.* (2000) and Erosheva *et al.* (2007), a Dirichlet distribution is used for P. The hyperparameters of the distribution P roughly express the proportion of subject in each profiles, which is usually unknown. Such hyperparameters can be learned directly from the data by means of a suitable hyperprior (e.g. Dirichlet), which parameters are updated through a Metropolis step. In some cases, unfortunately, this method might be slow to converge, mostly because of the correlation of the Dirichlet parameters.

In the context of topic discovery, because of the high dimension of the document data involved, the preferred methods is usually *Variational Bayes (VB)* approximation of the posterior. Details on how to estimate model parameters using VB can be found in Blei *et al.* (2003) for the Dirichlet case, and in Lafferty and Blei (2006) for the case of logit normal.

As in the case of LCA a critical issue is in the choice of the dimension of the latent space H. This is usually done by comparing alternative models by means of information criteria or cross validation, after choosing an appropriate loss function (e.g. Airoldi *et al.*, 2010; White *et al.*, 2012). Alternatively, it is possible to include inference on the value of H directly in the model by adapting priors on the dimension of the latent space. These approaches require sampling in spaces with changing dimensionality across MCMC iterations. In this class we find the reversible jump approaches as the one described in Lopes and West (2004), and nonparametric latent factor models priors such as Knowles and Ghahramani (2011) and Bhattacharya and Dunson (2011).

Tucker decomposition is however not just confined to Mixed Membership models and finds its place in many other statistical applications. Some recent examples include: Schein *et al.* (2016),Yang and Dunson (2016), Johndrow *et al.* (2017) and Xiong *et al.* (2018).

Chapter 2

Grouped Tensor Factorization

In this Chapter we consider a new class of *GROuped Tensor (GROT)* factorizations, which have superior compression performance if compared to standard Parafac approach, using relatively few components to represent the joint probability mass function of the data. While popular Parafac factorizations rely on mixing together independent components, GROT mixes together grouped factorizations, equivalent to replacing vector arms in Parafac with low-dimensional tensor arms. We consider a Bayesian approach to inference with Dirichlet priors on the mixing weights and arm components, to obtain a combined low-rank and sparse structure, while facilitating efficient posterior computation via Markov chain Monte Carlo. In Section 2.1 we introduce the model, highlighting some key properties, and providing an algorithm for posterior computation. In Section 2.2 we propose a prior on the space of partitions to effectively learn GROT arm structure from the data. In Section 2.3 we analyze simulated data presenting different dependence structures, highlighting the improved performances in low sample size scenarios. Finally, in Section 2.4 we apply our proposed approach to classify and identify starting bases of Escherichia Coli promoter nucleotide sequences.

2.1 Model specification and properties

Parafac factorization introduced in 1.7 is a popular model which theoretical proprieties have been studied in the literature, and partially summarized in Chapter 1. However, in some practical applications the p.m.f. is poorly characterized by this model. This usually occurs when the dependence structure is dense (e.g. presence of cliques) and the available sample size is low. Intuitively, Parafac model deals with dependence only through marginalization of a categorical latent variable. Such a “naive” assumption strongly simplifies estimation and inference, and works well in sparse contexts, but it might be over-simplistic in other cases.

A possible solution to overcome this issue is to induce a more refined characterization of the dependence in tensor arms, in such a way that part of the dependence is directly modeled, instead of just induced through marginalization. To include such structure in the model we generalize Parafac allowing tensor arms to have (variable) dimension greater than 1, as exemplified in Figure 2.1. Note that the p.m.f. is not necessarely decomposed by rank one tensor as in Parafac case.

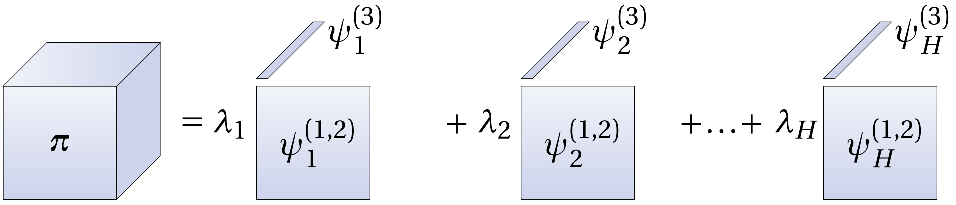


FIGURE 2.1: GROT tensor decomposition of a 3-dimensional tensor, having structure $\{\{1,2\},\{3\}\}$.

Let $B_1,\ldots,B_G$ be a partition of the set of indices $\{1,\ldots,p\}$, i.e. $B_j \cap B_l = \emptyset$ and $B_1 \cup \cdots \cup B_G = \{1,\ldots,p\}$, and $\{X_{ij} : j \in B_l\} = (X_{i1}^l,\ldots,X_{ip_l}^l)^T$ where $X_{ij}^l \in \{1,\ldots,d_j^l\}$ for $j = 1,\ldots,p_l = |B_l|$ and $i = 1,\ldots,n$. By convention we arrange the variable in each partition following their indices in increasing order. We propose the following specification for the joint probability density function

$$\pi_{c_1,\ldots,c_p} = \sum_{h=1}^{H} \lambda_h \prod_{l=1}^{G} \Psi^{(l)}_{h(c_1^l,\ldots,c_{p_l}^l)}, \tag{2.1}$$

where each arm $\mathbf{\Psi}_h^{(l)} = \{\Psi^{(l)}_{h(c_1,\ldots,c_{p_l})}, c_1 = 1,\ldots,d_1^l,\ldots,c_{p_l} = 1,\ldots,d_{p_l}^l\}$ is a probability tensor of dimension $\otimes_{\{j:j\in B_l\}} d_j$.

We can notice that each cell of the probability tensor π is modeled as mixture of multinomial distributions as in the Parafac case, that is obtained when $|B_l| = 1$ for $l = 1,\ldots,G$. Model specified in (2.1) shares flexibility in representation of Parafac and Tucker decompositions, as assured by Proposition 2.1.

Proposition 2.1. *Given a partition $B_1 \cup \cdots \cup B_G = \{1,\ldots,p\}$, having $B_j \cap B_l = \emptyset$, any probability tensor π, defined on the $(\prod_{j=1}^{p} d_j - 1)$-dimensional simplex can be represented as* (2.1).

Proof. The proposition can be proved by showing that equation (2.1) can be rewritten in Parafac form. Let $\varphi_h^{(l)} = \text{vec}(\mathbf{\Psi}_h^{(l)})$ be the vectorization of the l-th tensor-arm having element stacked from the last mode, then the p.m.f. derived from the GROT factorization can be expressed as $\pi = \sum_{h=1}^{H} \lambda_h \bigotimes_{l=1}^{G} \varphi_h^{(l)}$. Last expression is an instance of Parafac decomposition, hence by corollary one in Dunson and Xing (2009) there exist an H such that we can characterize any possible dependence among the G groups of variables. □

Proposition 2.1 is not necessarily restricted to probability tensors, and is indeed applicable in general to any form of tensor data and, more importantly, it grants that the choice of the partition, to be used in tensor arms, does not affect flexibility of representation (2.1).

Additionally, as stated in theorem (2.2), model (2.1) has the key benefit of giving a more parsimonious representation in term of latent classes respect to Parafac decomposition, while retaining its simplicity in estimation and posterior inference.

Theorem 2.2. *The number of latent classes H needed to fully characterize a probability tensor using GROT* (2.1) *is lower or equal than the one needed using Parafac* (1.3)*. Equality is reached iff all the variables in the same partition are independent.*

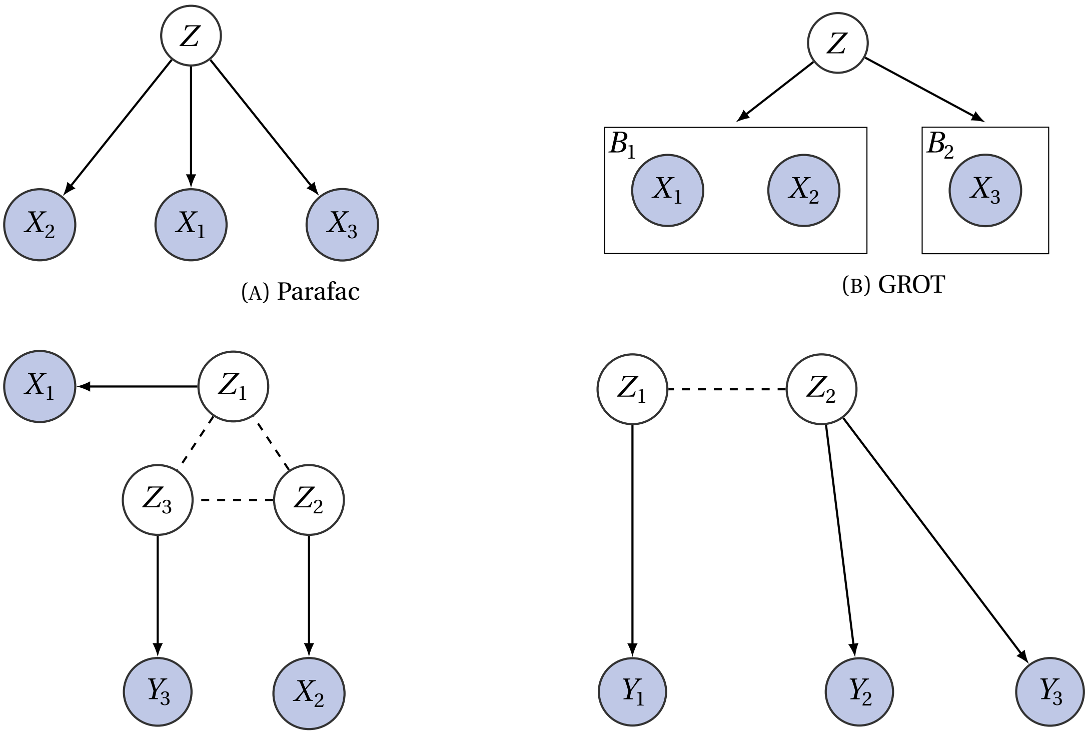
Proof. To prove the first part of the theorem we rewrite the probability tensor deriving from expression (2.1) in a Parafac form. By defining the tensors $\boldsymbol{\Phi}_h = \{\prod_{l=1}^{G} \Psi^{(l)}_{h c_1^l \dots c_p^l} : c_j^l = 1, \dots, d_j, j = 1, \dots d_{p_l}, l = 1, \dots, G\}$ for $h = 1, \dots, H$, we can write $\pi = \sum_{h=1}^{H} \lambda_h \boldsymbol{\Phi}_h$. Note that the tensors $\boldsymbol{\Phi}_h$ are defined on the $(\prod_{j=1}^{p} d_j - 1)$-dimensional simplex, and can have Parafac rank greater that one. Hence, there exist a sequence of ranks $K_1, ..., K_H$ with $K_h \geq 1$ such that the following equalities hold

$$\pi = \sum_{h=1}^{H} \lambda_h \boldsymbol{\Phi}_h = \sum_{h=1}^{H} \lambda_h \left[\sum_{k=1}^{K_h} \nu_k^{(h)} \bigotimes_{j=1}^{p} \boldsymbol{\vartheta}_{hk}^{(j)} \right] = \sum_{h=1}^{\bar{H}} \bar{\lambda}_h \bigotimes_{j=1}^{p} \bar{\boldsymbol{\vartheta}}_h^{(j)},$$

where $\bar{\lambda}_1 = \lambda_1 \nu_1^{(1)}, \bar{\lambda}_2 = \lambda_1 \nu_2^{(1)}, \dots, \bar{\lambda}_{\bar{H}} = \lambda_H \nu_{K_H}^{(H)}$, and $\bar{\boldsymbol{\vartheta}}_1^{(j)} = \boldsymbol{\vartheta}_{11}^{(j)}, \dots, \bar{\boldsymbol{\vartheta}}_{\bar{H}}^{(j)} = \boldsymbol{\vartheta}_{H K_H}^{(j)}$, for $j = 1, \dots, p$. Last expression is the Parafac reparametrization of (2.1), and since $\bar{H} \geq \sum_{h=1}^{H} K_h$, the first part of the theorem is proved. For the "only if" part it suffices to notice that a necessary condition for the equality to holds is that $K_h = 1$ for $h = 1, \dots, H$, which is true just if the variable in the same block of the partition are independent. □

The fact that GROT model is more parsimonious than Parafac from a latent class perspective does not imply that it is uniformly better in any application, since the performances would strongly depend on the "true" unknown dependence structure. In fact, by construction GROT model exploits a group structure to improve over standard Parafac factorization. When several groups of variables are present in the data, even with moderate across groups dependence, GROT is expected to perform better than standard Parafac, since the combined effect of multivariate kernels and latent variables is more effective in characterizing the dependence than simple marginalization. The presence of a block dependence is very common in the context of categorical variables, however, there are exceptions. A notable one is when, although some groups of dependent variables are present, the groups are partially overlapped, for example following a small order autoregressive model. These dependences might be effectively learned by a Parafac model with a small rank, and the additional number of parameters introduced in the multivariate kernels of GROT might not compensate for the reduction in the number of latent classes. In non-degenerate cases, however, this problem does not represent a real issue, and the loss of performances of GROT should be negligible.

Model (2.1) aims to provide a compact representation of π. This aim is similar in spirit to the one motivating the use of Tucker (e.g. Bhattacharya and Dunson, 2012) and C-Tucker (Johndrow *et al.*, 2017), but has the advantage of bearing a simpler functional form for π and its functionals provided by Parafac. A latent variable representation of the aforementioned decompositions is presented in Figure 2.2.



(A) Parafac

(B) GROT

(C) Tucker

(D) C-Tucker

FIGURE 2.2: Graphical representation of different tensor decompositions.

Although C-Tucker model shares a similar motivation to the proposed GROT decomposition, the two approaches produce advantages respect to standard Parafac in different settings. In fact, C-Tucker model complexity is dominated by Parafac when many small dependent groups are present. By contrast, the proposed GROT still needs a moderate to high number of latent classes when the underlying truth can be represented by means of marginally independent cliques of variables, having high cardinality. Both models, however, theoretically improve over standard Parafac whenever some non-sparse structured dependence is present in the data.

2.1.1 Posterior Computation

Once that the partition $\varrho = \{B_1, \dots, B_G\}$ has been fixed, model (2.1) can be seen as an instance of Parafac decomposition, i.e. a mixture of multinomial distributions. Hence in specifying conditionally conjugate Dirichlet priors for the tensor arms, and a finite Dirichlet prior for the mixture weights, we have an easy to implement Gibbs sampler. The model can be expressed in the following hierarchical form:

$$\begin{aligned}
\{X_{ij} : j \in B_l\} \mid Z_i = h &\sim \text{Multinomial}\Big(\{(1,\dots,1),\dots,(d^l_1,\dots,d^l_{p_l})\},(\psi^{(l)}_{h,(1\dots1)},\dots,\psi^{(l)}_{h,(d^l_1\dots d^l_{p_l})})\Big),\\
\boldsymbol{\Psi}^{(l)}_h &\sim \text{Dirichlet}\Big(\alpha^l_{11\dots1},\dots,\alpha^l_{d^l_1,\dots,d^l_{p_l}}\Big) \quad \text{for} \quad l=1,\dots,G; \quad h=1,\dots,H,\\
Z_i &\sim \text{Multinomial}((1,\dots,H),\lambda_1,\dots,\lambda_h),\\
(\lambda_1,\dots,\lambda_H)^T &\sim \text{Dirichlet}(1/H,\dots,1/H). \qquad (2.2)
\end{aligned}$$

Model specification (2.2) leads to the Gibbs sampler sketched in Algorithm 1.

```
Algorithm 1: Posterior GROT factorization with fix partition (2.2)
for r in 1:n_iterations do
    [1] update tensor arms for $h$ in $1:H$ & $l$ in $1:G$ do
        $\boldsymbol{\Psi}^{(l)}_h \mid - \sim \text{Dirichlet}\Big(\alpha^l_{1,\dots,1} + n^l_{h,(1,\dots,1)},\dots,\alpha^l_{d^l_1,\dots,d^l_{p_l}} + n^l_{h,(d^l_1,\dots,d^l_{p_l})}\Big)$,
        where $n^l_{h,(c_1,\dots,c_{p_l})} = \sum_{i:z_i=h} 1\{x^l_{i1} = c^l_1,\dots,x^l_{ip_l} = c^l_{p_l}\}$ is the number or subject in
        class $h$ presenting a certain configuration;
    [2] update $z_i$ for $i$ in $1:n$ do
        sample from the multinomial full conditional with
        $\text{pr}(z_i = h \mid -) = \frac{\lambda_h \prod_{l=1}^G \Psi^{(l)}_{h(x^l_{i1},\dots,x^l_{ip_l})}}{\sum_{k=1}^H \lambda_k \prod_{l=1}^G \Psi^{(l)}_{k(x^l_{i1},\dots,x^l_{ip_l})}}$, for $h = 1,\dots H$;
    [3] Update the mixture weights from their full conditional
        $\boldsymbol{\lambda} \mid - \sim \text{Dirichlet}(n_1 + 1/H,\dots,n_H + 1/H)$,
    where $n_h = \sum_{i=1}^H 1\{z_i = h\}$ is the number of subjects allocated in latent class $h$.
```

2.2 Group detection using product of partition models

In many real data applications we do not have concrete knowledge to chose a partition ϱ in model (2.1), hence it would be appealing to learn this structure directly from the data. Algorithm 1 also requires to simulate from Dirichlet distributions of size corresponding to the blocks partition. For this reason, in enhancing computational tractability and numerical stability, we define an upper bound m for the cardinality of the partition blocks. To the best of our knowledge – the problem of inferring groups structure with an upper bound on the cardinality of the groups has not been considered in the literature. This is not surprising since the presence of such bound is not typical in clustering applications. We propose here a Bayesian solution to this problem leveraging a model on the space of partitions.

Product Partition Models (PPM) (Hartigan, 1990; Barry and Hartigan, 1993; Crowley, 1997) have been deeply studied in literature, with a great focus on Dirichlet product mixture models, and applied in several tasks including density estimation/regression and clustering (e.g. Park and Dunson, 2010; Müller *et al.*, 2011). In clustering, the main goal is usually in contemporary detecting number and composition of the groups of subjects. With similar spirit we make use of a PPM to detect partitions in the variable space, taking into account uncertainty in their composition.

The basic idea of PPM is that the joint probability mass (or density) function can be expressed as a product. Dealing with categorical random variable, once we define a partition $\varrho = \{B_1, \dots, B_G\}$ the density of a PPM can be expressed as $\text{pr}(X_1 = x_1, \dots, X_p = x_p) = \prod_{l=1}^{G} \text{pr}(\{X_j = x_j : j \in B_l\})$.

To take into account uncertainty about the partition ϱ, a prior can be induced by means of a cohesion function, expressing the prior probability for each possible block composing a partition.

PPM prior leads to a conjugate model, meaning that the posterior is still in the class of PPM, having expression $\text{pr}(\varrho \mid -) = K \prod_{l=1}^{G} c(B_l)\text{pr}(\{X_j = x_j : j \in B_l\})$, where $c(\cdot)$ is the cohesion function. Theoretically, the normalization constant K is available in a simple analytic form, however, it implies a sum over the space of partition whose dimension is huge even for a small number of elements. Just think that if we have 12 objects the cardinality of the partition space exceeds 4 millions. However, inference for PPM can still be carried on by means of Monte Carlo methods.

PPM are very general and do not preclude any model specification for the probability mass function over each partition. Enhancing simplicity, we focus on the easiest model for categorical data, i.e. using a multinomial distribution on each possible partition.

Considering n independent replicates of $\boldsymbol{X}$, we indicate with $(X^l_{i1},\dots,X^l_{ip_l})^T = \{X_{ij} : j \in B_l\}$, where $X^l_{ij} \in \{1,\dots,d^l_j\}$ for $j = 1,\dots,p_l$ and $i = 1,\dots,n$. By convention we order the variables in each block B_l, following their indices in increasing order, however, the order is not important for estimation purposes.

For each block we have a multinomial likelihood with parameter $\boldsymbol{\Psi}^{(l)} = (\psi^{(l)}_{1\dots1},\dots,\psi^{(l)}_{d^l_1\dots d^l_{p_l}})^T$, having generic element $\psi^{(l)}_{(c_1,\dots,c_{p_l})} = \mathrm{pr}(X^l_1 = c_1,\dots,X^l_{p_l} = c_{p_l})$. For the parameters $\boldsymbol{\Psi}^{(l)}$ we consider conjugate Dirichlet priors having parameter $\boldsymbol{\alpha}^l = (\alpha_{1,\dots,1},\dots,\alpha_{d^l_1,\dots,d^l_{p_l}})^T$. This choice for the prior allows for analytic integration of the parameter $\boldsymbol{\Psi}^{(\boldsymbol{l})}$ leveraging the Dirichlet-multinomial conjugacy. Indicating with $\alpha_0 = \sum_{c_1=1}^{d^l_1}\cdots\sum_{c_{p_l}=1}^{d^l_{p_l}} \alpha_{c_1,\dots,c_{p_l}}$ and $n^l_{c^l_1,\dots,c^l_{p_l}} = \sum_{i=1}^{n} 1\{x^l_{i1} = c^l_1,\dots,x^l_{ip_l} = c^l_{p_l}\}$, the counts for the category $(c^l_1,\dots,c^l_{p_l})$ where $c^l_j \in \{1,\dots,d^l_j\}$, and $1\{\cdot\}$ is the indicator function, the corresponding posterior distribution for the model is given by

$$\mathrm{pr}(\varrho \mid -) = K \prod_{l=1}^{G} \left\{ c(B_l) \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \prod_{c_1=1}^{d^l_1} \cdots \prod_{c_{p_l}=1}^{d^l_{p_l}} \frac{\Gamma(\alpha_{c_1,\dots,c_{p_l}} + n^l_{c^l_1,\dots,c^l_{p_l}})}{\Gamma(\alpha_{c_1,\dots,c_{p_l}})} \right\}. \tag{2.3}$$

Expression (2.3) can be further simplified considering a symmetric Dirichlet distribution as prior, which is a sensible choice in absence of any concrete prior knowledge. As already mention computation of the normalizing constant K is prohibitive even in moderate dimension, however we can obtain approximate samples from (2.3) adapting algorithms for Bayesian non parametric priors (e.g. Neal, 2000). Steps to obtain approximate samples from the posterior are highlighted in Algorithm 2.

Algorithm 2: Posterior sampling for multinomial PPM (2.3)

```
To initialize the algorithm we need a starting partition $\varrho^{(0)} = \{B^{(0)}_1,\dots,B^{(0)}_{G^{(0)}}\}$ having
  positive probability according to (2.3). A possible default choice is in starting with
  the singleton partition (e.g. $|B_l| = 1$ for $l = 1,\dots,G$, and $G = p$).
for $r$ in 1:n_iterations do
    Set $R = \{1,\dots,p\}$ and $\varrho^{(r)} = \varrho^{(r-1)}$.
    while $|R| > 0$ do
        [1] Sample $j \in R$ with uniform probability $1/|R|$;
        [2] compute $\varrho^{(-j)}$ as the partition obtained by removing variable $j$ from $\varrho^{(r)}$;
        [3] if $|B^{(-j)}_l| > 0$ for $l = 1,\dots,G^{(-j)}$ then $G^{(-j)} = G^{(-j)} + 1$, with $B_{G^{(-j)}} = \{\varnothing\}$;
        [4] for $l$ in $1{:}|\varrho^{-j}|$ do
            set $\varrho^* = \varrho^{-j}$, $B^*_l = \{B^{-j}_l \cup j\}$, and store in $p^*_l$ the value of (2.3) for $\varrho^*$,
              omitting the normalizing constant;
        [5] assign $j$ to $B^*_l$ with probability proportional to $p^*_l$;
        [6] set $\varrho^{(r)} = \varrho^*/\{\varnothing\}$, $G^{(r)} = |\varrho^{(r)}|$ and $R = R/\{j\}$.
```

At each iteration, Algorithm 2 involves a random number of simple operations, namely $p \times G$, for $G \in \{1,\ldots,p\}$, to update the state of the partition. The computational cost might become prohibitive in ultra-high dimensional contexts. In such cases, instead of scanning all the possible groups, a single metropolis proposal can be used, adapting, for example, the split-and-merge algorithm (Green and Richardson, 2001; Jain and Neal, 2004)or the Chaperon algorithm (Miller *et al.*, 2015).

2.2.1 Choice of the cohesion function

The choice of a cohesion function has a deep impact on posterior inference, since it determines the clustering behavior. Intuitively, this prior sensitivity can be attributable to the huge dimension of the partition space, where the likelihood can support many possible configurations with equal probability. For some choices of the cohesion function, we obtain popular non parametric priors, such as Dirichlet process (Ferguson, 1973), Pitman-Yor (Pitman and Yor, 1997), and the more general class of Gibbs-type priors (Gnedin and Pitman, 2006; Blasi *et al.*, 2015).

Arguably, the most used prior belonging to this class is the Dirichlet process. Dirichlet process prior is characterized by the so called *richer-get-richer* property, leading to the detection of few big clusters which size decreases exponentially. This characteristic might be convenient in many applications, particularly when the main goal is in allowing local borrowing strength, in our context, however, we aim to find partitions with many small groups, supported by the data.

In accomplishing such a goal, we define a cohesion function giving the same probability to all the partitions having cardinality lower than a predefined upper bound m.

The proposed cohesion function can be expressed as:

$$c(B_l) \propto \begin{cases} 1 & \text{if} \quad |B_l| < m, \\ 0 & \text{if} \quad |B_l| \geq m. \end{cases} \quad , \quad \text{for } l = 1,\ldots,G. \tag{2.4}$$

Prior specification (2.4) has the main advantage of not enforcing a richer-get-richer behavior, while directly excluding partition with cardinality bigger that m. Equation (2.4) is a generalization of the uniform process, that can be obtained letting $m \to \infty$. Wallach *et al.* (2010) pointed out that uniform process is a suitable prior when the number of cluster is expected to be high with respect to the number of available objects. The proposed cohesion function shares the lack of exchangeability of the uniform process from which it derives, this however raises no concerns in our application where, the aim is in detecting constraint partitions.

In detecting dependence structure, ideally, we would like m to be larger than the maximum cardinality of the larger "true" block, however, this is often not feasible in practice. High dimensional dependence, in fact, are in general not easily detected, mainly when we have a very low sample size. In such contexts smaller sub-partitions are more easily identified, and might represent a good proxy of the real dependence structure.

Additionally, by putting a threshold parameter m we exclude usually unrealistic models such as all variable clustered together, while enhancing computational flexibility, since the choice of a small m deeply simplify partition updates, by reducing the cardinality of the correspondent space. Specifically by selecting a small m we reduce the space of admissible partitions in step [**4**] of Algorithm 2.

Although very simple the proposed partition model can give important insights on the dependence structures, successfully identifying subset of dependent variables that can be used for finer level inference at a second stage.

2.2.2 Learning arm structure in GROT factorization

The introduced PPM can be used *per se* as a learning techniques, for detecting structures and screening independent variables, however, it may lack the necessary flexibility in complex scenarios. We can leverage a PPM prior with cohesion function (2.4) to take into account uncertainty in the arm structure of model (2.2). This prior choice just introduce a small modification of Algorithm 1 to sample from the posterior. By using the same prior setting of model (2.2) we can, in fact, sample from the posterior relying on Algorithm 3

Algorithm 3: Posterior sampling for GROT factorization with unknown arm structure

```
To initialize the algorithm we need a starting partition $\varrho^{(0)} = \{B_1^{(0)}, \dots, B_{G^{(0)}}^{(0)}\}$. A
  possible default choice is in starting with the singleton partition (e.g. $|B_l| = 1$ for
  $l = 1, \dots, G$, and $G = p$).
for r in 1:n_iterations do
    [1] —[3] conditionally on the current partition $\varrho^{(r)}$ update $\boldsymbol{\Psi}$, $\boldsymbol{\lambda}$, and $z_i$ for
    $i = 1, \dots, n$, following Algorithm 1;
    [4] update the current partition $\varrho^{(r)}$, following the step in Algorithm 2, using the
    following conditional posterior in step [4]

    $$p_l^* = \prod_{h=1}^{H} \prod_{\{v:|B_v|<m\}} \left\{ \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_h)} \frac{\prod_{c_1=1}^{d_1^v} \cdots \prod_{c_{p_v}=1}^{d_{p_v}^v} \Gamma\left(\alpha_{c_1,\dots,c_{p_v}+n^v_{h(c_1^v,\dots,c_{p_v}^v)}}\right)}{\prod_{c_1=1}^{d_1^v} \cdots \prod_{c_{p_v}=1}^{d_{p_v}^v} \Gamma(\alpha_{c_1,\dots,c_{p_l}})} \right\},$$

    where $\alpha_0 = \sum_{c_1=1}^{d_1^v} \cdots \sum_{c_{p_v}=1}^{d_{p_v}^v} \alpha_{c_1,\dots,c_{p_v}}$ and
    $n^v_{h(c_1^v,\dots,c_{p_v}^v)} = \sum_{i:z_i=h} 1\{x_{i1}^v = c_1^v, \dots, x_{ip_v}^v = c_{p_v}^v\}$.
```

2.3 Simulation study

With the aim of testing the empirical performances of the proposed multinomial PPM and GROT factorization models, we consider two simulation studies, with an emphasis on learning dependence structures in the data. We focus on pairwise dependences, which are the easiest to characterize/interpret and, usually, the main object of inference in categorical variable applications. We highlight differences in performances with Parafac decomposition that represents the main competitor.

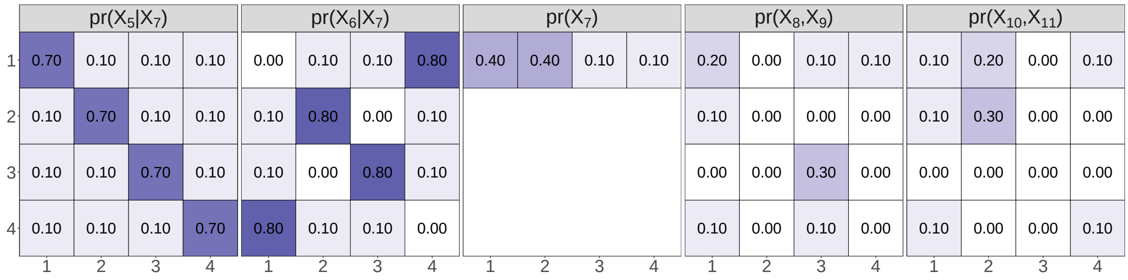
To effectively measure pairwise dependences a common strategy is in computing the discrepancy between the joint distribution and the product of the involved marginals, i.e. the joint distribution under independence assumption. When the two distributions are very similar we can conclude that there is high evidence of independence. Clearly, any distance can be used to measure such discrepancy, notable examples include model based Cramer V (e.g. Dunson and Xing, 2009; Dobra and Lenkoski, 2011) and normalized mutual information (e.g. Bhattacharya and Dunson, 2011).

The aforementioned distances have the main property of being defined in the $(0,1)$ interval, leading to a more clear interpretation of the strength of the dependence, however their computation might be subject to numerical instability when some components of the marginal probabilities are near zero. This instability is a known problem when measure of discrepancy for categorical variable are considered, for example when using the Pearson chi-square independence test, and consequently the Cramer V, (e.g. Agresti, 2013). To avoid such an issue in comparing models, we rely on L1-distance between the joint distribution and its equivalent under independence assumption.

2.3.1 Simulation settings

In a first simulation setting we focus on a scenario with $p = 30$ variables and, in order to introduce group dependence structure, we let the true p.m.f. to be decomposable in the following $G = 23$ blocks, $B_1^{(0)} = \{1,2,3,4\}$, $B_2^{(0)} = \{5,6,7\}$, $B_3^{(0)} = \{7,8\}$, $B_4^{(0)} = \{9,10\}$, $B_5^{(0)} = \{11,12\}$, $B_b^{(0)} = \{b\}$ for $b = 1,\dots,G$. Variables assigned to distinct blocks are considered as independents. We also fix the number of categories, $d_1,\dots,d_p = d = 4$ for all variables.

Inside each block we consider different generative mechanisms, specifically, for $B_1^{(0)}$ $\mathrm{pr}(X_1 = X_1 = X_3 = X_4 = d) = 0.0375$, we redistribute the residual of the 0.85 of the probability among other 5 random combination of the involved variables. For variables in $B_2^{(0)}$ we induce conditional dependence by letting $X_5 \perp\!\!\!\perp X_6 \mid X_7$. Probabilities to simulate variables in blocks $B_2^{(0)}$, $B_3^{(0)}$ and $B_4^{(0)}$ are presented in Figure 2.3. All the remaining variables are generated from discrete uniforms over $\{1,\dots,d\}$.



	pr(X5|X7)				pr(X6|X7)				pr(X7)				pr(X8,X9)				pr(X10,X11)			
1	0.70	0.10	0.10	0.10	0.00	0.10	0.10	0.80	0.40	0.40	0.10	0.10	0.20	0.00	0.10	0.10	0.10	0.20	0.00	0.10
2	0.10	0.70	0.10	0.10	0.10	0.80	0.00	0.10					0.10	0.00	0.00	0.00	0.10	0.30	0.00	0.00
3	0.10	0.10	0.70	0.10	0.10	0.00	0.80	0.10					0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00
4	0.10	0.10	0.10	0.70	0.80	0.10	0.10	0.00					0.10	0.00	0.10	0.00	0.10	0.00	0.00	0.10
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

FIGURE 2.3: Probability distributions for the variables belonging to blocks $B_2^{(0)}$, $B_3^{(0)}$ and $B_4^{(0)}$. Higher values have darker shade.

Theoretically, with the possible exception of the multinomial PPM, the considered models are able to retrieve any possible dependence structure in the data, provided that the sample size is large enough. Hence, we focus on scenarios with very low sample size compared to the effective number of parameters, namely $n = \{40,100\}$, to test how these models behave in recovering the dependence structure in these challenging settings.

In a second setting we consider sparse dependence induced by a Markovian structure of the simulated variables. We let X_j for $j = 2,3,4$, have the following transition probabilities $\mathrm{pr}(X_{j+1} \mid X_j = 1) = (0.3,0.7,0,0)^T$, $\mathrm{pr}(X_{j+1} \mid X_j = 2) = (0.2,0.1,0.7,0)^T$, $\mathrm{pr}(X_{j+1} \mid X_j = 3) = (0,0.2,0.1,0.7)^T$ and $\mathrm{pr}(X_{j+1} \mid X_j = 4) = (0,0,0.7,0.3)^T$, with initial distribution $\mathrm{pr}(X_1) = (0.5,0.2,0.2,0.1)^T$. Variable with indices from $j = 5,\dots,p = 30$ are instead generated from a discrete uniform over $\{1,2,3,4\}$.

The described scenario induces marginal dependence just in variable with indices $j = 1,2,3,4$, decaying with the distance of the indices. This kind of dependence may occur in practice when the variables' indices have some specific meaning, i.e. time or space. Again we consider a low sample size scenario, letting $n = 100$.

2.3.2 Prior settings

For the multinomial PPM model we make use of the cohesion function defined in 2.4. We set the upper bound $m = 3$ in such a way that it is lower than the highest cardinality of the groups composing the p.m.f., we use a symmetric Dirichlet prior with an small hyperparamter $\alpha = 0.01$, to favor sparsity. For the Parafac model we use the same setting proposed Dunson and Xing (2009), but we substitute the DP prior with a finite Dirichlet prior having hyperparamter $1/H$, setting the upper bound $H = 15$. For the GROT model we rely on Algorithm 3, using symmetric Dirichlet priors with hyperparamter set to $1/d$ for the tensor arms, fixing the upper bounds $H = 5$ and $m = 3$. For all the estimated models, we run the correspondent Gibbs sampling for 5000 iterations, discarding the first 2500 as burnin. In all cases trace plots inspection suggests convergence has been reached.

2.3.3 Simulation results

L1-distances between the bivariate distributions and the product of the corresponding marginals are shown in Figure 2.4, where the first row is for $n = 40$ while the second for $n = 100$. We can notice that, when there are very few data points, the multinomial PPM works well in identifying marginal dependence structures. The Parafac model, instead, fails to retrieve the correct structure in this example. This failure is probably due to the challenging scenario we proposed, and the model is expected to regain its flexibility when n increases. GROT model also fails to reconstruct the dependence structure in the first case. When $n = 100$, instead, we can notice that just variables sharing the same block present an L1-norm grater than 0. The model is perfectly able to detect the independent variables, that always present a L1-distance from the independence hypothesis approximately equal to 0.

Considering multinomial PPM and GROT models, we can also use as a proxy of the pairwise dependence the number of times variable share the same group. When using the GROT model groups are not necessarily index of the dependence since this last can be taken into account also by means of latent classes. Nevertheless, we empirically noted that strongly dependent variables are often in the same cluster. Results are shown in Figure 2.5.

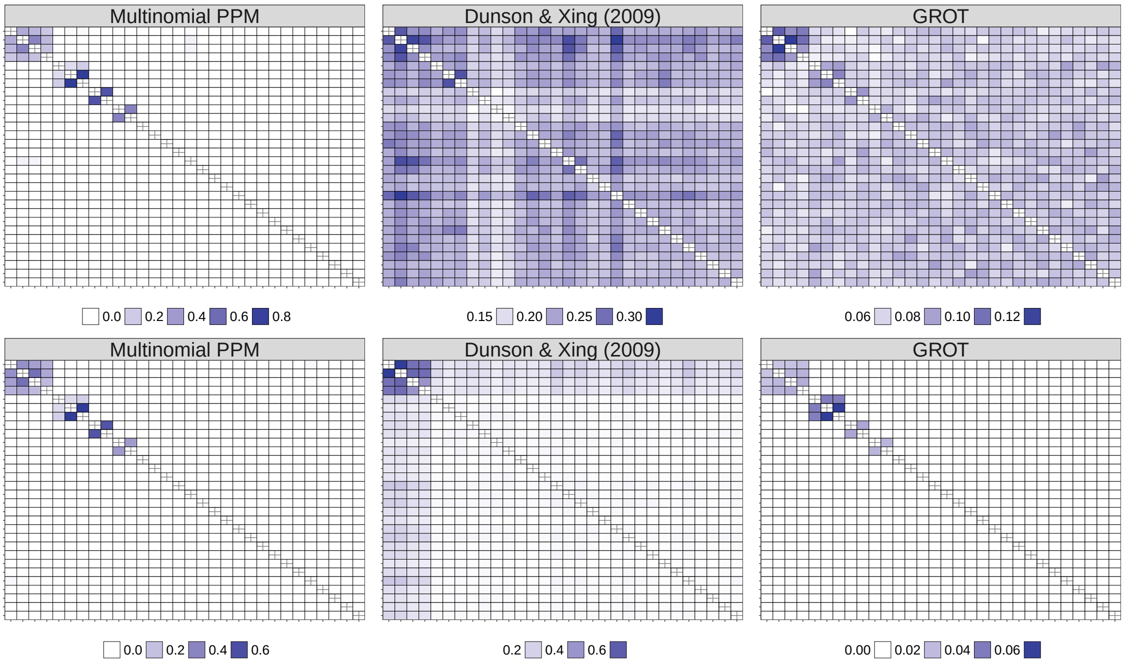


FIGURE 2.4: Posterior mean of L1-norm between joint bivariate distributions and the one obtained under independence assumption. First row is for sample size $n = 40$ and second row is for $n = 100$.

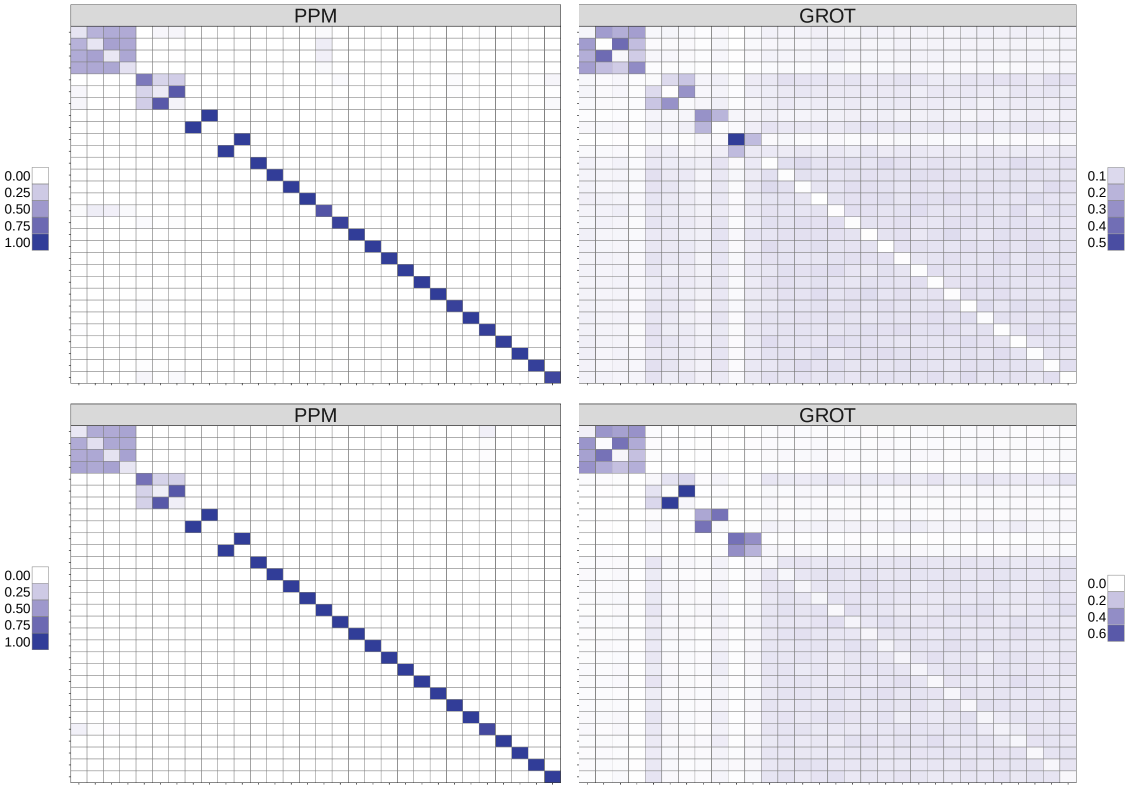


FIGURE 2.5: Proportion of time pair of variables are in the same clusters for multinomial PPM and GROT model. Value on the diagonal represent the number of times variables are in a singleton block. First row is for $n = 40$, second for $n = 100$

From Figure 2.5 we can notice that PPM models works amazingly well in identifying the dependence structure, including singleton blocks, even in cases where the sample size is very low. GROT factorization also produces a good proxy of the dependence block structure of the data, since dependent variables are clustered together with high probability. The model, however, fails to detect blocks with cardinality 1, that are grouped with other variables almost randomly.

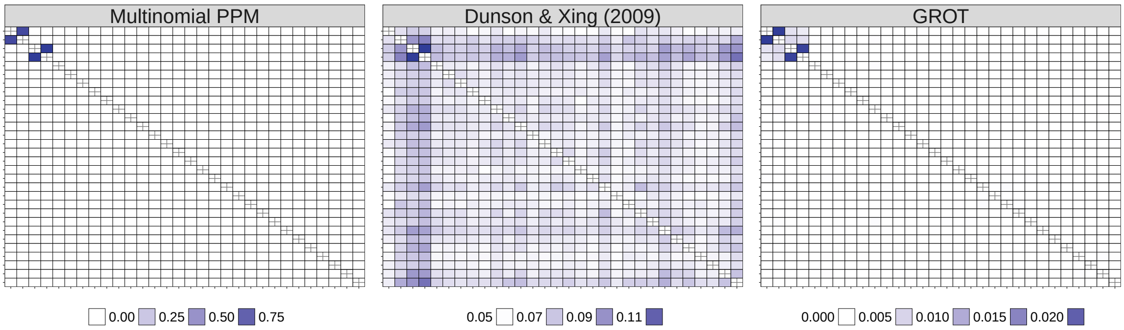


FIGURE 2.6: Posterior mean of L1-norm between joint bivariate distributions and the one obtained under independence assumption in the case of Markovian dependence.

When a well defined block dependence structure is present, the PPM model seems to have better screening ability than GROT factorization, since it is able to correctly identify the marginal dependence structure, including independent variables. This behavior is probably due to the fact that PPM proposal relies on a lower number of parameters than its competitors. Although the model is able to take into account dependence across blocks of the partition through prior (2.4), this approach may fail in correctly characterizing the dependence when it is due only to sparse subset of variables in a block. This behavior is indeed present in the second simulation scenario, in which PPM model totally misses association between some pairs of variables, e.g (X_2, X_3). All pairs of variable (X_j, X_k) for $j \neq k \in \{1,2,3,4\}$ are expected to be dependent, and correctly identified by GROT factorization, which produce L1-distance from the independence hypothesis different from 0 just for these variables (Figure 2.6). We can also notice how Parafac model gives some information on the dependence in the data, but results are noisy and there are many false positive.

We want to stress that we are considering scenarios with a very low signal to noise ratio, in fact when increasing the sample size (not shown) all models perform well in retrieving pairwise dependences.

2.4 Application to Promoter Data

In this Section we analyze the promoter data, publicly available at the UCI machine learning repository (Asuncion and Newman, 2007). A promoter is a genetic region which initiates the first step in the expression of a gene, known as transcription. The available data consist of $n = 106$ sequences of nucleotides A, C, G, T, observed at $p = 57$ locations, along with a binary response indicating if the sequence is a promoter or not. We can notice that the sample size is not very high compared to the length of the sequences, motivating the application of a compact model for the probability mass function.

In the data, there are 53 promoter and 53 non-promoter sequences of *Escherichia coli*, and the main aim is in identifying, if possible, the contact bases for the RNA polymerase, the enzyme responsible for starting the transcription, and in discriminating the promoter sequences.

Although the proposed model is introduced in an unsupervised fashion, to achieve such a goal we estimate the joint density of the binary response $Y_i \in \{0,1\}$, assuming value of 1 if the sequence i is a promoter, and the nucleotides sequence $\boldsymbol{X}_i$, obtaining an estimate of $\text{pr}(Y_i, \mathbf{X_i})$, and consequently the conditional p.m.f. $\text{pr}(Y_i \mid \mathbf{X_i})$. This approach has already been used in tensor factorization (e.g. Bhattacharya and Dunson, 2011) and it is akin to density regression in Bayesian nonparametric joint model for response and covariates (e.g. Wade *et al.*, 2014).

For posterior inference we consider the same setting of Section 2.3.2. In determining which locations can be a starting points for transcription, we measure the association between being a promoter and each of the 57 locations, computing the L1-distances from the independence assumption. Indeed, starting locations are expected to present similar nucleotide pattern across the sequences. Results are shown in Figure 2.7, indicating that nucleotides in position $j = 15, 16, 17$ are highly different in promoter and non promoter sequences respect to all other locations. In particular, $\text{pr}(X_j = T | Y_i = 1) \approx 0.406$ for $j = 15, 16$ and $\text{pr}(X_{17} = G | Y_i = 1) \approx 0.396$, while the same probabilities for non promoter sequences are all approximately 0.1. Hence, these bases are mainly preserved across the promoter sequences. This results is in accordance with the literature on this data, indicating that 17 and 16-base pair of basis, and the one in he immediate neighborhood, as possible staring points (O'Neill, 1989).

To test for predictive ability of our proposed approach, we compare the classification errors with the ones obtained trough standard classification models, namely Random forest and logistic lasso, implemented in R packages `randomForest` and `glmnet`, respectively. We also implement Dunson and Xing (2009) model as a direct competitor of our

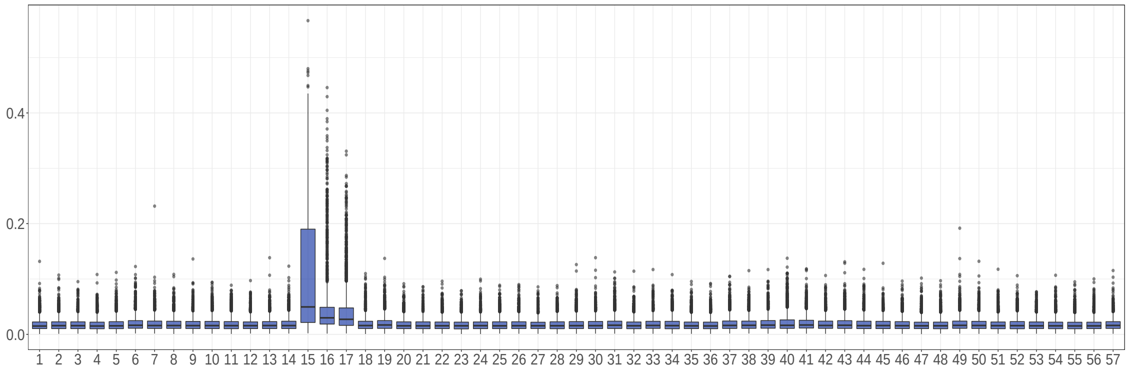


FIGURE 2.7: L1-norm between joint bivariate distribution of being promoter or not Y and each of the $p = 57$ locations X_j for $j = 1, \dots, p$, and the one obtained under independence assumption.

approach, relying on the same settings of Section 2.3.2.

For the comparison we make use of a 5 fold cross validation. While a pure Bayesian would undertake model assessment strictly on the basis of the posterior probabilities of competing Bayesian models, cross validation provides a practical way to compare Bayesian and non-Bayesian analyses of a particular data set. For a more theoretical justification of the use cross validation in a Bayesian settings one can refer to Alqallaf and Gustafson (2001).

We divided the dataset in 5 parts using in turn 4 for training the model and the remaining one to compute the out of sample classification errors. Classification errors from all the proposed models are shown in Table 2.1.

TABLE 2.1: 5 fold cross validation mean and standard deviation classification error for promoter sequences. First column is Total classification error, second is False Positive and third False negative.

Model	T Err.	FP Err.	FN Err.
Random Forest	**0.056(0.019)**	**0.035(0.049)**	0.070(0.068)
Lasso	0.094(0.066)	0.058(0.054)	0.118(0.120)
Parafac	0.321(0.189)	0.398(0.422)	0.164(0.200)
GROT	0.114(0.093)	0.204(0.207)	**0.015(0.034)**

We can notice that Random Forest produces the best classification error, with the exception of the false negative one. Our approach is close to lasso, if we consider the total classification error, but sensibly worse in terms of false positives. GROT factorization is uniformly better than the standard Parafac in this application. It is also worth noticing that our proposed method does not present dramatically worse performances than statistical learning tools, specifically designed for classification, while producing easily interpretable results.

Chapter 3

Multivariate mixed membership modeling: Inferring domain-specific risk profiles

In this Chapter we propose a novel multivariate generalization of mixed membership models, introduced in Chapter 1, which allows identification of correlated profiles related to different domains corresponding to separate groups of variables. Although we provide a broad new class of mixed membership models, which can be used in several contexts, we are initially motivated by the definition of survey based environmental and behavioral malaria risk profiles. Diffusion of malaria is a complex phenomenon evolving over time and space, driven by biological, behavioral and environmental factors acting together. We consider as a case study the Machadinho settlement project in Brazil, with the aim of defining survey based environmental and behavioral risk profiles and studying their interaction and evolution. To achieve this goal, we show that the use of correlated multiple membership vectors leads to interpretable inference requiring a lower number of profiles compared to standard formulations, while inducing a more compact representation of the population level model. We propose a novel multivariate logistic normal distribution for the membership vectors, which allows easy introduction of auxiliary information in the membership profiles leveraging a multivariate latent logistic regression. A Bayesian approach to inference, relying on Pólya gamma data augmentation, facilitates efficient posterior computation via Markov chain Monte Carlo. The proposed approach is shown to outperform the classical mixed membership model in simulations, and the malaria diffusion application.

In Section 3.2 we introduce our novel multivariate generalization of MM models, discussing some properties. Section 3.3 introduces a multivariate distribution defined on a product space of simplicies. In Section 3.4 we provide technical details on posterior computation. In Section 3.5 we study the performances of our model under different simulation scenarios, and in Section 3.6 we apply the model to the Machadinho malaria data.

3.1 Introduction and motivation

Malaria infection risk is largely driven by human behavior, especially in the later stages of disease diffusion, and its evaluation requires consideration of biological and economical aspects juxtaposed with behavioral and environmental factors. We focus on these lasts two aspects, with the aim of identifying social survey profiles, studying their evolution in time and space and their association with malaria diffusion. We consider the case of the Machadinho Settlement Project, located in the Rondônia state, Western Brazilian Amazon. The project was approved in 1982, with occupation starting in late 1984. The area was previously a forest sparsely inhabited by rubber tappers (Castro *et al.*, 2006).

Malaria became a problem in the area soon after occupation, with the proliferation of the *Anopheles Darlingi* mosquito, which is the principal vector for malaria diffusion in the Amazon area. Spread of malaria in frontier settlements is a complex phenomenon that can profoundly impact the ecosystem at different levels, including economic growth (Gallup and Sachs, 2001), land occupation (Zhang *et al.*, 2011), resource dilapidation due to direct costs of malaria prevention and treatment (e.g. Ettling and Shepard, 1991; Leighton and Foster, 1993), and labor productivity through working days lost to illness (Nur, 1993; Sauerborn *et al.*, 1995). Moreover, early childhood exposure to malaria can have long term consequences on physical and cognitive development (Jukes *et al.*, 2006).

For these reasons, the identification of risk profiles is very important in defining effective prevention and mitigation strategies to malaria diffusion. Crude measures of malaria risk (e.g. number of malaria cases reported in the households) are usually adopted to determine risk profiles through regression models. However, for the Machadinho settlement these indicators can lead to misleading results because of the presence of transient individuals, responsible for high malaria rates in zones presenting low risk conditions (see for example Castro *et al.*, 2006). Correcting for this effect would require integration of external ethnographic information and conjectures on land occupation, since highly migratory individuals are typically not included in the analysis, being not possible to trace. Unsupervised profiles based only on household features are not affected by this phenomenon, and can be linked to malaria risk, leading to more reliable findings for targeted stable populations in the settlement project.

Taking into account the distinction between behavioral variables (e.g. use of insecticide and presence/absence of crops in the settlement) and environmental variables (e.g. quality of roof, walls and housing), profiles are identified leveraging on a novel *Multivariate Mixed Membership Model (MMM)* which allows modeling different and possibly correlated latent traits.

Mixed membership models are used in a variety of contexts including for text analysis (Blei *et al.*, 2003), medicine (Erosheva *et al.*, 2007) and relational data (Airoldi *et al.*, 2005), among others. An extensive review can be found in Airoldi *et al.* (2014). Starting from multivariate observations, these admixture models can be used to estimate subject-specific latent scores corresponding to the weights in a discrete mixture model. The key idea is that the population is composed of H pure types, or profiles, and that each subject is endowed with a vector $\boldsymbol{\lambda_i} = (\lambda_{i1}, \ldots, \lambda_{iH})^T$, such that $0 < \lambda_{ih} < 1$ and $\sum_{h=1}^{H} \lambda_{ih} = 1$, indicating the degree of similarity of subject i with each of the H profiles. The vector $\boldsymbol{\lambda_i}$ induces a soft clustering of the subjects across the multiple profiles, which is particularly appealing when an exact grouping is difficult if not impossible to obtain, as for example in identification of disease risks (e.g. Chuit *et al.*, 2001; Castro *et al.*, 2006) or political ideology (e.g. Gross and Manrique-Vallier, 2012).

In many situations, interpretability is one of the most important features; hence it is desirable to employ a small number of profiles, ideally $H = 2$. For example, in epidemiology applications for $H = 2$, we can consider two profiles as corresponding to idealized healthy and unhealthy behavior, respectively. The weight vector $\boldsymbol{\lambda}_i$ then corresponds to values in $(0,1)$ summarizing the healthiness of subject i's behavior overall. This leads to a simple summary of risk. For example, in our malaria application in an unsupervised manner we can identity low and high risk profiles, can examine the differences in these profiles and also assign each household a risk score.

However, to accurately characterize the dependence structure in the data, usually a higher number of profiles is needed. Goodness-of-fit and interpretability are conflicting factors and, as noted in Singer and Castro (2014), when interpretability is the main concern, using separate models for different groups of variables might provide better insights. For example, one may define low and high risk profiles separately for behavioral and environmental variables; however, this simple approach ignores dependence across groups.

Our MMM framework addresses this issue by linking group-specific MM models through dependence in the membership scores. We show that this leads to a more compact representation of the joint probability mass function underlying the data, relaxing the constraints of the standard mixture membership model formulation. Additionally, we propose a novel joint distribution defined on a product space composed of simplices, leading to an easy-to-implement Gibbs sampler for posterior computation, based on Pólya gamma data augmentation (Polson *et al.*, 2013). The proposed framework allows simple inclusion of subject and group-specific covariates leveraging multiple latent logistic regression.

Motivated by the malaria risk application, we use the proposed framework to study environmental and behavioral membership scores evolving in time and space. The data consist of a 4 waves household survey conducted in Machadinho in 1985, 1986, 1987 and 1995. An occupied plot was defined as one at least partially cleared, or where settlers lived at least part-time. In 1985 70% of such plots were included in the survey, while in the remaining years 100% of such plots were considered; hence we do not take into account survey weights in the analysis. A core of 30 variables remained common to all the years, while some other questions were gradually added over time. Household spatial locations are also available and will be considered in the analysis. Although the survey was carefully administrated, the composition of the resulting data is highly heterogeneous across time and includes missing data. This is essentially due to the settlement process that was taking place in Machadinho, and the consequence of high mobility of the stable population in the area.

3.2 Multivariate mixed membership model

Considering the study of malaria risk profiles, standard application of the mixed membership model, briefly reviewed in Chapter 1, would fix $H = 2$ in order to have an easy to interpret model for the household risks. As discussed above, in this setting the use of standard mixed membership models, having a single score vector for all the variables, can lead to an unsatisfactory representation of the probability mass function due to the exchangeability assumption at the population level. Moreover, in the malaria context, the use of a single risk vector can mask the distinction between environmental and behavioral conditions, which can have a very different impact on spatio-temporal evolution of malaria diffusion (see for example Sawyer, 1992).

Motivated by these considerations, we generalize model (1.8), relaxing the exchangeability assumption by allowing group-specific mixed membership scores. We consider settings in which variables are naturally divided in advance into groups corresponding to different domains. Let $\boldsymbol{g} = (g_1, \dots, g_p)^T$ be an indicator vector of variable groups, where $g_j \in \{1, \dots, G\}$ for $j = 1, \dots, p$. Each subject is endowed with G membership score vectors $(\boldsymbol{\lambda}_i^{(1)^T}, \dots, \boldsymbol{\lambda}_i^{(G)^T})^T$ such that $\sum_{h=1}^{H} \lambda_{ih}^{(g)} = 1$ for $g = 1, \dots, G$. Note that the sum of the membership scores for the different domains is not equal to 1, i.e. $\sum_{g=1}^{G} \lambda_i^{(g)} \neq 1$.

The proposed model can be expressed in the following hierarchical form:

$$\begin{aligned} X_{ij} \mid Z_{ij} = h, \boldsymbol{\psi}_h^{(j)} &\sim \mathrm{Cat}(\psi_{h1}^{(j)}, \dots, \psi_{hd_j}^{(j)}), \\ Z_{ij} \mid \boldsymbol{\lambda}_i^{(g_j)} &\sim \mathrm{Cat}(\lambda_{i1}^{(g_j)}, \dots, \lambda_{iH}^{(g_j)}), \\ (\boldsymbol{\lambda}_i^{(1)^T}, \dots, \boldsymbol{\lambda}_i^{(G)^T})^T &\sim P. \end{aligned} \tag{3.1}$$

As in model (1.8), representation (3.1) relies on conditional independence of the observed variables given the profile labels; in fact the latent variables Z_{ij} are conditionally independent given the mixed membership scores $(\boldsymbol{\lambda}_i^{(1)}, \dots, \boldsymbol{\lambda}_i^{(G)})^T$:

$$\mathrm{pr}(X_{i1} = x_1, \dots, X_{ip} = x_p \mid \boldsymbol{\lambda}_i^{(1)}, \dots, \boldsymbol{\lambda}_i^{(G)}) = \sum_{h_1=1}^{H} \cdots \sum_{h_p=1}^{H} \lambda_{ih_1}^{(g_1)} \cdots \lambda_{ih_p}^{(g_p)} \prod_{j=1}^{p} \psi_{h_j x_j}^{(j)}. \tag{3.2}$$

Integrating out the scores from equation (3.2), we obtain the population level model:

$$\mathrm{pr}(X_1 = x_1, \dots, X_p = x_p \mid \boldsymbol{\psi}) = \sum_{h_1=1}^{H} \cdots \sum_{h_p=1}^{H} \bar{a}_{h_1 \dots h_p} \prod_{j=1}^{p} \psi_{h_j x_j}^{(j)}. \tag{3.3}$$

Although equation (3.3) seems identical to equation (1.10), the elements of the core tensors are different, as are the imposed constraints. The core tensor $\bar{\mathscr{A}} = \{\bar{a}_{h_1, \dots, h_p}, h_j = 1, \dots H; j = 1, \dots, p\}$ is not a symmetric tensor, but it has some equality constraints on the elements. Specifically, given the vector of indices $\boldsymbol{h} = (h_1, \dots, h_p)^T$, and a group indicator vector $\boldsymbol{g} = (g_1, \dots, g_p)^T$, we can define a group preserving permutation space $\mathfrak{S}_{\boldsymbol{h}}^{\boldsymbol{g}}$ such that the effect of $\bar{\sigma} \in \mathfrak{S}_{\boldsymbol{h}}^{\boldsymbol{g}}$ is to permute the elements of a vector within the groups, leaving the group structure unchanged. It immediately follows that $\mathfrak{S}_{\boldsymbol{h}}^{\boldsymbol{g}}$ is a well defined group since it is closed to composition, while also respecting associativity, identity and invertibility properties (e.g. Artin, 1991).

The core tensor $\bar{A}$ can be defined as a group symmetric tensor, meaning that given a multivariate index $\boldsymbol{h}$ we have $\bar{a}_{\boldsymbol{h}} = \bar{a}_{\bar{\sigma}(\boldsymbol{h})}$, for all $\bar{\sigma} \in \mathfrak{S}_{\boldsymbol{h}}^{\boldsymbol{g}}$. Note that a symmetric tensor, defined 1.1, can be viewed as group symmetric with only one group, or can be defined such that it is group symmetric for any possible group configuration $\boldsymbol{g}$.

The number of distinct elements in $\bar{\mathscr{A}}$ is given by $\prod_{g=1}^{G} H^{p_g}/p_g!$, which is considerably larger than in the symmetric tensor case, as can be seen from Figure 3.1.

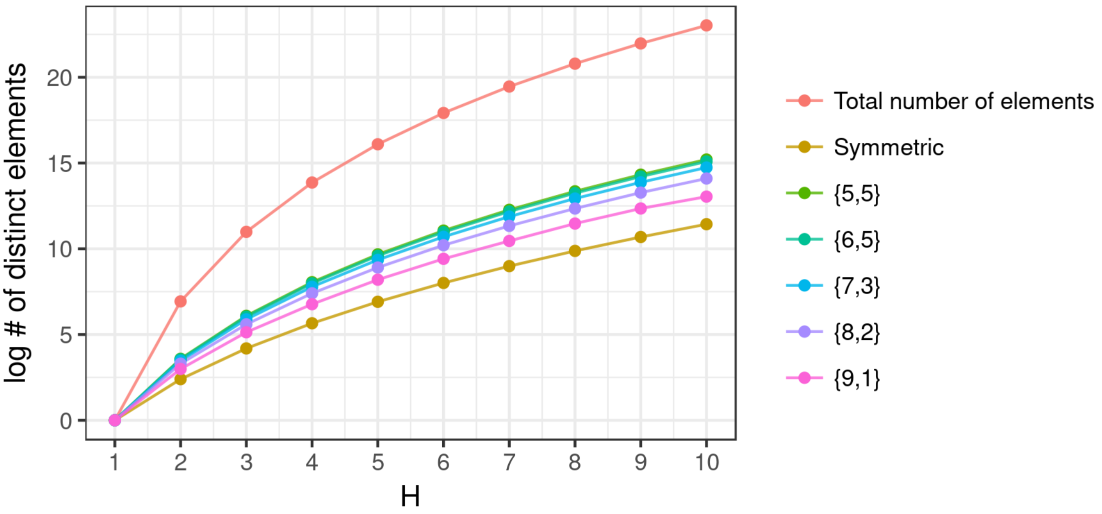


FIGURE 3.1: Logarithm of the number of distinct elements in a cubic tensor of dimension H^{10}, for symmetric and group symmetric tensors for all configurations of two groups.

Moreover, as a consequence of Theorem 3.1, equation (3.3) is a closer approximation of the 'true' probability mass function generating the data than equation (1.10) for any fixed H.

Theorem 3.1. *Let π_0 be a probability tensor of dimension $d_1 \times \ldots \times d_p$, π^{gsym} and π^{sym} be, respectively, the best group symmetric and symmetric multi-rank H approximations of π_0. Then $\|\pi_0 - \pi^{gsym}\|_F \leq \|\pi_0 - \pi^{sym}\|_F$, where $\|\cdot\|_F$ denote the Frobenius norm.*

Proof. The proof immediately follows after noticing that π^{gsym} by definition minimizes $\|\pi_0 - \pi^{\text{gsym}}\|_F$ under the constraint $\bar{a}_{\boldsymbol{h}} = \bar{a}_{\bar{\sigma}(\boldsymbol{h})}$ for all $\bar{\sigma} \in \mathfrak{S}^{\boldsymbol{g}}_{\boldsymbol{h}}$, and that π^{sym} can be obtained by solving the same problem with additional equality constraints to the element of the core tensor $\bar{\mathscr{A}}$ such that, $\bar{a}_{\boldsymbol{h}} = \bar{a}_{\boldsymbol{h}'}$ if $\boldsymbol{h} = \sigma(\boldsymbol{h}')$ for a $\sigma \in \mathfrak{S}$. □

Theorem 3.1 implies that incorporating group-specific membership scores provides a more compact representation of the true probability mass function π_0 generating the data, for any fixed H. Hence, if we fix $H = 2$ for interpretability, we will tend to produce a better fit to the data in using group-specific scores than in modeling a single global score vector. To complete a specification of the MMM model, it remains to choose an appropriate distribution P.

3.3 Multivariate Logistic Normal Distribution

Letting $S_H = \{\boldsymbol{x} \in [0,1]^H : \sum_{h=1}^{H} x_h = 1\}$ denote the $H-1$ probability simplex, we aim to define a joint distribution on the product space $S = S_{H_1} \otimes \cdots \otimes S_{H_G}$. To achieve this goal, we start from a distribution on $\mathbb{R}^{\sum_{g=1}^{G}(H_g-1)}$, mapping to S via an appropriate transformation. Potentially any continuous multivariate distribution can be used, but we focus on the multivariate Gaussian distribution to retain simplicity and flexibility.

Normal distribution, as linear regression, have been extensively used to parsimoniously model relationships in tensor data elements in different frameworks. Notable examples include Hoff (2015) in which he introduces a general multilinear tensor having Kronecker-structured regression parameters along with a separable covariance model; Guhaniyogi *et al.* (2017), that consider instead, a Bayesian approach to tensor regression, having a scalar response. A more general inference framework has been recently introduced in Lock (2018), in which he proposes a general tensor on tensor regression framework that includes several of the previous tensor regression models as special cases. We follow their lead to introduce a flexible distribution defined on the product of simplicies space S.

Specifically, let $\boldsymbol{Y} = (\boldsymbol{Y}^{(\mathbf{1})^T}, \ldots, \boldsymbol{Y}^{(\mathbf{G})^T})^T$ be a multivariate normal distribution of dimension $\sum_{g=1}^{G}(H_g - 1)$ with mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}^{(\mathbf{1})^T}, \ldots, \boldsymbol{\mu}^{(\mathbf{G})^T})^T$, where $\boldsymbol{\mu}^{(g)} \in \mathbb{R}^{H_g-1}$, and covariance matrix $\boldsymbol{\Sigma}$. We consider the transformed vector $\boldsymbol{X} = (\boldsymbol{X}^{(\mathbf{1})^T}, \ldots, \boldsymbol{X}^{(\mathbf{G})^T})^T$, whose elements can be defined as $X_h^{(g)} = \exp\{Y_h^{(g)}\}[1 + \sum_{k=1}^{Hg-1} \exp\{Y_k^{(g)}\}]^{-1}$ for $h = 1, \ldots, H_g - 1$ and $g = 1, \ldots, G$, with $X_{H_g}^{(g)} = [1 + \sum_{k=1}^{Hg-1} \exp\{Y_k^{(g)}\}]^{-1}$.

It is easy to show that $\boldsymbol{X} \in S$ and that the Jacobian matrix of the transformation is block diagonal having determinant given by $[\prod_{g=1}^{G} \prod_{h=1}^{H_g} X_h^{(g)}]^{-1}$. The probability density function of the resulting distribution is

$$f_X(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}^\star - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^\star - \boldsymbol{\mu})\right\}}{(2\pi)^{\sum_{g=1}^{G}(H_g-1)/2} |\boldsymbol{\Sigma}|^{1/2} \prod_{g=1}^{G} \prod_{h=1}^{H_g} x_h^{(g)}}, \tag{3.4}$$

where $\boldsymbol{x}^\star = \text{conc}\left(\left\{\log(x_h^{(g)}/x_{H_g}^{g}), \text{ for } h = 1, \ldots, (H_g - 1); g = 1, \ldots, G\right\}\right)$. Each of the group marginals $\boldsymbol{X}^{(\boldsymbol{v})}$ has a logistic normal distribution with parameters $\boldsymbol{\mu}^{(\boldsymbol{v})}$ and $\boldsymbol{\Sigma}^{(\boldsymbol{v})}$, where $\boldsymbol{\Sigma}^{(\boldsymbol{v})}$ is the block of the matrix $\boldsymbol{\Sigma}$ corresponding to the v-th group; for this reason, we refer to (3.4) as the *Multivariate Logistic Normal Distribution (MLND)*. Distribution (3.4) can be alternatively derived as a compound distribution from a collection of independent logistic normal distributions and a multivariate normal for the concatenation of the means, as stated in Proposition 2.

Proposition 2. Let $\boldsymbol{X} = (\boldsymbol{X}^{(\mathbf{1})^T}, \ldots, \boldsymbol{X}^{(\mathbf{G})^T})^T \in S$ such that $\boldsymbol{X}^{(\boldsymbol{g})} \mid \boldsymbol{\mu}^{(\boldsymbol{g})} \sim \text{LogitNormal}(\boldsymbol{\mu}^{(\boldsymbol{g})}, \boldsymbol{\Sigma}^{(\boldsymbol{g})})$ independently for $g = 1, \ldots, G$, and let $\boldsymbol{\mu} = (\boldsymbol{\mu}^{(\mathbf{1})^T}, \ldots, \boldsymbol{\mu}^{(\mathbf{G})^T}) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{0})$. Then $\boldsymbol{X} \sim \text{MLND}(\boldsymbol{\mu}_\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_\mathbf{0} + \text{block}(\boldsymbol{\Sigma}^{(\mathbf{1})}, \ldots, \boldsymbol{\Sigma}^{(\boldsymbol{G})})$.

Following Aitchison and Shen (1980), we consider a class of distribution preserving transformations, useful to maintain some invariance properties of the induced distribution. According to our problem, we additionally restrict our attention to the sub-class of group preserving transformations (e.g. group permutation defined in Section 3.2).

Proposition 3. Let $\boldsymbol{X} = (\boldsymbol{X}^{(1)^T}, \dots, \boldsymbol{X}^{(G)^T})^T \sim \text{MLND}(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{B}$ a $Q \times \sum_{g=1}^{G}(H_g - 1)$ block diagonal matrix, having diagonal blocks $\boldsymbol{B}^{(g)}$ of dimension $q_g \times (H_g - 1)$ for $g = 1 \dots, G$, then the $Q \times G$ dimensional vector $\mathbf{X}'$ whose elements are defined as

$$x'^{(g)}_q = \prod_{h=1}^{H_g-1} \left(\frac{x_h^{(g)}}{x_{H_g}^{(g)}} \right)^{b_{qh}^{(g)}} \left[1 + \sum_{k=1}^{q_g} \prod_{h=1}^{H_g-1} \left(\frac{x_h^{(g)}}{x_{H_g}^{(g)}} \right)^{b^{(g)}{kh}} \right]^{-1}, \quad q = 1, \dots, q_g, \quad g = 1, \dots G,$$

has distribution $\boldsymbol{X'} \sim \text{MLND}\left(\boldsymbol{B\mu}, \boldsymbol{B}\Sigma\boldsymbol{B}^T\right)$.

The diagonal block structure of matrix $\boldsymbol{B}$ in Proposition 3 assures that the transformation preserves the same group structure of the original vector, and it can be structured to accommodate changes of the baseline category, permutation of the labels, and merging of one or more categories within groups.

The proposed MLND distribution has finite moments, but these moments in general do not have a simple analytic form. However, we can obtain simple expressions for moments related to log-odds and odds ratios both between and across the groups.

For example, letting

$$\mathrm{m}_l(h,g;h',g') = \mathbb{E}\left[\log\left(\frac{X_h^{(g)}/X_{H_g}^{(g)}}{X_{h'}^{(g')}/X_{H_{g'}}^{(g')}}\right)\right] \quad \text{and} \quad \mathrm{m}_o(h,g;h',g') = \mathbb{E}\left[\left(\frac{X_h^{(g)}/X_{H_g}^{(g)}}{X_{h'}^{(g')}/X_{H_{g'}}^{(g')}}\right)\right],$$

we have

$$\begin{aligned} \mathrm{m}_l(h,g;h',g') &= \mu_h^{(g)} - \mu_{h'}^{(g')}, \\ \mathrm{m}_o(h,g;h',g') &= \exp\left\{\mu_h^{(g)} - \mu_{h'}^{(g')} + \frac{1}{2}\left[\Sigma_{hh}^{(g)} + \Sigma_{h'h'}^{(g')} - 2\Sigma_{hh'}^{(g,g')}\right]\right\}, \end{aligned} \tag{3.5}$$

where with an abuse of notation we indicate with $\Sigma_{hk}^{(g,v)}$ the element in position (h,k) of the non diagonal block corresponding to the groups g and v. Higher order moments can also be computed relying on normal and log-normal distribution properties.

From equation (3.5) we can notice that the log-odds of the elements in different groups are linearly related. Moreover, when applied to multivariate mixed membership models with $H_g = 2$, log-odds and odds ratios give important insights on which group is more important in characterizing high and low risk conditions.

Although having a simple expression, the proposed MLND adapt to different scenarios, characterizing many possible dependence structure across the profiles. Figure 3.2 shows contour plots of four different MLND distributions with $G = 2$ and $H_1 = H_2 = 2$, highlighting how for different values of the parameter the distribution changes.

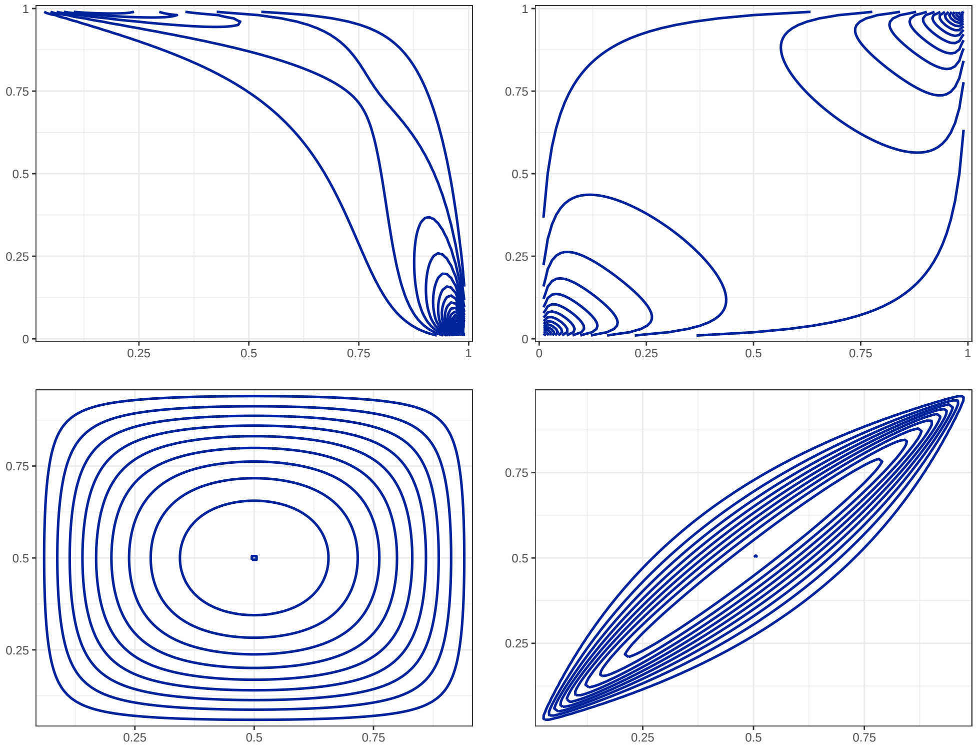


FIGURE 3.2: Bivariate MLND p.d.f. for different values of the parameters.

3.4 Posterior Computation

We propose an algorithm to simulate from the posterior of model (3.1), with $(\boldsymbol{\lambda}_i^{(1)^T}, \dots, \boldsymbol{\lambda}_i^{(G)^T})^T \sim \text{MLND}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ defined in (3.4).

Following the detailed motivation earlier in the Chapter, we focus on the special case in which $H_g = 2$, for $g = 1, \dots, G$. Generalization to more pure types can be obtained iterating the proposed Pólya gamma data augmentation on all the conditional log-odds (Polson and Scott, 2011), or alternatively relying on the stick breaking parametrization of the multinomial likelihood introduced in Linderman *et al.* (2015).

Enhancing computational flexibility, we specify conjugate prior distributions for all the parameters in the model. Specifically, for the kernel probabilities $\boldsymbol{\psi}_{\boldsymbol{h}}^{(j)} \sim$

$\mathrm{Dir}(\alpha_1^{(j)}, \dots, \alpha_{d_j}^{(j)})$, and for the hyperparameter $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$, and for the variance and covariance matrix $\boldsymbol{\Sigma} \sim IW(\nu_0, \boldsymbol{\Psi}_0)$. Parameters can be updated iterating the steps in Algorithm 4.

Algorithm 4: Posterior computation for the MMM model

```
for $r$ in 1:n_iterations do
    [1] Update the kernel probabilities
    for $j$ in $1:p$ & $h$ in $1:2$ do
        $$\boldsymbol{\psi}_{\boldsymbol{h}}^{(j)} \mid - \sim \mathrm{Dir}\Big(\alpha_1^{(j)} + \sum_{i:z_{ij}=h} 1\{x_{ij}=1\}, \dots, \alpha_{d_j}^{(j)} + \sum_{i:z_{ij}=h} 1\{x_{ij}=d_j\}\Big),$$
    where $1\{\cdot\}$ is the indicator function;
    [2] Considering the model $\mathrm{pr}(Z_{ij}=2 \mid \lambda_i^{(g_j)}) = \lambda_i^{(g_j)}$, we can sample the profile
    indicator with probability
    for $i$ in $1:n$ & $j$ in $1:p$ do
        $$\mathrm{pr}(Z_{ij}=2 \mid -) = \frac{\lambda_i^{(g_j)} \psi_{2x_{ij}}^{(j)}}{(1-\lambda_i^{(g_j)})\psi_{1x_{ij}}^{(j)} + \lambda_i^{(g_j)} \psi_{2x_{ij}}^{(j)}}.$$
    [3] We make use of Pólya gamma data augmentation to retrieve conjugacy
    between binomial and logistic normal distributions. We consider the augmented
    variables
    for $i$ in $1:n$ & $g$ in $1:G$ do
        $$\omega_i^{(g)} \mid - \sim \mathrm{PG}(p_g, \mathrm{logit}(\lambda_i^{(g)})),$$
        where $p_g = \sum_{j=1}^{p} 1\{g_j = g\}$ is the number of variables in $g$-th group for
        $g = 1, \dots, G$.
    [4] We define $k_i^{(g)} = \sum_{j=1}^{p_g} 1\{Z_{ij}=2\} - p_g/2$, and we have that the vector
    $(\boldsymbol{k_i}/\boldsymbol{\omega_i}) = (k_i^{(1)}/\omega_i^{(1)}, \dots, k_i^{(G)}/\omega_i^{(G)})^T \mid \boldsymbol{\lambda_i} \sim \mathcal{N}(\mathrm{logit}(\boldsymbol{\lambda_i}), \mathrm{diag}(1/\omega_i^{(1)}, \dots, 1/\omega_i^{(G)}))$ and
    we can update the membership scores from
    for $i$ in $1:n$ do
        $$\boldsymbol{\lambda_i} \mid \boldsymbol{\mu} \sim \mathrm{MLND}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star),$$
    where $\boldsymbol{\Sigma}^\star = \big(\mathrm{diag}(\omega_i^{(1)}, \dots, \omega_i^{(G)}) + \boldsymbol{\Sigma}^{-1}\big)^{-1}$ and $\boldsymbol{\mu}^\star = \boldsymbol{\Sigma}^\star(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{k_i})$.
    [5] We can update the vector $\boldsymbol{\mu}$ integrating out the membership scores vectors $\boldsymbol{\lambda_i}$,
    we have that the vector $(\boldsymbol{k_i}/\boldsymbol{\omega_i}) \mid \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Upsilon_i^{-1}})$, where
    $\boldsymbol{\Upsilon_i^{-1}} = (\mathrm{diag}(1/\omega_i^{(1)}, \dots, 1/\omega_i^{(G)}) + \boldsymbol{\Sigma})$, and hence the full conditional is given by
    $$\boldsymbol{\mu} \mid - \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),$$
    where $\boldsymbol{\Sigma}^* = \big(\sum_{i=1}^{n} \boldsymbol{\Upsilon_i} + \boldsymbol{\Sigma_0^{-1}}\big)^{-1}$ and $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*\big(\sum_{i=1}^{n} \boldsymbol{\Upsilon_i} \boldsymbol{k_i}/\boldsymbol{\omega_i} + \boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0}\big)$.
    [6]We finally update the covariance matrix and its parameters from the full
    conditional
    $$\boldsymbol{\Sigma} \mid - \sim IW\Big(\nu_0 + n, \boldsymbol{\Psi}_0 + \sum_{i=1}^{n} (\mathrm{logit}(\boldsymbol{\lambda_i}) - \boldsymbol{\mu})(\mathrm{logit}(\boldsymbol{\lambda_i}) - \boldsymbol{\mu})^T\Big).$$
```

In step 5 we update the mean vector of a multivariate Gaussian distribution; this step can be substituted with a multivariate regression in order to take into account covariate effects.

Potentially, for our MMM model, as in other MM models and more broadly for mixture models, we may encounter label switching. This occurs when the external profiles change their meaning across the MCMC iterations. In such a case post processing should be used to appropriately align the MCMC samples (see for example Stephens, 2002). However, such post processing was not applied in any of the simulated data or malaria analysis we report below, as there was no evidence of label switching in the MCMC samples.

3.5 Simulation study

We analyze different simulation scenarios in evaluating the empirical performance of our proposed approach. We consider different probability distribution functions for the membership scores P, relying on hierarchical representation (3.1) to generate the data. The goal in defining these scenarios is to assess whether the proposed model can characterize generative mechanisms having broadly different properties. We compare our results with the standard admixture formulation implemented in the R package `mixedMem`, using separate models for each group. We initially assume that $H_g = 2$ is the 'true' number of extreme profiles, presenting four different scenarios, while in a second section we consider the misspecified case $H_g > 2$.

3.5.1 Number of profiles correctly specified

We consider $G = 2$ groups, $n = 1000$ subjects, $p_g = 5$ categorical variables, having $d_j = d = 4$ modalities and $H_g = 2$ profiles for $g = 1,2$. We simulate data from categorical distributions, whose probabilities are drawn from a Dirichlet distribution with parameters $\varphi_h^{(g)}$ having values $\varphi_1^{(1)} = (10,3,2,1)^T$, $\varphi_2^{(1)} = (1,1,1,11)^T$, $\varphi_1^{(2)} = (5,5,1,0)^T$ and $\varphi_2^{(2)} = (1,1,1,8)^T$.

In the first simulation scenario, we let the probability density function for the joint distribution of the score vectors $(\lambda_i^{(1)}, \lambda_i^{(2)})^T$ be a bivariate truncated normal distribution over the unit square, having parameter $\boldsymbol{\mu} = (0.5, 0.5)^T$ and $\text{vec}(\boldsymbol{\Sigma}) = (\Sigma_{11}, \Sigma_{21}, \Sigma_{12}, \Sigma_{22})^T = (0.05, 0.02, 0.02, 0.05)^T$. This formulation induces positive dependence between the two scores with their distribution having ellipsoid contours truncated at the borders. In the second simulation scenario, we consider the distribution proposed in Section 3.3 with

$\boldsymbol{\mu} = (-1.2, 1)^T$ and $\text{vec}(\boldsymbol{\Sigma}) = (3.0, -2.4, -2.4, 3.5)^T$. In the third scenario, we rely on the generative mechanism (1.8), having profile distribution shared by all variables; we generate this profile from a uniform distribution. Finally, in the fourth simulation scenario, we consider P to be the product of two independent uniforms, forcing independence in the variables belonging to different groups, which translates into the case in which two separate models for the groups represents the correctly specified model.

We perform posterior inference under the proposed model (3.1) with priors defined in Section 3.4, setting $\alpha_1^{(j)} = \ldots, \alpha_{d_j}^{(j)} = 1/d_j$ for $j = 1, \ldots, p$, we consider $\boldsymbol{\mu_0} = (0,0)^T$, $\boldsymbol{\Sigma_0} = \boldsymbol{I}$, $\nu_0 = 2$ and $\Psi_0 = \boldsymbol{I}$. We maintained these default hyperparameters in all our simulation cases, collecting 5000 Gibbs samples from Algorithm 4. Trace plots suggest convergence is reached by a burn-in of 1000.

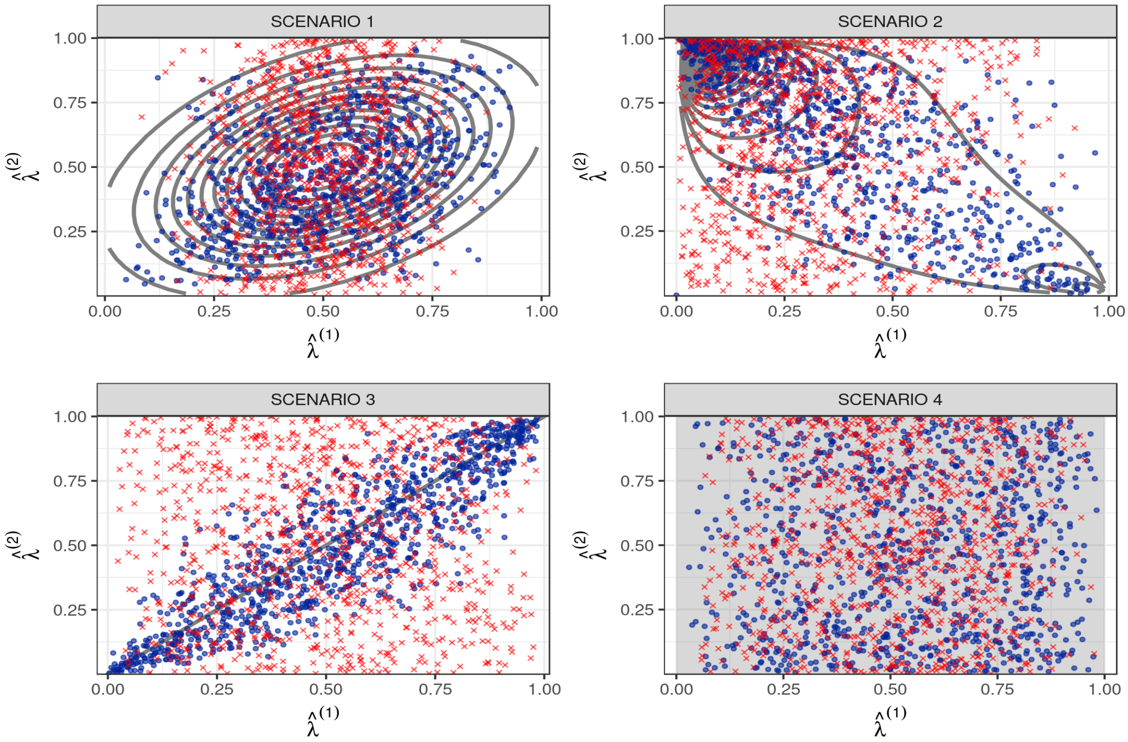


FIGURE 3.3: 1000 samples from estimated membership scores distribution from model (3.1) (blue dots) and `mixedMem` (orange crosses). Grey area represents the contour of the true profiles distribution.

Figure 3.3 shows the estimated profiles distribution P for all simulated scenarios, comparing results with the use of two separate MM models. Despite the challenging scenarios and the misspecification of the profile distribution, our proposed approach is able to reconstruct the latent mechanism underlying the profiles in a satisfactory way.

In evaluating subject-specific estimates of the scores $(\lambda_i^{(1)}, \lambda_i^{(2)})$, we rely on mean square error relative to the 'true' value. We obtain good results in retrieving the 'true' membership vectors in all simulation scenarios (Table 3.1) as the proposed approach always produces better or comparative results compared to the standard mixed membership model implemented in the package `mixedMem`.

TABLE 3.1: Mean square error for predicting individual membership scores $(\lambda_i^{(1)}, \lambda_i^{(2)})$ in all simulation scenarios.

	SCENARIO 1	SCENARIO 2	SCENARIO 3	SCENARIO 4
MMM g = 1	0.026(0.035)	0.027(0.035)	0.021(0.028)	0.038(0.046)
MMM g = 2	0.025(0.033)	0.027(0.045)	0.021(0.031)	0.036(0.051)
mixedMem g = 1	0.029(0.035)	0.034(0.046)	0.036(0.044)	0.041(0.046)
mixedMem g = 2	0.030(0.041)	0.037(0.050)	0.034(0.050)	0.036(0.048)

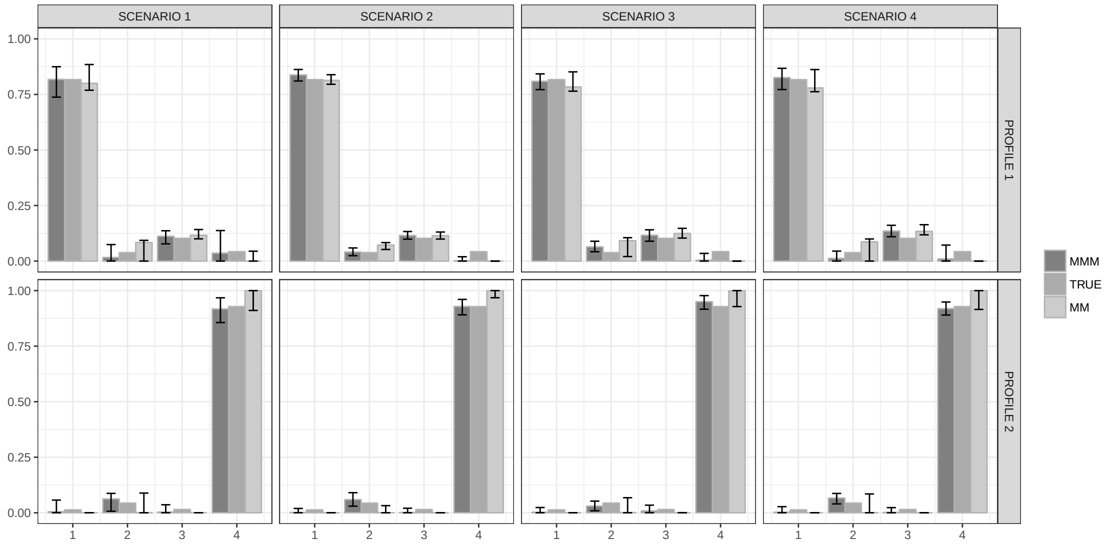


FIGURE 3.4: True values of the estimated profiles $\boldsymbol{\psi}_h^{(j)}$ for $h = 1,2$ of a representative variable in group $g_j = 1$. Bars represent 0.1 and 0.9 posterior quantiles for our MMM model and bootstrap 0.8 confidence intervals for the MM model estimated with the `MixedMem` package.

Figures 3.4 and 3.5 show posterior estimates and credible intervals for the kernel parameters $\boldsymbol{\psi}_h^{(j)}$ for selected variables in both scenarios. We notice that our proposed approach robustly estimates kernels in the different proposed scenarios. Contrarily in some cases, the MM model underestimates variability. This behavior is evident in the lower part of Figures 3.4 where MM produces confidence intervals for $\psi_{21}^{(j)}$, $\psi_{22}^{(j)}$, $\psi_{23}^{(j)}$ collapsing to 0 inappropriately.

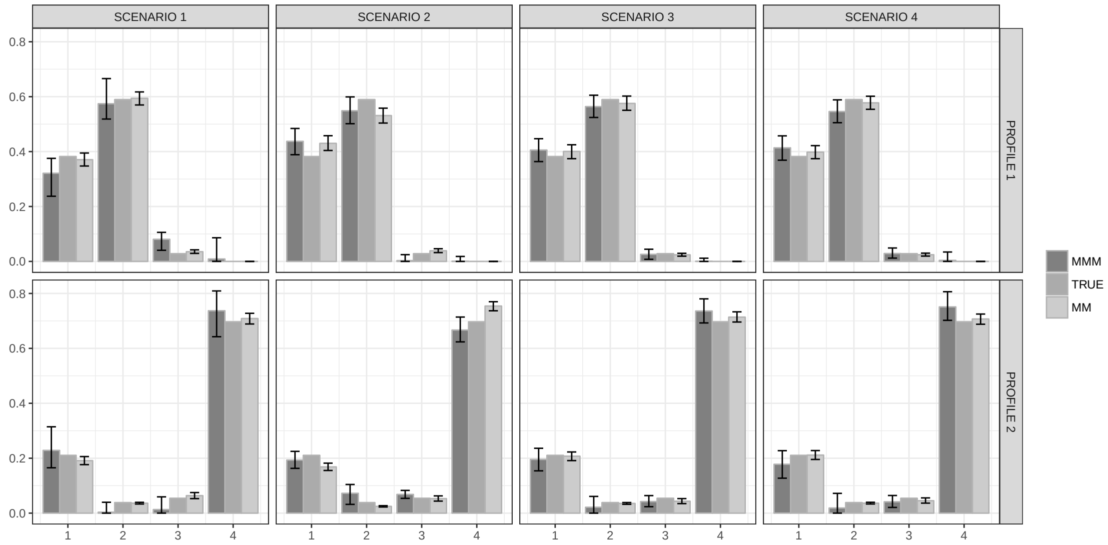


FIGURE 3.5: True values of the estimated profiles $\psi_h^{(j)}$ for $h = 1,2$ of a representative variable in group $g_j = 2$. Bars represent 0.1 and 0.9 posterior quantiles for our MMM model and bootstrap 0.8 confidence intervals for the MM model estimated with the `MixedMem` package.

3.5.2 Misspecification: more than two pure types

In this Section we consider a scenario in which generative model (3.1) has more than 2 pure types, while retaining the proposed inference model with $H_g = 2$ for $g = 1,2$. The key idea is to understand how the model is able to approximate the profiles in a lower dimensional space, and compare this approximation with that for the standard MM model. Specifically we consider as generative mechanism a $G = 2$ group model with $H_1^0 = 4$. Kernels for the first group are fixed as $\varphi_1^{(1)} = (0.85, 0.05, 0.05, 0.05)^T$, $\varphi_2^{(1)} = (0.05, 0.85, 0.05, 0.05)^T$, $\varphi_3^{(1)} = (0.05, 0.05, 0.85, 0.05)^T$ and $\varphi_4^{(1)} = (0.05, 0.05, 0.05, 0.85)^T$, while membership scores $(\lambda_{i1}^{(1)}, \dots \lambda_{i4}^{(1)})^T \sim \text{Dirichlet}(0.25, 0.25, 0.25, 0.25)$. For the second group we consider instead the same mechanism used in scenario 4 with $H_2^0 = 2$, enforcing no dependence in the scores distribution. This scenario is constructed to favour the use of two separate MM models, having no dependence across the groups and a Dirichlet distribution for the profiles.

Figure 3.6 shows the relations between each component of the 'true' unknown $\lambda_{i1}^{(1)}, \dots, \lambda_{i4}^{(1)}$ and the estimates from our proposed approach. As expected, the estimated score for both models is strongly correlated with more 'true' profiles. Some individual variability is lost in the process as we are projecting a 3-dimensional space to a 1-dimensional one.

By reducing the dimensionality we obtain 'mixed' pure types that can be considered as averages of the 'true' ones. For example, MMM model profile 2 is composed of subjects

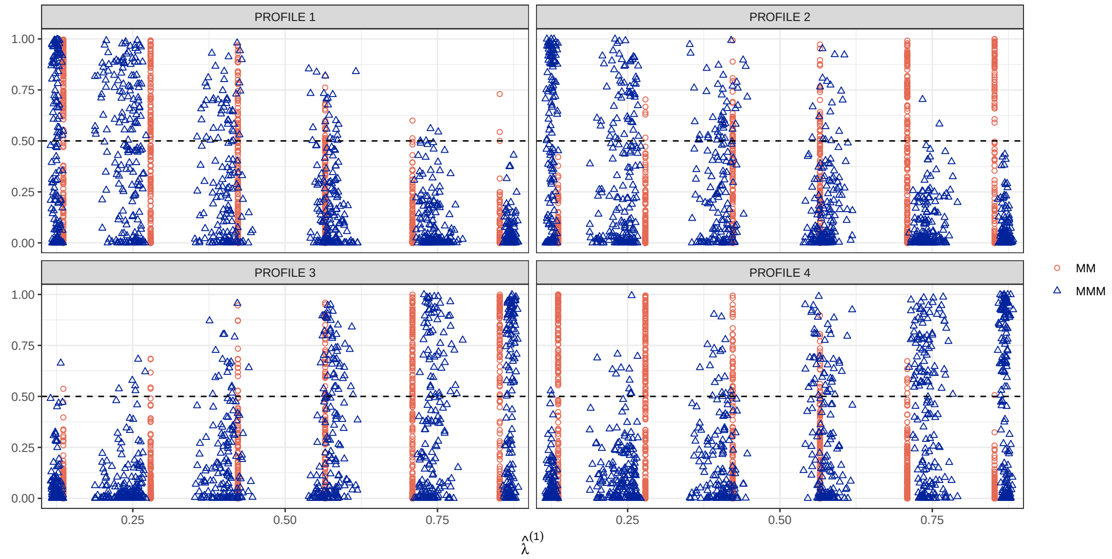


FIGURE 3.6: Estimated (x-axis) and 'true' (y-axis) values for the score vectors, relying on MMM and `MixedMem` models for $g = 1$.

with high values of $\lambda_{i3}^{(1)}$ and $\lambda_{i4}^{(1)}$, and low values of $\lambda_{i1}^{(1)}$ and $\lambda_{i2}^{(1)}$, while in the MM model profile 2 is composed of high values of $\lambda_{i2}^{(1)}$ and $\lambda_{i3}^{(1)}$ and low values of $\lambda_{i1}^{(1)}$ and $\lambda_{i4}^{(1)}$. This can be assessed by looking at the estimated kernels for a representative variable in group 1 (see Table 3.2).

TABLE 3.2: Estimated kernels for MMM and MM model for variable 1 in group $g_1 = 1$. Numbers in parenthesis are the 0.1 and 0.9 quantiles.

	1	2	3	4
MMM $\psi_1^{(1)}$	0.487(0.453;0.521)	0.491(0.458;0.523)	0.008(0.000;0.024)	0.013(0.000;0.037)
$1/2(\varphi_1^{(1)} + \varphi_2^{(1)})$	0.450	0.450	0.050	0.050
MixedMem $\psi_1^{(1)}$	0.504	0.000	0.000	0.496
$1/2(\varphi_1^{(1)} + \varphi_4^{(1)})$	0.450	0.050	0.050	0.450
MMM $\psi_2^{(1)}$	0.026(0.000;0.059)	0.011(0.000;0.032)	0.449(0.412;0.484)	0.515(0.478;0.550)
$1/2(\varphi_3^{(1)} + \varphi_4^{(1)})$	0.050	0.050	0.450	0.450
MixedMem $\psi_1^{(1)}$	0.000	0.533	0.467	0.000
$1/2(\varphi_2^{(1)} + \varphi_3^{(1)})$	00450	0.450	0.450	0.050

To additionally evaluate model performance, we compute the Frobenius norm between the 'true' probability tensor $\pi_0^{(1)} = \{\text{pr}(X_1 = x_1, \ldots, X_{p_1} = x_{p_1});$ for $x_j = 1, \ldots 4, j = 1, \ldots, p_1\}$, and $\pi_{\text{MM}}^{(1)}$ and $\pi_{\text{MMM}}^{(1)}$, denoting the estimates from the MMM and MM model, respectively. Leveraging equations (1.10) and (3.3), for MM we have a closed form expression of the core tensor, while for MMM we use 10^5 Monte Carlo replicates for estimation. To estimate uncertainty in the MM case, we rely on 1000 bootstrap replicates. We obtain a posterior mean of 0.131 for $||\pi_0^{(1)} - \pi_{\text{MMM}}^{(1)}||_F$ with a standard deviation of 0.048, and a bootstrap mean 0.133 for $||\pi_0^{(1)} - \pi_{\text{MM}}^{(1)}||_F$ with standard deviation 0.029. For comparison, we also estimate the Frobenius norm considering a correctly specified model with $H_1 = 4$ profiles for group 1, leading to a bootstrap mean of 0.089 with standard deviation of 0.032. This slightly improvement in estimation accuracy does not justify the greater complexity of interpretation in using the larger H_g value.

3.6 Application to malaria risk assessment

We apply the proposed approach to the survey described in Section 3.1, with some modification to accommodate the structure of the data, and to consider space and time evolution. The composition of the considered data is summarized in Table 3.3.

TABLE 3.3: Distribution of the number of subjects and variables included in the analysis.

Year	# Subjects	# Behavioral variables	# Environmental variables	# variables
1985	269	14	28	42
1986	575	16	24	40
1987	802	14	29	33
1995	1108	19	36	55
Total	2704	63	117	180

Missing data were considered informative for this application; hence we defined a missing category for each variable so that the missingness pattern can inform the analysis results.

To take into account space and time effects, we leverage the multivariate Gaussian model in step 5 of Algorithm 4 in Section 3.4. Different multivariate spatio-temporal models can be considered (see for example Banerjee *et al.*, 2014), and we rely on a separable model for time and space, assuming no interaction. This assumption leads to a simple and computationally efficient latent model, while accommodating non regular observations in space and time. Space-time dependence can be included in distribution (3.4) by letting

$$(\lambda_i^{(\mathrm{B})}, \lambda_i^{(\mathrm{E})})^T \quad \sim \quad \mathrm{MLND}(\boldsymbol{\beta}_{\boldsymbol{t_i}} + \boldsymbol{\zeta}_t(\boldsymbol{s}_i), \Sigma_t), \tag{3.6}$$

where $t_i \in \{1985, 1986, 1987, 1995\}$, and $\boldsymbol{s}_i = (s_{i1}, s_{i2})^T$, are respectively a time indicator and observed longitude and latitude for subject i.

We account for time dependence through a multivariate Gaussian hierarchical model with common hyperprior. Specifically, $\boldsymbol{\beta_t} = (\beta_t^{(\mathrm{B})}, \beta_t^{(\mathrm{E})}) \sim \mathcal{N}(\boldsymbol{\beta}, \Sigma)$, $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta_0}, \boldsymbol{\Sigma_0})$ and $\boldsymbol{\Sigma} \sim$

$\mathrm{IW}(\nu_\beta, \boldsymbol{\Psi}_{\boldsymbol{\beta}})$. This model does not impose a rigid time evolution, allowing borrowing of information across different years.

For the spatial effect $\zeta_t(\boldsymbol{s}_i)$ we specify a bivariate spatial model. A simple possibility would be to rely on a separable structure for the spatial cross covariance of the process (e.g. Banerjee *et al.*, 2000). However, this model would imply the same spatial effect for both the environmental and behavioral domain. We expect that behavioral and environmental profiles can have a very different spatial evolution, and for this reason we rely on a conditional Gaussian process $p(\boldsymbol{\zeta}|\boldsymbol{\Sigma}_t)$ for the components of $\zeta_t(\boldsymbol{s}_i) = (\zeta_t^{(\mathrm{B})}(\boldsymbol{s}_i), \zeta_t^{(\mathrm{E})}(\boldsymbol{s}_i))^T$. Specifically we consider $\boldsymbol{\Sigma_t} = \boldsymbol{L_t}\boldsymbol{L_t}^T$, where $\boldsymbol{L_t}$ is a lower triangular matrix obtained through Cholesky decomposition, and we let $\tilde{\zeta}_t(\boldsymbol{s}_i) = \boldsymbol{L}_t^{-1}\zeta_t(\boldsymbol{s}_i)$ and $\tilde{\zeta}_t^{(g)} \sim \mathrm{GP}(0, \boldsymbol{K}_t^{(g)})$. This formulation enforces no dependence across time and space for the spatial effect components.

Since we are considering standardized data, we parameterize the Gaussian processes in terms of correlation functions $K_t^{(g)}(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$, obtained by normalizing the square exponential form $\exp\{-1/2\sum_{d=1}^{2}\gamma_{td}^{(g)}\mathrm{d}^2(s_{id}, s_{i'd})\} + \tau 1\{i = i'\}$, where $\gamma_{td}^{(g)}$ are length scale parameters, $\mathrm{d}(\cdot,\cdot)$ is the Euclidean distance and τ is a nugget effect to limit numerical instability. The considered prior induces independent spatial effects for each domain and time, while leading to a computationally efficient model, since small matrices are involved in the Gaussian process computation.

The model can be easily implemented adapting Gibbs sampler Algorithm 4, with updating of the length scale parameters $\gamma_{td}^{(g)}$ relying on a Metropolis step.

3.6.1 Model checking

We assess goodness-of-fit of the assumed model to the observed data. We are particularly interested in whether the assumption of $H_g = 2$ leads to significant lack o fit.

One possibility is to compute posterior distributions for some statistics of the considered data and compare them with the corresponding empirical quantities (e.g. Gelman *et al.*, 2013). We consider as statistics the marginal and bivariate distributions, that can be obtained as:

$$\begin{aligned} \pi_{x_j}^{(j)} &= \mathrm{pr}(X_j = x_j \mid -) = \sum_{h=1}^{H} \bar{a}_h^{(g_j)} \psi_{h x_j}^{(j)}, \\ \pi_{x_j, x_k}^{(j,k)} &= \mathrm{pr}(X_j = x_j, X_k = x_k \mid -) = \sum_{h_j=1}^{H}\sum_{h_k=1}^{H} \bar{a}_{h_j h_k}^{(g_j, g_k)} \psi_{h_j x_j}^{(j)} \psi_{h_k x_k}^{(k)}, \end{aligned} \tag{3.7}$$

for $j = 1, \ldots, p$ and $k \neq j$, and where $\bar{a}_h^{(g_j)} = \mathbb{E}[\lambda_h^{(g_j)}]$ and $\bar{a}_{h_j h_k}^{(g_j, g_k)} = \mathbb{E}[\lambda_{h_j}^{(g_j)} \lambda_{h_k}^{(g_k)}]$.

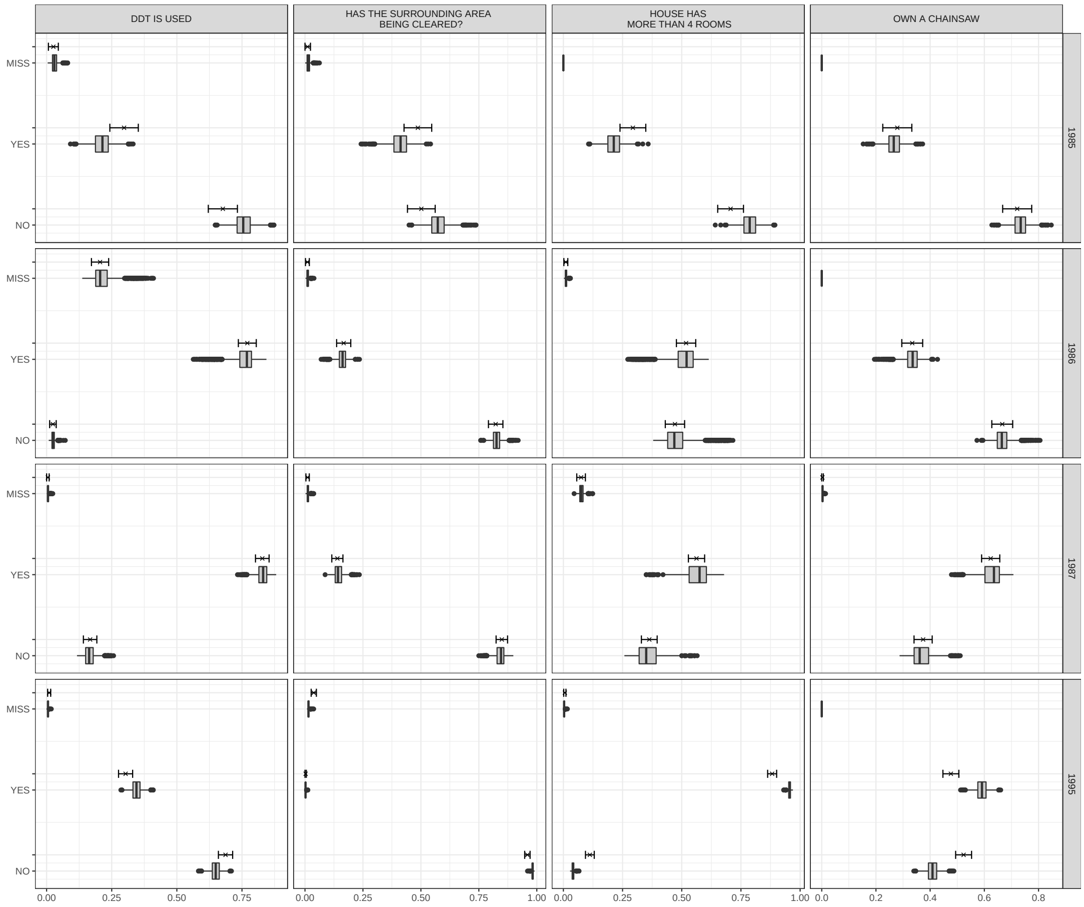
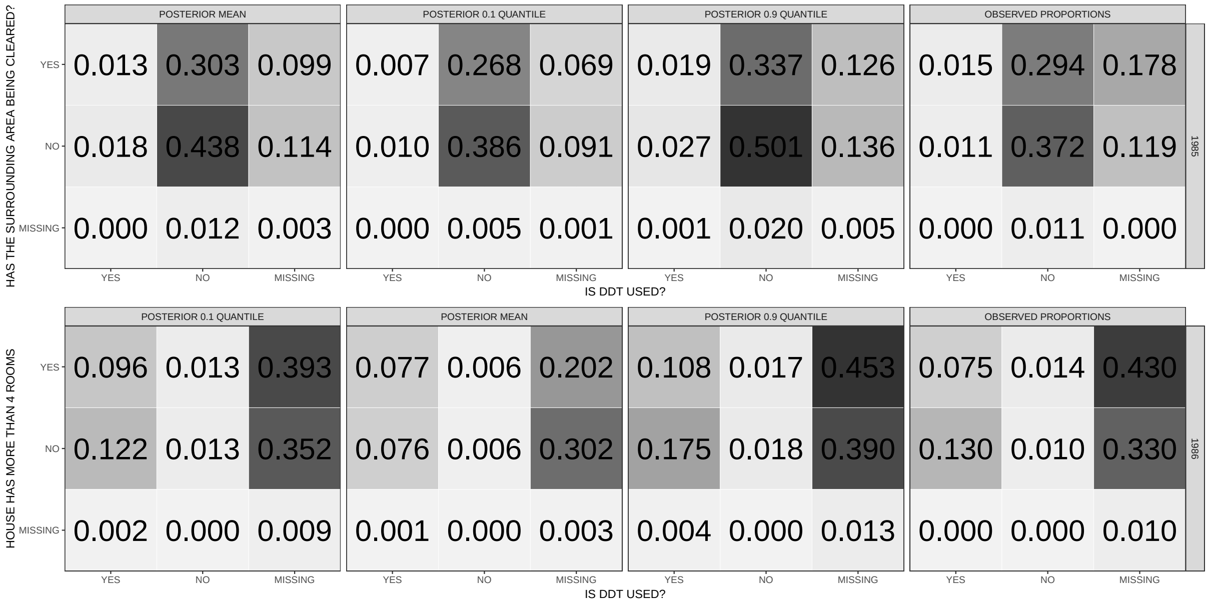


FIGURE 3.7: Box-plots of marginal distributions for 4 representative variables over the years. The 'x's and errors bands above the box-plots are the sample proportions and 0.95 Wald-type confidence intervals.

Figure 3.7 shows estimated marginal distributions from the proposed model obtained from 2500 Gibbs samples for 4 representative variables across the years, together with the sample proportion with the 0.95 level Wald type confidence interval. All the estimated marginals are compatible with the observed ones. We additionally estimated the posterior mean of the L1-norm between the empirical frequencies $\hat{f}_{jc} = n^{-1}\sum_{i=1}^{n} 1\{X_{ij} = c\}$ and the estimated ones $\hat{\pi}_{jc}$ obtained averaging 2500 MCMC samples of the expression in 3.7. The L1-norm has expression $\sum_{c=1}^{d_j} |\hat{\pi}_{jc} - \hat{f}_{jc}|$ and belongs to the interval $(0,2)$. Considering all the variables, we have an average L1-norm of about 0.071, and 0.1 and 0.9 quantiles of (0.024,0.14). These quantities also suggest a strong adherence of the estimated marginals with the empirical ones, comparing with the maximum attainable value of 2.

Figure 3.8 shows posterior distribution and quantiles for 2 bivariate distributions. Also in this case we observe a satisfactory adherence of the estimated quantities and the empirical ones.



HAS THE SURROUNDING AREA BEING CLEARED?	POSTERIOR MEAN			POSTERIOR 0.1 QUANTILE			POSTERIOR 0.9 QUANTILE			OBSERVED PROPORTIONS		
IS DDT USED?	YES	NO	MISSING	YES	NO	MISSING	YES	NO	MISSING	YES	NO	MISSING
YES	0.013	0.303	0.099	0.007	0.268	0.069	0.019	0.337	0.126	0.015	0.294	0.178
NO	0.018	0.438	0.114	0.010	0.386	0.091	0.027	0.501	0.136	0.011	0.372	0.119
MISSING	0.000	0.012	0.003	0.000	0.005	0.001	0.001	0.020	0.005	0.000	0.011	0.000

HOUSE HAS MORE THAN 4 ROOMS	POSTERIOR 0.1 QUANTILE			POSTERIOR MEAN			POSTERIOR 0.9 QUANTILE			OBSERVED PROPORTIONS		
IS DDT USED?	YES	NO	MISSING	YES	NO	MISSING	YES	NO	MISSING	YES	NO	MISSING
YES	0.096	0.013	0.393	0.077	0.006	0.202	0.108	0.017	0.453	0.075	0.014	0.430
NO	0.122	0.013	0.352	0.076	0.006	0.302	0.175	0.018	0.390	0.130	0.010	0.330
MISSING	0.002	0.000	0.009	0.001	0.000	0.003	0.004	0.000	0.013	0.000	0.000	0.010

FIGURE 3.8: Posterior mean and quantiles for 2 bivariate distributions, compared with the empirical frequencies in the data.

As with the marginals we compute the L1-norm between the empirical bivariate distributions and the estimated ones. We focus on pairs of variables in the same year. We obtain good results also in this case with a mean of 0.13 and 0.1 and 0.9 quantiles of 0.11 and 0.18.

As general tendency we observed that variables that are sensibly different across the estimated profiles are generally better reconstructed. These are also the most interesting from an interpretation point of view since they characterize the profiles.

3.6.2 Relation with malaria rate and prediction

Our estimated scores define simple and interpretable risk measures with respect to correlated latent traits for behavioral and environmental conditions at the household level. Being not directly based on malaria risk measures that are just used as external variables, the obtained scores do not suffer from the bias discussed in Section 3.1. We expect, however, that household classified as high behavioral and or environmental risk presentes in average higher malaria rates. We first inspect parameters trace plots, finding no evidence of label switching across MCMC iterations. Although our analysis is unsupervised we post process the classes so that the second pure profile is more reasonably viewed as corresponding to high risk. In this way we can consider $\lambda_i^{(B)}$ and $\lambda_i^{(E)}$ as high risk scores.

To asses the relationship between malaria risk and the estimated profiles, for each MCMC sample of each year, we assign households into low, moderate and high behavioral and environmental risk groups using tertiles, and we compute the average malaria rate for the obtained 9 groups. Posterior median and quantiles of the malaria rate distribution for the considered clusters are shown in Table 3.4. A similar strategy using tertiles of separate MM models has been used to determine risk profiles for Chagas disease-risk in northern Argentina (Chuit *et al.*, 2001). We expect the obtained distribution to be ordered in such a way that higher malaria rates corresponds to high environmental and behavioral risk clusters. Formally, considering a table as 3.4, having low risk clusters in the upper left corner, we expect the resulting table to be a double-gradient one, meaning that each row should be non decreasing from left-to-right and each column should be non decreasing from top-to-bottom.

TABLE 3.4: Median, 0.1 and 0.9 posterior quantiles of the malaria rates for low, median and high environmental and behavioral profiles. Groups are computed leveraging scores $\lambda_i^{(\mathrm{B})}$ and $\lambda_i^{(\mathrm{E})}$ tertiles. Values at bottom and right side of the table are obtained considering just behavioral or environmental scores respectively. Light grey values indicates violation in median of the double-gradient assumption, while dashes indicates that there are not enough subjects in the cluster to compute the median and the quantiles.

	1985			
	Low B	Med B	High B	
Low E	0.000(0.000;0.000)	0.045(0.000;0.197)	—	0.000(0.000;0.000)
Med E	0.067(0.000;0.310)	0.101(0.050;0.143)	0.115(0.000;0.250)	0.106(0.067;0.129)
High E	—	0.115(0.000;0.356)	0.117(0.100;0.125)	0.117(0.100;0.125)
	0.000(0.000;0.000)	0.091(0.050;0.125)	0.117(0.100;0.125)	

	1986			
	Low B	Mod B	High B	
Low E	0.228(0.206;0.250)	0.231(0.139;0.327)	—	0.228(0.206;0.250)
Mod E	0.238(0.167;0.299)	0.250(0.200;0.279)	0.232(0.000;0.658)	0.250(0.206;0.273)
High E	—	0.250(0.200;0.302)	0.250(0.250;0.286)	0.250(0.250;0.279)
	0.229(0.211;0.250)	0.250(0.217;0.275)	0.250(0.250;0.286)	

	1987			
	Low B	Mod B	High B	
Low E	0.167(0.133;0.180)	0.216(0.140;0.458)	—	0.167(0.133;0.180)
Mod E	0.177(0.146;0.207)	0.190(0.167;0.215)	0.180(0.082;0.243)	0.183(0.167;0.200)
High E	—	0.200(0.171;0.250)	0.201(0.193;0.227)	0.201(0.197;0.219)
	0.167(0.150;0.180)	0.197(0.181;0.209)	0.201(0.193;0.227)	

	1995			
	Low B	Mod B	High B	
Low E	0.028(0.021;0.042)	—	—	0.028(0.021;0.042)
Mod E	0.028(0.024;0.028)	—	—	0.028(0.024;0.028)
High E	0.028(0.024;0.033)	0.033(0.030;0.036)	0.036(0.029;0.042)	0.033(0.031;0.033)
	0.028(0.028;0.028)	0.033(0.030;0.036)	0.036(0.029;0.042)	

From Table 3.4 we notice that in defining groups with either environmental or behavioral scores we obtain groups sharing almost the same rate. However, a more detailed classification can be obtained leveraging both domains at the same time. In general, we observe higher malaria median rates in high behavioral and environmental risk zones, with only two violation of the expected double gradient assumption if we consider the median, in both cases, however upper quantiles are still increasing as expected.

We additionally consider an external validation of our model, checking its predictive ability with respect to the malaria rate. We rely on 5-fold cross validation, comparing the results with random forest and lasso prediction on the raw data, implemented in the R packages `randomForest` and `glmnet`, respectively.

Crude malaria rate and behavioral and environmental profiles measure different risks, hence although they are expected to be related, they might be not identical, nor expressed in the same scale. To overcome this issue, we rely on a simple model, assessing proportionality between malaria rate r_i, and behavioral and environmental risks. Formally we suppose:

$$r_i = c_1 \hat{\lambda}_i^{(\mathrm{B})} + c_2 \hat{\lambda}_i^{(\mathrm{E})},$$

where $\hat{\lambda}_i^{(\mathrm{B})}$ and $\hat{\lambda}_i^{(\mathrm{E})}$ are the posterior mean of the behavioral and environmental risks from our model, and we estimate c_1 and c_2 via least squares on a training sample. Clearly, more refined prediction models can be used, but simplicity and transparency are important in this application. To select tuning parameters and avoid over fitting we also include, for both lasso and random forest methods, a second cross validation layer for each fold. Moreover, since the response variable r_i belongs to the $(0,1)$ interval, we build all models for logit(r_i) mapping the prediction back via inverse transformation. Model performances are evaluated using Mean Square Error (MSE). Results are shown in Table 3.5.

TABLE 3.5: 5 fold cross validation mean square prediction error.

	1985	1986	1987	1995
Random Forest	0.104(0.026)	0.131(0.023)	0.108(0.011)	0.021(0.006)
lasso regression	0.104(0.026)	0.169(0.017)	0.123(0.005)	0.021(0.006)
MMM	0.099(0.024)	0.148(0.015)	0.105(0.003)	0.021(0.006)

Based on Table 3.5, our multivariate mixed membership model has comparable out-of-sample predictive performance to black box machine learning algorithm such as random forests. The ML algorithms and other supervised approaches are subject to the selection bias issue mentioned in Section 3.1. This result, combined with the goodness-of-fit assessment of Section 3.6.1, gives evidence that a model with 2 profiles for each group provides a reasonable low dimensional representation in our contest.

3.6.3 Time and space domain evolution

Leveraging equation (3.5) we can compute the distribution of the odds ratios of being high risk in behavioral and environmental domains. Specifically, for each iteration of the MCMC algorithm, we can compute the quantity $\exp\{\beta_t^{(\mathrm{B})} - \beta_t^{(\mathrm{E})} + 1/2(\Sigma_{t11} + \Sigma_{t22} - 2\Sigma_{t12})\}$, where $(\Sigma_{t11}, \Sigma_{t22}, \Sigma_{t12})$ are the elements of the covariance matrix $\boldsymbol{\Sigma}_t$ in equation (3.6). Boxplots summarizing their distribution are shown in Figure 3.9.

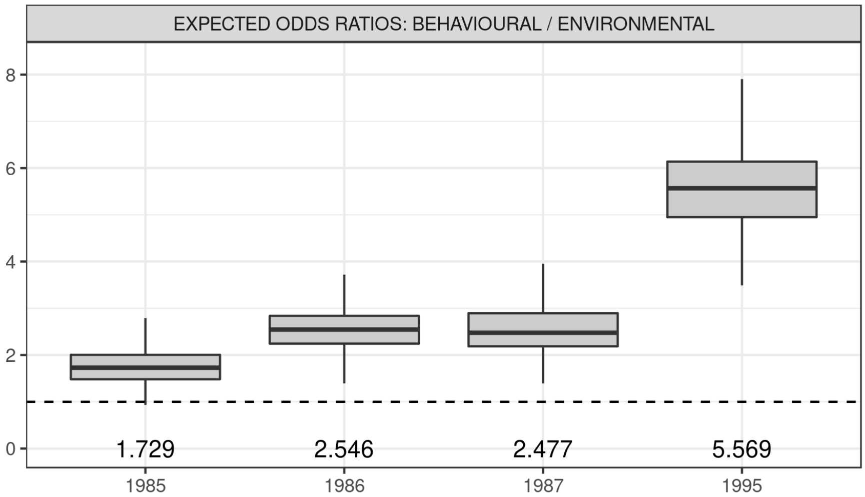


FIGURE 3.9: Odds ratios between environmental and behavioral risk scores. Numbers at the bottom of the plots are the value of the median.

We notice an increasing trend in the odds ratios across the years; specifically, in 1985 environmental and behavioral risks coexist, while starting from 1986 behavioral risk starts to gain more and more importance, as is evident from the fact that the posterior odds ratio is not significantly above one in 1985 and then it gradually increases. These results are in accordance with current literature on malaria risk, reporting that the risk is initially driven by favorable environmental conditions for malaria vectors to proliferate (e.g. Castro *et al.*, 2006). Soon after human settlement, there is a phase lasting for about 8 or 10 years, in which environmental risk is high but human behavior is starting to gradually become the predominant risk factor. In the last stage, called the endemic phase, the risk is far more related to behavioral causes.

From a spatial perspective, we can consider the posterior predictive distribution of the ζ_t evaluated over a regular grid of values (Figure 3.10). We notice that the behavioral risk distribution is constant both across time and space; hence the spatial variability is driven by environmental conditions. Environmental risk zones can be mostly explained in term of geographical characteristics of the area; in fact higher risk zones correspond to the forest fringe and the Machadinho river path.

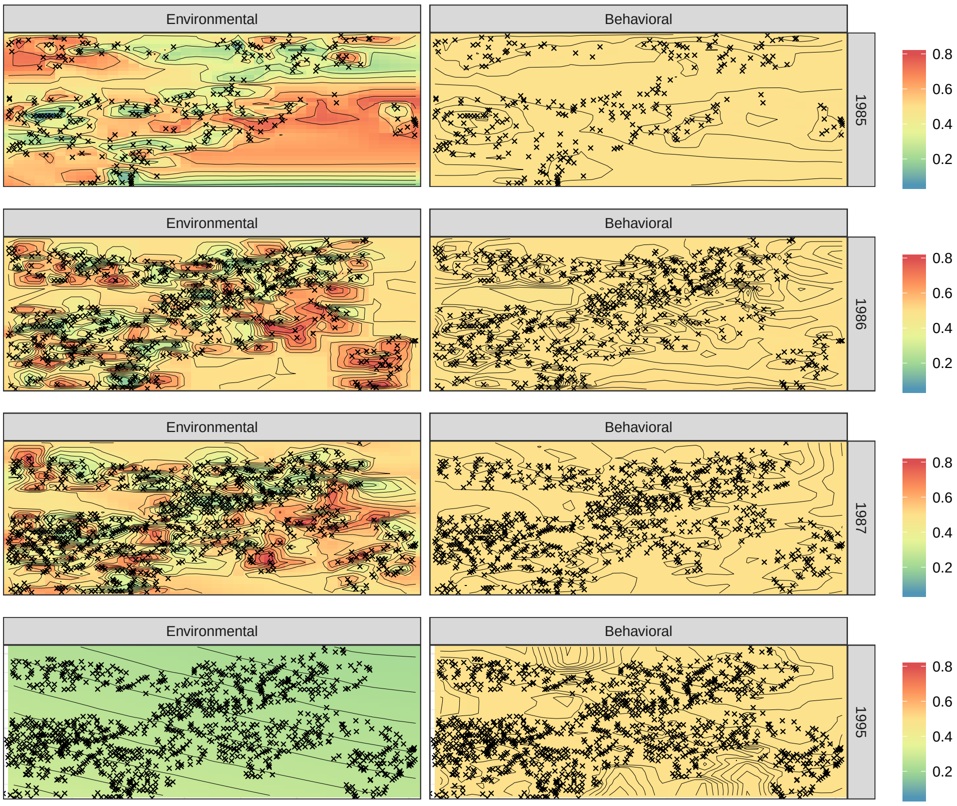


FIGURE 3.10: Spatial risk predictions for the Machadinho area, for both behavioral and environmental domain. Values are expressed on the probability scale.

3.6.4 Risk conditions

Studying the distribution of $\boldsymbol{\psi}_h^{(j)}$, we can have a general idea of behavioral and environmental risk profiles, their evolution in time and their similarity with malaria literature. Some conditions are indeed very stable while others can assume different meaning in different time periods. To select the most relevant variables composing the profiles, we rely on the notion of *admissibility* (Singer, 1989). The basic idea is to define a level of variable admissible for a given profiles, if such level occurs with probability close to 1 for the given profile. Such definition is however too restrictive to be applied in practice and can be replaced by considering as admissible a condition whose probability is substantially larger that the observed one. In the proposed Bayesian framework we can easily compute the posterior probability of being admissible as $\text{pr}[(\psi_{hc}^{(j)} - \hat{f}_c^{(j)})/\hat{f}_c^{(j)} > K]$, where $\hat{f}_c^{(j)}$ is the empirical frequency. We use as threshold $K = 0.70$ if $\hat{f}_c^{(j)} < 0.50$ and $K = 0.35$ otherwise, defining as admissible the conditions which posterior probability is greater than 0.50. Estimated posterior medians and 90% credible intervals of these parameters are shown in the Appendix 3.6.4. Variables have been divided into environmental and behavioral domains, and the order is such that the variables whose L1 norms across the profiles differs the most are on top of the tables, while admissible profiles are highlighted in grey.

As a general tendency, we find stable results in the first 3 years, while variation occurs in 1995. This is not surprising given the number of years gap, the general evolution of the settlement project and of malaria diffusion, which is starting its endemic phase in the last year. Generally we find environmental conditions to be more stable across time, especially variables concerning house conditions, distance from a water source and forest and topography of the area. High risk environmental profiles are well characterized by house conditions and composition, i.e. small houses with bad condition of walls and roof. This effect of housing condition on malaria risk infection has been considered in very few studies, a recent exception is Dlamini *et al.* (2017).

In 1985 the high risk group includes settlers who just arrived in Machadinho not owning proper tools for plantation and land clearance. This might reflect the fact that, at the beginning of the project, many settlers lacked the capacity to clear large areas thereby leaving them in close proximity to A. Darlingi habitats for many months. In 1986 and 1987 low risk profiles reflect the work done to clear the land and make the place livable and inhospitable for the malaria vector. These households are characterized by the presence of cultivations, use of DDT, and the owning of certain goods as planters and chainsaws. In 1995 we observe the opposite behavior, in fact planted and cleared areas are considered high risk zones. This might have been caused by different factors, including deforestation and changes in the plant life. Canonical epidemiological understanding of the deforestation effect is, in fact, that it increases malaria risk in Africa and the Americas and diminishes it in Southeast Asia (e.g. Guerra *et al.*, 2006).

The plantation effect has also been highlighted in several studies. For example, the negative effect of rice on malaria diffusion in Africa was noted by Ijumba and Lindsay (2001) who reported that malaria prevalence was four times lower in children living near irrigated rice, and that in Gambia there was less malaria near the rice fields than in other rural communities. We encountered a similar behavior in 1986, while in 1995 rice presence seems more related to high risk conditions. The reason for such a change, is not clear, but may be the ecological transformation occurring.

Another relevant behavioral condition is the time the settler has spent in the current house. Specifically, newcomers have been included in the high risk group, which is also coherent with malaria development, in fact, a stable population is more likely to develop premunition that protect them from successive infections (e.g. Bereczky *et al.*, 2007; Males *et al.*, 2008).

Although closeness to planted areas is a relevant risk condition, somehow surprisingly closeness to pasture areas does not seem to impact risk. It is indeed expected that the presence of live stock can influence malaria transmission. Presence of animals may divert the blood-seeking mosquito vectors from humans, thereby decreasing the biting on people, on the other hand, livestock can provide an additional blood-source, which can increase vector survival and density (e.g. Franco *et al.*, 2014).

Discussion and future directions of research

In the context of tensor factorization models, this Thesis introduced two novel factorizations dealing with multivariate categorical variables, with main goal being in defining parsimonious representations, in terms of latent space. In Chapter 2, we propose a class of grouped tensor factorization, that generalizes popular Parafac decomposition, by allowing tensor arm to have dimension greater than one. To learn the unknown arm structure directly from the data, we leverage a product partition model prior, having an *ad hoc* cohesion function that enhances computational tractability. We show that the proposed factorization retains flexibility of the standard Parafac, while leading to a more effective characterization of the dependence structure, especially in applications with low sample size. We applied the proposed factorization to promoter DNA sequence data, showing how the model is able to identify potential starting point for the transcription.

In Chapter 3, we introduced a novel multivariate generalization of mixture membership models to deal with domain specific membership scores. We showed that when fixing the number of profiles we gain a better characterization of the probability mass function in linking different MM models, if compared to the standard approaches. Empirical finding highlighted robustness respect to misspecification of the profile distribution, showing that kernels estimates do not drastically change under different scenarios. When a higher number of profiles would be needed to explaining the phenomenon under exam, the use of $H = 2$, in many application, might still be a valid approximation leading to important finding, provided that coherence with the observed data is carefully checked.

We applied the proposed approach to survey data collected during the Machadinho settlement project in Brazil, showing that the obtained profiles retain important information, and can accurately predict malaria rate. We also highlighted that profiles identified as low/high risks, as well as their space and time evolution, match several epidemiological studies. The proposed approach is however much more general and can be applied in several contexts in which variable domain knowledge is available.

A promising direction for both models would be to extend the proposed unsupervised approaches to the case of density regression with many categorical predictors. In this framework, a nice adaptation of Tucker decomposition, in the case of classification, is provided by Yang and Dunson (2016). Their approach works amazingly well in setting where just a sparse subset of the predictors have effect on the response variable. Exploiting the group structure of our GROT and/or MMM models, however, we can efficiently learn the effect of separate groups of categorical predictors and their interactions on the response variable.

Another important direction is in developing new efficient algorithms for posterior inference, to substitute MCMC steps. This would allow application of the proposed methodologies to high and ultra-high dimensional tensor data, provided, for example, by new DNA sequences technologies and neuroimaging data. Possible solutions are to adapt online variational methods (Hoffman *et al.*, 2010), or a generalized method of moments, as the one provided for Dirichlet latent variable models in Zhao *et al.* (2018).

An interesting extension of GROT model would be in replacing Dirichlet-multinomial kernels for tensor arms with more parsimonious representations, in order to accommodate sparse group structures. For example, one can reparametrize tensor arms in the corresponding log-linear model form, and includes a subset of the interactions up to the second or third order (e.g. Massam *et al.*, 2009).

A sensible generalization of MMM model would be in including continuous and/or mixed type variables, while retaining a simple representation in the profile space, this can be accomplished recasting the proposed framework in the alternative class of Partial Membership Models (e.g. Heller *et al.*, 2008).

Malaria risk conditions

TABLE A.1: 1985: posterior median and 0.1 and 0.9 quantiles for behavioral and environmental variables.

Behavioral	$\psi_1^{(j)}$	$\psi_2^{(j)}$
Plant Cassava: NO	0.042(0.006;0.107)	0.990(0.965;0.999)
Plant Cassava: YES	0.933(0.869;0.972)	0.003(0.000;0.025)
Plant Cassava: MISSING	0.021(0.004;0.045)	0.003(0.000;0.017)
Lavoura branca: NO	0.105(0.042;0.178)	0.993(0.977;0.999)
Lavoura branca: YES	0.888(0.814;0.950)	0.002(0.000;0.014)
Lavoura branca: MISSING	0.005(0.000;0.021)	0.002(0.000;0.013)
DDT is used: NO	0.296(0.211;0.383)	0.952(0.910;0.976)
DDT is used: YES	0.687(0.598;0.774)	0.007(0.000;0.049)
DDT is used: MISSING	0.013(0.002;0.040)	0.035(0.016;0.060)
Plan to build a new house within a year: NO	0.562(0.470;0.658)	0.006(0.000;0.040)
Plan to build a new house within a year: YES	0.343(0.246;0.439)	0.976(0.936;0.997)
Plan to build a new house within a year: MISSING	0.092(0.052;0.141)	0.011(0.000;0.038)
Arrived in Machadino before 1985: NO	0.655(0.564;0.745)	0.034(0.003;0.093)
Arrived in Machadino before 1985: YES	0.345(0.255;0.436)	0.966(0.907;0.997)
Own a planter: NO	0.138(0.059;0.230)	0.743(0.674;0.810)
Own a planter: YES	0.862(0.770;0.941)	0.257(0.190;0.326)
Do you own other proprieties: NO	0.316(0.213;0.411)	0.761(0.694;0.827)
Do you own other proprieties: YES	0.684(0.589;0.787)	0.239(0.173;0.306)
Lived in current house for more that 1m: NO	0.580(0.502;0.653)	0.986(0.963;0.997)
Lived in current house for more that 1m: YES	0.412(0.339;0.488)	0.002(0.000;0.020)
Lived in current house for more that 1m: MISSING	0.003(0.000;0.024)	0.009(0.001;0.024)
Knowledge of malaria vector: NO	0.726(0.627;0.819)	0.330(0.261;0.401)
Knowledge of malaria vector: YES	0.231(0.147;0.330)	0.495(0.425;0.569)
Knowledge of malaria vector: MISSING	0.029(0.000;0.096)	0.172(0.125;0.223)
Plant cocoa: NO	0.636(0.559;0.702)	0.998(0.987;1.000)
Plant cocoa: YES	0.364(0.298;0.441)	0.002(0.000;0.013)
Own more the 4 goods: NO	0.153(0.072;0.242)	0.514(0.445;0.589)
Own more the 4 goods: YES	0.847(0.758;0.928)	0.486(0.411;0.555)
Plant coffee: NO	0.662(0.585;0.728)	0.993(0.979;0.999)
Plant coffee: YES	0.332(0.267;0.407)	0.001(0.000;0.009)
Plant coffee: MISSING	0.002(0.000;0.017)	0.004(0.000;0.014)
Do you often go to surrounding cities: NO	0.003(0.000;0.035)	0.328(0.274;0.383)
Do you often go to surrounding cities: YES	0.975(0.940;0.992)	0.669(0.613;0.723)
Do you often go to surrounding cities: MISSING	0.017(0.004;0.038)	0.001(0.000;0.009)
HH has high level of education: NO	0.682(0.581;0.784)	0.411(0.338;0.487)
HH has high level of education: YES	0.302(0.199;0.403)	0.558(0.480;0.630)
HH has high level of education: MISSING	0.010(0.000;0.042)	0.029(0.010;0.053)
Own chickens and/or porks: NO	0.681(0.590;0.765)	0.929(0.876;0.981)
Own chickens and/or porks: YES	0.311(0.227;0.401)	0.067(0.015;0.120)
Own chickens and/or porks: MISSING	0.005(0.000;0.021)	0.002(0.000;0.012)
HH wife has high level of education: < 4 yr	0.551(0.451;0.655)	0.452(0.380;0.524)
HH wife has high level of education: > 4 yr	0.279(0.190;0.375)	0.476(0.407;0.549)
HH wife has high level of education: NO-WIFE	0.013(0.000;0.062)	0.036(0.006;0.065)
HH wife has high level of education: MISSING	0.141(0.075;0.208)	0.028(0.000;0.078)
More than 4 people in the house: NO	0.652(0.545;0.753)	0.418(0.344;0.491)
More than 4 people in the house: YES	0.348(0.247;0.455)	0.582(0.509;0.656)
Spray insecticide: NO	0.647(0.553;0.738)	0.841(0.779;0.897)
Spray insecticide: YES	0.353(0.262;0.447)	0.159(0.103;0.221)
Get malaria from dirty water: NO	0.333(0.240;0.430)	0.472(0.399;0.546)
Get malaria from dirty water: YES	0.590(0.496;0.689)	0.427(0.351;0.497)
Get malaria from dirty water: MISSING	0.070(0.031;0.124)	0.100(0.064;0.143)
Use plant to cure malaria: NO	0.700(0.604;0.794)	0.634(0.564;0.705)
Use plant to cure malaria: YES	0.088(0.013;0.171)	0.235(0.175;0.302)
Use plant to cure malaria: MISSING	0.206(0.138;0.287)	0.127(0.081;0.180)
Own a chainsaw: NO	0.650(0.558;0.736)	0.770(0.705;0.828)
Own a chainsaw: YES	0.350(0.264;0.442)	0.230(0.172;0.295)
Use a bednet: NO	0.847(0.758;0.934)	0.762(0.695;0.823)
Use a bednet: YES	0.136(0.048;0.222)	0.220(0.162;0.286)
Use a bednet: MISSING	0.014(0.000;0.044)	0.015(0.001;0.034)
Use repellent: NO	0.984(0.940;0.999)	0.862(0.820;0.900)
Use repellent: YES	0.016(0.001;0.060)	0.138(0.100;0.180)
Part of family did not come: NO	0.653(0.561;0.752)	0.640(0.568;0.711)
Part of family did not come: YES	0.317(0.222;0.409)	0.351(0.280;0.422)
Part of family did not come: MISSING	0.026(0.002;0.055)	0.004(0.000;0.027)
Do you ever go to urban area?: NO	0.003(0.000;0.027)	0.090(0.062;0.124)
Do you ever go to urban area?: YES	0.979(0.949;0.995)	0.905(0.869;0.934)
Do you ever go to urban area?: MISSING	0.013(0.001;0.034)	0.002(0.000;0.015)
Arrived in Rondonia before 1985: NO	0.943(0.897;0.984)	0.975(0.940;0.994)
Arrived in Rondonia before 1985: YES	0.002(0.000;0.013)	0.005(0.000;0.016)
Arrived in Rondonia before 1985: MISSING	0.053(0.013;0.097)	0.017(0.001;0.051)
Before coming was your occupation rural: NO	0.001(0.000;0.008)	0.001(0.000;0.006)
Before coming was your occupation rural: YES	0.994(0.975;1.000)	0.982(0.963;0.994)
Before coming was your occupation rural: MISSING	0.003(0.000;0.020)	0.016(0.005;0.034)
Are there rubber tree: NO	0.982(0.960;0.995)	0.998(0.988;1.000)
Are there rubber tree: YES	0.018(0.005;0.040)	0.002(0.000;0.012)

Environmental	$\psi_1^{(j)}$	$\psi_2^{(j)}$
Roof has good quality: NO	0.135(0.031;0.273)	0.982(0.947;0.996)
Roof has good quality: YES	0.856(0.720;0.960)	0.005(0.000;0.040)
Roof has good quality: MISSING	0.003(0.000;0.022)	0.009(0.001;0.023)
Walls have good quality: NO	0.207(0.080;0.315)	0.983(0.952;0.996)
Walls have good quality: YES	0.785(0.678;0.911)	0.004(0.000;0.034)
Walls have good quality: MISSING	0.003(0.000;0.022)	0.009(0.001;0.023)
House has more than 4 rooms: NO	0.235(0.118;0.350)	0.987(0.949;0.999)
House has more than 4 rooms: YES	0.765(0.650;0.882)	0.013(0.001;0.051)
Has the surrounding area being cleared: NO	0.109(0.018;0.213)	0.736(0.674;0.801)
Has the surrounding area being cleared: YES	0.885(0.782;0.972)	0.246(0.180;0.308)
Has the surrounding area being cleared: MISSING	0.002(0.000;0.017)	0.016(0.006;0.033)
Good water source available: NO	0.422(0.316;0.528)	0.886(0.812;0.961)
Good water source available: YES	0.571(0.463;0.676)	0.103(0.026;0.177)
Good water source available: MISSING	0.003(0.000;0.022)	0.009(0.001;0.024)
Do you have close neighbours (<500mt): NO	0.620(0.510;0.730)	0.203(0.136;0.273)
Do you have close neighbours (<500mt): YES	0.380(0.270;0.490)	0.797(0.727;0.864)
More that 100mt from a forest: NO	0.412(0.317;0.518)	0.755(0.688;0.817)
More that 100mt from a forest: YES	0.582(0.474;0.676)	0.239(0.178;0.305)
More that 100mt from a forest: MISSING	0.003(0.000;0.017)	0.004(0.000;0.014)
Distant from stagnant water(no culvert): NO	0.010(0.000;0.059)	0.187(0.143;0.236)
Distant from stagnant water(no culvert): YES	0.982(0.934;0.999)	0.775(0.724;0.825)
Distant from stagnant water(no culvert): MISSING	0.002(0.000;0.017)	0.035(0.018;0.057)
More than 600mt from a culvert: NO	0.033(0.001;0.098)	0.167(0.122;0.218)
More than 600mt from a culvert: YES	0.959(0.892;0.995)	0.791(0.736;0.841)
More than 600mt from a culvert: MISSING	0.003(0.000;0.022)	0.040(0.022;0.065)
More than 600mt from a river: NO	0.498(0.393;0.606)	0.571(0.496;0.642)
More than 600mt from a river: YES	0.492(0.384;0.595)	0.420(0.349;0.496)
More than 600mt from a river: MISSING	0.006(0.000;0.030)	0.007(0.000;0.022)
Anybody cleared the area before HH: NO	0.981(0.946;0.997)	0.878(0.838;0.912)
Anybody cleared the area before HH: YES	0.006(0.000;0.040)	0.119(0.085;0.158)
Anybody cleared the area before HH: MISSING	0.007(0.000;0.025)	0.001(0.000;0.011)
More that 10km from an hospital: NO	0.055(0.005;0.134)	0.151(0.101;0.201)
More that 10km from an hospital: YES	0.945(0.866;0.995)	0.849(0.799;0.899)
Sealing has good quality: NO	0.910(0.859;0.944)	0.987(0.970;0.997)
Sealing has good quality: YES	0.083(0.052;0.130)	0.001(0.000;0.011)
Sealing has good quality: MISSING	0.003(0.000;0.023)	0.009(0.001;0.024)
Good bathing place is available: NO	0.979(0.940;0.998)	0.974(0.950;0.992)
Good bathing place is available: YES	0.012(0.000;0.049)	0.014(0.001;0.034)
Good bathing place is available: MISSING	0.003(0.000;0.022)	0.010(0.001;0.024)

TABLE A.2: 1986: posterior median and 0.1 and 0.9 quantiles for behavioral and environmental variables.

Behavioral	$\psi_1^{(j)}$	$\psi_2^{(j)}$
Plant coffee: NO	0.115(0.068;0.161)	0.957(0.915;0.977)
Plant coffee: YES	0.882(0.837;0.930)	0.005(0.000;0.052)
Plant coffee: MISSING	0.001(0.000;0.006)	0.031(0.017;0.050)
Cultivate rice: NO	0.007(0.000;0.041)	0.743(0.661;0.835)
Cultivate rice: YES	0.988(0.956;0.999)	0.213(0.115;0.294)
Cultivate rice: MISSING	0.001(0.000;0.009)	0.044(0.026;0.068)
Own a planter: NO	0.060(0.008;0.111)	0.710(0.624;0.805)
Own a planter: YES	0.940(0.889;0.992)	0.290(0.195;0.376)
Own chickens and/or porks: NO	0.001(0.000;0.012)	0.566(0.498;0.635)
Own chickens and/or porks: YES	0.997(0.985;1.000)	0.408(0.338;0.479)
Own chickens and/or porks: MISSING	0.001(0.000;0.006)	0.023(0.011;0.041)
DDT is used: NO	0.024(0.005;0.042)	0.022(0.001;0.061)
DDT is used: YES	0.969(0.948;0.991)	0.399(0.321;0.467)
DDT is used: MISSING	0.003(0.000;0.020)	0.574(0.508;0.650)
More than 4 people in the house: NO	0.723(0.671;0.772)	0.148(0.062;0.239)
More than 4 people in the house: YES	0.277(0.228;0.329)	0.852(0.761;0.938)
Own more the 4 goods: NO	0.032(0.002;0.078)	0.601(0.522;0.693)
Own more the 4 goods: YES	0.968(0.922;0.998)	0.399(0.307;0.478)
Plant cocoa: NO	0.424(0.379;0.468)	0.971(0.930;0.989)
Plant cocoa: YES	0.573(0.529;0.619)	0.006(0.000;0.052)
Plant cocoa: MISSING	0.001(0.000;0.007)	0.017(0.006;0.033)
Are there rubber tree: NO	0.528(0.485;0.570)	0.980(0.960;0.993)
Are there rubber tree: YES	0.469(0.428;0.512)	0.001(0.000;0.015)
Are there rubber tree: MISSING	0.001(0.000;0.008)	0.016(0.005;0.031)
Before coming was your occupation rural: NO	0.820(0.770;0.871)	0.364(0.283;0.449)
Before coming was your occupation rural: YES	0.176(0.127;0.227)	0.633(0.548;0.714)
Before coming was your occupation rural: MISSING	0.002(0.000;0.007)	0.001(0.000;0.008)
Do you own other proprieties: NO	0.401(0.350;0.455)	0.851(0.773;0.926)
Do you own other proprieties: YES	0.599(0.545;0.650)	0.149(0.074;0.227)
Lived in current house for more that 1m: NO	0.524(0.478;0.571)	0.931(0.863;0.964)
Lived in current house for more that 1m: YES	0.473(0.427;0.519)	0.024(0.000;0.095)
Lived in current house for more that 1m: MISSING	0.001(0.000;0.007)	0.041(0.025;0.063)
Plan to build a new house within a year: NO	0.689(0.635;0.740)	0.267(0.183;0.354)
Plan to build a new house within a year: YES	0.268(0.219;0.320)	0.627(0.540;0.716)
Plan to build a new house within a year: MISSING	0.041(0.010;0.074)	0.103(0.053;0.162)
HH wife has high level of education: < 4 yr	0.525(0.474;0.575)	0.235(0.150;0.316)
HH wife has high level of education: > 4 yr	0.454(0.405;0.506)	0.348(0.264;0.429)
HH wife has high level of education: NO-WIFE	0.010(0.000;0.022)	0.008(0.000;0.034)
HH wife has high level of education: MISSING	0.003(0.000;0.029)	0.406(0.343;0.471)
Working in the plot from more than 1 month: NO	0.001(0.000;0.014)	0.392(0.335;0.453)
Working in the plot from more than 1 month: YES	0.966(0.945;0.983)	0.568(0.504;0.630)
Working in the plot from more than 1 month: MISSING	0.030(0.014;0.048)	0.037(0.012;0.072)
Part of family did not come: NO	0.750(0.698;0.799)	0.348(0.265;0.432)
Part of family did not come: YES	0.247(0.199;0.299)	0.570(0.487;0.654)
Part of family did not come: MISSING	0.001(0.000;0.007)	0.081(0.057;0.109)
Own a chainsaw: NO	0.540(0.488;0.594)	0.888(0.813;0.963)
Own a chainsaw: YES	0.460(0.406;0.512)	0.112(0.037;0.187)
Spray insecticide: NO	0.549(0.496;0.601)	0.795(0.709;0.874)
Spray insecticide: YES	0.444(0.391;0.496)	0.192(0.109;0.279)
Spray insecticide: MISSING	0.007(0.000;0.017)	0.011(0.001;0.029)
Knowledge of malaria vector: NO	0.472(0.421;0.523)	0.386(0.305;0.469)
Knowledge of malaria vector: YES	0.298(0.251;0.346)	0.406(0.329;0.488)
Knowledge of malaria vector: MISSING	0.229(0.186;0.273)	0.206(0.140;0.275)
Arrived in Machadino before 1985: NO	0.228(0.191;0.270)	0.085(0.032;0.147)
Arrived in Machadino before 1985: YES	0.772(0.730;0.809)	0.915(0.853;0.968)
Get malaria from dirty water: NO	0.417(0.369;0.467)	0.332(0.252;0.409)
Get malaria from dirty water: YES	0.535(0.486;0.583)	0.639(0.562;0.721)
Get malaria from dirty water: MISSING	0.047(0.027;0.068)	0.025(0.001;0.058)
Arrived in Rondonia before 1985: NO	0.860(0.833;0.889)	0.884(0.838;0.918)
Arrived in Rondonia before 1985: YES	0.001(0.000;0.006)	0.096(0.070;0.128)
Arrived in Rondonia before 1985: MISSING	0.137(0.110;0.165)	0.014(0.000;0.057)
Use plant to cure malaria: NO	0.814(0.770;0.858)	0.918(0.847;0.989)
Use plant to cure malaria: YES	0.177(0.134;0.219)	0.076(0.006;0.147)
Use plant to cure malaria: MISSING	0.009(0.002;0.017)	0.002(0.000;0.015)
HH has high level of education: NO	0.522(0.472;0.570)	0.428(0.350;0.505)
HH has high level of education: YES	0.473(0.424;0.523)	0.559(0.481;0.636)
HH has high level of education: MISSING	0.003(0.000;0.014)	0.011(0.000;0.028)

Environmental	$\psi_1^{(j)}$	$\psi_2^{(j)}$
House has more than 4 rooms: NO	0.043(0.001;0.097)	0.991(0.970;0.999)
House has more than 4 rooms: YES	0.940(0.888;0.983)	0.002(0.000;0.022)
House has more than 4 rooms: MISSING	0.015(0.006;0.027)	0.003(0.000;0.016)
Walls have good quality: NO	0.003(0.000;0.015)	0.891(0.818;0.964)
Walls have good quality: YES	0.997(0.985;1.000)	0.109(0.036;0.182)
Roof has good quality: NO	0.020(0.001;0.071)	0.763(0.696;0.829)
Roof has good quality: YES	0.980(0.929;0.999)	0.237(0.171;0.304)
Good water source available: NO	0.183(0.126;0.242)	0.864(0.778;0.946)
Good water source available: YES	0.815(0.755;0.872)	0.132(0.050;0.218)
Good water source available: MISSING	0.001(0.000;0.006)	0.002(0.000;0.009)
Anybody cleared the area before HH: NO	0.982(0.945;0.997)	0.498(0.441;0.558)
Anybody cleared the area before HH: YES	0.009(0.000;0.048)	0.484(0.426;0.541)
Anybody cleared the area before HH: MISSING	0.005(0.000;0.017)	0.016(0.003;0.033)
Do you have close neighbours (<500mt): NO	0.630(0.567;0.692)	0.301(0.220;0.384)
Do you have close neighbours (<500mt): YES	0.370(0.308;0.433)	0.699(0.616;0.780)
Far from permanent water: NO	0.439(0.376;0.496)	0.735(0.663;0.812)
Far from permanent water: YES	0.544(0.488;0.607)	0.261(0.184;0.333)
Far from permanent water: MISSING	0.016(0.007;0.028)	0.001(0.000;0.013)
More that 100mt from a forest: NO	0.691(0.650;0.732)	0.977(0.932;0.997)
More that 100mt from a forest: YES	0.307(0.266;0.347)	0.020(0.000;0.064)
More that 100mt from a forest: MISSING	0.001(0.000;0.007)	0.002(0.000;0.009)
Is topography bottom: NO	0.659(0.596;0.717)	0.406(0.323;0.482)
Is topography bottom: YES	0.335(0.277;0.399)	0.566(0.490;0.647)
Is topography bottom: MISSING	0.003(0.000;0.014)	0.027(0.013;0.044)
Has the surrounding area being cleared: NO	0.722(0.679;0.767)	0.956(0.905;0.981)
Has the surrounding area being cleared: YES	0.276(0.231;0.319)	0.019(0.000;0.071)
Has the surrounding area being cleared: MISSING	0.001(0.000;0.008)	0.022(0.011;0.037)
Near big planted area: NO	0.832(0.799;0.864)	0.957(0.927;0.978)
Near big planted area: YES	0.161(0.132;0.193)	0.006(0.000;0.033)
Near big planted area: MISSING	0.003(0.000;0.019)	0.033(0.015;0.053)
Sealing has good quality: NO	0.879(0.854;0.903)	0.995(0.980;1.000)
Sealing has good quality: YES	0.114(0.092;0.139)	0.002(0.000;0.015)
Sealing has good quality: MISSING	0.005(0.000;0.012)	0.001(0.000;0.009)
More that 10km from an hospital: NO	0.149(0.110;0.189)	0.045(0.008;0.093)
More that 10km from an hospital: YES	0.851(0.811;0.890)	0.955(0.907;0.992)
Good bathing place is available: NO	0.871(0.838;0.903)	0.967(0.929;0.992)
Good bathing place is available: YES	0.127(0.095;0.160)	0.024(0.001;0.061)
Good bathing place is available: MISSING	0.001(0.000;0.005)	0.007(0.002;0.017)
Far from temporary water: NO	0.903(0.868;0.933)	0.943(0.903;0.977)
Far from temporary water: YES	0.087(0.059;0.118)	0.031(0.004;0.069)
Far from temporary water: MISSING	0.009(0.000;0.027)	0.023(0.003;0.044)
Near big pasture area: NO	0.994(0.979;0.999)	0.963(0.944;0.980)
Near big pasture area: YES	0.002(0.000;0.009)	0.005(0.000;0.014)
Near big pasture area: MISSING	0.002(0.000;0.017)	0.031(0.014;0.048)

TABLE A.3: 1987: posterior median and 0.1 and 0.9 quantiles for behavioral and environmental variables.

Behavioral	$\psi_1^{(j)}$	$\psi_2^{(j)}$
Plant coffee: NO	0.002(0.000;0.014)	0.883(0.799;0.936)
Plant coffee: YES	0.994(0.981;0.999)	0.043(0.001;0.132)
Plant coffee: MISSING	0.002(0.000;0.009)	0.068(0.042;0.099)
Plant banana: NO	0.001(0.000;0.008)	0.664(0.588;0.756)
Plant banana: YES	0.998(0.990;1.000)	0.248(0.145;0.328)
Plant banana: MISSING	0.000(0.000;0.004)	0.088(0.062;0.121)
Own more the 4 goods: NO	0.002(0.000;0.012)	0.750(0.669;0.833)
Own more the 4 goods: YES	0.998(0.988;1.000)	0.250(0.167;0.331)
Own a chainsaw: NO	0.206(0.174;0.242)	0.944(0.833;0.995)
Own a chainsaw: YES	0.792(0.756;0.824)	0.051(0.001;0.161)
Own a chainsaw: MISSING	0.002(0.000;0.006)	0.003(0.000;0.015)
More than 4 people in the house: NO	0.621(0.590;0.652)	0.004(0.000;0.041)
More than 4 people in the house: YES	0.298(0.267;0.331)	0.990(0.953;0.999)
More than 4 people in the house: MISSING	0.080(0.066;0.095)	0.001(0.000;0.014)
Plan to build a new house within a year: NO	0.822(0.787;0.858)	0.219(0.115;0.324)
Plan to build a new house within a year: YES	0.172(0.136;0.208)	0.740(0.634;0.843)
Plan to build a new house within a year: MISSING	0.004(0.000;0.014)	0.042(0.014;0.071)
Own chickens and/or porks: NO	0.001(0.000;0.005)	0.413(0.351;0.478)
Own chickens and/or porks: YES	0.999(0.994;1.000)	0.399(0.323;0.471)
Own chickens and/or porks: MISSING	0.000(0.000;0.002)	0.187(0.147;0.231)
HH wife has high level of education: < 4 yr	0.570(0.533;0.607)	0.088(0.002;0.190)
HH wife has high level of education: > 4 yr	0.398(0.365;0.434)	0.309(0.217;0.398)
HH wife has high level of education: NO-WIFE	0.001(0.000;0.008)	0.051(0.027;0.077)
HH wife has high level of education: MISSING	0.027(0.003;0.054)	0.547(0.462;0.629)
Plant cocoa: NO	0.452(0.421;0.483)	0.942(0.914;0.962)
Plant cocoa: YES	0.547(0.516;0.577)	0.002(0.000;0.015)
Plant cocoa: MISSING	0.000(0.000;0.004)	0.053(0.034;0.078)
Do you go often to urban area: NO	0.429(0.393;0.468)	0.802(0.695;0.892)
Do you go often to urban area: YES	0.535(0.496;0.573)	0.108(0.009;0.215)
Do you go often to urban area: MISSING	0.035(0.021;0.050)	0.090(0.048;0.138)
Do you own other proprieties: NO	0.487(0.453;0.523)	0.887(0.798;0.971)
Do you own other proprieties: YES	0.513(0.477;0.547)	0.113(0.029;0.202)
Arrived in Rondonia before 1985: NO	0.956(0.934;0.976)	0.574(0.496;0.656)
Arrived in Rondonia before 1985: YES	0.023(0.006;0.044)	0.369(0.293;0.451)
Arrived in Rondonia before 1985: MISSING	0.020(0.009;0.032)	0.053(0.019;0.094)
DDT is used: NO	0.091(0.066;0.119)	0.419(0.335;0.508)
DDT is used: YES	0.907(0.880;0.933)	0.560(0.471;0.646)
DDT is used: MISSING	0.000(0.000;0.004)	0.019(0.006;0.036)
Before coming was your occupation rural: NO	0.800(0.765;0.836)	0.452(0.354;0.555)
Before coming was your occupation rural: YES	0.200(0.164;0.235)	0.548(0.445;0.646)
Use plant to cure malaria: NO	0.448(0.410;0.485)	0.782(0.676;0.884)
Use plant to cure malaria: YES	0.538(0.499;0.576)	0.198(0.099;0.307)
Use plant to cure malaria: MISSING	0.015(0.005;0.024)	0.011(0.000;0.049)
Use protective clothes: NO	0.855(0.823;0.884)	0.514(0.412;0.603)
Use protective clothes: YES	0.137(0.108;0.167)	0.294(0.211;0.392)
Use protective clothes: MISSING	0.005(0.000;0.021)	0.190(0.137;0.245)
Spray insecticide: NO	0.603(0.569;0.640)	0.917(0.815;0.991)
Spray insecticide: YES	0.397(0.360;0.431)	0.083(0.009;0.185)
Are there rubber tree: NO	0.693(0.667;0.719)	0.942(0.913;0.963)
Are there rubber tree: YES	0.306(0.280;0.331)	0.001(0.000;0.013)
Are there rubber tree: MISSING	0.000(0.000;0.004)	0.054(0.035;0.079)
Part of family did not come: NO	0.639(0.603;0.674)	0.359(0.260;0.458)
Part of family did not come: YES	0.358(0.324;0.394)	0.612(0.515;0.711)
Part of family did not come: MISSING	0.001(0.000;0.006)	0.027(0.013;0.048)
Use a bednet: NO	0.789(0.755;0.822)	0.559(0.462;0.657)
Use a bednet: YES	0.204(0.172;0.238)	0.250(0.151;0.342)
Use a bednet: MISSING	0.004(0.000;0.018)	0.192(0.140;0.243)
Plant guarana: NO	0.769(0.745;0.792)	0.916(0.886;0.941)
Plant guarana: YES	0.229(0.206;0.253)	0.001(0.000;0.011)
Plant guarana: MISSING	0.000(0.000;0.004)	0.081(0.056;0.110)
HH has high level of education: NO	0.631(0.596;0.667)	0.415(0.313;0.513)
HH has high level of education: YES	0.365(0.329;0.400)	0.524(0.428;0.626)
HH has high level of education: MISSING	0.002(0.000;0.012)	0.059(0.029;0.091)
Get malaria from dirty water: NO	0.467(0.429;0.503)	0.289(0.189;0.393)
Get malaria from dirty water: YES	0.497(0.460;0.536)	0.655(0.548;0.755)
Get malaria from dirty water: MISSING	0.035(0.021;0.051)	0.054(0.014;0.102)
Do you ever go to city through BR364: NO	0.576(0.538;0.612)	0.602(0.503;0.698)
Do you ever go to city through BR364: YES	0.407(0.370;0.444)	0.287(0.187;0.380)
Do you ever go to city through BR364: MISSING	0.017(0.003;0.032)	0.110(0.065;0.168)
Arrived in Machadino before 1985: NO	0.168(0.144;0.191)	0.025(0.002;0.084)
Arrived in Machadino before 1985: YES	0.832(0.809;0.856)	0.975(0.916;0.998)
Knowledge of malaria vector: NO	0.504(0.466;0.540)	0.452(0.348;0.550)
Knowledge of malaria vector: YES	0.327(0.292;0.360)	0.332(0.243;0.430)
Knowledge of malaria vector: MISSING	0.169(0.141;0.197)	0.215(0.139;0.294)
Worked in rural area for more tha 1 year: NO	0.032(0.006;0.049)	0.059(0.011;0.141)
Worked in rural area for more tha 1 year: YES	0.967(0.950;0.993)	0.892(0.807;0.942)
Worked in rural area for more tha 1 year: MISSING	0.000(0.000;0.003)	0.047(0.029;0.071)
Lived in rural area for more than 1 year: NO	0.028(0.003;0.045)	0.046(0.001;0.129)
Lived in rural area for more than 1 year: YES	0.972(0.954;0.997)	0.902(0.819;0.953)
Lived in rural area for more than 1 year: MISSING	0.000(0.000;0.002)	0.049(0.030;0.074)
Own a planter: NO	0.656(0.619;0.691)	0.698(0.596;0.799)
Own a planter: YES	0.343(0.308;0.380)	0.297(0.198;0.397)
Own a planter: MISSING	0.000(0.000;0.003)	0.004(0.000;0.014)

Environmental	$\psi_1^{(j)}$	$\psi_2^{(j)}$
House has more than 4 rooms: NO	0.089(0.049;0.132)	0.905(0.851;0.954)
House has more than 4 rooms: YES	0.841(0.799;0.884)	0.003(0.000;0.024)
House has more than 4 rooms: MISSING	0.069(0.043;0.095)	0.086(0.042;0.137)
Walls have good quality: NO	0.166(0.123;0.213)	0.916(0.836;0.987)
Walls have good quality: YES	0.832(0.785;0.875)	0.082(0.011;0.163)
Walls have good quality: MISSING	0.001(0.000;0.005)	0.001(0.000;0.006)
Roof has good quality: NO	0.003(0.000;0.026)	0.671(0.603;0.743)
Roof has good quality: YES	0.995(0.971;0.999)	0.325(0.253;0.394)
Roof has good quality: MISSING	0.001(0.000;0.004)	0.002(0.000;0.009)
Near big planted area: NO	0.542(0.507;0.577)	0.698(0.641;0.749)
Near big planted area: YES	0.444(0.411;0.478)	0.001(0.000;0.018)
Near big planted area: NO-PLOT	0.012(0.000;0.030)	0.091(0.055;0.128)
Near big planted area: MISSING	0.000(0.000;0.002)	0.204(0.168;0.246)
Good water source available: NO	0.241(0.200;0.282)	0.647(0.558;0.724)
Good water source available: YES	0.758(0.717;0.799)	0.349(0.271;0.438)
Good water source available: MISSING	0.000(0.000;0.003)	0.003(0.000;0.010)
Do you have close neighbours (<500mt): NO	0.666(0.619;0.718)	0.305(0.208;0.396)
Do you have close neighbours (<500mt): YES	0.334(0.282;0.381)	0.695(0.604;0.792)
Has the surrounding area being cleared: NO	0.950(0.927;0.978)	0.618(0.544;0.685)
Has the surrounding area being cleared: YES	0.048(0.022;0.072)	0.345(0.282;0.416)
Has the surrounding area being cleared: MISSING	0.000(0.000;0.004)	0.035(0.021;0.053)
Is topography bottom: NO	0.435(0.395;0.478)	0.152(0.080;0.222)
Is topography bottom: YES	0.528(0.485;0.570)	0.801(0.722;0.875)
Is topography bottom: MISSING	0.036(0.021;0.054)	0.047(0.018;0.081)
Near big pasture area: NO	0.950(0.934;0.965)	0.721(0.667;0.769)
Near big pasture area: YES	0.048(0.034;0.064)	0.010(0.000;0.044)
Near big pasture area: NO-PLOT	0.000(0.000;0.005)	0.051(0.034;0.072)
Near big pasture area: MISSING	0.000(0.000;0.002)	0.210(0.172;0.253)
More that 100mt from a forest: NO	0.803(0.778;0.829)	0.960(0.924;0.980)
More that 100mt from a forest: YES	0.194(0.168;0.220)	0.009(0.000;0.048)
More that 100mt from a forest: MISSING	0.002(0.000;0.008)	0.026(0.013;0.043)
Sealing has good quality: NO	0.836(0.813;0.857)	0.993(0.977;0.999)
Sealing has good quality: YES	0.164(0.142;0.185)	0.002(0.000;0.017)
Sealing has good quality: MISSING	0.000(0.000;0.003)	0.003(0.000;0.010)
Distance from coop >200mt: NO	0.990(0.976;0.999)	0.895(0.860;0.929)
Distance from coop >200mt: YES	0.005(0.000;0.019)	0.065(0.035;0.094)
Distance from coop >200mt: MISSING	0.002(0.000;0.011)	0.040(0.023;0.061)
More that 10km from an hospital: NO	0.128(0.103;0.152)	0.030(0.002;0.072)
More that 10km from an hospital: YES	0.872(0.848;0.897)	0.970(0.928;0.998)
Good bathing place is available: NO	0.925(0.909;0.940)	0.994(0.978;0.999)
Good bathing place is available: YES	0.073(0.059;0.089)	0.003(0.000;0.018)
Good bathing place is available: MISSING	0.001(0.000;0.004)	0.001(0.000;0.008)

TABLE A.4: 1995: posterior median and 0.1 and 0.9 quantiles for behavioral and environmental variables.

Behavioral	$\psi_1^{(j)}$	$\psi_2^{(j)}$
Planted Corn: NO	0.969(0.949;0.980)	0.041(0.003;0.087)
Planted Corn: YES	0.002(0.000;0.018)	0.958(0.912;0.996)
Planted Corn: MISSING	0.027(0.018;0.038)	0.000(0.000;0.002)
Plant Cassava: NO	0.951(0.932;0.967)	0.090(0.046;0.137)
Plant Cassava: YES	0.001(0.000;0.012)	0.906(0.859;0.951)
Plant Cassava: MISSING	0.045(0.030;0.062)	0.002(0.000;0.010)
Plant banana: NO	0.942(0.876;0.966)	0.187(0.147;0.232)
Plant banana: YES	0.015(0.000;0.085)	0.812(0.767;0.852)
Plant banana: MISSING	0.038(0.027;0.052)	0.000(0.000;0.003)
Cultivate rice: NO	0.740(0.670;0.823)	0.001(0.000;0.005)
Cultivate rice: YES	0.231(0.144;0.304)	0.999(0.994;1.000)
Cultivate rice: MISSING	0.027(0.018;0.040)	0.000(0.000;0.002)
Planted Bean: NO	0.967(0.951;0.978)	0.330(0.287;0.370)
Planted Bean: YES	0.001(0.000;0.009)	0.669(0.629;0.711)
Planted Bean: MISSING	0.030(0.020;0.044)	0.001(0.000;0.004)
Own a chainsaw: NO	0.947(0.869;0.998)	0.298(0.250;0.346)
Own a chainsaw: YES	0.053(0.002;0.131)	0.702(0.654;0.750)
Active in community organization: NO	0.990(0.965;0.998)	0.401(0.359;0.438)
Active in community organization: YES	0.005(0.000;0.030)	0.597(0.561;0.640)
Active in community organization: MISSING	0.003(0.000;0.009)	0.001(0.000;0.004)
More than 4 people in the house: NO	0.167(0.091;0.249)	0.681(0.634;0.730)
More than 4 people in the house: YES	0.833(0.751;0.909)	0.319(0.270;0.366)
Own a planter: NO	0.461(0.410;0.517)	0.001(0.000;0.007)
Own a planter: YES	0.539(0.483;0.590)	0.999(0.993;1.000)
Plant coffee: NO	0.403(0.355;0.457)	0.001(0.000;0.006)
Plant coffee: YES	0.566(0.511;0.617)	0.999(0.993;1.000)
Plant coffee: MISSING	0.029(0.019;0.042)	0.000(0.000;0.002)
HH wife has high level of education: < 4 yr	0.148(0.077;0.210)	0.468(0.428;0.505)
HH wife has high level of education: > 4 yr	0.503(0.438;0.572)	0.502(0.462;0.540)
HH wife has high level of education: NO-WIFE	0.008(0.001;0.016)	0.001(0.000;0.005)
HH wife has high level of education: MISSING	0.343(0.279;0.407)	0.027(0.000;0.061)
Own more the 4 goods: NO	0.340(0.300;0.385)	0.001(0.000;0.008)
Own more the 4 goods: YES	0.660(0.615;0.700)	0.999(0.992;1.000)
Plant cocoa: NO	0.965(0.946;0.977)	0.660(0.630;0.688)
Plant cocoa: YES	0.002(0.000;0.017)	0.339(0.312;0.370)
Plant cocoa: MISSING	0.030(0.021;0.043)	0.000(0.000;0.002)
Are there rubber tree: NO	0.975(0.959;0.985)	0.680(0.652;0.707)
Are there rubber tree: YES	0.002(0.000;0.016)	0.319(0.293;0.348)
Are there rubber tree: MISSING	0.020(0.012;0.031)	0.000(0.000;0.002)
Arrived in Rondonia before 1985: NO	0.543(0.476;0.605)	0.839(0.803;0.876)
Arrived in Rondonia before 1985: YES	0.396(0.334;0.461)	0.150(0.117;0.184)
Arrived in Rondonia before 1985: MISSING	0.063(0.034;0.089)	0.009(0.000;0.026)
Use plant to cure malaria: NO	0.853(0.795;0.918)	0.577(0.536;0.617)
Use plant to cure malaria: YES	0.145(0.080;0.203)	0.422(0.382;0.463)
Use plant to cure malaria: MISSING	0.001(0.000;0.006)	0.001(0.000;0.003)
Plant guarana: NO	0.979(0.942;0.998)	0.737(0.709;0.766)
Plant guarana: YES	0.021(0.002;0.058)	0.263(0.234;0.291)
DDT is used: NO	0.822(0.758;0.881)	0.613(0.575;0.652)
DDT is used: YES	0.149(0.092;0.215)	0.386(0.347;0.424)
DDT is used: MISSING	0.027(0.017;0.040)	0.000(0.000;0.003)
Do you own other proprieties: NO	0.784(0.712;0.851)	0.565(0.522;0.611)
Do you own other proprieties: YES	0.216(0.149;0.288)	0.435(0.389;0.478)
Got a loan for pasture: NO	0.963(0.947;0.974)	0.788(0.765;0.809)
Got a loan for pasture: YES	0.001(0.000;0.009)	0.211(0.189;0.234)
Got a loan for pasture: MISSING	0.034(0.024;0.049)	0.000(0.000;0.003)
Planted Nut: NO	0.947(0.920;0.964)	0.801(0.776;0.824)
Planted Nut: YES	0.003(0.000;0.026)	0.197(0.174;0.221)
Planted Nut: MISSING	0.046(0.032;0.063)	0.001(0.000;0.006)
Own chickens and/or porks: NO	0.188(0.158;0.221)	0.001(0.000;0.003)
Own chickens and/or porks: YES	0.812(0.779;0.842)	0.999(0.997;1.000)
Knowledge of malaria vector: NO	0.310(0.240;0.379)	0.461(0.419;0.504)
Knowledge of malaria vector: YES	0.573(0.504;0.645)	0.474(0.430;0.516)
Knowledge of malaria vector: MISSING	0.117(0.072;0.163)	0.064(0.041;0.091)
Planted Pepper: NO	0.975(0.962;0.984)	0.842(0.822;0.861)
Planted Pepper: YES	0.001(0.000;0.008)	0.157(0.138;0.178)
Planted Pepper: MISSING	0.022(0.014;0.033)	0.000(0.000;0.002)
Get malaria from dirty water: NO	0.507(0.433;0.582)	0.379(0.335;0.421)
Get malaria from dirty water: YES	0.462(0.388;0.535)	0.602(0.560;0.645)
Get malaria from dirty water: MISSING	0.028(0.007;0.055)	0.020(0.006;0.033)
Got a loan for agriculture: NO	0.963(0.948;0.975)	0.855(0.835;0.872)
Got a loan for agriculture: YES	0.001(0.000;0.006)	0.144(0.127;0.164)
Got a loan for agriculture: MISSING	0.034(0.023;0.049)	0.000(0.000;0.003)
Spray insecticide: NO	0.850(0.789;0.909)	0.712(0.674;0.749)
Spray insecticide: YES	0.148(0.089;0.210)	0.286(0.250;0.324)
Spray insecticide: MISSING	0.001(0.000;0.005)	0.001(0.000;0.003)
Do you go often to urban area: NO	0.492(0.405;0.580)	0.618(0.567;0.666)
Do you go often to urban area: YES	0.497(0.409;0.582)	0.380(0.332;0.431)
Do you go often to urban area: MISSING	0.010(0.003;0.019)	0.001(0.000;0.006)
Lived in rural area for more than 1 year: NO	0.140(0.101;0.191)	0.038(0.013;0.059)
Lived in rural area for more than 1 year: YES	0.847(0.797;0.888)	0.960(0.939;0.986)
Lived in rural area for more than 1 year: MISSING	0.011(0.004;0.021)	0.000(0.000;0.004)
Arrived in Machadino before 1985: NO	0.001(0.000;0.013)	0.096(0.081;0.112)
Arrived in Machadino before 1985: YES	0.968(0.943;0.991)	0.891(0.870;0.910)
Arrived in Machadino before 1985: MISSING	0.028(0.007;0.051)	0.012(0.001;0.026)
Use a bednet: NO	0.880(0.835;0.920)	0.963(0.941;0.984)
Use a bednet: YES	0.104(0.068;0.146)	0.029(0.008;0.048)
Use a bednet: MISSING	0.014(0.001;0.033)	0.008(0.000;0.018)
Have another rural plot: NO	0.839(0.779;0.898)	0.763(0.727;0.800)
Have another rural plot: YES	0.159(0.100;0.218)	0.236(0.200;0.272)
Have another rural plot: MISSING	0.001(0.000;0.006)	0.001(0.000;0.003)
HH has high level of education: NO	0.392(0.313;0.462)	0.441(0.401;0.486)
HH has high level of education: YES	0.597(0.527;0.675)	0.557(0.513;0.597)
HH has high level of education: MISSING	0.010(0.003;0.020)	0.001(0.000;0.005)
Go to main urban area from treatment: NO	0.199(0.131;0.277)	0.165(0.122;0.206)
Go to main urban area from treatment: YES	0.790(0.711;0.859)	0.834(0.794;0.877)
Go to main urban area from treatment: MISSING	0.010(0.004;0.017)	0.000(0.000;0.002)
Go to secondary urban area from treatment: NO	0.829(0.753;0.898)	0.829(0.789;0.871)
Go to secondary urban area from treatment: YES	0.158(0.089;0.233)	0.170(0.129;0.211)
Go to secondary urban area from treatment: MISSING	0.013(0.006;0.022)	0.000(0.000;0.002)
Got a loan for equipment: NO	0.963(0.949;0.975)	0.981(0.973;0.987)
Got a loan for equipment: YES	0.000(0.000;0.005)	0.018(0.012;0.025)
Got a loan for equipment: MISSING	0.036(0.024;0.049)	0.000(0.000;0.003)

Environmental	$\psi_1^{(j)}$	$\psi_2^{(j)}$
House has more than 4 rooms: NO	0.923(0.781;0.969)	0.001(0.000;0.010)
House has more than 4 rooms: YES	0.024(0.000;0.184)	0.996(0.988;1.000)
House has more than 4 rooms: MISSING	0.040(0.013;0.072)	0.001(0.000;0.006)
More that 10km from an hospital: NO	0.568(0.432;0.710)	0.007(0.000;0.024)
More that 10km from an hospital: YES	0.432(0.290;0.568)	0.993(0.976;1.000)
Anybody cleared the area before HH: NO	0.006(0.000;0.050)	0.332(0.311;0.354)
Anybody cleared the area before HH: YES	0.697(0.609;0.779)	0.667(0.645;0.687)
Anybody cleared the area before HH: MISSING	0.285(0.212;0.370)	0.001(0.000;0.004)
Has the surrounding area being cleared: NO	0.673(0.576;0.760)	0.996(0.990;0.999)
Has the surrounding area being cleared: YES	0.006(0.000;0.031)	0.002(0.000;0.005)
Has the surrounding area being cleared: MISSING	0.316(0.232;0.407)	0.001(0.000;0.007)
Do you have close neighbours (<500mt): NO	0.427(0.234;0.615)	0.661(0.630;0.690)
Do you have close neighbours (<500mt): YES	0.367(0.188;0.551)	0.309(0.280;0.339)
Do you have close neighbours (<500mt): MISSING	0.205(0.095;0.324)	0.030(0.016;0.046)
Is topography bottom: NO	0.937(0.799;0.997)	0.688(0.664;0.711)
Is topography bottom: YES	0.057(0.001;0.193)	0.305(0.283;0.329)
Is topography bottom: MISSING	0.002(0.000;0.018)	0.007(0.004;0.011)
Is road quality good: NO	0.005(0.000;0.047)	0.094(0.082;0.107)
Is road quality good: YES	0.534(0.330;0.730)	0.644(0.612;0.677)
Is road quality good: MISSING	0.452(0.257;0.646)	0.261(0.231;0.291)
Near big pasture area: NO	0.801(0.735;0.858)	0.998(0.994;1.000)
Near big pasture area: YES	0.010(0.000;0.029)	0.001(0.000;0.003)
Near big pasture area: MISSING	0.187(0.131;0.251)	0.000(0.000;0.005)
Distant from stagnant water: NO	0.796(0.708;0.873)	0.975(0.965;0.984)
Distant from stagnant water: YES	0.002(0.000;0.024)	0.018(0.013;0.024)
Distant from stagnant water: MISSING	0.196(0.122;0.278)	0.006(0.000;0.016)
More than 600mt from a river: NO	0.558(0.345;0.772)	0.635(0.602;0.667)
More than 600mt from a river: YES	0.363(0.149;0.580)	0.365(0.332;0.397)
More than 600mt from a river: MISSING	0.076(0.048;0.116)	0.000(0.000;0.002)
Roof has good quality: NO	0.200(0.120;0.282)	0.008(0.001;0.019)
Roof has good quality: YES	0.800(0.718;0.880)	0.992(0.981;0.999)
Sealing has good quality: NO	0.499(0.291;0.686)	0.641(0.609;0.671)
Sealing has good quality: YES	0.501(0.314;0.709)	0.359(0.329;0.391)
Distance from coop >200mt: NO	0.802(0.719;0.860)	0.908(0.895;0.922)
Distance from coop >200mt: YES	0.013(0.000;0.083)	0.091(0.078;0.104)
Distance from coop >200mt: MISSING	0.174(0.126;0.233)	0.000(0.000;0.003)
Walls have good quality: NO	0.387(0.211;0.598)	0.262(0.232;0.290)
Walls have good quality: YES	0.613(0.402;0.789)	0.738(0.710;0.768)
Distant from to well: NO	0.852(0.689;0.929)	0.878(0.860;0.900)
Distant from to well: YES	0.052(0.001;0.218)	0.120(0.099;0.138)
Distant from to well: MISSING	0.085(0.049;0.130)	0.001(0.000;0.005)
Good water source available: NO	0.217(0.049;0.400)	0.210(0.184;0.238)
Good water source available: YES	0.777(0.593;0.945)	0.789(0.761;0.815)
Good water source available: MISSING	0.003(0.000;0.016)	0.000(0.000;0.002)
More that 100mt from a forest: NO	0.917(0.809;0.962)	0.854(0.836;0.871)
More that 100mt from a forest: YES	0.031(0.000;0.145)	0.146(0.128;0.164)
More that 100mt from a forest: MISSING	0.045(0.024;0.075)	0.000(0.000;0.002)
Good bathing place is available: NO	0.985(0.924;0.999)	0.873(0.857;0.888)
Good bathing place is available: YES	0.015(0.001;0.076)	0.127(0.112;0.143)
More than 500mt from health unit: NO	0.086(0.016;0.181)	0.037(0.024;0.049)
More than 500mt from health unit: YES	0.914(0.819;0.984)	0.963(0.951;0.976)

Bibliography

Agresti, A. (2013) Categorical data analysis, 3rd edn. *New Jersey: John Willey & Sons* .

Airoldi, E., Blei, D., Xing, E. and Fienberg, S. (2005) A latent mixed membership model for relational data. *Proceedings of the 3rd International Workshop on Link Discovery(LinkKDD '05). New York, NY, USA: ACM* pp. 82–89.

Airoldi, E. M., Blei, D., Erosheva, E. A. and Fienberg, S. E. (2014) *Handbook of Mixed Membership Models and their Applications.* New York: Chapman and Hall/CRC.

Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T. and Shringarpure, S. (2010) Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences* .

Aitchison, J. and Shen, S. M. (1980) Logistic-normal distributions: Some properties and uses. *Biometrika* **67**(2), 261–272.

Alqallaf, F. and Gustafson, P. (2001) On cross-validation of Bayesian models. *The Canadian Journal of Statistics* **29**(2), 333–340.

Artin, M. (1991) *Algebra.* Prentice Hall.

Asuncion, A. and Newman, D. (2007) Uci machine learning repository.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical modeling and analysis for spatial data.* New York: Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. E. and Polasek, W. (2000) Geostatistical modelling for spatial interaction data with application to postal service performance. *Journal of Statistical Planning and Inference* **90**(1), 87 – 105.

Barry, D. and Hartigan, J. A. (1993) A Bayesian analysis for change point problems. *Journal of the American Statistical Association* **88**(421), 309–319.

Bereczky, S., Liljander, A., Rooth, I., Faraja, L., Granath, F., Montgomery, S. M. and Färnert, A. (2007) Multiclonal asymptomatic Plasmodium falciparum infections predict a reduced risk of malaria disease in a Tanzanian population. *Microbes and Infection* **9**(1), 103 – 110.

Berkman, L., Singer, B. and Manton, K. (1989) Black/white differences in health status and mortality among the elderly. *Demography* **26**(4), 661–678.

Bhattacharya, A. and Dunson, D. B. (2011) Sparse Bayesian infinite factor models. *Biometrika* pp. 291–306.

Bhattacharya, A. and Dunson, D. B. (2012) Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association* **107**(497), 362–377.

Blasi, P. D., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015) Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(2), 212–229.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

Castro, M. C., Monte-Mór, R. L., Sawyer, D. O. and Singer, B. H. (2006) Malaria risk on the Amazon frontier. *Proceedings of the National Academy of Sciences* **103**(7), 2452–2457.

Chuit, R., Gurtler, R. E., Mac Dougall, L., Segura, E. L. and Singer, B. (2001) Chagas disease-risk assessment by an environmental approach in northern Argentina. *Revista de Patologia Tropical* **30**(2), 193–208.

Crowley, E. M. (1997) Product partition models for normal means. *Journal of the American Statistical Association* **92**(437), 192–198.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39**(1), 1–38.

Dlamini, N., Hsiang, M. S., Ntshalintshali, N., Pindolia, D., Allen, R., Nhlabathi, N., Novotny, J., Kang Dufour, M.-S., Midekisa, A., Gosling, R., LeMenach, A., Cohen, J., Dorsey, G., Greenhouse, B. and Kunene, S. (2017) Low-quality housing is associated with increased risk of malaria infection: A national population-based study from the low transmission setting of Swaziland. *Open Forum Infectious Diseases* **4**(2), ofx071.

Dobra, A. and Lenkoski, A. (2011) Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics* **5**(2A), 969–993.

Dunson, D. B. and Xing, C. (2009) Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**(487), 1042–1051.

Durante, D., Canale, A. and Rigon, T. (2019) A nested expectation–maximization algorithm for latent class models with covariates. *Statistics & Probability Letters* **146**, 97 – 103.

Eckart, C. and Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika* **1**(3), 211–218.

Erosheva, E. A. and Fienberg, S. E. (2005) Bayesian mixed membership models for soft clustering and classification. *Classification – The Ubiquitous Challenge, Springer* pp. 11–26.

Erosheva, E. A., Fienberg, S. E. and Joutard, C. (2007) Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics* **1**(2), 502–537.

Ettling, M. B. and Shepard, D. S. (1991) Economic cost of malaria in Rwanda. *Tropical Medicine and Parasitology: official organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ)* **42**(3), 214–218.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**(2), 209–230.

Franco, A. O., Gomes, M. G. M., Rowland, M., Coleman, P. G. and Davies, C. R. (2014) Controlling malaria using livestock-based interventions: A one health approach. *PLOS ONE* **9**(7), 1–17.

Gallup, J. and Sachs, J. (2001) The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene* **64**(1), 85–96.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis.* CRC Press.

Gnedin, A. and Pitman, J. (2006) Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* **138**(3), 5674–5685.

Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**(2), 215–231.

Green, P. J. and Richardson, S. (2001) Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**(2), 355–375.

Gross, J. H. and Manrique-Vallier, D. (2012) A mixed-membership approach to the assessment of political ideology from survey responses. *Handbook of mixed membership models and their applications* pp. 119–140.

Guerra, C., Snow, R. and Hay, S. (2006) A global assessment of closed forests, deforestation and malaria risk. *Annals of tropical Medicine and Parasitology* **100**(3), 189–204.

Guhaniyogi, R., Qamar, S. and Dunson, D. B. (2017) Bayesian tensor regression. *Journal of Machine Learning Research* **18**(79), 1–31.

Hartigan, J. A. (1990) Partition models. *Communications in statistics-Theory and methods* **19**(8), 2745–2756.

Heller, K. A., Williamson, S. and Ghahramani, Z. (2008) Statistical models for partial membership. *Proceedings of the 25th International Conference on Machine Learning* pp. 392–399.

Hoff, P. D. (2015) Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics* **9**(3), 1169–1193.

Hoffman, M. D., Blei, D. M. and Bach, F. (2010) Online learning for latent Dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 856–864. USA: Curran Associates Inc.

Ijumba, J. and Lindsay, S. (2001) Impact of irrigation on malaria in Africa: paddies paradox. *Medical and Veterinary Entomology* **15**(1), 1–11.

Ishwaran, H. and Zarepour, M. (2008) Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics* **30**(2), 269–283.

Jain, S. and Neal, R. M. (2004) A split-merge Markov chain monte carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics* **13**(1), 158–182.

Johndrow, J. E., Bhattacharya, A. and Dunson, D. B. (2017) Tensor decompositions and sparse log-linear models. *Annals of statistics* **45**(1), 1.

Jukes, M. C. H., Pinder, M., Grigorenko, E. L., Smith, H. B., Walraven, G., Bariau, E. M., Sternberg, R. J., Drake, L. J., Milligan, P., Cheung, Y. B., Greenwood, B. M. and Bundy, D. A. P. (2006) Long-term impact of malaria chemoprophylaxis on cognitive abilities and educational attainment: Follow-up of a controlled trial. *PLOS Clinical Trials* **1**(4), 1–8.

Kim, Y.-D. and Choi, S. (2007) Nonnegative Tucker decomposition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8.

Knowles, D. and Ghahramani, Z. (2011) Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics* **5**(2B), 1534–1552.

Kolda, T. G. and Bader, B. W. (2009) Tensor decompositions and applications. *SIAM review* **51**(3), 455–500.

Kunihama, T. and Dunson, D. B. (2013) Bayesian modeling of temporal dependence in large sparse contingency tables. *Journal of the American Statistical Association* **108**(504), 1324–1338.

Lafferty, J. D. and Blei, D. M. (2006) Correlated topic models. *Advances in Neural Information Processing Systems* **18**, 147–154.

Landis, J. R., Miller, M. E., Davis, C. S. and Koch, G. G. (1988) Some general methods for the analysis of categorical data in longitudinal studies. *Statistics in medicine* **7**(1-2), 109–137.

Lazarsfeld, P. F. (1950) In Measurement and prediction, S. Stouffer, ed. Princeton University Press, Princeton, N.J. **Chapters 10 – 11**.

Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent structure analysis.* New York, NY: Houghton, Mifflin.

Leighton, C. and Foster, R. (1993) Economic impacts of malaria in Kenya and Nigeria. *Health Financing and Sustainability Project* **6**.

Linderman, S., Johnson, M. and Adams, R. P. (2015) Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems* **28**, 3456–3464.

Lock, E. F. (2018) Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics* **27**(3), 638–647.

Lopes, H. F. and West, M. (2004) Bayesian model assessment in factor analysis. *Statistica Sinica* **14**(1), 41–67.

Males, S., Gaye, O. and Garcia, A. (2008) Long-term asymptomatic carriage of Plasmodium falciparum protects from malaria attacks: a prospective study among Senegalese children. *Clinical Infectious Diseases* **46**(4), 516–522.

Manton, K. G., Stallard, E., Woodbury, M. A., Tolley, H. D. and Yashin, A. I. (1987) Grade-of-membership techniques for studying complex event history processes with unobserved covariates. *Sociological Methodology* **17**, 309–346.

Massam, H., Liu, J. and Dobra, A. (2009) A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics* **37**(6A), 3431–3467.

Miller, J., Betancourt, B., Zaidi, A., Wallach, H. and Steorts, R. C. (2015) Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *Bayesian Nonparametrics: The Next Generation workshop, NIPS* .

Mimno, D., Hoffman, M. D. and Blei, D. M. (2012) Sparse stochastic inference for latent Dirichlet allocation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1515–1522.

Müller, P., Quintana, F. and Rosner, G. L. (2011) A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20**(1), 260–278.

Muthen, B. and Christoffersson, A. (1981) Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* **46**(4), 407–419.

Nardi, Y., Rinaldo, A. *et al.* (2012) The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18**(3), 945–974.

Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**(2), 249–265.

Ntzoufras, I., Forster, J. J. and Dellaportas, P. (2000) Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation* **68**(1), 23–37.

Nur, E. T. M. (1993) The impact of malaria on labour use and efficiency in the Sudan. *Social Science & Medicine* **37**(9), 1115–1119.

O'Neill, M. C. (1989) Escherichia coli promoters. i. Consensus as it relates to spacing class, specificity, repeat substructure, and three-dimensional organization. *Journal of Biological Chemistry* **264**(10), 5522–5530.

Papathomas, M. and Richardson, S. (2016) Exploring dependence between categorical variables: Benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms. *Journal of Statistical Planning and Inference* **173**, 47 – 63.

Park, J.-H. and Dunson, D. B. (2010) Bayesian generalized product partition model. *Statistica Sinica* **20**(3), 1203–1226.

Pitman, J. and Yor, M. (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**(2), 855–900.

Polson, N. G. and Scott, J. G. (2011) Default Bayesian analysis for multi-way tables: a data-augmentation approach. *arXiv preprint arXiv:1109.4180* .

Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American statistical Association* **108**(504), 1339–1349.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**(2), 945–959.

Rousseau, J. and Mengersen, K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B* **73**(5), 689–710.

Russo, M., Durante, D. and Scarpa, B. (2018) Bayesian inference on group differences in multivariate categorical data. *Computational Statistics & Data Analysis* **126**, 136 – 149.

Santos, L. M., Amorim, L. D. A., Santos, D. N. and Barreto, M. L. (2015) Measuring the level of social support using latent class analysis. *Social science research* **50**, 139–146.

Sauerborn, R., Ibrango, I., Nougtara, A., Borchert, M., Hien, M., Benzler, J., Koob, E. and Diesfeld, H. (1995) The economic costs of illness for rural households in Burkina Faso. *Tropical Medicine and Parasitology* **46**(1), 54–60.

Sawyer, D. R. Sawyer, D. O. (1992) In *Advancing health in developing countries: the role of social research*, ed. L. C. Chen, pp. 105–122. Auburn House.

Schein, A., Zhou, M., Blei, D. M. and Wallach, H. (2016) Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *International Conference on Machine Learning*.

Singer, B. (1989) Grade of membership representations: Concepts and problems. *Probability, Statistics, and Mathematics* pp. 317 – 334.

Singer, B. H. and Castro, M. C. (2014) Interpretability constraints and trade-offs in using mixed membership models. *Handbook of mixed membership models and their applications* pp. 159–172.

Stephens, M. (2002) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B* **62**(4), 795–809.

Tucker, L. R. (1963) Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change* **15**, 122–137.

Wade, S., Dunson, D. B., Petrone, S. and Trippa, L. (2014) Improving prediction from Dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research* **15**(1), 1041–1071.

Wallach, H., Jensen, S., Dicker, L. and Heller, K. (2010) An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 892–899.

Wang, H. and Ahuja, N. (2005) Rank-r approximation of tensors using image-as-matrix representation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pp. 346–353.

White, A., Chan, J., Hayes, C. and Murphy, T. (2012) Mixed membership models for exploring user roles in online fora. In *The 6th International AAAI Conference on Weblogs and Social Media*, pp. 599–602.

Woodbury, M. A., Clive, J. and Garson, A. (1978) Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers and Biomedical Research* **11**(3), 277 – 298.

Xiong, S., Fu, Y. and Ray, A. (2018) Bayesian nonparametric modeling of categorical data for information fusion and causal inference. *Entropy* **20**(6), 396.

Yang, Y. and Dunson, D. B. (2016) Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association* **111**(514), 656–669.

Zhang, S., Castro, M. C., Canning, D. *et al.* (2011) The effect of malaria on settlement and land use: Evidence from the Brazilian Amazon. Technical report, Program on the Global Demography of Aging.

Zhao, S., Engelhardt, B. E., Mukherjee, S. and Dunson, D. B. (2018) Fast moment estimation for generalized latent Dirichlet models. *Journal of the American Statistical Association* **0**(0), 1–13.

Zhou, J., Bhattacharya, A., Herring, A. H. and Dunson, D. B. (2015) Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association* **110**(512), 1562–1576.

Zhou, J., Herring, A. H., Bhattacharya, A., Olshan, A. F. and Dunson, D. B. a. (2016) Nonparametric Bayes modeling for case control studies with many predictors. *Biometrics* **72**(1), 184–192.