

Bayesian inference on group differences in multivariate categorical data

Massimiliano Russo, Daniele Durante & Bruno Scarpa
June, 08 2017



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- **Multivariate categorical variables** are common in many field of applications including social science, psychology and biostatistic.
- We usually observe a vector of categorical data for each subject along with a discrete variable indicating **group membership** (sex, party affiliation, case/control...)

Main goal: assess evidence of group differences in the considered categorical variables.

We analyzed data from **American national election survey** 2012 (ANES) freely available at <http://electionstudies.org/>.

We focus on a scenario of potential interest for a democratic candidate.

- A democratic candidate would be particularly interested in studying differences in opinions for voters affiliated to his party and the unaffiliated ones.
- Unaffiliated voters are more sensitive to his campaign than republicans and therefore are worth being targeted to increase their positive opinions.

Data consists in

- **2 groups:** 256 **independent democrats** + 94 **unaffiliated voters** = 350 subjects.
- 12 questions, 5 levels scale, on Obama and Romney personality including morality, leadership, care, knowledgeability, intelligence and honesty.
- 5 questions, 4 levels scale, on Obama political decisions in previous presidential term: job, economy, foreign affairs, health and war.

We consider n subject and we observe

- $y_i = (y_{i1}, \dots, y_{ip})^T \in \mathcal{Y} = (1, \dots, d_1) \times \dots \times (1, \dots, d_p)$
vector of categorical data for the subject i
- $x_i \in \mathcal{X} = (1, \dots, k)$ group indicator for the subject i

and we indicate with

$$\pi_{Y,X}(y, x) = \mathbb{P}[Y = y, X = x], \quad \text{for } y \in \mathcal{Y} \quad \text{and} \quad x \in \mathcal{X}$$

the joint probability mass function underling the data
 $(y_1, x_1), \dots, (y_n, x_n)$.

Our main aim is to assess **global association** between the random variables X and Y this formally require testing

$$H_0 : \pi_{Y,X}(y, x) = \pi_Y(y)\pi_X(x)$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ versus

$$H_1 : \pi_{Y,X}(y, x) \neq \pi_Y(y)\pi_X(x)$$

for some $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ versus

Rejection of H_0 imply that variables changes across groups.

A possibility is to rely on independent tests

- 1 separately assess evidence of group differences in each categorical variable via chi-square tests
- 2 account for multiple testing via **false discovery rate control**.

This approach does not consider the dependence structure in the data (**loss of power**).

Pesarin & Salmaso (2010) addressed this issue via nonparametric permutation tests

These methods cannot capture differences that go beyond changes in the marginals providing inaccurate insights when the group differences are in higher-order structure.

To address previous robustness issues one can rely on a flexible representation of the p.m.f.

- **Log-linear models**: preferred tool but characterized by an **explosion in the number of parameters** even for a moderate p
- **Tensor factorization**: avoids pre-specifying dependence structure defining a **flexible and computationally tractable** representation for p.m.f. The considered test can be incorporated directly in this model by use of suitable prior.

We characterize the probability mass function $\pi_{Y,X}$ using the decomposition

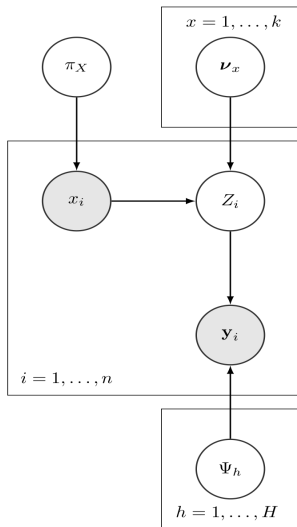
$$\pi_{Y,X}(y, x) = \pi_{Y|X=x}(y) \pi_X(x)$$

where $y = (y_1, \dots, y_p) \in \{1, \dots, d_1\} \times \dots \times \{1, \dots, d_p\}$ and $x \in \{1, \dots, k\}$.

Since $\pi_X(x)$ is simply the p.m.f. of a discrete random variable, a crucial point is to propose a **flexible model** for $\pi_{Y|X=x}(y)$.

We use the following **mixture of tensors**

$$\pi_{Y|X=x} = \sum_{h=1}^H \nu_{hx} \Psi_h, \quad \Psi_h = \psi_h^{(1)} \otimes \dots \otimes \psi_h^{(p)}$$



- $x_i \in \{1, \dots, k\}$ group indicator generated from a discrete random variable with pmf π_X
- p -variate response vector y_i is generated conditionally on x_i through a mixture representation.
- Given the group indicator $x_i = x$ we choose the mixture component $z_i \in \{1, \dots, H\}$ with probabilities $\nu_x = (\nu_{x1}, \dots, \nu_{xH})$
- The j element of y_i is generated from discrete distribution $\psi_h^{(j)}$ element of the tensor Ψ_h .

Given representations has several benefits

- Allows inference on changes in the multivariate random variable Y across the groups defined by X , with the conditional p.m.f. $\pi_{Y|X=x}(y)$ for each group $x = 1, \dots, k$
- Conditional independence of Y_j within the each mixture components $h = 1, \dots, H$ provides a **parsimonious and tractable formulation**, while introducing dependence using group-specific mixing probabilities ν_x .
- Accounting for group-dependence only in the mixing probabilities allows **borrowing of information** across units.
- The proposed model can represent any possible dependence structure

Only the mixture weights depends on the group variable X hence

$$H_0 : \nu_1 = \dots = \nu_k \quad \text{versus} \quad H_1 : \nu_x \neq \nu_{x'} \quad \text{for some } x, x'.$$

Following what proposed in Lock & Dunson (2015) we can incorporate this test directly in the model by choosing the following prior

$$\begin{aligned} \nu_x &= (1 - T)u + Tu_x \\ u &\sim \text{Dir}\{\gamma_1, \dots, \gamma_H\}, \quad u_x \sim \text{Dir}\{\gamma_1, \dots, \gamma_H\}, \quad x = 1, \dots, k \\ T &\sim \text{Ber}\{\mathbb{P}[H_1]\} \end{aligned}$$

This specification simplify inference and induce a **full support** on the probability simplex (key result for accurate inference).

- Rejection of the global null hypothesis provides evidence of group differences in the multivariate categorical random variable Y
- Such changes may be attributable to several structures and we want to provide interpretable inference.
- To do so we consider each group differences in each marginal Y_j of Y and pairwise dependence between pairs $(Y_j, Y_{j'})$

To assess group difference in each marginal we rely on a model based version of the Cramer's V coefficient proposed in Dunson & Xing (2009)

- Measuring the association between Y_j and X with $\rho_j \in [0, 1]$ provides a convenient choice for interpretation.
- A formal test can be approximated using a small interval null hypothesis $H_{0j} : \rho_j \leq \epsilon$.

To measure conditional pairwise association we consider again the model based Cramer's V .

Quantities needed for Cramer's V coefficient can be directly calculated from the model using the fact that

Given a subset of indices $\mathcal{J} \subset (1, \dots, p)$ such that $\mathcal{J} \cup \mathcal{J}^c = (1, \dots, p)$ we have

$$\pi_{Y_{\mathcal{J}}|X=x}(y_{\mathcal{J}}) = \sum_{h=1}^H \nu_{hx} \prod_{j \in \mathcal{J}} \pi_{hj}(y_j)$$

and

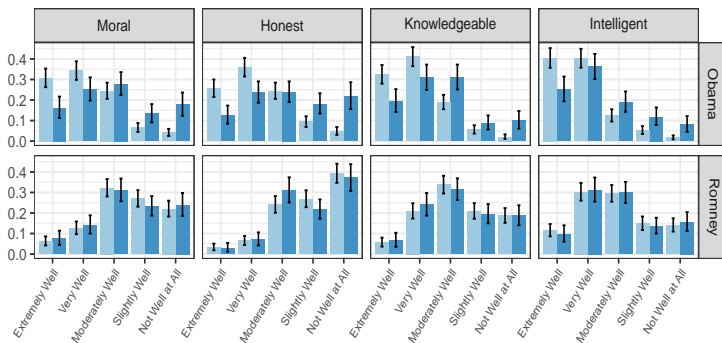
$$\pi_{Y_{\mathcal{J}}}(y_{\mathcal{J}}) = \sum_{x \in \mathcal{X}} \pi_X(x) \left\{ \sum_{h=1}^H \nu_{hx} \prod_{j \in \mathcal{J}} \pi_{hj}(y_j) \right\}$$

where $Y_{\mathcal{J}}$ is random vector containing the variables with indices in set \mathcal{J} .

Some Results I

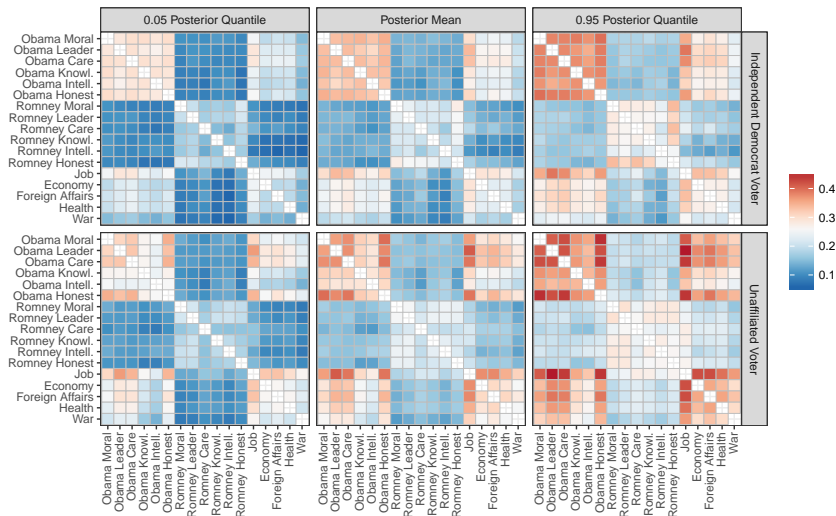


- We reject the global null hypothesis ($\mathbb{P}[H_1 \mid \text{data}] > 0.99$)
- Setting $\epsilon = 0.1$ and rejecting if $\mathbb{P}[H_{1k} \mid \text{data}]$ is > 0.99 , we have that the variables concerning Obama vary, while those concerning Romney do not.



Posterior mean and 0.90 credible interval for the opinion on Romney and Obama for independents voters (dark) and independent democrat (light).

Some Results II







Pairwise Cramér V mean and quantiles for the two groups.

We proposed

- flexible and easy to implement model for multivariate categorical variables
- procedures for inference and testing on global and local group differences.

The application to the data suggest

- Opinions about Romney do not influence opinions about Obama
- Opinions about Obama personality and its political decision are related
- Personal (as political) opinions about Obama are strongly associated while opinions about Romney are less associated.

-  RUSSO, M., DURANTE, D & SCARPA B. (2017)
Bayesian inference on group differences in multivariate categorical data
[arXiv:1606.09415](#)
-  DUNSON, D. B. & XING, C. (2009).
Nonparametric Bayes modeling of multivariate categorical data.
J. Am. Statist. Assoc. **104**, 1042–1051.
-  LOCK, E. F. & DUNSON, D. B. (2015).
Shared kernel Bayesian screening.
Biometrika **102**, 829–842.
-  PESARIN, F. & SALMASO, L. (2010).
Permutation tests for complex data: Theory, applications and software.
Wiley.