Bayesian testing of group differences in multivariate categorical data

Massimiliano Russo, Daniele Durante & Bruno Scarpa
russo@stat.unipd.it

Department of Statistical Sciences, University of Padua, Italy

Università degli Studi di Padova

Introduction & motivation

Motivation: studying global and local differences in voters' opinions across party affiliation groups during the 2012 American national elections.

Data type: vector of categorical variables for each subject along with a qualitative variable indicating membership to a specific group, common in many application.

Aim: testing group differences in the whole set of measured variables.

Some used approaches are

- Latent class analysis: useful simplification but introduce systematic bias.
- Non parametric combination tests: use the dependence structure but can not consider changes beyond marginals.

We propose a **hierarchical Bayesian model** that allows

- **Global test**: all set of measured variables varies across groups?
- **Local test**: which variables are responsible for such variation?
- **Conditional pairwise association**: what is the pairwise association among the variables in each group?

Dependent mixture of tensor factorization

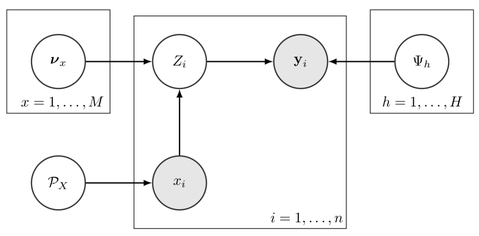

Figure 1.
Graphical representation of the probabilistic mechanism generating the data (y_i, x_i) under factorization 1 and 2.

- $x_i \in \{1, \dots, M\}$ group indicator generated from a discrete random variable with pmf p_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 k element of y_i is generated from a discrete distribution $\psi_h^{(k)}$ element of the tensor Ψ_h.

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

$$p_{\mathcal{Y},X}(y, x) = p_{\mathcal{Y}|X}(y)p_X(x) = \text{pr}(Y_1 = y_1, \dots, Y_p = y_p \mid x)\text{pr}(X = x) \quad (1)$$

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

A crucial point is to propose a **flexible model** for the conditional pmf $p_{\mathcal{Y}|X}(y)$ and we use the following **mixture of tensors**

$$p_{\mathcal{Y}|X} = \sum_{h=1}^{H} \nu_{hx}\Psi_h, \quad \Psi_h = \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)} \quad (2)$$

Any collection of group dependent probability mass function $p_{\mathcal{Y}|X}$ can be expressed as 2 for some H, and this ensures that any joint distribution $p_{\mathcal{Y},X}$ can be characterized using 1 and 2.

Prior specification & tests

We specify independent prior distributions for the group component p_X and the for the conditional probability tensor $p_{Y|X}$ and we use all conjugate priors to have a simple Gibbs sampling.

In our specification, global hypothesis reduces to

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

We include the test directly in the model through the following prior specification [2]

$$\begin{aligned} \nu_x &= (1 - T)u + Tu_x \\ u &\sim \text{Dirichlet}\{\gamma_1, \dots, \gamma_H\}, \quad u_x \sim \text{Dirichlet}\{\gamma_1, \dots, \gamma_H\}, \quad x = 1, \dots, M \\ T &\sim \text{Ber}\{\text{pr}(H_1)\} \end{aligned}$$

where, T is a hypothesis indicator, with $T = 0$ for H_0 and $T = 1$ for H_1.

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

For local and conditional tests we use a model based Cramér V [1]

$$\rho_k^2 = \frac{1}{\min\{M, d_k\} - 1} \sum_{x=1}^{M} p_X(x) \sum_{y=1}^{d_k} \frac{(p_{\mathcal{Y}|X}(y) - p_{\mathcal{Y}}(y))^2}{p_{\mathcal{Y}}(y)}.$$

We approximate the point null hypothesis of independence using a small interval hypothesis

$$H_{0k} : \rho_k \leq \epsilon \quad \text{vs} \quad H_{1k} : \rho_k > \epsilon$$

Data

We analyzed data from American national election survey **ANES 2012 face-to-face interview**.

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

Application to election data

- We reject the global null hypothesis ($\hat{\text{pr}}(H_1 \mid \text{data}) > 0.99$)
- Setting $\epsilon = 0.1$ and rejecting the local null hypothesis if the posterior probability of H_{1k} is > 0.99, we have that all the variables concerning Obama vary across the groups while those concerning Romney do not.

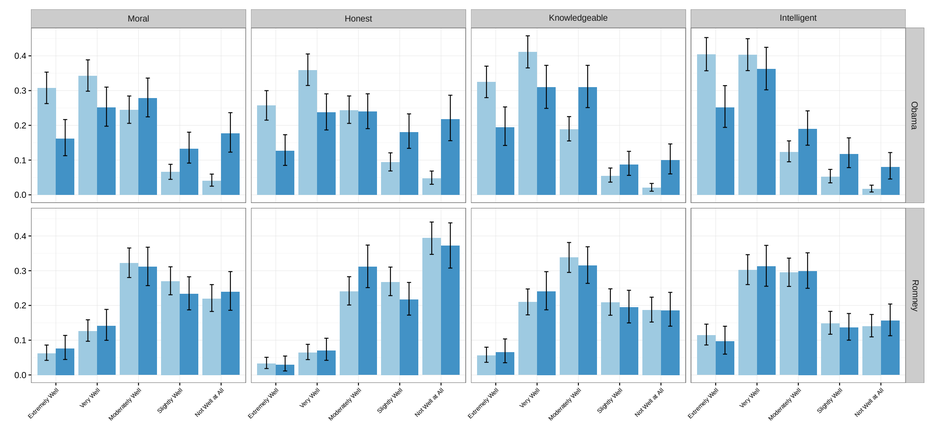

Figure 2. Posterior mean and 0.90 credible interval for the opinion on Romney and Obama for independents voters (dark) and independent democrat (light).

- Both groups have neutral or bad opinion about Romney.
- Independent democrats have better opinions about Obama personality and political decisions than unaffiliated voters.

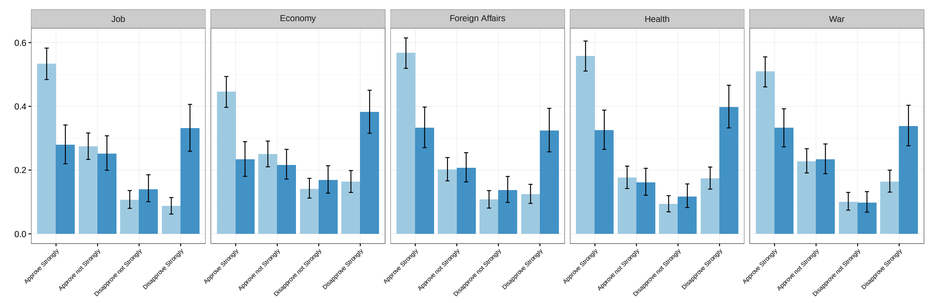

Figure 3. Posterior mean and 0.90 credible interval for the opinion abut Obama previous term for independents voters (dark) and independent democrat (light).

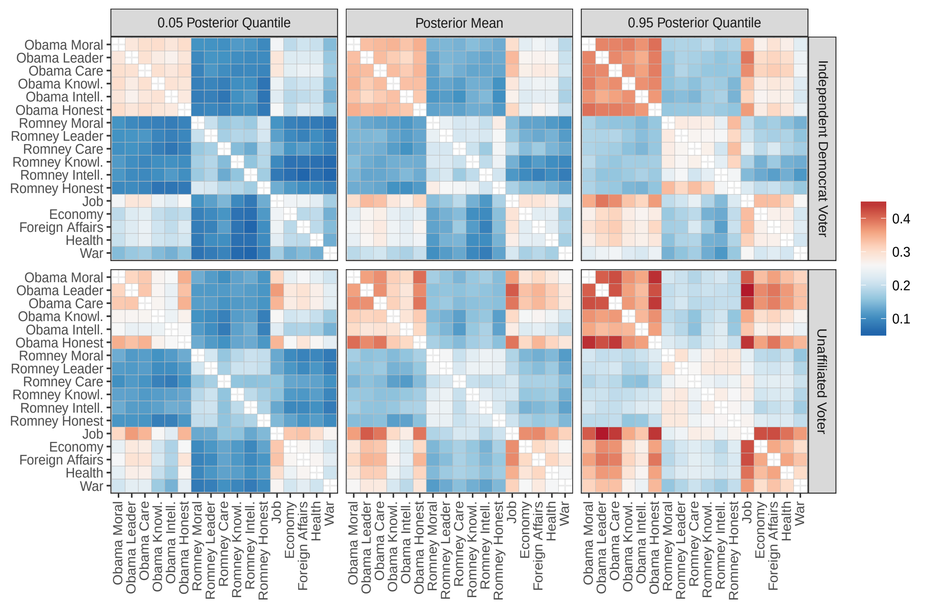

Figure 4. Pairwise Cramér V mean and quantiles for the two groups.

- Opinions about Romney do not influence opinions about Obama and do not seem to discriminate democratic and unaffiliated voters.
- Opinions about Obama personality and its political decision are related, especially for unaffiliated voters.

References

[1] Dunson, D. B. & Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**, 1042–1051.

[2] Lock, E. F. & Dunson, D. B. (2015). Shared kernel bayesian screening. *Biometrika* **102**, 829–842.

[3] Pesarin, F. & Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

[4] Russo, M., Durante, D. & Scarpa, B. (2017). Bayesian inference on group differences in multivariate categorical data. *arXiv preprint arXiv:1606.09415* .

[5] Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved + three-step approaches. *Political analysis* , 450–469.