Computational Statistics and Data Analysis 126 (2018) 136–149

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian inference on group differences in multivariate categorical data

Massimiliano Russo [a,*], Daniele Durante [b], Bruno Scarpa [a]

[a] Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy
[b] Department of Decision Sciences and Bocconi Institute for Data Science and Analytics, Bocconi University, Via Roentgen 1, 20136 Milano, Italy

ARTICLE INFO

Article history:
Received 9 August 2017
Received in revised form 19 January 2018
Accepted 24 April 2018
Available online 9 May 2018

Keywords:
Bayesian hypothesis testing
Election poll
Multivariate categorical data
Tensor factorization

ABSTRACT

Multivariate categorical data are common in many fields. An illustrative example is provided by election polls studies assessing evidence of changes in voters' opinions with their candidates preferences in the 2016 United States Presidential primaries or caucuses. Similar goals arise in routine applications, but current literature lacks a general methodology which combines flexibility, efficiency, and tractability in testing for group differences in multivariate categorical data at different – potentially complex – scales. This contribution addresses such goal by leveraging a Bayesian representation, which factorizes the joint probability mass function for the group variable and the multivariate categorical data as the product of the marginal probabilities for the groups and the conditional probability mass function of the multivariate categorical data, given the group membership. To enhance flexibility, the conditional probability mass function of the multivariate categorical data is defined via a group-dependent mixture of tensor factorizations which facilitates dimensionality reduction and borrowing of information, while providing tractable procedures for computation, and accurate tests assessing global and local group differences. The proposed methods are compared with popular competitors, and the improved performance is outlined in simulations and in American election polls studies.

1. Introduction

Multivariate categorical data arise frequently in relevant fields of application. Notable examples include epidemiology (e.g. Landis et al., 1988), psychology (e.g. Muthen and Christoffersson, 1981), social science (e.g. Santos et al., 2015), and business intelligence (e.g. Bijmolt et al., 2004)—among others. In such settings it is increasingly common to observe a vector of categorical responses for each subject, along with a qualitative variable indicating membership to a specific group. For example, in psychological studies a vector of categorical traits is typically measured for each individual, and the focus is on studying differences in these traits across groups, such as gender or level of education (e.g. Shao et al., 2014). We are specifically motivated by election polls studies measuring changes in voters' opinions with their preferences for the Presidential candidates, expressed in the primaries or caucuses of the 2016 United States Presidential elections. These elections have attracted a considerable interest by the political scientists – mainly due to the striking and partially unpredicted results – thereby motivating ongoing attempts to understand the determinants underlying the final outcomes. Most of the available political analyses provide qualitative explanations for the effect of the media, and the effectiveness of

[*] Corresponding author.
E-mail address: russo@stat.unipd.it (M. Russo).

https://doi.org/10.1016/j.csda.2018.04.010
0167-9473/© 2018 Elsevier B.V. All rights reserved.

Table 1
Opinions on several political topics collected from voters during the 2016 American national elections, along with their preference for Hillary Clinton or Bernie Sanders in the 2016 Democratic Presidential primaries.

	VOTER 15	VOTER 16	. . .
Preference primaries x_i	`Hillary Clinton`	`Bernie Sanders`	. . .
Political opinions $\boldsymbol{y}_i = (y_{i1}, \dots, y_{ip})^{\mathrm{T}}$			
Clinton made you FEEL ANGRY	`Never`	`Never`	. . .
. . .	. . .	. . .	. . .
Clinton made you FEEL DISGUSTED	`Never`	`Never`	. . .
How well the expression "has a strong LEADERSHIP" describes Clinton	`Extremely well`	`Very well`	. . .
. . .	. . .	. . .	. . .
How well the expression " SPEAKS MIND" describes Clinton	`Extremely well`	`Very well`	. . .
Trump made you FEEL ANGRY	`Always`	`About half the time`	. . .
. . .	. . .	. . .	. . .
Trump made you FEEL DISGUSTED	`Always`	`Always`	. . .
How well the expression "has a strong LEADERSHIP" describes Trump	`Not well at all`	`Not well at all`	. . .
. . .	. . .	. . .	. . .
How well the expression " SPEAKS MIND" describes Trump	`Extremely well`	`Very well`	. . .

the different campaigns and supported policies—among others. Refer to Lilleker et al. (2016) for a careful summary of the most valuable studies and comments.

Although all the above explanations allow important insights, quantitative assessments providing empirical evidence of the suggested conclusions in the light of the observed polls data are fundamental to improve the current understanding of the determinants underlying the 2016 United States Presidential elections. However, such contributions are still lacking. This is mainly due to the only recent availability of relevant datasets, along with the broad variability of the research interests characterizing the 2016 United States Presidential elections. In this contribution, our overarching goal is to assess evidence of differences in political opinions between the subset of voters who chose Hillary Clinton as Presidential candidate, and the one opting for Bernie Sanders in the 2016 Democratic Presidential primaries. There is, in fact, a common perception in the media that Bernie Sanders may have been a more effective candidate for the Democratic party in the Presidential campaign against Donald Trump (e.g. Lilleker et al., 2016).

We address the above goal with a main interest on how the voters' feelings toward Hillary Clinton and Donald Trump, along with their evaluations on specific personality traits of the two Presidential candidates, change between Hillary Clinton and Bernie Sanders voters in the 2016 Democratic Presidential primaries. The data are obtained from the American National Election Studies, and comprise five different feelings along with five specific personality traits for each of the two Presidential candidates, thereby providing a total of $p = 20$ categorical opinions measured for $n = 953$ potential voters. The opinions of each voter i are collected in a vector $\boldsymbol{y_i} = (y_{i1}, \dots, y_{ip})^{\mathrm{T}}$, jointly with a group indicator $x_i \in (1, 2)$, where $x_i = 1$ indicates a preference for Hillary Clinton and $x_i = 2$ for Bernie Sanders, for $i = 1, \dots, n$. Specifically, there are $n_1 = 567$ voters who expressed their preference for Hillary Clinton and $n_2 = 386$ who chose Bernie Sanders. Letting $\boldsymbol{Y}$ and X be the random variables generating data $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ and $x_1, \dots, x_n$, respectively, our overarching focus is on providing inference on the dependence between $\boldsymbol{Y}$ and X, and on learning how the conditional distribution of $\boldsymbol{Y}$ given $X = x$ changes with x. Table 1 provides an overview of the data under study which are publicly available at http://electionstudies.org/, along with the corresponding questionnaire and codebook files. According to Table 1, it is not clear – a priori – whether there exist group differences in the voters' opinions, and, if present, whether these differences are found in the entire vector of the $p = 20$ categorical variables, or only on a subset of the marginals or higher-order structures—including the bivariates, and more complex joint combinations. Obtaining statistical evidence of these differences at multiple scales can provide interesting insights on how marginal, bivariate, or more complex joint opinions of the voters change with their preference for Hillary Clinton or Bernie Sanders in the 2016 Democratic Presidential primaries.

There is a wide interest in studying differences in political opinions across groups of voters defined by gender (e.g. Atkeson and Rapoport, 2003), race (e.g. Brown, 2009), and affiliation party (e.g. Finkel and Scarrow, 1985)—among others. In accomplishing this goal, a widely used approach proceeds by summarizing the multivariate categorical data into a single latent class membership variable, while testing for group differences in these latent classes (Bolck et al., 2004). Although latent class analysis provides a useful simplification, the procedures required to perform the above test are subject to systematic bias, and it is still an active area of research to improve this method (e.g. Vermunt, 2010).

An alternative procedure is to avoid data reduction by assessing evidence of group differences in each categorical variable via separate χ^2 tests, while accounting for multiple testing via false discovery rate control (e.g Benjamini and Hochberg, 1995). These methodologies do not incorporate dependence structures among the p categorical variables, and therefore have low power. Pesarin and Salmaso (2010) addressed this issue via permutation tests preserving the dependence structure in the multivariate categorical data. Although this contribution provides a possible solution, the proposed methods cannot