Accepted Manuscript

A family of block-wise one-factor distributions for modeling high-dimensional binary data

Matthieu Marbac, Mohammed Sedki

PII: S0167-9473(17)30093-2
DOI: http://dx.doi.org/10.1016/j.csda.2017.04.010
Reference: COMSTA 6458

To appear in: *Computational Statistics and Data Analysis*

Received date: 18 July 2016
Revised date: 26 April 2017
Accepted date: 27 April 2017

Please cite this article as: Marbac, M., Sedki, M., A family of block-wise one-factor distributions for modeling high-dimensional binary data. *Computational Statistics and Data Analysis* (2017), http://dx.doi.org/10.1016/j.csda.2017.04.010

A Family of Block-wise One-Factor Distributions for modeling High-Dimensional Binary Data

Matthieu Marbac[†] and Mohammed Sedki[*]
[†] Modal project-team, INRIA Lille, France
[*] UMR Inserm-1181/Université Paris-Sud

April 26, 2017

Abstract

A new family of one-factor distributions for modeling high-dimensional binary data is introduced. The model provides an explicit probability for each event, thus avoiding the numeric approximations often made by existing methods. Model interpretation is easy, because each variable is described by two continuous parameters (corresponding to the marginal probability and to the strength of dependency with the other variables) and by one binary parameter (defining if the dependencies are positive or negative). This model is extended by splitting the variables into independent blocks, where each block follows the new one-factor distribution. Finally, a parsimonious version of the model, forcing some equality constraints between the dependency parameters, is proposed. Parameter estimation is carried out by an inference margin procedure, where the second step is achieved by an expectation-maximization algorithm. Model selection is performed by a deterministic approach, which strongly reduces the number of competing models. This consistent approach uses a hierarchical ascendant classification of the variables which selects a narrow subset of models. This selection is based on the empirical version of Cramer's V. The new model is evaluated on numerical experiments and on a real data set. The procedure is implemented in the R package MvBinary.

Keywords: Binary data, EM algorithm, High-dimensional data, IFM procedure, Model selection, One-factor copulas.

1 Introduction

Binary data (Cox and Snell, 1989; Collett, 2002) are increasingly emerging in various research fields, like in economics (Perez et al., 2015; Hoderlein and Sherman, 2015), psychometrics (Sorensen, 2015; Brusco and Steinley, 2006) or in life sciences (Marbac et al., 2016). Binary datasets often contain many variables because they are easy to access and

[*]Corresponding adress: U1181/Bât 15/16, 16 avenue P.V. Couturier, 94807 Villejuif, France. E-mail adress: mohammed.sedki@u-psud.fr.