Accepted Manuscript

A general algorithm for covariance modeling of discrete data

Gordana C. Popovic, Francis K.C. Hui, David I. Warton

PII: S0047-259X(17)30752-2
DOI: https://doi.org/10.1016/j.jmva.2017.12.002
Reference: YJMVA 4311

To appear in: *Journal of Multivariate Analysis*

Received date : 1 April 2015

Please cite this article as: G.C. Popovic, F.K.C. Hui, D.I. Warton, A general algorithm for covariance modeling of discrete data, *Journal of Multivariate Analysis* (2017), https://doi.org/10.1016/j.jmva.2017.12.002

A general algorithm for covariance modeling of discrete data

Gordana C. Popovic[a,1,*], Francis K.C. Hui[b], David I. Warton[a,c,2]

[a]School of Mathematics and Statistics, The University of New South Wales, NSW 2052, Australia
[b]Mathematical Sciences Institute, The Australian National University, Acton, ACT 2601, Australia
[c]Evolution and Ecology Research Centre, The University of New South Wales, NSW 2052, Australia

Abstract

We propose an algorithm that generalizes to discrete data any given covariance modeling algorithm originally intended for Gaussian responses, via a Gaussian copula approach. Covariance modeling is a powerful tool for extracting meaning from multivariate data, and fast algorithms for Gaussian data, such as factor analysis and Gaussian graphical models, are widely available. Our algorithm makes these tools generally available to analysts of discrete data and can combine any likelihood-based covariance modeling method for Gaussian data with any set of discrete marginal distributions. Previously, tools for discrete data were generally specific to one family of distributions or covariance modeling paradigm, or otherwise did not exist. Our algorithm is more flexible than alternate methods, takes advantage of existing fast algorithms for Gaussian data, and simulations suggest that it outperforms competing graphical modeling and factor analysis procedures for count and binomial data. We additionally show that in a Gaussian copula graphical model with discrete margins, conditional independence relationships in the latent Gaussian variables are inherited by the discrete observations. Our method is illustrated with a graphical model and factor analysis on an overdispersed ecological count dataset of species abundances.

Keywords: Factor analysis, Gaussian copula, graphical model, overdispersed count data, species interaction.

1. Introduction

Models for covariance give us valuable information about the structure of multivariate data when there are a large number of response variables, and the literature on such tools for Gaussian data is quite advanced. Gaussian graphical models [2, 13, 28, 36, 41] for example, describe conditional independence relationships between variables, which can be used to distinguish between direct and indirect relationships among variables. Factor analysis models can identify latent factors which drive the covariance between variables [9]. In addition, covariance modeling of Gaussian data is a fast moving field, with interesting algorithms continually being developed, including sparse factor analysis [4] and latent variable graphical model [29]. These and other covariance modeling methods were developed in the context of Gaussian data, and equivalent algorithms for discrete data are often limited or do not exist. In this article, we aim to develop a flexible method to apply covariance models to discrete data, with particular focus on overdispersed counts, our motivating example.

Covariance modeling of discrete data has been advanced separately for each covariance modeling paradigm, and these advances generally allow only a narrow class of discrete distributions. In the context of factor analysis, for binomial and multinomial data, item response theory allows limited latent variable modeling [15]. Counts and categorical outcomes can be modeled using, for example, generalized latent variable models [18, 37], a flexible covariance modeling method. These models combine generalized linear mixed models and structural equation models into a unifying

*Corresponding author
Email addresses: g.popovic@unsw.edu.au (Gordana C. Popovic), fhui28@gmail.com (Francis K.C. Hui), david.warton@unsw.edu.au (David I. Warton)

[1]GCP is supported by an Australian postgraduate award from the University of New South Wales
[2]DIW is supported by Australian Research Council Discovery Projects and Future Fellow funding schemes (project number DP130102131 and FT120100501)