ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian multidimensional scaling procedure with variable selection

L. Lin [a,*], D.K.H. Fong [b]

[a] Department of Statistics, The Pennsylvania State University, University Park, USA
[b] Smeal College of Business, The Pennsylvania State University, University Park, USA

ARTICLE INFO

Article history:
Received 18 December 2017
Received in revised form 14 July 2018
Accepted 15 July 2018
Available online xxxx

Keywords:
Bayesian multidimensional scaling
Variable selection
Model selection
Markov chain Monte Carlo

ABSTRACT

Multidimensional scaling methods are frequently used by researchers and practitioners to project high dimensional data into a low dimensional space. However, it is a challenge to integrate side information which is available along with the dissimilarities to perform such dimension reduction analysis. A novel Bayesian integrative multidimensional scaling procedure, namely Bayesian multidimensional scaling with variable selection, is proposed to incorporate external information on the objects into the analysis through the use of a latent multivariate regression structure. The proposed Bayesian procedure allows the incorporation of covariate information into the dimension reduction analysis through the use of a variable selection strategy. An efficient computational algorithm to implement the procedure is also developed. A series of simulation experiments and a real data analysis are conducted, and the proposed model is shown to outperform several benchmark models based on some measures commonly used in the literature.

1. Introduction

Multidimensional scaling (MDS) is arguably one of the most commonly used dimension reduction methods that projects high dimensional data into a low dimensional space through the use of (dis)similarity matrix (Torgerson, 1952; Kruskal, 1964; Kruskal and Wish, 1978). The method greatly facilitates data visualization and exploratory studies. More importantly, MDS allows investigators to uncover the underlying dimensions that contribute to the judgment of (dis)similarity and as a result to gain significant insights through the analysis. In this paper, we study the problem of integrating the (dis)similarity matrix with information from a set of external variables to perform dimension reduction. Such integrative analysis problems can be found in many application areas. For example, in clinical studies, "lab results" (primary data) such as subjects' microbiota profiles or their transcriptomic data are commonly used to define pairwise dissimilarities among subjects in a MDS analysis (e.g., Lahti et al. (2014); Zhang et al. (2016)). Furthermore, in most of the cases, there are also side information (external variables) on each subject such as gender, ethnicity, age, Body Mass Index, etc. that a researcher would like to include in the analysis to obtain the MDS solution. These external variables typically have different measurement scales from the primary data and many of them may not be discriminative factors to differentiate subjects. As an example, subjects cannot be diagnosed solely based on their gender. However, these variables can be helpful to better stratify the subjects when performing dimension reduction based on the primary data (e.g., Chaturvedi (2003); Gottlieb et al. (2004); O'Hare et al. (2007); Ngo et al. (2014)). Note, the task of incorporating such side information into the analysis is different from the problem dealing with multiple input distance matrices. Under the context of our example here, the latter is concerned with

* Correspondence to: Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA.
E-mail address: llin@psu.edu (L. Lin).

https://doi.org/10.1016/j.csda.2018.07.007
0167-9473/

the procedure of combining both microbiota profiles and transcriptomic data to form the distance matrix (e.g., Ibba et al. (2010); Konukoglu et al. (2013); Bai et al. (2016)).

MDS has found many applications for scientific studies. For example, in psychological research, MDS has been used to construct personality profiles where similarities between individuals were analyzed to uncover how (groups of) individuals differ with respect to their profiles (Kim et al., 2004; Allik and McCrae, 2004; Ding, 2006). In bioinformatics, MDS has been applied to visualize different biological data types so that differences between different biological conditions can be easily identified and visualized (Chen and Meltzer, 2002; Tzeng et al., 2008; Park et al., 2012; Malaspinas et al., 2014). In addition to visualization, MDS can also help in uncovering the underlying dimensionality associated with the dissimilarities. For example, MDS can help in marketing research to answer the most common question: what product attributes and features most contribute to a product's position in the market share (Carroll and Green, 1997).

More specifically, MDS aims to find a mapping in low dimensional space that preserves distances between pairs of data points in high dimensional space. That is MDS utilizes the information about the global dissimilarity or similarity of the data points, and then selects the low-dimensional space that is closest to the pairwise distances (dissimilarities) in the original space in terms of euclidean distance. In certain applications, there is no strict metric on the data points. For example, ranking in survey data is more important than its absolute value. Non-metric MDS, which was first suggested by Shepard (1962), then tries to find the low-dimensional space that follows closely the rank orders of the original data points. Both metric and non-metric MDS analyses output a spatial configuration, in which the data points in the original space are represented as lower-dimensional points. The points are arranged such that similar data points are represented by points that are close to each other, and dissimilar data points are represented by points that are further apart. For more discussions and references, please see Borg and Groenen (2005) for useful textbook accounts.

The earliest practical MDS method, which is called Classical Multidimensional Scaling (CMDS), is proposed by Torgerson (1952). CMDS assumes the distances to be Euclidean, and utilizes the principal components of the double-centered distance matrix to obtain the projected low dimensional space. A more popular method, proposed by Kruskal (1964), was defined in terms of minimizing a loss function called "STRESS". The stress function is a weighted sum of squared errors between the observed distances and the inter-vector distances in the low dimensional space. This estimation technique, which amounts to minimizing least squares, can be sensitive to outliers, because such loss function assumes constant Gaussian errors (e.g., Spence and Lewandowsky (1989)).

Oh and Raftery (2001) have developed a Bayesian MDS (BMDS), which assumes a Gaussian distribution with constant variance for the dissimilarity data. They provide a Bayesian solution for the projected low dimensional space by using a Markov chain Monte Carlo (MCMC) method. It is found that BMDS is more accurate in fitting some data than did CMDS, especially when the data contain significant measurement errors, or when the Euclidean distance which is assumed by CMDS is not satisfied, or when the latent dimensionality was incorrectly specified. One appeal of the Bayesian modeling approach is that it provides great flexibility in incorporating external information through prior specification. The priors can be either informed by historical or other relevant data ("informative prior") or left diffuse ("uninformative prior"). However, BMDS was not developed specifically to incorporate additional covariates (external information) into the analysis.

In this paper, we develop a more general Bayesian MDS procedure, namely Bayesian MDS with Variable Selection (BMDS-VS). First, the assumption of constant Gaussian measurement error variance in the observed dissimilarity is removed. Hence, we allow different error variances for different pairwise dissimilarities which can accommodate different degrees of uncertainty associated with the observed dissimilarity. Second, our model allows the incorporation of additional covariate information with a variable selection option to select the most important covariates that help to better explore low dimensional spaces. We believe our method provides a realistic modeling approach for real data analysis. Lastly, most existing MDS methods do not provide an explicit function that maps data from the original space to the latent map, hence it is not possible to directly make prediction on new data set. BMDS-VS can potentially learn a mapping to the latent space by leveraging on existing side information. This enhancement could lead to a more accurate prediction. We apply BMDS-VS to two simulation studies and one real data analysis, and demonstrate the ability of BMDS-VS to yield better solutions compared with some benchmark models.

The rest of this paper is organized as follows. In Section 2, we introduce notations and provide an overview of existing methods that are most relevant to our proposed model. In Section 3, we describe the proposed Bayesian MDS model and the corresponding estimation procedures: Section 3.1 defines the model and the prior, Section 3.2 outlines an MCMC algorithm for the model update, and describes posterior summary procedures. Section 3.3 describes the dimension selection criteria for determining the optimal number of dimensions. In Section 4, we present the results of simulation studies to examine the performance of several competing models with known data structures and parameters. Section 5 gives empirical application analyzing one microbiome data set. Finally, we summarize our contributions and discuss future research directions in Section 6.

2. Preliminaries

We start by representing data in a general form, with the following notations and definitions. Consider a set of pairwise dissimilarities d_{ij} between n measured objects, where $d_{ij} \in \mathbb{R}$ measures the dissimilarity between objects i and j, for $i, j \in \{1, \ldots, n\}$. There are three properties associated with d_{ij}: (1) Non-negativity: $d_{ij} \geq 0$; (2) Symmetry: $d_{ij} = d_{ji}$; and (3) $d_{ii} = 0$ for all i. Hence, the total number of distinct dissimilarities among n objects is $m = n(n-1)/2$. Let D denote

Please cite this article in press as: Lin L., Fong D.K.H., Bayesian multidimensional scaling procedure with variable selection. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.07.007.