Computational Statistics and Data Analysis 124 (2018) 220–234

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Overfitting Bayesian mixtures of factor analyzers with an unknown number of components

Panagiotis Papastamoulis

University of Manchester, Faculty of Biology, Medicine and Health, Division of Informatics, Imaging and Data Sciences, Michael Smith building, Oxford Road, M13 9PL, Manchester, UK

ARTICLE INFO

Article history:
Received 4 June 2017
Received in revised form 26 February 2018
Accepted 8 March 2018
Available online 27 March 2018

Keywords:
Factor analysis
Mixture models
Clustering
MCMC

ABSTRACT

Recent advances on overfitting Bayesian mixture models provide a solid and straightforward approach for inferring the underlying number of clusters and model parameters in heterogeneous datasets. The applicability of such a framework in clustering correlated high dimensional data is demonstrated. For this purpose an overfitting mixture of factor analyzers is introduced, assuming that the number of factors is fixed. A Markov chain Monte Carlo (MCMC) sampler combined with a prior parallel tempering scheme is used to estimate the posterior distribution of model parameters. The optimal number of factors is estimated using information criteria. Identifiability issues related to the label switching problem are dealt by post-processing the simulated MCMC sample by relabeling algorithms. The method is benchmarked against state-of-the-art software for maximum likelihood estimation of mixtures of factor analyzers using an extensive simulation study. Finally, the applicability of the method is illustrated in publicly available data.

1. Introduction

Factor Analysis (FA) is a popular statistical model that aims to explain correlations in a high-dimensional space by dimension reduction. This is typically achieved by expressing the observed multivariate data as a linear combination of a smaller set of hypothetical and uncorrelated variables known as factors. The factors are not observed, so they are treated as missing data. The reader is referred to Kim and Mueller (1978) and Bartholomew et al. (2011) for an overview of factor analysis models, estimation techniques and applications.

However, when the observed data is not homogeneous, the typical FA model will not adequately fit the data. In such a case, a Mixture of Factor Analyzers (MFA) can be used in order to take into account the underlying heterogeneity. Thus, MFA models jointly treat two inferential tasks: model-based density estimation for high dimensional data as well as dimensionality reduction. Estimation of MFA models is straightforward by using the Expectation–Maximization (EM) algorithm (Dempster et al., 1977; Ghahramani and Hinton, 1996; McLachlan and Peel, 2000; McLachlan et al., 2003, 2011). The family of parsimonious Gaussian mixture models (PGMM) is introduced in McNicholas and Murphy (2008), McNicholas et al. (2010) and McNicholas and Murphy (2010), which is based on Gaussian mixture models with parsimonious factor analysis like covariance structures. Under a Bayesian setup, Fokoué and Titterington (2003) estimate the number of mixture components and factors by simulating a continuous-time stochastic birth–death point process using a Birth–Death MCMC algorithm (Stephens, 2000a). Their algorithm is shown to perform well in small to moderately scaled multivariate data.

Fully Bayesian approaches to estimate the number of components in a mixture model include the Reversible jump MCMC (RJMCMC) (Green, 1995; Richardson and Green, 1997; Dellaportas and Papageorgiou, 2006; Papastamoulis and Iliopoulos, 2009), Birth–death MCMC (BDMCMC) (Stephens, 2000a) and allocation sampling (Nobile and Fearnside, 2007) algorithms.

E-mail address: panagiotis.papastamoulis@manchester.ac.uk.

https://doi.org/10.1016/j.csda.2018.03.007
0167-9473/© 2018 Elsevier B.V. All rights reserved.

In recent years there is a growing progress on the usage of overfitted mixture models in Bayesian analysis (Rousseau and Mengersen, 2011; van Havre et al., 2015). An overfitting mixture model consists of a number of components which is much larger than its true (and unknown) value. Under a frequentist approach, overfitting mixture models is not a recommended practice. In this case, the true parameter lies on the boundary of the parameter space and identifiability of the model is violated due to the fact that some of the component weights can be equal to zero or some components may have equal parameters. Consequently, standard asymptotic Maximum Likelihood theory does not apply in this case (Li et al., 1988). Choosing informative prior distributions that bound the posterior away from unidentifiability sets can increase the stability of the MCMC sampler, however these informative priors tend to force too many distinct components and the possibility of reducing the overfitting mixture to the true model is lost (see section 4.2.2 in Frühwirth-Schnatter (2006)). Under suitable prior assumptions introduced by Rousseau and Mengersen (2011), it has been shown that asymptotically the redundant components will have zero posterior weight and force the posterior distribution to put all its mass in the sparsest way to approximate the true density. Therefore, the inference on the number of mixture components can be based on the posterior distribution of the "alive" components of the overfitted model, that is, the components which contain at least one allocated observation.

The simplicity of this approach is in stark contrast with the fully Bayesian approach of treating the number of clusters as a random variable. For example, in the RJMCMC algorithm the researcher has to design sophisticated move types that bridge models with different number of clusters. On the other hand, the allocation sampler is only applicable to cases where the model parameters can be analytically integrated out. Even in such cases though, the design of proper Metropolis–Hastings moves on the space of latent allocation variables of the mixture model is required to obtain a reasonable mixing of the simulated MCMC chain (see Nobile and Fearnside (2007); Papastamoulis and Rattray (2017)).

The contribution of this study is to utilize recent advances on overfitting mixture models (van Havre et al., 2015) to the context of Bayesian MFA (Fokoué and Titterington, 2003). We use a Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) which is embedded in a prior parallel tempering scheme in order to improve the mixing of the algorithm. In addition, we explore the usage of information criteria for estimating the number of factors. After estimating the number of clusters and factors, we perform inference on the chosen model by dealing with identifiability issues related to the label switching problem (Papastamoulis, 2016). Our results indicate that overfitting Bayesian MFA models provide a simple and efficient approach to estimate the number of clusters in correlated high-dimensional data.

The rest of the paper is organized as follows. Section 2.1 reviews the basics of FA models. Finite mixtures of FA models are presented in Section 2.2 and a brief review of previous frequentist approaches is given in Section 2.3. The Bayesian formulation is presented in Section 2.4. The overfitting MFA model is introduced in Section 2.5. Section 2.6 deals with estimating the number of factors using information criteria. Section 2.7 presents the prior parallel tempering scheme which is incorporated into the MCMC sampler. Identifiability issues related to the label switching phenomenon are discussed in Section 2.8 and further details of the overall implementation are given in Section 2.9. Our method is illustrated and compared against the EM algorithm in Section 3 using a simulation study (Section 3.1) as well as three publicly available datasets (Sections 3.2–3.4). The paper concludes in Section 4. Further technical details and simulation results are provided in the Appendix.

2. Methodology

At first we introduce some conventional guidelines that will be followed in our notation throughout this paper, unless explicitly stated otherwise. We will use bold face for vectors and matrices. The notation α_k will correspond to the kth member of a vector $\boldsymbol{\alpha}$. In addition, $\mathbf{A}_k$ will denote the kth member of a vector $\mathbf{A}$ whose elements are matrices. The (i, j) element of a matrix $\boldsymbol{\Sigma}$ will be denoted by the corresponding lower case letter, that is, σ_{ij}. The transpose matrix of $\boldsymbol{\Sigma}$ will be denoted as $\boldsymbol{\Sigma}^T$. We will not differentiate the notation between random variables and their specific realizations. We use $f(x|y)$ to denote the probability mass or density function of x given y. For a discrete random variable z, the notation $\mathrm{P}(z = k)$ will be also used to denote the probability of the event $\{z = k\}$. The $p \times p$ identity matrix is denoted as $\mathbf{I}_p$, $p \in \mathbb{N}$.

2.1. Factor analysis model

Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ denote a random sample of p dimensional observations with $\boldsymbol{x}_i \in \mathbb{R}^p$; $i = 1, \ldots, n$. We assume that $\boldsymbol{x}_i$ is expressed as a linear combination of a latent vector (factors) $\boldsymbol{y}_i \in \mathbb{R}^q$

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_i + \boldsymbol{\varepsilon}_i. \tag{1}$$

The unobserved random vector $\boldsymbol{y}_i$ lies on a lower dimensional space, that is, $q < p$ and it consists of uncorrelated features $y_{i1}, \ldots, y_{iq}$. In particular, we assume that

$$\boldsymbol{y}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}_q), \tag{2}$$

independent for $i = 1, \ldots, n$ and $\mathbf{0}$ denotes a vector of zeros. The $p \times q$ dimensional matrix $\boldsymbol{\Lambda} = (\lambda_{rj})$ contains the factor loadings, while the p-dimensional vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ contains the marginal mean of $\boldsymbol{x}_i$. For the error term $\boldsymbol{\varepsilon}_i$ assume that

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \tag{3}$$