Journal of Multivariate Analysis 154 (2017) 112–134

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Nonparametric estimation of a latent variable model

Augustin Kelava[a], Michael Kohler[b], Adam Krzyżak[c,*], Tim Fabian Schaffland[a]

[a] *Wirtschafts-und Sozialwissenschaftliche Fakultät, Hector-Institut für Empirische Bildungsforschung, Universität Tübingen, Europastraße 6, 72072 Tübingen, Germany*

[b] *Fachbereich Mathematik, Technische Universität Darmstadt, Schloßgartenstraße 7, 64289 Darmstadt, Germany*

[c] *Department of Computer Science and Software Engineering, Concordia University, 1455, boul. de Maisonneuve ouest, Montréal, Québec, Canada H3G 1M8*

ARTICLE INFO

Article history:
Received 17 April 2015
Available online 1 November 2016

AMS subject classifications:
primary 62G08
secondary 62G20

Keywords:
Common factor analysis
Latent variables
Nonparametric regression
Consistency

ABSTRACT

In this paper a nonparametric latent variable model is estimated without specifying the underlying distributions. The main idea is to estimate in a first step a common factor analysis model under the assumption that each manifest variable is influenced by at most one of the latent variables. In a second step nonparametric regression is used to analyze the relation between the latent variables. Theoretical results concerning consistency of the estimates are presented, and the finite sample size performance of the estimates is illustrated by applying them to simulated data.

1. Introduction

Latent variable models provide statistical tool for explaining and analyzing underlying structure of multivariate data by using the idea that observable phenomena are influenced by underlying factors which cannot be observed or measured directly. They have applications in various areas including psychology, social sciences, education or economics, where theoretical concepts such as intelligence, desirability or welfare cannot be measured directly but instead observable indicators (or manifest variables) are given.

One possibility to fit latent variable models to data is to assume that the underlying distribution is Gaussian, and therefore it is uniquely determined by its covariance structure. Then the maximum likelihood principle together with structural assumptions on the underlying latent variable model can be used to fit the latent variable model to observed data.

In contrast in this paper we try to avoid any assumption on the class of the underlying distributions. Given multivariate random variables X and Y, we approximate them by linear combinations of suitable latent variables Z_1 and Z_2 and then use nonparametric regression to study the relation between Z_1 and Z_2. In this way the whole procedure splits into two separate problems: In a first step we fit a common factor analysis model to X and Y. And then we apply suitable nonparametric regression techniques to analyze the relation between the latent variables in this model.

[*] Corresponding author.
E-mail addresses: augustin.kelava@uni-tuebingen.de (A. Kelava), kohler@mathematik.tu-darmstadt.de (M. Kohler), krzyzak@cs.concordia.ca (A. Krzyżak), tim.schaffland@gmail.com (T.F. Schaffland).

http://dx.doi.org/10.1016/j.jmva.2016.10.006
0047-259X/

The main trick in estimation of the common factor analysis model is to estimate the values of (Z_1, Z_2) in such a way that the corresponding empirical distribution asymptotically satisfies the conditions that characterize the distribution of (Z_1, Z_2) uniquely. This primarily requires independence of (Z_1, Z_2) of the random errors occurring in the manifest variables, and we ensure this by minimizing some kind of distance between the empirical cumulative distribution function of all these random variables and the product of the marginal cumulative distribution functions.

Our main theoretical result is that the empirical distribution of the estimated values of (Z_1, Z_2) converges weakly, with probability 1, to the distribution of (Z_1, Z_2). We use this result to define the least squares estimates of the regression function of (Z_1, Z_2). We show that our regression estimate is strongly consistent whenever the regression function is Lipschitz-continuous and bounded. The finite sample size performance of our estimates is illustrated by applying them to simulated data.

1.1. Discussion of related results

Surveys on latent variables and its applications can be found, e.g., in [1,2].

One way to determine latent variable models is the use of principal component analysis; see, e.g., Section 14.5 in Hastie, Tibshirani and Friedman [3]. There the manifest variables are approximated by the best linear approximation of a given rank. The obvious drawback is that in this case the sum of the latent variable and its random error is approximated. The classical factor analysis model takes into account these random errors. If we assume that all random variables are Gaussian, then the model can be fitted by maximum likelihood; see, e.g., Section 14.7 in Hastie, Tibshirani and Friedman [3]. In the independent component analysis (described, e.g., in [4]) the latent variables are assumed to be independent, which resolves any identifiability problem in the above approaches. However, this assumption is often not realistic in applications and cannot be used in the context of regression estimation. Identifiability conditions for latent parameters in hidden Markov models and random graph mixture models have been discussed in [5–7].

Independent factor analysis model which is often used for dimensionality reduction assumes that random variables are generated by a linear model containing latent independent components and perturbed by an additive gaussian noise. The density of observed variables has been estimated by a kernel estimate by Amato et al. [8]. A linear latent variable model where observed variables depend linearly on unobservable latent variables has been analyzed in [9]. Under normality assumption the covariance structure of the model is estimated by maximum likelihood and its asymptotic normality is established. For ordered categorical data the latent variable model has been investigated by Breslaw and McIntosh [10] and by Gebregziabher and DeSantis [11] for missing categorical data. It has been applied to finance by Bai and Ng [12]. A generalized linear latent variable model (GLLVM) has been estimated using Laplace approximation by Bianconcini and Cagnone [13]. Similar model with semi-nonparametric specification of distribution of latent variables has been analyzed by Irincheeva, Cantoni and Genton [14]. Bartolucci, Pennoni and Francis [15] considered latent Markov model and estimated its parameters using EM algorithm and Bartolucci [16] applied it to detecting patterns of criminal activity.

A mixture of latent variables model was applied to clustering, classification and discriminant analysis; see [17]. Parsimonious Gaussian mixture models (PGMMs) are recently introduced model-based clustering techniques generalizing mixtures of factor analyzers model and are based on a latent Gaussian mixture model. McNicholas [18] used PGMM and Bayesian information criteria to perform model-based classification. A general latent variable model incorporating spatial correlation and shifted dependencies has been analyzed by Christensen and Amemiya [19]. Colombo et al. [20] applied latent variables to learning of high-dimensional acyclic graphs. In longitudinal data analysis one often encounters non-Gaussian data. Hall et al. [21] used a latent Gaussian process model for prediction by means of functional principal component analysis (PCA). The PCA approach has also been used to estimate latent variable models by Lynn and McCulloch [22]. In a model where the number of manifest variables is the same for all latent variables, and where this number and the number of observations of each of them increase, Bai and Ng [23] estimate the number of latent variables using an asymptotic principal component analysis.

The previous works on regression estimation in the context of latent variables were confined to parametric models, often formulated with so-called structural equations models; for surveys, see, e.g., Skrondal and Rabe-Hesketh [2] or Schumacker and Marcoulides [24]. In [25] a high-dimensional linear regression problem is considered, where a low dimensional latent variable model determines the response variable. Principal component analysis is used to estimate the underlying latent variables, and it is assumed that all variables have a Gaussian distribution. A generalization of Gaussian latent variable models to the case that the manifest variables are indirect observations of normal underlying variables can be done via generalized linear latent variable models; see, e.g., [26].

Our results generalize previously known results in so far that we do not need to impose any parametric structure on the regression function considered and that we do not restrict the class of error distributions occurring in the model. Our estimation of the common factor model is related to errors-in-variables models. In fact our estimation principle is based on generalization of the uniqueness result for such models presented in [27].

Nonparametric regression estimation has been studied in the literature for a long time. The most popular estimates for random design regression include kernel regression estimate (see, e.g., [28–33]), partitioning regression estimate (see, e.g., [34,35]), nearest neighbor regression estimate (see, e.g., [36–39]), least squares estimates (see, e.g., [40]) or smoothing spline estimates (see, e.g., [41]). The main theoretical results are summarized in the monograph by Györfi et al. [42]. To the best of the authors' knowledge, the application of nonparametric regression in the context of latent variables is new.