Nonparametric estimation of a latent variable model *

Augustin Kelava[1], Michael Kohler[2] and Adam Krzyżak[3†]

[1] *Wirtschafts-und Sozialwissenschaftliche Fakultät, Institut für Erziehungswissenschaft, Universität Tübingen, Europastr. 6, 72072 Tübingen, Germany, email: augustin.kelava@uni-tuebingen.de*

[2] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de, dennisweinbender@googlemail.com*

[3] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

February 24, 2014

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.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: Common factor analysis, latent variables, nonparametric regression, consistency.

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

*Running title: *Nonparametric latent variables*

†Corresponding author: Tel. +1 514 848 2424, ext. 3007, Fax. +1 514 848 2830

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.

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 one 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.

1.1 Discussion of related results

Surveys on latent variables and its applications can be found, e.g., in Bollen (2002) and Skrondal and Rabe-Hesketh (2007).

One way to determine latent variable models is the use of principal component analysis (c.f., e. g., Hastie, Tibshirani and Friedman (2009), ch. 14.5). 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 (c.f., e. g., Hastie, Tibshirani and Friedman (2009), ch. 14.7). In the independent component analysis (described e.g. in Montanari and Viroli (2010)) 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 the applications and cannot be used in context of regression estimation. Identifiability conditions for latent parameters in hidden Markov models and random graph mixture models have been discussed in Kruskal (1976, 1977) and Allman, Matias and Rhodes (2009). 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. (2010). A linear latent variable model where observed variables depend linearly on unobservable latent variables has been analyzed by Anderson (1989). Under normality assumptions 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 (1998) and by Gebregziabher and DeSantis (2010) for missing categorical data. It has been applied to finance by Bai and Ng (2006). A generalized linear latent variable model (GLLVM) has been estimated using Laplace approximation by Bianconcini and Cagnone (2012). Similar model with semi-nonparametric specification of distribution of latent variables has been analyzed by Irincheeva, Cantoni and Genton (2012). Bartolucci (2006) considered latent Markov model and estimated its parameters using EM algorithm and applied it to detecting patterns of criminal activity, see Bartolucci, Pennoni and Francis (2007). A mixture of latent variables model was applied to clustering, classification and discriminant analysis, see Browne and McNicholas (2012). 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 (2010) 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 (2002). Colombo et al. (2012) applied latent variables to learning of high dimensional acyclic graphs. In longitudinal data analysis one often encounters non-Gaussian data. Hall et al. (2008) used latent Gaussian process model for prediction by means of functional principal component analysis (PCA). PCA approach has also been used to estimate latent variable models by Lynn and McCuloch (2000). 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 (2002) 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., Marsh, Wen and Hau (2004) or Schumacker and Marcoulides (1998). In Paul et al. (2008) 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 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, cf., e.g., Conne, Ronchetti and Victoria-Feser (2010).

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 Li (2002).

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 (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977) or Devroye and Krzyżak (1989)), partitioning regression estimate (cf., e.g., Györfi (1981) or Beirlant and Györfi (1998)), nearest neighbor regression estimate (cf., e.g., Devroye (1982), Devroye, Györfi, Krzyżak and Lugosi (1994), Mack (1981) or Zhao (1987)), least squares estimates (cf., e.g., Lugosi and Zeger (1995)) or smoothing spline estimates (cf., e.g., Kohler and Krzyżak (2001)). The main theoretical results are summarized in the monograph by Györfi et al. (2002). To the best of authors' knowledge, the application of nonparametric regression in the context of latent variables is new.

1.2 Notation

Throughout this paper we use the following notation: the sets of integers, rational numbers and real numbers are denoted by $\mathbb{N}$, $\mathbb{Q}$ and $\mathbb{R}$, respectively. For $k \in \mathbb{N}$ and subsets $B_1, \dots, B_k$ of $\mathbb{R}^d$ we write

$$\prod_{i=1}^{k} B_i = \{(x_1, \dots, x_k) \quad : \quad x_i \in B_i \quad (i = 1, \dots, k)\}$$

for the Cartesian product of the sets. 1_B is the indicator of the set B. If X is $\mathbb{R}^d$-valued random variable then

$$\varphi_X(u) = \mathbf{E}\{e^{i \cdot u^T X}\}$$

is its characteristic function. For $f : D \to \mathbb{R}$ we write

$$x = \arg\min_{z \in D} f(z)$$

in case that

$$x \in D \quad \text{and} \quad f(x) = \min_{z \in D} f(z).$$

1.3 Outline

The estimate of the common factor analysis model is described in Section 2. In Section 3 we use techniques of nonparametric regression to analyze the relationship between the latent variables. The proofs are given in Section 4.

2 Estimation of a common factor analysis model

In the sequel X and Y are $\mathbb{R}^{d_X}$- and $\mathbb{R}^{d_Y}$-valued observable random variables (manifest variables). In order to analyze the relation between X and Y we assume that they depend linearly on some hidden and unobservable variables Z_1 and Z_2, where Z_1 and Z_2 are d_{Z_1}- and d_{Z_2}-dimensional random vectors, resp. Here we assume $d_{Z_1} < d_X$ and $d_{Z_2} < d_Y$. More precisely we assume that X and Y satisfy the following common factor analysis models

$$X = A \cdot Z_1 + \epsilon \tag{1}$$

and

$$Y = B \cdot Z_2 + \delta, \tag{2}$$

where A and B are $d_X \times d_{Z_1}$ and $d_Y \times d_{Z_2}$-dimensional matrices, resp., and ϵ and δ are d_X- and d_Y-dimensional random vectors where all components are independent and have mean zero, furthermore we assume that (Z_1, Z_2), ϵ and δ are independent. Given a sample

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

of independent and identically distributed copies of (X, Y), we want to estimate A, B and the corresponding values of the latent variables $Z_{1,i}$ and $Z_{2,i}$ corresponding to X_i and Y_i $(i = 1, \ldots, n)$. In the next section we will apply nonparametric regression to the estimated sample

$$\{(\hat{z}_{1,1}, \hat{z}_{2,1}), \ldots, (\hat{z}_{1,n}, \hat{z}_{2,n})\}$$

of (Z_1, Z_2) in order to analyze the relation between Z_1 and Z_2.

$$\begin{pmatrix} X_{1,1} \\ X_{1,2} \\ \vdots \\ X_{1,l_1} \\ \vdots \\ X_{d_{Z_1},1} \\ X_{d_{Z_1},2} \\ \vdots \\ X_{d_{Z_1},l_{d_{Z_1}}} \\ Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,k_1} \\ \vdots \\ Y_{d_{Z_2},1} \\ Y d_{Z_2}, 2 \\ \vdots \\ Y_{d_{Z_2},k_{d_{Z_2}}} \end{pmatrix} = \begin{pmatrix} 1 \cdot Z_{1,1} \\ a_{1,2} \cdot Z_{1,1} \\ \vdots \\ a_{1,l_1} \cdot Z_{1,1} \\ \vdots \\ 1 \cdot Z_{d_{Z_1},1} \\ a_{d_{Z_1},2} \cdot Z_{d_{Z_1},1} \\ \vdots \\ a_{d_{Z_1},l_{d_{Z_1}}} \cdot Z_{d_{Z_1},1} \\ 1 \cdot Z_{1,2} \\ b_{1,2} \cdot Z_{1,2} \\ \vdots \\ b_{1,k_1} \cdot Z_{1,2} \\ \vdots \\ 1 \cdot Z_{d_{Z_2},2} \\ b_{d_{Z_2},2} \cdot Z_{d_{Z_2},2} \\ \vdots \\ b_{d_{Z_2},k_{d_{Z_2}}} \cdot Z_{d_{Z_2},2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{l_1,1} \\ \vdots \\ \epsilon_{1,d_{Z_1}} \\ \epsilon_{2,d_{Z_1}} \\ \vdots \\ \epsilon_{l_{d_{Z_1}},d_{Z_1}} \\ \delta_{1,1} \\ \delta_{2,1} \\ \vdots \\ \delta_{k_1,1} \\ \vdots \\ \delta_{1,d_{Z_2}} \\ \delta_{2,d_{Z_2}} \\ \vdots \\ \delta_{k_{d_{Z_2}},d_{Z_2}} \end{pmatrix} \tag{3}$$

In this section we describe how to estimate the common factor analysis model described by (1) and (2). Here we assume that some a priori information on the structure of the matrices is given. More precisely, we assume a simple structure in terms of a single cause of variation (i.e., a single latent variable) for each manifest variables. In other words, each of the components of the manifest variables is influenced by at most one of the components of the latent variables, so that each row of A and B contains at most one nonzero entry. By rescaling the columns of the matrices and the latent variables we can assume furthermore that one of the entries in each column is one (which enables us to show that the model is uniquely defined, cf. Lemma 1 below). If this is true we can rewrite our model by (3), where we assume that $l_1, \ldots, l_{d_{Z_1}}, k_1, \ldots, k_{d_{Z_2}} \geq 3$.

In order to simplify the notation we assume throughout this paper $d_{Z_1} = d_{Z_2} = 1$, and consequently we can rewrite the model (1) and (2) in the form:

$$\begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(d)} \\ Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(l)} \end{pmatrix} = \begin{pmatrix} 1 \cdot Z_1 \\ a_2 \cdot Z_1 \\ \vdots \\ a_d \cdot Z_1 \\ 1 \cdot Z_2 \\ b_2 \cdot Z_2 \\ \vdots \\ b_l \cdot Z_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_d \\ \delta_1 \\ \delta_2 \\ \vdots \\ \delta_l \end{pmatrix} \tag{4}$$

where we assume that the coefficients are all nonzero, that $d, l \geq 3$, and that Z_1, Z_2, ϵ_1, $\ldots \epsilon_d$, δ_1, $\ldots$, δ_l are real random variables with the property that (Z_1, Z_2), ϵ_1, $\ldots \epsilon_d$, δ_1, $\ldots$, δ_l are independent and that satisfy $\mathbf{E}\{\epsilon_j\} = \mathbf{E}\{\delta_k\} = 0$.

Our first result shows that under the additional assumption that the characteristic function of

$$(X, Y) = (X^{(1)}, \ldots, X^{(d)}, Y^{(1)}, \ldots, Y^{(l)})$$

does not vanish at any point the distribution of (X, Y) determines uniquely the (joint) distribution of all other random variables occurring in the above model.

Lemma 1. *Assume that in the model (4) the random variables $X^{(1)}, \ldots, X^{(d)}$, $Y^{(1)}, \ldots, Y^{(l)}$ are in L_2, that Z_1, Z_2, ϵ_1, $\ldots$, ϵ_d, δ_1, $\ldots$, δ_l are in L_1, that (Z_1, Z_2), ϵ_1, $\ldots$, ϵ_d, δ_1, $\ldots$, δ_l are independent, that*

$$\mathbf{E}\{\epsilon_1\} = \mathbf{E}\{\epsilon_2\} = \ldots = \mathbf{E}\{\epsilon_d\} = \mathbf{E}\{\delta_1\} = \mathbf{E}\{\delta_2\} = \ldots = \mathbf{E}\{\delta_l\} = 0,$$

that $\mathbf{E}\{Z_k^2\} > 0$ $(k \in \{1, 2\})$ and that $a_2, \ldots, a_d, b_2, \ldots, b_l \in \mathbb{R}$ and $d, l \geq 3$ and $a_2 \neq 0, a_3 \neq 0, b_2 \neq 0$ and $b_3 \neq 0$. Assume furthermore, that the characteristic function of (X, Y) does not vanish at any point.

If $\tilde{Z}_1$, $\tilde{Z}_2$, $\tilde{\epsilon}_1$, $\ldots$, $\tilde{\epsilon}_d$, $\tilde{\delta}_1$, $\ldots$, $\tilde{\delta}_l$ are in L_1, $\tilde{a}_2$, $\ldots$, $\tilde{a}_d$, $\tilde{b}_2$, $\ldots$, $\tilde{b}_l$ are in $\mathbb{R}$ and $\tilde{Z}_1$, $\ldots$, $\tilde{b}_l$ satisfy

$$\begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(d)} \\ Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(l)} \end{pmatrix} = \begin{pmatrix} 1 \cdot \tilde{Z}_1 \\ \tilde{a}_2 \cdot \tilde{Z}_1 \\ \vdots \\ \tilde{a}_d \cdot \tilde{Z}_1 \\ 1 \cdot \tilde{Z}_2 \\ \tilde{b}_2 \cdot \tilde{Z}_2 \\ \vdots \\ \tilde{b}_l \cdot \tilde{Z}_2 \end{pmatrix} + \begin{pmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_d \\ \tilde{\delta}_1 \\ \tilde{\delta}_2 \\ \vdots \\ \tilde{\delta}_l \end{pmatrix}$$

where the equality above holds in distribution,

$$\mathbf{E}\{\tilde{\epsilon}_1\} = \mathbf{E}\{\tilde{\epsilon}_2\} = \ldots = \mathbf{E}\{\tilde{\epsilon}_d\} = \mathbf{E}\{\tilde{\delta}_1\} = \mathbf{E}\{\tilde{\delta}_2\} = \ldots = \mathbf{E}\{\tilde{\delta}_l\} = 0$$

and $(\tilde{Z}_1, \tilde{Z}_2)$, $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_d$, $\tilde{\delta}_1, \ldots, \tilde{\delta}_l$ *are independent, then* $\tilde{a}_j = a_j$ $(j = 1, \ldots, d)$, $\tilde{b}_k = b_k$ $(k = 1, \ldots, l)$, $\mathbf{P}_{(\tilde{Z}_1,\tilde{Z}_2)} = \mathbf{P}_{(Z_1,Z_2)}$, $\mathbf{P}_{\tilde{\epsilon}_1} = \mathbf{P}_{\epsilon_1}$, $\ldots$, $\mathbf{P}_{\tilde{\epsilon}_d} = \mathbf{P}_{\epsilon_d}$ *and* $\mathbf{P}_{\tilde{\delta}_1} = \mathbf{P}_{\delta_1}$, $\ldots$, $\mathbf{P}_{\tilde{\delta}_l} = \mathbf{P}_{\delta_l}$.

Hence under the above assumptions $a_2, \ldots, a_d, b_2, \ldots, b_l$, *and the distributions of* (Z_1, Z_2), $\epsilon_1, \ldots, \epsilon_d$, $\delta_1, \ldots, \delta_l$ *are uniquely determined by the distribution of* (X, Y).

Remark 1. In case $d = 2$ and $l = 2$ the model (4) is not unique. For instance if Z, ϵ_1 and ϵ_2 are independent normally distributed with mean zero then the distribution of

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} Z + \epsilon_1 \\ a \cdot Z + \epsilon_2 \end{pmatrix}$$

does not uniquely determine the distribution of $Z, \epsilon_1, \epsilon_2$. For instance take $a = 1, Z \sim N(0,1), \epsilon_1 \sim N(0,1), \epsilon_2 \sim N(0,4)$ or $a = 4, Z \sim N(0,1/4), \epsilon_1 \sim N(0,7/4), \epsilon_2 \sim N(0,1)$. By computing covariance matrices it is easy to see that in both cases the distributions of (X_1, X_2) are the same.

Remark 2. A generalization of the proof of Lemma 1 shows that if we assume the model (3) in case $d_{z_1} > 1$ or $d_{z_2} > 1$, then our independence assumption together with the assumption that the characteristic function does not vanish imply that the distribution of (X, Y) uniquely determines the joint distribution of all other variables occurring in the model and all coefficients $a_{i,l}$ and $b_{j,k}$.

In the sequel we want to estimate the above latent variable model from the independent and identically distributed observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

The crucial property which allows us to show that the above model is uniquely determined is independence of the random variables. In the sequel we use this property for estimation of the model by determining estimates of the values of the latent variables in such a way that the corresponding empirical distributions satisfy asymptotically this independence assumption.

We start with definition of the estimate of the above model by estimating the coefficients a_j and b_k. Here we use

$$a_2 = \frac{\mathbf{E}\{X^{(2)} \cdot X^{(3)}\}}{\mathbf{E}\{X^{(1)} \cdot X^{(3)}\}} \quad \text{and} \quad a_j = \frac{\mathbf{E}\{X^{(2)} \cdot X^{(j)}\}}{\mathbf{E}\{X^{(1)} \cdot X^{(2)}\}}$$

and

$$b_2 = \frac{\mathbf{E}\{Y^{(2)} \cdot Y^{(3)}\}}{\mathbf{E}\{Y^{(1)} \cdot Y^{(3)}\}} \quad \text{and} \quad b_k = \frac{\mathbf{E}\{Y^{(2)} \cdot Y^{(k)}\}}{\mathbf{E}\{Y^{(1)} \cdot Y^{(2)}\}}$$

for $j, k > 2$ (cf., proof of Lemma 1) and set $\hat{a}_1 = \hat{b}_1 = 1$ and

$$\hat{a}_2 = \frac{\frac{1}{n}\sum_{i=1}^n X_i^{(2)} \cdot X_i^{(3)}}{\frac{1}{n}\sum_{i=1}^n X_i^{(1)} \cdot X_i^{(3)}} \quad \text{and} \quad \hat{a}_j = \frac{\frac{1}{n}\sum_{i=1}^n X_i^{(2)} \cdot X_i^{(j)}}{\frac{1}{n}\sum_{i=1}^n X_i^{(1)} \cdot X_i^{(2)}}$$

and

$$\hat{b}_2 = \frac{\frac{1}{n}\sum_{j=1}^n Y_j^{(2)} \cdot Y_j^{(3)}}{\frac{1}{n}\sum_{j=1}^n Y_j^{(1)} \cdot Y_j^{(3)}} \quad \text{and} \quad \hat{b}_k = \frac{\frac{1}{n}\sum_{j=1}^n Y_j^{(2)} \cdot Y_j^{(k)}}{\frac{1}{n}\sum_{j=1}^n Y_j^{(1)} \cdot Y_j^{(2)}}$$

for $j, k > 2$.

Next we try to determine estimates $(\hat{z}_{1,i}, \hat{z}_{2,i})$ of $(Z_{1,i}, Z_{2,i})$ for $i = 1, \ldots, n$. As soon we have available such estimates, we also have available estimates of the values of $\epsilon_j = X^{(j)} - a_j \cdot Z_1$ and $\delta_k = Y^{(k)} - b_k \cdot Z_2$, namely

$$\hat{\epsilon}_{j,i} = X_i^{(j)} - \hat{a}_j \cdot \hat{z}_{1,i} \quad \text{and} \quad \hat{\delta}_{k,i} = Y_i^{(k)} - \hat{b}_k \cdot \hat{z}_{2,i}$$

$(i = 1, \ldots, n)$, so we have available an estimated sample of the joint distribution of

$$((Z_1, Z_2), \epsilon_1, \ldots, \epsilon_d, \delta_1, \ldots \delta_l).$$

The basic idea is to consider the empirical distribution μ_n belonging to this estimated sample and to determine the estimates of the values of the latent variables in such a way that this empirical distribution satisfies approximately the independence condition of Lemma 1 and $\mathbf{E}\{\epsilon_j\} = \mathbf{E}\{\delta_k\} = 0$ which ensure uniqueness of the latent variable model.

More precisely, for values $\kappa_1, \ldots, \kappa_n$ in $\mathbb{R}^p$ let μ_{n,κ_1^n} be the empirical distribution of $\kappa_1, \ldots, \kappa_n$, i.e.,

$$\mu_{n,\kappa_1^n}(B) = \frac{1}{n}\sum_{i=1}^n 1_B(\kappa_i) \quad (B \subseteq \mathbb{R}^p).$$

Let $\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$ be the empirical distribution to

$$\begin{aligned}&((\hat{z}_{1,i}, \hat{z}_{2,i}), \hat{\epsilon}_{1,i}, \ldots, \hat{\epsilon}_{d,i}, \hat{\delta}_{1,i}, \ldots, \hat{\delta}_{l,i})\\ &= ((\hat{z}_{1,i}, \hat{z}_{2,i}), X_i^{(1)} - \hat{a}_1 \cdot \hat{z}_{1,i}, \ldots, X_i^{(d)} - \hat{a}_d \cdot \hat{z}_{1,i}, Y_i^{(1)} - \hat{b}_1 \cdot \hat{z}_{2,i}, \ldots, Y_i^{(l)} - \hat{b}_l \cdot \hat{z}_{2,i})\end{aligned}$$

$(i \in \{1, \ldots, n\})$, i.e.,

$$\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} = \mu_{n,((\hat{z}_1,\hat{z}_2),\hat{\epsilon}_1,\ldots,\hat{\epsilon}_d,\hat{\delta}_1,\ldots,\hat{\delta}_l)_1^n}.$$

The distribution μ of $((Z_1, Z_2), \epsilon_1, \ldots, \epsilon_d, \delta_1, \ldots, \delta_l)$ satisfies

$$\mu(\prod_{i=1}^{1+d+l} B_i) = \mu(B_1 \times \prod_{j=2}^{1+d+l} \mathbb{R}) \cdot \prod_{i=2}^{1+d+l} \mu(\mathbb{R}^2 \times \prod_{j=2}^{i-1} \mathbb{R} \times B_i \times \prod_{j=i+1}^{1+d+l} \mathbb{R})$$

for any $B_1 \in \mathcal{B}_2$, $B_2 \in \mathcal{B}, \ldots, B_{1+d+l} \in \mathcal{B}$ because of the independence assumption. It follows from probability theory that if this relation holds for all intervals of the form $(-\infty, x]$, then μ has independent components. We choose our estimated values such that this is approximately true for the empirical distribution $\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$. In order to be able to compute the estimate, we use here a sigmoidal approximation of the indicator function of an interval.

More precisely, we choose a continuous sigmoidal function $\sigma : \mathbb{R} \to \mathbb{R}$, i.e., a continuous monotone function $\sigma : \mathbb{R} \to [0,1]$ satisfying

$$\sigma(x) \to 0 \quad \text{as} \quad x \to -\infty \quad \text{and} \quad \sigma(x) \to 1 \quad \text{as} \quad x \to \infty,$$

probability weights $(p_r)_{r\in\mathbb{N}}$, $\alpha_{r,1}, \alpha_{r,2}, \beta_{r,j}, \gamma_{r,k} \in \mathbb{Q}$ such that

$$\mathbb{Q}^{2+d+l} = \{(\alpha_{r,1}, \alpha_{r,2}, \beta_{r,1}, \dots, \beta_{r,d}, \gamma_{r,1}, \dots, \gamma_{r,l}) \quad : \quad r \in \mathbb{N}\}$$

and $N_n \in \mathbb{N}$ satisfying $N_n \to \infty$ $(n \to \infty)$, and define our values of $(\hat{z}_1, \hat{z}_2)$ by minimizing

$$\begin{aligned}
T_n &:= \\
&\sum_{r=1}^{N_n} \Bigg| \frac{1}{n} \sum_{i=1}^{n} \sigma(-n \cdot (\hat{z}_{1,i} - \alpha_{r,1})) \cdot \sigma(-n \cdot (\hat{z}_{2,i} - \alpha_{r,2})) \cdot \prod_{j=1}^{d} \sigma(-n \cdot (\hat{\epsilon}_{j,i} - \beta_{r,j})) \\
&\qquad \cdot \prod_{k=1}^{l} \sigma(-n \cdot (\hat{\delta}_{k,i} - \gamma_{r,k})) \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \sigma(-n \cdot (\hat{z}_{1,i} - \alpha_{r,1})) \cdot \sigma(-n \cdot (\hat{z}_{2,i} - \alpha_{r,2})) \cdot \prod_{j=1}^{d} \frac{1}{n} \sum_{i=1}^{n} \sigma(-n \cdot (\hat{\epsilon}_{j,i} - \beta_{r,j})) \\
&\qquad \cdot \prod_{k=1}^{l} \frac{1}{n} \sum_{i=1}^{n} \sigma(-n \cdot (\hat{\delta}_{k,i} - \gamma_{r,k})) \Bigg|^2 \cdot p_r \\
&+ \sum_{j=1}^{d} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_{j,i} \right)^2 + \sum_{k=1}^{l} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_{k,i} \right)^2
\end{aligned}$$

subject to the constraints

$$\frac{1}{n} \sum_{i=1}^{n} \hat{z}_{1,i}^2 \leq 1 + \frac{1}{n} \sum_{i=1}^{n} (X_i^{(1)})^2 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{2,i}^2 \leq 1 + \frac{1}{n} \sum_{i=1}^{n} (Y_i^{(1)})^2. \tag{5}$$

Our main result is the following theorem.

Theorem 1. *Assume that the assumptions of Lemma 1 are satisfied, and let the estimate* $\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$ *of the distribution* μ *of*

$$((Z_1, Z_2), \epsilon_1, \dots, \epsilon_d, \delta_1, \dots, \delta_l)$$

be defined as above. Then with probability one

$$\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} \to \mu \quad \textit{weakly},$$

i.e.,

$$\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}(A) \to \mu(A) \quad (n \to \infty)$$

for all sets A *such that the boundary* ∂A *satisfies* $\mu(\partial A) = 0$.

Remark 3. It is straightforward to extend our estimate to the case of model (3) with $d_{Z_1} > 1$ or $d_{Z_2} > 1$: To do this, one just needs to replace the empirical distribution of

$$((\hat{z}_{1,i}, \hat{z}_{2,i}), \hat{\epsilon}_{1,i}, \dots, \hat{\epsilon}_{d,i}, \hat{\delta}_{1,i}, \dots, \hat{\delta}_{l,i})$$

by the empirical distribution of the vector of all latent variables and all estimated error terms in model (3) and adjust the definition of T_n.

Remark 4. In our definition of the estimate we minimize T_n subject to constraint (5). It follows from the proof of Theorem 1 that we can impose even more restrictions in the above minimization problems, as long as the values of the latent variables satisfy them with probability one for large n. For instance, in the next section we will assume $\mathbf{E}\{|Y^{(1)}|^4\} < \infty$. Since Z_2 and δ_1 are independent, $Y^{(1)} = Z_2 + \delta_1$ and $\mathbf{E}\{\delta_1\} = 0$, this implies

$$\begin{aligned} \mathbf{E}\left\{|Y^{(1)} - \mathbf{E}\{Y^{(1)}\}|^4\right\} &= \mathbf{E}\left\{|Z_2 - \mathbf{E}\{Z_2\} + \delta_1|^4\right\} \\ &\geq \mathbf{E}\left\{|Z_2 - \mathbf{E}\{Z_2\}|^4\right\}, \end{aligned}$$

hence

$$\begin{aligned} \mathbf{E}\{Z_2^4\} &\leq 2^4 \cdot \mathbf{E}\{(Z_2 - \mathbf{E}Z_2)^4\} + 2^4 \cdot |\mathbf{E}\{Y^{(1)}\}|^4 \\ &\leq 256 \cdot \mathbf{E}\{|Y^{(1)}|^4\} + 272 \cdot |\mathbf{E}\{Y^{(1)}\}|^4. \end{aligned}$$

Consequently, if we impose in this case the additional constraint

$$\frac{1}{n}\sum_{i=1}^{n} \hat{z}_{2,i}^4 \leq 1 + 256 \cdot \frac{1}{n}\sum_{i=1}^{n}(Y_i^{(1)})^4 + 272 \cdot \left(\frac{1}{n}\sum_{i=1}^{n} Y_i^{(1)}\right)^4 \tag{6}$$

in the above minimization problem, then the assertion of Theorem 1 still holds.

3 Estimation of the regression function corresponding to latent variables

In this section we estimate the regression function corresponding to the latent variables Z_1 and Z_2 in model (4), i.e., we estimate

$$m : \mathbb{R} \to \mathbb{R}, \quad m(x) = \mathbf{E}\{Z_2 | Z_1 = x\},$$

from the data

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}.$$

The basic idea is to use the data as in Section 2 to construct the sample

$$(\hat{z}_{1,1}, \hat{z}_{1,2}), \dots, (\hat{z}_{n,1}, \hat{z}_{n,2})$$

of (Z_1, Z_2) and to apply a regression estimate to this data.

By Theorem 1 we know that in case that we assume that all occurring random variables are bounded

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{z}_{i,2}-f(\hat{z}_{i,1})|^2-\mathbf{E}|Z_2-f(Z_1)|^2=\int|z_2-f(z_1)|^2d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)}-\int|z_2-f(z_1)|^2d\mu\to 0 \quad a.s.$$

for all bounded and continuous functions $f:\mathbb{R}\to\mathbb{R}$. We will see in the proof of Theorem 2 below that in case that we impose the additional constraint (6) in the definition of our estimate, then this result also holds for unbounded random variables provided that $\mathbf{E}\{|Y^{(1)}|^4\}<\infty$.

Since

$$\mathbf{E}\{|Z_2-m(Z_1)|^2\}=\min_{f:\mathbb{R}\to\mathbb{R}}\mathbf{E}\{|Z_2-f(Z_1)|^2\}$$

(cf., e.g., Section 1.1 in Györfi et al. (2002)) this motivates to estimate the regression function m by the well-known least squares estimate

$$m_n(\cdot)=\arg\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|\hat{z}_{i,2}-f(\hat{z}_{i,1})|^2, \tag{7}$$

where $\mathcal{F}_n$ is a suitable defined set of functions consisting of continuous and bounded functions $f:\mathbb{R}\to\mathbb{R}$ depending on the sample size n. For notational simplicity we assume here and in the sequel that the minimum above exists. Our main result is the following theorem.

Theorem 2. *Assume that in the model (4) the random variables Z_1, Z_2, ϵ_1, $\dots$, ϵ_d, δ_1, $\dots$, δ_l are in L_1, that (Z_1,Z_2), ϵ_1, $\dots$, ϵ_d, δ_1, $\dots$, δ_l are independent, that*

$$\mathbf{E}\{\epsilon_1\}=\mathbf{E}\{\epsilon_2\}=\ldots=\mathbf{E}\{\epsilon_d\}=\mathbf{E}\{\delta_1\}=\mathbf{E}\{\delta_2\}=\ldots=\mathbf{E}\{\delta_l\}=0,$$

that $\mathbf{E}\{Z_k^2\}>0$ $(k\in\{1,2\})$ and that $a_2,\dots,a_d,b_2,\dots,b_l\in\mathbb{R}$ and $d,l\geq 3$ and $a_2\neq 0, a_3\neq 0, b_2\neq 0$ and $b_3\neq 0$. Assume furthermore, that the characteristic function of (X,Y) does not vanish at any point, that $X^{(1)},\dots,X^{(d)}$, $Y^{(1)},\dots,Y^{(l)}$ are in L_2 and that $\mathbf{E}\{|Y^{(1)}|^4\}<\infty$.

Let $\mathcal{F}_n$ be sets of functions $f:\mathbb{R}^d\to\mathbb{R}$ which are bounded by some constant $L>0$ and assume that

$$\cup_{n=1}^{\infty}\mathcal{F}_n \quad \textit{is a equicontinuous set of functions.} \tag{8}$$

Let the least squares estimate m_n be defined as above, where we impose the condition (6) as additional constraint in the minimization problem. Then

$$\inf_{f\in\mathcal{F}_n}\int|f(x)-m(x)|^2\mathbf{P}_{Z_1}(dx)\to 0 \quad (n\to\infty) \tag{9}$$

implies

$$\int|m_n(x)-m(x)|^2\mathbf{P}_{Z_1}(dx)\to 0 \quad a.s.$$

In the sequel we choose $\mathcal{F}_n$ as suitably defined space of polynomial splines and show that in the case of bounded and Lipschitz continuous regression functions the corresponding least squares estimate (7) is strongly consistent.

Let $M \in \mathbb{N}$ be arbitrary. For $j \in \mathbb{Z}$ and $K \in \mathbb{N}$ let $B_{j,M}^K : \mathbb{R} \to \mathbb{R}$ be the B–spline with degree M, knot sequence $\{i/K : i \in \mathbb{Z}\}$ and support $[j/K, (j+M+1)/K]$ (cf., e.g., de Boor (1978), Schumaker (1981) or Chapter 14 in Györfi et al. (2002)). One well-known property of B–splines is that they are nonnegative and sum up to one (see de Boor (1978), pp. 109, 110). Furthermore,

$$\left\{ \sum_{i=-M}^{K-1} a_i \cdot B_{i,M}^K \; : \; a_i \in \mathbb{R} \right\}$$

is on $[0,1]$ equal to the set of all piecewise polynomials of degree M with respect to a partition of $[0,1]$ consisting of K equidistant intervals, which are $(M-1)$–times continuously differentiable on $[0,1]$. For $K_n \in \mathbb{N}$, $c_1 > 0$ and $c_2 > 0$ set

$$\mathcal{F}_n = \left\{ \sum_{j=-M}^{K_n-1} a_j \cdot B_{j,M}^{K_n} \; : \; |a_j - a_{j-1}| \leq \frac{c_1}{K_n} \text{ and } |a_j| \leq c_2 \quad (j \in \mathbb{Z}) \right\} \tag{10}$$

and define the estimate m_n by (7). Then the following result holds:

Corollary 1. *Assume that the assumptions of Theorem 1 are valid, and, in addition, that* $m(x) = \mathbf{E}\{Z_2|Z_1 = x\}$ *is Lipschitz continuous and bounded in absolute value. Assume furthermore that* $Z_1 \in [0,1]$ *a.s. and that we enforce in the definition of the estimate in Section 2* $\hat{z}_{i,1} \in [0,1]$ $(i = 1, \ldots, n)$*. Let the least squares estimate* m_n *be defined as in Theorem 2 for some* $K_n > 0$ *satisfying*

$$K_n \to \infty \quad (n \to \infty).$$

Then for c_1 *and* c_2 *sufficiently large we have*

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_{Z_1}(dx) \to 0 \quad a.s.$$

Proof. The functions in $\mathcal{F}_n$ are Lipschitz continuous with Lipschitz constant c_1 (cf., e.g., Lemma 14.6 in Györfi et al. (2002)), hence $\cup_{n=1}^{\infty} \mathcal{F}_n$ is equicontinuous. Furthermore, they are all bounded in absolute value by L (cf., e.g., Lemma 14.2 and Lemma 14.4 in Györfi et al. (2002)). Since

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_{Z_1}(dx) \leq \inf_{f \in \mathcal{F}_n} \sup_{x \in [0,1]} |f(x) - m(x)|^2 \to 0 \quad (n \to \infty)$$

(which follows because of m Lipschitz continuous and c_1 and c_2 sufficiently large from $K_n \to \infty$ $(n \to \infty)$, cf., e.g., Györfi et al. (2002), p. 271) the result follows from Theorem 2. □

Remark 5. Any application of the above estimate requires a data-dependent choice of all parameters of the functions space, in particular of the bounds on the coefficients and the differences of the coefficients. One way of doing this is to use splitting of the sample. It is an open problem whether in this case the above consistency result still holds, or (in case that it is not valid) there exist another method for a data-dependent choice of the parameters leading to consistent estimates.

4 Proofs

4.1 Proof of Lemma 1.

The proof is an extension of the proof of Lemma 2.1 in Li (2002).

Set $a_1 = b_1 = 1 = \tilde{a}_1 = \tilde{b}_1$. For $j, k = 1, \dots, d$, $j \neq k$, we have

$$\mathbf{E}\{X^{(j)} \cdot X^{(k)}\} = \mathbf{E}\left\{(a_j \cdot Z_1 + \epsilon_j) \cdot (a_k \cdot Z_1 + \epsilon_k)\right\} = a_j \cdot a_k \cdot \mathbf{E}\{Z_1^2\}$$

(where the last equality follows from the independence assumption and $\mathbf{E}\{\epsilon_k\} = 0$ $(k \in \{1, \dots, d\})$), and similarly

$$\mathbf{E}\{X^{(j)} \cdot X^{(k)}\} = \tilde{a}_j \cdot \tilde{a}_k \cdot \mathbf{E}\{\tilde{Z}_1^2\}.$$

Since a_2, a_3 and $\mathbf{E}\{Z_1^2\}$ are nonzero, $\tilde{a}_2, \tilde{a}_3$ and $\mathbf{E}\{\tilde{Z}_1^2\}$ share this property. Hence for $j = 2$ we have

$$a_2 = \frac{a_2 \cdot a_3 \cdot \mathbf{E}\{Z_1^2\}}{1 \cdot a_3 \cdot \mathbf{E}\{Z_1^2\}} = \frac{\mathbf{E}\{X^{(2)} \cdot X^{(3)}\}}{\mathbf{E}\{X^{(1)} \cdot X^{(3)}\}} = \frac{\tilde{a}_2 \cdot \tilde{a}_3 \cdot \mathbf{E}\{\tilde{Z}_1^2\}}{1 \cdot \tilde{a}_3 \cdot \mathbf{E}\{\tilde{Z}_1^2\}} = \tilde{a}_2$$

and for $j = 3, \dots, d$ we get

$$a_j = \frac{a_2 \cdot a_j \cdot \mathbf{E}\{Z_1^2\}}{1 \cdot a_2 \cdot \mathbf{E}\{Z_1^2\}} = \frac{\mathbf{E}\{X^{(2)} \cdot X^{(j)}\}}{\mathbf{E}\{X^{(1)} \cdot X^{(2)}\}} = \frac{\tilde{a}_2 \cdot \tilde{a}_j \cdot \mathbf{E}\{\tilde{Z}_1^2\}}{1 \cdot \tilde{a}_2 \cdot \mathbf{E}\{\tilde{Z}_1^2\}} = \tilde{a}_j$$

Similarly we get

$$b_2 = \frac{\mathbf{E}\{Y^{(2)} \cdot Y^{(3)}\}}{\mathbf{E}\{Y^{(1)} \cdot Y^{(3)}\}} = \tilde{b}_2, \quad \text{and} \quad b_k = \frac{\mathbf{E}\{Y^{(2)} \cdot Y^{(k)}\}}{\mathbf{E}\{Y^{(1)} \cdot Y^{(2)}\}} = \tilde{b}_k$$

for $k = 3, \dots, l$.

Using (4) and the independence assumption we see that the characteristic function $\varphi_{(X,Y)}$ of (X, Y) is given by

$$\begin{aligned}
&\varphi_{(X,Y)}(u_1, \dots, u_d, v_1, \dots, v_l) \\
&= \mathbf{E}\left\{\exp\left(i \cdot \sum_{j=1}^{d} u_j \cdot X^{(j)} + i \cdot \sum_{k=1}^{l} v_k \cdot Y^{(k)}\right)\right\}
\end{aligned}$$

$$= \mathbf{E}\left\{\exp\left(i\cdot\sum_{j=1}^{d}u_j\cdot(a_j\cdot Z_1+\epsilon_j)+i\cdot\sum_{k=1}^{l}v_k\cdot(b_k\cdot Z_2+\delta_k)\right)\right\}$$

$$= \mathbf{E}\Bigg\{\exp\left(i\cdot(\sum_{j=1}^{d}u_j\cdot a_j\cdot Z_1+\sum_{k=1}^{l}v_k\cdot b_k\cdot Z_2)\right)\cdot\prod_{j=1}^{d}\exp\left(i\cdot u_j\cdot\epsilon_j\right)$$

$$\cdot\prod_{k=1}^{l}\exp\left(i\cdot v_k\cdot\delta_k\right)\Bigg\}$$

$$= \varphi_{(Z_1,Z_2)}\left(\sum_{j=1}^{d}u_j\cdot a_j,\sum_{k=1}^{l}v_k\cdot b_k\right)\cdot\prod_{j=1}^{d}\varphi_{\epsilon_j}(u_j)\cdot\prod_{k=1}^{l}\varphi_{\delta_k}(v_k).$$

Since we know that the characteristic function of (X,Y) does not vanish at any point, we can conclude that also $\varphi_{(Z_1,Z_2)}$, φ_{ϵ_j} and φ_{δ_k} share this property. Furthermore, using

$$\varphi_{\epsilon_j}(0)=\varphi_{\delta_k}(0)=1\quad(j=2,\ldots,d,k=2,\ldots,l)$$

and

$$\varphi'_{\epsilon_2}(0)=i\cdot\mathbf{E}\epsilon_2=0=\varphi'_{\delta_2}(0)$$

we get

$$\varphi_{(X,Y)}(u_1,0,\ldots,0,v_1,0,\ldots,0)=\varphi_{(Z_1,Z_2)}(u_1,v_1)\cdot\varphi_{\epsilon_1}(u_1)\cdot\varphi_{\delta_1}(v_1),$$

$$\begin{aligned}&\frac{\partial}{\partial u_2}\varphi_{(X,Y)}(u_1,0,\ldots,0,v_1,0,\ldots,0)\\&=a_2\cdot\frac{\partial}{\partial z_1}\varphi_{(Z_1,Z_2)}(u_1,v_1)\cdot\varphi_{\epsilon_1}(u_1)\cdot\varphi_{\delta_1}(v_1)+\varphi_{(Z_1,Z_2)}(u_1,v_1)\cdot\varphi_{\epsilon_1}(u_1)\cdot\varphi_{\delta_1}(v_1)\cdot\varphi'_{\epsilon_2}(0)\\&=a_2\cdot\frac{\partial}{\partial z_1}\varphi_{(Z_1,Z_2)}(u_1,v_1)\cdot\varphi_{\epsilon_1}(u_1)\cdot\varphi_{\delta_1}(v_1)\end{aligned}$$

and

$$\frac{\partial}{\partial v_2}\varphi_{(X,Y)}(u_1,0,\ldots,0,v_1,0,\ldots,0)=b_2\cdot\frac{\partial}{\partial z_2}\varphi_{(Z_1,Z_2)}(u_1,v_1)\cdot\varphi_{\epsilon_1}(u_1)\cdot\varphi_{\delta_1}(v_1).$$

We conclude

$$\begin{aligned}&\varphi_{(Z_1,Z_2)}(u,v)\\&=\exp\left((\log\varphi_{(Z_1,Z_2)}(u,v)-\log\varphi_{(Z_1,Z_2)}(u,0))\right)\\&\qquad\cdot\exp\left((\log\varphi_{(Z_1,Z_2)}(u,0)-\log\varphi_{(Z_1,Z_2)}(0,0))\right)\\&=\exp\left(\int_0^v\frac{1}{b_2}\cdot\frac{\frac{\partial}{\partial v_2}\varphi_{(X,Y)}(u,0,\ldots,0,s,0,\ldots,0)}{\varphi_{(X,Y)}(u,0,\ldots,0,s,0,\ldots,0)}\,ds\right)\\&\qquad\cdot\exp\left(\int_0^u\frac{1}{a_2}\cdot\frac{\frac{\partial}{\partial u_2}\varphi_{(X,Y)}(t,0,\ldots,0,0,0,\ldots,0)}{\varphi_{(X,Y)}(t,0,\ldots,0,0,0,\ldots,0)}\,dt\right).\end{aligned}$$

We have considered the integrals above as parametrization of complex curve integrals of the function $z \mapsto 1/z$ and split them into finitely many integrals such that $\log z$ is well defined for each integral. (Here the number of intervals is finite since the curves in the integrals above have finite length and a positive distance to the origin.) This results in additional factor $\exp(i \cdot s \cdot 2\pi) = 1$ for some $s \in \mathbb{N}$. Similarly we get

$$\begin{aligned}
&\varphi_{(\tilde{Z}_1,\tilde{Z}_2)}(u,v) \\
&= \exp\left(\int_0^v \frac{1}{\tilde{b}_2} \cdot \frac{\frac{\partial}{\partial v_2}\varphi_{(X,Y)}(u,0,\ldots,0,s,0,\ldots,0)}{\varphi_{(X,Y)}(u,0,\ldots,0,s,0,\ldots,0)}\,ds\right) \\
&\qquad \cdot \exp\left(\int_0^u \frac{1}{\tilde{a}_2} \cdot \frac{\frac{\partial}{\partial u_2}\varphi_{(X,Y)}(t,0,\ldots,0,0,0,\ldots,0)}{\varphi_{(X,Y)}(t,0,\ldots,0,0,0,\ldots,0)}\,dt\right)
\end{aligned}$$

and from $a_2 = \tilde{a}_2$ and $b_2 = \tilde{b}_2$ we conclude $\varphi_{(Z_1,Z_2)} = \varphi_{(\tilde{Z}_1,\tilde{Z}_2)}$. But from $\varphi_{(Z_1,Z_2)}$ and $a_1,\ldots,a_d,b_1,\ldots,b_l$ we can determine φ_{ϵ_j} and φ_{δ_k} via

$$\varphi_{(X,Y)}(0,\ldots,0,u_j,0,\ldots 0,0,\ldots,0) = \varphi_{(Z_1,Z_2)}\left(u_j \cdot a_j, 0\right) \cdot \varphi_{\epsilon_j}(u_j)$$

and

$$\varphi_{(X,Y)}(0,\ldots,0,0,\ldots,0,v_k,0,\ldots 0) = \varphi_{(Z_1,Z_2)}\left(0, v_k \cdot b_k\right) \cdot \varphi_{\delta_k}(v_k).$$

Using the same relation for $\varphi_{(\tilde{Z}_1,\tilde{Z}_2)}$, $\varphi_{\tilde{\epsilon}_j}$ and $\varphi_{\tilde{\delta}_k}$ we see that

$$\varphi_{\epsilon_j} = \varphi_{\tilde{\epsilon}_j} \quad \text{and} \quad \varphi_{\delta_k} = \varphi_{\tilde{\delta}_k},$$

which implies the assertion. □

4.2 Proof of Theorem 1.

Throughout the proof we will use the abbreviation

$$\begin{aligned}
&\int f((u_1,u_2),v_1,\ldots,v_d,w_1,\ldots,w_l)\,d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} \\
&= \int f((u_1,u_2),v_1,\ldots,v_d,w_1,\ldots,w_l)\,\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}(d((u_1,u_2),v_1,\ldots,v_d,w_1,\ldots,w_l)),
\end{aligned}$$

so, e.g.,

$$\int \sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} = \frac{1}{n}\sum_{i=1}^n \sigma(-n\cdot(\hat{z}_{1,i}-\alpha_{r,1}))\cdot\sigma(-n\cdot(\hat{z}_{i,2}-\alpha_{r,2}))$$

and

$$\int v_j\,d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_{j,i}.$$

The proof is divided into nine steps.

In the first step of the proof we show that $(\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n})_{n\in\mathbb{N}}$ is tight with probability one, i.e., with probability one we find for each $\epsilon > 0$ a compact set $K \subseteq \mathbb{R}^2 \times \mathbb{R}^{d+l}$ such that

$$\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}(K^c) \leq \epsilon \quad \text{for all } n \in \mathbb{N}.$$

By the strong law of large numbers we know that with probability one

$$\frac{1}{n}\sum_{i=1}^n (X_i^{(j)})^2 \to \mathbf{E}\{(X^{(j)})^2\} < \infty \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n (Y_i^{(k)})^2 \to \mathbf{E}\{(Y^{(k)})^2\} < \infty, \tag{11}$$

so by definition of the estimate we may assume w.l.o.g.

$$\frac{1}{n}\sum_{i=1}^n (X_i^{(j)})^2 \leq c, \frac{1}{n}\sum_{i=1}^n (Y_i^{(k)})^2 \leq c, \frac{1}{n}\sum_{i=1}^n \hat{z}_{i,1}^2 \leq c \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \hat{z}_{i,2}^2 \leq c \tag{12}$$

for all $n \in \mathbb{N}$ for some $c > 0$ with probability one. Furthermore because of

$$\hat{a}_j \to a_j \quad (n \to \infty) \quad \text{and} \quad \hat{b}_k \to b_k \quad (n \to \infty) \tag{13}$$

with probability one we may assume in addition that

$$|\hat{a}_j| \leq c \quad \text{and} \quad |\hat{b}_k| \leq c$$

with probability one. By Markov inequality we get

$$\begin{aligned}
&\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\left(([-M,M]^{2+d+l})^c\right) \\
&\leq \hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\{|u_1| > M\} + \hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\{|u_2| > M\} + \sum_{j=1}^d \hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\{|v_j| > M\} \\
&\qquad + \sum_{k=1}^l \hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\{|w_k| > M\} \\
&\leq \frac{\int |u_1|^2 d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}}{M^2} + \frac{\int |u_2|^2 d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}}{M^2} + \sum_{j=1}^d \frac{\int |v_j|^2 d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}}{M^2} + \sum_{k=1}^d \frac{\int |w_k|^2 d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}}{M^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n \hat{z}_{1,i}^2}{M^2} + \frac{\frac{1}{n}\sum_{i=1}^n \hat{z}_{2,i}^2}{M^2} + \sum_{j=1}^d \frac{\frac{1}{n}\sum_{i=1}^n (X_i^{(j)} - \hat{a}_j \cdot \hat{z}_{1,i})^2}{M^2} \\
&\qquad + \sum_{k=1}^l \frac{\frac{1}{n}\sum_{i=1}^n (Y_i^{(k)} - \hat{b}_k \cdot \hat{z}_{2,i})^2}{M^2} \\
&\leq \frac{c}{M^2} + \frac{c}{M^2} + d \cdot \frac{2c + 2c^3}{M^2} + l \cdot \frac{2c + 2c^3}{M^2} \leq \epsilon
\end{aligned}$$

for M sufficiently large.

In the second step of the proof we show

$$T_n \to 0 \quad a.s. \tag{14}$$

Let $\tilde{T}_n$ and $\hat{\mu}_n^{(Z_1,Z_2)_1^n}$ be defined as T_n and $\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$, resp., with $(\hat{z}_{i,1}, \hat{z}_{i,2})$ be replaced by $(Z_{1,i}, Z_{2,i})$ $(i = 1, \dots, n)$. Because of

$$\mathbf{E}\{(X^{(1)})^2\} = \mathbf{E}Z_1^2 + \mathbf{E}\epsilon_1^2$$

we have $\mathbf{E}Z_1^2 \leq \mathbf{E}\{(X^{(1)})^2\} < \infty$, so by the strong law of large numbers we get

$$\frac{1}{n}\sum_{i=1}^n Z_{1,i}^2 \to \mathbf{E}Z_1^2 \leq \mathbf{E}\{(X^{(1)})^2\} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (X_i^{(1)})^2 \quad a.s.,$$

hence with probability one for n large enough

$$\frac{1}{n}\sum_{i=1}^n Z_{1,i}^2 \leq 1 + \frac{1}{n}\sum_{i=1}^n (X_i^{(1)})^2.$$

Similarly we see that with probability one we have for n large enough

$$\frac{1}{n}\sum_{i=1}^n Z_{2,i}^2 \leq 1 + \frac{1}{n}\sum_{i=1}^n (Y_i^{(1)})^2.$$

Then by definition of T_n we have with probability one for n large enough

$$T_n \leq \tilde{T}_n,$$

so it suffices to show

$$\tilde{T}_n \to 0 \quad a.s.$$

Since $(p_r)_{r\in\mathbb{N}}$ are probability weights and since σ is bounded this in turn follows from

$$\left(\int v_j d\hat{\mu}_n^{(Z_1,Z_2)_1^n}\right)^2 \to 0 \quad a.s. \quad (j = 1, \dots, d), \tag{15}$$

$$\left(\int w_k d\hat{\mu}_n^{(Z_1,Z_2)_1^n}\right)^2 \to 0 \quad a.s. \quad (k = 1, \dots, l) \tag{16}$$

and

$$\Bigg| \int \sigma(-n\cdot(u_1 - \alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\cdot\prod_{j=1}^d \sigma(-n\cdot(v_j - \beta_{r,j}))$$
$$\cdot \prod_{k=1}^l \sigma(-n\cdot(w_k - \gamma_{r,k}))d\hat{\mu}_n^{(Z_1,Z_2)_1^n}$$

$$
\begin{aligned}
-\int \sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))d\hat{\mu}_n^{(Z_1,Z_2)_1^n} \\
\cdot\prod_{j=1}^{d}\int \sigma(-n\cdot(v_j-\beta_{r,j}))d\hat{\mu}_n^{(Z_1,Z_2)_1^n} \\
\cdot\prod_{k=1}^{l}\int \sigma(-n\cdot(w_k-\gamma_{r,k}))d\hat{\mu}_n^{(Z_1,Z_2)_1^n}\Bigg|^2 \to 0 \quad a.s. \qquad (17)
\end{aligned}
$$

for any $r\in\mathbb{N}$.

Let $\bar{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$ and $\bar{\mu}_n^{(Z_1,Z_2)_1^n}$ be the empirical measures which we get if we replace in the definition of $\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$ and $\hat{\mu}_n^{(Z_1,Z_2)_1^n}$ the estimated coefficients by the true coefficients, respectively. The proof of step 1 implies that $(\bar{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n})_{n\in\mathbb{N}}$ and $(\bar{\mu}_n^{(Z_1,Z_2)_1^n})_{n\in\mathbb{N}}$ are tight with probability one, too. Since the estimated coefficients converge by the strong law of large numbers almost surely to the true coefficients, we conclude that we have for any bounded, uniformly continuous function f

$$
\int f\,d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}-\int f\,d\bar{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\to 0\ a.s. \text{ and } \int f\,d\hat{\mu}_n^{(Z_1,Z_2)_1^n}-\int f\,d\bar{\mu}_n^{(Z_1,Z_2)_1^n}\to 0\ a.s. \qquad (18)
$$

Here we have used that because of the tightness of the measures w.l.o.g. we can integrate (18) over some compact set, so that all occurring variables are bounded.

Furthermore, since $\bar{\mu}_n^{(Z_1,Z_2)_1^n}$ is in fact an empirical distribution to independent and identically distributed data, we know again by the strong law of large numbers that we have in addition

$$
\int f\,d\bar{\mu}_n^{(Z_1,Z_2)_1^n}\to\int f\,d\mu \quad a.s.,
$$

so altogether we know that we have for all bounded, uniformly continuous functions f

$$
\int f\,d\hat{\mu}_n^{(Z_1,Z_2)_1^n}\to\int f\,d\mu \quad a.s.
$$

Because of our independence assumption, which implies

$$
\begin{aligned}
&\mathbf{E}\Bigg\{\sigma(-n\cdot(Z_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(Z_2-\alpha_{r,2}))\cdot\prod_{j=1}^{d}\sigma(-n\cdot(\epsilon_j-\beta_{r,j})) \\
&\qquad\qquad\cdot\prod_{k=1}^{l}\sigma(-n\cdot\delta_k-\gamma_{r,k}))\Bigg\} \\
&=\mathbf{E}\left\{\sigma(-n\cdot(Z_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(Z_2-\alpha_{r,2}))\right\}\cdot\prod_{j=1}^{d}\mathbf{E}\left\{\sigma(-n\cdot(\epsilon_j-\beta_{r,j}))\right\} \\
&\qquad\qquad\cdot\prod_{k=1}^{l}\mathbf{E}\left\{\sigma(-n\cdot\delta_k-\gamma_{r,k}))\right\},
\end{aligned}
$$

from this we conclude (17). Relation (15) follows from $\mathbf{E}\epsilon_j = 0$ and the strong law of large numbers, which implies

$$\int v_j \, d\hat{\mu}_n^{(Z_1,Z_2)_1^n} = \frac{1}{n}\sum_{i=1}^{n}(X_i^{(j)} - \hat{a}_j \cdot Z_{1,i}) \to \mathbf{E}\{X^{(j)} - a_j \cdot Z_1\} = \mathbf{E}\epsilon_j \quad a.s.$$

Similarly we conclude (16) from $\mathbf{E}\delta_k = 0$.

In the third step of the proof we set

$$S_j(x_1, \dots, x_{2+d+l}) = a_j \cdot x_1 + x_{j+2}$$

for $j \in \{1, \dots, d\}$ and

$$S_j(x_1, \dots, x_{2+d+l}) = b_{j-d} \cdot x_2 + x_{j+2}$$

for $j \in \{d+1, \dots, d+l\}$ and show that we have with probability one

$$\left(\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\right)_{(S_1,\dots,S_{d+l})} \to \mathbf{P}_{(X^{(1)},\dots,X^{(d)},Y^{(1)},\dots,Y^{(l)})} \quad weakly. \tag{19}$$

To see this, we set

$$\bar{\epsilon}_{j,i} = X_i^{(j)} - a_j \cdot \hat{z}_{1,i} \quad \text{and} \quad \bar{\delta}_{k,i} = Y_i^{(k)} - b_k \cdot \hat{z}_{2,i}$$

and observe that our estimates of the random variables satisfy trivially the equations

$$X_i^{(j)} = a_j \cdot \hat{z}_{1,i} + X_i^{(j)} - a_j \cdot \hat{z}_{1,i} = S_j(\hat{z}_{1,i}, \hat{z}_{2,i}, \bar{\epsilon}_{1,i}, \dots, \bar{\epsilon}_{d,i}, \bar{\delta}_{1,i}, \dots, \bar{\delta}_{l,i})$$

and

$$Y_i^{(k)} = b_k \cdot \hat{z}_{2,i} + Y_i^{(k)} - b_k \cdot \hat{z}_{2,i} = S_{d+k}(\hat{z}_{1,i}, \hat{z}_{2,i}, \hat{\epsilon}_{1,i}, \dots, \bar{\epsilon}_{d,i}, \bar{\delta}_{1,i}, \dots, \bar{\delta}_{l,i}),$$

from which we conclude

$$\left(\bar{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}\right)_{(S_1,\dots,S_{d+l})} = \mu_{n,(X,Y)_1^n},$$

where the distribution on the right-hand side is the empirical distribution to (X_1, Y_1), $\dots$, (X_n, Y_n). But this distribution converges weakly to $\mathbf{P}_{(X,Y)}$, and together with (18) and the continuity of $S_1, \dots, S_{d+l}$ this implies (19).

In the fourth step of the proof we show that with probability one there exists a subsequence $(n_r)_r$ of $(n)_n$ and a measure μ satisfying

$$\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}} \to \mu \quad weakly \tag{20}$$

and

$$\mu_{(S_1,\dots,S_{d+l})} = \mathbf{P}_{(X,Y)}. \tag{21}$$

To see this, observe that by the first step of the proof the measures $\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n}$ are tight, and hence according to the theorem of Prohorov (cf., e.g., Theorem 6.1 in Billingsley (1968)) relatively compact, so (20) holds. Since $S_1, \ldots, S_{d+l}$ are continuous, this implies

$$\left(\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}}\right)_{(S_1,\ldots,S_{d+l})} \to \mu_{(S_1,\ldots,S_{d+l})} \quad \textit{weakly},$$

from which we get (21) by (19) and the uniqueness of the limit distribution in the case of weak convergence.

In the fifth step of the proof we show by an approximation of indicator functions of intervals by suitable neural networks that because of (14) the components of μ corresponding to (Z_1, Z_2), $\epsilon_1, \ldots, \epsilon_d$, $\delta_1, \ldots, \delta_l$ are independent with probability one. Let F be the distribution function of μ, i.e.,

$$\begin{aligned}&F((x_1,x_2),e_1,\ldots,e_d,d_1,\ldots,d_l)\\&=\mu\{u_1\le x_1,u_2\le x_2,v_1\le e_1,\ldots,v_d\le e_d,w_1\le d_1,\ldots,w_l\le d_l\},\end{aligned}$$

and set

$$F_{(Z_1,Z_2)}(x_1,x_2)=\mu\{u_1\le x_1,u_2\le x_2\},$$
$$F_{\epsilon_j}(e_j)=\mu\{v_j\le e_j\}$$

and

$$F_{\delta_k}(d_k)=\mu\{w_k\le d_k\}.$$

We have to show that

$$F((x_1,x_2),e_1,\ldots,e_d,d_1,\ldots,d_l)=F_{(Z_1,Z_2)}(x_1,x_2)\cdot\prod_{j=1}^{d}F_{\epsilon_j}(e_j)\cdot\prod_{k=1}^{l}F_{\delta_k}(d_k) \tag{22}$$

for all x_1, x_2, $e_1, \ldots, e_d$,$d_1, \ldots, d_l \in \mathbb{R}$.

Since distribution functions are right continuous, it suffices to show (22) for x_1, x_2, $e_1, \ldots, e_d$, $d_1, \ldots, d_l$ in some dense subset of $\mathbb{R}$, which we choose as

$$D=\mathbb{R}\setminus\left\{x\in\mathbb{R}\,:\,\mu\{u_1=x\}+\mu\{u_2=x\}+\sum_{j=1}^{d}\mu\{v_j=x\}+\sum_{k=1}^{l}\mu\{w_k=x\}>0\right\}$$

(which is dense in $\mathbb{R}$ since $\{\ldots\}$ is countable).

Let x_1, x_2, $e_1, \ldots, e_d$,$d_1, \ldots, d_l \in D$. For any $x \in \mathbb{R}$ and any $\epsilon > 0$ we can find $\alpha \in \mathbb{Q}$ satisfying for sufficiently large n

$$-n\cdot(z-\alpha)\quad\text{is sufficiently large for } z<x-\epsilon$$

and

$$-n\cdot(z-\alpha)\quad\text{is sufficiently small for } z>x-\epsilon$$

such that

$$\left|1_{(-\infty,x]}(z)-\sigma(-n\cdot(z-\alpha))\right|\le\epsilon$$

for $z < x - \epsilon$ or $z > x + \epsilon$ in case n sufficiently large. Furthermore, for any $x_1, x_2 \in \mathbb{R}$ and any $\epsilon > 0$ we can find $\alpha_1, \alpha_2 \in \mathbb{Q}$ satisfying

$$\left|1_{(-\infty,x_1]\times(-\infty,x_2]}(z_1, z_2) - \sigma(-n \cdot (z_1 - \alpha_1)) \cdot \sigma(-n \cdot (z_2 - \alpha_2))\right| \leq \epsilon \tag{23}$$

in case that $z_1 < x_1 - \epsilon$ or $z_1 > x_1 + \epsilon$, and that $z_2 < x_2 - \epsilon$ or $z_2 > x_2 + \epsilon$, for n sufficiently large. To see this, fix x_1, $x_2 \in \mathbb{R}$ and $\epsilon > 0$. Choose α_1, $\alpha_2 \in \mathbb{Q}$ such that

$$\left|1_{(-\infty,x_1]}(z) - \sigma(-n \cdot (z - \alpha_1))\right| \leq \frac{\epsilon}{2}$$

for $z < x_1 - \epsilon$ or $z > x_1 + \epsilon$, and such that

$$\left|1_{(-\infty,x_2]}(z) - \sigma(-n \cdot (z - \alpha_2))\right| \leq \frac{\epsilon}{2}$$

for $z < x_2 - \epsilon$ or $z > x_2 + \epsilon$. Then it is easy to see that (23) holds if one considers separately the four cases $z_1 < x_1 - \epsilon$ and $z_2 < x_2 - \epsilon$, $z_1 > x_1 + \epsilon$ and $z_2 < x_2 - \epsilon$, $z_1 < x_1 - \epsilon$ and $z_2 > x_2 + \epsilon$, and $z_1 > x_1 - \epsilon$ and $z_2 > x_2 - \epsilon$.

Consequently for suitably chosen r we see by expanding the terms below in a telescoping sum that we have

$$\begin{aligned}
&\Bigg| F((x_1, x_2), e_1, \ldots, e_d, d_1, \ldots, d_l) - \\
&\qquad \int \sigma(-n \cdot (u_1 - \alpha_{r,1})) \cdot \sigma(-n \cdot (u_2 - \alpha_{r,2})) \cdot \prod_{j=1}^{d} \sigma(-n \cdot (v_j - \beta_{r,j})) \\
&\qquad\qquad \cdot \prod_{k=1}^{l} \sigma(-n \cdot (w_k - \gamma_{r,k})) d\mu \Bigg| \\
&\leq (d + l + 1) \cdot \epsilon + \mu\{x_1 - \epsilon \leq z_1 \leq x_1 + \epsilon\} + \mu\{x_2 - \epsilon \leq z_2 \leq x_2 + \epsilon\} \\
&\qquad + \sum_{j=1}^{d} \mu\{e_j - \epsilon \leq v_j \leq e_j + \epsilon\} + \sum_{k=1}^{l} \mu\{d_k - \epsilon \leq w_k \leq d_k + \epsilon\}
\end{aligned}$$

and

$$\begin{aligned}
&\Bigg| F_{(Z_1,Z_2)}(x_1, x_2) \cdot \prod_{j=1}^{d} F_{\epsilon_j}(e_j) \cdot \prod_{k=1}^{l} F_{\delta_k}(d_k) - \\
&\qquad \int \sigma(-n \cdot (u_1 - \alpha_{r,1})) \cdot \sigma(-n \cdot (u_2 - \alpha_{r,2}))\, d\mu \cdot \prod_{j=1}^{d} \int \sigma(-n \cdot (v_j - \beta_{r,j}))\, d\mu \\
&\qquad\qquad \cdot \prod_{k=1}^{l} \int \sigma(-n \cdot (w_k - \gamma_{r,k}))\, d\mu \Bigg| \\
&\leq (d + l + 1) \cdot \epsilon + \mu\{x_1 - \epsilon \leq z_1 \leq x_1 + \epsilon\} + \mu\{x_2 - \epsilon \leq z_2 \leq x_2 + \epsilon\}
\end{aligned}$$

$$+\sum_{j=1}^{d}\mu\{e_j-\epsilon\le v_j\le e_j+\epsilon\}+\sum_{k=1}^{l}\mu\{d_k-\epsilon\le w_k\le d_k+\epsilon\}.$$

For $x_1,\ x_2,\ e_1,\ldots,e_d\ ,d_1,\ldots,d_l\in D$ the right-hand side above converges to zero for $\epsilon\to 0$, so it suffices to show that we have for any r

$$\begin{aligned}
&\int\sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\cdot\prod_{j=1}^{d}\sigma(-n\cdot(v_j-\beta_{r,j}))\\
&\qquad\qquad\cdot\prod_{k=1}^{l}\sigma(-n\cdot(w_k-\gamma_{r,k}))d\mu\\
&=\int\sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\,d\mu\cdot\prod_{j=1}^{d}\int\sigma(-n\cdot(v_j-\beta_{r,j}))\,d\mu\\
&\qquad\qquad\cdot\prod_{k=1}^{l}\int\sigma(-n\cdot(w_k-\gamma_{r,k}))\,d\mu.
\end{aligned}$$

But this in turn follows from (20), since

$$\begin{aligned}
&\int\sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\cdot\prod_{j=1}^{d}\sigma(-n\cdot(v_j-\beta_{r,j}))\\
&\qquad\qquad\cdot\prod_{k=1}^{l}\sigma(-n\cdot(w_k-\gamma_{r,k}))d\mu\\
&\quad-\int\sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\,d\mu\cdot\prod_{j=1}^{d}\int\sigma(-n\cdot(v_j-\beta_{r,j}))\,d\mu\\
&\qquad\qquad\cdot\prod_{k=1}^{l}\int\sigma(-n\cdot(w_k-\gamma_{r,k}))\,d\mu\\
&=\lim_{l\to\infty}\Bigg(\int\sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\cdot\prod_{j=1}^{d}\sigma(-n\cdot(v_j-\beta_{r,j}))\\
&\qquad\qquad\cdot\prod_{k=1}^{l}\sigma(-n\cdot(w_k-\gamma_{r,k}))d\hat{\mu}_{n_l}^{(\hat{z}_1,\hat{z}_2)_1^{n_l}}\\
&\quad-\int\sigma(-n\cdot(u_1-\alpha_{r,1}))\cdot\sigma(-n\cdot(u_2-\alpha_{r,2}))\,d\hat{\mu}_{n_l}^{(\hat{z}_1,\hat{z}_2)_1^{n_l}}\\
&\qquad\cdot\prod_{j=1}^{d}\int\sigma(-n\cdot(v_j-\beta_{r,j}))\,d\hat{\mu}_{n_l}^{(\hat{z}_1,\hat{z}_2)_1^{n_l}}\cdot\prod_{k=1}^{l}\int\sigma(-n\cdot(w_k-\gamma_{r,k}))d\hat{\mu}_{n_l}^{(\hat{z}_1,\hat{z}_2)_1^{n_l}}\Bigg)\\
&=0\quad a.s.
\end{aligned}$$

by (14) and $N_n\to\infty$ $(n\to\infty)$.

In the sixth step of the proof we show that the components of μ are with probability one in L_1. By Portmanteau theorem (cf. Billingsley (1968)) and $\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}} \to \mu$ weakly with probability one we have with probability one

$$\begin{aligned} \int |u_1| \, d\mu &= \int_0^\infty \mu\{|u_1| > t\} \, dt \\ &\leq \int_0^\infty \liminf_{r\to\infty} \hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}} \{|u_1| > t\} \, dt \\ &\leq \int_0^\infty \liminf_{r\to\infty} \frac{\int |u_1|^2 d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}}}{t^2} \, dt < \infty, \end{aligned}$$

since by definition of the estimate we have with probability one

$$\begin{aligned} &\liminf_{r\to\infty} \int |u_1|^2 d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}} = \liminf_{r\to\infty} \frac{1}{n_r} \sum_{i=1}^{n_r} \hat{z}_{1,i}^2 \\ &\leq \liminf_{r\to\infty} (1 + \frac{1}{n_r} \sum_{i=1}^{n_r} (X_i^{(1)})^2) = 1 + \mathbf{E}\{(X_i^{(1)})^2\} < \infty. \end{aligned}$$

Furthermore

$$\int |v_j| \, d\mu \leq \int_0^\infty \liminf_{r\to\infty} \frac{\int |v_j|^2 d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}}}{t^2} \, dt < \infty \quad a.s.,$$

since we have with probability one

$$\begin{aligned} &\liminf_{r\to\infty} \int |v_j|^2 d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}} = \liminf_{r\to\infty} \frac{1}{n_r} \sum_{i=1}^{n_r} \left(X_i^{(j)} - \hat{a}_j \cdot \hat{z}_{1,i}\right)^2 \\ &\leq \liminf_{r\to\infty} \left(2 \cdot \frac{1}{n_r} \sum_{i=1}^{n_r} \left(X_i^{(j)}\right)^2 + 2 \cdot \hat{a}_j^2 \cdot \frac{1}{n_r} \sum_{i=1}^{n_r} \hat{z}_{1,i}^2\right) \\ &= 2 \cdot \mathbf{E}\{(X_i^{(1)})^2\} + 2 \cdot a_j^2 \cdot (1 + \mathbf{E}\{(X_i^{(1)})^2\}) < \infty. \end{aligned}$$

Similar arguments for the other components yield the desired result.

In the seventh step of the proof we show that we have with probability one

$$\int v_j \, d\mu = \int w_k \, d\mu = 0 \quad \text{for } j \in \{1, \ldots, d\} \text{ and } k \in \{1, \ldots, l\}. \tag{24}$$

To do this, we observe that because of (14) we have with probability one

$$\int v_j \, d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} \to 0 \quad (n \to \infty) \quad \text{and} \quad \int w_k \, d\hat{\mu}_n^{(\hat{z}_1,\hat{z}_2)_1^n} \to 0 \quad (n \to \infty)$$

for $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, l\}$. Using the arguments of the sixth step of the proof we see that we have

$$\int |x_j| \cdot 1_{\{|x_j|>L\}} d\mu(x_j) \to 0 \quad (L \to \infty)$$

and

$$\int |x_j| \cdot 1_{\{|x_j|>L\}} d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}}(x_j)$$
$$\leq \frac{1}{L} \cdot \int |x_j|^2 d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}}(x_j) \to 0 \quad (L \to \infty).$$

Consequently we may replace

$$(x_1, \ldots, x_{2+d+l}) \mapsto x_j$$

by a bounded and continuous function in the integrals below, hence $\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}} \to \mu$ weakly implies

$$\int x_j \, d\mu(x_j) = \lim_{r\to\infty} \int x_j \; d\hat{\mu}_{n_r}^{(\hat{z}_1,\hat{z}_2)_1^{n_r}}(x_j) = 0.$$

In the eighth step of the proof we show that we have with probability one

$$\mu = \mathbf{P}_{((Z_1,Z_2),\epsilon^{(1)},\ldots,\epsilon^{(d)},\delta^{(1)},\ldots,\delta^{(l)})}. \tag{25}$$

This follows directly of the uniqueness of the distribution of

$$((Z_1, Z_2), \epsilon^{(1)}, \ldots, \epsilon^{(d)}, \delta^{(1)}, \ldots, \delta^{(l)})$$

shown in Lemma 1 and the properties of the distribution μ proven in the previous four steps.

In the ninth and final step of the proof we show the assertion of the theorem.

Let f be an arbitrary bounded and continuous function. We have to show that with probability one for all such functions

$$\int f \, d\hat{\mu}_n \to \int f \, d\mathbf{P}_{((Z_1,Z_2),\epsilon^{(1)},\ldots,\epsilon^{(d)},\delta^{(1)},\ldots,\delta^{(l)})} \quad (n \to \infty).$$

To show this, it suffices to show that with probability one for any subsequence $(n_r)_r$ of $(n)_n$ and all such functions there exists a subsubsequence $(n_{r_k})_k$ with the property

$$\int f \, d\hat{\mu}_{n_{r_k}} \to \int f \, d\mathbf{P}_{((Z_1,Z_2),\epsilon^{(1)},\ldots,\epsilon^{(d)},\delta^{(1)},\ldots,\delta^{(l)})} \quad (k \to \infty). \tag{26}$$

Let $(n_r)_r$ be an arbitrary subsequence of $(n)_n$. According to steps 1 till 8 above applied to $(n_r)_r$ instead of $(n)_n$ there exists a subsequence $(n_{r_k})_k$ of $(n_r)_r$ with the property

$$\hat{\mu}_{n_{r_k}} \to \mathbf{P}_{((Z_1,Z_2),\epsilon^{(1)},\ldots,\epsilon^{(d)},\delta^{(1)},\ldots,\delta^{(l)})} \quad \textit{weakly}.$$

Here the weak convergence holds whenever (11), (12) and (13) hold. But this implies (26), and the proof is complete. □

4.3 Proof of Theorem 2.

Choose $f_n \in \mathcal{F}_n$ such that

$$\int |f_n(z) - m(z)|^2 \mathbf{P}_{Z_1}(dz) \to 0 \quad (n \to \infty).$$

Then

$$\begin{aligned} 0 &\leq \int |m_n(z) - m(z)|^2 \mathbf{P}_{Z_1}(dz) \\ &= \int |m_n(z_1) - z_2|^2 d\mu - \int |m(z_1) - z_2|^2 d\mu \\ &= \int |m_n(z_1) - z_2|^2 d\mu - \int |f_n(z_1) - z_2|^2 d\mu + \int |f_n(z) - m(z)|^2 \mathbf{P}_{Z_1}(dz), \end{aligned}$$

hence it suffices to show

$$\limsup_{n\to\infty} \int |m_n(z_1) - z_2|^2 d\mu - \int |f_n(z_1) - z_2|^2 d\mu \leq 0 \quad a.s.$$

Since by definition of m_n

$$\begin{aligned} &\int |m_n(z_1) - z_2|^2 d\mu - \int |f_n(z_1) - z_2|^2 d\mu \\ &\leq \int |m_n(z_1) - z_2|^2 d\mu - \frac{1}{n}\sum_{i=1}^n |m_n(\hat{z}_{i,1}) - \hat{z}_{i,2}|^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^n |f_n(\hat{z}_{i,1}) - \hat{z}_{i,2}|^2 - \int |f_n(z_1) - z_2|^2 d\mu \end{aligned}$$

this in turn follows from

$$\int |m_n(z_1) - z_2|^2 d\mu - \frac{1}{n}\sum_{i=1}^n |m_n(\hat{z}_{i,1}) - \hat{z}_{i,2}|^2 \to 0 \quad a.s. \tag{27}$$

and

$$\frac{1}{n}\sum_{i=1}^n |f_n(\hat{z}_{i,1}) - \hat{z}_{i,2}|^2 - \int |f_n(z_1) - z_2|^2 d\mu \to 0 \quad a.s. \tag{28}$$

For $\beta > 0$ and $z \in \mathbb{R}$ set $T_\beta z = \max\{\min\{z, \beta\}, -\beta\}$. We have

$$\int |z_2|^2 \cdot 1_{\{|z_2|>\beta\}} d\mu \to 0 \quad (\beta \to \infty)$$

by dominated convergence and

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n |\hat{z}_{i,2}|^2 \cdot 1_{\{|\hat{z}_{i,2}|>\beta\}}$$

$$
\begin{aligned}
&\leq \limsup_{n\to\infty} \frac{1}{\beta^2} \cdot \frac{1}{n} \sum_{i=1}^{n} |\hat{z}_{i,2}|^4 \\
&\leq \frac{1}{\beta^2} \cdot \limsup_{n\to\infty} \left(1 + 256 \cdot \frac{1}{n} \sum_{i=1}^{n} (Y_i^{(1)})^4 + 272 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} Y_i^{(1)} \right)^4 \right) \\
&\to 0 \quad (\beta \to \infty)
\end{aligned}
$$

a.s. by (6) and the strong law of large numbers. Hence in order to prove (27) it suffices to show

$$
\int |m_n(z_1) - T_\beta z_2|^2 d\mu - \frac{1}{n} \sum_{i=1}^{n} |m_n(\hat{z}_{i,1}) - T_\beta \hat{z}_{i,2}|^2 \to 0 \quad a.s.
$$

for all $\beta > 0$.

Let $\beta > 0$ be arbitrary. It suffices to show: With probability one any subsequence $(n_k)_k$ from $(n)_n$ contains a subsubsequence n_{k_r} such that

$$
\int |m_{n_{k_r}}(z_1) - T_\beta z_2|^2 d\mu - \frac{1}{n_{k_r}} \sum_{i=1}^{n_{k_r}} |m_{n_{k_r}}(\hat{z}_{i,1}) - T_\beta \hat{z}_{i,2}|^2 \to 0 \quad (r \to \infty).
$$

In the sequel we condition on the event that

$$
\hat{\mu}_n^{(\hat{z}_1, \hat{z}_2)_1^n} \to \mu \quad weakly, \tag{29}
$$

which has probability one because of Theorem 1. Let $(n_k)_k$ be an arbitrary subsequence of $(n)_n$. By the Theorem of Arzela-Ascoli (cf., Dunford and Schwartz (1958)) the sequence m_{n_k} of equicontinuous functions contains a (random) subsequence $m_{n_{k_r}}$ which converges in supremum norm to some (random) function $\bar{m}$. Since the functions $m_{n_{k_r}}$ are continuous and bounded, $\bar{m}$ has this property, too. By (29) we know

$$
\int |\bar{m}(z_1) - T_\beta z_2|^2 d\mu - \frac{1}{n_{k_r}} \sum_{i=1}^{n_{k_r}} |\bar{m}(\hat{z}_{i,1}) - T_\beta \hat{z}_{i,2}|^2 \to 0 \quad (r \to \infty).
$$

Using

$$
\begin{aligned}
&\left| \int |m_{n_{k_r}}(z_1) - T_\beta z_2|^2 d\mu - \int |\bar{m}(z_1) - T_\beta z_2|^2 d\mu \right| \\
&= \left| \int (m_{n_{k_r}}(z_1) - \bar{m}(z_1)) \cdot (m_{n_{k_r}}(z_1) + \bar{m}(z_1) - 2 \cdot T_\beta z_2) d\mu \right| \\
&\leq (2L + 2\beta) \cdot \|m_{n_{k_r}} - \bar{m}\|_\infty
\end{aligned}
$$

and

$$
\begin{aligned}
&\left| \frac{1}{n_{k_r}} \sum_{i=1}^{n_{k_r}} |m_{n_{k_r}}(\hat{z}_{i,1}) - T_\beta \hat{z}_{i,2}|^2 - \frac{1}{n_{k_r}} \sum_{i=1}^{n_{k_r}} |\bar{m}(\hat{z}_{i,1}) - T_\beta \hat{z}_{i,2}|^2 \right| \\
&\leq (2L + 2\beta) \cdot \|m_{n_{k_r}} - \bar{m}\|_\infty
\end{aligned}
$$

we see that this implies (27). In the same way we can also prove (28), which completes the proof. □

5 Acknowledgment

The authors would like to thank the referees for their constructive comments.

References

[1] Allman, E. S., Matias, C. and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, **37**, pp. 3099–3132.

[2] Amato, U., Antoniadis, A., Samarov, A. and Tsybakov, A. B. (2010). Noisy independent factor analysis model for density estimation and classification. *Electronic Journal of Statistics*, **4**, pp. 707–736.

[3] Anderson, T. W. (1989). Linear latent variable models and covariance structures. *Journal of Econometrics*, **41**, pp. 91–119.

[4] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70**, pp. 191–221.

[5] Bai, J. and Ng, S. (2006). Evaluating latent and observed factors in microeconomics and finance. *Journal of Econometrics*, **131**, pp. 507–537.

[6] Bartolucci, F.,Pennoni, F. and Francis, B. (2006). Likelihood inference for a class of latent Markov models under linear hypothesis on the transition probabilities. *Journal of Royal Statistical Society B*, **68**, pp. 155–178.

[7] Bartolucci, F. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of Royal Statistical Society A*, **170**, pp. 115–132.

[8] Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, **71**, pp. 93–107.

[9] Bianconcini, S. and Cagnone S. (2012). Estimation of generalized linear latent variable models via fully exponential Laplace approximation. *Journal of Multivariate Analysis*, in press.

[10] Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley & Sons, New York.

[11] Bollen, K. A. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, New York.

[12] Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology,*, **53**, pp. 605-634.

[13] de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.

[14] Breslaw, J. A. and McIntosh, J. (1998). Simulated latent variable estimation of models with ordered categorical data. *Journal of Econometrics*, **87**, pp. 25–47.

[15] Browne, R. P. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, **142**, pp. 2976–2984.

[16] Colombo, D., Maathuis, M. H., Kalisch, M. and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, **40**, pp. 294–321.

[17] Connes, D., Ronchetti, E., and Victoria-Feser, M.-P. (2010). Goodness of fit for generalized linear latent variables models. *Journal of the American Statistical Association*, **105**, pp. 1126-1134.

[18] Christensen, W. F. and Amemiya, Y. (2002). Latent variable analysis of multivariate spatial data. *Journal of the American Statistical Association*, **97**, pp. 302–317.

[19] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467–481.

[20] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**, pp. 1371–1385.

[21] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, pp. 71–82.

[22] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231–239.

[23] Dunford, N. and Schwartz, J. T. (1958). *Linear Operators, Volume 1*. Wiley-Interscience, New York.

[24] Gebregziabher, M. and DeSantis, S. M. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference*, **140**, pp. 3252–3262.

[25] Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, **10**, pp. 43–52.

[26] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression.* Springer-Verlag, New York.

[27] Hall, P., Müller, H.-G. and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of Royal Statistical Society B*, **70**, pp. 703–723.

[28] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Second Edition, Springer-Verlag, New York.

[29] Holzinger, K. J. and Swineford, F. (1939). A Study in Factor Analysis: The Stability of a Bi-factor Solution. *Supplementary Educational Monographs.* Chicago, Ill.: The University of Chicago.

[30] Hu, L. and Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, **3**, pp. 424-453.

[31] Irincheeva, I., Cantoni, E. and Genton, M. G. (2012). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, pp. 1–18.

[32] Klein, A. G. and Moosbrugger, H. Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, **65**, 457-474.

[33] Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling.* Third Edition. Guilford Press, New York.

[34] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054-3058.

[35] Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, **41**, pp. 281–293.

[36] Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, **18**, pp. 95–138.

[37] Li, T. (2002). Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics*, **110**, pp. 1–26.

[38] Lynn, H.S. and McCulloch, C. E. (2000). Using principal component analysis and correspondence analysis for estimation in latent variable models. *Journal of the American Statistical Association*, **95**, pp. 561–572.

[39] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677–687.

[40] Mack, Y. P. (1981). Local properties of k–nearest neighbor regression estimates. *SIAM Journal on Algebraic and Discrete Methods*, **2**, pp. 311–323.

[41] Marcoulides, G. A. and Schumacker, R. E. (Eds). (1996). *Advanced Structural Equation Modeling: Issues and Techniques.* Lawrence Erlbaum Associates, Hillsdale, NJ.

[42] Marsh, H. W., Wen, Z. and Hau, K.-T. (2006). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock & R.O. Mueller (Eds.) *Structural Equation Modeling: A Second Course*, pp. 225-265. Information Age Publishing.

[43] McNicholas, P. D. (2010). Model-based classifcation using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, **140**, pp. 1175–1181.

[44] Montanari, A. and Viroli, C. (2010). The independent factor analysis approach to latent variable modelling. *Statistics*, **44**, pp. 397-416.

[45] Muthén, L. K. and Muthén, B. O. (1998-2011). *Mplus User's Guide.* Sixth Edition. Muthén and Muthén, Los Angeles.

[46] Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability and its Applications*, **9**, pp. 141–142.

[47] Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, **15**, pp. 134–137.

[48] Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *Annals of Statistics*, **36**, pp. 1595-1618.

[49] Schumacker, R. and Marcoulides, G. (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling.* Lawrence Erlbaum Associates, Hillsdale, NJ.

[50] Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

[51] Skrondal, A. and Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, **34**, pp. 712-745.

[52] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, pp. 595–645.

[53] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.

[54] Zhao, L. C. (1987). Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *Journal of Multivariate Analysis*, **21**, pp. 168–178.