Journal of Multivariate Analysis 163 (2018) 1–14

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Inverse regression approach to robust nonlinear high-to-low dimensional mapping

Emeline Perthame [a], Florence Forbes [a,*], Antoine Deleforge [b]

[a] Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France
[b] Inria, Rennes, France

ARTICLE INFO

Article history:
Received 2 February 2017
Available online 7 October 2017

Keywords:
EM algorithm
Inverse regression
Mixture of regressions
Nonlinear regression
High dimension
Robust regression
Student distribution

ABSTRACT

The goal of this paper is to address the issue of nonlinear regression with outliers, possibly in high dimension, without specifying the form of the link function and under a parametric approach. Nonlinearity is handled via an underlying mixture of affine regressions. Each regression is encoded in a joint multivariate Student distribution on the responses and covariates. This joint modeling allows the use of an inverse regression strategy to handle the high dimensionality of the data, while the heavy tail of the Student distribution limits the contamination by outlying data. The possibility to add a number of latent variables similar to factors to the model further reduces its sensitivity to noise or model misspecification. The mixture model setting has the advantage of providing a natural inference procedure using an EM algorithm. The tractability and flexibility of the algorithm are illustrated in simulations and real high-dimensional data with good performance that compares favorably with other existing methods.

1. Introduction

A large amount of applications deal with relating explanatory variables (or covariates) to response variables through a regression-type model. In many circumstances, assuming a linear regression model is inadequate and more sensible models are likely to be nonlinear. Other complexity sources include the necessity to take into account a large number of covariates and the possible presence of outliers or influential observations in the data. Estimating a function defined over a large number of covariates is generally difficult because standard regression methods have to estimate a large number of parameters. Then, even in moderate dimension, outliers can result in misleading values for these parameters and predictions may no longer be reliable. In this work, we address these three complication sources by proposing a tractable model that is able to perform nonlinear regression from a high-dimensional space while being robust to outlying data.

A natural approach for modeling nonlinear mappings is to approximate the target relationship by a mixture of linear regression models. Mixture models and paradoxically also the so-called mixture of regression models [10,17,20] are mostly used to handle clustering issues and few papers refer to mixture models for actual regression and prediction purposes. Conventional mixtures of regressions are used to add covariates information to clustering models. For high-dimensional data, some penalized approaches of mixtures of regressions have been proposed such as the Lasso regularization [12,36] but these methods are not designed for prediction and do not deal with outliers. For moderate dimensions, more robust mixtures of regressions have been proposed using Student t distributions [33] possibly combined with trimming [44]. However, in general, conventional mixtures of regressions are inadequate for regression because they assume *assignment*

* Corresponding author.
E-mail addresses: emeline.perthame@pasteur.fr (E. Perthame), florence.forbes@inria.fr (F. Forbes), antoine.deleforge@inria.fr (A. Deleforge).

https://doi.org/10.1016/j.jmva.2017.09.009
0047-259X/

independence [21]. This means that the assignments to each of the regression components are independent of the covariate values. In contrast, in the method we propose, the covariate value is expected to be related to the membership to one of the linear regressions. Each linear regression is mostly active in a specific region of the covariate space.

When extended with assignment dependence, models in the family of mixtures of regressions are more likely to be suitable for regression application. This is the case of the so-called Gaussian Locally Linear Mapping (GLLiM) model [11] that assumes Gaussian noise models and is in its unconstrained version equivalent to a joint Gaussian mixture model (GMM) on both responses and covariates. GLLiM includes a number of other models in the literature. It may be viewed as an affine instance of mixture of experts as formulated in [43] or as a Gaussian cluster-weighted model (CWM) [19] except that the response variable can be multivariate in GLLiM while only scalar in CW models. There have been a number of useful extensions of CW models. The CWt model of [22] deals with non Gaussian distributions and uses Student t distributions for an increased robustness to outliers. The work of [37] uses a factor analyzers approach (CWFA) to deal with CW models when the number of covariates is large. The idea is to overcome the high dimensionality issue by imposing constraints on the covariance matrix of the high-dimensional variable. Incrementally, [38] combines then the Student and Factor analyzers extensions in a so-called CWtFA model. As an alternative to heavy-tailed distributions, some approaches propose to deal with outliers by removing them from the estimation using trimming. Introducing trimming into CWM has then been investigated in [18] but for a small number of covariates and a small number of mixture components. All these CW variants have been designed for clustering and have not been assessed in terms of regression performance.

In contrast, we consider an approach dedicated to regression. To handle the high dimensionality, we adopt an *inverse regression* strategy in the spirit of GLLiM which consists of exchanging the roles of responses and covariates. Doing so, we bypass the difficulty of high-to-low regression by considering the problem the other way around, i.e., low-to-high. We build on the work in [11] by considering mixtures of Student distributions that are able to better handle outliers. As an advantage over the CWtFA approach, our model can deal with response variables of dimension greater than 1. In addition, CWtFA involves the computation of a large empirical covariance matrix of the size of the higher dimension. Furthermore, under our approach, the observed response variables can be augmented with unobserved latent responses. This is interesting for solving regression problems in the presence of data corrupted by irrelevant information for the problem at hand. It has the potential of being well suited in many application scenarios, namely whenever the response variable is only partially observed, because it is neither available, nor observed with appropriate sensors. Moreover, used in combination with the inverse regression trick, the augmentation of the response variables with latent variables acts as a factor analyzer modeling for the noise covariance matrix in the forward regression model. The difference between our approach and CWtFA is further illustrated in Appendix B.

The present paper is organized as follows. The proposed model is presented in Section 2 under the acronym SLLiM for Student Locally Linear Mapping. Its use for prediction is also specified in the same section. Section 3 presents an EM algorithm for the estimation of the model parameters with technical details postponed in Appendix A. Proposals for selecting the number of components and the number of latent responses are described in Section 4. The SLLiM model properties and performance are then illustrated in simulations in Section 5 and real high-dimensional data in Section 6. Section 7 ends the paper with a discussion and some perspectives.

2. Robust mixture of linear regressions in high dimension

We consider the following regression problem. For $n \in \{1, \ldots, N\}$, $\mathbf{y}_n \in \mathbb{R}^L$ stands for a vector of response variables with dimension L and $\mathbf{x}_n \in \mathbb{R}^D$ stands for a vector of explanatory variables or covariates with dimension D. These vectors are assumed to be independent realizations of two random variables $\mathbf{Y}$ and $\mathbf{X}$. It is supposed that $L \ll D$ and the number of observations N can be smaller than D. The objective is to estimate the regression function g that we will also call *forward* regression that maps a set of covariates $\mathbf{x}$ to the response variable space, $g(\mathbf{x}) = \mathrm{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$.

Inverse regression strategy. When the number D of covariates is large, typically more than hundreds, estimating g is difficult because it relies on the exploration of a large dimensional space. A natural approach is therefore to, prior to regression, reduce the dimension of the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and this preferably by taking into account the responses $\mathbf{y}_1, \ldots, \mathbf{y}_N$. Methods like partial least squares (PLS), sliced inverse regression (SIR) and principal component based methods [1,9,27,35,42] follow this approach, in the category of semi- or non-parametric approaches. When considering parametric models, the issue is usually coming from the necessity to deal with large covariance matrices. A common solution is then to consider parsimonious modeling of these matrices either by making an oversimplistic independence assumption or using structured parameterization based on eigenvalues decomposition [6] or factor modeling [37]. In this work, we follow a third approach based on the concept of *inverse* regression while remaining parametric as described in [11]. The idea is to bypass the difficulty of estimating a high-to-low dimensional mapping g by estimating instead the other-way-around relationship, namely the low-to-high or *inverse* mapping from $\mathbf{Y}$ to $\mathbf{X}$. This requires then to focus first on a model of the distribution of $\mathbf{X}$ given $\mathbf{Y}$ and implies the definition of a joint model on $(\mathbf{Y}, \mathbf{X})$ to go from one conditional distribution to the other. The reference to a joint distribution is already present in the mixture of experts (MoE) model of [43] in the Gaussian case. However, inversion is not addressed and generally not tractable in non-Gaussian MoE such as those proposed in [8].

Mixture of linear regressions. Because $\mathbf{Y}$ is of moderate dimension, typically less than 10, the inverse regression is likely to be much easier to estimate. However, it is still likely to be nonlinear. An attractive approach for modeling nonlinear data is to use a mixture of linear models; see [11,19,43]. Focusing on the modeling of the inverse regression, we consider that each