Inverse regression approach to robust non-linear high-to-low dimensional mapping

Emeline Perthame, Florence Forbes, Antoine Deleforge

▶ **To cite this version:**

Emeline Perthame, Florence Forbes, Antoine Deleforge. Inverse regression approach to robust non-linear high-to-low dimensional mapping. 2016. hal-01347455v1

HAL Id: hal-01347455

https://hal.science/hal-01347455v1

Preprint submitted on 21 Jul 2016 (v1), last revised 30 Nov 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Inverse regression approach to robust non-linear high-to-low dimensional mapping

Emeline Perthame, Florence Forbes and Antoine Deleforge

July 20, 2016

Abstract The goal of this paper is to address the issue of non linear regression with outliers possibly in high dimension, without specifying the form of the link function and under a parametric approach. Non linearity 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 modelling 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 mispecification. The mixture model setting has the advantage to provide a natural inference procedure using an EM algorithm. The tractability and flexibility of the algorithm are illustrated on real high dimensional data with good performance that compares favorably with other existing methods.

Keywords Robust regression, non linear regression, mixture of regressions, inverse regression, high dimension, Generalized Student distribution, EM algorithm.

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 non-linear. 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 the three complication sources by proposing a tractable model that is able to perform non-linear regression from a high-dimensional space while being robust to outlying data.

A natural approach for modeling non-linear mappings is to approximate the target relationship by a piecewise linear function. In the method we propose, non linearity is handled via a mixture of locally linear regression models. Mixture models and paradoxically also the so-called mixture of regression models [de Veaux, 1989, Goldfeld and Quandt, 1973, Frühwirth-Schnatter, 2006] 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 [Städler et al., 2010, Devijver, 2015] 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 t-distributions [Peel and McLachlan, 2000] possibly combined with trimming [Yao et al., 2014]. However, in general, conventional mixture of regressions are inadequate for regression because they assume *assignment independence* [Hennig, 2000]. This means that the assignments to each of the regression components are independent of the covariate values. In contrast, in piecewise linear regression the covariate value is expected to be related to the membership to one of the local linear regressions.

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 [Deleforge et al., 2015] 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 [Xu et al., 1995] or as a Gaussian cluster-weighted model (CWM) [Gershenfeld, 1997] 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 [Ingrassia et al., 2012] deals with non Gaussian distributions and uses Student-t distributions for an increased robustness to outliers. The work of [Subedi et al., 2013] 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, [Subedi et al., 2015] combine 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 [Garcia-Escudero et al., 2015] but for a small number of covariates and 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 [Deleforge et al., 2015] 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 one and can be estimated with a standard EM algorithm while CWtFA is implemented via an AECM algorithm which 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 modelling for the noise covariance matrix in the forward regression model.

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. Two important issues are discussed, namely initialization and model selection. 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 on real high dimensional data in Section 5. Section 6 ends the paper with a discussion and some perspectives.

2 Robust piecewise linear regression in high dimension

We consider the following regression problem where the usual notation is reversed for reasons that will become clearer below. For $n \in \{1, \dots, N\}$, $\mathbf{x}_n \in \mathbb{R}^L$ stands for a vector of response variables with dimension L and $\mathbf{y}_n \in \mathbb{R}^D$ stands for a vector of explanatory variables or covariates with dimension D. These vectors are assumed to be independent realizations of two random variables $\mathbf{X}$ and $\mathbf{Y}$. 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 $\mathbf{y}$ to the response variable space, $g(\mathbf{y}) = \mathbb{E}[X_n|\mathbf{Y} = \mathbf{y}]$.

Inverse regression strategy. When the number D of covariates is large, estimating g is difficult because it relies on the exploration of a large dimensional space. A natural approach is therefore to, prior to regression, first reduce the dimension of the covariates $\{\mathbf{y}_n\}_{n=1}^N$ and this preferably by taking into account the responses $\{\mathbf{x}_n\}_{n=1}^N$. Methods like partial least squares (PLS), sliced inverse regression (SIR) and Principal component based methods [Rosipal and Krämer, 2006, Li, 1991, Wu, 2008, Cook, 2007,

Adragni and Cook, 2009] follow this approach, in the category of non or semi-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 modelling of these matrices either by making oversimplistic independence assumption or using structured parameterization based on eigenvalues decomposition [Bouveyron et al., 2007] or factor modelling [Subedi et al., 2013]. In this work, we follow a third approach based on the concept of *inverse* regression while remaining parametric as described in [Deleforge et al., 2015]. 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 $\mathbf{X}$ to $\mathbf{Y}$. This requires then to focus first on a model of the distribution of $\mathbf{Y}$ given $\mathbf{X}$ (therefore the proposed reversed notation where the low-dimensional variable $\mathbf{X}$ is the regressor) and implies the definition of a joint model on $(\mathbf{X}, \mathbf{Y})$ 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 [Xu et al., 1995] in the Gaussian case. However, inversion is not addressed and generally not tractable in non Gaussian MoE such as those proposed in [Chamroukhi, 2015].

Piecewise linear regression. Because $\mathbf{X}$ is of moderate dimension, the inverse regression is likely to be much easier to estimate. However, it is still likely to be non linear. An attractive approach for modeling non-linear data is to use a mixture of locally linear models (*e.g.* [Xu et al., 1995, Gershenfeld, 1997, Deleforge et al., 2015]). Focusing on the modelling of the inverse regression, we consider that each $\mathbf{y}$ is the noisy image of $\mathbf{x}$ obtained from a K-component locally-affine transformation. This is modeled by introducing the latent variable $Z \in \{1, ..., K\}$ such that:

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbb{I}(Z = k)(\boldsymbol{A}_k \mathbf{X} + \boldsymbol{b}_k + \boldsymbol{E}_k) \tag{1}$$

where $\mathbb{I}$ is the indicator function, matrix $\boldsymbol{A}_k \in \mathbb{R}^{D \times L}$ and vector $\boldsymbol{b}_k \in \mathbb{R}^D$ define an affine transformation and $\boldsymbol{E}_k \in \mathbb{R}^D$ is an error term not correlated with $\mathbf{X}$ capturing both the observation noise in $\mathbb{R}^D$ and the reconstruction error due to the local affine approximation. $\boldsymbol{E}_k \in \mathbb{R}^D$ is assumed to be zero-mean. For the forward regression $p(\mathbf{X}|\mathbf{Y})$ to be easy to derive from $p(\mathbf{Y}|\mathbf{X})$, it is important to control the nature of the joint $p(\mathbf{X}, \mathbf{Y})$. Once a family of tractable joint distributions is chosen, we can look for one that is compatible with (1). When $\boldsymbol{E}_k$ is assumed to be Gaussian, such a piecewise linear inverse regression strategy has been proposed and successfully applied to high dimensional problems in [Deleforge et al., 2015] using Gaussian distributions.

Outliers accommodation. The tractability and stability properties of the Gaussian distributions are very convenient and appropriate to the manipulation of conditional and marginal distributions. However, Gaussian models are limited in their ability to handle atypical data due to their short tails. Student t-distributions are heavy-tailed alternatives that have the advantage to remain tractable. For a joint model of $\mathbf{X}$ and $\mathbf{Y}$, we therefore

consider a mixture of K *generalized* Student distributions with the following *local* $L+D$ dimensional generalized Student distributions:

$$p(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y}|Z=k)=\mathcal{S}_{L+D}([\mathbf{x},\mathbf{y}]^T;\boldsymbol{m}_k,\boldsymbol{V}_k,\alpha_k,1) \tag{2}$$

where $[\mathbf{x},\mathbf{y}]^T$ denotes the transpose of the vector $[\mathbf{x},\mathbf{y}]$, $\boldsymbol{m}_k$ is a $L+D$ dimensional mean vector, $\boldsymbol{V}_k$ is a $(D+L)\times(D+L)$ scale matrix and α_k a positive scalar. A generalized version of the standard t-distribution is considered. It is also referred to as the Arellano-Valle and Bolfarine's Generalized t distribution in [Kotz and Nadarajah, 2004] p. 94, Section 5.5. In contrast to the standard t-distribution, the generalized t-distribution is stable by conditioning and marginalizing. The probability density function of a M-dimensional generalized t is given by:

$$\begin{aligned}\mathcal{S}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\alpha,\gamma)&=\int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/u)\;\mathcal{G}(u;\alpha,\gamma)\;\mathrm{d}u \\ &=\frac{\Gamma(\alpha+M/2)}{|\boldsymbol{\Sigma}|^{1/2}\,\Gamma(\alpha)\,(2\pi\gamma)^{M/2}}\;[1+\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})/(2\gamma)]^{-(\alpha+M/2)},\end{aligned} \tag{3}$$

where $\mathcal{N}_M(\,.\,;\boldsymbol{\mu},\boldsymbol{\Sigma}/u)$ denotes the M-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/u$ and $\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})=(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ is the square of the Mahalanobis distance between $\mathbf{y}$ and $\boldsymbol{\mu}$. The first order moment exists for $\alpha>1/2$ and the mean is $\boldsymbol{\mu}$ in this case but $\boldsymbol{\Sigma}$ is not strictly speaking the covariance matrix of the t-distribution which is $\gamma\boldsymbol{\Sigma}/(\alpha-1)$ when $\alpha>1$. For identifiability reason, we assume in addition that $\gamma=1$ as the expression above depends on γ and $\boldsymbol{\Sigma}$ only through the product $\gamma\boldsymbol{\Sigma}$. The first equality in (3) shows a useful representation of the distribution as a Gaussian scale mixture which involves an additional Gamma distributed[1] positive scalar latent variable U.

In applying the inverse regression strategy, the key point is to account for (1) into the parameterization of $\boldsymbol{m}_k$ and $\boldsymbol{V}_k$. Given $Z=k$, it follows from (2) that $\mathbf{X}$ is Student distributed and $\mathbf{X}$ can be assumed to have a mean $\boldsymbol{c}_k\in\mathbb{R}^L$ and a scale matrix $\boldsymbol{\Gamma}_k\in\mathbb{R}^{L\times L}$. Then using (1), it comes straightforwardly that

$$\begin{aligned}\boldsymbol{m}_k&=\begin{bmatrix}\boldsymbol{c}_k\\ \boldsymbol{A}_k\boldsymbol{c}_k+\boldsymbol{b}_k\end{bmatrix},\\ \boldsymbol{V}_k&=\begin{bmatrix}\boldsymbol{\Gamma}_k & \boldsymbol{\Gamma}_k\boldsymbol{A}_k^T\\ \boldsymbol{A}_k\boldsymbol{\Gamma}_k & \boldsymbol{\Sigma}_k+\boldsymbol{A}_k\boldsymbol{\Gamma}_k\boldsymbol{A}_k^T\end{bmatrix}.\end{aligned} \tag{4}$$

With no constraints on the parameters, parametrization (4) is general in the sense that all admissible values of $\boldsymbol{m}_k$ and $\boldsymbol{V}_k$ can be written this way (see a proof in Appendix A of [Deleforge et al., 2015] or for $D=1$ in [Ingrassia et al., 2012]). Counting the number of parameters, we get for the joint model (2) $1/2(D(D-1)+L(L-1))+DL+D+L$,

[1] the Gamma distribution when the variable is X is denoted by $\mathcal{G}(x;\alpha,\gamma)=x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\gamma x)\gamma^\alpha$ where Γ denotes the Gamma function

which is not suprisingly symmetric in D and L. The inverse regression parameterization becomes interesting when adding constraints. Typically one may add constraints on the largest covariance matrix, namely $\mathbf{\Sigma}_k$. If $\mathbf{\Sigma}_k$ is assumed diagonal, the number of parameters reduces to $1/2(L(L-1))+DL+2D+L$. It is then clear that proceeding the other way around, that is assuming $\mathbf{\Gamma}_k$ diagonal instead, would not reduce the number of parameters as drastically. As an example, for $D = 500$ and $L = 2$, the model has 2003 parameters using the inverse strategy and 126,254 using a forward parameterization. But this gain in complexity would not be useful if the forward regression $p(\mathbf{X}|\mathbf{Y}, Z = k)$ was not easy to derive. The advantage of the joint distribution defined by (2) and (4) is that all conditionals and marginals can be derived and remain Student. More specifically, we obtain the following Student distributions (see [Kotz and Nadarajah, 2004] Section 5.5 p. 94)

$$p(\mathbf{X} = \mathbf{x}|Z = k) = \mathcal{S}_L(\mathbf{x}; \boldsymbol{c}_k, \mathbf{\Gamma}_k, \alpha_k, 1) \tag{5}$$
$$p(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, Z = k) = \mathcal{S}_D(\mathbf{y}; \boldsymbol{A}_k\mathbf{x} + \boldsymbol{b}_k, \mathbf{\Sigma}_k, \alpha_k^y, \gamma_k^y), \tag{6}$$

$$\text{with} \quad \alpha_k^y = \alpha_k + L/2$$
$$\gamma_k^y = 1 + \frac{1}{2}\delta(\mathbf{x}, \boldsymbol{c}_k, \mathbf{\Gamma}_k) \; .$$

and

$$p(\mathbf{Y} = \mathbf{y}|Z = k) = \mathcal{S}_D(\mathbf{y}; \boldsymbol{c}_k^*, \mathbf{\Gamma}_k^*, \alpha_k, 1) \tag{7}$$
$$p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}, Z = k) = \mathcal{S}_L(\mathbf{x}; \boldsymbol{A}_k^*\mathbf{y} + \boldsymbol{b}_k^*, \mathbf{\Sigma}_k^*, \alpha_k^x, \gamma_k^x), \tag{8}$$

$$\text{with} \quad \alpha_k^x = \alpha_k + D/2$$
$$\gamma_k^x = 1 + \frac{1}{2}\delta(\mathbf{y}, \boldsymbol{c}_k^*, \mathbf{\Gamma}_k^*) \; .$$

and

$$\boldsymbol{c}_k^* = \boldsymbol{A}_k\boldsymbol{c}_k + \boldsymbol{b}_k, \tag{9}$$
$$\mathbf{\Gamma}_k^* = \mathbf{\Sigma}_k + \boldsymbol{A}_k\mathbf{\Gamma}_k\boldsymbol{A}_k^T, \tag{10}$$
$$\boldsymbol{A}_k^* = \mathbf{\Sigma}_k^*\boldsymbol{A}_k^T\mathbf{\Sigma}_k^{-1}, \tag{11}$$
$$\boldsymbol{b}_k^* = \mathbf{\Sigma}_k^*(\mathbf{\Gamma}_k^{-1}\boldsymbol{c}_k - \boldsymbol{A}_k^T\mathbf{\Sigma}_k^{-1}\boldsymbol{b}_k), \tag{12}$$
$$\mathbf{\Sigma}_k^* = (\mathbf{\Gamma}_k^{-1} + \boldsymbol{A}_k^T\mathbf{\Sigma}_k^{-1}\boldsymbol{A}_k)^{-1}. \tag{13}$$

It is interesting to consider the structure of the $D \times D$ scale matrix $\mathbf{\Gamma}_k^*$ in (10). If a direct parameterization of the forward regression was considered, $\mathbf{\Gamma}_k^*$ would be the high dimensional covariance parameter. When $\mathbf{\Sigma}_k$ is assumed diagonal, we recover a factor

analyzer structure. The factor decomposition of $\mathbf{\Gamma}_k^*$ shows some similarity with the cluster-weighted modelling approach of [Subedi et al., 2015] which assumes a factor decomposition of the high-dimensional covariates covariance matrix in a forward modelling. However some differences can be pointed out. Our parameterization is more parsimonious with qD less parameters if q is the number of factors used in [Subedi et al., 2015]. Then, the joint model used in [Subedi et al., 2015] (*e.g.* their eq. (4)) is not a joint Student model because the degrees of freedom parameters in their joint probability density function decomposition are the same for the conditional and marginal pdf. Typically a joint Student model would imply instead a degree of freedom parameter that depends on $\mathbf{y}$ in the expression of $p(\mathbf{x}|\mathbf{y})$. Note that the notation for $\mathbf{x}, \mathbf{y}$ is reversed compare to [Subedi et al., 2015]. One consequence of that is that [Subedi et al., 2015] cannot use a regular EM for parameter estimation but have to use an AECM algorithm. In terms of performance, we could not really assess the performance of their approach on very high dimensional data as the code of this recent work is not available. However for comparison, we applied our method on the two real data sets used in [Subedi et al., 2015] for which $L = 1$ and $D = 6$ and 13 respectively. The results are reported in Appendix A and confirm that the two models yield to similar results on simple well separated cluster examples and different results on more complex data.

Low-to-high mapping and prediction. Defining π_k as the probability $p(Z = k)$, equations from (5) to (13), show that the whole model is entirely defined by a set of parameters denoted by $\boldsymbol{\theta} = \{\boldsymbol{c}_k, \mathbf{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \mathbf{\Sigma}_k, \alpha_k, \pi_k\}_{k=1}^K$ and an inverse regression from $\mathbb{R}^L$ (low-dimensional space) to $\mathbb{R}^D$ (high-dimensional space) can be obtained using the following *inverse conditional density*:

$$p(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}_L(\mathbf{x}; \boldsymbol{c}_k, \mathbf{\Gamma}_k, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j \mathcal{S}_M(\mathbf{x}; \boldsymbol{c}_j, \mathbf{\Gamma}_j, \alpha_j, 1)} \mathcal{S}_D(\mathbf{y}; \boldsymbol{A}_k\mathbf{x} + \boldsymbol{b}_k, \mathbf{\Sigma}_k, \alpha_k^y, \gamma_k^y). \tag{14}$$

Also, more importantly, the forward regression of interest, *i.e.*, from $\mathbb{R}^D$ (the high dimension) to $\mathbb{R}^L$ (the low dimension), is obtained from the *forward conditional density*:

$$p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}^*) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}_D(\mathbf{y}; \boldsymbol{c}_k^*, \mathbf{\Gamma}_k^*, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j \mathcal{S}_D(\mathbf{y}; \boldsymbol{c}_j^*.\mathbf{\Gamma}_j^*, \alpha_j, 1)} \mathcal{S}_L(\mathbf{x}; \boldsymbol{A}_k^*\mathbf{y} + \boldsymbol{b}_k^*, \mathbf{\Sigma}_k^*, \alpha_k^x, \gamma_k^x). \tag{15}$$

The latter involves parameters $\{\pi_k, \alpha_k\}_{k=1}^K$ and the *forward regression* parameters $\{\boldsymbol{c}_k^*, \mathbf{\Gamma}_k^*, \boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \mathbf{\Sigma}_k^*\}_{k=1}^K$ that can be analytically derived from the *inverse regression parameters* $\boldsymbol{\theta}$ with a drastic reduction of the model size, making tractable its estimation. Indeed, if we consider isotropic equal $\mathbf{\Sigma}_k$, the dimension of the learned parameter vector $\boldsymbol{\theta}$ is $\mathcal{O}(DL+L^2)$, while it would be $\mathcal{O}(DL + D^2)$ using a forward model[2].

[2]Recall that $L \ll D$.

Then, when required, a prediction of response $\mathbf{x}$ corresponding to an input $\mathbf{y}$ can be proposed using the expectation of $p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta})$ in (15) that can be obtained using:

$$\mathbb{E}[\mathbf{X}|\mathbf{y}; \boldsymbol{\theta}^*] = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}_D(\mathbf{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j \mathcal{S}_D(\mathbf{y}; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*, \alpha_j, 1)} (\boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*). \tag{16}$$

Response augmentation. In some applications, it is known that some confusing factors could interfere with the responses without being measured. This phenomenon can be modeled by assuming that the response $\mathbf{X}$ is partially observed. The $\mathbf{X}$ vector is therefore decomposed as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{T} \\ \mathbf{W} \end{bmatrix}$$

where $\mathbf{T} \in \mathbb{R}^{L_t}$ is the observed part and $\mathbf{W} \in \mathbb{R}^{L_w}$ is not observed and is considered as latent. Accordingly, the dimension of the response is $L = L_t + L_w$ where L_t is the number of observed responses and L_w is the number of unobserved factors. To account for this decomposition, we introduce the notations below

$$\boldsymbol{c}_k = \begin{bmatrix} \boldsymbol{c}_k^t \\ \boldsymbol{c}_k^w \end{bmatrix}, \boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_k^t & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_k^w \end{bmatrix}$$

and

$$\boldsymbol{A}_k = [\boldsymbol{A}_k^t, \boldsymbol{A}_k^w]$$

with $\boldsymbol{A}_k^t$ (resp. $\boldsymbol{A}_k^w$) a $p \times L_t$ (resp. $p \times L_w$) matrix. For identifiability, $\boldsymbol{c}_k^w$ and $\boldsymbol{\Gamma}_k^w$ must be fixed and are usually set to $\boldsymbol{c}_k^w = \mathbf{0}$ and $\boldsymbol{\Gamma}_k^w = \mathbb{1}_{L_w}$ ($\mathbb{1}_M$ denotes the $M \times M$ identity matrix). This model in which the response variable is augmented with latent factors is the one we refer to as Student Locally Linear Mapping (SLLiM) in the following of the paper.

Remark. Considering the addition of these factors $\mathbf{W}$, one can show that the conditional distribution function of $\mathbf{Y}$ given the observed $\mathbf{T}$ is:

$$p(\mathbf{Y} = \mathbf{y}|\mathbf{T} = \mathbf{t}, Z = k) = \mathcal{S}_D(\mathbf{y}, \boldsymbol{A}_k[\mathbf{t}, \boldsymbol{c}_k^w]^T + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k + \boldsymbol{A}_k^w \boldsymbol{\Gamma}_k^w \boldsymbol{A}_k^{w\top}, \alpha_k^y, \gamma_k^y).$$

Therefore, compare to the non augmented version of our model, the high dimensional matrix $\boldsymbol{\Sigma}_k$ is replaced by $\boldsymbol{\Sigma}_k + \boldsymbol{A}_k^w \boldsymbol{\Gamma}_k^w \boldsymbol{A}_k^{w\top}$ which corresponds to a factor model with L_w factors. When $\boldsymbol{\Sigma}_k$ is set to a diagonal matrix, this allows for more general dependence structures that can account for some additional dependencies while remaining tractable in high dimension. Similarly in the forward model, $\boldsymbol{\Gamma}_k^* = \boldsymbol{\Sigma}_k + \boldsymbol{A}_k^w \boldsymbol{\Gamma}_k^w \boldsymbol{A}_k^{w\top} + \boldsymbol{A}_k^t \boldsymbol{\Gamma}_k^t \boldsymbol{A}_k^{t\top}$ is also augmented with a L_w factor structure.

3 Estimation procedure

In contrast to the Gaussian case, no closed-form solution exists for the maximum likelihood estimation of the parameters for a t-distribution but tractability is maintained, both in the univariate and multivariate case, via the Gaussian scale mixture representation introduced in Equation (3) [Peel and McLachlan, 2000, Bishop and Svensen, 2005, Archambeau and Verleysen, 2007]. An efficient closed-form EM algorithm provides maximum likelihood estimates of the parameters $\boldsymbol{\theta} = \{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k \, \pi_k, \alpha_k\}_{k=1}^K$. The class of EM algorithms consists in updating parameters by iteratively maximizing the conditional expectation of the complete data log-likelihood, given the observed training data $(\mathbf{y}, \mathbf{t})_{1:N} = \{\mathbf{y}_n, \mathbf{t}_n\}_{n=1}^N$, and the last update $\boldsymbol{\theta}^{(i)}$. More generally, we will write $(\ )_{1:N}$ to indicate that a N-sample of the argument is considered. At iteration $(i+1)$, we look for the new set $\boldsymbol{\theta}^{(i+1)}$ that verifies:

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}[\log p((\mathbf{y}, \mathbf{t}, \mathbf{W}, U, Z)_{1:N}; \boldsymbol{\theta}) | (\mathbf{y}, \mathbf{t})_{1:N}; \boldsymbol{\theta}^{(i)}],$$

Using that responses $\mathbf{T}$ and latent variables $\mathbf{W}$ are independent given hidden variables Z and U and that $\boldsymbol{c}_k^w$ and $\boldsymbol{\Gamma}_k^w$ are fixed, the expected log likelihood to be maximized splits into the two following parts :

$$\mathbb{E}_{\widetilde{r}_Z} \mathbb{E}_{\widetilde{r}_{U|Z}} \mathbb{E}_{\widetilde{r}_{W|Z,U}} \left[\log p((\mathbf{y})_{1:N} | (\mathbf{t}, \mathbf{W}, U, Z)_{1:N}; \boldsymbol{\theta})\right] + \mathbb{E}_{\widetilde{r}_Z} \mathbb{E}_{\widetilde{r}_{U|Z}} \left[\log p((\mathbf{t}, U, Z)_{1:N}; \boldsymbol{\theta})\right] \quad (17)$$

where $\widetilde{r}_Z$, $\widetilde{r}_{U|Z}$ and $\widetilde{r}_{W|Z,U}$ denote the following posterior conditional distribution functions at iteration $(i+1)$:

$$\begin{aligned}
\widetilde{r}_Z &= p((Z)_{1:N} | (\mathbf{y}, \mathbf{t})_{1:N}; \boldsymbol{\theta}^{(i)}), \\
\widetilde{r}_{U|Z} &= p((U)_{1:N} | (\mathbf{y}, \mathbf{t}, Z)_{1:N}; \boldsymbol{\theta}^{(i)}), \\
\widetilde{r}_{W|Z,U} &= p((\mathbf{W})_{1:N} | (\mathbf{y}, \mathbf{t}, Z, U)_{1:N}; \boldsymbol{\theta}^{(i)}).
\end{aligned}$$

The proposed EM algorithm iterates over the following E and M steps.

3.1 Expectation step

The expectation step splits into three steps in which the distributions of $\widetilde{r}_Z$, $\widetilde{r}_{U|Z}$ and $\widetilde{r}_{W|Z,U}$ are specified. Some of the computation is similar to that of the Gaussian case and details can be found in [Deleforge et al., 2015].

E-W step. The distribution $\widetilde{r}_{W|Z,U}$ is fully specified by computing the $N \times K$ functions $p(\mathbf{w}_n | \mathbf{y}_n, \mathbf{t}_n, Z_n = k, U_n = u_n; \boldsymbol{\theta}^{(i)})$ which are all Gaussian distribution functions with expectation $\widetilde{\boldsymbol{\mu}}_{nk}^w$ and variance $\widetilde{\boldsymbol{S}}_k^w / u_n$ defined as:

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_{nk}^w &= \widetilde{\boldsymbol{S}}_k^w \left(\boldsymbol{A}_k^{w(i)\top} \boldsymbol{\Sigma}_k^{(i)-1} \left(\mathbf{y}_n - \boldsymbol{A}_k^{t(i)} \mathbf{t}_n - \boldsymbol{b}_k^{(i)} \right) + \boldsymbol{\Gamma}_k^{w(i)-1} \boldsymbol{c}_k^{w(i)} \right) \\
\widetilde{\boldsymbol{S}}_k^w &= \left(\boldsymbol{\Gamma}_k^{w(i)-1} + \boldsymbol{A}_k^{w(i)\top} \boldsymbol{\Sigma}_k^{(i)-1} \boldsymbol{A}_k^{w(i)} \right)^{-1}.
\end{aligned}$$

E-U step. Similarly, $\widetilde{r}_{U|Z}$ is fully defined by computing for $n = 1 : N$ and $k = 1 : K$, the distribution $p(u_n|\mathbf{y}_n, \mathbf{t}_n, Z_n = k; \boldsymbol{\theta}^{(i)})$, which is the density function of a Gamma distribution $\mathcal{G}(u_n, \alpha_k^{t(i)}, \gamma_k^{t(i)})$ with parameters:

$$\alpha_k^{t(i+1)} = \alpha_k^{(i)} + \frac{L_t + D}{2}$$
$$\gamma_k^{t(i+1)} = 1 + \frac{1}{2}\left(\delta\left(\mathbf{y}_n, \boldsymbol{A}_k^{(i)}[\mathbf{t}_n, \boldsymbol{c}_k^{w(i)}]^T + \boldsymbol{b}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)} + \boldsymbol{A}_k^{w(i)}\boldsymbol{\Gamma}_k^{w(i)}\boldsymbol{A}_k^{w(i)\top}\right) + \delta\left(\mathbf{t}_n, \boldsymbol{c}_k^{t(i)}, \boldsymbol{\Gamma}_k^{t(i)}\right)\right).$$

The values $\alpha_k^{t(i+1)}$ and $\gamma_k^{t(i+1)}$ are deduced using that $p(u_n|\mathbf{y}_n, \mathbf{t}_n, Z_n = k; \theta^{(i)})$ is proportional to $p(\mathbf{y}_n|\mathbf{t}_n, Z_n = k, U_n = u_n; \boldsymbol{\theta}^{(i)})p(\mathbf{t}_n|Z_n = k, U_n = u_n; \boldsymbol{\theta}^{(i)})p(u_n|Z_n = k; \boldsymbol{\theta}^{(i)})$ and by noticing that the Gamma distribution is a conjuguate prior for a Gaussian likelihood. As in traditional Student mixtures (see eg [Peel and McLachlan, 2000]), the E-$\boldsymbol{U}$ step actually reduces to the computation of the conditional expectation $\mathbb{E}[U_n|\mathbf{t}_n, \mathbf{y}_n, Z_n = k; \boldsymbol{\theta}^{(i)}]$:

$$\bar{u}_{nk}^{(i+1)} = \mathbb{E}[U_n|\mathbf{t}_n, \mathbf{y}_n, Z_n = k; \boldsymbol{\theta}^{(i)}] = \frac{\alpha_k^{t(i+1)}}{\gamma_k^{t(i+1)}}$$
$$= \frac{\alpha_k^{(i)} + (L_t + D)/2}{1 + 1/2\left(\delta\left(\mathbf{y}_n, \boldsymbol{A}_k^{(i)}[\mathbf{t}_n, \boldsymbol{c}_k^{w(i)}]^T + \boldsymbol{b}_k^{(i)}, \boldsymbol{\Sigma}_k^{(i)} + \boldsymbol{A}_k^{w(i)}\boldsymbol{\Gamma}_k^{w(i)}\boldsymbol{A}_k^{w(i)\top}\right) + \delta\left(\mathbf{t}_n, \boldsymbol{c}_k^{t(i)}, \boldsymbol{\Gamma}_k^{t(i)}\right)\right)}.$$

When $\mathbf{y}_n$ gets away from $\boldsymbol{A}_k^{(i)}\mathbf{x}_n + \boldsymbol{b}_k^{(i)}$ or when $\mathbf{t}_n$ gets away from $\boldsymbol{c}_k^{t(i)}$ or both, then the Mahalanobis distances in the denominator increase and $\bar{u}_{nk}^{(i+1)}$ decreases. $\bar{u}_{nk}^{(i+1)}$ acts as a weight. A low $\bar{u}_{nk}^{(i+1)}$ downweights the impact of $\mathbf{t}_n$ and $\mathbf{y}_n$ in the parameters estimations (see below). In the following, the covariance matrix $\boldsymbol{\Sigma}_k^{(i)} + \boldsymbol{A}_k^{w(i)}\boldsymbol{\Gamma}_k^{w(i)}\boldsymbol{A}_k^{w(i)\top}$ is denoted by $\widetilde{\boldsymbol{S}}_k^u$.

E-Z step. Characterizing $\widetilde{r}_Z$ is equivalent to compute each $r_{nk}^{(i+1)}$ defined as the posterior probability that $(\mathbf{t}_n, \mathbf{y}_n)$ belongs to the kth component of the mixture given the current estimates of the mixture parameters $\boldsymbol{\theta}^{(i)}$:

$$r_{nk}^{(i+1)} = \frac{\pi_k^{(i)} p(\mathbf{t}_n, \mathbf{y}_n|Z_n = k; \boldsymbol{\theta}^{(i)})}{\sum_{j=1}^{K} \pi_j^{(i)} p(\mathbf{t}_n, \mathbf{y}_n|Z_n = j; \boldsymbol{\theta}^{(i)})} \tag{18}$$

where the joint distribution $p(t_n, y_n|Z_n = k; \boldsymbol{\theta}^{(i)})$ is a $L_t + D$ dimensional generalized Student distribution denoted by $\mathcal{S}_{L_t+D}([\mathbf{t}_n, \mathbf{y}_n]^T; \boldsymbol{m}_k^{t(i)}, \boldsymbol{V}_k^{t(i)}, \alpha_k^{(i)}, 1)$ with $\boldsymbol{m}_k^{t(i)}$ and $\boldsymbol{V}_k^{t(i)}$ defined as

$$\boldsymbol{m}_k^{t(i)} = \begin{bmatrix} \boldsymbol{c}_k^{t(i)} \\ \boldsymbol{A}_k^{(i)}\boldsymbol{c}_k^{(i)} + \boldsymbol{b}_k^{(i)} \end{bmatrix}, \boldsymbol{V}_k^{t(i)} = \begin{bmatrix} \boldsymbol{\Gamma}_k^{t(i)} & \boldsymbol{\Gamma}_k^{t(i)}\boldsymbol{A}_k^{t(i)\top} \\ \boldsymbol{A}_k^{t(i)}\boldsymbol{\Gamma}_k^{t(i)} & \boldsymbol{\Sigma}_k^{(i)} + \boldsymbol{A}_k^{(i)}\boldsymbol{\Gamma}_k^{(i)}\boldsymbol{A}_k^{(i)\top} \end{bmatrix}.$$

3.2 Maximization step

The update of the parameters decomposes into three parts. Update equations for $\{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \alpha_k\}$ are derived from the second part of the expected log-likelihood in Expression (17) and can be straightforwardly derived from previous work on Student mixtures, *e.g.* [Peel and McLachlan, 2000]. The update of $\{\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}$ is deduced from the first part of Expression (17) and generalizes the corresponfing step in [Deleforge et al., 2015]. The Student case involves classically some *double weights* accounting for the introduction of an extra latent variable U:

$$\widetilde{r}_{nk}^{(i+1)} = r_{nk}^{(i+1)} \bar{u}_{nk}^{(i+1)} \ .$$

We use also the following notation $\widetilde{r}_{k}^{(i+1)} = \sum_{n=1}^{N} \widetilde{r}_{nk}^{(i+1)}$ and $r_{k}^{(i+1)} = \sum_{n=1}^{N} r_{nk}^{(i+1)}$.

M-$(\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ step. We recover the Student mixture formula. For this part the model behaves as a Student mixture on the $\{\mathbf{t}_n\}_{n=1}^{N}$, which gives the following updates:

$$\begin{aligned}
\pi_k^{(i+1)} &= \frac{r_k^{(i+1)}}{N} \\
\boldsymbol{c}_k^{t(i+1)} &= \sum_{n=1}^{N} \frac{\widetilde{r}_{kn}^{(i+1)}}{\widetilde{r}_k^{(i+1)}} \mathbf{t}_n \\
\boldsymbol{\Gamma}_k^{(i+1)} &= \sum_{n=1}^{N} \frac{\widetilde{r}_{kn}^{(i+1)}}{r_k^{(i+1)}} (\mathbf{t}_n - \boldsymbol{c}_k^{t(i+1)})(\mathbf{t}_n - \boldsymbol{c}_k^{t(i+1)})^\top
\end{aligned}$$

M-α_k step. The estimates do not exist in closed form, but can be computed by setting the following expression to 0 (see [Forbes and Wraith, 2014] for details):

$$\begin{aligned}
&- \Upsilon(\alpha_k) + \Upsilon\left(\alpha_k^{(i)} + \frac{L_t + D}{2}\right) \\
&- \frac{1}{r_k^{(i+1)}} \sum_{n=1}^{N} r_{nk}^{(i+1)} \log\left(1 + \frac{1}{2}\left(\delta(\mathbf{y}_n, \boldsymbol{A}_k^{(i)}[\mathbf{t}_n, \boldsymbol{c}_k^{w(i)}]^T + \boldsymbol{b}_k^{(i)}, \widetilde{\boldsymbol{S}}_k^{u}) + \delta(\mathbf{t}_n, \boldsymbol{c}_k^{t(i)}, \boldsymbol{\Gamma}_k^{t(i)})\right)\right)
\end{aligned}$$

which gives that α_k is estimated by numerically computing:

$$\begin{aligned}
\alpha_k^{(i+1)} = \Upsilon^{-1}\Bigg(&\Upsilon\left(\alpha_k^{(i)} + \frac{L_t + D}{2}\right) \\
&- \frac{1}{r_k^{(i+1)}} \sum_{n=1}^{N} r_{nk}^{(i+1)} \log\left(1 + \frac{1}{2}\left(\delta(\mathbf{y}_n, \boldsymbol{A}_k^{(i)}[\mathbf{t}_n, \boldsymbol{c}_k^{w(i)}]^T + \boldsymbol{b}_k^{(i)}, \widetilde{\boldsymbol{S}}_k^{u})) + \delta(\boldsymbol{t}_n, \boldsymbol{c}_k^{t(i)}, \boldsymbol{\Gamma}_k^{t(i)})\right)\right)\Bigg)
\end{aligned}$$

where Υ is the Digamma function that verifies $E[\log W] = \Upsilon(\alpha) - \log\gamma$ when W follows a $\mathcal{G}(\alpha, \gamma)$ distribution. The Digamma function also satisfies $\frac{d\log\Gamma(\alpha)}{d\alpha} = \Upsilon(\alpha)$.

M-($\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k$) step. The updating of the mapping parameters $\{\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ is also in closed-form and is obtained by maximizing the first part in (17). It is easy to see that the results in [Deleforge et al., 2015] can be used by replacing $r_{nk}^{(i+1)}$ with $\widetilde{r}_{nk}^{(i+1)}$ to account for the extra $\boldsymbol{U}$ variables. It follows,

$$\boldsymbol{A}_k^{(i+1)} = \widetilde{\mathbf{Y}}_k \widetilde{\mathbf{X}}_k^\top (\widetilde{\boldsymbol{S}}_k^x + \widetilde{\mathbf{X}}_k \widetilde{\mathbf{X}}_k^\top)^{-1} \tag{19}$$

$$\boldsymbol{b}_k^{(i+1)} = \sum_{n=1}^N \frac{\widetilde{r}_{kn}^{(i+1)}}{\widetilde{r}_k^{(i+1)}} (\mathbf{y}_n - \boldsymbol{A}_k^{(i+1)} \widetilde{\mathbf{x}}_{nk}) \tag{20}$$

$$\boldsymbol{\Sigma}_k^{(i+1)} = \boldsymbol{A}_k^{w(i+1)} \widetilde{\boldsymbol{S}}_k^w \boldsymbol{A}_k^{w(i+1)\top} + \sum_{n=1}^N \frac{\widetilde{r}_{kn}^{(i+1)}}{r_k^{(i+1)}} (\mathbf{y}_n - \boldsymbol{A}_k^{(i+1)} \widetilde{\mathbf{x}}_{nk} - \boldsymbol{b}_k^{(i+1)})(\mathbf{y}_n - \boldsymbol{A}_k^{(i+1)} \widetilde{\mathbf{x}}_{nk} - \boldsymbol{b}_k^{(i+1)})^\top \tag{21}$$

where

$$\begin{aligned}
\widetilde{\mathbf{x}}_{nk} &= [\mathbf{t}_n, \widetilde{\boldsymbol{\mu}}_{nk}^w]^T \\
\widetilde{\mathbf{Y}}_k &= \frac{1}{\sqrt{r_k}} \left[\sqrt{\widetilde{r}_{1k}} (\mathbf{y}_1 - \widetilde{\mathbf{y}}_k), \ldots, \sqrt{\widetilde{r}_{Nk}} (\mathbf{y}_N - \widetilde{\mathbf{y}}_k) \right] \\
\widetilde{\mathbf{y}}_k &= \sum_{n=1}^N \frac{\widetilde{r}_{kn}}{r_k} \mathbf{y}_n \\
\widetilde{\mathbf{X}}_k &= \frac{1}{\sqrt{r_k}} \left[\sqrt{\widetilde{r}_{1k}} (\widetilde{\mathbf{x}}_{1k} - \widetilde{\mathbf{x}}_k), \ldots, \sqrt{\widetilde{r}_{Nk}} (\widetilde{\mathbf{x}}_{Nk} - \widetilde{\mathbf{x}}_k) \right] \\
\widetilde{\mathbf{x}}_k &= \sum_{n=1}^N \frac{\widetilde{r}_{kn}}{r_k} \widetilde{\mathbf{x}}_{nk} \\
\widetilde{\boldsymbol{S}}_k^x &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{S}}_k^w . \end{bmatrix}
\end{aligned}$$

3.3 Constrained estimations

The **E** and **M** steps above are given for general $\boldsymbol{\Sigma}_k$. However in practice, for high D, a great gain in complexity can be achieved by imposing some simplifying constraints on $\boldsymbol{\Sigma}_k$. When $\boldsymbol{\Sigma}_k$ is assumed diagonal, it can be estimated by the diagonal of $\boldsymbol{\Sigma}_k^{(i+1)}$ given by (21). In the isotropic case, $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbb{1}_D$, we just need to compute

$$\sigma_k^{2(i+1)} = \frac{trace(\boldsymbol{\Sigma}_k^{(i+1)})}{D}. \tag{22}$$

In the isotropic and equal case, $\mathbf{\Sigma}_k = \sigma^2 \mathbb{I}_D$ for all k, the unique variance parameter is then updated by

$$\sigma^{2(i+1)} = \sum_{k=1}^{K} \pi_k^{(i+1)} \sigma_k^{2(i+1)},$$

with the expression (22) just above.

3.4 Initialization

Initial values for the **E-U** and **E-Z** steps are natural: the $\bar{u}_{nk}^{(0)}$'s can be set to 1 while the $r_{nk}^{(0)}$'s can be set to the values obtained with a standard EM algorithm for a K-component Gaussian mixture on $(\mathbf{Y}, \mathbf{T})$. Steps that involve $\mathbf{W}$ are less straightforward to initialize. Therefore, one solution is to consider a marginal EM algorithm in which the latent variable $\mathbf{W}$ is integrated out. Considering Expression (17), one can see that the **E-Z** and **E-U** steps are unchanged and that the **E-W** step is removed. With $\boldsymbol{c}_k^w$ and $\mathbf{\Gamma}_k^w$ fixed to $\mathbf{0}_{L_w}$ and $\mathbb{I}_{L_w}$ respectively, the estimation of $(\pi_k, \boldsymbol{c}_k^t, \mathbf{\Gamma}_k^t)$ and α_k is unchanged in the M-step. The estimation of $(\boldsymbol{A}_k, \boldsymbol{b}_k, \mathbf{\Sigma}_k)$ only involves the observed data $(\mathbf{t}_n)_{n=1:N}$ and can be performed in two steps, a regression step for $(\boldsymbol{A}_k^t, \boldsymbol{b}_k)$ and a PPCA-like step for $(\boldsymbol{A}_k^w, \mathbf{\Sigma}_k)$:

M-$(\boldsymbol{A}_k^t, \boldsymbol{b}_k)$ step.

$$\begin{aligned} \boldsymbol{A}_k^{t(i+1)} &= \widetilde{\mathbf{Y}}_k \widetilde{\mathbf{T}}_k^\top (\widetilde{\mathbf{T}}_k \widetilde{\mathbf{T}}_k^\top)^{-1} \\ \boldsymbol{b}_k^{(i+1)} &= \sum_{n=1}^{N} \frac{\widetilde{r}_{nk}}{\widetilde{r}_k} (\mathbf{y}_n - \boldsymbol{A}_k^{t(i+1)} \mathbf{t}_n) \end{aligned} \tag{23}$$

where

$$\begin{aligned} \widetilde{\mathbf{T}}_k &= \left[\frac{\sqrt{\widetilde{r}_{1k}}}{\sqrt{\widetilde{r}_k}} (\mathbf{t}_1 - \widetilde{\mathbf{t}}_k), \ldots, \frac{\sqrt{\widetilde{r}_{Nk}}}{\sqrt{\widetilde{r}_k}} (\mathbf{t}_N - \widetilde{\mathbf{t}}_k) \right] \\ \widetilde{\mathbf{t}}_k &= \sum_{n=1}^{N} \frac{\widetilde{r}_{nk}}{\widetilde{r}_k} \mathbf{t}_n \end{aligned}$$

M-$(\boldsymbol{A}_k^w, \mathbf{\Sigma}_k)$ step. Updates are obtained by minimizing the following criterion:

$$\begin{aligned} Q_k(\mathbf{\Sigma}_k, \boldsymbol{A}_k^w) = & \log |\mathbf{\Sigma}_k + \boldsymbol{A}_k^w \boldsymbol{A}_k^{w\top}| \\ & + \sum_{n=1}^{N} \frac{\widetilde{r}_{nk}}{r_k} (\mathbf{y}_n - \boldsymbol{A}_k^{t(i+1)} \boldsymbol{t}_n - \boldsymbol{b}_k^{(i+1)})^\top (\mathbf{\Sigma}_k + \boldsymbol{A}_k^w \boldsymbol{A}_k^{w\top})^{-1} (\boldsymbol{y}_n - \boldsymbol{A}_k^{t(i+1)} \mathbf{t}_n - \boldsymbol{b}_k^{(i+1)}) \ . \end{aligned}$$

More details on the practical resolution are given in [Deleforge et al., 2015]. In practice, only one iteration of this marginal EM is run to initialize the complete EM.

4 Model selection issues

The SLLiM model relies on the preliminary choice of two numbers, K the number of local linear regressions and L_w the number of additional latent variables. We mention below simple ways to select such values. A more thorough study of this issue would be useful but it is out of the scope of the present paper.

Determining the number of clusters K. This number can be equivalently interpreted as the number of affine regressions or as the number of mixture components. In this latter case, it is known that regularity conditions do not hold for the Chi-squared approximation used in the likelihood ratio test statistic to be valid. As an alternative, penalized likelihood criteria like the Bayesian Information Criterion (BIC) are often used for clustering issues because interpretation of the results may strongly depend on the number of clusters. In our piecewise regression context, the specific value of K may be less important. As pointed in [Deleforge et al., 2015], the number of affine approximations K can be arbitrarily set to a large enough value (*e.g.* $K = 50$) as the results are not very sensitive to the choice of this parameter. In the same manner for SLLiM, K can be set to an arbitrary value large enough to catch non linear relationships in a D-dimensional space, while being vigilant that the number of observations is large enough to allow a stable fit of all K components. In practice, we will compare this solution to the one using BIC to select K.

Determining the number of latent variables L_w. The selection of L_w is similar to the issue of selecting the number of factors in a factor analyzer model [Baek et al., 2010]. Regularity conditions usually hold for tests on the number of factors but like in [Baek et al., 2010], we rather investigate the use of BIC for choosing L_w. When K is not fixed, BIC can be computed for varying couples (K, L_w) but the available sample size usually limits the range of values that can be tested with reliable BIC computation. For this reason, if necessary we will rather fix K to some value not too large so as to be able to investigate a larger range of L_w values.

5 Experiments on real data

5.1 An illustrative example: Paris' subway air quality

We first consider a low dimensional example with $D = L = 1$ to illustrate the ability of Student distributions to reduce the effect of outliers in a regression context. This is a simple case that implies $L_w = 0$ but it has the advantage to provide results that can be understood and displayed graphically. Rather than using simulated data, we consider a data set provided by the RATP group which is a public operator in charge of public transportation in Paris. This institution also studies the air quality in the subway network. Gas concentrations and meteorological features are measured since 2013 every day,

all day long. The dataset of interest contains the measures for three stations including Châtelet, Franklin Roosevelt and Auber[3]. We focus here on the Châtelet station on line 4 in Paris. Every day at each hour, a number of features are measured including some usual meteorological parameters (temperature, humidity), air exchange (Carbon dioxide) and air quality (Nitrogen oxides and particles). Although the issue is usually to predict pollutants concentrations over time or as function of meteorology, we do not focus here on this aspect because meteorological and time variables are discretized in this data set. We focus instead on the relationship between the Nitrogen monoxide concentration (NO, measured in μg/m^3) and Nitrogen dioxide concentrations (NO_2 measured in μg/m^3). They are both toxic air pollutants whose respective concentrations exhibit an interesting non linear relationship (grey points in Figure 1 (a)) relevant for our illustration purpose. The dataset consists in 341 measures during March 2015.

SLLiM is compared to its Gaussian counterpart GLLiM [Deleforge et al., 2015]. Figure 1 (a) shows the regression curves fitted by GLLiM (blue) and SLLiM (black) for K=2. The $K = 2$ case clearly illustrates that although the model is built on only 2 underlying linear transformations, the obtained regression curves are not piecewise affine but show some clear nonlinearity. More importantly, Figure 1 (a) illustrates that the SLLiM fitted curve is less affected by outliers (extreme large values of NO). We identified as outliers 6 data points with a NO concentration greater than 200. We then considered a data set without these 6 observations. Figure 1 (b) shows the obtained regression curves after removing these 6 points. The estimated GLLiM curve gets closer to the SLLiM one confirming the sensitivity of GLLiM to oultiers.

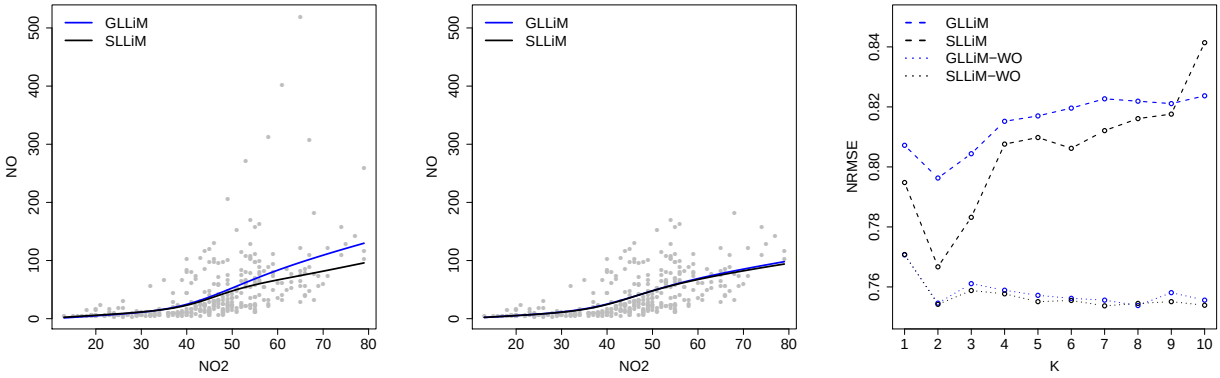
The prediction error is assessed using the normalized root mean squared error (NRMSE[4]). The NRMSE is a normalized version of the RMSE in which we compare the prediction rate to the one reached by predicting all responses by the mean of the training responses, independently of the covariates. A NRMSE equal to 1 means that the method performs as well as one that would set all predictions to the training responses mean. The smaller the NRMSE the better. A 100-fold cross-validation is performed by randomly sampling data into sets of 300 training observations using the 41 remaining ones (resp. 35 if outliers are removed) for testing.

Figure 1 (c) displays the average NRMSE computed by cross-validation for SLLiM (black) and GLLiM (blue) on the complete data (long dashed lines) and data without 6 outliers (dotted lines), for 10 values of K from 1 to 10. This plot shows that $K = 2$ is optimal in terms of the reconstruction with a minimum average NRMSE at this value. Note that we found that $K = 8$ provided a slightly lower BIC than $K = 2$. As already observed in [Baek et al., 2010] in a clustering context, it seems that BIC does not always lead to a choice of K that simultaneously minimizes BIC and the prediction error. The figure

[3]It can be downloaded at
`http://data.ratp.fr/explore/dataset/qualite-de-lair-mesuree-dans-la-station-chatelet`

[4]NRMSE $= \sqrt{\frac{\sum_i (t_i - \hat{t}_i)^2}{\sum_i (t_i - \bar{t}_{\text{train}})^2}}$

Figure 1: Nitrogen oxide with respect to Nitrogen dioxide during March 2015



(a) complete data (b) after removing 6 outliers (c) average NRMSE

also illustrates the robustness of the proposed model as SLLiM achieves a better prediction rate than GLLiM when performed on data with outliers (long-dashed lines). The difference between GLLiM and SLLiM reduces when some outliers are removed (dotted lines).

5.2 Application on real high dimensional data

In this section, the performance of the proposed method is assessed on two datasets. The following subsection illustrates the properties of the model on data with a small number of observations regarding to the number of variables and the second subsection investigates the contribution of SLLiM over the Gaussian version (GLLiM) on a dataset for which GLLiM is already performing well.

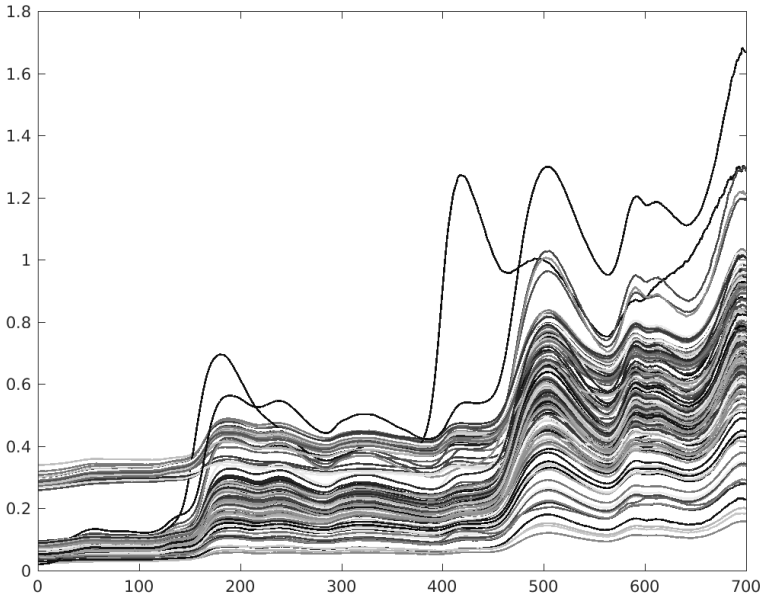
5.2.1 Orange juice dataset

The proposed method is now applied to the Orange juice public dataset in order to illustrate that SLLiM is competitive in high dimensional settings with $D \approx N$. The goal is to assess the efficiency of SLLiM in such a setting and to illustrate that latent factors $\mathbf{W}$ introduced in the model are useful to catch dependency among features.

Data. The data contains near-infrared spectra measured on $N = 218$ orange juices[5]. The length of each spectrum is 700 and the aim is to model the relationship between the level of sucrose ($L = 1$) and the spectra. Figure 2 shows the N spectra. The curves are quite similar, even if some spectra appear to have extreme values and exhibit isolated peaks.

[5]It can be downloaded at `http://www.ucl.ac.be/mlg/index.php?page=DataBases` or from the open-source `cggd R` package available on the CRAN in the object `data(OJ)`.

Figure 2: Curves of orange juice spectra



Method. First, spectra are decomposed on a splines basis using the `smooth.splines` function of the `R` software. $D = 134$ knots are retained. Reducing the data to the splines coefficients makes the variables exchangeable and is also convenient to reduce the dimension while preserving information. The following methods are then compared:

- Two versions of the proposed model of robust non linear regression (SLLiM). In a first version (denoted by "SLLiM (BIC)"), the numbers of clusters K and latent variables L_w are both estimated by BIC. In a second version (denoted by "SLLiM (K=10)"), the number of clusters is set to $K = 10$, which is large enough to catch non linear relationships regarding to the dimension of the data. The number of latent factors L_w is estimated by BIC.
- A Gaussian version of our model (GLLiM [Deleforge et al., 2015]) using the Matlab toolbox available at `http://team.inria.fr/perception/gllim_toolbox/`. Numbers K and L_w are chosen as above.
- Random forests [Breiman, 2001] performed using the default options of the R package `randomForest`.
- Multivariate Adaptive Regression Splines [Friedman, 1991] using the `mars` function of the `mda R` package.
- Support Vector Machine (SVM [Vapnik, 1998]) performed with several kernels (linear, Gaussian and polynomial) as in the R package `e1071` [Karatzoglou et al., 2006].
- Sliced Inverse Regression (SIR [Li, 1991]) followed by a polynomial regression of degree 3 performed on the SIR components. We assess predictions with 1 to 10 directions, using the `dr` function of the `dr R` package which implements dimension reduction methods.
- Relevant Vector Machine (RVM)[6] which is known to perform better in some cases than SVM [Tipping, 2001]. We compare results achieved by several kernels (Gaussian, linear and Cauchy which is a heavy-tailed kernel) for 60 values of the scale parameter from 0.1 to 6 with a 0.1 increment.

The prediction accuracy is evaluated using a Leave-One-Out cross-validation (LOO-CV). The model is estimated on training data sets of size 217 and a normalized prediction error is computed by predicting the sucrose level of the 1 remaining observation. Each method is therefore assessed 218 times. As the number of observations is small regarding to the number of variables, the presence of outliers in the testing dataset in the CV generates artificially bad predictions which are not absorbed by the size of the testing sample. For these reasons, the computed NRMSE is affected by outliers: large prediction errors are observed on outlying data points. We therefore compute the median instead of the mean of the NRMSE values to get a better insight on the respective methods prediction performance.

[6] We use the Matlab code available at `http://mi.eng.cam.ac.uk/ at315/MVRVM.htm`.

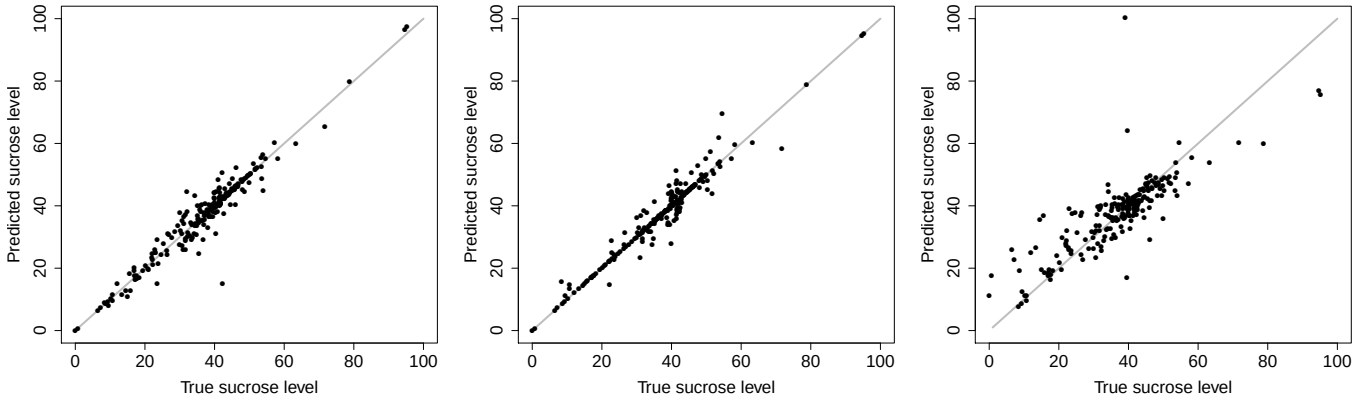
Table 1: LOO-CV results for Orange juice data after decomposition on splines. Median NRMSE and % of outliers in parenthesis.

Procedure	Median NRMSE (% outliers)
SLLiM (BIC)	**0.420 (22.93)**
SLLiM (K=10)	**0.388 (27.06)**
SLLiM–0 (K=10)	0.885 (45.87)
GLLiM (BIC)	0.623 (34.86)
GLLiM (K=10)	0.466 (29.36)
GLLiM–0 (K=10)	1.022 (50.46)
Random forests	0.589 (31.19)
MARS	0.629 (33.03)
SVM	**0.425 (24.77)**
SIR	1.020 (51.83)
RVM	0.536 (33.49)

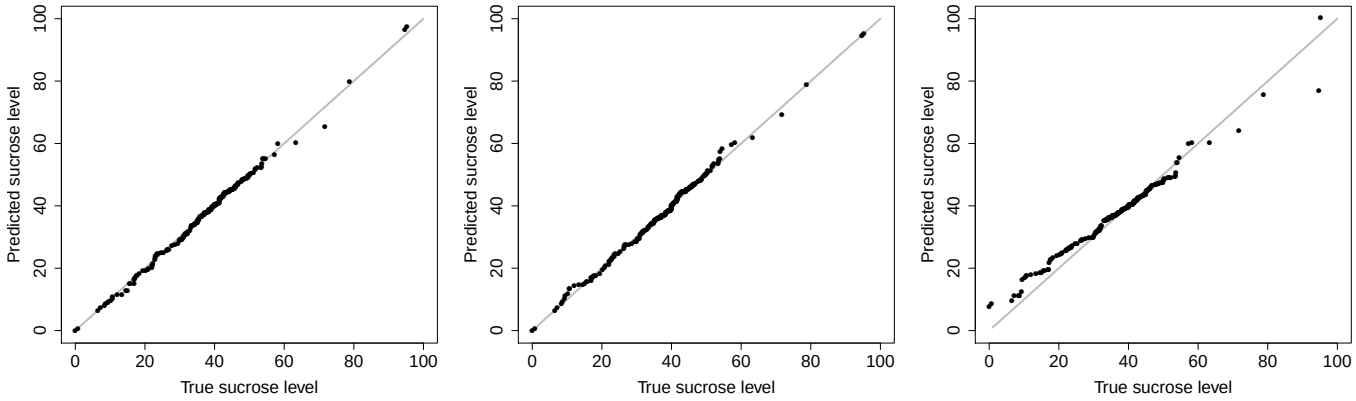
Results. Table 1 shows the median of the NRMSE and percentage of outliers for the compared methods. Outliers are defined as runs leading to an error greater than the error obtained using the training data set mean as predictor. Results are presented with parameters values leading to the best results, namely linear kernel for SVM, 1 direction for SIR and Gaussian kernel with scale parameter set to 0.70 for RVM. For both GLLiM and SLLiM, setting the number of clusters to 10 and choosing the number of latent factors with BIC leads to the best prediction. For SLLiM, BIC criterion retained 9 to 12 latent factors for 91% of the CV-runs (similar proportions for GLLiM). Selecting K by BIC leads to values between 8 and 10 for 96 % of the CV-runs. For the two different ways to select K and L_w, SLLiM always outperforms its Gaussian counterpart. In Table 1, GLLiM-0 (resp. SLLiM-0) denotes the results for $K = 10$ and $L_w = 0$. It shows worse predictions for both GLLiM and SLLiM and illustrates the advantage of adding latent factors to the model. Results for $L_w = 0$ and K selected by BIC are not presented but are similar. Among the other compared methods, the best prediction error is obtained using SVM with a linear kernel. When choosing both K and L_w with BIC, SLLiM achieves the same prediction rate. However, SLLiM performs better than SVM when K is fixed to 10 and L_w chosen by BIC. RVM, SIR, MARS and random forests are not competitive on this example.

Figure 3 presents the adjustment quality for SVM (linear kernel), SLLiM and GLLiM ($K = 10$, L_w estimated using BIC). The first row shows the predicted sucrose levels against the true ones and the second row shows quantile vs quantile plots (QQ-plots) of the predicted sucrose levels as a function of the observed ones. These plots illustrate graphically that SLLiM achieves the best adjustment of the observed responses in particular compare to SVM.

Figure 3: Adjustment on training data



(a) GLLiM (b) SLLiM (c) SVM



(d) GLLiM-QQ-plot (e) SLLiM-QQ-plot (f) SVM-QQ-plot

5.2.2 Hyperspectral data from Mars

We now investigate the effect of additional robustness on a dataset already studied in [Deleforge et al., 2015] and for which good results were already observed with GLLiM.

Data. As described into more details in [Deleforge et al., 2015], the dataset corresponds to hyperspectral images of the Mars planet. Spectra are acquired at several locations on the planet and the goal is to recover from each spectrum some physical characteristics of the surface at each location. To do so, a training dataset is available made of $N = 6983$ spectra of length $D = 184$ synthesized from surface characteristics using a radiative transfer model designed by experts. A testing dataset is then available corresponding to spectra acquired on the south polar cap of Mars and for which the true surface characteristics are not known. More specifically, we focus on recovering two quantities, the proportion of CO_2 ice and the proportion of dust. Our observed response variable is therefore bivariate with $L_t = 2$.

Method. The same methods as in the previous section are compared, except SVM and random forests which cannot handle multivariate responses. For GLLiM and SLLiM, the number of clusters K is estimated by BIC or fixed to $K = 10$. A smaller value of K was chosen compare to [Deleforge et al., 2015] for several reasons. We observe that for larger values of K the likelihood exhibits some noisy behaviour and is not regularly increasing as it should. We therefore suspect that the number of training data may be too small when the number of parameters increases. Then, as mentioned earlier, we are rather interested in investigating the choice of L_w and for the previous reason, it may not be reliable to both increase K and L_w considering the available sample size in this example. The number of additional latent variables L_w is chosen using BIC. The prediction accuracy is first evaluated on the training set using the NRMSE and a cross-validation setting. 100 datasets of size 6000 are randomly sampled among the $N = 6983$ observations and NRMSE are computed by predicting simultaneously the proportions of dust and CO_2 ice on the 983 remaining observations.

Results on training data. Table 2 presents the prediction accuracy achieved by the different tested methods. For SIR, best prediction rates are achieved for 10 directions. For RVM, optimal results are obtained with a Cauchy kernel (heavy-tailed kernel) and a scale parameter set to 1. SLLiM performs better predictions than GLLiM. Among other methods, regression splines (MARS) achieves the best predictions but slightly worse than SLLiM. SIR achieves good prediction rates for the proportion of dust but not for the proportion of CO_2 ice and RVM provides the worst results in this example.

Application to Mars surface properties retrieval from hyperspectral images. The training of the radiative model database can then be used to predict proportions of

Table 2: Mars data: average NRMSE and standard deviations in parenthesis for proportions of CO_2 ice and dust over 100 runs.

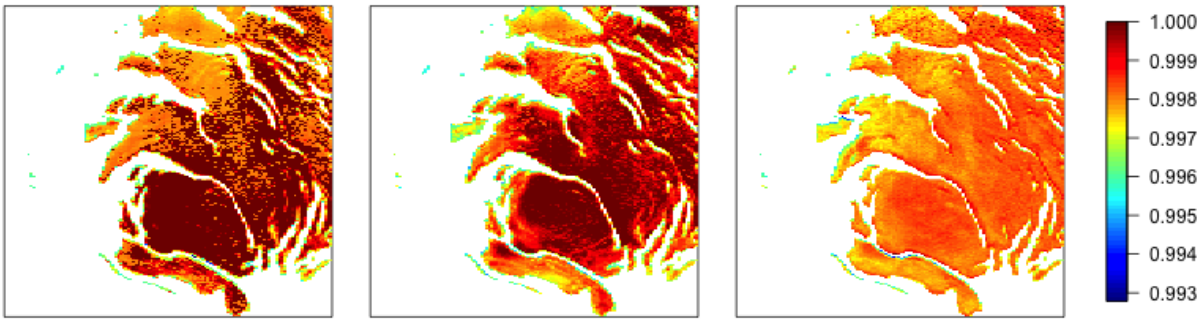
Method	Prop. of CO_2 ice	Prop. of dust
SLLiM (BIC)	0.258 (0.035)	0.257 (0.043)
SLLiM (K=10)	**0.168 (0.019)**	**0.145 (0.020)**
GLLiM (BIC)	0.197 (0.024)	0.173 (0.022)
GLLiM (K=10)	**0.180 (0.023)**	**0.155 (0.023)**
MARS	**0.173 (0.016)**	**0.160 (0.021)**
SIR	0.243 (0.025)	0.157 (0.016)
RVM	0.299 (0.021)	0.275 (0.034)

interest from real observed spectra acquired as images. In particular, we focus on a dataset of Mars South polar cap corresponding to a 128×265 image [Bernard-Michel et al., 2009]. Since no ground truth is currently available for the physical properties of Mars polar regions, we propose a qualitative evaluation using the three best performing methods among the tested ones, namely SLLiM and GLLiM with $K = 10$ and MARS. Figure 4 shows the obtained images for CO_2 ice and dust proportions. All methods appear to match satisfyingly the expected results from planetology experts. Indeed, the observed region is expected to be made of CO_2 ice with increasing amount of dust at the borders with non icy regions. All retrieved images in Figure 4 show satisfyingly this dust proportion variation. The main difference between the methods lies in the proportions ranges. SLLiM provides CO_2 ice proportions much higher in the central part of the cap, while MARS provides smoother values all over the cap. According to SLLiM the CO_2 ice would be purer with almost no dust in the central part.

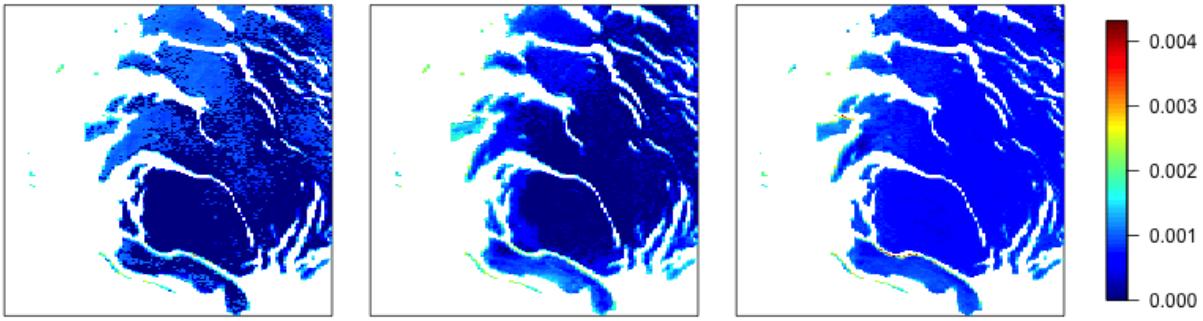
6 Conclusion

We proposed a new robust non linear regression model for high dimensional data based on Student mixture distributions. Non linearity is captured via a piecewise linear modelling into a number of simple linear regressions while extra robustness to noise is expected from the possibility to augment the regression setting with a number of latent variables. In high dimension, the tractability of our model lies upon an inverse regression strategy. In this context, the use of Student distributions has at least two advantages. They allow the random vectors to be heavy tailed and have tail dependence making the model less sensitive to outliers while maintaining its tractability thanks to their representation as scale mixture of Gaussians. These extra flexibilities make our proposal suitable for modeling real world data for which high dimension and outliers contamination may combine and complicate the analysis. Experiments on synthetic and real world data have been conducted to illustrate

Figure 4: Orbit 41



(a) GLLiM - CO2 ice (b) SLLiM - CO2 ice (c) Regression splines - CO2 ice



(d) GLLiM - Dust (e) SLLiM - Dust (f) Regression splines - Dust

the empirical usefulness of the proposed method. In practice, real applications raise the issue of selecting appropriate numbers for the number of linear regressions and the number of latent variables. We proposed to use a standard BIC to deal with this issue with satisfying results but this is certainly one aspect that could be further investigated, in particular in a context where the number of available training data may not be large enough for selection criteria to be theoretically reliable. Then, in this paper the main target was to address the fact that an outlier may distort the derivation of the local linear mappings so as to fit the outlier well, and therefore may result in wrong parameter estimation. Outlier contamination may then pose distortion to model building and subsequent prediction. While we could indeed check on low dimensional synthetic examples that our modelling was effective in dealing with this issue, in actual high dimensional examples it is often only possible to check the better adjustment of our model in terms of prediction errors. The concept of outlier in a high dimensional space is not obvious and it would be interesting to investigate the use of our model for actual outlier detection, examining for instance the intermediate weight variables computed in the EM algorithm. Also, our derivations would be similar for other members in the scale mixture of Gaussians family or among other elliptical distributions. It would be interesting to study in a regression context, the use of distributions with even more flexible tails such as multiple scale Student distributions [Forbes and Wraith, 2014] or various skew-t [Lee and McLachlan, 2014, Lin, 2010] or Normal Inverse Gaussian distributions [O'Hagan et al., 2014, Wraith and Forbes, 2015]. At last, another interesting direction of research would be to further complement the model with sparsity inducing penalties in particular for situations where interpreting the influential covariates is important.

A Comparison with a cluster-weighted modelling approach

Although our model was focused on regression aspects, it shares some similarity with the so-called CWtFA clustering technique of [Subedi et al., 2015] that uses factor decompositions of the high dimensional covariance matrices. To illustrate the difference, we apply SLLiM to the data sets used in [Subedi et al., 2015] which are however not high dimensional: the `f.voles` data from the `Flury` R package and the `UScrime` data from the `MASS` package. The first data set is made of 86 observations divided into two known species of female voles Microtus californicus (41 individuals) and M. ochrogaster (45 individuals). The goal is to predict age ($L = 1$) on the basis of skull measurements ($D = 6$). The second data set contains aggregate measurements on 47 states of the USA and the goal is to investigate the relationship between the crime rate ($L = 1$) and a number of covariates ($D = 13$). An additional grouping variable is available that indicates the 16 Southern states.

Table 3 shows the clustering results for CWtFA and SLLiM. As the purpose of this study is to compare the clustering returned by these two methods, we set K to 2. We consider a model for SLLiM equivalent to the one considered in [Subedi et al., 2015] with

Table 3: SLLiM and CWtFA clustering results for `f.voles` (a) and `UScrime` (b) data sets. The goal is to assess how the two estimated clusters (columns) fit the two known ones (lines).

		Cluster	
		1	2
CWtFA-CCCU	M. Ochrogaster	43	2
	M. Californicus	0	41
SLLiM-K=2-Lw=1	M. Ochrogaster	43	2
	M. Californicus	0	41

(a) Clustering on `f.voles` dataset library `Flury`

		Cluster	
		1	2
CWtFA-UUUU	Not Southern	30	1
	Southern	3	13
SLLiM-K=2-Lw=1	Not Southern	9	22
	Southern	0	16

(b) Clustering on `UScrime` dataset library `MASS`

the same assumptions: full covariance matrix for covariates, constant across groups for the `f.voles` data and unconstrained for the `UScrime` data. For the `f.voles` data, as for the CWtFA model, BIC selects 1 latent factor. The clustering returned by SLLiM is similar to the one returned by CWtFA. On this dataset, the specie appears to be a relevant discriminant variable between individuals as the clustering separates subjects according to this variable (except 2 errors). For the `UScrime` data, as for CWtFA, BIC selects 1 latent factor. The clustering returned by SLLiM is different from the one returned by CWtFA which illustrates that SLLiM and CWtFA are not equivalent. Moreover, in contrast to the CWtFA result which finds 2 clusters, the other methods compared in [Subedi et al., 2015] indicate that the estimated number of clusters is usually larger (3 clusters). This suggests that the variable of interest (Southern states) is not the only discriminant variable between states. For example, we suspect differences could exist between East and West states but state labels are not available and results cannot be further analyzed.

References

[Adragni and Cook, 2009] Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A*, 367(1906):4385–4405.

[Archambeau and Verleysen, 2007] Archambeau, C. and Verleysen, M. (2007). Robust Bayesian clustering. *Neural Networks*, 20(1):129–138.

[Baek et al., 2010] Baek, J., McLachlan, G. J., and Flack, L. K. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1298–1309.

[Bernard-Michel et al., 2009] Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L., and Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research: Planets*, 114(E6).

[Bishop and Svensen, 2005] Bishop, C. M. and Svensen, M. (2005). Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252.

[Bouveyron et al., 2007] Bouveyron, C., Girard, S., and Schmid, C. (2007). High dimensional data clustering. *Comput. Statist. Data Analysis*, 52(1):502–519.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Chamroukhi, 2015] Chamroukhi, F. (2015). Non-normal mixtures of experts. *ArXiv e-prints*.

[Cook, 2007] Cook, D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26.

[de Veaux, 1989] de Veaux, R. D. (1989). Mixtures of linear regressions. *Computational Statistics and Data Analysis*, 8(3):227–245.

[Deleforge et al., 2015] Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911.

[Devijver, 2015] Devijver, E. (2015). Finite mixture regression: A sparse variable selection by model selection for clustering. *Electronic Journal of Statistics*, 9:2642–2674.

[Forbes and Wraith, 2014] Forbes, F. and Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. *Statistics and Computing*, 24(6):971–984.

[Friedman, 1991] Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19(1):1–141.

[Frühwirth-Schnatter, 2006] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics.

[Garcia-Escudero et al., 2015] Garcia-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., and Mayo-Iscar, A. (2015). Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *ArXiv e-prints*.

[Gershenfeld, 1997] Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, 808(1):18–24.

[Goldfeld and Quandt, 1973] Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1(1):3 – 15.

[Hennig, 2000] Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17:273–296.

[Ingrassia et al., 2012] Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of classification*, 29(3):363–401.

[Karatzoglou et al., 2006] Karatzoglou, A., Meyer, D., and Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, 15(9):1–28.

[Kotz and Nadarajah, 2004] Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University press.

[Lee and McLachlan, 2014] Lee, S. and McLachlan, G. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24:181–202.

[Li, 1991] Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

[Lin, 2010] Lin, T. (2010). Robust mixture modelling using multivariate skew-t distribution. *Statistics and Computing*, 20:343–356.

[O'Hagan et al., 2014] O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P., and Karlis, D. (2014). Clustering with the multivariate Normal Inverse Gaussian distribution . *Computational Statistics and Data Analysis.*

[Peel and McLachlan, 2000] Peel, D. and McLachlan, G. (2000). Robust mixture modeling using the t distribution. *Statistics and Computing*, 10(4):339–348.

[Rosipal and Krämer, 2006] Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In Saunders, C., Grobelnik, M., Gunn, S., and Shawe-Taylor, J., editors, *Subspace, Latent Structure and Feature Selection*, volume 3940 of *Lecture Notes in Computer Science*, pages 34–51. Springer.

[Städler et al., 2010] Städler, N., Bühlmann, P., and vande Geer, S. (2010). L1-penalization for mixture regression models. *TEST*, 19(2):209–256.

[Subedi et al., 2013] Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, 7(1):5–40.

[Subedi et al., 2015] Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods and Applications*, 24(4):623–649.

[Tipping, 2001] Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research 1*, pages 211–244.

[Vapnik, 1998] Vapnik, V. (1998). Statistical learning theory.

[Wraith and Forbes, 2015] Wraith, D. and Forbes, F. (2015). Location and scale mixtures of gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. *Computational Statistics and Data Analysis*, 90:61 – 73.

[Wu, 2008] Wu, H. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610.

[Xu et al., 1995] Xu, L., Jordan, M., and Hinton, G. (1995). An alternative model for mixtures of experts. *Advances in neural information processing systems*, pages 633–640.

[Yao et al., 2014] Yao, W., Wei, Y., and Yu, C. (2014). Robust mixture regression using the t-distribution. *Computational Statistics and Data Analysis*, 71:116–127.