



Non linear robust regression in high dimension

Emeline Perthame, Florence Forbes, Brice Olivier, Antoine Deleforge

► **To cite this version:**

Emeline Perthame, Florence Forbes, Brice Olivier, Antoine Deleforge. Non linear robust regression in high dimension. The XXVIIIth International Biometric Conference, Jul 2016, Victoria, Canada. 2016. <hal-01423622>

HAL Id: hal-01423622

<https://hal.archives-ouvertes.fr/hal-01423622>

Submitted on 30 Dec 2016

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.

Non linear robust regression in high dimension

E. Perthame*, F. Forbes*, B. Olivier*, A. Deleforge**

*MISTIS, INRIA Grenoble, France, **PANAMA, INRIA Rennes, France

1 - Non linear mapping problem

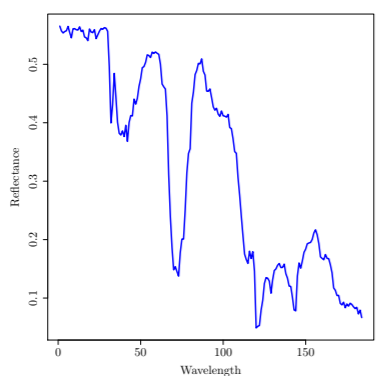
- The goal is to retrieve \mathbf{X} from \mathbf{Y} through a **non linear** regression function g

$$\mathbb{E}(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = g(\mathbf{y})$$

with $Y \in \mathbb{R}^D, X \in \mathbb{R}^L, D \gg L$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_D \end{pmatrix} \xrightarrow{g(\mathbf{y})} \begin{pmatrix} x_1 \\ \vdots \\ x_L \end{pmatrix} = \mathbf{x}$$

- For example, \mathbf{Y} is a reflectance spectrum ($D = 184$) measured at a specific location of the Mars planet and \mathbf{X} is the composition of the ground at this location ($L = 3$)



prop. of dust
prop. of CO₂ ice
prop. of water ice

2 - Difficulties

- High dimension ($D \gg L$) → Inverse regression strategy

$$\mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = f(\mathbf{x})$$

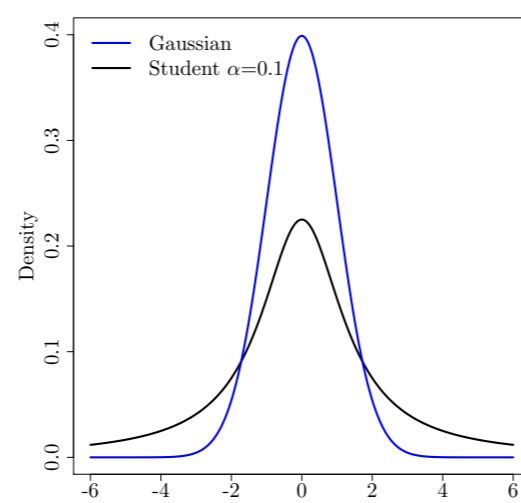
- Non linear mapping → Piecewise linear approximation of f (and g)

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

with $\mathbb{E}(\mathbf{E}_k^2) \propto \Sigma_k$ and Z multinomial latent variable

$$\mathbb{P}(Z = k) = \pi_k$$

- Dealing with outliers → Heavy tail distribution
→ Generalized Student distribution



$$\mathcal{S}_M(\mathbf{y}; \boldsymbol{\mu}, \Sigma, \alpha, \gamma) = \frac{\Gamma(\alpha + M/2)}{|\Sigma|^{1/2} \Gamma(\alpha) (2\pi\gamma)^{M/2}} [1 + \delta(\mathbf{y}, \boldsymbol{\mu}, \Sigma)/(2\gamma)]^{-(\alpha + M/2)}$$

→ Gaussian scale mixture representation (using weight variable U distributed according to a Gamma distribution)

$$\mathcal{S}_M(\mathbf{y}; \boldsymbol{\mu}, \Sigma, \alpha, \gamma) = \int_0^\infty \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \Sigma/u) \mathcal{G}(u; \alpha, \gamma) du$$

→ Parameters estimation is tractable by a general EM algorithm

3 - SLLiM model

- A mixture of Student distributions encodes the piecewise linear regressions

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

with

$$\mathbf{m}_k = \begin{bmatrix} \mathbf{c}_k \\ \mathbf{A}_k \mathbf{c}_k + \mathbf{b}_k \end{bmatrix} \text{ and } \mathbf{V}_k = \begin{bmatrix} \Gamma_k & \Gamma_k \mathbf{A}_k^T \\ \mathbf{A}_k \Gamma_k & \Sigma_k + \mathbf{A}_k \Gamma_k \mathbf{A}_k^T \end{bmatrix}$$

- Therefore, the joint density (\mathbf{X}, \mathbf{Y}) is a mixture of Student regressions

$$p(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = \sum_{k=1}^K \pi_k \mathcal{S}_{L+D}([\mathbf{x}, \mathbf{y}]^T; \mathbf{m}_k, \mathbf{V}_k, \alpha_k, 1)$$

- We denote by $\boldsymbol{\theta} = (\mathbf{c}_k, \Gamma_k, \mathbf{A}_k, \mathbf{b}_k, \Sigma_k, \pi_k, \alpha_k)_{1 \leq k \leq K}$ the set of parameters
- Extension to partially observed responses

$$\mathbf{X} = [\mathbf{T}, \mathbf{W}]^T$$

with T observed and W hidden variables

→ Allow to account for dependence among covariates and reduce the sensitivity of the method to model misspecification

References

[1] A. Deleforge, F. Forbes, and R. Horaud. High-dimensional regression with Gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 2015.

[2] E. Perthame, F. Forbes, and A. Deleforge. Inverse regression approach to robust non-linear high-to-low dimensional mapping. *Submitted*, 2016.

[3] Link to RATP (subway) data: <http://data.ratp.fr/explore/dataset/qualite-de-lair-mesuree-dans-la-station-chatelet>

4 - Inverse regression strategy

- Forward strategy ($\mathbf{x} = g(\mathbf{y})$), conditionals are

$$p(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{S}_L(\mathbf{x}; \mathbf{c}_k, \Gamma_k, \alpha_k, 1)$$

$$p(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{S}_D(\mathbf{y}; \mathbf{A}_k \mathbf{x} + \mathbf{b}_k, \Sigma_k, \alpha_k^y, \gamma_k^y)$$

→ $D = 500, L = 2, \Gamma_k$ diagonal → 126 254 parameters

- Inverse strategy ($\mathbf{y} = f(\mathbf{x})$)

$$p(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}^*) = \sum_{k=1}^K \pi_k \mathcal{S}_D(\mathbf{y}; \mathbf{c}_k^*, \Gamma_k^*, \alpha_k, 1)$$

$$p(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}^*) = \sum_{k=1}^K \pi_k \mathcal{S}_L(\mathbf{x}; \mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*, \Sigma_k^*, \alpha_k^x, \gamma_k^x)$$

with $\boldsymbol{\theta}^* = (\mathbf{c}_k^*, \Gamma_k^*, \mathbf{A}_k^*, \mathbf{b}_k^*, \Sigma_k^*, \pi_k, \alpha_k)_{1 \leq k \leq K}$ and

$$\mathbf{c}_k^* = \mathbf{A}_k \mathbf{c}_k + \mathbf{b}_k; \quad \Gamma_k^* = \Sigma_k + \mathbf{A}_k \Gamma_k \mathbf{A}_k^T;$$

$$\mathbf{A}_k^* = \Sigma_k^* \mathbf{A}_k^T \Sigma_k^{-1}; \quad \mathbf{b}_k^* = \Sigma_k^* (\Gamma_k^{-1} \mathbf{c}_k - \mathbf{A}_k^T \Sigma_k^{-1} \mathbf{b}_k); \quad \Sigma_k^* = (\Gamma_k^{-1} + \mathbf{A}_k^T \Sigma_k^{-1} \mathbf{A}_k)^{-1}$$

→ $D = 500, L = 2, \Sigma_k$ diagonal → 2 003 parameters

→ Our approach reduces the number of parameters to estimate

- Prediction : The regression function of interest g is approached by \tilde{g}

$$\tilde{g}(\mathbf{y}) = \mathbb{E}(\mathbf{X} | \mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}^*) = \sum_{k=1}^K \frac{\pi_k \mathcal{S}_D(\mathbf{y}; \mathbf{c}_k^*, \Gamma_k^*, \alpha_k, 1)}{\sum_{j=1}^K \pi_j \mathcal{S}_D(\mathbf{y}; \mathbf{c}_j^*, \Gamma_j^*, \alpha_j, 1)} (\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*)$$

5 - Estimation of $\boldsymbol{\theta}$ by EM algorithm

- E-step

- E-U step: Update of weight of each data point $\mathbb{E}[U | \mathbf{x}, \mathbf{y}, Z = k; \boldsymbol{\theta}^{(i)}]$
- E-Z step: Update posterior probabilities $\mathbb{P}(Z = k | \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}^{(i)})$

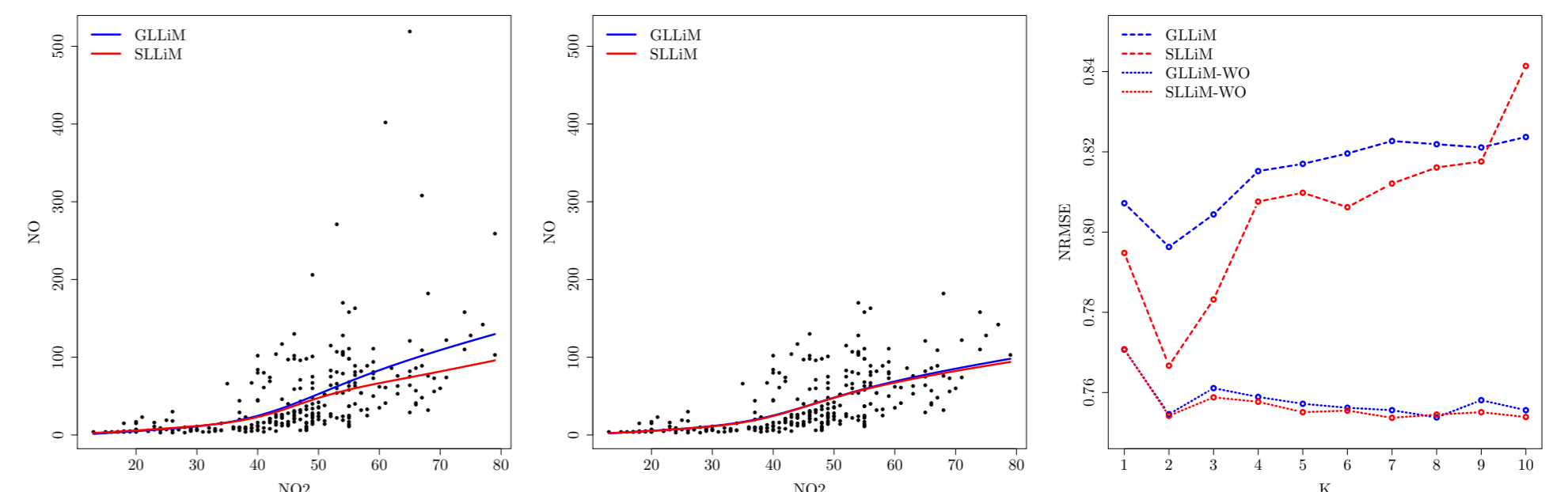
- M-step

- $(\pi_k, \mathbf{c}_k, \Gamma_k) \rightarrow$ Estimation is like a standard Student mixture
- $(\mathbf{A}_k, \mathbf{b}_k, \Sigma_k) \rightarrow$ Estimation is "Linear regression-like"
- $\alpha_k \rightarrow$ Not in closed-form but standard

6 - Application to air quality in the subway in Paris

- Prediction of NO (L=1) from NO₂ (D=1) in Châtelet station in Paris during March 2015 ($N = 341$ measures)
- SLLiM achieves better prediction ^a than its Gaussian counterpart (GLLiM) on complete data
- SLLiM is equivalent to GLLiM when no outliers (removed)

Estimated regression functions with 7 outliers (left panel) and no outliers (center panel) and prediction error rates (right panel)



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

7 - Other applications

- Application when $D \gg L$

- Hyperspectral data on Mars

- * $D=184, L=3, N=6983$
- * K fixed to 10, number of latent variables \mathbf{W} estimated by BIC
- * Prediction of proportion of CO₂ ice and dust from spectra

- Near-infrared spectra on orange juices

- * $D=134, L=1, N=218$
- * Prediction of sucrose level of each orange juice from its spectra

→ Comparison with other non linear regression methods

Prediction error rates for Mars data: average NRMSE (standard deviations) for proportions of CO₂ ice and dust over 100 cross validation runs

Method	Prop. of CO ₂ ice	Prop. of dust
SLLiM (K=10)	0.168 (0.019)	0.145 (0.020)
GLLiM (K=10)	0.180 (0.023)	0.155 (0.023)
Regression splines	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)