Non linear robust regression in high dimension

Emeline Perthame, Florence Forbes, Brice Olivier, Antoine Deleforge

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

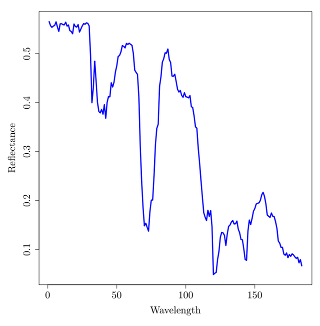
- The goal is to retrieve $\mathbf{X}$ from $\mathbf{Y}$ through a non linear regression function g

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

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

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

- 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 ($L = 3$)



$$\longrightarrow \begin{bmatrix} \text{prop. of dust} \\ \text{prop. of } CO_2 \text{ ice} \\ \text{prop. of water ice} \end{bmatrix}$$

2 - Difficulties

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

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

- Non linear mapping $\rightarrow$ Piecewise linear approximation of f (and g)

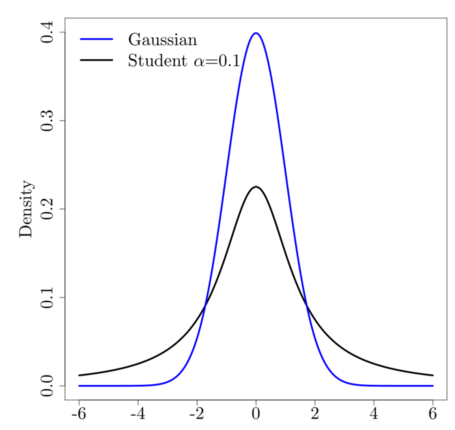
$$\mathbf{Y} = \textstyle\sum_{k=1}^K (\mathbb{I}_{Z=k}) \boldsymbol{A}_k \mathbf{X} + \boldsymbol{b}_k + \boldsymbol{E}_k$$

with $\mathbb{E}(E_k^2) \propto \boldsymbol{\Sigma}_k$ and Z multinomial latent variable

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

- Dealing with outliers $\rightarrow$ Heavy tail distribution

$\rightarrow$ Generalized Student distribution



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

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

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

$\rightarrow$ Parameters estimation is tractable by a general EM algorithm

3 - SLLiM model

- A mixture of Student distributions encodes the piecewise linear regressions

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

with

$$\boldsymbol{m}_k = \begin{bmatrix} \boldsymbol{c}_k \\ \boldsymbol{A}_k \boldsymbol{c}_k + \boldsymbol{b}_k \end{bmatrix} \text{ and } \boldsymbol{V}_k = \begin{bmatrix} \boldsymbol{\Gamma}_k & \boldsymbol{\Gamma}_k \boldsymbol{A}_k^T \\ \boldsymbol{A}_k \boldsymbol{\Gamma}_k & \boldsymbol{\Sigma}_k + \boldsymbol{A}_k \boldsymbol{\Gamma}_k \boldsymbol{A}_k^T \end{bmatrix}$$

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

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

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

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

with T observed and W hidden variables

$\rightarrow$ Allow to account for dependence among covariates and reduce the sensitivity of the method to model mispecification

4 - Inverse regression strategy

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

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

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

$\rightarrow$ $D = 500$, $L = 2$, $\boldsymbol{\Gamma}_k$ diagonal $\rightarrow$ 126 254 parameters

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

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

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

with $\boldsymbol{\theta}^* = (\boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*, \pi_k, \alpha_k)_{1 \le k \le K}$ and

$$\boldsymbol{c}_k^* = \boldsymbol{A}_k \boldsymbol{c}_k + \boldsymbol{b}_k; \qquad \boldsymbol{\Gamma}_k^* = \boldsymbol{\Sigma}_k + \boldsymbol{A}_k \boldsymbol{\Gamma}_k \boldsymbol{A}_k^T;$$

$$\boldsymbol{A}_k^* = \boldsymbol{\Sigma}_k^* \boldsymbol{A}_k^T \boldsymbol{\Sigma}_k^{-1};\ \boldsymbol{b}_k^* = \boldsymbol{\Sigma}_k^*(\boldsymbol{\Gamma}_k^{-1} \boldsymbol{c}_k - \boldsymbol{A}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{b}_k);\ \boldsymbol{\Sigma}_k^* = (\boldsymbol{\Gamma}_k^{-1} + \boldsymbol{A}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{A}_k)^{-1}$$

$\rightarrow$ $D = 500$, $L = 2$, $\boldsymbol{\Sigma}_k$ diagonal $\rightarrow$ 2 003 parameters

$\rightarrow$ Our approach reduces the number of parameters to estimate

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

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

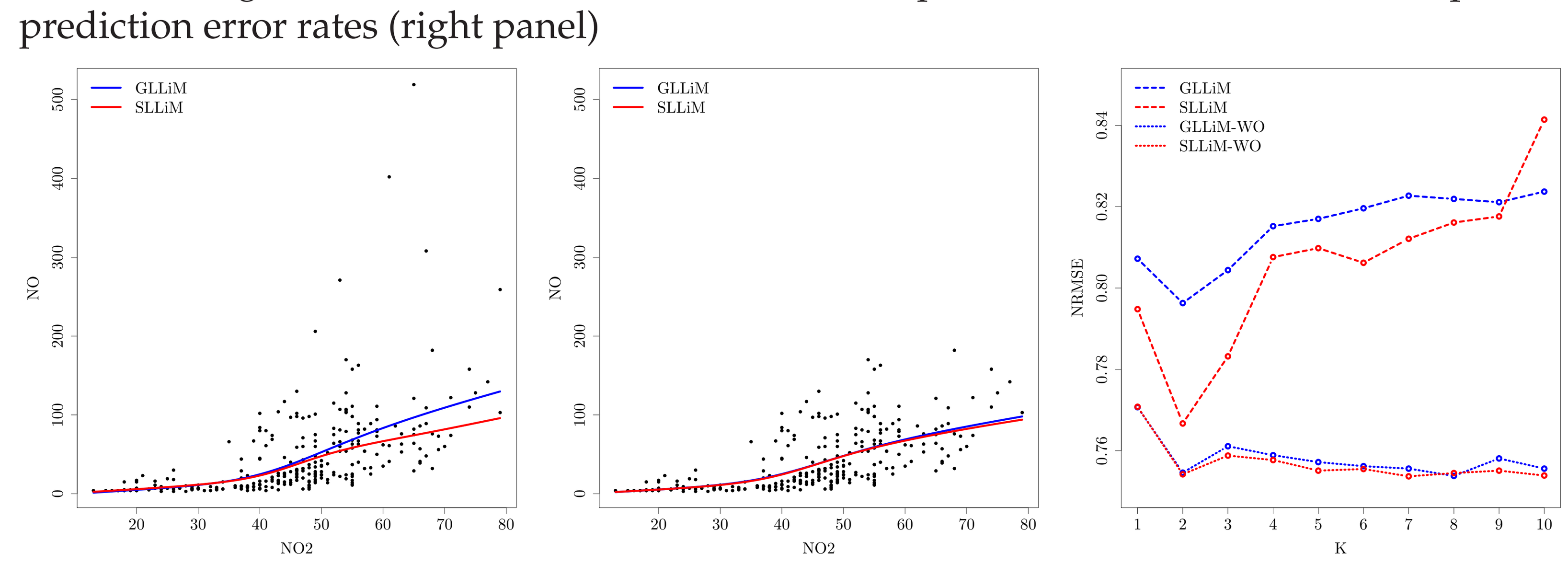
5 - Estimation of $\boldsymbol{\theta}$ by EM algorithm

- 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
  - $(\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ $\rightarrow$ Estimation is like a standard Student mixture
  - $(\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k)$ $\rightarrow$ Estimation is "Linear regression-like"
  - α_k $\rightarrow$ Not in closed-form but standard

6 - Application to air quality in the subway in Paris

- Prediction of NO (L=1) from NO_2 (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] $\text{NRMSE} = \sqrt{\frac{\sum_i (t_i - \hat{t}_i)^2}{\sum_i (t_i - \bar{t}_{\text{train}})^2}}$

7 - Other applications

- Application when $D \gg L$
  - Hyperspectral data on Mars
    - D=184, L=2, N=6983
    - K fixed to 10, number of latent variables $\boldsymbol{W}$ estimated by BIC
    - Prediction of proportion of CO_2 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

$\rightarrow$ Comparison with other non linear regression methods

Prediction error rates for Mars data: average NRMSE (standard deviations) for proportions of CO_2 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)

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`