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.
1 - Non linear mapping problem

* The goal is to retrieve $X$ from $Y$ through a non linear regression function $g$

$$E(X|Y = y) = g(y)$$

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

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

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

2 - Difficulties

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

$$E(Y|X = x) = f(x)$$

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

$$Y = \sum_{c=1}^{C} \alpha_c X + b_c + \epsilon_c$$

with $E(\epsilon_c^2) \propto \Sigma_c$ and $Z$ multinomial latent variable

$$P(Z = k) = n_k$$

* Dealing with outliers → Heavy tail distribution

→ Generalized Student distribution

$$S_U(y; \mu, \Sigma, \alpha, \gamma) = \frac{\Gamma((\mu + M/2)/2)}{\Gamma(M/2) \Gamma((\mu)/2)} \left[ 1 + \frac{d(y; \mu, \Sigma)}{2\gamma} \right]^{-(\mu + M/2)}$$

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

$$S_U(y; \mu, \Sigma|u, \alpha, \gamma) = \int_0^{\infty} N_U(y; \mu, \Sigma/u) g(u; \alpha, \gamma) \, du$$

→ Parameters estimation is tractable by a general EM algorithm

3 - SLLiM model

* A mixture of Student distributions encodes the piecewise linear regressions

$$p(X = x, Y = y|Z = k) = S_{L,D}(x, y; m_k, V_k, \alpha_k, 1)$$

with

$$m_k = \begin{bmatrix} c_k \\ A_k c_k + b_k \end{bmatrix} \quad \text{and} \quad V_k = \begin{bmatrix} \Gamma_k & \Gamma_k A_k^T \\ A_k \Gamma_k & \Sigma_k + A_k \Gamma_k A_k^T \end{bmatrix}$$

* Therefore, the joint density $(X, Y)$ is a mixture of Student regressions

$$p(X = x, Y = y) = \sum_{k=1}^{K} \pi_k S_{L,D}(x, y; m_k, V_k, \alpha_k, 1)$$

* We denote by $\theta = (c_k, \alpha_k, \Gamma_k, A_k, b_k, \Sigma_k, \pi_k, \alpha_k)_{1 \leq k \leq K}$ the set of parameters

* Extension to partially observed responses

$$X = [T, W]^T$$

with $T$ observed and $W$ hidden variables

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

4 - Inverse regression strategy

* Forward strategy ($x = g(y)$), conditionals are

$$p(X = x; \theta) = \sum_{k=1}^{K} \pi_k S_{L}(x; c_k, \alpha_k, 1)$$

$$p(Y = y|X = x; \theta) = \sum_{k=1}^{K} \pi_k S_L(y; A_k x + b_k, \Sigma_k, \alpha_k, \gamma_k)$$

$$D = 500, L = 2, \Gamma_k \text{ diagonal} \rightarrow 126 \times 254 \text{ parameters}$$

* Inverse strategy ($y = f(x)$)

$$p(Y = y; \theta^*) = \sum_{k=1}^{K} \pi_k S_0(y; c_k^*, \Gamma_k^*, \alpha_k, 1)$$

$$p(X = x|Y = y; \theta^*) = \sum_{k=1}^{K} \pi_k S_L(x; A_k^* y + b_k^*, \Sigma_k^*, \alpha_k^*, \gamma_k^*)$$

with $\theta^* = (c_k^*, \Gamma_k^*, A_k^*, b_k^*, \Sigma_k^*, \alpha_k^*, \gamma_k^*)_{1 \leq k \leq K}$ and

$$A_k^* = \Sigma_k^* A_k^T \Sigma_k^{-1}; \quad b_k^* = \Sigma_k^* (\Gamma_k^* c_k^* - A_k^T \Sigma_k^{-1} b_k); \quad \Sigma_k^* = (\Gamma_k^* + A_k^T \Sigma_k^{-1} A_k)^{-1}$$

$$D = 500, L = 2, \Sigma_k \text{ diagonal} \rightarrow 2 \times 003 \text{ parameters}$$

→ Our approach reduces the number of parameters to estimate

* Prediction: The regression function of interest $g$ is approached by $\hat{g}$

$$\hat{g}(y) = E(X|Y = y; \theta^*) = \sum_{k=1}^{K} \pi_k S_0(y; c_k^*, \Gamma_k^*, \alpha_k, 1) (A_k^* y + b_k^*)$$

5 - Estimation of $\theta$ by EM algorithm

* E-step

→ E-U step: Update of weight of each data point $E[U|x, y, Z; \theta^{(t)}]$

→ E-Z step: Update posterior probabilities $P(Z = k|x, y; \theta^{(t)})$

* M-step

→ $(c_k, \alpha_k, \Gamma_k) \rightarrow$ Estimation is like a standard Student mixture

→ $(A_k, b_k, \Sigma_k)$ → Estimation is “Linear regression-like”

→ $\alpha_k$ → Not in closed-form but standard

6 - Application to air quality in the subway in Paris

* Prediction of NO (L=1) from NO

→ Inverse strategy

→ Application when $D \gg L$

→ Hyperspectral data on Mars

→ $D = 184, L = 3, N = 6983$

→ $K$ fixed to 10, number of latent variables $W$ 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

→ Comparison with other non linear regression methods

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

7 - Other applications

* Application when $D \gg L$

→ 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 (0.052) 0.157 (0.065)

RVM 0.299 (0.021) 0.275 (0.034)

References


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