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

Emeline Perthame

Joint work with Florence Forbes

INRIA, team MISTIS, Grenoble

LMNO, Caen
October 27, 2016

Outlines

1. Non linear mapping problem

2. GLLiM/SLLiM: inverse regression approach

3. Estimation of parameters

4. Results and conclusion

Outlines

1. Non linear mapping problem

2. GLLiM/SLLiM: inverse regression approach

3. Estimation of parameters

4. Results and conclusion

A non linear mapping problem

- A non linear mapping problem

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

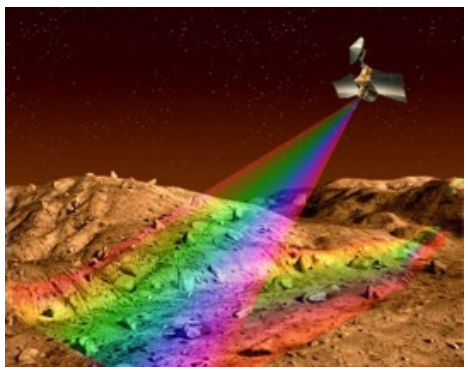
- Prediction of X from Y through a non linear regression function g

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

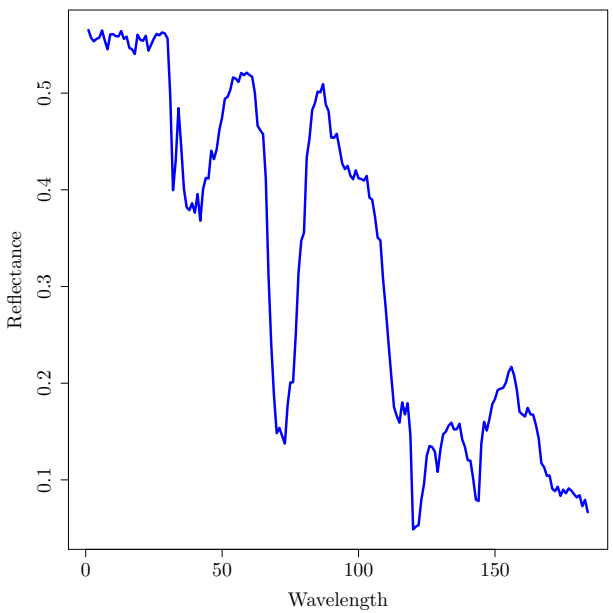
with $Y \in \mathbb{R}^D, X \in \mathbb{R}^L, D \gg L$

A non linear mapping problem

- Application: Ω mission on Mars $\rightarrow$ launch of a spectrometer around Mars
- Problem: Retrieving physical properties from hyperspectral images
  - Y: spectrum (D=184)
  - X: composition of the ground (L=3)



Mars Express - Omega (2004)
[http://geops.geol.u-psud.fr/]



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

Some approaches

- Difficulty: D large $\rightarrow$ curse of dimensionality
- Solutions: via dimensionality reduction
  - Reduce dimension of y before regression: eg. PCA on y

  $\rightarrow$ Risk: poor prediction of x
  - Take x into account: PLS, SIR, Kernel SIR, PC based methods

  $\rightarrow$ Two steps approaches not expressed as a single optimization problem

$\rightarrow$ Our approach: inverse regression to reduce dimension

Outlines

Proposed Method: An inverse regression strategy

- $x \in \mathbb{R}^L$ low-dimensional space,
- $y \in \mathbb{R}^D$ high-dimensional space,
- (y, x) are realizations of $(Y, X) \sim p(Y, X; \theta)$, θ parameters

Inverse conditional density: $p(Y \mid X; \theta)$

- Y is a noisy function of X
- Modeled via mixtures $\rightarrow$ Tractable θ estimation

Forward conditional density: $p(X \mid Y; \theta^*)$, with $\theta^* = f(\theta)$

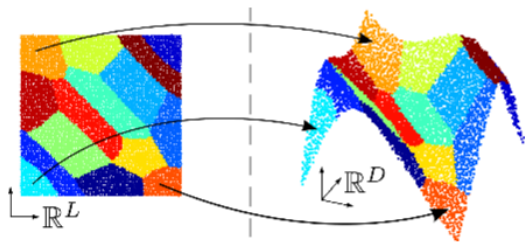
$\rightarrow$ High-to-low prediction, eg. $\hat{X} = \mathbb{E}[X \mid Y = Y; \theta^*]$

Student Locally-linear Mapping (SLLiM)

A piecewise affine model:

- Introduce a missing variable $Z \rightarrow Z = k \Leftrightarrow Y$ is the image of X by an affine transformation

$$Y = \sum_{k=1}^{K} \mathbb{I}(Z = k)(A_k X + b_k + E_k)$$



Definition of SLLiM

$$p(Y|X, Z = k; \theta) = \mathcal{S}(Y; A_k X + b_k, \Sigma_k, \alpha_k^y, \gamma_k^y)$$

- Affine transformations are local: mixture of K Student laws

$$\begin{aligned} p(X|Z = k; \theta) &= \mathcal{S}(X; c_k, \Gamma_k, \alpha_k, 1) \\ p(Z = k; \theta) &= \pi_k \end{aligned}$$

- The set of all model parameters is:

$$\theta = \{\pi_k, c_k, \Gamma_k, A_k, b_k, \Sigma_k, \alpha_k, k = 1 \dots K\}$$

Why a Student mixture ?

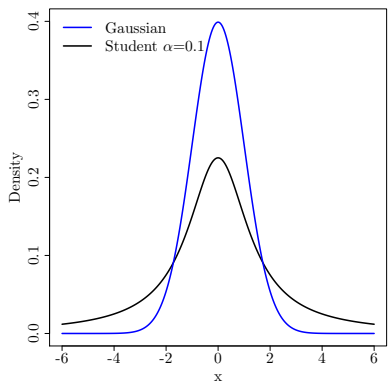
- Dealing with outliers $\rightarrow$ Generalized Student distribution for the joint density of (X, Y)

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

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

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

- Parameters estimation is tractable by an EM algorithm



Low-to-high (Inverse) Regression

- If X and Y are both observed
  - The parameter vector, θ, can be estimated in closed-form using an EM inference procedure
  - This yields the inverse conditional density which is a Student mixture:

$$p(Y|X;\theta) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}(X; c_k, \Gamma_k, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j \mathcal{S}(X; c_j, \Gamma_j, \alpha_j, 1)} \mathcal{S}(Y; A_k X + b_k, \Sigma_k \alpha_k^y, \gamma_k^y)$$

- Both densities are Student mixtures parameterized by θ. Therefore, to obtain:
  - A low-to-high inverse regression function:

$$\mathbb{E}[Y|X = x;\theta] = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}(x; c_k, \Gamma_k, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j \mathcal{S}(x; c_j, \Gamma_j, \alpha_k, 1)} (A_k x + b_k),$$

High-to-low (Forward) Regression

- The forward conditional density is a Student mixture as well:

$$p(X|\,Y;\theta^*) = \sum_{k=1}^{K} \frac{\pi_k^* \mathcal{S}(\,Y;\, c_k^*, \Gamma_k^*, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j^* \mathcal{S}(\,Y;\, c_j^*, \Gamma_j^*, \alpha_j, 1)} \mathcal{S}(X;\, A_k^*\, Y + b_k^*, \Sigma_k^*, \alpha_k^x, \gamma_k^x)$$

- The forward parameter vector, θ^* has an analytic expression as a function of θ
- Both densities are Student mixtures parameterized by θ. Therefore, to obtain:
  - A high-to-low forward regression function:

$$\mathbb{E}[X|\,Y = y;\theta] = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}(y;\, c_k^*, \Gamma_k^*, \alpha_k, 1)}{\sum_{j=1}^{K} \pi_j \mathcal{S}(y;\, c_j^*, \Gamma_j^*, \alpha_j, 1)} (A_k^* y + b_k^*).$$

The forward parameter vector θ^* from θ

$$
\begin{aligned}
c_k^* &= A_k\, c_k + b_k, \\
\Gamma_k^* &= \Sigma_k + A_k \Gamma_k A_k^T, \\
A_k^* &= \Sigma_k^* A_k^T \Sigma_k^{-1}, \\
b_k^* &= \Sigma_k^* (\Gamma_k^{-1}\, c_k - A_k^T \Sigma_k^{-1}\, b_k), \\
\Sigma_k^* &= (\Gamma_k^{-1} + A_k^T \Sigma_k^{-1} A_k)^{-1}.
\end{aligned}
$$

A joint model approach to reduce the number of parameters

- Joint model

$$p(X = x,\, Y = y | Z = k) = \mathcal{S}_{L+D}\left(\begin{bmatrix} x \\ y \end{bmatrix}; m_k,\, V_k, \alpha_k, 1\right)$$

with

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

- Reduce the number of parameters to estimate
  - Forward strategy + Γ_k diagonal
    - nb. par. $= \frac{1}{2} D(D-1) + DL + 2L + D$
    - $D = 500, L = 2 \rightarrow 126\ 254$ parameters
  - Inverse strategy + Σ_k diagonal
    - nb. par. $\frac{1}{2} L(L-1) + DL + 2D + L$
    - $D = 500, L = 2 \rightarrow 2\ 003$ parameters

Extension to partially observed responses

- Incorporate a latent component into the low-dimensional variable:

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

  where $T \in \mathbb{R}^{L_\mathrm{t}}$ is observed and $W \in \mathbb{R}^{L_\mathrm{w}}$ is latent $(L = L_\mathrm{t} + L_\mathrm{w})$
- Example on Mars data: lighting ? temperature ? grain size ?
- Observed pairs $\{(y_n, T_n), n = 1 \dots N\}$ $(T \in \mathbb{R}^{L_\mathrm{t}})$
- Additional latent variable W $(W \in \mathbb{R}^{L_\mathrm{w}})$
- Assuming the independence of T and W given Z :

$$p(X = (T, W)^\top \mid Z = k) = \mathcal{S}_L((T, W)^\top; c_k, \Gamma_k, \alpha_k, 1)$$

$$\text{with} \quad c_k = \begin{bmatrix} c_k^\mathrm{t} \\ 0 \end{bmatrix}, \; \Gamma_k = \begin{bmatrix} \Gamma_k^\mathrm{t} & 0 \\ 0 & \mathbb{I}_{L_w} \end{bmatrix}$$

Extension to partially observed responses

- Extension of SLLiM to more general covariance structure
- With $A_k = \left[\begin{array}{cc} A_k^{\mathrm{t}} & A_k^{\mathrm{w}} \end{array} \right]$,

$$Y = \sum_{k=1}^{K} \mathbb{I}(Z = k)(A_k^{\mathrm{t}} T + A_k^{\mathrm{w}} W + b_k + E_k)$$

rewrites

$$Y = \sum_{k=1}^{K} \mathbb{I}(Z = k)(A_k^{\mathrm{t}} T + b_k + E'_k)$$

$$\text{with } Var(E'_k) \propto \Sigma_k + A_k^{\mathrm{w}} A_k^{\mathrm{w}\top}$$

  - Diagonal $\Sigma_k \longrightarrow$ Factor analysis with L_{w} factors (at most)
  - A compromise between full $O(D^2)$ and diagonal $O(D)$ covariances

Outlines

1. Non linear mapping problem

2. GLLiM/SLLiM: inverse regression approach

3. Estimation of parameters

4. Results and conclusion

Estimation of $\theta = (c_k, \Gamma_k, A_k, b_k, \Sigma_k, \pi_k, \alpha_k)_{1 \leq k \leq K}$ by EM algorithm

- E-step
  - Update posterior probabilities

    (E_Z) $p(Z = k|t, y, \theta^{(i)}) \rightarrow$ "SMM-like"

    (E_W) $p(W|Z = k, t, y, \theta^{(i)}) \rightarrow$ Probabilistic PCA or Factor Analysis like

    (E_U) $\mathbb{E}(U|Z = k, t, y, \theta^{(i)}) \rightarrow$ Down-weighting extreme/atypic values in estimators $\rightarrow$ More robust

- M-step

  (M_X) $(\pi_k, c_k, \Gamma_k) \rightarrow$ "SMM-like"

  $(M_{Y|X})$ $(A_k, b_k, \Sigma_k) \rightarrow$ Hybrid between linear regression and PPCA/FA

$$\widetilde{A}_k = \widetilde{Y}_k \widetilde{X}_k^T \left(\begin{bmatrix} 0 & 0 \\ 0 & \widetilde{S}_k^w \end{bmatrix} + \widetilde{X}_k \widetilde{X}_k^T \right)^{-1}$$

  (M_α) $\alpha_k \rightarrow$ Not in closed-form but standard (specific to Student)

Outlines

1. Non linear mapping problem

2. GLLiM/SLLiM: inverse regression approach

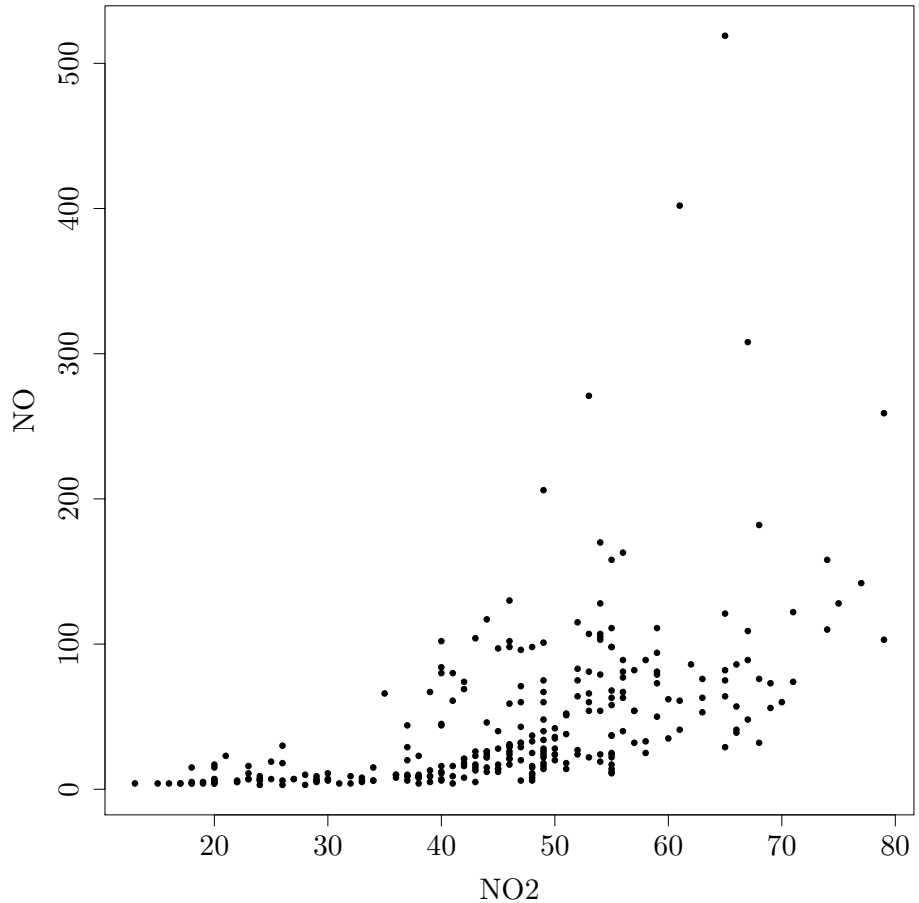
3. Estimation of parameters

4. Results and conclusion

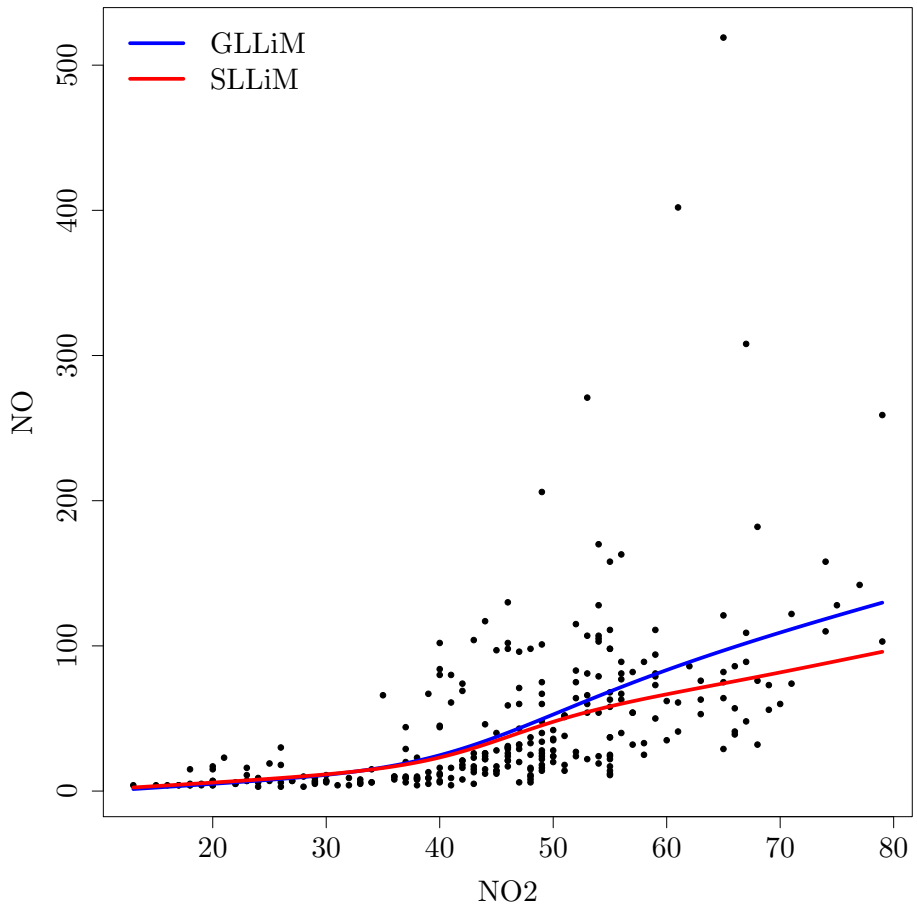
Application $L = D = 1$

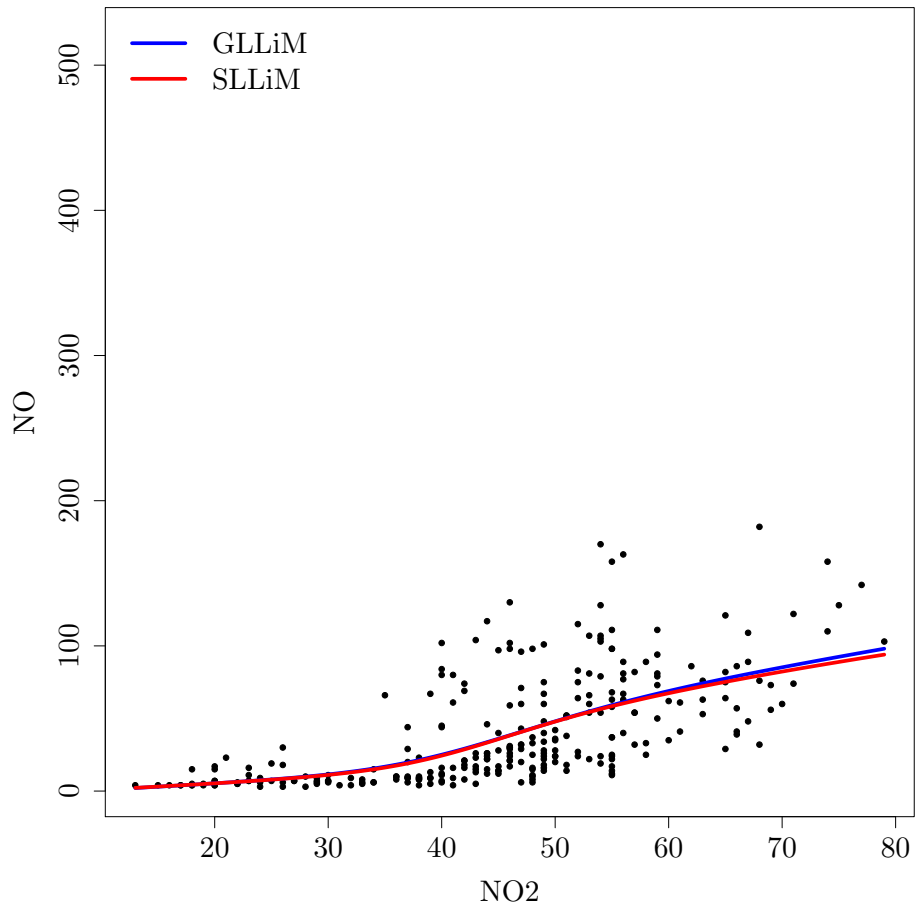
- RATP $\rightarrow$ Subway in Paris
- Measure of air quality at Châtelet station, line 4
- March 2015 $\rightarrow$ $N = 341$ measures
- Prediction of NO (L=1) from NO_2 (D=1)

$\rightarrow$ Robustness of SLLiM



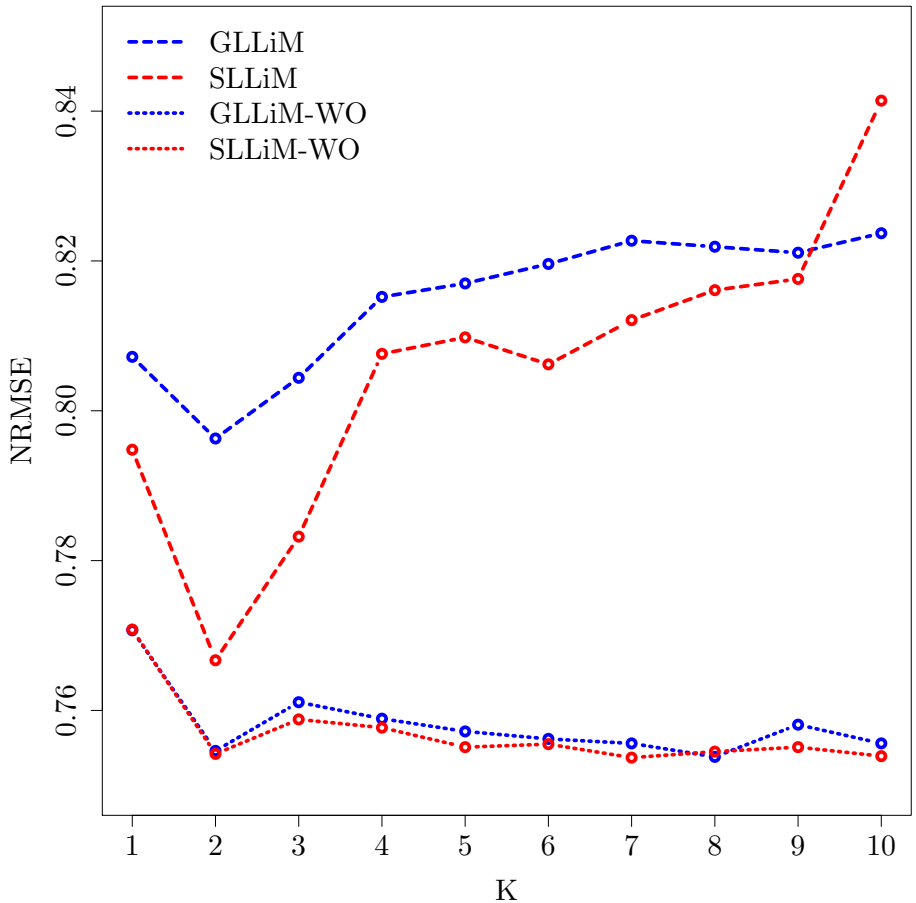
Application $L = D = 1$ / SLLiM compared to GLLiM





$\rightarrow$ Illustration of robustness of the proposed model

Application $L = D = 1$ / SLLiM compared to GLLiM



→ SLLiM achieves better prediction rates than GLLiM on complete data

→ SLLiM becomes equivalent to GLLiM when outliers are removed

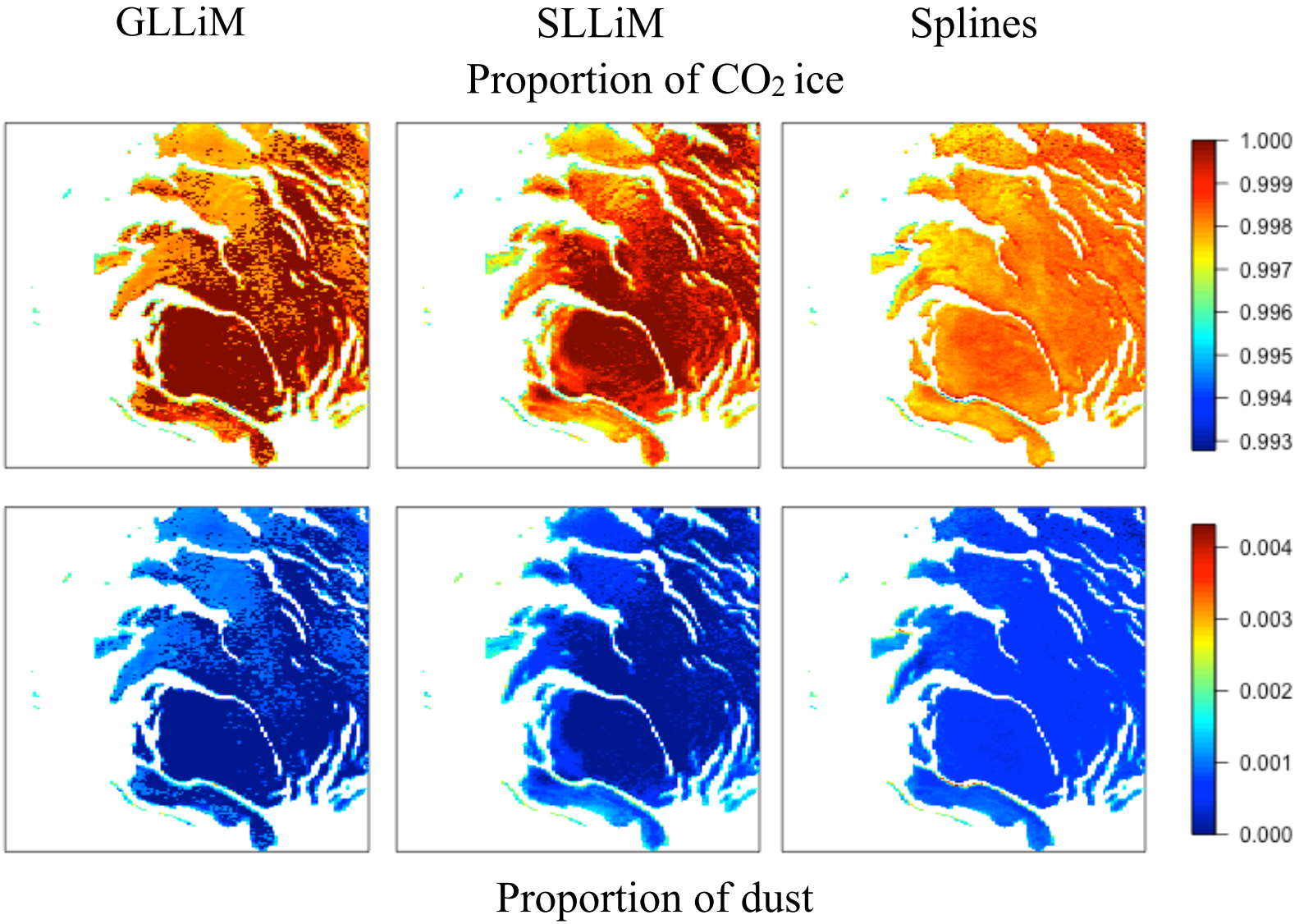
Other applications and augmented version of SLLiM

- Application when $D \gg L$
  - Hyperspectral data on Mars (D=184, L=2, N=6983)
  - $\rightarrow$ Comparison with other non linear regression methods

TABLE: 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 (K=10)	**0.168 (0.019)**	**0.145 (0.020)**
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)

Results - Application to hyperspectral image analysis



Conclusion and future work

- Mixture model used for prediction
- Addition of latent variables of partially observed responses

- Selection of K and L_w
  - K fixed ? Or selected by BIC ?
  - L_w selected by BIC ?

Thank you for your attention ! Any questions ?