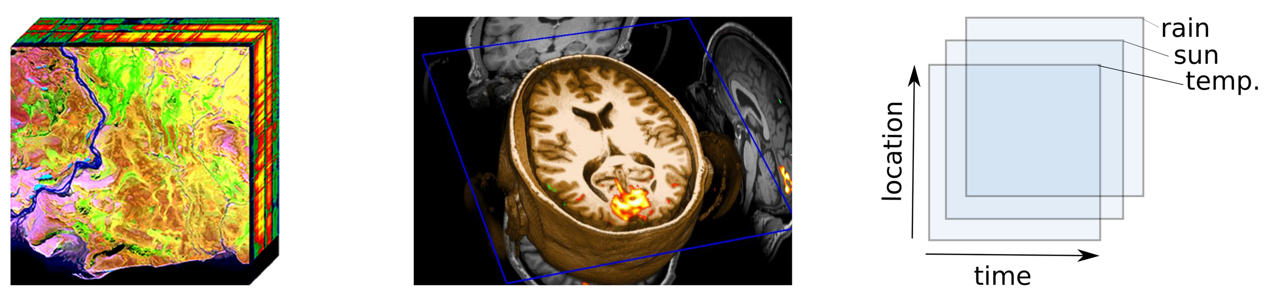
Low-Rank Regression with Tensor Responses

Guillaume Rabusseau[1,2] Hachem Kadri[1] [1]Aix-Marseille University [2]McGill University

Overview

- Reduced-Rank Regression extension to the multilinear setting.
- Regression problem with tensor structured outputs.

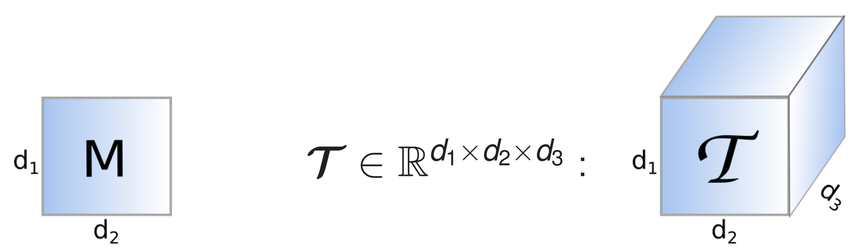


- Approximation algorithm rather than convex relaxation.
- Fast and efficient algorithm with strong theoretical guarantees.

Tensors

- Higher-order generalization of vectors and matrices:

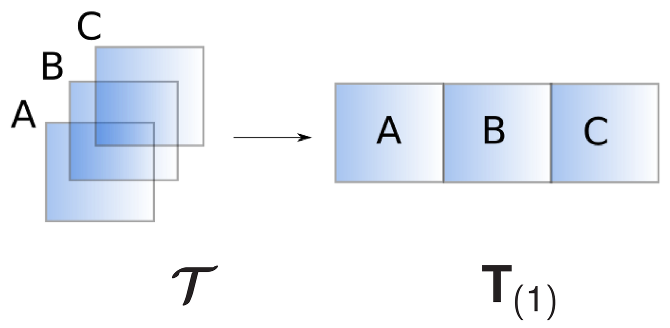
$\mathbf{M} \in \mathbb{R}^{d_1\times d_2}$: $\mathcal{T} \in \mathbb{R}^{d_1\times d_2\times d_3}$:



Tensors: Matricizations

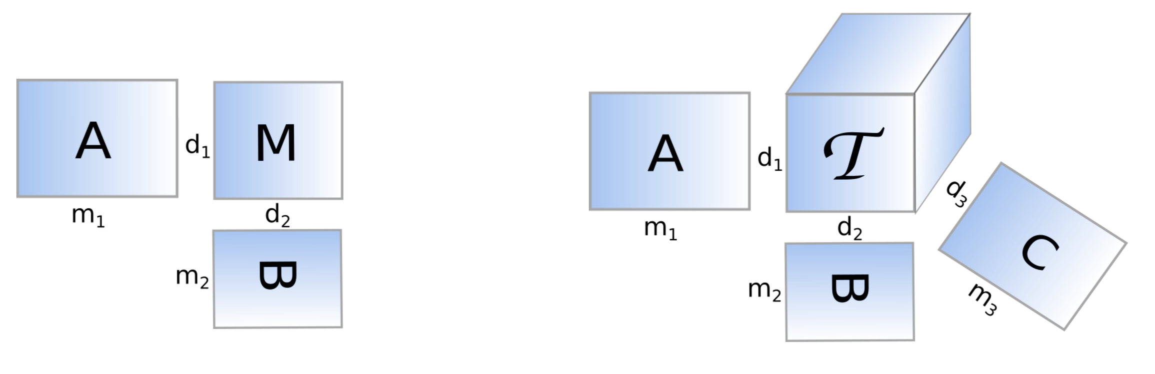
- $\mathcal{T} \in \mathbb{R}^{d_1\times d_2\times d_3}$ can be reshaped into a matrix as

$\mathbf{T}_{(1)} \in \mathbb{R}^{d_1\times d_2 d_3}$
$\mathbf{T}_{(2)} \in \mathbb{R}^{d_2\times d_1 d_3}$
$\mathbf{T}_{(3)} \in \mathbb{R}^{d_3\times d_1 d_2}$



$\mathcal{T}$ $\mathbf{T}_{(1)}$

Tensors: Multiplication with Matrices



$\mathbf{AMB}^\top \in \mathbb{R}^{m_1\times m_2}$ $\quad$ $\mathcal{T}\times_1 \mathbf{A}\times_2 \mathbf{B}\times_3 \mathbf{C} \in \mathbb{R}^{m_1\times m_2\times m_3}$

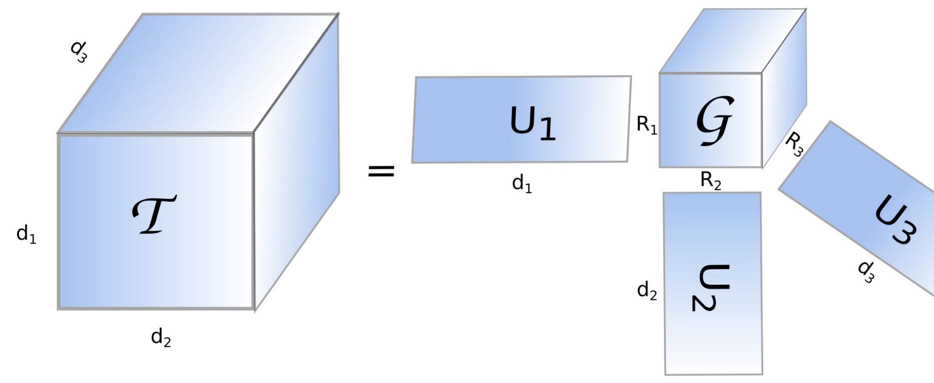
- For vectors, we note $\mathcal{T}\bullet_n \mathbf{v} = \mathcal{T}\times_n \mathbf{v}^\top$

Tucker Decomposition and Multilinear Rank

- Multilinear rank:

$$\text{rank}_{ml}(\mathcal{T}) = (R_1, R_2, R_3) \Leftrightarrow R_i = rank(\mathbf{T}_{(i)}) \text{ for } i = 1,2,3$$

- Tucker decomposition:



$$\mathcal{T} = \mathcal{G}\times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \text{ with } \mathbf{U}_i^\top\mathbf{U}_i = \mathbf{I} \text{ for all } i. \quad (1)$$

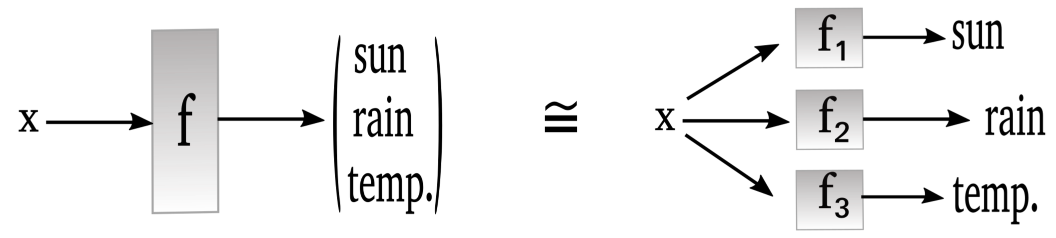
- Multilinear rank = smallest (R_1, R_2, R_3) such that (1) holds.

Multivariate Regression

- Learn $f : \mathbb{R}^d \to \mathbb{R}^p$ from $\{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^N$ where $\mathbf{y}^{(n)} \simeq f(\mathbf{x}^{(n)})$.
- Linear model: $f(\mathbf{x}) = \mathbf{W}^\top\mathbf{x}$ $\quad (\mathbf{W}\in\mathbb{R}^{d\times p})$
- Ordinary Least Squares:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathbb{R}^{d\times p}}{\arg\min} \|\mathbf{XW}-\mathbf{Y}\|_F^2 \quad (\mathbf{X}\in\mathbb{R}^{N\times d}, \mathbf{Y}\in\mathbb{R}^{N\times p})$$

$\Rightarrow$ Equivalent to perform p independent linear regressions!



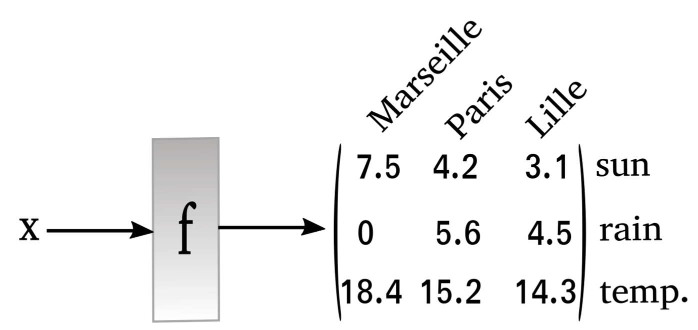
How can we capture linear dependencies in the output?

- Reduced Rank Regression [3]:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathbb{R}^{d\times p}}{\arg\min} \|\mathbf{XW}-\mathbf{Y}\|_F^2 \text{ s.t. } \text{rank}(\mathbf{W}) \leq R$$

Tensor-valued Regression

- Learn $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1\times d_2}$ from $\{(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\}_{n=1}^N$ where $\mathbf{Y}^{(n)} \simeq f(\mathbf{x}^{(n)})$.



- Linear model: $f(\mathbf{x}) = \mathcal{W}\bullet_1 \mathbf{x}$ $\quad (\mathcal{W}\in\mathbb{R}^{d_0\times d_1\times d_2})$
- *Low-Rank Regression for Tensor Structured Responses:*

$$\underset{\mathcal{W}\in\mathbb{R}^{d_0\times d_1\times d_2}}{\arg\min} \sum_{n=1}^N \mathcal{L}(\mathcal{W}\bullet_1 \mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) \text{ s.t. } \text{rank}_{ml}(\mathcal{W}) \leq (R_0, R_1, R_2)$$

- For the squared error loss we obtain:

Problem 1.

$$\underset{\mathcal{W}\in\mathbb{R}^{d_0\times d_1\times d_2}}{\arg\min} \|\mathcal{W}\times_1 \mathbf{X} - \mathcal{Y}\|_F^2 \quad s.t.\ \text{rank}_{ml}(\mathcal{W}) \leq (R_0, R_1, R_2)$$

where $\mathbf{X}\in\mathbb{R}^{N\times d_0}$ *and* $\mathcal{Y}\in\mathbb{R}^{N\times d_1\times d_2}$ *with* $\mathbf{X}_{n,:} = \mathbf{x}^{(n)}$ *and* $\mathcal{Y}_{n,:,:} = \mathbf{Y}^{(n)}$.

Solving the Minimization Problem

- Multilinear rank constraint implies that Problem 1 is equivalent to

$$argmin\|\mathcal{G}\times_1 \mathbf{XU}_0 \times_2 \mathbf{U}_1 \times_3 \mathbf{U}_2 - \mathcal{Y}\|_F^2 \quad (2)$$

w.r.t. $\mathcal{G}\in\mathbb{R}^{R_0\times R_1\times R_2}$, $\mathbf{U}_i\in\mathbb{R}^{d_i\times R_i}$ s.t. $\mathbf{U}_i^\top\mathbf{U}_i = \mathbf{I}$ for $0\leq i\leq 2$

Theorem 1.
For given column-wise orthogonal matrices $\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2$ *the tensor* $\mathcal{G}$ *that minimizes* (2) *is given by*

$$\mathcal{G} = \mathcal{Y}\times_1 (\mathbf{U}_0^\top\mathbf{X}^\top\mathbf{XU}_0)^{-1}\mathbf{U}_0^\top\mathbf{X}^\top \times_2 \mathbf{U}_1^\top \times_3 \mathbf{U}_2^\top.$$

$\Rightarrow$ Problem 1 is equivalent to:

$$\underset{\mathbf{U}_0\in\mathbb{R}^{d_0\times R_0},\ \mathbf{U}_1\in\mathbb{R}^{d_1\times R_1},\ \mathbf{U}_2\in\mathbb{R}^{d_2\times R_2}}{\arg\min} \|\mathcal{Y}\times_1 \mathbf{\Pi}_0 \times_2 \mathbf{\Pi}_1 \times_3 \mathbf{\Pi}_2 - \mathcal{Y}\|_F^2$$

$$\text{subject to} \quad \begin{cases} \mathbf{U}_i^\top\mathbf{U}_i = \mathbf{I} \text{ for } i = 0,1,2 \\ \mathbf{\Pi}_i = \mathbf{U}_i\mathbf{U}_i^\top \text{ for } i = 1,2 \\ \mathbf{\Pi}_0 = \mathbf{XU}_0\left(\mathbf{U}_0^\top\mathbf{X}^\top\mathbf{XU}_0\right)^{-1}\mathbf{U}_0^\top\mathbf{X}^\top \end{cases}$$

- Find 3 low-dimensional subspaces U_0, U_1, U_2 such that projecting $\mathcal{Y}$ onto the spaces $\mathbf{X}U_0, U_1, U_2$ is close to $\mathcal{Y}$.
- NP-hard... Solve each $\arg\min_{\mathbf{U}_i} \|\mathcal{Y}\times_{i+1} \mathbf{\Pi}_i - \mathcal{Y}\|_F^2$ independently instead.

Higher-Order Low-Rank Regression

Problem 2.

$$\underset{\mathcal{W}\in\mathbb{R}^{d_0\times d_1\times\cdots\times d_p}}{\arg\min} \|\mathcal{W}\times_1 \mathbf{X} - \mathcal{Y}\|_F^2 + \gamma\|\mathcal{W}\|_F^2 \quad s.t.\ \text{rank}_{ml}(\mathcal{W}) \leq (R_0, \cdots, R_p)$$

Algorithm (HOLRR).

Input: $\mathbf{X}\in\mathbb{R}^{N\times d_0}$, $\mathcal{Y}\in\mathbb{R}^{N\times d_1\times\cdots\times d_p}$, *rank* $(R_0, R_1, \cdots, R_p)$.

$\mathbf{U}_0 \leftarrow$ *top* R_0 *eigenvectors of* $(\mathbf{X}^\top\mathbf{X}+\gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}_{(1)}\mathbf{Y}_{(1)}^\top\mathbf{X}$
for $i = 1$ ***to*** p ***do***
$\quad\mathbf{U}_i \leftarrow$ *top* R_i *eigenvectors of* $\mathbf{Y}_{(i+1)}\mathbf{Y}_{(i+1)}^\top$
end for
$\mathbf{M} \leftarrow \left(\mathbf{U}_0^\top(\mathbf{X}^\top\mathbf{X}+\gamma\mathbf{I})\mathbf{U}_0\right)^{-1}\mathbf{U}_0^\top\mathbf{X}^\top$
$\mathcal{G} \leftarrow \mathcal{Y}\times_1 \mathbf{M}\times_2 \mathbf{U}_1^\top \times_3 \cdots \times_{p+1} \mathbf{U}_p^\top$
return $\mathcal{G}\times_1 \mathbf{U}_0 \times_2 \cdots \times_{p+1} \mathbf{U}_p$

Approximation Guarantees

- HOLRR is a quasi-optimal algorithm.

Theorem 2.
Let $\mathcal{W}^*$ *be a solution of Problem 2, let* $\hat{\mathcal{W}}$ *be the regression tensor returned by HOLRR, and let* $\mathcal{J}$ *denote the loss function. Then,*

$$\mathcal{J}(\hat{\mathcal{W}}) \leq (p+1)\mathcal{J}(\mathcal{W}^*).$$

Statistical Guarantees

- HOLRR recovers the low-rank regression (LRR) solution.

Theorem 3.

- *Training sample* $\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^N \subset \mathbb{R}^{d_0}\times\mathbb{R}^{d_1}$
- *Regularization parameters:* $1\leq R\leq d_0$ *and* $\gamma\geq 0$
- $\mathbf{W}_{HOLRR}\in\mathbb{R}^{d_0\times d_1}$*: HOLRR estimator with rank constraint* $(R_0, R_1) = (R, d_1)$ *and ridge parameter* γ
- $\mathbf{W}_{LRR}\in\mathbb{R}^{d_0\times d_1}$*: LRR estimator with rank constraint R and ridge parameter* γ

Then, $\mathbf{W}_{HOLRR} = \mathbf{W}_{LRR}$.

- HOLRR is statistically consistent.

Theorem 4.

- $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *i.i.d. in* $\mathbb{R}^{d_0}$
- $\xi^{(1)}, \cdots, \xi^{(N)} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ *i.i.d. in* $\mathbb{R}^{d_1\times\cdots\times d_p}$
- $\mathcal{W}^*\in\mathbb{R}^{d_0\times\cdots\times d_p}$ *a tensor s.t.* $\text{rank}_{ml}(\mathcal{W}) = (R_0, \cdots, R_p)$

$$\mathcal{Y}^{(n)} = \mathcal{W}^*\bullet_1 \mathbf{x}^{(n)} + \xi^{(n)} \text{ for all } n\in[N].$$

- *Let* $\mathcal{W}_N$ *be the estimator returned by HOLRR with training sample* $\{(\mathbf{x}^{(n)}, \mathcal{Y}^{(n)})\}_{n=1}^N$*, rank parameter* $(R_0, \cdots, R_p)$ *and regularization parameter* $\gamma = 0$

Then, for any $\varepsilon > 0$ *we have*

$$\lim_{N\to\infty} \mathbb{P}[\|\mathcal{W}^* - \mathcal{W}_N\|_F > \varepsilon] = 0.$$

- Generalization bound for the class of functions

$$\mathcal{F}_{ml} = \{\mathbf{x}\mapsto \mathcal{W}\bullet_1 \mathbf{x} \ : \ \text{rank}_{ml}(\mathcal{W}) = (R_0, \cdots, R_p)\}.$$

Theorem 5.
Let $\mathcal{L} : \mathbb{R}^{d_1\times\cdots\times d_p}\to\mathbb{R}$ *be a loss function bounded by M.*
For all $h\in\mathcal{F}_{ml}$*, for any* $\delta > 0$*, with probability at least* $1-\delta$*:*

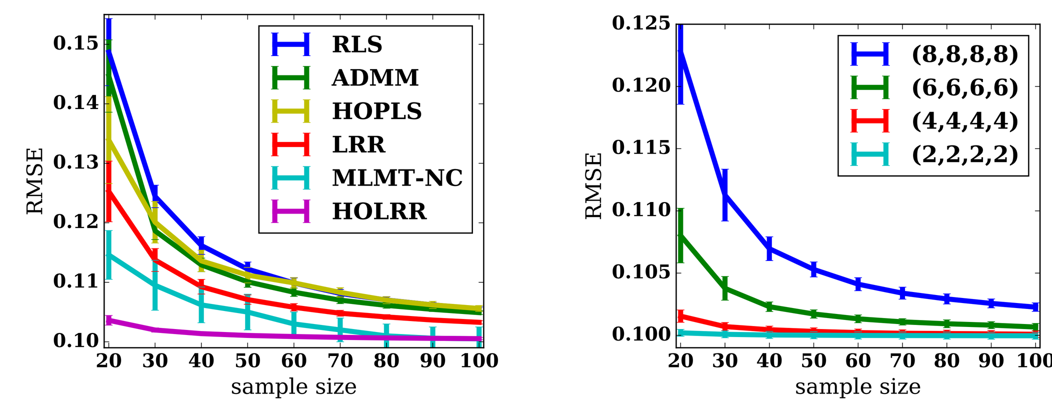
$$R(h) \leq \hat{R}(h) + M\sqrt{\frac{2D\log\left(\frac{4e(p+2)d_0d_1\cdots d_p}{\max_{i\geq 0} d_i}\right)\log N}{N}} + M\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2N}}$$

where $D = R_0R_1\cdots R_p + \sum_{i=0}^p R_i d_i$.

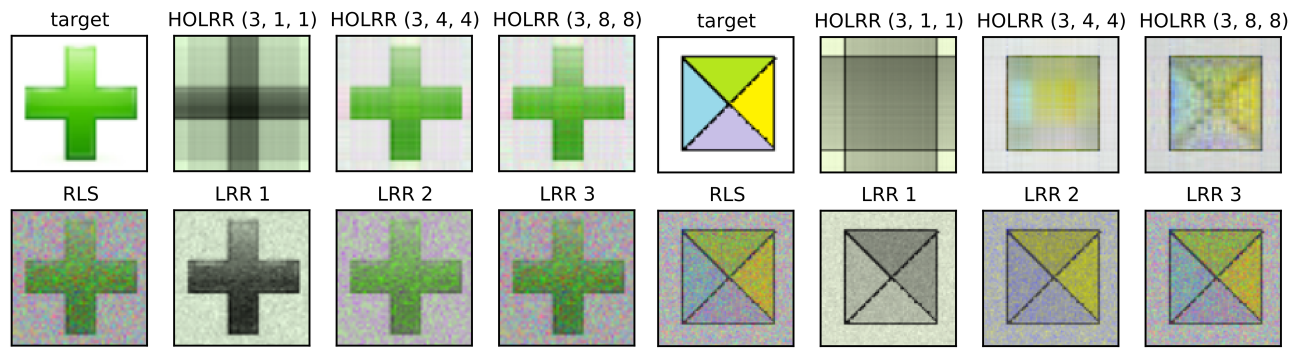
$\rightarrow$ Without low-rank constraint: $\mathcal{O}\left(\sqrt{d_0d_1\cdots d_p}\right)$

Experiments

- Comparison of HOLRR with:
  - **RLS**: Regularized least squares.
  - **LRR**: Low-rank regression.
  - **ADMM**: Convex relaxation (trace norms) [2, 4].
  - **MLMT-NC**: Nonconvex multilinear multitask learning [4].
  - **Greedy**: Greedy approach for spatio-temporal forecasting [1].
  - **HOPLS**: Higher-order partial least squares [5].
- (left) Synthetic data. (right) Effect of over-estimating the rank.



- Matrix vs. tensor rank regularization.



- Real data: forecasting task
  - **CCDS:** 17 variables across 125 locations from 1990 to 2001.
  - **Foursquare:** check-in records by 121 users in 15 categories over 1200 time intervals.
  - **Meteo-UK:** 5 variables across 16 locations from 1960 to 2000.

Data set	ADMM	Greedy	HOPLS	HOLRR	K-HOLRR (poly)	K-HOLRR (rbf)
CCDS	0.8448	0.8325	0.8147	0.8096	0.8275	**0.7913**
Foursquare	0.1407	**0.1223**	**0.1224**	0.1227	**0.1223**	0.1226
Meteo-UK	0.6140	–	0.625	0.5971	0.6107	**0.5886**

Table: Average RMSE over 10 splits train/test data sets (90%/10%)

Data set	MLMT-NC	ADMM	Greedy	HOPLS	HOLRR	K-HOLRR (poly)	K-HOLRR (rbf)
Synthetic	945.79	12.92	–	0.12	0.04	0.53	–
CCDS	–	235.73	75.47	121.28	100.94	0.46	0.61
Foursquare	–	33.83	37.70	22.3	14.41	19.20	19.67
Meteo UK	–	40.23	–	2.12	1.67	1.57	1.66

Table: Average running times in seconds

Future Works

- Sample complexity analysis.
- Minimax lower bounds for low rank regression.
- Take the tensor structure of the input into account.
- Extend to other loss functions.

References

[1] M. Bahadori, R. Yu, and Y. Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *NIPS*. 2014.
[2] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2), 2011.
[3] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2), 1975.
[4] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *ICML*, 2013.
[5] Q. Zhao, D. Caiafa, C.and Mandic, Z. Chao, Y. Nagasaka, N. Fujii, L. Zhang, and A. Cichocki. Higher order partial least squares (hopls). *PAMI*, 2013.

Emails: guillaume.rabusseau@mail.mcgill.ca, hachem.kadri@lif.univ-mrs.fr