Multiway high-dimensional lasso-penalized analysis with imputation of missing data applied to postgenomic data in an Ebola vaccine trial

Hadrien Lorenzo[1], Jérôme Saracco[2], Rodolphe Thiebaut[1]

[1]SISTM (Inserm, U1219, Bordeaux Population Health and Inria, Talence, France) and Vaccine Research Institute, Creteil, France.
[2]CQFD (INRIA Bordeaux Sud-Ouest, France), CNRS (UMR5251)

SMPGD, January 12th, 2018

Context

rVSV-ZEBOV Ebola Vaccine phase I dose escalation trial

- First vaccine to show efficiency during the Ebola outbreak [Henao-Restrepo et al., *The Lancet*, 2017],

Hamburg vaccination dataset content

- 3 types of responses :

  Antibody response Cellular functionnality Genomic expression
- 18 participants divided in 2 vaccination groups :

  $3 \cdot 10^6 pfu$ $20 \cdot 10^6 pfu$

Context

rVSV-ZEBOV Ebola Vaccine phase I dose escalation trial

- First vaccine to show efficiency during the Ebola outbreak [Henao-Restrepo et al., *The Lancet*, 2017],

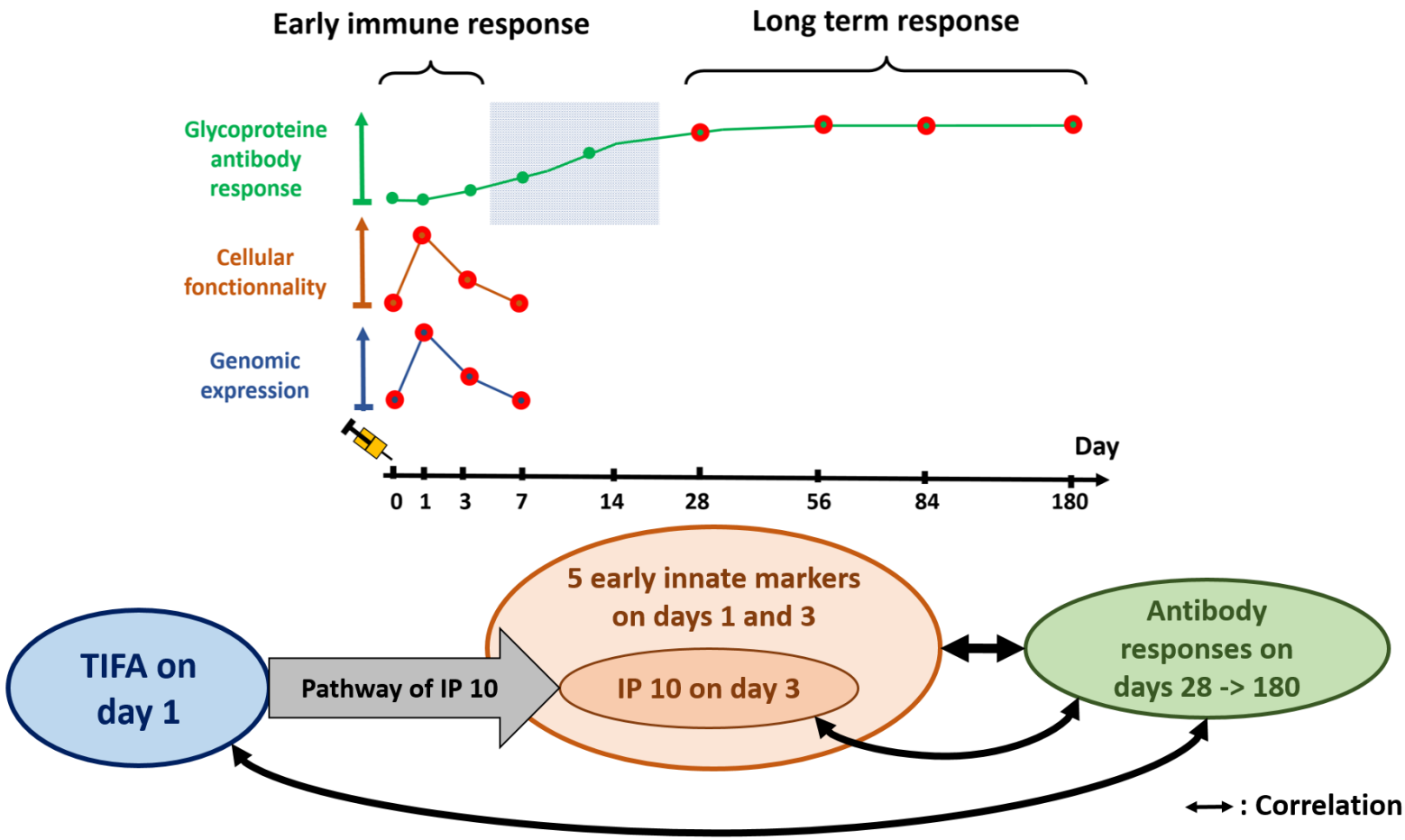
Hamburg vaccination dataset content

- 3 types of responses :

  Antibody response Cellular functionnality Genomic expression
- 18 participants divided in 2 vaccination groups :

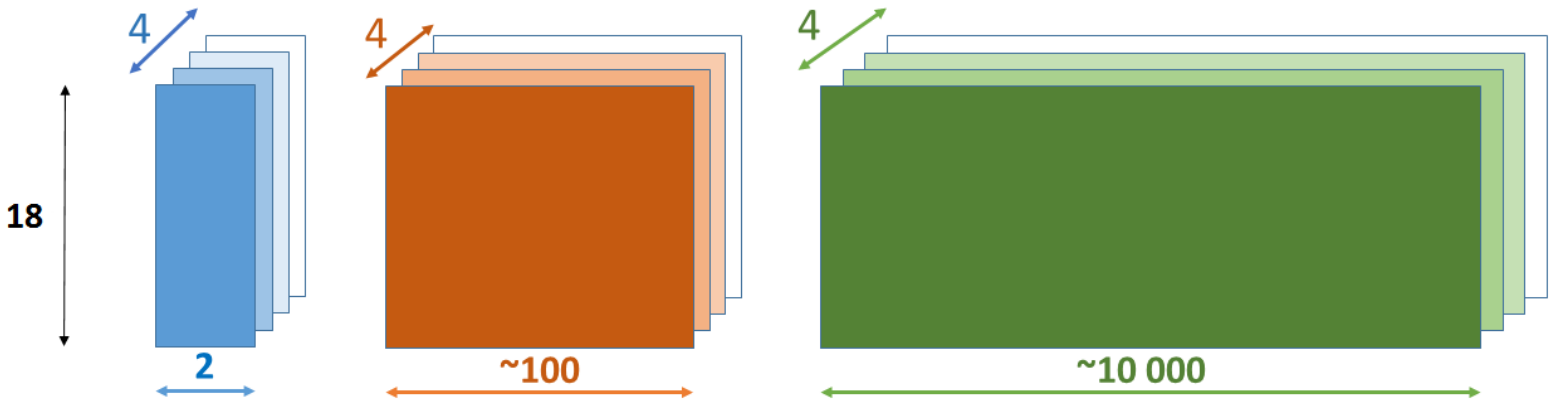
$$3 \cdot 10^6 pfu \qquad\qquad 20 \cdot 10^6 pfu$$

System vaccinology approach to examine the early innate immune response to Ebola rVSV vaccine, see [Rechtien et al., *Cell reports*, 2017]



rVSV -ZEBOV Ebola Vaccine phase I datasets

3 blocks of longitudinal data



Antibody response	Cellular fonctionnality	Genomic expression
Days 28, 56, 84, 180	Days 0, 1, 3, 7	Days 0, 1, 3, 7

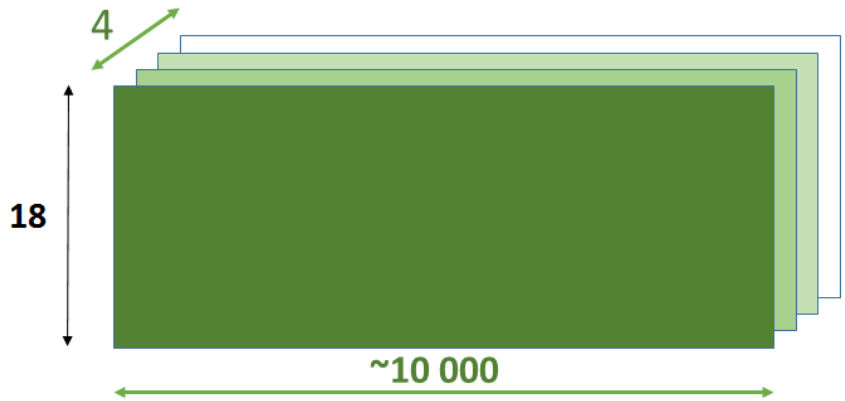
Genomic expression analysis : high dimensional problem

$$n = 18, p = 18301, T = 4$$

T : number of time measurement or "ways" $\Longrightarrow$ **multiway**

rVSV -ZEBOV Ebola Vaccine phase I datasets

3 blocks of longitudinal data



Antibody response	Cellular fonctionnality	Genomic expression
Days 28, 56, 84, 180	Days 0, 1, 3, 7	Days 0, 1, 3, 7

Genomic expression analysis : high dimensional problem

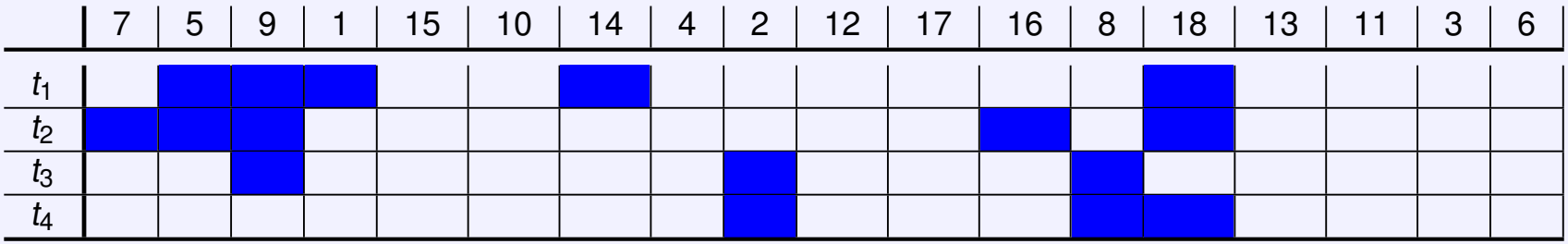
$$n = 18, p = 18301, T = 4$$

T : number of time measurement or "ways" $\Longrightarrow$ **multiway**

Missing origins

Poor sample qualities in case of :

- Low RNA integrity number (RIN)
- Insufficient library concentration
- Low sequencing depth



	7	5	9	1	15	10	14	4	2	12	17	16	8	18	13	11	3	6
t_1		■	■	■			■							■				
t_2	■	■	■									■		■				
t_3			■						■				■					
t_4									■				■	■				

Table: Missing path in the Ebola rVSV-ZEBOV RNA-Seq dataset where $t_1 = day_0$, $t_2 = day_1$, $t_3 = day_3$ and $t_4 = day_7$. Columns for participants.

Preliminar observations

- 30% of missing samples/values,
- Missing structure, parallel to time structure

Related & imagined methods

- **S/RGCCA**, from [Tenenhaus and Tenenhaus, 2011] :
  Multiway (Canonical Correlation) Regul. $\mathcal{L}_2$, $\mathcal{L}_1$ No imputation
  Applications : MRI Imaging, micro-array, heterogeneous datasets
- **softImpute** [Hastie et al., 2015] :
  Uniway (PCA) Regul. $\mathcal{L}_2$ Imputation
  Applications : Netflix (17770×480189, 99% of NA)
- **imputeMFA** in **missMDA** [Husson and Josse, 2013] :
  Multiway (WPCA) Regul. $\mathcal{L}_2$ automatic Imputation
  Applications : Sensory datasets
  The authors : *Efficient on highly correlated datasets.*

Our objectives and the chosen solutions

- Dimension reduction $\Longrightarrow$ multi-axes method,
- Link the latency variables $\Longrightarrow$ covariance criterion,
- Accuracy in prediction and interpretability $\Longrightarrow$ Lasso,
- Include Differential Expression (DE) information $\Longrightarrow$ weighted Lasso.

Related & imagined methods

- **S/RGCCA**, from [Tenenhaus and Tenenhaus, 2011] :
  Multiway (Canonical Correlation) Regul. $\mathcal{L}_2$, $\mathcal{L}_1$ No imputation
  Applications : MRI Imaging, micro-array, heterogeneous datasets
- **softImpute** [Hastie et al., 2015] :
  Uniway (PCA) Regul. $\mathcal{L}_2$ Imputation
  Applications : Netflix (17770×480189, 99% of NA)
- **imputeMFA** in **missMDA** [Husson and Josse, 2013] :
  Multiway (WPCA) Regul. $\mathcal{L}_2$ automatic Imputation
  Applications : Sensory datasets
  The authors : *Efficient on highly correlated datasets.*

Our objectives and the chosen solutions

- Dimension reduction $\Longrightarrow$ multi-axes method,
- Link the latency variables $\Longrightarrow$ covariance criterion,
- Accuracy in prediction and interpretability $\Longrightarrow$ Lasso,
- Include Differential Expression (DE) information $\Longrightarrow$ weighted Lasso.

Objective criterion approximation formulation

Approximed chosen criterion in semi-lagrangian notation

$$\underset{(x_t)_{t\in[\![1,T]\!]}}{\text{minimize}} \quad \sum_{t\in[\![1,T]\!],s>t} \frac{1}{2}\left|\left|\frac{X_t^{\star T}}{\sigma_1(X_t^\star)}\frac{X_s^\star}{\sigma_1(X_s^\star)} - x_t x_s^T\right|\right|_F^2 + \sum_{t\in[\![1,T]\!]} \frac{\lambda_t}{2}(||x_t||_2^2 - 1) + \mu_t(||D(\theta, X^\star)x_t||_1 - \eta_t)$$

$$\text{s.t. } \forall t \in [\![1,T]\!], \quad \eta_t = arg \min_{\eta\in\mathbb{R}_+, ||x_t||_0 \leqslant keep_X} abs(||D(\theta, X^\star)x_t||_1 - \eta),$$

$$(\lambda_t, \mu_t) > 0$$

$\theta \in]0,1]$ a parameter giving importance to DE genes ($\theta \to 0^+$) or no importance ($\theta = 1$).

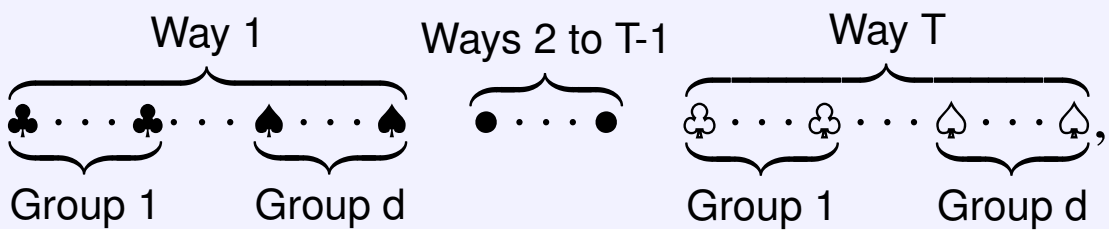
T the number of ways/time measurements,
$X^\star = (X_t^\star)_{t=1..T}$ the imputed matrices,
$\sigma_1(.)$ the largest eigenvalue of ".",

$keep_X$ the max number of genes to keep,
D diagonal matrix with weights giving power to the DE genes.

Simulations

Principle of the simulations

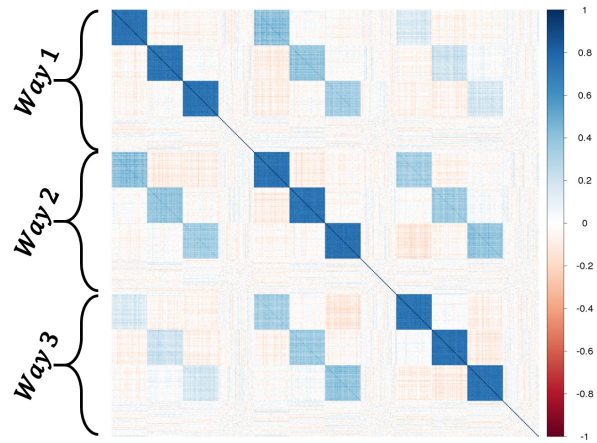
d groups of variables, $\forall j \in 1..d, p_j$: number of variables in group j.



- Inter-way correlation coefficients : $(\rho_j)_{j=1..d}$.
- Intra-way correlation structure : $AR(1)$ with coefficient ρ_t.

Example :

$T = 3, d = 4, p_j = 40, \rho_j = 0.8, \rho_t = 0.5$

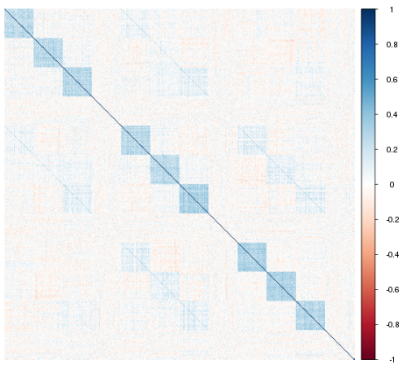


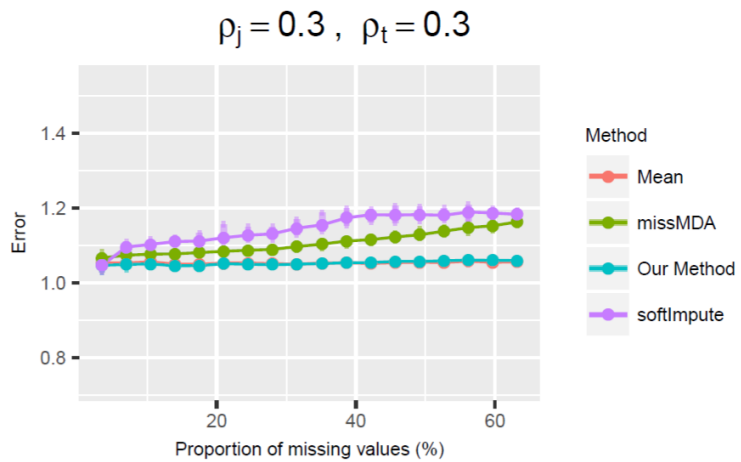
Simulations

Comparisons with *SoftImpute*, *missMDA* and *Mean imputation* in a high-dimensional context $n << p$

Fix $T = 3, d = 4, n = 200, p_j = 400 (\implies p = 1600)$, with criterion

$$\text{RMSE} = f(prop_{NA}, \rho_j = 0.3, \rho_t = 0.3)$$





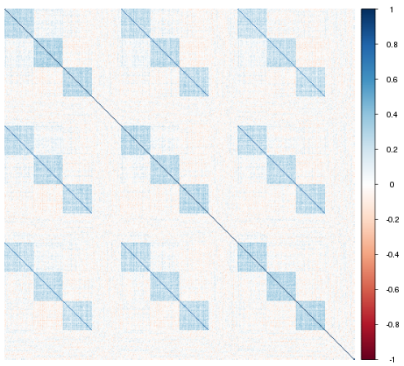
Our Method limits errors, equivalent to Mean.

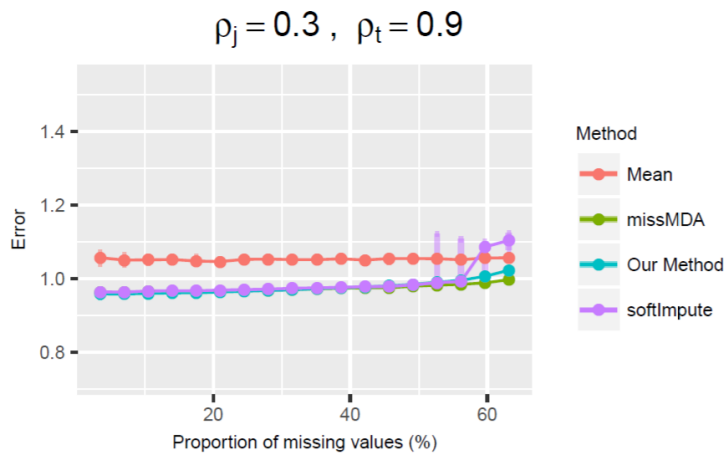
Simulations

Comparisons with *SoftImpute*, *missMDA* and *Mean imputation* in a high-dimensional context $n << p$

Fix $T = 3, d = 4, n = 200, p_j = 400 (\implies p = 1600)$, with criterion

$$\text{RMSE} = f(prop_{NA}, \rho_j = 0.3, \rho_t = \mathbf{0.9})$$





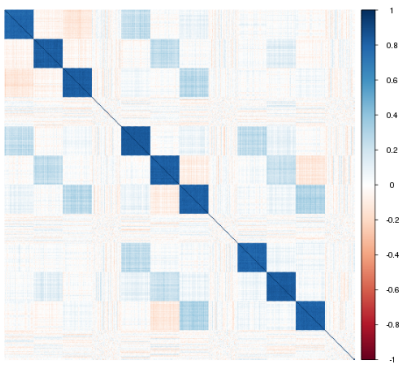
Mean is wronger than others except for high missing values where softImpute is the worst.
All methods seem to learn from temporal structure.

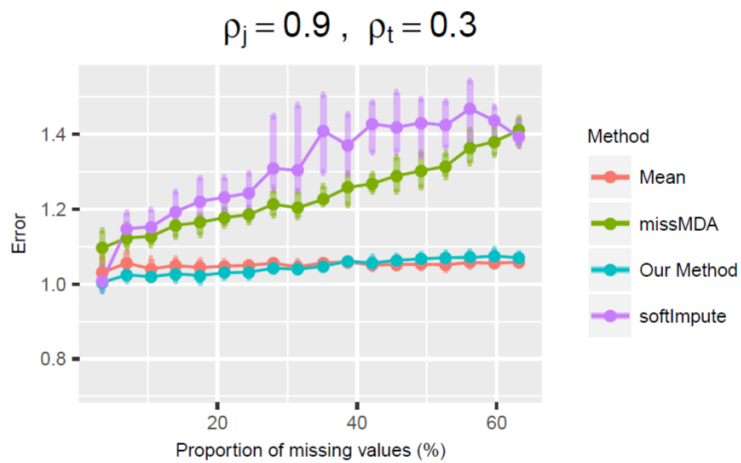
Simulations

Comparisons with *SoftImpute*, *missMDA* and *Mean imputation* in a high-dimensional context $n << p$

Fix $T = 3, d = 4, n = 200, p_j = 400 (\implies p = 1600)$, with criterion

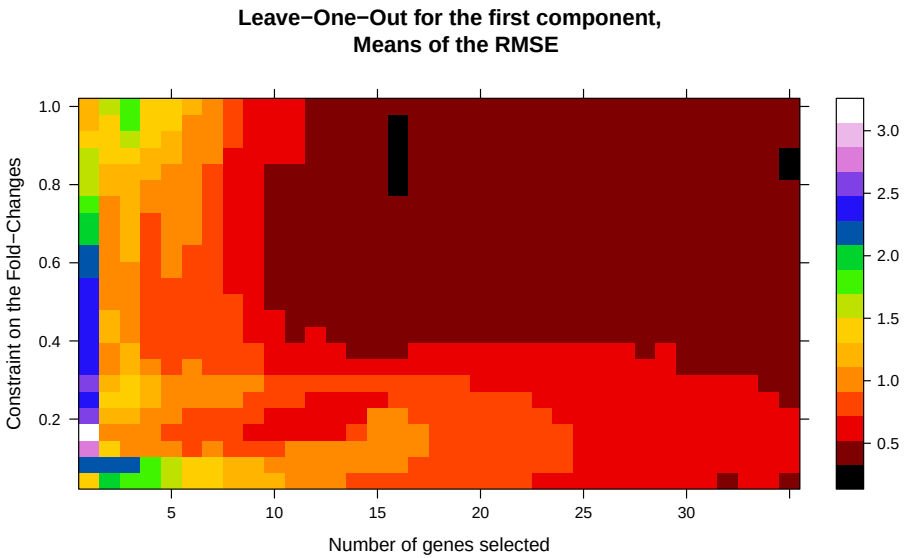
$$\text{RMSE} = f(prop_{NA}, \rho_j = \mathbf{0.9}, \rho_t = 0.3)$$





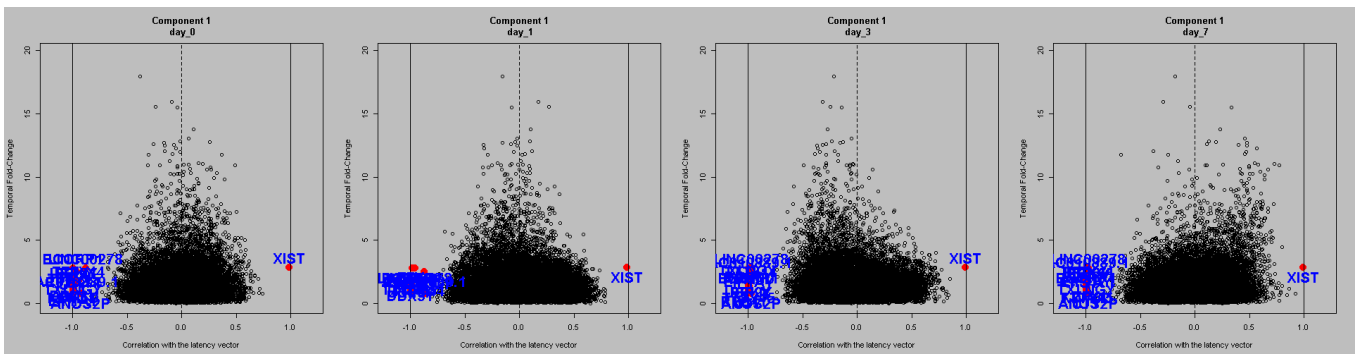
Our Method and Mean did well in comparison to the other methods.

Analysis of rVSV-ZEBOV RNA-Seq dataset, 1^{st} axis, Leave-One-Out on $(keep_X, \theta)$



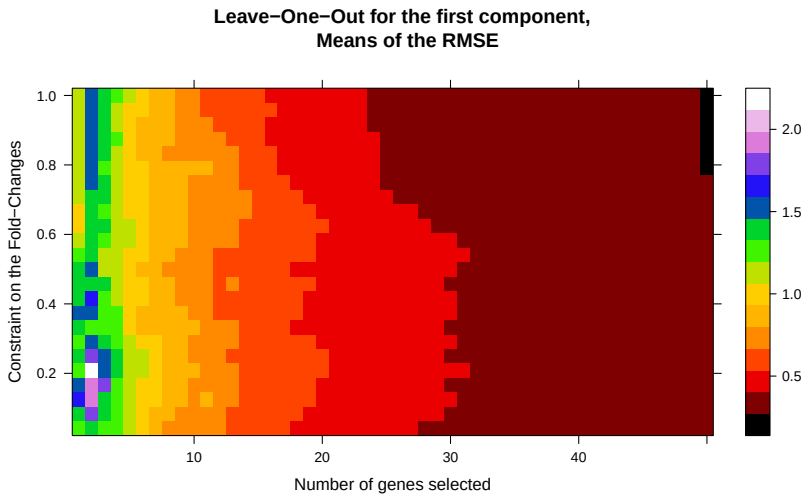
Gene name	
XIST	RPS4Y1
ZFY	LINC00278
PRKY	TTTY15
USP9Y	DDX3Y
UTY	ANOS2P
TTTY14	BCORP1
TXLNGY	AC010889.1
KDM5D	EIF1

Minimum for $keep_X = 16$, detectable using θ parameter.

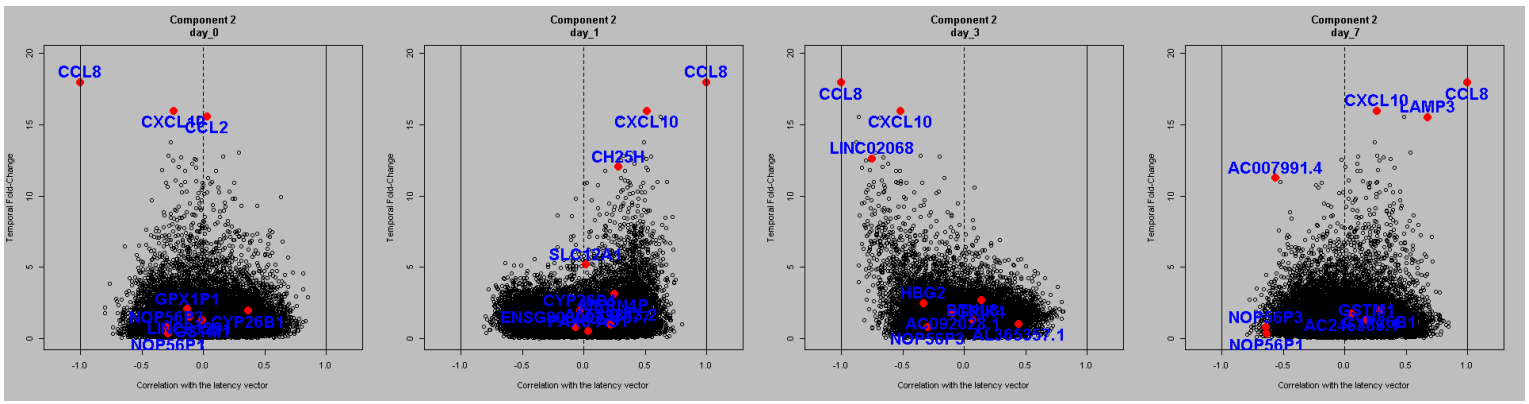


Genes linked to the sex and not to the vaccination.

Analysis of rVSV-ZEBOV RNA-Seq dataset, 2^{nd} axis, Leave-One-Out on $(keep_X, \theta)$ no minimum



Take $\theta = 5 \cdot 10^{-3}$ and $keep_X = 9$ for observations.



Genes reacting to vaccination are selected, cf [Rechtien et al., *Cell reports*, 2017]

Conclusion

- Ongoing work,
- Method able to learn from intra-time and longitudinal structure,
- Selection based on a trade-off between correlation and differential expression,
- Add other datasets for data-heterogeneous analysis.

hadrien.lorenzo@u-bordeaux.fr

References


Trevor Hastie et al. “Matrix completion and low-rank svd via fast alternating least squares”. In: *J. Mach. Learn. Res* 16.1 (2015), pp. 3367–3402.


Ana Maria Henao-Restrepo et al. “Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!)” In: *The Lancet* 389.10068 (2017), pp. 505–518.


François Husson and Julie Josse. “Handling missing values in multiple factor analysis”. In: *Food quality and preference* 30.2 (2013), pp. 77–85.


Anne Rechtien et al. “Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV”. In: *Cell reports* 20.9 (2017), pp. 2251–2261.


Arthur Tenenhaus and Michel Tenenhaus. “Regularized generalized canonical correlation analysis”. In: *Psychometrika* 76.2 (2011), pp. 257–284.

Algorithm

The algorithm use is a alternating least square algorithm with soft-thresholding solution to the Lasso constraint.
The imputation is performed based on linear regression on the projected matrices such as, where $\forall t \in 1..T, Z_t = \frac{X_t^\star}{\sigma_1(X_t)}$:

$$Z_t = Z_t x_t x_t^T + Z_t(\mathbb{I}_p - x_t x_t^T), \tag{1}$$

and $Z_t x_t x_t^T$ can be approximated with the actual estimations of $(x_s)_{s \neq t}$

$$Z_t x_t x_t^T = \sum_{s=1, s \neq t}^{T} \beta_{t,s} Z_s x_s x_s^T + \epsilon, \tag{2}$$

where $\forall (t,s), \beta_{t,s} \in \mathbb{R}$ and ϵ is a matrix with a norm negligeable against $Z_t^\star x_t x_t^T$.
Scalar projections onto $\forall s \neq t, x_s$ permits to estimate each $\beta_{t,s}$.
$X_t^\star$ is then uploaded with the estimated values but only the missing values are changed. A normalization is then applied to restart the algorithm if the criterion is not small enough.

Weighted Lasso

$$\forall g \in [\![1, G]\!], d_g = \theta + (1 - \theta)\gamma_g, \tag{3}$$

$$\gamma_g = 1 - \frac{DE_g}{\max_{h \in [\![1,G]\!]}(DE_h)}, \tag{4}$$

$$DE_g = \sum_{t,s=1..T}^{s>t} |(\mu_g^{(t)} - min_{h \in [\![1,G]\!]}(\mu_h^{(t)})) - (\mu_g^{(s)} - min_{h \in [\![1,G]\!]}(\mu_h^{(s)}))|, \tag{5}$$

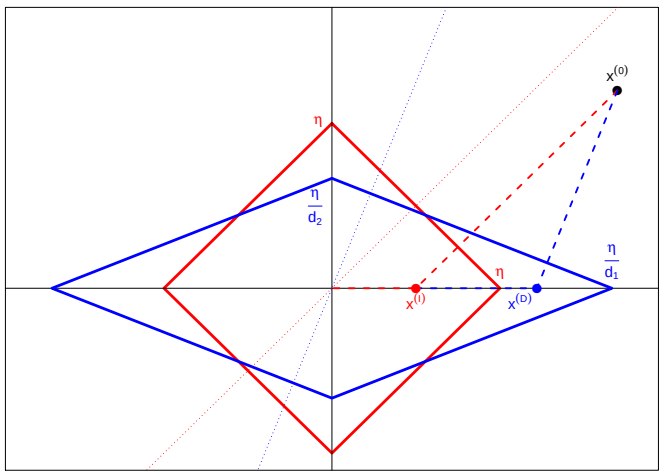


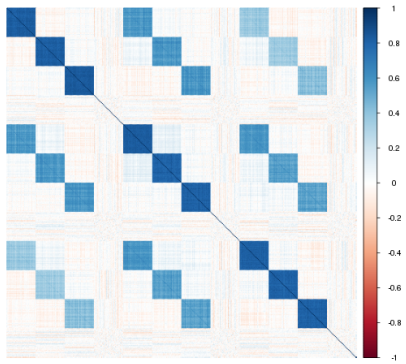
Figure: Lasso and weighted Lasso behavior representations

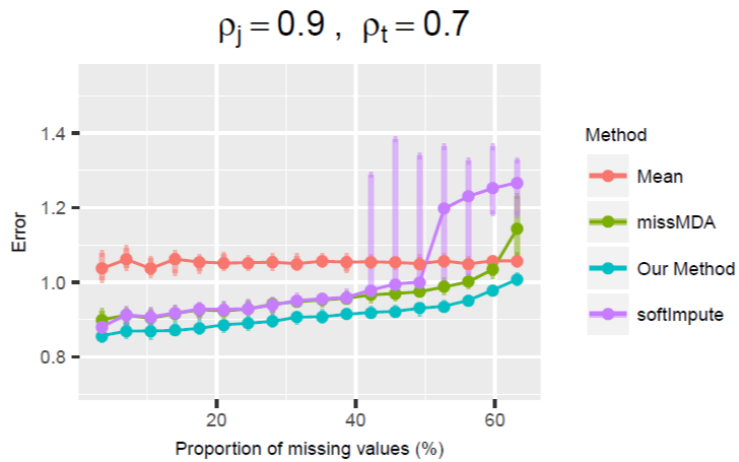
Simulations

Comparisons with *SoftImpute*, *missMDA* and *Mean imputation* in a high-dimensional context $n << p$

Fix $T = 3, d = 4, n = 200, p_j = 400 (\implies p = 1600)$, with criterion

$$\text{RMSE} = f(prop_{NA}, \rho_j = \mathbf{0.9}, \rho_t = \mathbf{0.7})$$





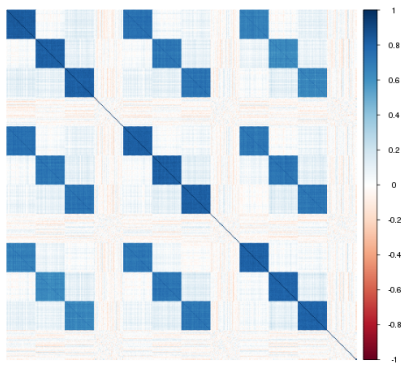
All methods seem to learn from temporal structure, even better since ρ_j is high. Divergence for softImpute for high proportions of missing values.

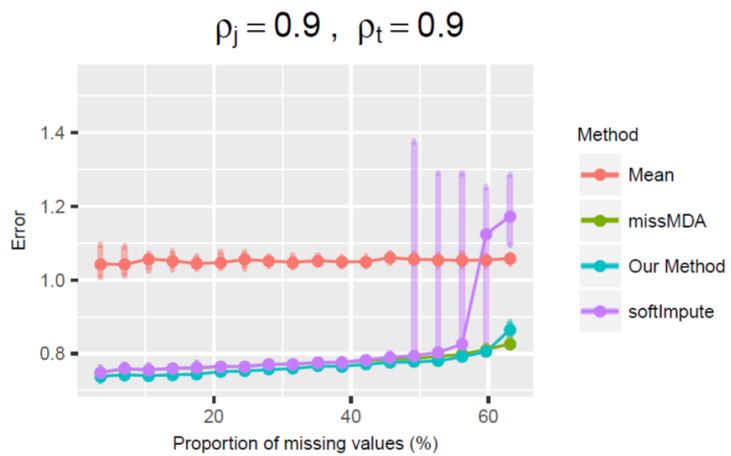
Simulations

Comparisons with *SoftImpute*, *missMDA* and *Mean imputation* in a high-dimensional context $n << p$

Fix $T = 3, d = 4, n = 200, p_j = 400 (\implies p = 1600)$, with criterion

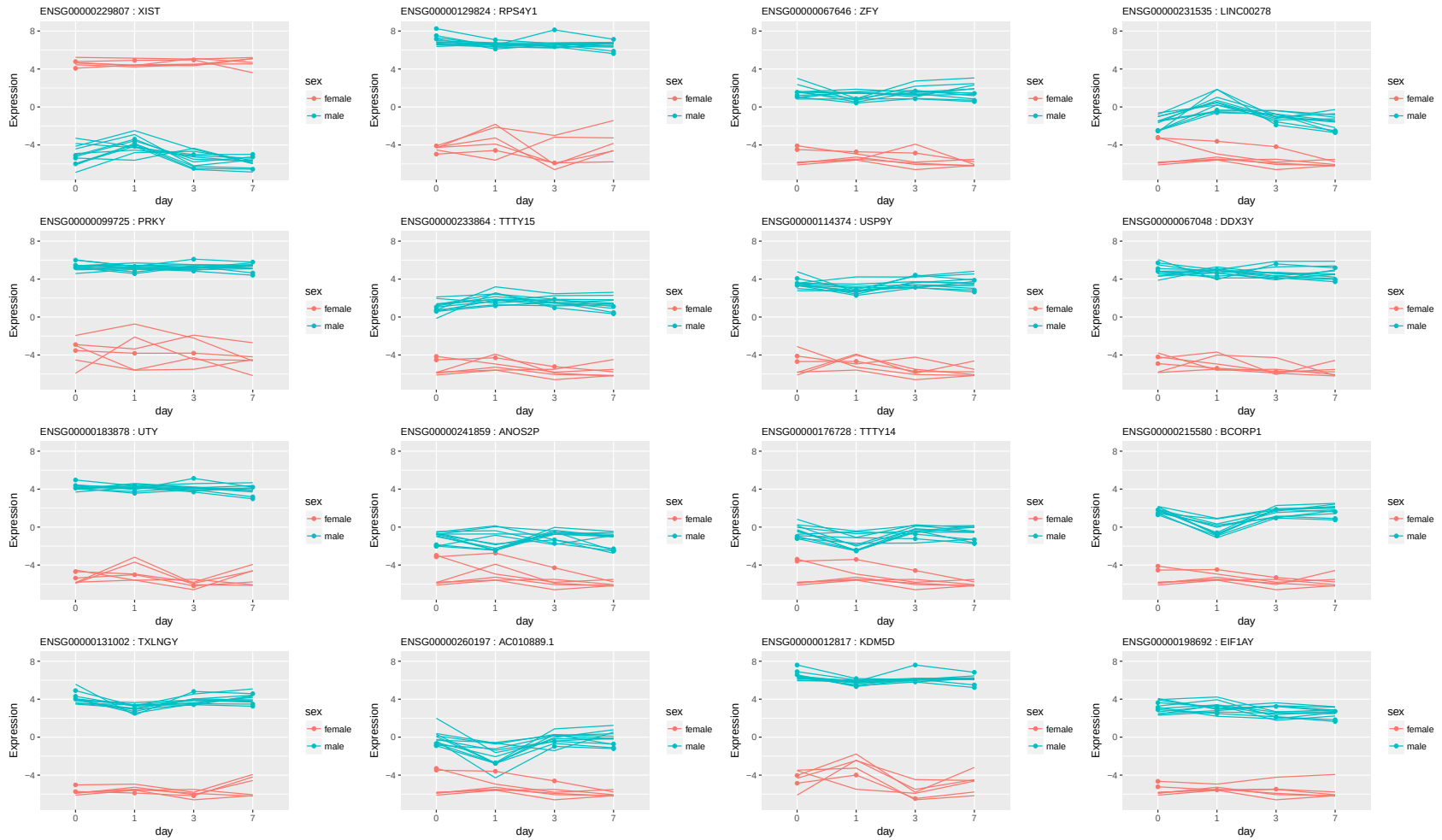
$$\text{RMSE} = f(prop_{NA}, \rho_j = \mathbf{0.9}, \rho_t = \mathbf{0.9})$$





missMDA is the most efficient method, especially in the large proportion of missing values.
All methods seem to learn from temporal structure.

Genes selected along the first component



Genes selected along the second component