2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)

High-dimensional multi-block analysis of factors associated with thrombin generation potential

Hadrien Lorenzo*, Misbah Razzaq†, Jacob Odeberg‡, Pierre-Emmanuel Morange§,
Jérôme Saracco¶, David-Alexandre Trégouët† and Rodolphe Thiébaut*

* Univ. Bordeaux, Inria BSO, Inserm U1219 Bordeaux Population Health Research Center, SISTM team, France,
Emails: hadrien.lorenzo@u-bordeaux.fr, rodolphe.thiebaut@u-bordeaux.fr
† Univ. Bordeaux, Inserm U1219 Bordeaux Population Health Research Center, VINTAGE team, France,
Emails: misbah.razzaq@inserm.fr, david-alexandre.tregouet@inserm.fr
‡Department of Proteomics, School of Biotechnology, Science for Life Laboratory,
KTH Royal Institute of Technology, Stockholm, Sweden,
Email: jacob.odeberg@scilifelab.se
§Laboratory of Haematology, La Timone Hospital, Marseille, France,
Email: pierre.morange@ap-hm.fr
¶ Inria BSO, CQFD team, CNRS UMR5251 Institut Mathmatique de Bordeaux, ENSC Bordeaux INP, Talence, France,
Email: jerome.saracco@inria.fr

***Abstract*—The identification of novel biological factors associated with thrombin generation, a key biomarker of the coagulation process, remains a relevant strategy to disentangle pathophysiological mechanisms underlying the risk of venous thrombosis (VT). As part of the MARseille THrombosis Association Study (MARTHA), we measured whole blood DNA methylation levels, plasma levels of 300 proteins, 3 thrombin generation biomarkers (endogeneous thrombin potential, peak and lagtime), clinical and genetic data in 700 patients with VT. The application of a novel high-dimensional multi-levels statistical methodology we recently developed, the data driven sparse Partial Least Square method (`ddsPLS`), on the MARTHA datasets enabled us 1/ to confirm the role of a known mutation of the variability of endogenous thrombin potential and peak, 2/ to identify a new signature of 7 proteins strongly associated with lagtime.**

***Index Terms*—Multi-Omics, High Dimensional Data, Missing Data, SVD, Partial Least Square, Variable Selection, Multi-Block Analysis, Machine Learning, Thrombine Generation**

I. Introduction

Venous thrombosis (VT) is a complex disease characterized by the formation of a blood clot in a deep vein that can later break free and travel to the lung to provoke pulmonary embolism.

In this process, thrombin is a key molecule and individuals that have a strong capacity to produce thrombin are at higher risk for VT. The thrombin generation potential (TGP) of an individual can be measured by thrombin generation assays that capture the complete dynamics of the coagulation process following clot formation. TGP is generally summarized by the use of three associated parameters, the $LagTime$, time after the lag-phase that follows trigger of coagulation until the initiation of thrombin generation, the $Peak$, the maximum amount of thrombin that can be produced, and the ETP that corresponds to the area under the thrombogram curve (i.e. amount of thrombin generated), see Figure 1. While ETP and $Peak$ are highly correlated, the $LagTime$ variable generally shows moderate correlation with the two other markers [1].

We had previously demonstrated that the $F2$ G20210A mutation was the main genetic factor contributing to ETP and $Peak$ plasma variability without impacting $LagTime$, see [1]. In addition, using a methylation-wide association strategy, we reported that DNA methylation marks in whole blood did not strongly associate with TGP biomarkers [2]. Motivated by the search for novel molecular determinants of TGP biomarkers, plasma samples of MARTHA participants were profiled for 200 proteins using a recent high-throughput technology [3]. In the current work, these proteomics data were studied for association with ETP, $Peak$ and $LagTime$. Using a recently developed high-dimensional multi-*omics* algorithm, we jointly studied the association between ETP, $Peak$ and $LagTime$ variables through a multivariate analysis with known TGP determinants, the proteomics data and DNA methylation that was available only in a subsample of participants. The proposed methodology uses the data driven sparse Partial Least Square method (`ddsPLS`) to predict each of the three TGP biomarkers in a multivariate fashion taking into account possible missing values in the covariates using information in the response matrix while keeping sparsity in the context of high-dimensional data where the number of predictors p is in the same order or superior to the number of individuals n. `ddsPLS` has been implemented in **R** and **Python**, and is available on **CRAN** [1] and on **PyPi** [2], respectively.

The present work is divided in four parts. The first part describes the data set. The second part describes the method

[1]See https://cran.r-project.org/package=ddsPLS.
[2]See https://pypi.org/project/py_ddspls/.

2372-9198/19/$31.00 ©2019 IEEE
DOI 10.1109/CBMS.2019.00094

and the third one describes the results applying the method to the data set. The fourth part concludes and gives perspectives.

II. Notation

Lets denote the matrices in bold upper cases and vectors in bold lower cases. In the case of greek letter objects, matrices are represented with underlined greek letters and no difference in the notation in those case between matrices and column vector matrices. $\mathbf{X}$ denoted matrices account for predictor associated matrices and $\mathbf{Y}$ for response matrices. Indices are used when necessary. Without further indication, all matrices are standardized, zero mean and unit variance for each column. n is used to represent the number of rows of a matrix, and also the number of individuals. The number of columns of a $\mathbf{X}$ matrix, resp. $\mathbf{Y}$ matrix, is represented by the letter p, resp. by the letter q. The r^{th} column vector of a given matrix $\mathbf{U}$ is denoted $\mathbf{u}^{(r)}$. Matrix sets are denoted as $\mathbb{R}^{n\times p}$ and correspond to n rows and p columns matrices. $\mathbb{I}_n$ denotes identity matrix with n columns. $||.||_2$ denotes the $\mathcal{L}_2$-norm of a given vector and $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{n\times 1^2} \rightarrow \mathbf{x}^T\mathbf{y}$ the associated $\mathcal{L}_2$-cross-product where the transpose operator is symbolized by ".^{T}". Let the proportion of variance of a matrix $\mathbf{Y}$ explained by a matrix $\mathbf{X}$ be expressed as

$$\frac{||\mathbf{XB}(\mathbf{Y},\mathbf{X})||_F^2}{||\mathbf{Y}||_F^2} \times 100,$$

where $||.||_F$ is the Frobenius norm and $\mathbf{B}(\mathbf{Y},\mathbf{X})$ is the multivariate coefficient regression matrix which solves the Ordinary Least Square (OLS) problem $\max_{\mathbf{B}} ||\mathbf{Y} - \mathbf{XB}||_F^2$.
Let be denoted by R the number of dimensions built in the following.

III. Materials

This work was based on the MARseille THrombosis Association Study (MARTHA) cohort including patients with VT recruited at the Thrombophilia center of La Timone hospital (Marseille, France) between January 1994 and October 2005. This study has been extensively described in previous works [1], [4].

A. Biological measurements

Thrombin generation potential (TGP) was measured in platelet-poor plasma (PPP) of 705 individuals using the CAT method as described in [1]. Plasma levels of these individuals were profiled for 384 antibodies (referred thereafter to as HPAs as they were selected from the Human Protein Atlas) targeting 234 proteins using high-affinity bead array technology. These proteins have been selected because of their potential role in the coagulation and fibrinolysis cascades or because they have been reported to be associated with cardiovascular traits. From this sample of participants, 350 have been epityped for whole blood DNA methylation (referred to as mDNA) using the Illumina H450K array as described in [4]. For the current application, we only used the 3,174 most variable and relevant

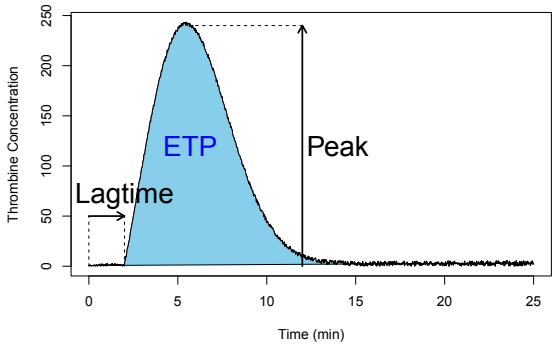

Fig. 1. Thrombin generation test curve and it main features.

CpG sites [3] among the ≈380,000 measured CpG sites.
All patients were genotyped for the F2 G20210A mutation and measured for Protein S (PS), Protein C (PC) and Antithrombin (AT), the three main natural coagulation inhibitors, as previously described [4]. From the initial set of 705 participants with TGP measurements, 9 were excluded from the final analyses because they exhibited extreme outlier values for $LagTime$.

In the studied population, the correlations between ETP and $LagTime$, ETP and $Peak$ and $LagTime$ and $Peak$ were 0.17, 0.77 and 0.013, respectively.

B. Missing sample structure

All studied variables were divided in seven matrices, ie. blocks, according to the nature of the data. Blocks were of unequal sizes because of missing values. Table I shows the division of missing samples in each data set. Only five blocks showed samples with missing data. The number of samples missing is provided according to each couple of blocks ($Block_{row}$, $Block_{column}$). Colors of rows and columns symbolize the same blocks respectively and so the diagonal represents the number of missing samples for the single corresponding block. The 696 participants selected for the present analysis had no missing information on the $\mathbf{Y}$ block defined by the three TGP biomarkers (ETP, $Peak$, $LagTime$). As a consequence, working on available data only would lead to the exclusion of 470 individuals, see Table I.

IV. Methods

Many methods are available for high-dimensional data analysis. For example, the sparse PLS, see [5], is popular but deal with missing values in a two-steps approach and not in a supervised framework. Non multi-block and non supervised methods such as softImpute [6] can also deal with missing values but not in the context of supervised analysis. imputeMFA, see [7], deals with missing values but is not

[3]We removed the CpG sites for which InterQuartiles Range was lower than 0.05 and for which maximum absolute correlation with any of the three TGP biomarkers was below 0.25.

TABLE I
MISSING INFORMATION ACCORDING TO THE TYPE OF DATA IN THE 696 PARTICIPANTS INCLUDED IN THE STUDY. ONLY TWO SETS OF INDIVIDUALS, CORRESPONDING TO [a] AND [b] IN THE FOLLOWING TABLE, ARE MISSING IN THREE BLOCKS.

BLOCK	HPAs	PS	F2_G20210A	PC	AGE	BMI	AT	mDNA
HPAs	4	0	0	0	0	0	0	4
PS	0	3	0	2	0	0	0	1
F2_G20210A	0	0	0	0	0	0	0	2
PC	0	2	0	4	0	0	0	4
AGE	0	0	0	0	0	0	0	0
BMI	0	0	0	0	0	2	0	18
AT	0	0	0	0	0	0	0	0
mDNA	4	1	2	4	0	18	0	433
Total	8	6 +6[a]	2	10 +8[a,b]	0	20 +2[b]	0	462 +8[a,b]
	8	12	2	18	0	22	0	470

[a] **6** individuals are missing for {**PS**,**PC**,**mDNA**} consequently.
[b] **2** individuals are missing for {**BMI**,**PC**,**mDNA**} consequently.

supervised and shows poor results in high dimensional setting. Support Vector Machine (SVM), introduced by [8] and neural networks do not allow variable selection and generally require a very large number of samples to achieve efficiency and accuracy. Aggregative methods such as random forests [9] are also very attractive methodologies but are computationally time demanding. We here propose to apply a data driven sparse Partial Least Square (`ddsPLS`, see [10]) that has the multiple advantages of addressing high-dimensional multi-block supervised problems, multivariate regression or classification, in the presence of missing data with regularization and variable selection, and in a time effective manner.
A variance-covariance matrix soft-thresholding algorithm inspired from `ddsPLS` tools allows regularization and variable selection while missing data imputation is performed thanks to the `Koh-Lanta` algorithm that the authors developed. Both of those aspects are described below.

A. A `PLS` (Partial Least Square) inspired method

`PLS` looks for common structure, through singular value decomposition (`SVD`) decomposition, to $\mathbf{X}$ and $\mathbf{Y}$ maximizing $(\mathbf{YX})^T(\mathbf{YX})$. Only the first principal vector is built through the **weight** vector $\mathbf{u}$, resp. $\mathbf{v}$, for the $\mathbf{X}$ part, resp. the $\mathbf{Y}$ part, corresponding to **component** vector $\mathbf{t} = \mathbf{Xu}$, resp. $\mathbf{s} = \mathbf{Yv}$. A technical step, denoted as **deflation** allows to remove the information carried by that **component** to both of the matrices. Classically, once that deflation is performed, the same procedure is done on the residual matrices, building a second then a third up to build R components, as ordered by the user. Deflation is not performed in our method for reasons exposed in [10].

B. `ddsPLS`: a three steps algorithm

The first step permits to extract the marginal R-dimensional common structure of each block t and the block $\mathbf{Y}$. The second step finds a R-dimensional common structure to the T different R-dimensional components and the block $\mathbf{Y}$. The last step builds the linear regression matrix predicting block $\mathbf{Y}$.
The soft-thresholding operation, $\forall\lambda \in [0,1]$, denoted as S_λ, applied to any matrix to each of its coefficient such as $S_\lambda : x \to sign(x)(|x| - \lambda)_+$, where $sign$ gives the sign of a real, $|.|$ denotes the absolute value and $(.)_+$ the max between its argument and 0.
That operator is applied to the different variance-covariance matrices, let say the multi-block data set is built on T blocks and so, $\forall t \in \{1, \cdots, T\}$, the following `SVD` decompositions are performed

$$\max_{\mathbf{U}_t \in \mathbb{R}^{p_t \times R}} \sum_{r=1}^{R} ||S_\lambda\Big(\frac{\mathbf{Y}^T\mathbf{X}_t}{n-1}\Big)\mathbf{u}_t^{(r)}||_2^2$$
$$\text{s.t.} \quad \mathbf{U}_t^T\mathbf{U}_t = \mathbb{I}_R,$$

where each $\mathbf{u}^{(r)}$ is the r^{th} **weight** associated with the block $\mathbf{X}_t$ and the corresponding r^{th} **component** is denoted as $\mathbf{t}_t^{(r)} = \mathbf{X}_t\mathbf{u}_t^{(r)} \in \mathbb{R}^{n\times 1}$. The soft-thresholding operation implies that some coefficients are naturally put to 0 in the resulting matrix permitting efficient and sparse **weight** extraction. In the following example $\mathbf{X}$, resp. $\mathbf{Y}$, is defined through $p = 3$, resp. $q = 2$, variables and the **weight** matrix is indeed sparse

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 \\ 0.5 & 0.2 \\ 0.6 & 0.1 \end{bmatrix}}_{\frac{(\mathbf{Y}^T\mathbf{X})^T}{n-1}} \underset{\lambda=0.2}{\Longrightarrow} \underbrace{\begin{bmatrix} 0 & 0.7 \\ 0.3 & 0 \\ 0.4 & 0 \end{bmatrix}}_{S_\lambda\left(\frac{(\mathbf{Y}^T\mathbf{X})^T}{n-1}\right)} \underset{\text{SVD}}{\Longrightarrow} \mathbf{U} = \Bigg[\underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{\mathbf{u}^{(1)}}, \underbrace{\begin{bmatrix} 0 \\ 0.6 \\ 0.8 \end{bmatrix}}_{\mathbf{u}^{(2)}}\Bigg].$$

Once each t structure is defined, a common structure to the T blocks is build thanks to another R-dimensional `SVD` decomposition applied to the concatenation of the T different R-dimensional descriptions $S_\lambda\Big(\frac{\mathbf{Y}^T\mathbf{X}_t}{n-1}\Big)\mathbf{U}_t$. The part of the built weights corresponding to block t is denoted as **super-weight** and is symbolized by $\underline{\boldsymbol{\beta}}_t \in \mathbb{R}^{R\times R}$. It corresponds to the impact of the previously selected variables of every blocks on the **super** description of the block $\mathbf{Y}$ for each **super-component** $\mathbf{X}_t\mathbf{U}_t\underline{\boldsymbol{\beta}}_t \in \mathbb{R}^{n\times R}$. The **scaled super-weights** $\mathbf{U}_t\underline{\boldsymbol{\beta}}_t \in \mathbb{R}^{n\times R}$ permit to interpret the effect of one variable of a given block on the considered **super-component** since its absolute value is inferior or equal to 1 and, for a given **super-component**, variables are ordered in term of importance by their **scaled super-weights**.
A last step builds a regression model such as

$$\mathbf{Y} \approx \sum_{t=1}^{T} \mathbf{X}_t\mathbf{B}_t \in \mathbb{R}^{n\times q},$$

using Moore-Penrose pseudo-inverse in the case of regression and a linear discriminant analysis model (`LDA`) is built on the basis of the R **super-components** to predict new individual classes.

C. The Koh-Lanta algorithm

Missing values might appear in a $train$ or in a $test$ data set but the model is built on the $train$ part and tested on the $test$ part. The way of dealing with the missing values must therefore be different in both cases, which is the objective of the *Koh-Lanta* algorithm developped in [10]. The *Tribe Stage* permits imputation in the train data set and must enrich the being built model while the *Reunification Stage* estimates missing values in the $test$ data set with no modification of the model.

The *Tribe Stage* is an iterative procedure which uses, considering a given block for which samples are missing, pieces of information present for the other samples in the block $\mathbf{Y}$ [4] are used to build a `ddsPLS` model on non missing valued data sets and then estimates potential positions of missing samples. At each iteration of the algorithm, only previously selected variables are taken into account and others are removed from the analysis, this is the *Tribe Stage* of *Koh Lanta*. Convergence is controlled with a maximum number of iterations and a minimum variation of the Moore-Penrose description of the $\mathbf{X}$ part, according to the $\mathcal{L}_2$-cross-product.

The *Reunification Stage* uses the final model of the *Tribe Stage* to predict the missing values of the $test$ data set in a single loop.

D. A new parametrization of the `ddsPLS`

So far `ddsPLS` models depend on two user tunable parameters which are

- $R \in \mathbb{N}^\star$: The number of dimensions to be built.
- $\lambda \in [0, 1]$: The correlation threshold above which an interaction between a variable of a $\mathbf{X}$ block variable and a $\mathbf{Y}$ variable is not taken into account.

The `ddsPLS` has been modified and parameter λ has been replaced with parameter $L_0 \in \mathbf{N}^\star$ which represents the maximum number of $\mathbf{X}$ variables to be selected in the model. But for a given L_0, λ is no unique. The chosen rule was to consider the model corresponding to the smallest λ for a given L_0 because it would eventually give the same degree of sparsity but gathers more information since soft-thresholding operation removes less information in that case.

Also this solution would certainly be efficient in the cases of data sets with variance-covariance matrices particularly sensible to down-sampling but this has not been explored yet.

E. A new initialization of the missing values in the `ddsPLS`

The mean imputation for initializing missing values was so far considered. This drives the algorithm to bias the correlations between the $\mathbf{Y}$ variables and the predictors. In the context of many missing samples, which is the case in the $Methylation$ block ($\approx 67.5\%$ of missing samples) that bias implies sub-optimal choice over the soft-thresholding operation. The most correlated $Methylation$ CpG site is $cg08719422$ and its highest absolute correlation with one of the three TGP biomarkers is equal to:

- ≈ 0.404 on the present individuals,
- ≈ 0.232 if missing samples have been imputed to mean.

In that context it has been decided to slightly modify the missing sample imputations at the initialization step. And so, $\forall t \in \{1, \cdots, T\}$ such as some rows of that block are missing, the others are not missing. The algorithm builds a `ddsPLS` with $Block_t$-non missing samples as a response matrix and the $\mathbf{Y}$ matrix (which is the "official" response matrix) for the $Block_t$-non missing samples as the predictor matrix. Then it computes the predicted values on the $\mathbf{Y}$ matrix for the $Block_t$-missing samples. Those are the initialization missing samples. The regularization and the number of components are taken accordingly to the choice of the user.

V. Results

A. Cross validation

The two parameters of the `ddsPLS` model were tuned with 40-folds cross-validation using a unitary step on L_0. The error criterion chosen is the Mean Square Error in Prediction (MSEP), R is upper bounded by the number of columns in the $\mathbf{Y}$ block, which is equal to 3. While the identified first and second components (referred thereafter to as super-components) substantially contributed to the model by explaining around 24% and 11% of the total variance of $\mathbf{Y}$, respectively, the third component added little information with less than 2% of the variance explained. The MSEP curve displayed on Figure 2 clearly showed that $Peak$ was badly predicted in cross-validation since the error was always above 1, the upper-bound limit for prediction when the outcome variable is standardized. The ETP showed a minimum MSEP for $L_0 = 2$ which then dramatically increased for higher L_0. By contrast, $LagTime$ reached a clear minimal MSEP for $L_0 = 12$. For the following results, we will focus on the model identified with $L_0 = 12$ that provided the best predictions for $LagTime$ and denoted as $\mathcal{M}_{TGP}$ in the following.

TABLE II
SELECTED VARIABLES AND MSEP ERRORS OF $\mathcal{M}_{TGP}$

	Variables selected			MSEP			
L_0	***HPA***	***Bio.***	***mDNA***	***Lagtime***	***Peak***	***ETP***	***Mean***
12	7	2[c]	3	0.796	1.14	1.05	0.994

[c] $F2\_G20210A$ and AGE are selected.

B. Description of the selected model

Model $\mathcal{M}_{TGP}$, see Section V-A, selects a total of 12 variables. In Figure 5 are shown the values, per **super-component**, of the **scaled super-weights**, as described in IV-B, as well as the variance they explain for each of the three TGP biomarkers. Each row of Figure 5 describes the variables that compose one of the two identified **super-components**

[4] Only the $\mathbf{Y}$ is used as a covariate matrix, through its **super-scores**, to use only dimensions linked to the current model.

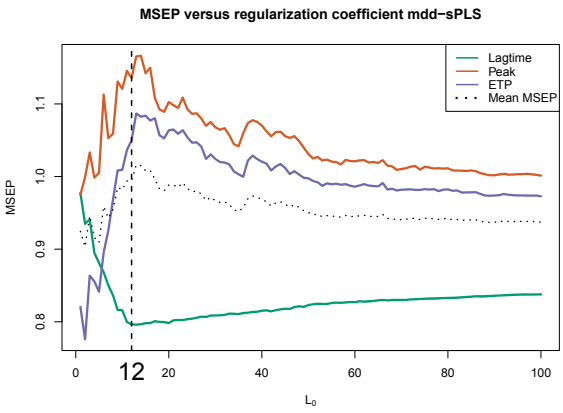

Fig. 2. Mean Square Error in Prediction (MSEP) for $R = 2$ with mean error curves and the mean of both of those curves. The dashed line represents the chosen mode, for $L_0 = 12$. The **Y** variables were standardized before application of the cross validation procedure.

with their corresponding **scaled super-weights**. The second column shows the percentage of variance explained by the **super-component** for each of the three TGP biomarkers. It is important to note that those variances are computed on the imputed data sets, once the model is built.

The first **super-component** explains 24% of the total **Y** joint variance of the three TGP biomarkers, with a decreasing contribution of $Peak$, ETP and $Lagtime$ (40% of explained variance, 29% and 1.9%, respectively). Interestingly, the first **super-component** was mainly driven by one CpG site, `cg08719422`, with some minor contribution of the $F2$ G20210 mutation and two other CpG, `cg18876487` and `cg11015505`. The pattern of correlation between selected features is given on Figure 3.

The three selected CpG sites were correlated with each other and the correlation between `cg08719422` and $Peak$ was 0.64. While this first **super-component** demonstrated good descriptive characteristics with high percentages of variance explained for $Peak$ and ETP, its predictive properties were relatively poor as illustrated by the high values of the corresponding MSEP that were slightly greater than 1.

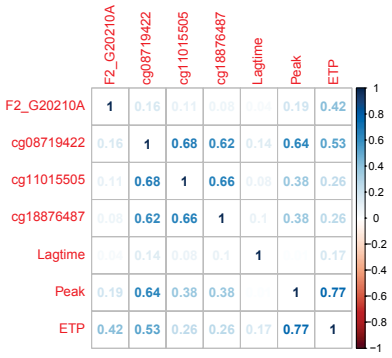

Fig. 3. Correlation structures of super-component 1 of the model

The second **super-component** was mainly driven by Age and seven antibodies, and was mainly associated with $LagTime$. It explained about 23% of the $LagTime$ variance but only 5.8% and 4.3% for ETP and $Peak$ variability, respectively. Figure 4 shows strong correlations between proteins but also with Age, which helps imputation in $test$ data sets. Four of these seven antibodies were targeting proteins of the complement cascade (C5, C9) and two were antibodies targeting proteins (C4BPA, PROS1) associated with the Protein S pathway [11].

The same model including Age and these seven antibodies was identified when the `ddsPLS` algorithm was applied on the $LagTime$ variable only (data not shown).

In order to replicate the observed association of the second super-component with $LagTime$, we investigated it in an independent sample of 133 MARTHA patients measured for $LagTime$ and for the seven antibodies using the same technique but without $Methylation$ data available. In this independent population, we observed a trend for a positive correlation between the seven antibodies signature (plus age) and $LagTime$ (r = 0.16, p = 0.069).

VI. Clinical Interests and Future Works

Using `ddsPLS`, a recently developed method for multi-$omics$ data analysis, we identified a biomarker signature composed of antibodies targeting seven distinct proteins that explain about 20% of the plasma variability of $LagTime$. Interestingly, this signature is enriched in proteins belonging to the complement cascade adding support to the emerging link between the complement and coagulation pathways. Further works are needed to assess the relevance of this signature with respect to the risk of VT.

`ddsPLS` has also permitted to extract information from $Methylation$ data set filed with $\approx 70\%$ of missing values, sadly `ddsPLS` does not yet permit to correctly predict information on $test$ data sets with such a proportion of missing values but this will drive further researches.

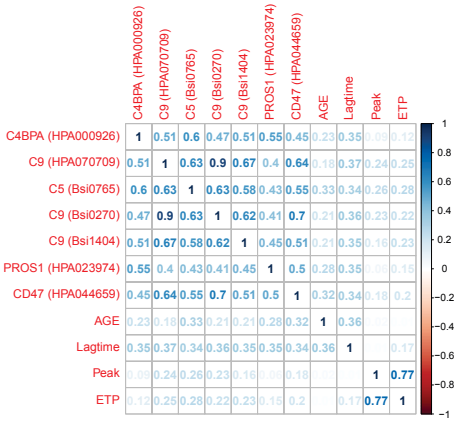

Fig. 4. Correlation structures of super-component 2 of the model

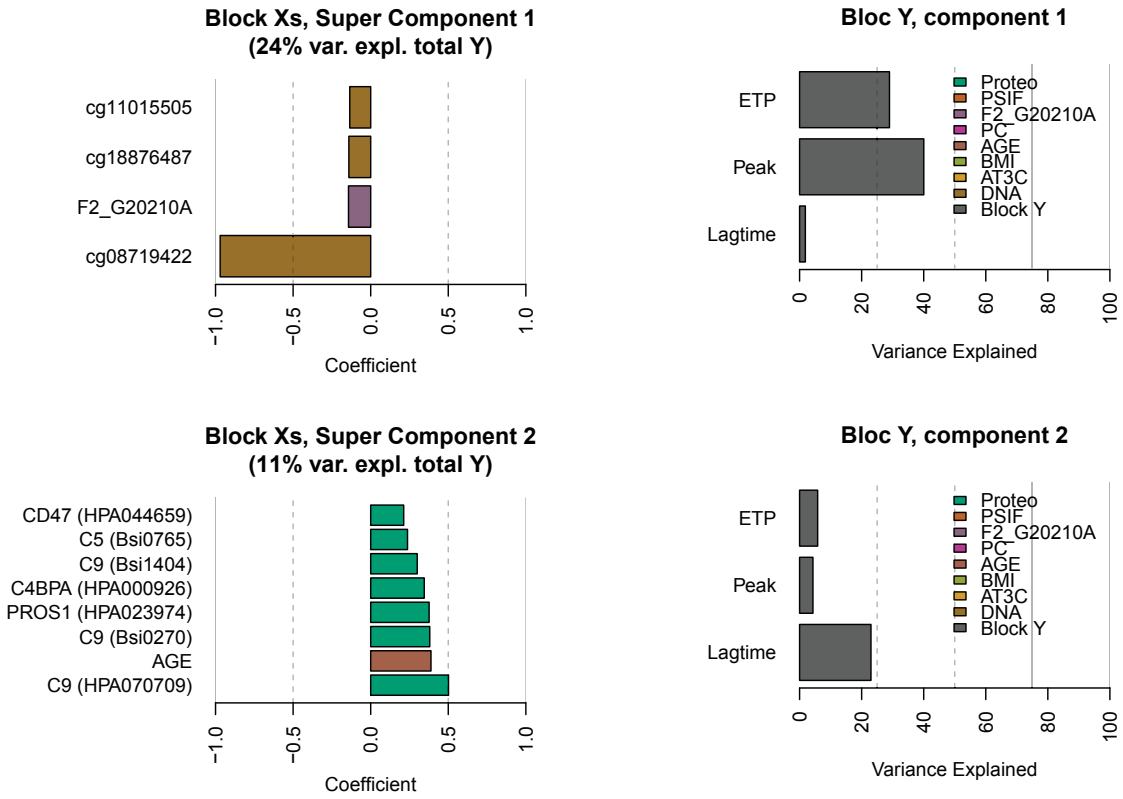

Fig. 5. Scaled super-weights per super-component and variance explained per response variable per component.

Further developments are also needed for widening the use of the `ddsPLS` method in the context of mutli-*omics* epidemiological cohorts. These include the handling of missing data in the response block, denoted as $\mathbf{Y}$, and the possibility of integrating millions of genetic variables produced by high-throughput genotyping/sequencing instruments, the latter raising some computational challenges.

ACKNOWLEDGMENT

Hadrien Lorenzo is supported by a 2016 Inria-Inserm thesis grant *Médecine Numérique* (for *Digital Medicine*). Misbah Razzaq is supported by a grant from the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013]. David-Alexandre Trégouët is supported by the EPIDEMIOM-VTE Senior Chair from the Initiative of Excellence of the University of Bordeaux.

REFERENCES

[1] A. Rocanin-Arjo, W. Cohen, L. Carcaillon, C. Frère, N. Saut, L. Letenneur, M. Alhenc-Gelas, A.-M. Dupuy, M. Bertrand, M.-C. Alessi *et al.*, "A meta-analysis of genome-wide association studies identifies orm1 as a novel gene controlling thrombin generation potential," *Blood*, vol. 123, no. 5, pp. 777–785, 2014.

[2] A. Rocañín-Arjó, J. Dennis, P. Suchon, D. Aïssi, V. Truong, D.-A. Trégouët, F. Gagnon, and P.-E. Morange, "Thrombin generation potential and whole-blood dna methylation," *Thrombosis research*, vol. 135, no. 3, pp. 561–564, 2015.

[3] K. Drobin, P. Nilsson, and J. M. Schwenk, "Highly multiplexed antibody suspension bead arrays for plasma protein profiling," in *The Low Molecular Weight Proteome*. Springer, 2013, pp. 137–145.

[4] T. Oudot-Mellakh, W. Cohen, M. Germain, N. Saut, C. Kallel, D. Zelenika, M. Lathrop, D.-A. Trégouët, and P.-E. Morange, "Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein c anticoagulant pathway: the martha project," *British journal of haematology*, vol. 157, no. 2, pp. 230–239, 2012.

[5] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse pls for variable selection when integrating omics data," *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008.

[6] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3367–3402, 2015.

[7] J. Josse and F. Husson, "missmda: a package for handling missing values in multivariate data analysis," *Journal of Statistical Software*, vol. 70, no. 1, pp. 1–31, 2016.

[8] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] H. Lorenzo, J. Saracco, and R. Thiébaut, "Supervised learning for multi-block incomplete data," *arXiv preprint arXiv:1901.04380*, 2019.

[11] A. Buil, D.-A. Trégouët, J. C. Souto, N. Saut, M. Germain, M. Rotival, L. Tiret, F. Cambien, M. Lathrop, T. Zeller *et al.*, "C4bpb/c4bpa is a new susceptibility locus for venous thrombosis with unknown protein s independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies," *Blood*, pp. blood–2010, 2010.