DU BDSI 2019
ENSC

Hadrien Lorenzo
hadrien.lorenzo@u-bordeaux.fr
https://hadrienlorenzo.netlify.com/

3 juillet 2019

ENSC + INSERM + INRIA = ☺

Cursus

- **2009** $\rightarrow$ **2015** : Prépas + Ecole d'ingénieur.
- **2015** $\rightarrow$ **2016** : Ingénieur de recherche au sein de l'équipe SISTM.
- **Depuis 2016** : Thèse de biostatistiques, 3^{me} année, U1219 Bordeaux Population Health, Bourse Inserm-Inria.

Un vaccin contre Ebola et des données à disposition

Essai rVSV-ZEBOV Ebola de phase 1 avec doses échelonnées

- **Premier vaccin** à présenter une efficacité depuis la survenue de la maladie [Henao-Restrepo et al., *The Lancet*, 2017]

Données issues d'un essai vaccinal :

- 18 participants,
- Différents types de données :
- Données répétées,
- Echantillons manquants (30%),
- Problème supervisé

Premier travail

Résultat : [Rechtien et al., *Cell reports*, 2017]

$\longrightarrow$ Faiblesse au niveau de la prise en compte des données manquantes

Travail actuel :

Développement d'une méthode supervisée gérant les données manquantes dans le contexte multi-block de grande dimension avec sélection de variables

Méthodes non supervisées

ACP

Critère d'optimisation :

$$\max_{u} ||Xu||_2^2,$$

où X doit être centrée. On standardise souvent les variables aussi, pour ne pas en privilégier certaines. C'est en fait la décomposition en valeurs propres de la matrices de variance-covariance $X'X$.
Cette méthode est très performante lorsque les interactions sont linéaires. Elle permet alors de définir des variances expliquées, par axe.
Le nombre d'axe étant fixé par l'importance apportée par chaque nouvel axe (critère du coude).
Cette variance expliquée permet de voir à quel point l'information visuelle est représentative de l'ensemble du jeu de données.

ACP

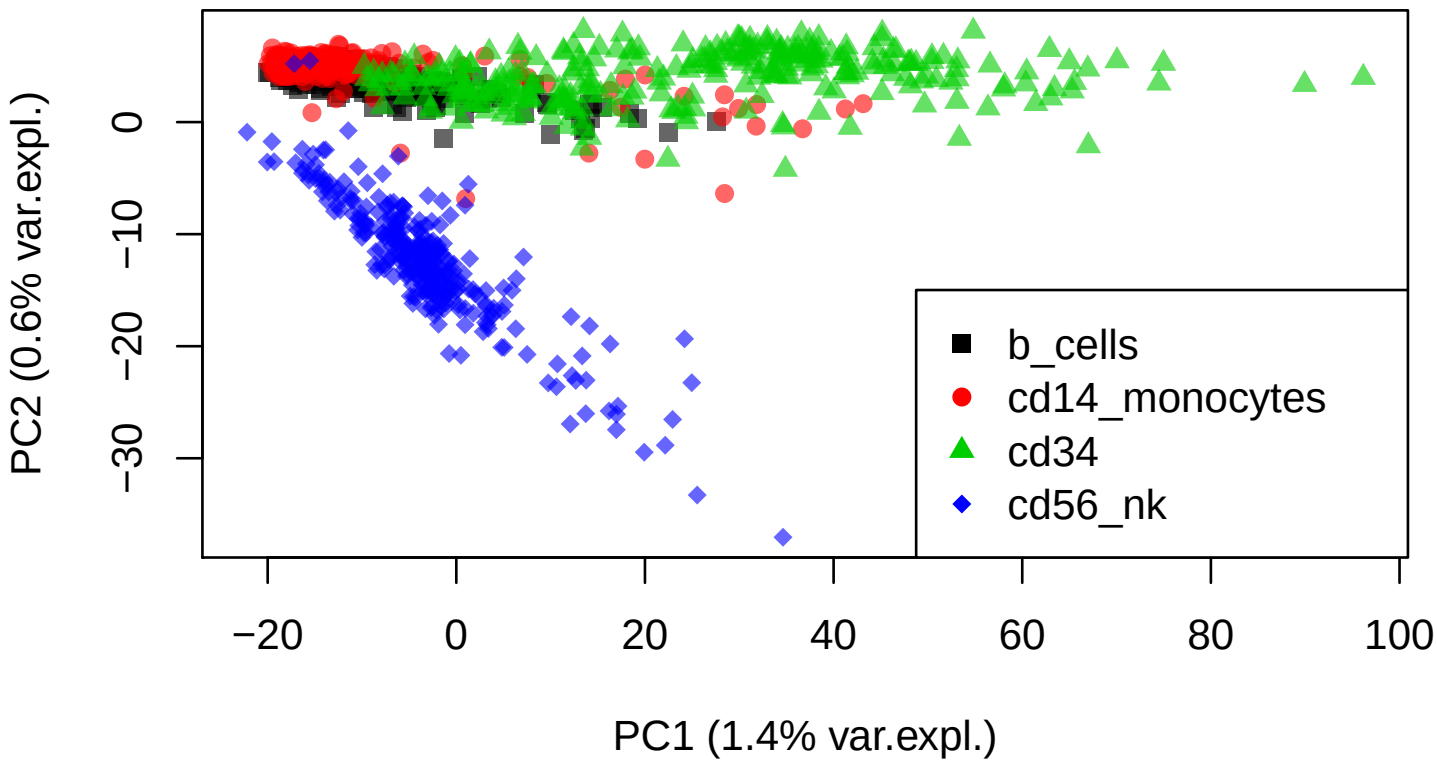


FIGURE – Représentation des deux premiers axes de l'ACP appliquée à un sous-échantillonage des données de 10X [ZHENG et al., *Nature communications*, 2017]

ACP et ses menus soucis

- Soucis de communication si l'on a de l'infos sur plusieurs axes (plus que 2 ou 3).
- Difficultés de l'algorithme si structure complexe : non linéarités par exemple.

Il a été pensé différentes méthodes, plus ou moins fiables, on va rapidemnt présenter **UMAP**.

UMAP, package umap

UMAP, voir [McInnes et Healy, *arXiv preprint arXiv :1802.03426*, 2018], est une méthode de représentation qui recherche des structures globales (comme l'**ACP**) mais aussi locales (comme les **k-plus proches voisins**).
Elle a fait beaucoup parler d'elle car elle répond à l'utilisation de **t-SNE**, voir [Maaten et Hinton, *Journal of machine learning research*, 2008], une méthode très puissante de réduction de dimension mais qui souffre de plusieurs défauts.
L'appli https://distill.pub/2016/misread-tsne/ permet d'entrevoir les problèmes de **t-SNE**.

UMAP

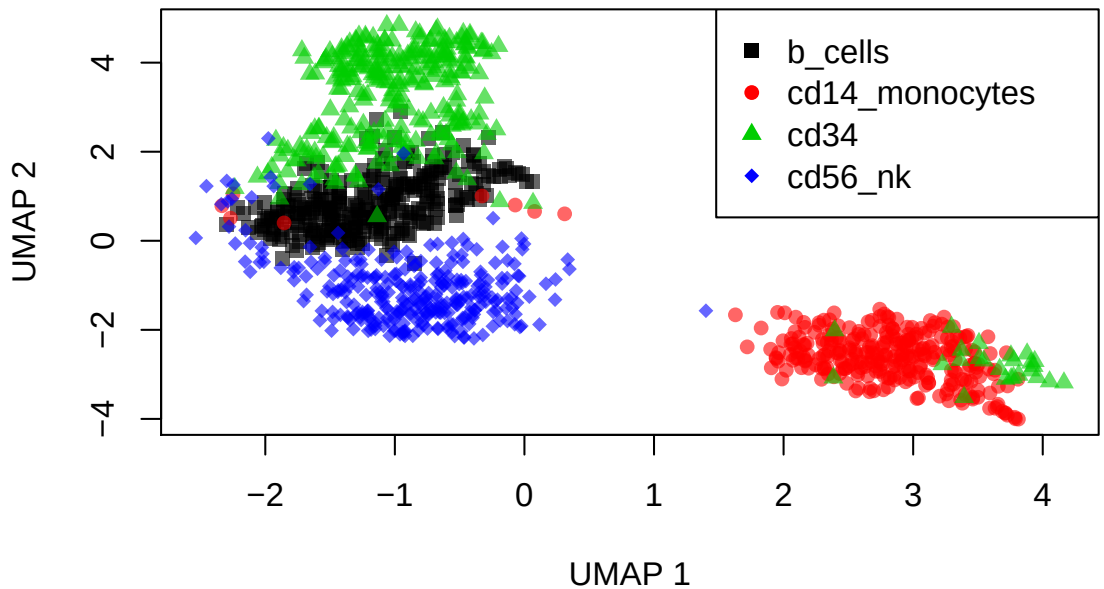


FIGURE – Représentation UMAP d'un sous-échantillonage des données de 10X [ZHENG et al., *Nature communications*, 2017]

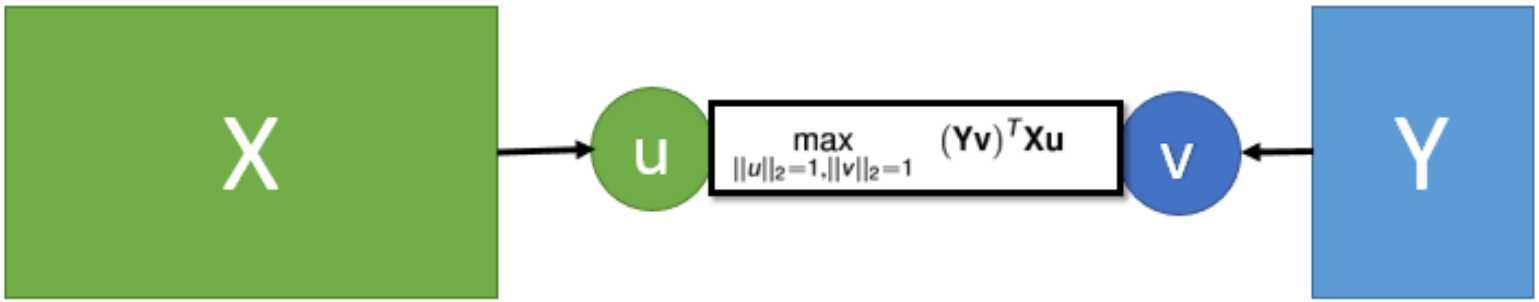
Groupes séparés. Individus d'un même groupe ensemble.
Distributions patates.
Jolie visualisation.

Les données

Single-cell 10X
4 types cellulaires différents.

Méthodes supervisées

Approches PLS [Wold père et fils, 1983]



Équivalent à une recherche de sous-espaces propres (SVD). On appelle :

- **Poids** ou **weights** ou **loadings** u et v : importance donnée d'une variable de X, via u, et de Y, via v.
- **Scores** ou **variates** Xu et Yv : projections de X et de Y dans les sous-espaces définis par u et v.

$\Longrightarrow$ Rechercher dans X l'information qui est très liée à Y.

Résolution du problème de PLS

Utilisation du formalisme lagrangien :

$$\max_{u,v,\alpha_x,\alpha_y} (\mathbf{Y}v)^T \mathbf{X}u - \alpha_x/2(||u||_2^2 - 1) - \alpha_y/2(||v||_2^2 - 1),$$

$\mathbf{X}$ et $\mathbf{Y}$ les matrices échantillons, centrées, des covariables et des variables à prédire. α_x et α_y les coefficients de Lagrange. Alors :

Système :

$$\begin{cases} \partial_u . : & \alpha_x u = \mathbf{X}^T \mathbf{Y} v \\ \partial_v . : & \alpha_y v = \mathbf{Y}^T \mathbf{X} u \\ \partial_{\alpha_x} . : & ||u||_2^2 = 1 \\ \partial_{\alpha_y} . : & ||v||_2^2 = 1 \end{cases}$$

Optimisation :

1. $u \leftarrow \mathbf{X}^T \mathbf{Y} v$
2. $u \leftarrow u/||u||_2$
3. $v \leftarrow \mathbf{Y}^T \mathbf{X} u$
4. $v \leftarrow v/||v||_2$

Régression :

$$\mathbf{Y} \approx \mathbf{X}\mathbf{B}$$
$$\mathbf{B} = \frac{v^T \mathbf{Y}^T \mathbf{X} u}{||\mathbf{X}u||_2^2} u v^T$$

Classification :
LDA sur $(\mathbf{X}u, \mathbf{Y})$

Matrice de variance-covariance

Elle est au centre des approches PLS, via $\mathbf{Y}^T\mathbf{X}$!

La sélection de variables en PLS → sparse PLS

Principe, intérêt et pistes explorées

- Peu de mesures biologiques nécessaires en prédiction.
- Pénalisations $\mathcal{L}_1$ des poids

  $\Longrightarrow$ Sélection des variables et régularisation des données.

Des PLS parcimonieuses, package `mixOmics`

- [Lê Cao et al., 2008], 2 paramètres/axe :

$$\min_{u,v} ||\mathbf{Y}^T\mathbf{X} - vu^T||_F^2 + \lambda_x||u||_1 + \lambda_y||v||_1$$

- [Chun et Keleş, 2010], $M = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$, 3 paramètres/axe :

$\min_{w,c} -\kappa w^T M w + (1-\kappa)(c-w)^T M(c-w) + \lambda_1||c||_1 + \lambda_2||c||_2$
subj. to $w^T w = 1$,

Le Lasso

Formulation

Vous avez sûrement vu la formulation suivante

$$\mathcal{L}(\beta) = ||Y - \beta_0 - X\beta||_2^2 + \lambda|\beta| \tag{1}$$

où λ est le coefficien de Lasso, à tuner par validation croisée.

Cette formulation permet de casser le fléau de la dimension tout en sélectionnant les variables d'intérêt : TOP !
Par contre, comment faire pour un problème de classification ?

Régression logistique pénalisée Lasso !

Régression logistique, package `glmnet`

Soit Y binaire (0 ou 1). On modélise la probabilité conditionnelle $p(Y = 1|X = x)$ telle que

$$ln\frac{p(Y = 1|X = x)}{1 - p(Y = 1|X = x)} = \beta_0 + \sum_{j=1..p} x_j\beta_j,$$

ou de façon équivalente

$$\pi(x) = p(Y = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1..p} x_j\beta_j}}{1 + e^{\beta_0 + \sum_{j=1..p} x_j\beta_j}}$$

que l'on optimise par maximum de vraisemblance, en notant la vraisemblance

$$\mathcal{L}_{logit} = \prod_{i=1}^{n} p(Y_i = y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i},$$

en supposant les n échantillons *iid* suivant une loi binomiale de probabilité $p(Y = 1|X = x_i)$.

Régression logistique multinomiale

Dans le cas multinomiale à K classes, on écrit $K-1$ modèles binaires indépendants tels que, $\forall k \in [1,..,K-1]$, en prenant le groupe K comme référence :

$$ln\frac{p(Y=k|X=x)}{p(Y=K|X=x)} = \beta_{0,k} + \sum_{j=1..p} x_j\beta_{j,k},$$

avec, $p(Y=K|X=x) = 1 - \sum_{k=1}^{K-1} p(Y=k|X=x)$, car c'est une probabilité et en factorisant :

$$\begin{aligned} p(Y=K|X=x) &= 1 - p(Y=K|X=x)\sum_{k=1}^{K-1} e^{\beta_{0,k}+\sum_{j=1..p} x_j\beta_{j,k}} \\ &= \frac{1}{1+\sum_{k=1}^{K-1} e^{\beta_{0,k}+\sum_{j=1..p} x_j\beta_{j,k}}}, \end{aligned}$$

Régression logistique multinomiale

en écrivant alors $\pi_k(x) = p(Y = k|X = x)$, il vient la vraisemblance suivante

$$\mathcal{L}_{logit,multi} = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k(x_i)^{y_{i,k}},$$

avec $\mathbf{y} \in \mathbb{R}^{n,K}$ indicatrice de la classe d'appartenance de l'échantillon i. Que l'on réécrit grâce aux développements précédents.

Pénalisation des méthodes logistiques

Il est possible de pénaliser les méthodes de régression logistique au travers de la **log-vraisemblance**. Pour le cas binomial, par exemple, il vient

$$\mathcal{LL}_{log,pen} = -ln\mathcal{L}_{logit} + \lambda\mathcal{P}(\beta),$$

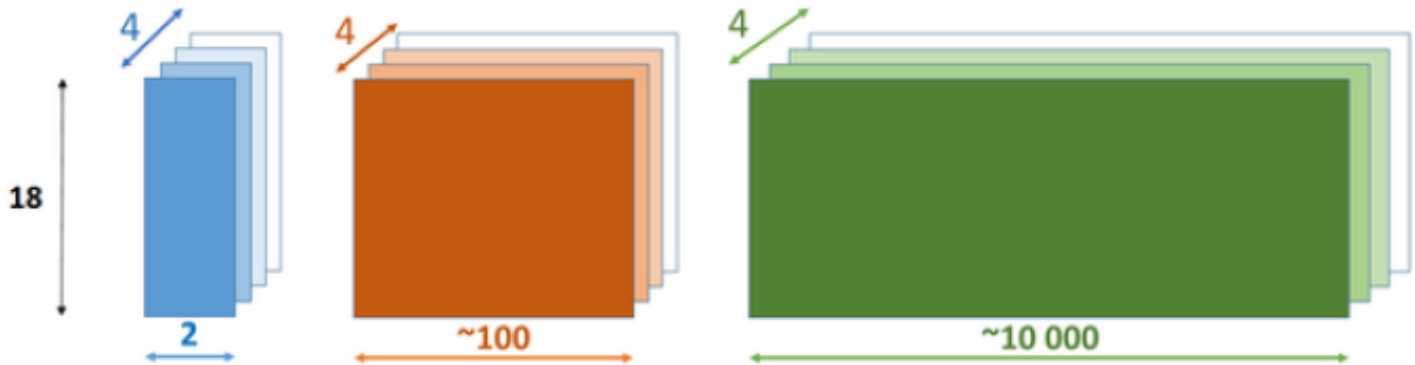
où λ est le coefficient de pénalisation et $\mathcal{P}$ la fonction de pénalisation. Classiquement, on a différentes pénalisations :

- Lasso : $\mathcal{P}(\beta) = |\beta|$
- Ridge : $\mathcal{P}(\beta) = ||\beta||_2^2$,
- Elastic-net : $\mathcal{P}(\beta) = \alpha||\beta||_2^2/2 + (1-\alpha)|\beta|$,

Le nouveau coefficient α permet de gérer la préférence entre le côté Lasso et le côté Ridge. La Ridge ne permet pas de faire de sélection de variable mais il gomme des spécificités individuelles au profit de spécificités de groupe : diminution du sur-apprentissage.

Applications

Motivation



Réponse anticorps	Fonctionnalité cellulaire	Expression génétique
Jours 28, 56, 84, 180	Jours 0, 1, 3, 7	Jours 0, 1, 3, 7

Echantillons manquants : données génétiques

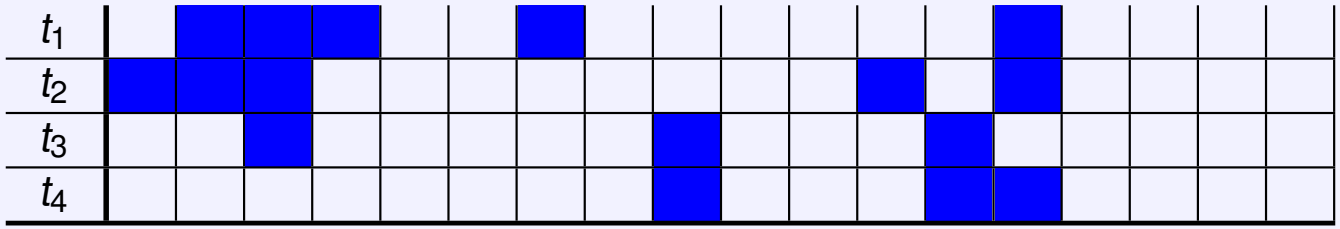


TABLE – Missing path du dataset Ebola rVSV-ZEBOV RNA-Seq où $t_1 = jour_0$, $t_2 = jour_1$, $t_3 = jour_3$ et $t_4 = jour_7$. Colonnes pour les participants.

- 30% de données/échantillons manquants,
- Lien "Missing structure"/"time structure"

Objectif

Prédire la réponse anticorps de façon parcimonieuse en gérant efficacement les données manquantes

Approches existantes

Modèle général, algorithme EM

Combinent des alternances d'estimation :

0. Initialiser les valeurs pour les données manquantes,
1. Estimer une factorisation des données complétées,
2. Estimer les données manquantes,
3. Recommencer en 1. jusqu'à convergence.

... en attente de stabilisation.
$\longrightarrow$ D'autant plus vrai dans le cas de modèles parcimonieux.
Côté utilisateur : difficile à optimiser

Contrainte majeure

Très peu d'individus : la stabilisation est plus difficile à trouver.

... Nous en reparlerons cet après-midi.

Tendreté de la viande

Prédire la tendreté de 5 morceaux de viande pour 10 vaches différentes en fonction de 20 protéines [voir Ellies-Oury et al., *Foods*, 2019]

- Pas de `NA`.
- Peu d'individus.
- Pas d'information en analyse marginale.

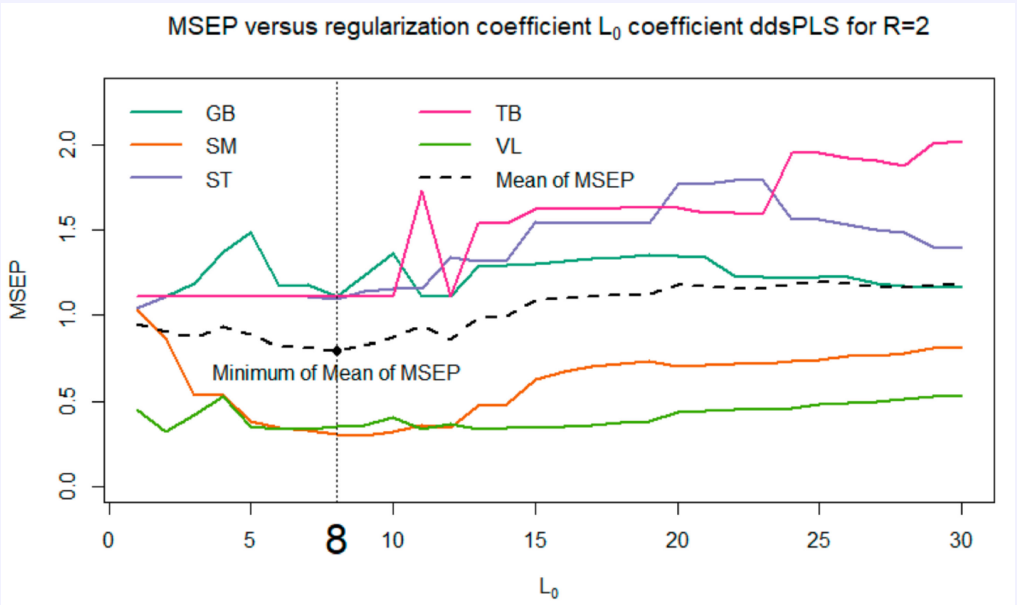


Figure – Courbes de validation croisée, **que pouvez-vous dire ?**

Thrombose veineuse et génétique

Prédire le niveau de thrombine (3 descripteurs) grâce à plusieurs marqueurs (ADN, Protéines, Age, Sexe, ...) [voirLorenzo et al., *CBMS*, 2019]

- Beaucoup de `NA`, surtout pour l'ADN
- Beaucoup d'individus.
- Pas d'information en analyse marginale.

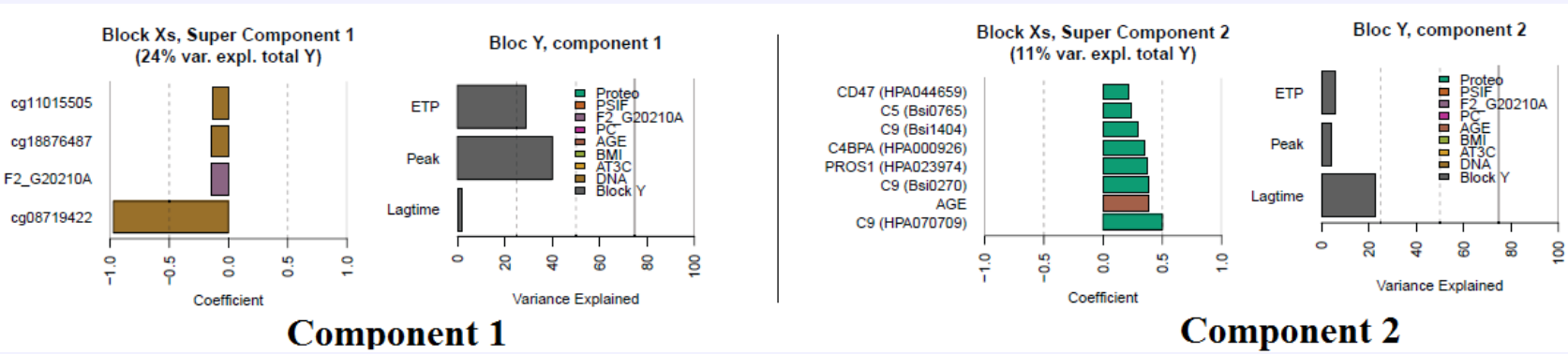


Figure – Modèle final, **que pouvez-vous dire ?**

Thrombose veineuse et génétique (suite)

Un petit quelque chose d'étrange ?

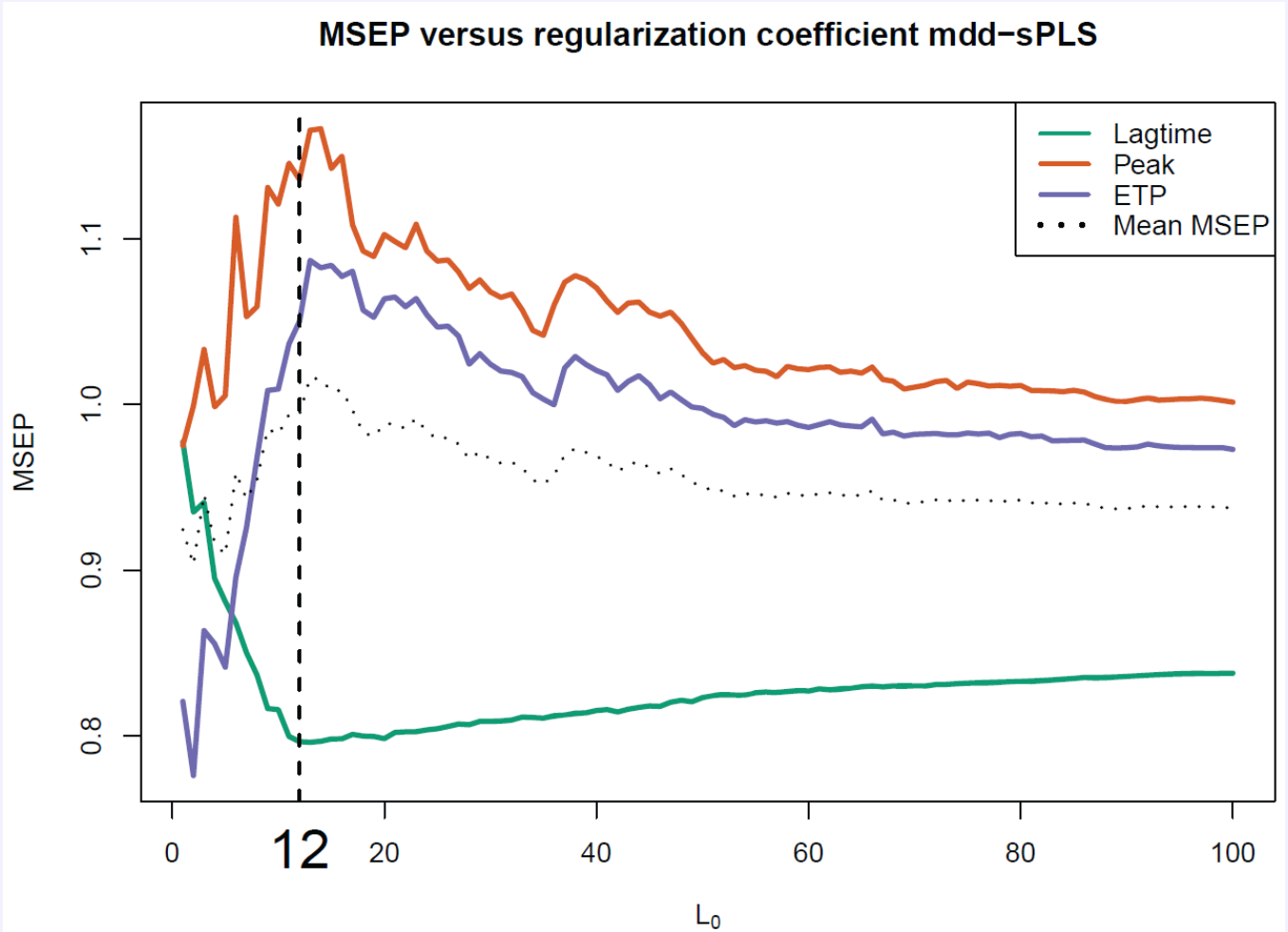


FIGURE – Validation croisée, **que pouvez-vous dire ? A comparer avec le modèle finale**

References

Hyonho Chun et Sündüz Keleş. “Sparse partial least squares regression for simultaneous dimension reduction and variable selection”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 72.1 (2010), p. 3–25.

Marie-Pierre Ellies-Oury et al. “An Original Methodology for the Selection of Biomarkers of Tenderness in Five Different Muscles”. In : *Foods* 8.6 (2019), p. 206.

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), p. 505–518.

Kim-Anh Lê Cao et al. “A sparse PLS for variable selection when integrating omics data”. In : *Statistical applications in genetics and molecular biology* 7.1 (2008).

Hadrien Lorenzo et al. “High-dimensional multi-block analysis of factors associated with thrombin generation potential”. In : *CBMS* 1.1 (2019), p. 1.

Laurens van der Maaten et Geoffrey Hinton. “Visualizing data using t-SNE”. In : *Journal of machine learning research* 9.Nov (2008), p. 2579–2605.

Leland McInnes et John Healy. “Umap : Uniform manifold approximation and projection for dimension reduction”. In : *arXiv preprint arXiv :1802.03426* (2018).

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), p. 2251–2261.