


Projet Génétique ENSAI

Hadrien Lorenzo

hadrien.lorenzo@u-bordeaux.fr

https://hadrienlorenzo.netlify.com/

10 décembre 2018

ENSAI + INSERM Bordeaux + INRIA BSO = 



- ▶ **Depuis 2016** : Thèse de biostatistiques, 3^{me} année, U1219 Bordeaux Population Health, Bourse Inserm-Inria.
- ▶ **2015** → **2016** : Ingénieur de recherche au sein de l'équipe SISTM.
- ▶ **2012** → **2015** : Etudiant de l'école Supélec, Mathématiques Appliquées, Gif-sur-Yvette.
- ▶ **2009** → **2012** : Etudiant MPSI puis PSI lycée Masséna, Nice.

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 [HENAÛ-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

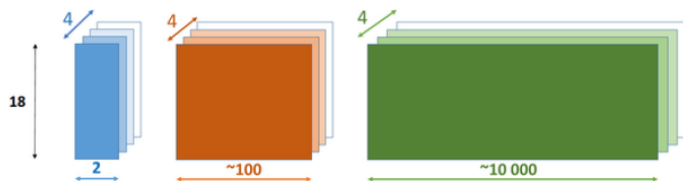
Résultat : [RECHTIEN et al., *Cell reports*, 2017]

→ 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

Motivation



Réponse
anticorps

Jours 28, 56, 84, 180

Fonctionnalité
cellulaire

Jours 0, 1, 3, 7

Expression
génétique

Jours 0, 1, 3, 7

Echantillons manquants : données génétiques

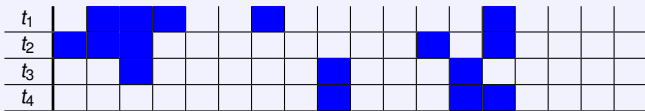


TABLE – Missing path du dataset Ebola rVSV-ZEBOV RNA-Seq où $t_1 = jour_0$, $t_2 = jour_1$, $t_3 = jour_3$ et $t_4 = jour_7$. Colonnes pour les participants.

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

Objectif

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

Modèle général

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.

→ D'autant plus vrai dans le cas de modèles parcimonieux.

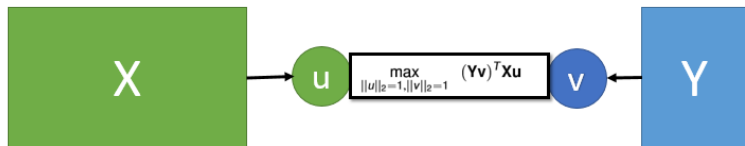
Côté utilisateur : difficile à optimiser

Contrainte majeure

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

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 .

⇒ 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 :

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

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
⇒ Sélection des variables et régularisation des données.

Des PLS parcimonieuses, package `mixOmics`

- ▶ [LÊ CAO et al., 2008], **2 paramètres/axe** :

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

- ▶ [CHUN et KELEŞ, 2010], $M = \mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$, **3 paramètres/axe** :

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

subj. to $w^T w = 1$,

Formulation

Vous avez sûrement vu la formulation suivante

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

où λ est le coefficient 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

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

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 , on écrit $K - 1$ modèles binaires indépendants tels que, $\forall k \in [1, \dots, K - 1]$, 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

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

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.

Méthodes non supervisées

Très proche de la PLS, elle maximise la covariance :

$$\max_u \|Xu\|_2^2,$$

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.

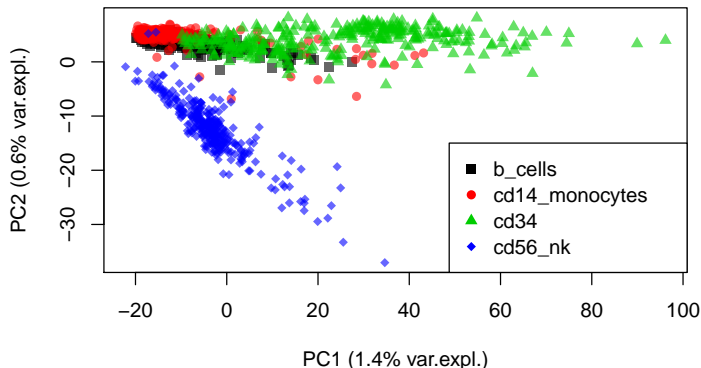


FIGURE – Représentation des deux premiers axes de l'ACP appliquée à un sous-échantillonnage des données de 10X [ZHENG et al., *Nature communications*, 2017]

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 rapidement 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.

Ces méthodes sont complexes d'un point de vue maths et il ne vous ai pas demandé de les comprendre, sauf si ça vous dit !

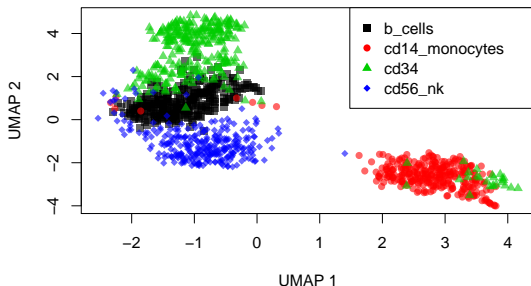


FIGURE – Représentation UMAP d'un sous-échantillonnage des données de 10X [ZHENG et al., *Nature communications*, 2017]

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

La Recherche Publique

La recherche publique

- ▶ Thésard (3 ans ou +) + PostDoc (0 à 25ans) + Poste fixe
- ▶ Ingénieur de recherche

Moi j'ai choisi le premier, qui passe par

- ▶ Bibliographie,
- ▶ Communication :
 - ▶ Articles
 - ▶ Conférences (posters, talks)

Le gros + : on travaille sur ce qui nous plaît. Pas seulement au tout début de la thèse, mais surtout après, de plus en plus ! L'autre gros + : plus haut diplôme universitaire internationalement reconnu.

Dans SISTM

On bosse à la fois avec INSERM : cliniciens, problématiques concrètes, données de première ligne (Ebola, VIH,...).

Mais aussi avec INRIA : développement de méthodes à la pointe, comme :

- ▶ Vaccination [PASIN et al., *arXiv preprint arXiv :1801.06227*, 2018]
- ▶ Clustering non paramétrique [HEJBLUM et al., *arXiv preprint arXiv :1702.04407*, 2017]
- ▶ Analyse de données RNA-Seq par groupes de gènes [AGNIEL et HEJBLUM, *Biostatistics*, 2017]

Projet Génétique

Travail demandé

Réfléchir au problème suivant

Discrimination des types cellulaires immunitaires grâce à leur expression génétique et interprétabilité.

Pour discuter ce problème on utilisera les données générées par la compagnie **10X** [ZHENG et al., *Nature communications*, 2017].

Ce travail doit jongler entre deux critères principaux qui sont :

- ▶ Efficacité de la méthode (minium en erreur de prédiction).
- ▶ Interprétabilité du modèle (sélection de variables).

Précision

Il n'y a pas de mauvaise conclusion à ce travail si votre chemin a été pavé avec discernement !

Contraintes pour le travail

- ▶ 3 groupes de 3 et 1 groupe de 4.
- ▶ Une partie introduction des données de single cell RNA-Seq en général et intérêt en immunologie.
- ▶ Une partie Analyse, divisée en 2 sous-parties :
 - ▶ *Non supervisée* : Analyses ACP et UMAP.
 - ▶ *Supervisée* : Analyses Lasso, sPLS, elastic-net (ou autres ???)

Dans chaque sous-partie, introduire succinctement la méthode en décrivant les paramètre à tuner et comparer : efficacité, temps de calcul, interprétabilité, reproductibilité,...

Contraintes pour le travail 2

Très important !

Il faut, dès le début, mettre de côté un échantillon de **10%**.

C'est l'échantillon de **test** !

Ce sous-échantillon doit représenter l'échantillon initial (on pense à la proportion de chaque classe...).

Les **90%** restants servent à l'estimation des hyperparamètres (coefficient de régression, paramètres de régularisation).

Les résultats, tels que :

- ▶ Visualisation des méthodes (supervisées et/ou non supervisées) se font sur l'échantillon de **test**.
- ▶ Erreurs de prédiction (supervisées) se font sur l'échantillon de **test**.

Contraintes pour le compte-rendu du travail

- ▶ Un compte-rendu sous forme d'un article en utilisant le template \LaTeX **JMLR**¹.
- ▶ Ne pas modifier les paramètres de mise en page (taille de la police, largeur des marges,...).
- ▶ Titre, Abstract, Introduction, Analyse, Résultats, Discussion, référence.
- ▶ Le résumé (Abstract) est forcément en anglais, pour le reste c'est au choix.
- ▶ **8 pages maximum** sans compter les appendices (où il y a un minimum de texte).
- ▶ Un fichier **Rmarkdown** avec l'ensemble des commandes dont les résultats sont commentés dans le compte rendu. Ce dernier doit être annoté clairement.

1. <https://fr.overleaf.com/latex/templates/template-for-journal-of-machine-learning-research-jmlr-with-jmlr2e-dot-sty/vjcpvhvztrjn>

Astuces pour le travail

- ▶ Sous-échantillonner les données dans un premier temps.
- ▶ Supprimer rapidement les gènes qui ne s'expriment pas **sur toute les classes**.
- ▶ Supprimer les gènes qui ne s'expriment pas suffisamment **sur toute les classes**, en fonction des capacités calculatoires et du temps restant.
- ▶ **logger** les données.
- ▶ NE PAS négliger les temps de calcul dans ce projet : ~**Big Data**.
- ▶ En discussion : En plus du reste, discuter de l'intérêt d'une méthode hiérarchique pour notre problème. Ou un autre type de méthode?!?

Soumission du dossier

Au plus tard le **mardi 22 janvier à 23h59**.

A envoyer par mail à **hadrien.lorenzo@u-bordeaux.fr** Dossier qui comprend :

- ▶ Un fichier **pdf** de l'article.
 - ▶ Un fichier **Rmarkdown** des commandes utilisées.
-
- ▶ Soutenance le **30 janvier 2019**,
 - ▶ 20 minutes de présentation,
 - ▶ 10 minutes de questions,
 - ▶ Présentation en **anglais**,
 - ▶ Chacun a le même temps de parole.

Les données

Import possible via le lien dropbox

https://www.dropbox.com/s/iwhzcuayvga0kqp/dataset_10X.zip?dl=0

ou via la clé USB.

10 types cellulaires différents.

Stages et thèses

Stage de fin d'étude et thèse, encadré par Cécile Proust-Lima :

- ▶ modélisation dynamique de phénomènes de santé,
- ▶ prédiction dynamique individualisée d'événements cliniques (diagnostic, rechute, décès)
- ▶ modèles de causalité à partir de données répétées dans le temps.

Un stage encadré par mes soins sur les problématiques liées aux données de single cell en immunologie, supervisé et non supervisé.

References



Denis AGNIEL et Boris P HEJBLUM. “Variance component score test for time-course gene set analysis of longitudinal RNA-seq data”. In : *Biostatistics* 18.4 (2017), p. 589–604.



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.



Boris P HEJBLUM et al. “Sequential dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data”. In : *arXiv preprint arXiv :1702.04407* (2017).



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).



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).



Chloé PASIN et al. “Controlling IL-7 injections in HIV-infected patients”. In : *arXiv*