

# Application d'une analyse de type *Machine Learning* en épidémiologie

Identification des dimensions de l'activité physique et du comportement sédentaire qui prédisent le risque de mortalité chez les personnes âgées, basée sur le développement d'un modèle sur les données d'accéléromètre de la cohorte Whitehall



# Sommaire

- Partie 1 : Présentation du contexte de l'étude, des objectifs de la méthode, des principes, et de l'interprétation
- **Partie 2 : Présentation du code R**

# Régression linéaire multiple

- Suppose que le nombre d'observation  $>$  nombre de variables explicatives
- Suppose que les variables ne soient pas trop corrélées entre elles

Contraintes peu satisfaites en pratique (notamment pour les variables dérivées des données d'accéléromètre)

# Partial Least Square (PLS)

PLS

- Alternative à la régression linéaire multiple en cas de forte colinéarité entre les variables ou lorsque le nombre de variables >> nombre d'observations (Tenenhaus 1999; Wold 1975; Wold, Martens, and Wold 1983)
- But :
  - construire des « composantes latentes » combinaisons linéaires de l'ensemble des  $X_j$
  - composantes latentes utilisées comme nouvelles variables explicatives dans une analyse de régression standard

Note : Différences avec l'analyse en composantes principales (ACP) :

- PLS utilise aussi  $Y$  pour construire les « composantes latentes ».
- PCA extrait des composantes qui expliquent la variance dans les  $X_j$  tandis que PLS extrait des composantes qui ont une forte covariance avec  $y$

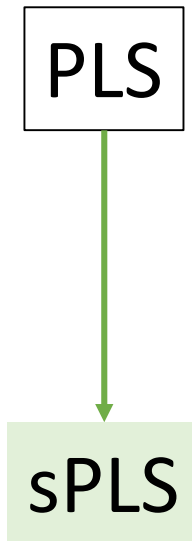
# Partial Least Square (PLS)

PLS

- Limites : en cas de très grand nombre de prédicteurs, notamment en cas de nombreuses variables non pertinentes : estimation de coefficients incohérents

→ Besoin de sélectionner les variables avant d'effectuer la régression PLS

# Sparse Partial Least Square (sPLS)



- Sélection des variables simultanément à la réduction dimensionnelle (Chun and Keleş 2010)
- Application d'une pénalité sur les poids des  $X_j$  dans les composantes latentes

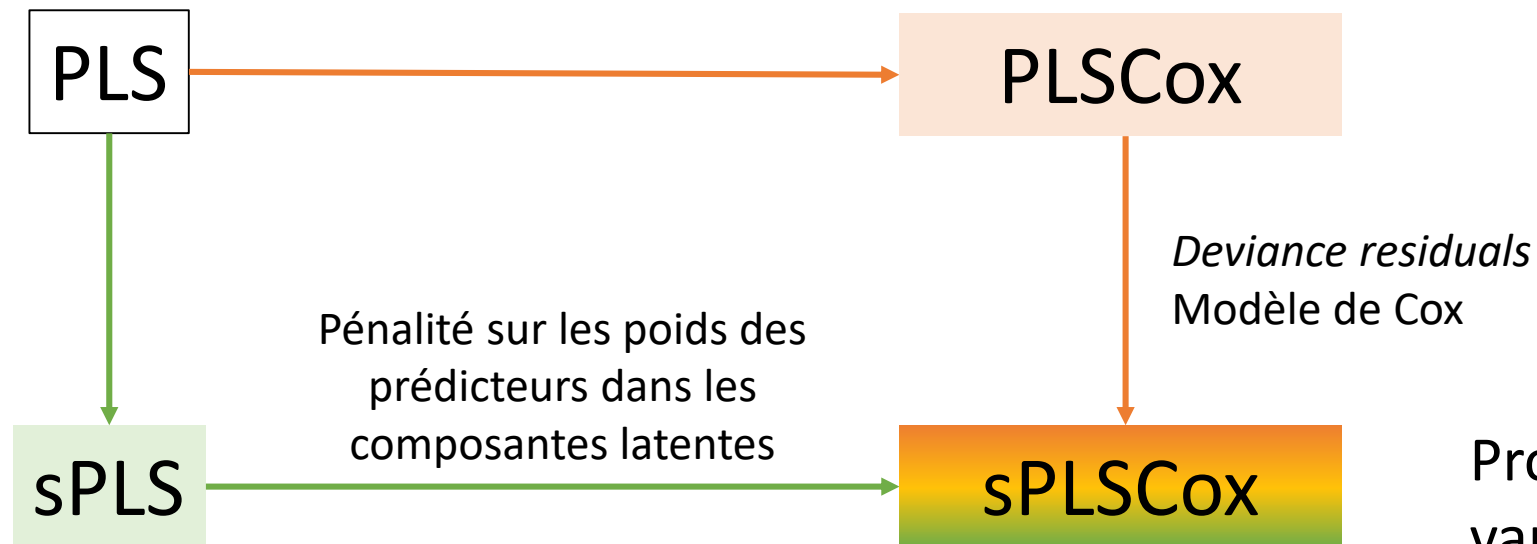
$$t = a_1 X_1 + a_2 X_2 + \dots$$

# PLS pour analyse de données de survie (PLSCox)



- PLS adaptée aux données censurées
- Utilisation du modèle de Cox pour dériver les composantes latentes (Bastien and Tenenhaus 2001)
- Méthode basée sur les « *deviance residuals* » ( = mesure l'excès de risque de survenue de l'évènement) (Bastien 2008)

# sPLS pour analyse de données de survie (sPLSCox)



Procédure de sélection de variables et de réduction dimensionnelle adaptée pour l'analyse de données de survie (Bastien et al. 2014; Bertrand and Maumy-Bertrand 2021)

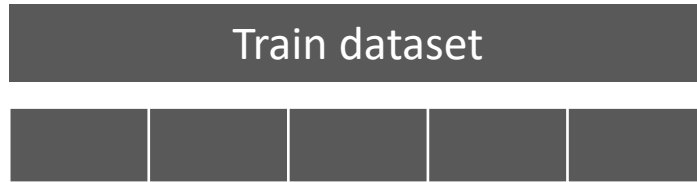
La valeur optimale du seuil de pénalité **eta** et le nombre de composantes latentes **ncomp** obtenue par validation croisée



# Procédure de validation croisée

Fixer une valeur de **eta** et une valeur de **ncomp**

Par exemple :  $\eta = 0.5$  et  $n_{comp} = 2$

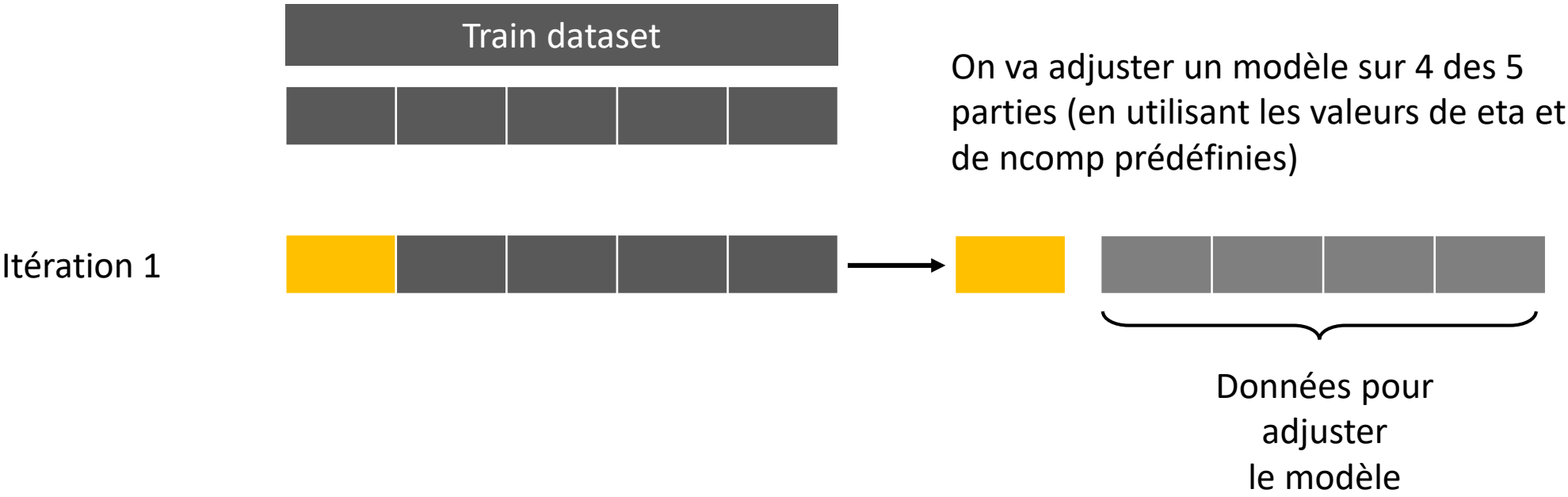


$K = 5$ :

diviser l'ensemble de données en 5 parties égales et indépendantes (“folds”)

# Procédure de validation croisée

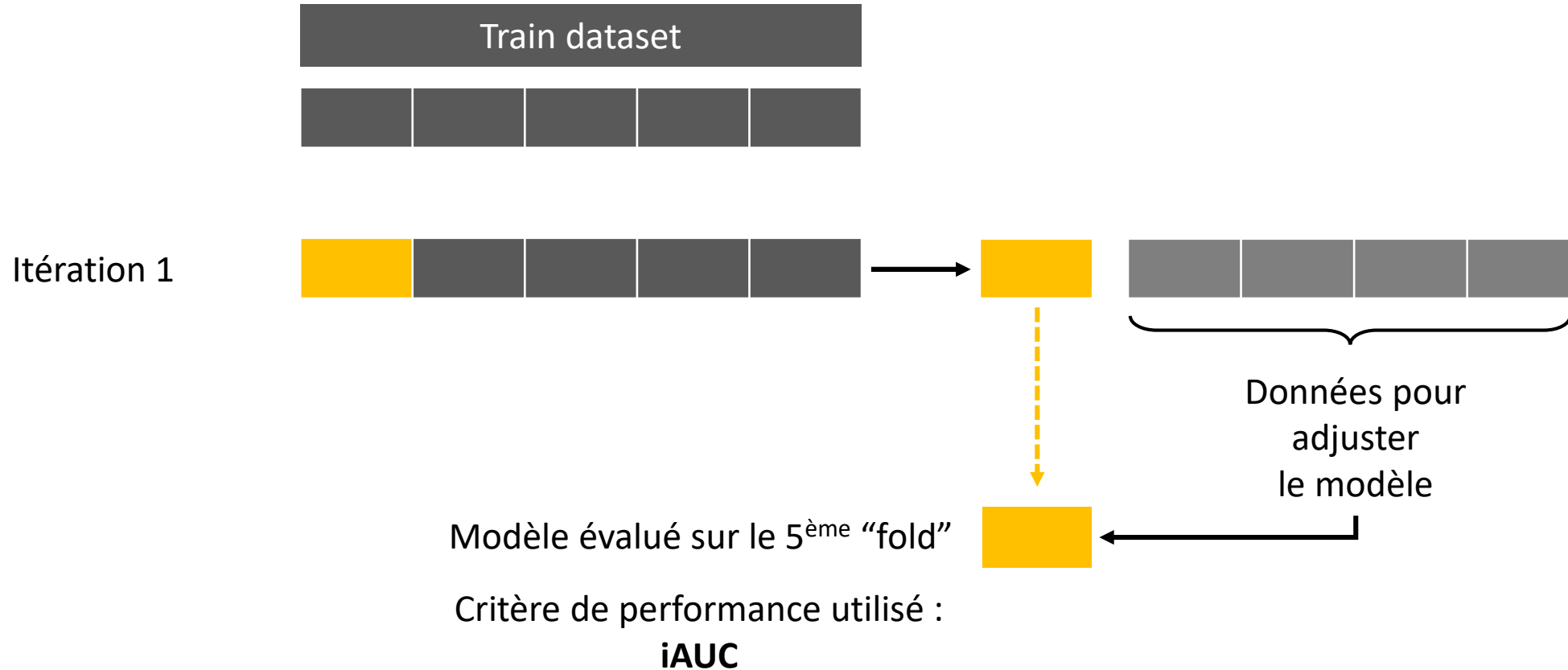
Fixer une valeur de **eta** et une valeur de **ncomp**  
Par exemple : eta = 0.5 et ncomp = 2



# Cross-validation procedure to select hyperparameters values

Fixer une valeur de **eta** et une valeur de **ncomp**

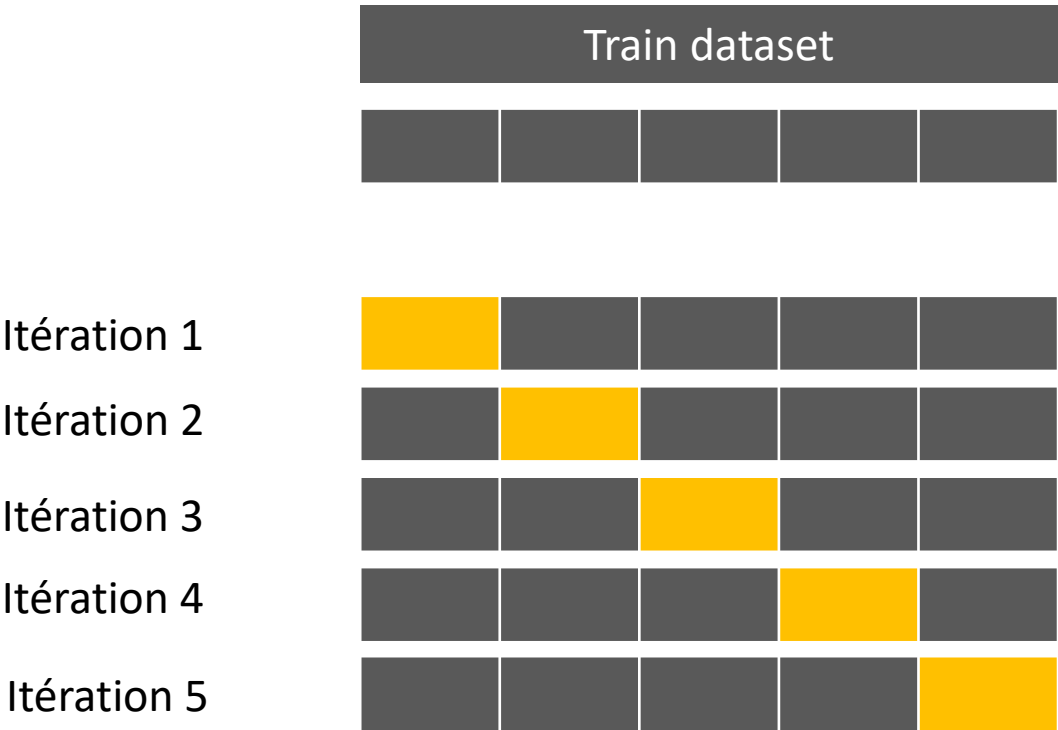
Par exemple :  $\eta = 0.5$  et  $n_{comp} = 2$



(Bertrand and Maumy-Bertrand 2021)

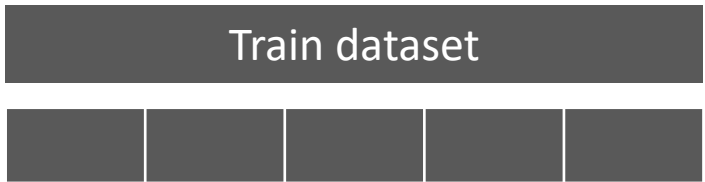
# Procédure de validation croisée

Fixer une valeur de **eta** et une valeur de **ncomp**  
Par exemple : eta = 0.5 et ncomp = 2



On répète la procédure à chaque itération, en changeant la part des données utilisées pour l'évaluation du modèle

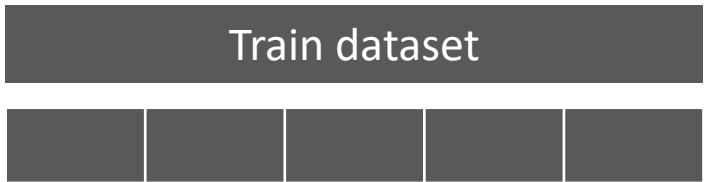
# Procédure de validation croisée



Répéter l'opération pour toutes les combinaisons de eta\* ncomp possibles

	eta=0.1	eta=0.15	...	eta=0.90	eta=0.95
Ncomp = 0					
Ncomp = 1					
Ncomp = 2					
Ncomp = 3					
Ncomp = 4					
Ncomp = 5					

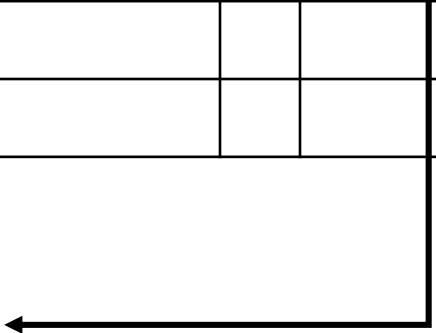
# Procédure de validation croisée



Répéter l'opération pour toutes les combinaisons de eta\* ncomp possibles

	eta=0.1	eta=0.15	...	eta=0.90	eta=0.95
Ncomp = 0					
Ncomp = 1					
Ncomp = 2					
Ncomp = 3					
Ncomp = 4					
Ncomp = 5					

Combinaison avec la performance la plus élevée  
Les valeurs sélectionnées seront utilisées pour ajuster le modèle sPLSCox



# Procédures

- Préparation des données
- Procédure de validation croisée pour choisir la valeur des hyper-paramètres du modèle sPLSCox
- Ajustement du modèle sPLSCox et calcul des composantes latentes
- Utilisation des composantes latentes dans un modèle de Cox

# Application dans R

- Utilisation du package *plsRCox* pour calculer les composantes latentes et ajuster le modèle



# Plus de details sur plsRcox

## plsRcox, Cox-Models in a High Dimensional Setting in R

Frédéric Bertrand and Myriam Maumy-Bertrand

The goal of plsRcox is provide Cox models in a high dimensional setting in R.

plsRcox implements partial least squares Regression and various regular, sparse or kernel, techniques for fitting Cox models in high dimensional settings <https://doi.org/10.1093/bioinformatics/btu660>, Bastien, P., Bertrand, F., Meyer N., Maumy-Bertrand, M. (2015), Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data, *Bioinformatics*, 31(3):397-404. Cross validation criteria were studied in [arXiv:1810.02962](https://arxiv.org/abs/1810.02962), Bertrand, F., Bastien, Ph. and Maumy-Bertrand, M. (2018), Cross validating extensions of kernel, sparse or regular partial least squares regression models to censored data.

The package was presented at the [User2014!](http://user2014.r-project.org/abstracts/posters/177_Bertrand.pdf) conference. Frédéric Bertrand, Philippe Bastien, Nicolas Meyer and Myriam Bertrand (2014). "plsRcox, Cox-Models in a high dimensional setting in R", *book of abstracts*, User2014!, Los Angeles, page 177, [http://user2014.r-project.org/abstracts/posters/177\\_Bertrand.pdf](http://user2014.r-project.org/abstracts/posters/177_Bertrand.pdf).

The plsRcox package contains an original allelotyping dataset from "Allelotyping identification of genomic alterations in rectal chromosomally unstable tumors without preoperative treatment", Benoît Romain, Agnès Neuville, Nicolas Meyer, Cécile Brigand, Serge Rohr, Anne Schneider, Marie-Pierre Gaub and Dominique Guenot (2010), *BMC Cancer*, 10:561, <https://doi.org/10.1186/1471-2407-10-561>.

Support for parallel computation and GPU is being developped.

The package provides several modelling techniques related to penalized Cox models or extensions of partial least squares to Cox models. The first two were new algorithms.



<https://fbertran.github.io/plsRcox/>

### Links

[View on CRAN](#)

[Browse source code](#)

[Report a bug](#)

### License


[GPL-3](#)

### Citation

[Citing plsRcox](#)

### Developers

Frederic Bertrand

Maintainer, author 

Myriam Maumy-Bertrand

Author 

### Dev status

lifecycle 

repo status 


[R-CMD-check](#)

 codecov 

CRAN 

downloads 

 Stars 

DOI 

# Plus de details sur plsRcox

- **coxplsDR** and **cv.coxplsDR** (Philippe Bastien, Frederic Bertrand, Nicolas Meyer, and Myriam Maumy-Bertrand (2015), "Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data", *Bioinformatics*, **31**(3):397-404, <https://doi.org/10.1093/bioinformatics/btu660>),
- **coxDKplsDR** and **cv.coxDKplsDR** (Philippe Bastien, Frederic Bertrand, Nicolas Meyer, and Myriam Maumy-Bertrand (2015), "Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data", *Bioinformatics*, **31**(3):397-404, <https://doi.org/10.1093/bioinformatics/btu660>),
- **coxDKplsDR** and **cv.coxDKplsDR** (Philippe Bastien (2008), "Deviance residuals based PLS regression for censored data in high dimensional setting", *Chemometrics and Intelligent Laboratory Systems*, **91**:78–86, <https://doi.org/10.1016/j.chemolab.2007.09.009>),
- **coxpls** and **cv.coxpls** (Nguyen, D.V., Rocke, D.M. (2002), "Partial least squares proportional hazard regression for application to DNA microarray survival data", *Bioinformatics*, **18**(12):1625–1632),
- **coxplsDR** and **cv.coxplsDR** (Philippe Bastien (2008), "Deviance residuals based PLS regression for censored data in high dimensional setting", *Chemometrics and Intelligent Laboratory Systems*, **91**:78–86, <https://doi.org/10.1016/j.chemolab.2007.09.009>),
- **DKplsRcox**,
- **larsDR** and **cv.larsDR** (Segal, M.R. (2006), "Microarray Gene Expression Data with Linked Survival Phenotypes: Diffuse large-B-Cell Lymphoma Revisited", *Biostatistics*, **7**:268-285, <https://doi.org/10.1093/biostatistics/kxj006>),
- **plsRcox** and **cv.plsRcox** (Philippe Bastien, Vincenzo Esposito Vinzi, and Michel Tenenhaus (2005), "PLS generalised linear regression", *Computational Statistics & Data Analysis*, **48**(1):17–46, <https://doi.org/10.1016/j.csda.2004.02.005>),
- **autoplsRcox** and **cv.autoplsRcox** (Philippe Bastien, Vincenzo Esposito Vinzi, and Michel Tenenhaus (2005), "PLS generalised linear regression", *Computational Statistics & Data Analysis*, **48**(1):17–46, <https://doi.org/10.1016/j.csda.2004.02.005>),

# Giuthub du projet

Plus d'informations sur l'ensemble des analyses effectuées pour l'article

Packages utilisés

Workflow

Scripts

## Identification of physical activity and sedentary behaviour dimensions that predict mortality risk in older adults: development of a machine learning model in the Whitehall II accelerometer sub-study and external validation in the CoLaus study

This repository contains scripts supporting a project aiming to identify accelerometer-derived dimensions of movement behaviours that predict mortality risk in older populations.

All analyses were undertaken using R version 4.1.2 (<http://www.r-project.org>), analyses required the specific packages:

- *GGIR* for accelerometer data processing (version 2.3-3, <https://cran.r-project.org/web/packages/GGIR/vignettes/GGIR.html>)
- *plsRcox* for sparse Partial Least Square regression (version 1.7.6, <https://cran.r-project.org/web/packages/plsRcox/index.html>)
- *Epi* (version 2.47, <https://cran.r-project.org/web/packages/Epi/index.html>), *aod* (version 1.3.2, <https://cran.r-project.org/web/packages/aod/index.html>), and *rms* (version 6.3-0, <https://cran.r-project.org/web/packages/rms/index.html>) for examine the association bewtween movement behaviour compopsite score and mortality
- *timeROC* for time-Dependent ROC Curve(version 0.4, <https://cran.r-project.org/web/packages/timeROC/index.html>)
- *boot* for performing bootstrap analyses (version 1.3-28, <https://cran.r-project.org/web/packages/boot/index.html>)
- *coxed* for computing 95% bias-corrected and accelerated confidence intervals for bootstrapped estimates (version 0.3.3, <https://cran.r-project.org/web/packages/coxed/index.html>)

Lien : <https://github.com/MathildeChen/PA-SB-dimensions-mortality-Whitehall>

DOI: <https://zenodo.org/badge/latestdoi/531462028>

# Références

- Bastien, Philippe. 2008. 'Deviance residuals based PLS regression for censored data in high dimensional setting', *Chemometrics and Intelligent Laboratory Systems*, 91: 78-86.
- Bastien, Philippe, Frédéric Bertrand, Nicolas Meyer, and Myriam Maumy-Bertrand. 2014. 'Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data', *Bioinformatics*, 31: 397-404.
- Bastien, Philippe, and M. Tenenhaus. 2001. "PLS generalized linear regression. Application to the analysis of life time data." In *Proceedings of the PLS'01 International Symposium*, Anacapri (Italy), 131-40. X, France.
- Bertrand, Frédéric, and Myriam Maumy-Bertrand 2021. 'Fitting and Cross-Validating Cox Models to Censored Big Data With Missing Values Using Extensions of Partial Least Squares Regression Models', *Frontiers in Big Data*, 4.
- Chun, Hyonho, and Sündüz Keleş. 2010. 'Sparse partial least squares regression for simultaneous dimension reduction and variable selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 3-25.
- Tenenhaus, M. 1999. 'L'approche PLS', *Revue de Statistique Appliquée*, 47: 5-40.
- Wold, Herman. 1975. 'Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach', *Journal of Applied Probability*, 12: 117-42.
- Wold, S., H. Martens, and H. Wold. 1983. 'The multivariate calibration problem in chemistry solved by the PLS method.' in, *Matrix Pencils*.