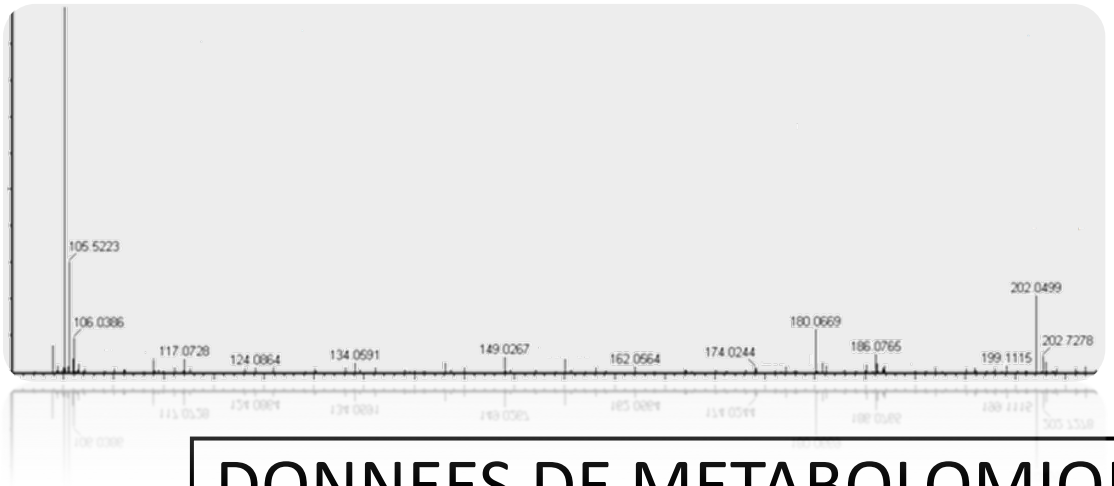
CATI 3G

eMPrEInTE

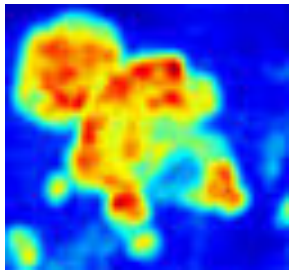


DONNEES DE METABOLOMIQUE : PRETRAITEMENT DES SPECTRES BRUTS ET ANALYSES STATISTIQUES

M. Pétéra, J.F. Martin & M. Tremblay-Franco

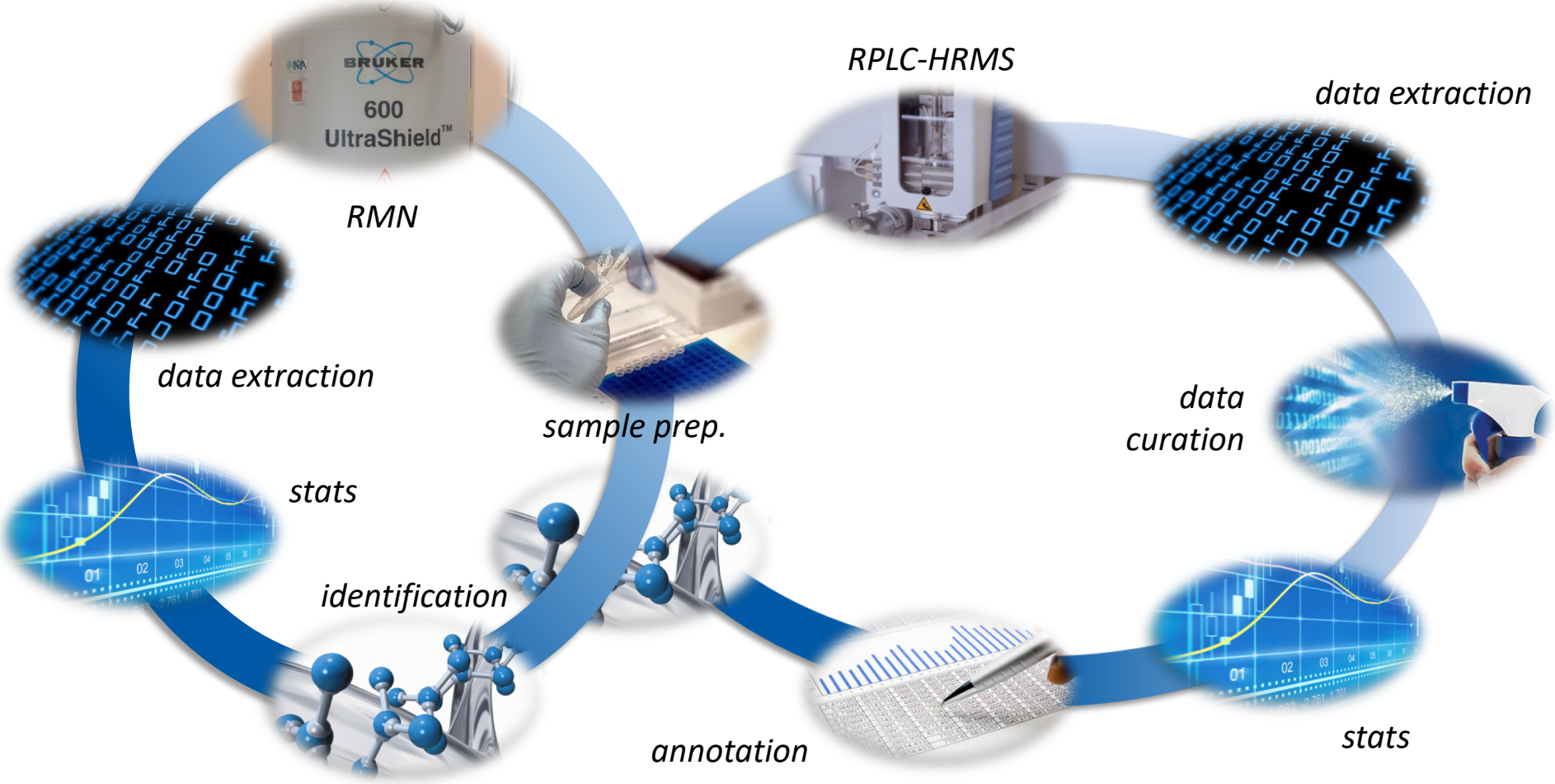
Workshop Inter CATIs, Paris, 17/10/2019

INRA SCIENCE & IMPACT



Analyses globales

analyse sans apriori, avec quantification relative de composés connus & inconnus



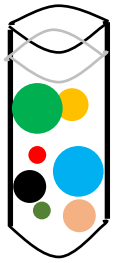
analyse de cellules, tissus, fluides par RMN

et/ou par RPLC-HRMS

Spectrométrie de masse haute résolution (HRMS)

- HRMS couplée à une chromatographie liquide (LC) ou gazeuse (GC)
- Technique analytique utilisée pour le dosage ciblé de substances (produit dopant, pesticides, etc…)
- En métabolomique, elle est utilisée dans un mode dit « fullscan » qui permet de détecter un ensemble de molécules sur une gamme de masse atomique définie (70..1000 daltons)

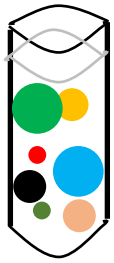
LC-HRMS Principe



Echantillon : fluide biologique (urine, plasma, foie, eau fécale, culture cellulaire…)

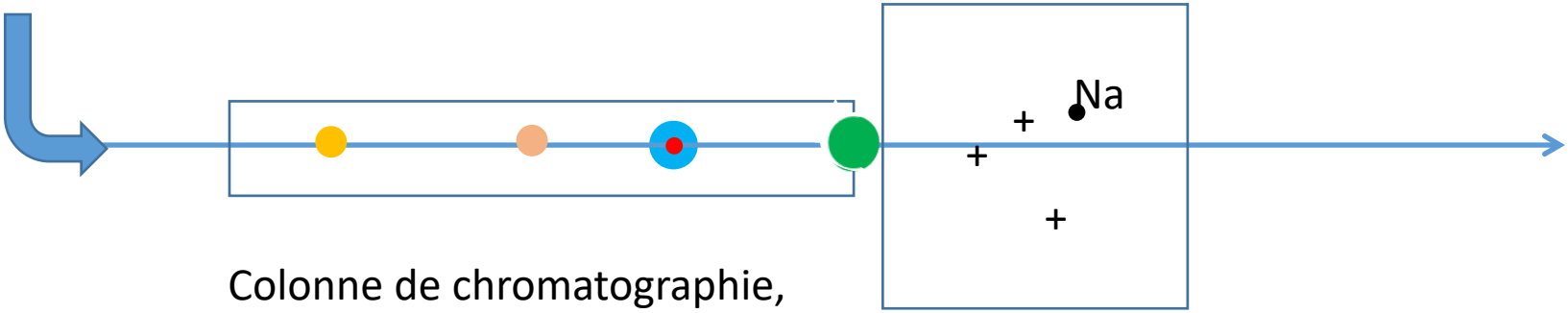
Chromatographie liquide ou gazeuse, séparation temporelle des molécules => temps de rétention

LC-HRMS Principe



Echantillon : fluide biologique (urine, plasma, foie, eau fécale, culture cellulaire…)

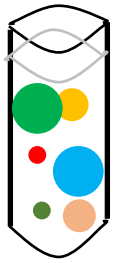
Source d'Ionisation



Colonne de chromatographie, séparation temporelle des molécules => temps de rétention

Ions moléculaires
+ isotope
+ fragments
+ adduits (Na)

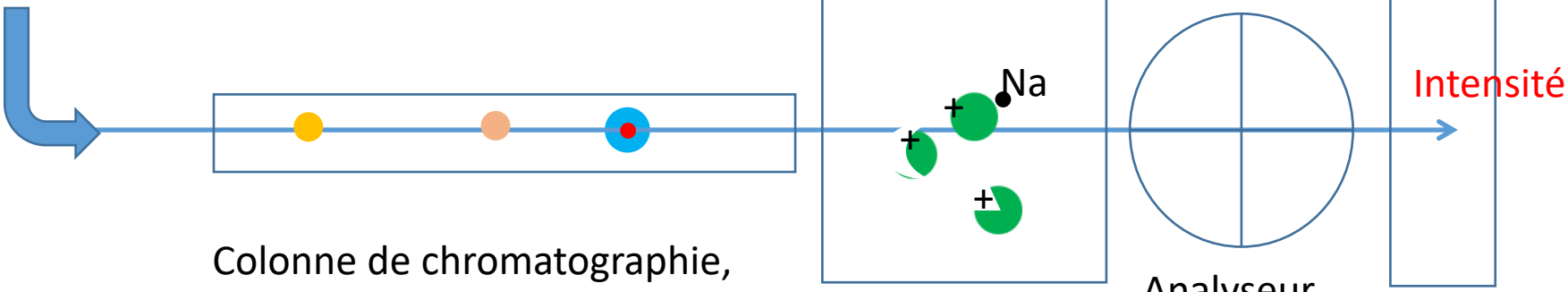
LC-HRMS Principe



Echantillon : fluide biologique (urine, plasma, foie, eau fécale, culture cellulaire…)

Détecteur: courant ionique => courant électrique

Source d'Ionisation

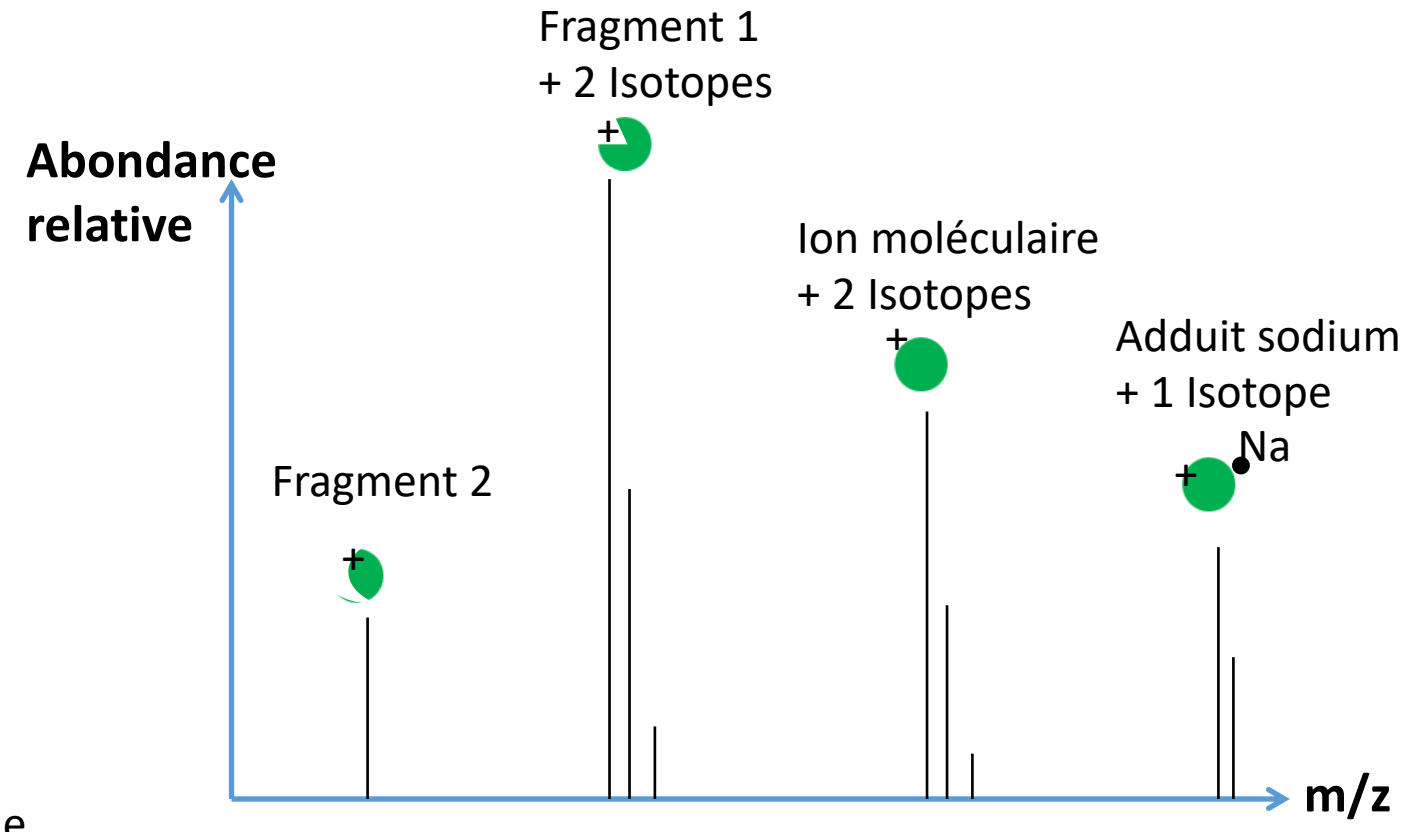


Colonne de chromatographie, séparation temporelle des molécules => temps de rétention

Analyseur, TOF, calcul de masse atomique

Transformation des molécules dans leur état naturel, en ions à l'état gazeux, puis tri des ions suivant le rapport masse / charge

Spectre de masse



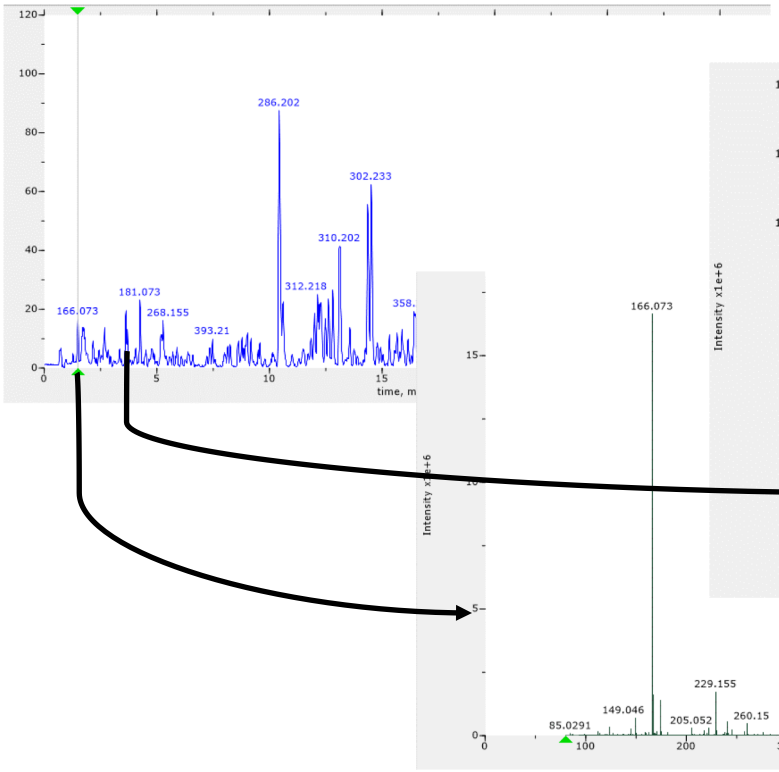
Exemple:
1 molécule
donne 9 ions

m : unité de masse atomique (uma) ou dalton (Da) :
= 1/12 masse du carbone 12 ($1.66054.10^{-27}$ kg)

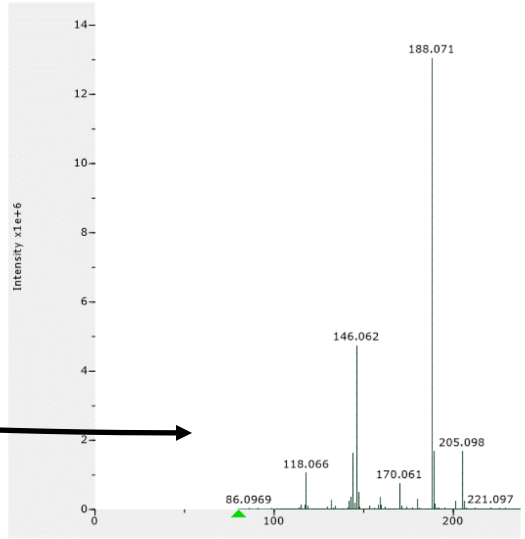
z : nombre de charge portées par l'ion
= n fois la charge de l'électron = n x $1.622177.10^{-19}$ C

Extraction des données

Chromatogramme



Spectres MS



Logiciels:

- xcms
- MSDial
- Progenesis

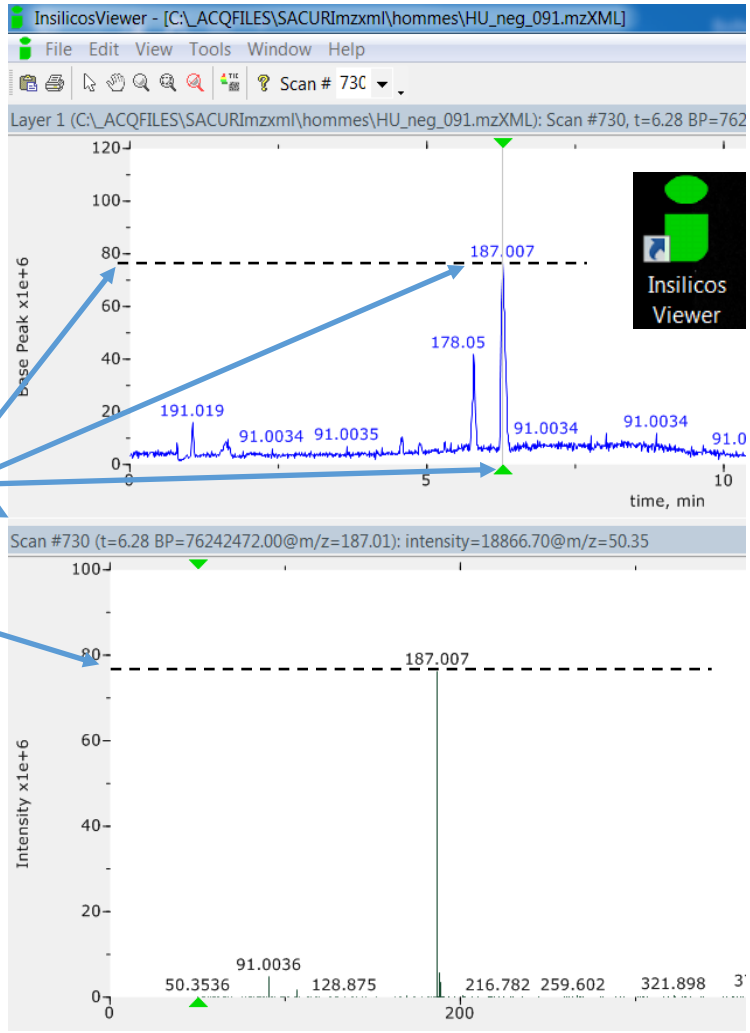
Matrice de données

ions	RetTime	Mass	T38CT05N	T38CT05N
114.067T1.5	1.5	114.067	9206.7362	4014.3652
137.072T1.5	1.5	137.072	2083.1412	3437.6839
212.853T1.5	1.5	212.853	0	2095.7974
196.88T1.5	1.5	196.88	0	1531.1653
162.114T1.5	1.5	162.114	1985.5564	267.3418
201.937T1.5	1.5	201.937	1934.2631	2295.2461
141.067T1.5	1.5	141.067	1656.8438	1182.8188
229.119T1.5	1.5	229.119	676.5843	688.6075
152.026T1.5	1.5	152.026	1002.5317	372.6582
407.186T6.1	6.1	407.186	183.2912	588.2105
359.059T6.1	6.1	359.059	36.4557	0
211.11T6.1	6.1	211.11	117.1308	175.5949

mzXML raw file

mzXML in a text editor
1 scan

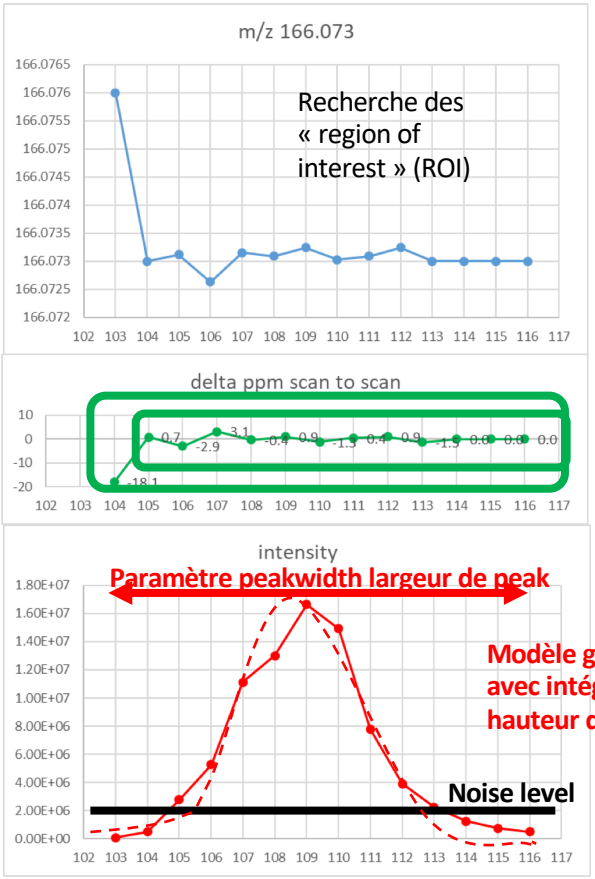
```
<scan num="730"
      scanEvent="1"
      scanType="FULL"
      centroided="1"
      msLevel="1"
      peaksCount="390"
      polarity="-"
      retentionTime="PT376.906S"
      lowMz="50.35359954834"
      highMz="989.62841796875"
      basePeakMz="187.00694279023"
      basePeakIntensity="7.6242472e07"
      totIonCurrent="1.10045032e08"
      msInstrumentID="IC1">
  <peaks compressedLen="0"
         precision="32"
         byteOrder="network"
         pairOrder="m/z-int">QklqFkaTZWhCUlS
</scan>
```



Extraction des données : package R xcms

Scan par scan :

```
fichier mzXML
  <scan num="263"
      scanType="Full"
      centroided="1"
      msLevel="1"
      peaksCount="10453"
      polarity="+"
      retentionTime="PT215.853S"
      basePeakMz="180.089111"
      basePeakIntensity="1.2813312e07"
      totIonCurrent="7.1073584e07"
      msInstrumentID="1">
    <peaks compressionType="none"
        compressedLen="0"
        precision="64"
        byteOrder="network"
  </scan>
```



Paramètre ppm

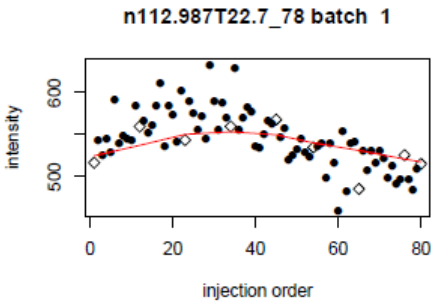
Algorithme centwave

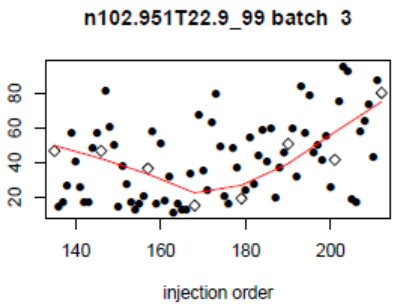
1 ion in 1 sample :
- m/z
- Retention time
- Intensity

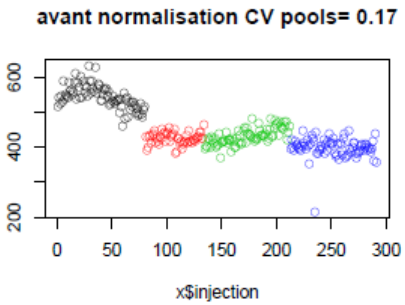
Alignement de tous les ions dans tous les échantillons

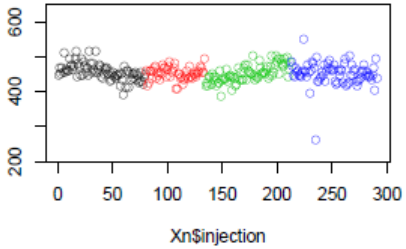
Normalisation pour la correction de dérives analytiques

- Permet de corriger des différences de signal entre séries d'échantillons (correction inter batch). Permet également de corriger l'évolution du signal en cours de batch (correction intra batch)
- Régression linéaire ou loess utilisant des échantillons quality control (QC) qui sont des pools de tous les échantillons de l'étude. L'évolution du signal est modélisée dans ces pools et le modèle obtenu permet de corriger l'ensemble des échantillons.









Matrice de données obtenues

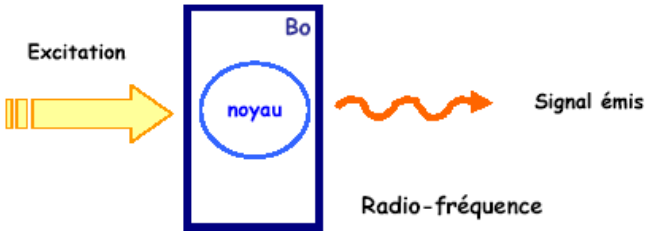
ions	BL1	BL10	BL11	BL12	BL13	BL14	BL15	BL16	BL17	BL18	BL19
n100.0398T5.6	18368.4688	14716.9531	15623.1484	15100.4609	16108.1641	16508.0781	13587.4141	13625.125	12681.1719	12607.5703	12180.07
n100.0762T1.1	891.882324	714.700195	751.878418	568	881.003418	697.046387	1008.396	991.26709	594	980.572266	6
n100.0762T1.6	65		.225586	599.729981	748.222168	758.146484	874.744629	927.121582	1003.88672	844.188965	7
n100.0762T3			.563965	730.700195	1148.12793	818.666504	1226	814	1130.26563	1038.34473	1251.215
n100.0762T3.8	12		.663574	1506.19336	1073.39063	866.657715	989.710449	1213.94531	1224.50684	1066.33594	1279.083
n100.0762T4.2	1		2.42969	891.325684	1122.11133	824.326172	1129.18848	1320	1050	1045	1203.568
n100.1125T2.4	96		6.90039	956	1008.12647	1257.0752	1072.99512	936.111328	1554	808.248535	1033.682
n100.9745T0.7	48		923.625	92608.8125	113604.625	96050.75	103613.125	103659.813	129088.125	121262.688	126879.6
n101.0093T0.7	48091.5313	104026.25	105923.625	92608.8125	113604.625	96050.75	103613.125	103659.813	129088.125	121262.688	126879.6
n101.0603T1.3	319.084229	464.531006	294.249023	304.04419	895.47168	392.012207	302	316.334961	440	426.32251	258.9064
n101.0714T11.5	1126.57617	951.02832	1186.11426	771	959.723633	1399.82715	917.661133	1066.15625	1725.875	634.616211	1197.17
n101.0714T5.6	535.116211	504.375732	643.516602	814.908691	523.386719	641.021973	592.533691	593.582031	905.007324	494.810547	756.3242
n101.0714T6.1	480.384277	652.623535	736.609375	755.845215	495.092529	579.353027	862.8320				
n101.0714T7.4	461	651.316895	526.250977	470.678467	543.924805	634.773926					
n101.0714T8.6	820.276856	932.563477	871.17627	1144.12793	893	716.029785	589.				
n101.0715T5.2	554.935059	594.411621	598.15625	776	745.694824	492.197754	604.135				
n102.0105T0.7	6553.55078	15509.0859	14393.1719	13492.125	15598.5156	16622.6563	14204.35				
n102.0554T7	1398	1120.2959	1162.17481	1267.37988	910.255371	787.980957	9				
n102.0555T6.7	1332.82617	1248.91016	948.468262	989	973.114746	926.985352	759.8085				
n102.0556T5.2	1400.45215	1109.02832	798.820313	1703.73535	1299.26953	1074.71582	1123.638				
n102.9687T0.7	29500.3594	51520.5	52947.7813	47987.5313	46368.6875	42674.6563	46539.				
n103.007T0.7	29500.3594	51520.5	52947.7813	47987.5313	46368.6875	42674.6563	46539.75	45159.4375	60585.4375	55670.25	53592.65
n103.0395T4.3	17438.4531	14693	16983.0469	13181.2734	16778.2031	13946.2422	13645.1172	12721.9766	13783.4453	10322.2344	11222.85
n103.0396T5.4	20193.9844	15949.8125	16633.0938	14389.2422	17040.4844	15182.8984	15089.5313	13884.0313	14364.3203	12712.25	12808.92
n103.0396T6.4	21768.9375	17229.875	18440.9688	16704.5781	18438.6563	15852.2734	15649.5156	16800.3281	15971.8828	13614.9688	13092.64
n103.0396T7.3	22385.1563	19841.9063	18431.9688	17168.125	19544.9844	18047.125	17335.3125	17313.3594	16550.1719	16617.6406	13888.18
n103.0396T8.1	25094.6563	19596.2031	19874.0469	18327.0781	21487.7031	17830.5781	17887.1563	17883.6563	17462.375	16352.5078	16086.76
n103.0548T2.6	493.284668	760.148438	705.457031	724.962402	932.781738	739.549805	695.007813	758.22168	798.960449	702.297363	431.2011
n103.0548T9.3	1028.6416	983.861328	1055.44434	1320.87598	1492.31738	1176.74414	1046	1258.47266	2016.23438	1254.88965	1158.372
n104.0501T3.5	6635.20703	6759.23047	7592.86328	6230.50781	7245.86719	6986.80078	5728.18359	6143.79688	3974.03711	6491.74219	7545.398
n104.0581T9.3	1503.80469	974.754883	688.983398	503.926514	774.052734	596.734375	797	690.936035	696	692.918457	764.2587
n104.0711T1.1	1025.06836	430.439453	255	446	321.367676	441.646973	242	229	347	408	2

ID ions : m/z & Temps de rétention

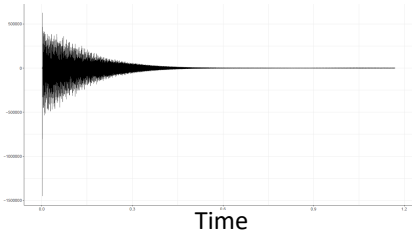
Intensité = abondance relative
Proportionnelle à la concentration mais aussi à la capacité d'ionisation de la molécule

Analyses globales : RMN du proton

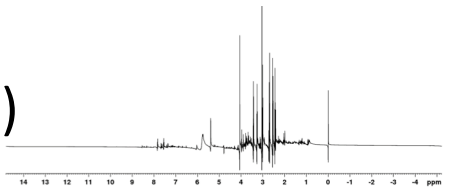
- Principe



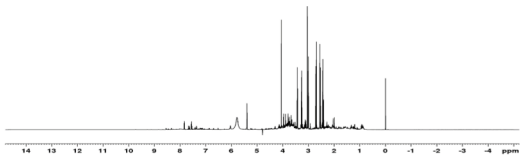
- Données brutes
  - Free Induction Decay : fonction du temps



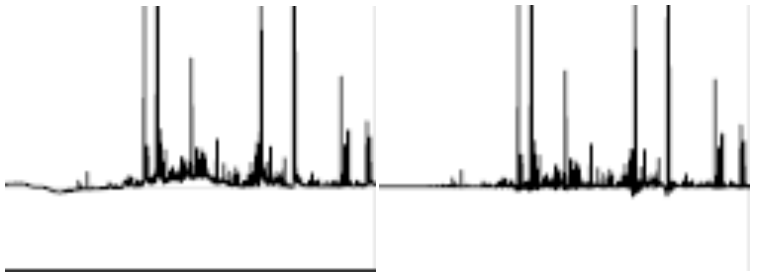
- Preprocessing
  - Fourier Transformation : fonction de la fréquence (Hz)
    - Axe des abscisses = Déplacement chimique (ppm)



  - Phase correction



  - Shift referencing
  - Baseline correction



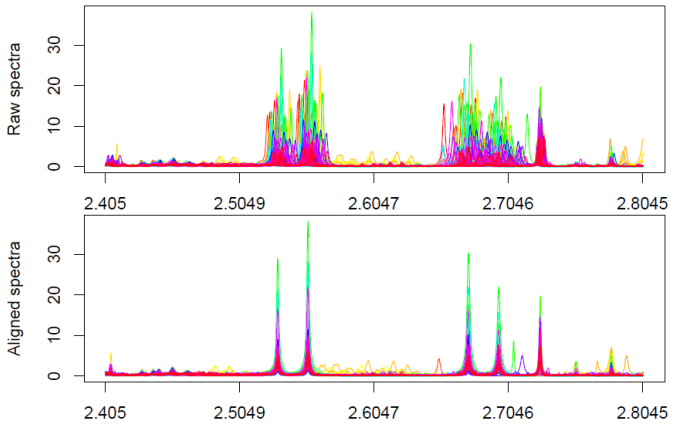
Analyses globales : RMN du proton

- Preprocessing
  - Matrice de données: couples de points

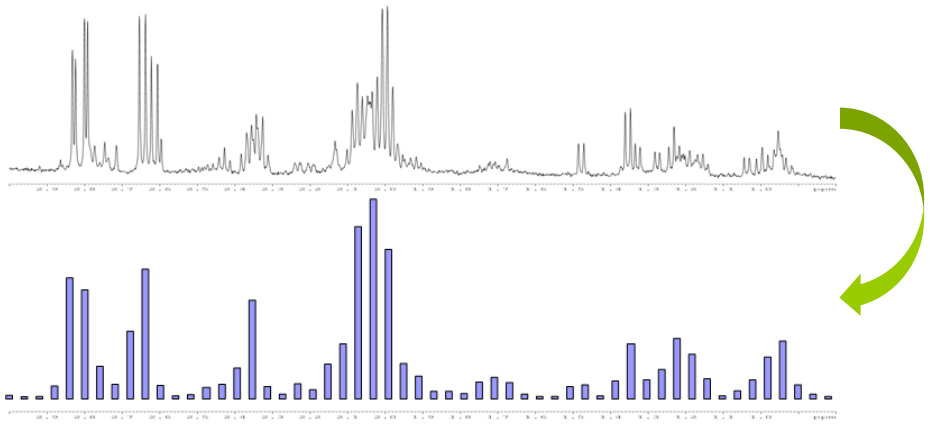
Ppm	ADG10003u_007	ADG10003u_008	ADG10003u_009	ADG10003u_010	ADG10003u_015	ADG10003u_016	ADG10003u_017	ADG10003u_021
14.7758802295849	794167.470036	1037744.972859	1139703.216949	1051843.510276	504964.179023	943052.263662	955099.994018	180498.765123
14.7752695046023	790024.068359	1032771.056669	1135901.875688	1046934.547379	503356.501625	939013.251365	951218.807188	179791.327966
14.7746587796196	785880.269598	1027796.621607	1132099.964577	1042025.05856	501748.571746	934973.767542	947337.142808	179083.80056
14.774048054637	781735.281657	1022820.632414	1128296.34581	1037113.994431	500139.885224	930932.87116	943454.047718	178376.002759
14.7734373296544	777587.919501	1017841.539934	1124489.315538	1032199.784589	498529.687026	926889.153699	939568.095091	177667.664877
14.7728266046718	773436.607227	1012857.283601	1120676.605758	1027280.340076	496916.972045	922840.741174	935677.386372	176958.428038
14.7722158796892	0	1007865.293927	1116855.386225	1022353.055834	495300.485914	918785.296154	931779.553221	176247.844535
14.7716051547065	0	1002862.494995	1113022.266355	1017414.813164	493678.725803	914720.019787	927871.759461	175535.378178
14.7709944297239	0	0	1109173.297138	0	492049.941236	910641.653831	923950.703025	0
14.7703837047413	0	0	1105303.973051	0	490412.134894	906546.48268	920012.617908	0
14.7697729797587	0	0	1101409.233984	0	488763.063435	0	916053.276132	0
14.769162254776	0	0	1097483.467177	0	487100.23831	0	0	0
14.7685515297934	0	0	1093520.509164	0	0	0	0	4493.864828
14.7679408048108	23452.486544	0	0	4785.954219	0	0	0	9847.076937
14.7673300798282	64991.222079	7919.34174	0	44818.589571	0	0	0	0
14.7667193548455	80352.629626	11009.853226	0	0	0	0	0	0
14.7661086298629	72527.324758	21361.643782	0	33933.50853	0	1971.623714	0	19578.922466
...	...	...	...	...	...	...	...	...
-5.203987576932	80024.469638	67212.127595	80358.031801	66336.451773	38554.798821	49938.974409	71707.866503	27276.12461
-5.20459830191462	46115.373335	64874.562258	83161.770869	101108.309625	14584.224021	82422.103969	15368.935974	13823.49713
-5.20520902689725	84821.619971	74275.160493	79226.294924	102756.846739	32775.393105	64278.990743	62281.333065	19098.386372
-5.20581975187987	12324.662047	84261.827582	64331.644833	52061.900808	44083.05321	100339.030476	74143.665911	10773.811414
-5.20643047686249	16016.847179	83083.482042	65583.890885	67010.334107	54453.586112	56370.14245	74172.355161	30490.038609
-5.20704120184512	46492.661082	83249.048368	87785.832189	101119.55271	37918.403946	104367.890085	79655.042232	22586.689804
-5.20765192682774	70646.316775	71137.190317	62622.689417	51768.88919	42283.722101	43280.642205	29505.003976	15190.091449
-5.20826265181036	79806.604066	63752.367913	94510.279475	87468.787976	41665.076768	120121.84998	59291.407603	15794.08022
-5.20887337679299	80645.568069	70183.534085	92758.788239	64604.586853	33233.002235	102261.980198	32505.191746	28258.123875
-5.20948410177561	33426.92407	74815.802359	86687.046779	89693.051589	27227.389563	95533.327921	20562.385331	17431.506286
-5.21009482675823	57162.165836	95813.711784	88748.950447	85428.149123	49935.702769	121885.97558	71581.943749	9824.495627
-5.21070555174086	90676.320408	90615.836285	78668.388006	138974.010535	42122.805034	54895.939413	68028.447631	13823.534413
-5.21131627672348	108017.4268	80525.134129	93681.414002	85465.526656	17878.676417	82702.173311	84373.711305	31090.379112
-5.2119270017061	39930.651957	82312.7961	92594.530223	112321.156858	33968.604423	69864.246856	69580.249541	20339.207857
-5.21253772668873	65594.179686	77522.198151	101890.026413	64213.002541	6186.619567	104314.764649	47497.79318	0
-5.21314845167135	69080.790665	99541.508254	80687.127045	59185.311284	18031.168822	112521.450202	98233.852776	11740.895167
-5.21375917665397	55241.416586	82652.163378	107742.066611	86305.697345	52621.564342	73085.850125	71992.920517	28941.598739

Analyses globales : RMN du proton

- Processing
  - Alignement



  - Bucketing / Intégration



  - Normalisation

- Matrice de données finales = intensité relative mesurée dans chaque bucket

Analyses globales : analyses statistiques multivariées

Common problems of omic data

Number of samples << number of variables

Noise in data

Often strong biological or technical collinearity in data

- Construction de « nouvelles » variables : combinaisons linéaires des variables originales (buckets RMN, profils lipidiques, …) :

$$t_h = \sum_{j=1}^{p} X_j a_{jh}$$

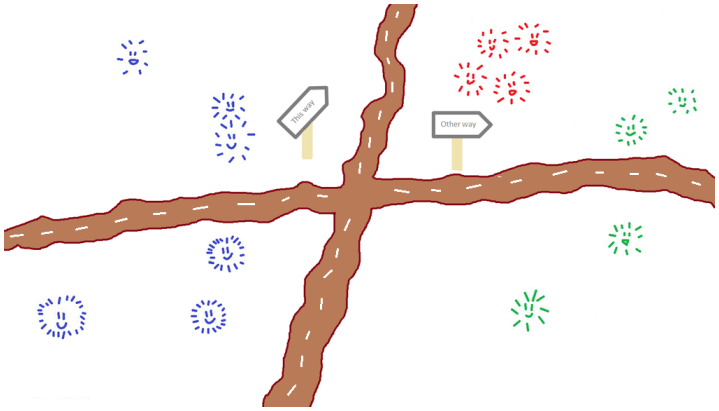
a_{jh} : poids (loading) de la variable j pour la composante h

- Buts méthodes multivariées
  - Réduction dimensionnalité : choix d'un « petit » nombre de combinaisons linéaires
  - « Décorrélation » variables : contrainte d'orthogonalité pour construire les combinaisons linéaires

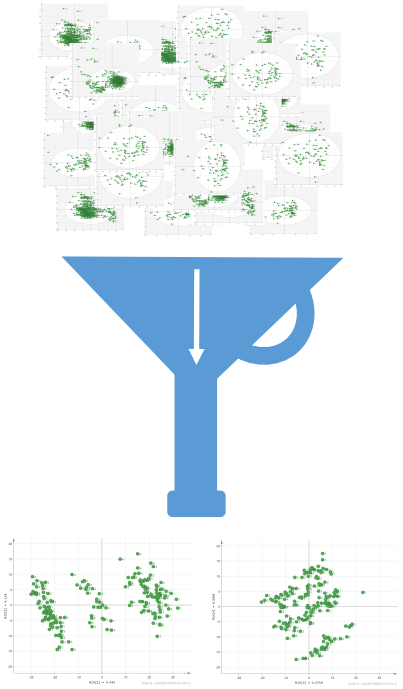
Analyse en Composantes Principales (1)

What we try to do making a PCA

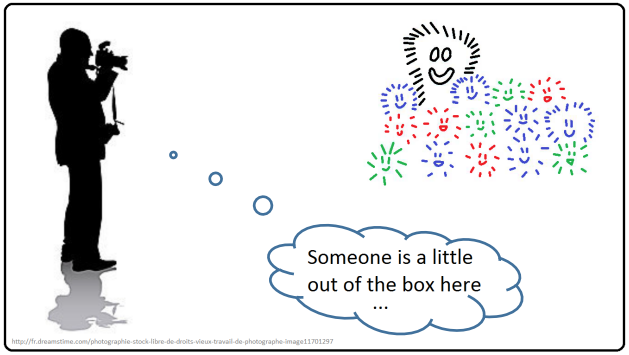
Internal structure detection



Convenient data viewing



Outlier detection



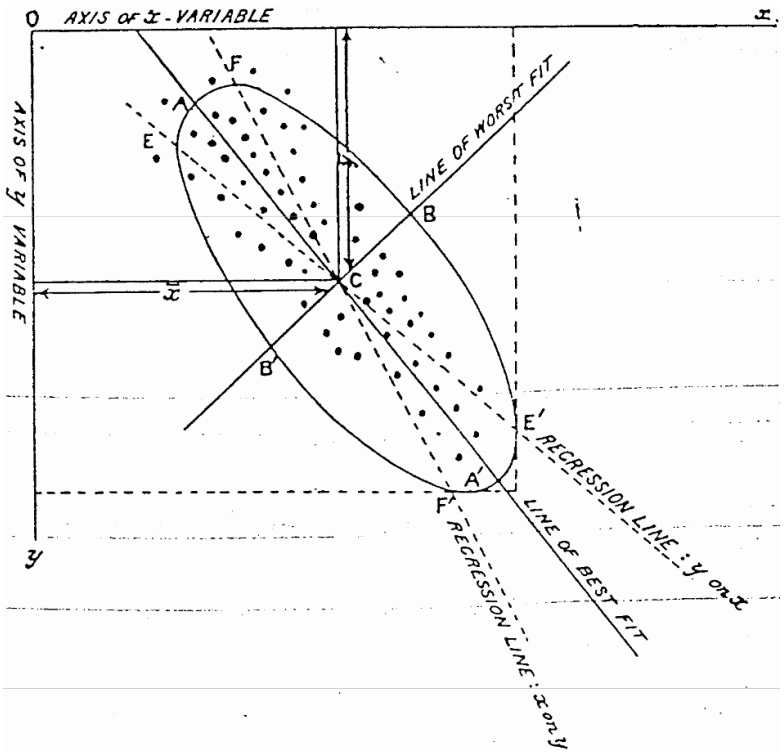
http://fr.dreamstime.com/photographie-stock-libre-de-droits-vieux-travail-de-photographe-image11701297

Analyse en Composantes Principales (2)

- Construction

Il s'agit de déterminer des axes, orthogonaux entre eux, qui maximisent l'inertie projetée sur ceux-ci. Plus simplement dit, on construit des composantes permettant de disperser le plus les individus.

La construction se fait axe par axe, le premier étant celui qui maximise le plus et le dernier le moins.



Pearson, K. 1901.
On lines and planes of closest fit to systems of points in space.
Philosophical Magazine 2:559-572.

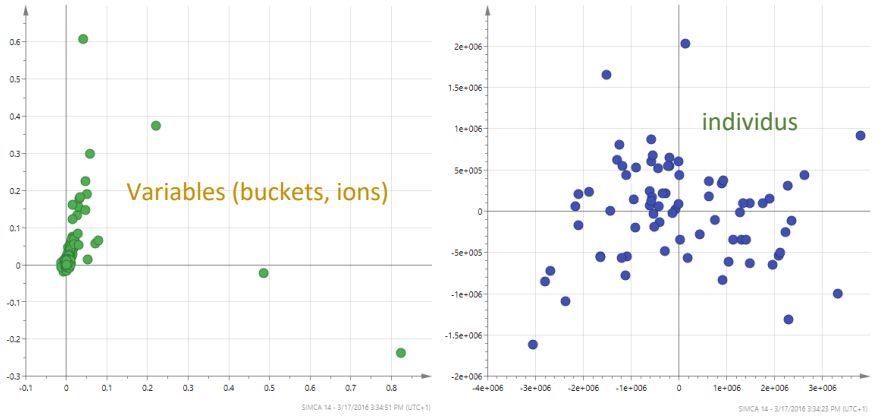
Analyse en Composantes Principales (3)

- Remarque sur les données

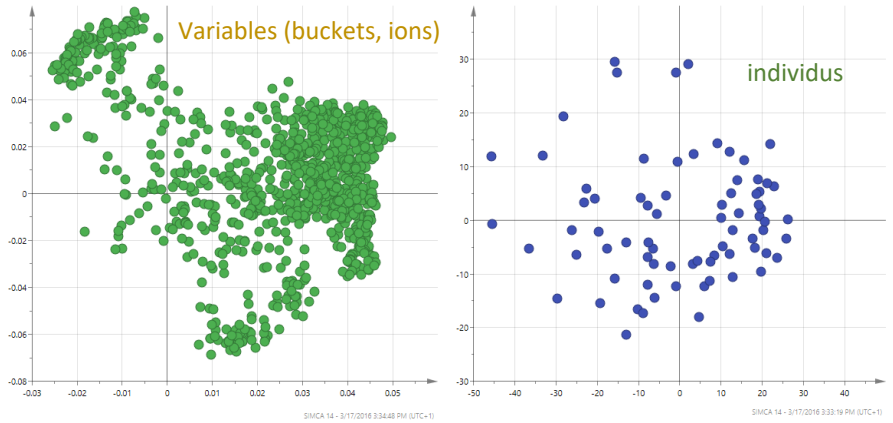
Lorsqu'on réalise une ACP, on centre à minima les données (facilité de calcul et de représentation).

On peut de plus réaliser d'autres transformations, comme la réduction qui a pour avantage de limiter l'avantage qu'ont les variables les plus dispersées dans la construction (par exemple du fait d'unités de mesure différentes)

Données centrées



Données centrées - réduites



Partial Least Square (PLS) regression (Moindres Carrés Partiels)

- One of the most employed methods in metabolomics
  - Worley, B. and R. Powers (2013). Multivariate Analysis in Metabolomics. Curr Metabolomics 1(1): 92-107.
- A method that has produced good results in untargeted metabolomics, especially in discriminant analysis
- A method that matches metabolomic data particularities

PLS regression

- As for PCA, it is based on component construction
  - This time components are calculated to be predictive of a defined set of interesting variables (Y)
  - In the Discriminant Analysis version (PLS-DA), Y is composed of dummy variables representing groups of interest

The idea

- To build a model to predict Y based on constructed components

The procedure

- Select the number of retained components
- Evaluate the goodness of the model
- Check if the model is valid

DATA FILTERING

- “Biological” noise (confounding factors) :
  - Many variability sources: experiment, physiology, instrument…
  - No information for model fitting
- When variability due to factor of interest < variability due to confounding factors, impossible to discriminate individuals according to the factor of interest

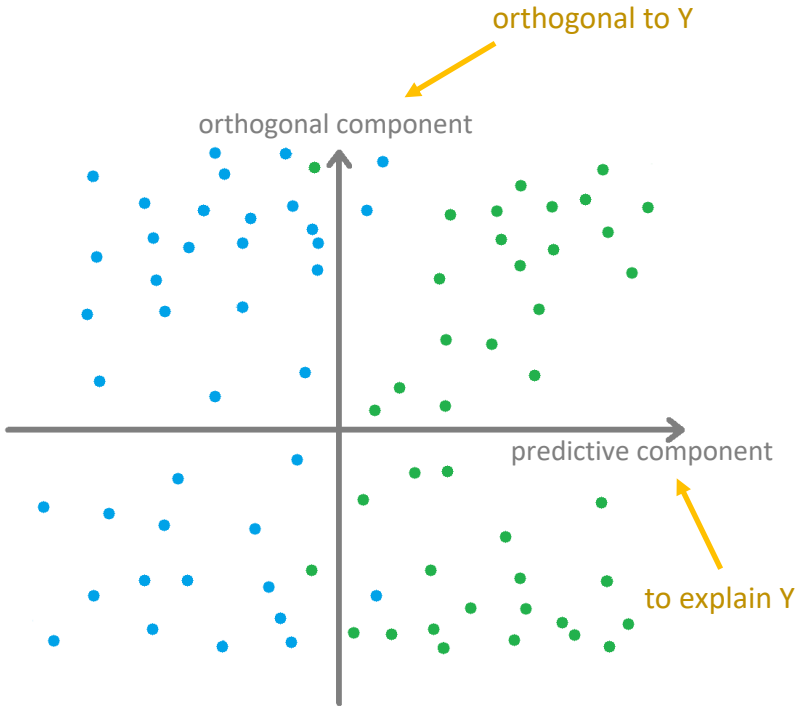
O-PLS (1)

Also exits in Discriminant Analysis derived method (O-PLS-DA)

O-PLS = Orthogonal Projection to Latente Structure

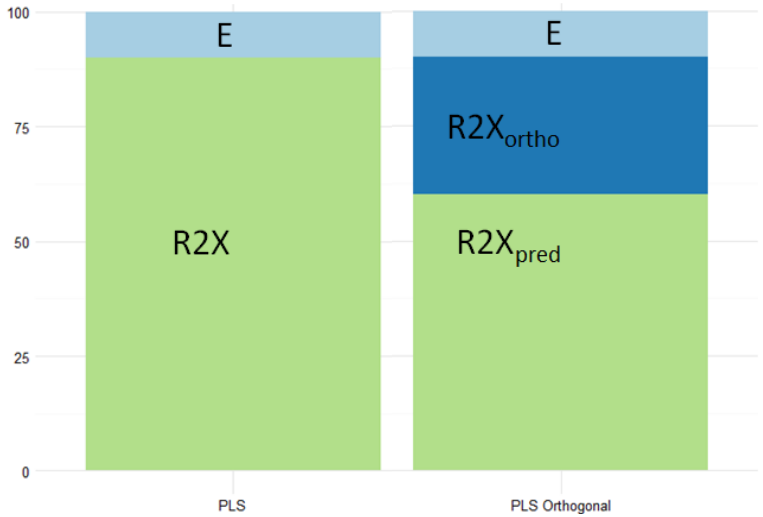
Idea

To simplify interpretation by separating variation among the extracted data: modelling apart what is correlated to Y and what is orthogonal to it



O-PLS (2)

- **O-PLS**: Orthogonal – PLS (Trygg et al. 2002)
  - Division of the X variability
    - in two parts during the PLS process
    - in three parts during the O-PLS process



$$X = T_p P_p' + T_o P_o' + E$$

  - Simultaneous PLS modelling of the predictive and orthogonal variabilities:
  - Enhancement of interpretation of the components but not the overall predictive performance (Q^2)
  - Avoid too many components to avoid overfitting

Sélection des variables discriminantes

- Cas des analyses PLS
  - VIP (Variable Importance in the Projection)
    - Importance de chaque variable dans le modèle
    - Tri des variables selon leur capacité à discriminer les observations
- Croisement avec des analyses univariées
  - Test de Kruskal-Wallis
    - Equivalent non paramétrique ANOVA
    - Basé sur les rangs des observations (pas sensible aux valeurs extrêmes)
    - Récupération de p-values, corrigées des tests multiples (FDR)

Annexes

Analyses statistiques multivariées (2)

- Matrices résultats :
  - **Scores**
    - Coordonnées individus sur nouvelles variables
    - Projection observations dans sous espace formé par nouvelles variables : visualisation des *observations similaires* (proches) / différentes (opposés sur un des axes de projection)
  - **Loadings**
    - Poids variables originales dans combinaisons linéaires
    - Identification des variables importantes dans la construction des nouvelles variables

Moindres Carrés Partiels (1)

- PLS-DA (Régression des moindres carrés partiels – Discriminant Analysis) : modélisation de la relation entre deux matrices de données Y – variables à expliquer (stimulation, pathologie, …) et X – variables explicatives (buckets RMN, profils lipidiques, …)
  - Y = variable qualitative
  - Création de dummy variables (indicatrices) en fonction des modalités de Y

Moindres Carrés Partiels (2)

- Etapes
  - Construction de « nouvelles » variables (variables latentes) :
    - Calcul de ajh (poids de la variable j pour la composante h) : maximisation covariance entre Y et X, sous contrainte d'orthogonalité
  - Régression linéaire de Y sur les variables latentes

$$y = \beta_1 t_1 + \cdots + \beta_p t_p + \varepsilon$$

- Choix du nombre de variables latentes à inclure dans la régression : validation croisée 7-fold
- Validation du modèle : critères de qualité
  - R^2Y : proportion de variance expliquée par le modèle PLS-DA
  - Q^2 : capacité prédictive (modèle valide : $Q^2 > 0.4$)

Clues for interpretation

Cross-validation: recalculation of particular indices using turning subsets of data

- Explained inertia and predictive performance

R^2Y

Y inertia explained by the model

$0 \leq R^2Y \leq 1$

The higher the number of components, the higher the R^2Y

Q^2Y

Indice based on cross-validation, closer to predictive performance estimation

$Q^2Y \leq R^2Y$

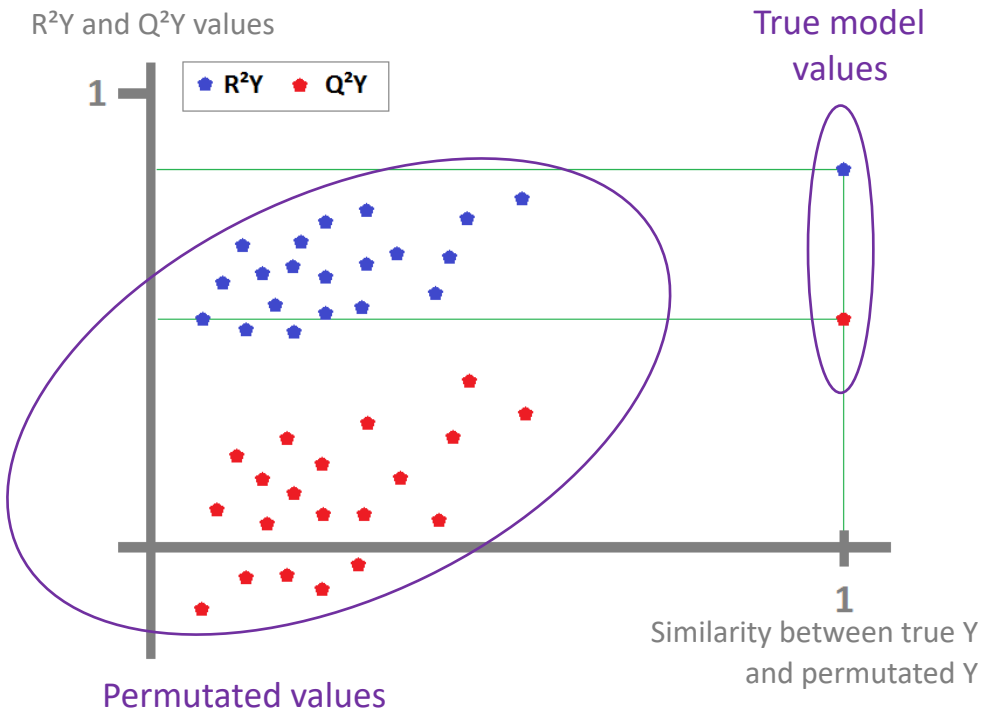
Due to overfitting coming with too many components, Q^2Y is limited by a data-related maximum, and can fall to negative values

Clues for interpretation

- Permutation analysis

This allows to assess if the model seems meaningful or not

- Idea: test how well would do models from the same variable dataset with the same number of components but with false Y variables
- False Y variables are generated by randomly permuting values of true Y



Moindres Carrés Partiels (3)

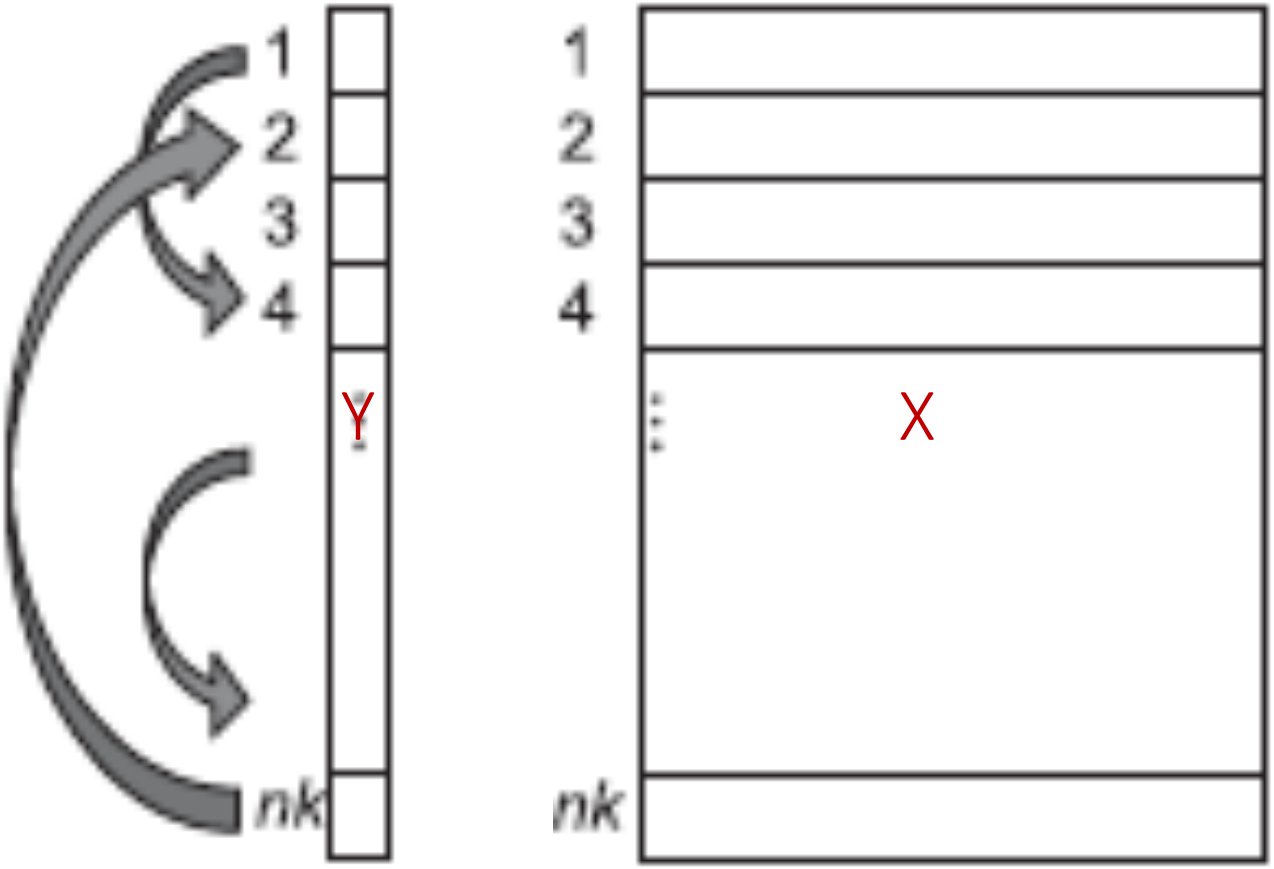
- Robustesse modèle : test des permutations
  - Valeur de Q^2 élevée pas suffisante pour que capacité prédictive modèle bonne
  - Idée : la randomisation des classes va conduire à des modèles de classification mauvais, incapables de distinguer les classes du facteur biologique

$\Rightarrow$ le modèle construit est-il meilleur que le hasard pour classer les observations?

- Hypothèses
  - H_0 : le facteur d'intérêt (BPA, régime alimentaire, …) n'a aucun effet sur le métabolome (pas de différence entre les groupes)
  - H_1 : le facteur d'intérêt a un effet sur le métabolome

Test des permutations

- Principe
  - Permutation de Y (mélange aléatoire des valeurs) sans modification de la matrice X



Test des permutations

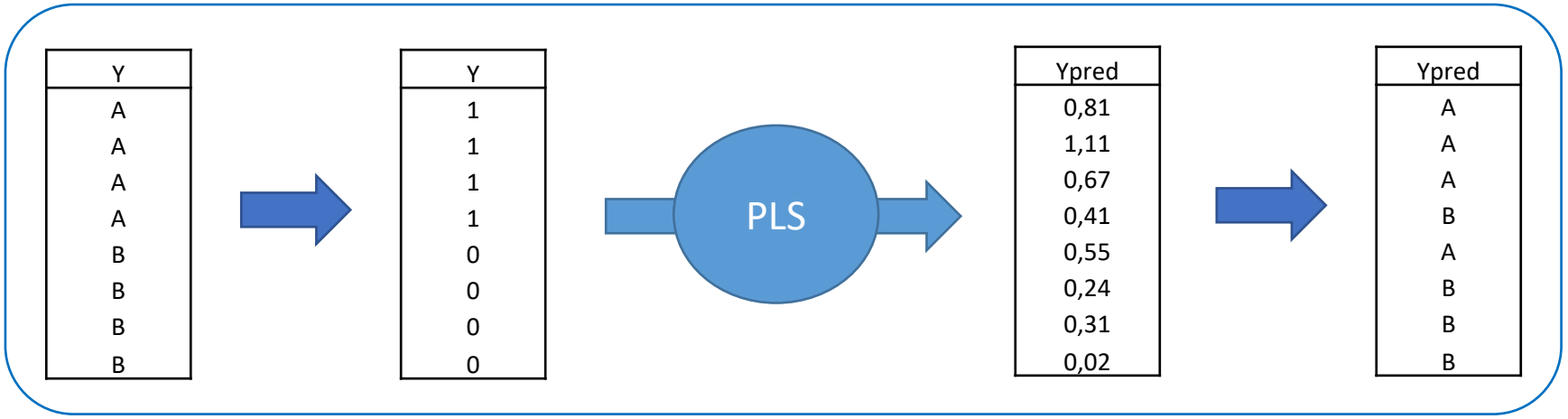
- Principe
  - Permutation de Y (mélange aléatoire des valeurs) sans modification de la matrice X
  - Construction modèle PLS-DA sur données permutées : calcul Q^2
  - Répétition permutation un grand nombre de fois
  - Comparaison de la qualité des modèles « permutés » à la qualité du modèle « brut »
  - Représentation des Q^2 en fonction de la corrélation entre Y « brut » et Y « permuté »
    - Modèle robuste : ordonnée à l'origine de la droite de régression qui passe dans nuage de points < 0

PLS – DA (Discriminant Analysis)

- Adaptation of PLS for discriminant analysis

Y quantitative → Y qualitative

General idea



OSC

- Orthogonal Signal Correction (Wold et al. 1998)
  - Pre-filtering method used prior to fitting a PLS model aiming at removing variability in the X spectral data orthogonal (unrelated) to the Y biological factor
  - Based on a two-steps algorithm, using the PLS model
    - Computation of linear combinations $t_h^o = \sum_{j=1}^{p} X_j a_{hj}^o$ such that $\mathrm{cov}(Y, t_h^o) = 0$
    - Subtraction of the orthogonal information included in the X matrix $\tilde{X} = X - t_h^o p'$
  - Iteration of these two steps until there is no orthogonal information in the X matrix
  - Prone to overfitting

O-PLS (1)

- **O-PLS**: Orthogonal – PLS (Trygg et al. 2002)
  - Division of the X variability
    - in two parts during the PLS process
    - and in three parts during the O-PLS process