# Datamining 5: La grande dimension
# La regression PLS (Partial Least Square)

M2 STD

November 5, 2015

# L'idée de base

Revenons dans un premier temps a de la regression linéaire. Le modele de regression linéaire multiple est :

$$Y = XA + B + \varepsilon$$

ou $Y \in \mathcal{M}_{n,1}(\mathbb{R})$, $X \in \mathcal{M}_{n,d}(\mathbb{R})$, sont les "matrices" de variables a expliquer et explicatives.
$A \in \mathcal{M}_{d,1}(\mathbb{R})$ est le vecteur des coefficient, $B = b(1, \ldots, 1)'$ est la "constante" et $\varepsilon$ le vercteur d'erreur.
Lorsque les différentes variables de $X$ sont indépendantes, le bruit gaussien et indépendant de $X$ la méthode des moindres carrés fonctionne très bien et on a des test statistiques explicites de significativité des coefficients
Lorsque les différentes variables de $X$ sont corrélées, la méthode des moindres carrés fonctionne très bien **(on a un $R^2$ "cible")** mais **aucun des coefficient n'est interprétable**

# Exemple de données corrélées : le cas "parfait"

$X_2 = X_1$ et le "vrai modele" est $Y = X_1 + X_2$
Alors toute combinaison : $Y = \lambda X_1 + (2 - \lambda) X_2$ est solution des moindres carrées
On ne peut en aucun cas interpréter $\lambda$

# Exemple de données corrélées : noter l'importance des résidus !!!!

$X_2 = X_1 + \varepsilon_1$ ($\varepsilon_1$ indépendant de $X_1$)
Le "vrai modele" est $Y = X_1 + X_2 + \varepsilon$ ($\varepsilon$ indépendant de $X_1$ et $X_2$) peut donc s'ecrire aussi :
$Y = 2X_1 + \varepsilon_1 + \varepsilon$ ($\varepsilon$ indépendant de $X_1$ et $\varepsilon_1$)
Pour retrouver les coefficients on peut donc:

1. regresser $X_2$ sur $X_1$ pour obtenir $X_2 = \hat{a}_2 X_1 + \hat{\varepsilon}_1$
2. regresser $Y$ sur $X_1$ et $\hat{\varepsilon}_1$ pour obtenir $Y = \hat{a} X_1 + \hat{b}\hat{\varepsilon_1} + \hat{\varepsilon}$
3. retrouver $Y = \alpha X_1 + \beta X_2 + \varepsilon$ en identifiant $\beta = \hat{b}$ et $\alpha = \hat{a} - \hat{b}$

discuter des stat, tests et interprétations des coefficients

# Hypothèse de travail

Dans toute la suite on supposera que les variables explicatives et a expliquer sont centrées réduites (i.e. entre autre qu'on travaille sans constante).

# Les solutions "naives"

- Utiliser une pseudo-inverse plutot qu'une inverse
- Travailler sur les $k < d$ composantes principales d'une $ACP$ (discuter des problemes)

# La regression sur les composantes principales

On détermine a l'aide de l'ACP une matrice $W \in \mathcal{M}_{d,k}(\mathbb{R})$ et on regresse $Y$ sur $T = XW$.

# point commun PLS et regression sur composantes principales:

Dans les deux cas on va rechercher une matrice $W \in \mathcal{M}_{d,k}(\mathbb{R})$ telle que "les colonnes de $XW$ soient indépendantes" et regresser $Y$ sur $XW$

1. Dans le cas de la regression sur composantes principales $W$ est calculée a partir de $X$
2. Dans le cas de la regression PLS $W$ est calculée a partir de $X$ et $Y$

# Un peu de vocabulaire

Dans la relation $T = XW$

- $T$ est appelée matrice des scores
- $W$ est appelée matrice des poids (ou loading)

# Le modèle PLS

Dans le modèle *ACP* l'idée est de chercher un nombre réduit $k$ de transformations linéaires de $X$ qui sont indépendantes et qui "expliquent" au mieux $X$

Dans le modèle *PLS* l'idée est de chercher un nombre réduit $k$ de transformations linéaires de $X$ qui sont indépendantes et qui "expliquent" au mieux $Y$

# L'algorithme de PLS

Pour $h = 1 \ldots k$ on cherche les $w_h$ tel que $t_h = Xw_h$:

- Maximise $Cov(t_h, Y)$ (explique au mieux $Y$) sous les contraintes :
  - $||w_h|| = 1$ (le vecteur est normé)
  - pour tout $i < h$ : $\langle w_h, w_i \rangle = 0$ (il est orthogonal aux précédentes aux précédents vecteurs trouvés).

# L'algorithme de PLS: premiere étape

1. $t_1 = X.w_1$ où $w_1 = (cov(X_1, Y), \ldots cov(X_d, Y))$ (maximisation de la covariance par projection)
2. $w_1 := w_1/||w_1||$
3. $Y = c_1 t_1 + Y^{(1)}$ (regression de $Y$ sur $t_1$)
4. $X = t_1 P_1' + X^{(1)}$ (regression de $X$ sur $t_1$)

Notez que les variables sont supposées réduites et qu'ainsi la covariance est une corrélation ! Notez aussi que l'axe 1 jouera "positivement" sur $Y$

# L'algorithme de PLS: deuxième étape

1. $t_2 = X^{(1)} w_2$ où $w_2 = (cov(X_1^{(1)}, Y^{(1)}), \ldots cov(X_d^{(1)}, Y^{(1)}))$
2. $w_2 := w_2 / ||w_2||$
3. $Y^{(1)} = c_2 t_2 + Y^{(2)}$
4. $X^{(1)} = t_2 P_2' + X^{(2)}$

Notez l'apparition des résidus de la regression précédente On itère ce processus jusqu'a la $k$eme étape Rq : a l'étape $h$ la somme des $Y(i+1)^2$ (ECQ) est notée $Ress(h)$ (Residual Sum of Square) et cette même somme mais "cross validée" est notée $Press(h)$ (Predicted Residual Sum of Square)

# Au final on peut ecrire

- $T = XW$
- $X = X_2P + E$ (i.e. $X_2$ permet de reconstruire $X$ a l'erreur $E$ pret, on peut voire la matrice $P$ comme une pseudo inverse)
- $Y = X_2Q + \varepsilon$ ($X_2$ permet de prédire linéairement $Y$ a l'erreur $\varepsilon$ pret

# Lecture comme regression

On a ainsi un prédicteur de $Y$ mais aussi, du fait de l'aspect assez linéaire des indicateurs d'aide a l'analyse. On peut ainsi determiner l'importance de la variable numéro $i$ ($X_i$) dans la regression a $k$ composantes via le $VIP$ (Variable Importance in the Prediction)

$$VIP_{k,i} = \sqrt{\frac{d \sum_{h=1}^{k} R^2(y, t_h) w_{h,j}^2}{\sum_{h=1}^{k} R^2(y, t_h)}}$$

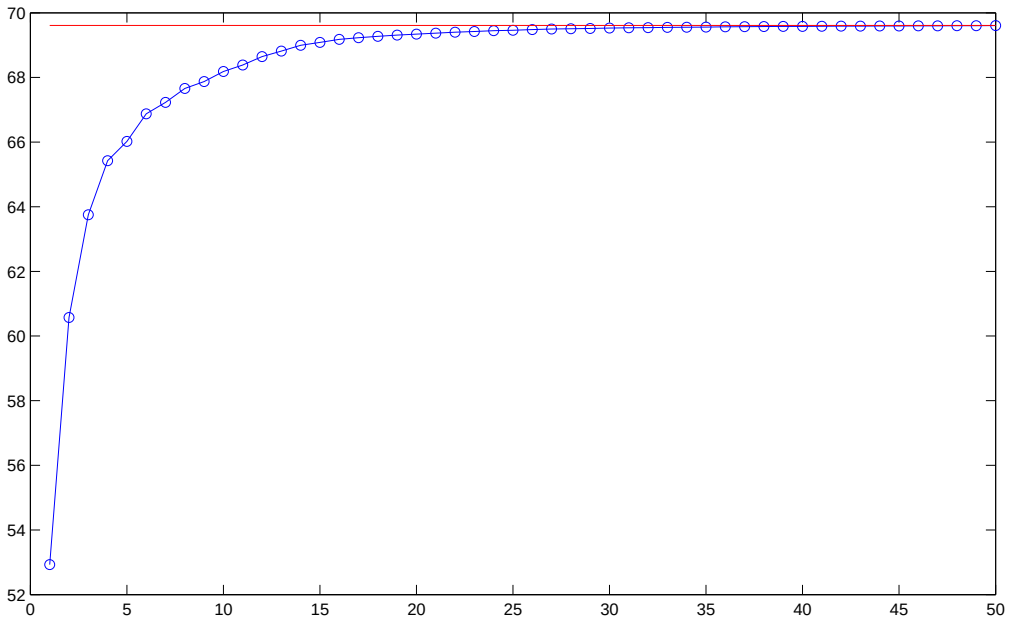
Usuellement on demande a une variable d'avoir un $VIP \geq 0.8$

# Choix de $k$ (nombre de variables)

- On peut bien sur réaliser une validation croisée.
- Ou le critère $Press_h \leq 0.95 Ress_{h-1}$
- Ou encore définir le $Q^2_{cum}(h) = 1 - \prod_{k=1}^{h} \frac{Press_k}{Ress_{k-1}}$ et retenir les composantes dont le $Q^2$ est nettement superieur a "celui d'après" (idem ACP).

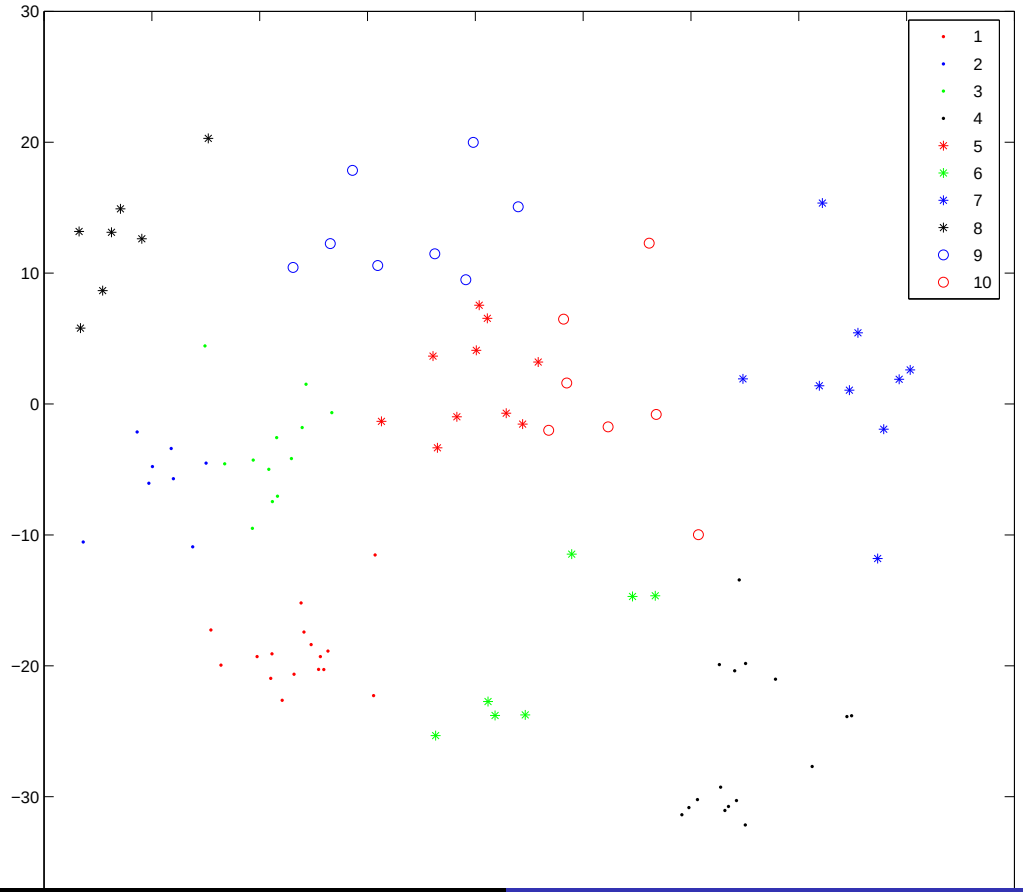
# exemple

Retour sur la criminalité aux états unis On a 100 variables explicatives avec un grand nombre de correlations, un $R^2$ de regression linéaire a 0.69 et 27 variables qui sont significatives dans la regression :



# exemple

Retour sur la criminalité aux états unis On a 100 variables explicatives avec un grand nombre de correlations, un $R^2$ de regression linéaire a 0.69 et 27 variables qui sont significatives dans la regression :



# Le cas de $Y$ multivarié

C'est la $PLS2$ en fait ce qu'on vient de voir s'appelle officiellement $PLS1$

# L'algorithme de PLS1 VS PLS2

Pour $h = 1 \dots k$ on cherche les $w_h$ tel que $t_h = Xw_h$ et les $u_h = Yv_h$

- Maximise $Cov(t_h, u_h)$ sous les contraintes :
  - $||w_h|| = 1$ (le vecteur est normé) $||v_h|| = 1$
  - pour tout $i < h$ : $\langle w_h, w_i \rangle = 0$ (il est orthogonal aux précédentes aux précédents vecteurs trouvés).

Donne type a exactment le même type d'algorithme de résolution que $PLS1$.

# Au dela du modèle predictif “numérique”

En plus de “prediction” numérique on a des aides graphiques a l'interprétation (type ACP) qui aident a comprendre le lien entre les variables (et les positionnement des individus) Rappelons qu'on obtient $T = XW$ une “réduction de dimension” de $X$ (en BON). On peut donc :

- Construire des cercles de corrélations commen en ACP (on regrade les corrélations des $X_i$ avec les $T_j$ ainsi que les $Y_i$ avec les $T_j$).
- Avoir une projection des individus (mais c'est moins important car on perd $Y$)

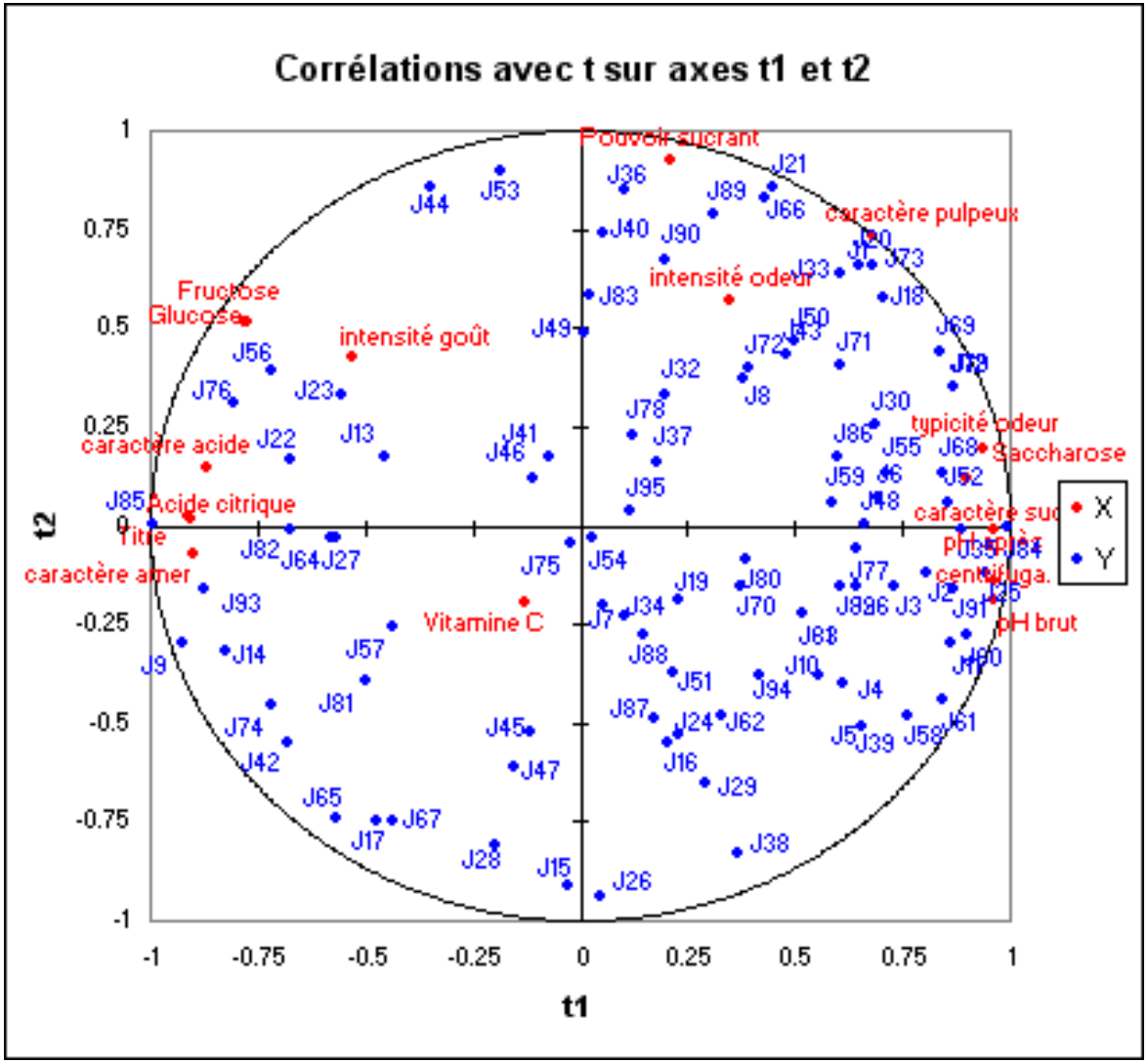
# exemple classique de la dégustation de jus de fruits

- On a 5 jus d'oranges individus (qualifiés par leur marque)
- Pour chaque jus de fruit on a 16 variables explicatives "objectives" quantitatives (quantite de glucose de fructose.....)
- Pour chaque jus de fruit on a les notes attribues par 96 juges (96 variables a expliquer)

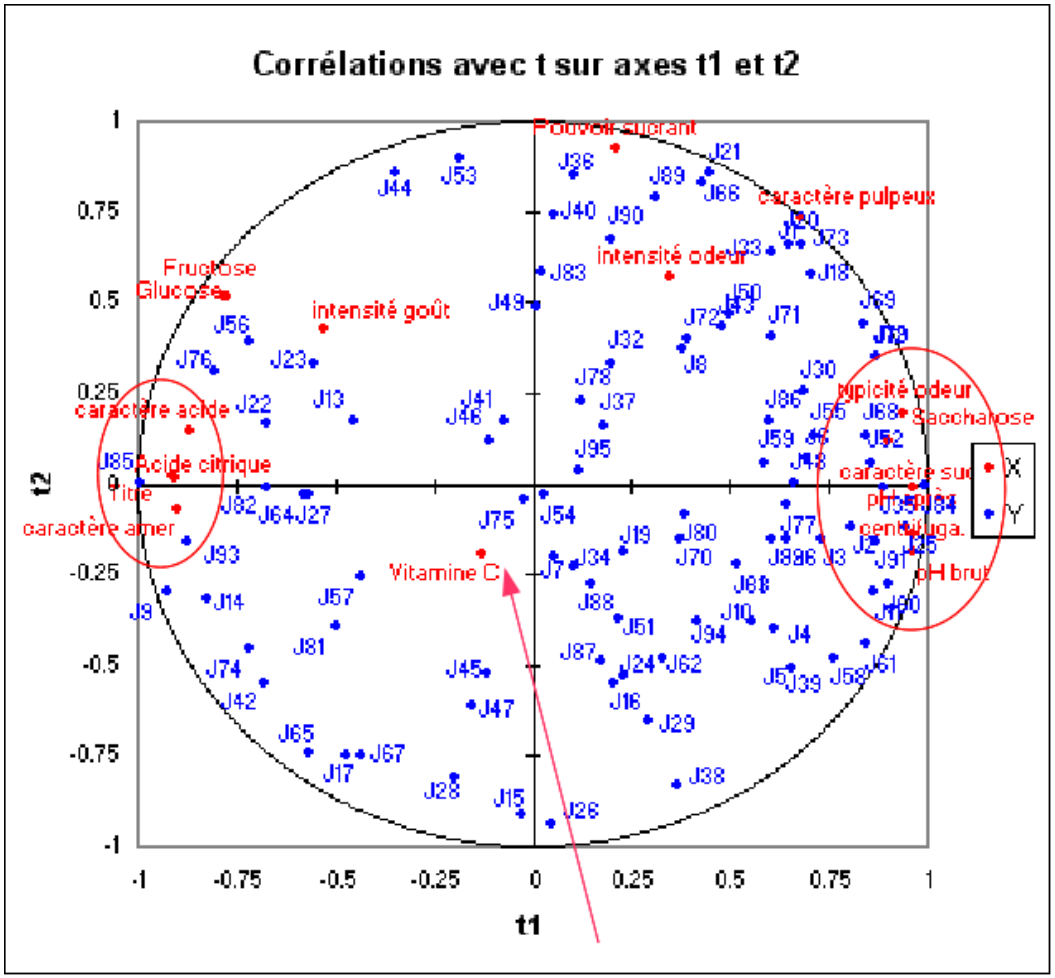
# exemple classique de la dégustation de jus de fruits

On pourrait faire 96 regressions $PLS1$ pour expliquer le comportement de chacun des juges (ce qui donnerais d'asses bons résultats pour tous les juges) ... ou... essayer de tout faire d'un coup, on a une explication plus "globale".

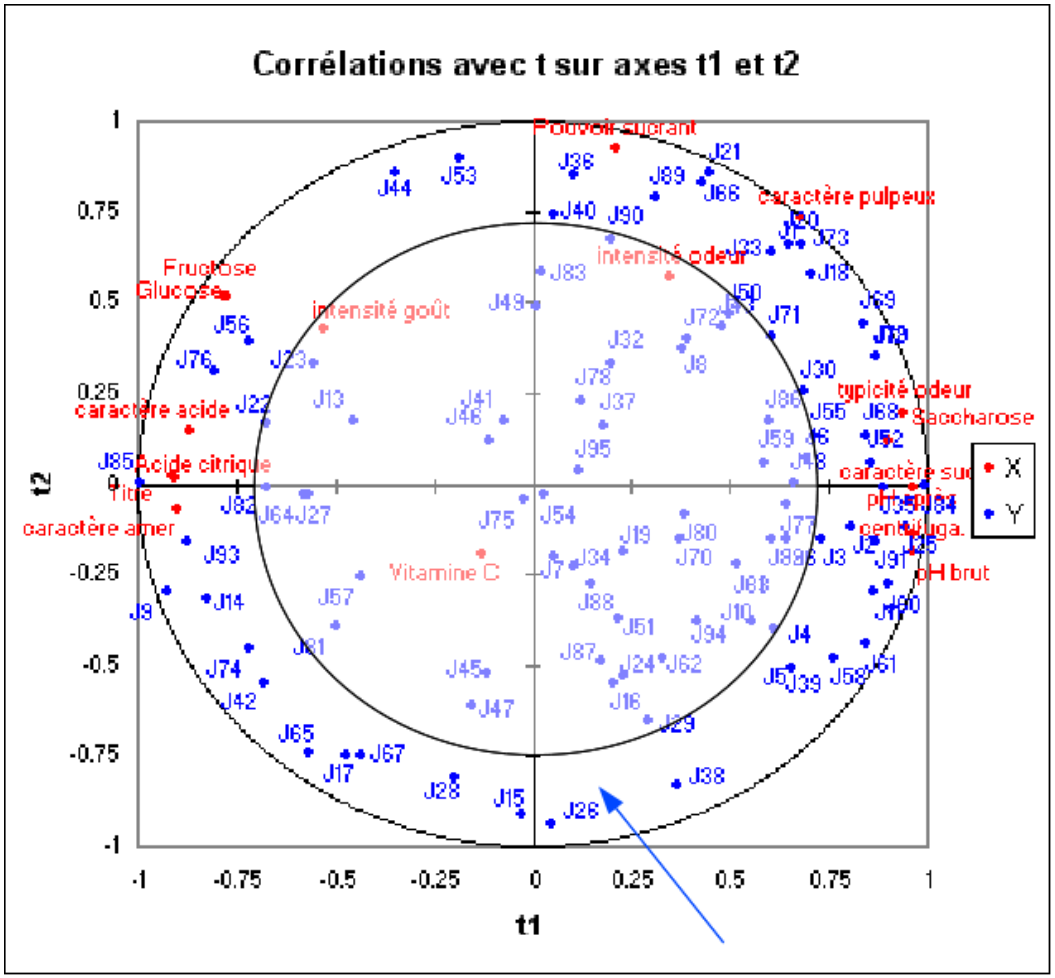
# Le “cercle des corrélations”



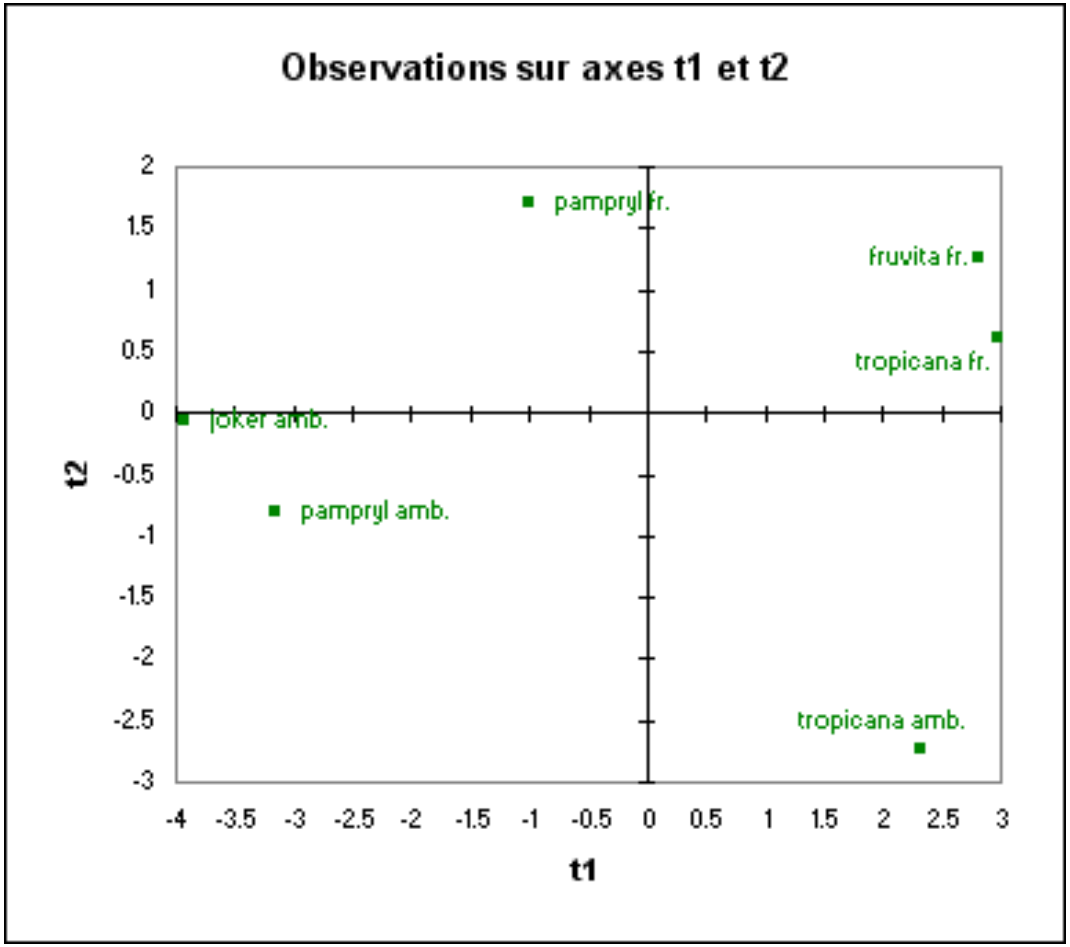
# un gros groupe de variables très corrélées



# Interprétation "juge"



# La position des individus



# Les liens entre regressions et classification

Supposons que $Y$ la variable a prédire soit binaire et ait pour modalité 0 et 1. Alors le problème de regression est équivalent au problème de classification (l'erreur quadratique est exactement le taux de mal classé). Ceci nous permet d'utiliser aussi la regression *PLS* dans ce cas ! Attention des qu'il y a plus de deux classes cela ne marche plus !