MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

# Partial Least Squares (2-blocks PLS)

**Michel BAYLAC**
*UMR 5202-UMS 2700 plate-forme Morphométrie MNHN*

# Partial Least Squares (2-blocks PLS) Semin-R MNHN 23 janvier 2009

M. Baylac UMR 5202-UMS 2700 plate-forme Morphométrie

**PLS : appartient aux méthodes d'analyses des relations multi-tableaux**

**pleine expansion : vaste champ avec de multiples variantes !**

**On se limitera au seul cas des PLS appliqués à deux tableaux (2B-PLS) dont un au moins correspondant à des données morphométriques**

**Dans ce cadre : comparaison avec les corrélations canoniques (CC) et la régression multiple**

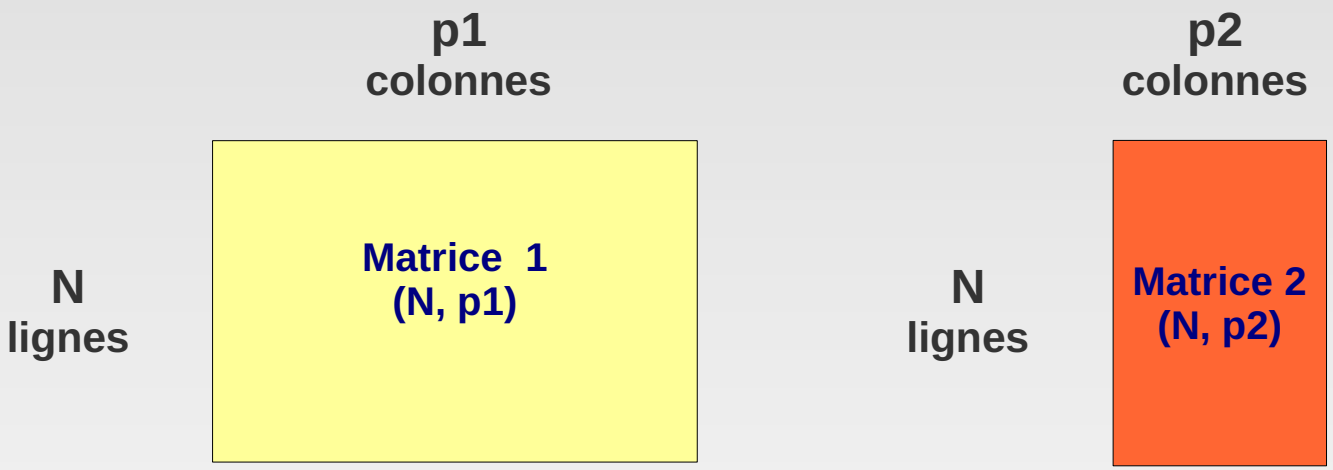
**illustrera l'intérêt relativement propre au contexte morphométrique de la richesse interprétative, impossible avec CC**

Ref principale : Rohlf & Corti, 2000. Systematic Biology, 49(4):740–753.

# Partial Least Squares (2-blocks PLS) Semin-R MNHN 23 janvier 2009

M. Baylac UMR 5202-UMS 2700 plate-forme Morphométrie

## Données




**Implique une forme d'appariement entre lignes :**

- Deux séries de caractéristiques différentes relevées sur les mêmes objets
- Géméllité vraie

- Dimorphisme sexuel intra (portée, population ...) ?
- Dimorphisme sexuel entre moyennes de populations ou espèces ou taxons supraspécifiques ?
-

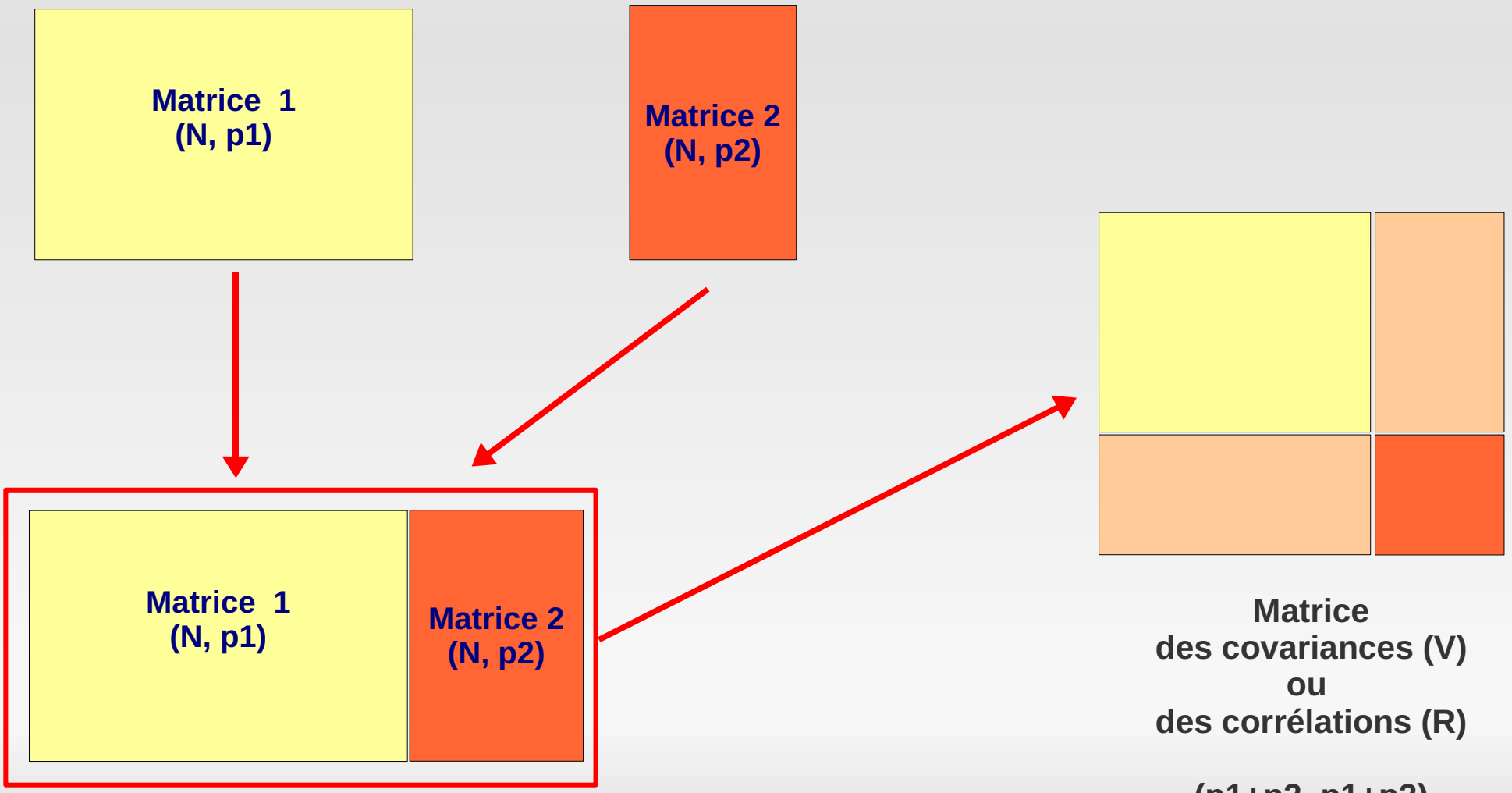
## Données et appariement

- Deux séries de caractéristiques différentes relevées sur les mêmes objets
  - Formes / couleurs / caractéristiques biochimiques / moléculaires / physico-chimiques
  - ...
- Evolution temporelle d'un même ensemble ?
- Géméllité fausse / vraie ?
- Dimorphisme sexuel intra- portée, population ...) ?
- Dimorphisme sexuel inter- (moyennes de populations ou espèces ou taxons supraspécifiques) ?

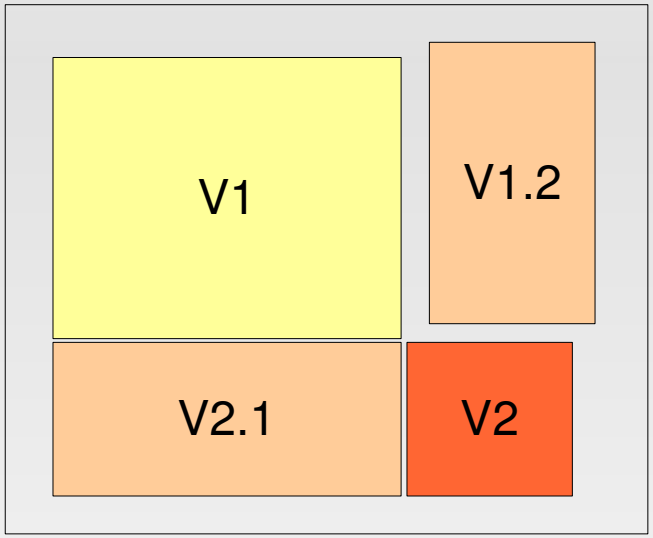
# Partial Least Squares (2-blocks PLS) Semin-R MNHN 23 janvier 2009

M. Baylac UMR 5202-UMS 2700 plate-forme Morphométrie

## Principes



## Principe des 2B-PLS




V (resp. R) est une matrice partitionnée
de dimensions (p1+p2, p1+p2)

Si on prend V :

- V1 = matrice des covariances bloc 1
- V2 = matrice des covariances bloc 2

- V1.2 (ou V2.1 sa transposée : V1.2 = V2.1')
  = matrice des covariances inter blocs

L'analyse porte uniquement sur la matrice des covariances inter blocs

et classiquement sur V1.2

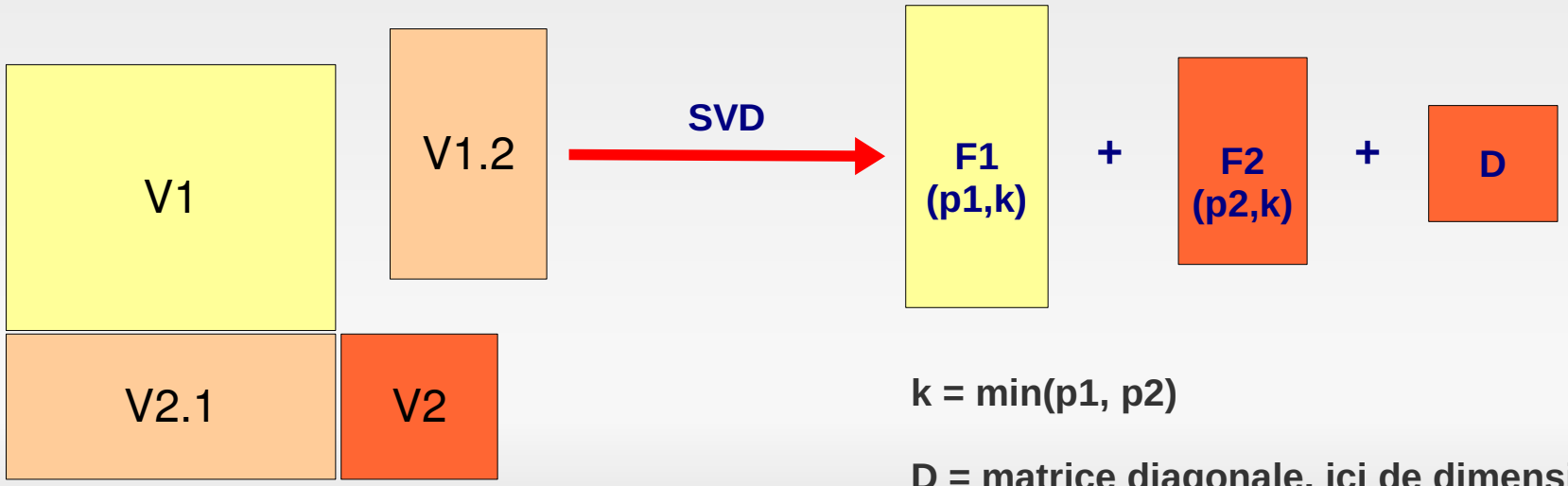
## Principe des 2B-PLS : analyse factorielle de V1.2 (ou R1.2)

V1.2 n'est pas symétrique et n'est ordinairement pas carrée

=> pas de diagonalisation possible

=> décomposition en valeurs singulières (Singular Value Decomposition = SVD)

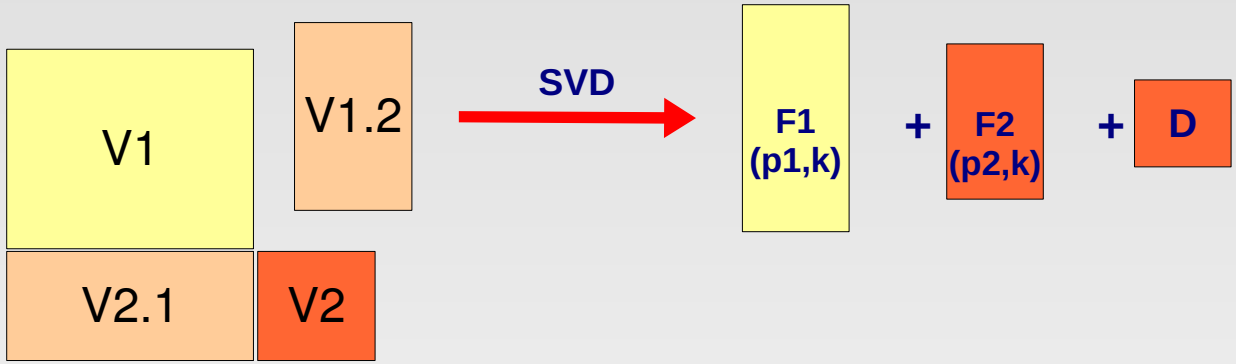
**F1 . D . F2' = V1.2**




k = min(p1, p2)

D = matrice diagonale, ici de dimensions (p2, p2)

## Principe des 2B-PLS : analyse factorielle de V1.2 (R1.2)

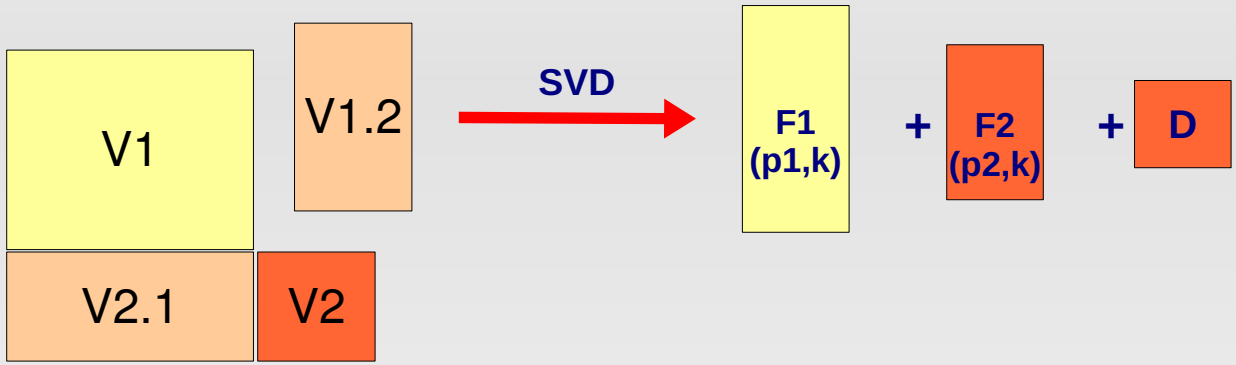



D = vecteur des valeurs singulières $\lambda_i$
= racines-carrées des variances communes 1-2

F1, F2 = deux ensembles fournissant les poids (saliences) affectés aux variables de F1 et F2 :

- F1 = saliences pour M1
- F2 = " " M2

## Principe des 2B-PLS : analyse factorielle de V1.2 (R1.2)



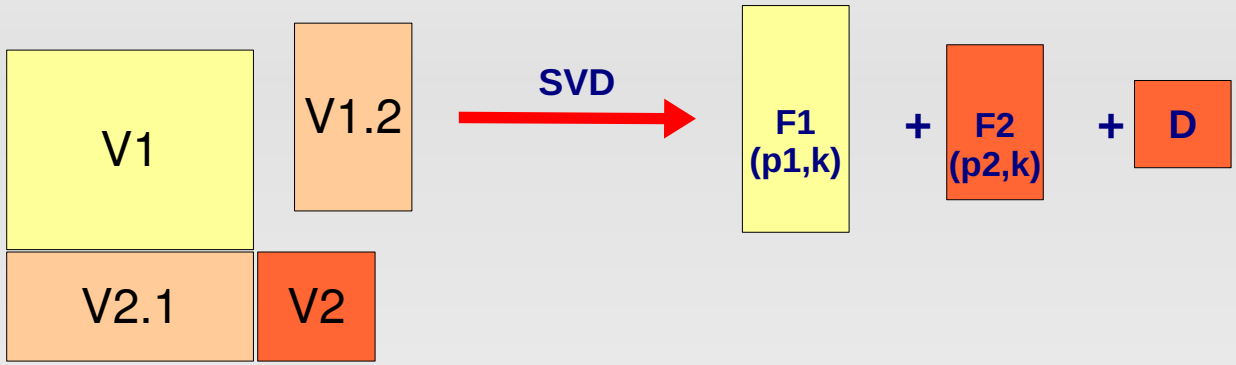

**D = vecteur des valeurs singulières $\lambda_i$ = racines-carrées des variances communes 1-2**

**$\lambda^2_i$ / $\sum^{1..k} \lambda^2_i$ : estimation de la proportion de covariance totale pour la ième combinaison**

**Dans le cas de R1.2, la covariance maximale possible = p1 * p2 (= somme des élements tous égaux à 1)**

**Dans ce cas seul, $\sum^{1..k} \lambda^2_i$ fournit une mesure de la covariance globale entre les deux blocs**

## Principe des 2B-PLS : analyse factorielle de V1.2 (R1.2)

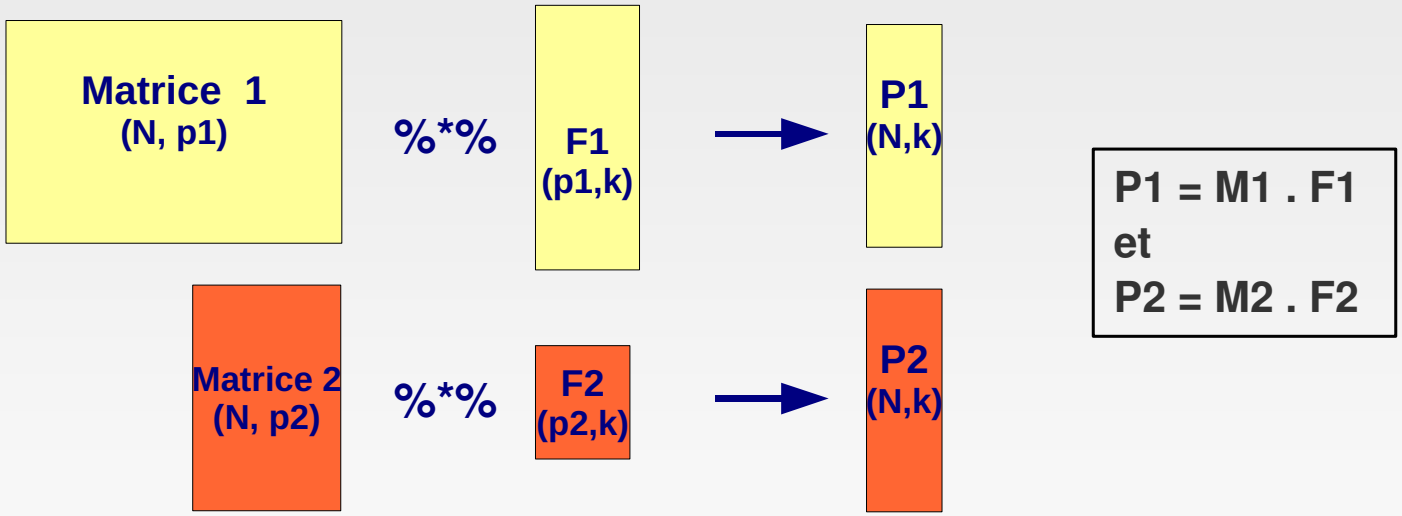



**Interprétations de F1 et F2**

**Saliences sont normées (unitaires) : leur comparaison pour des vecteurs de rangs différents implique quelles soient au préalable multipliées par leurs valeurs singulières**

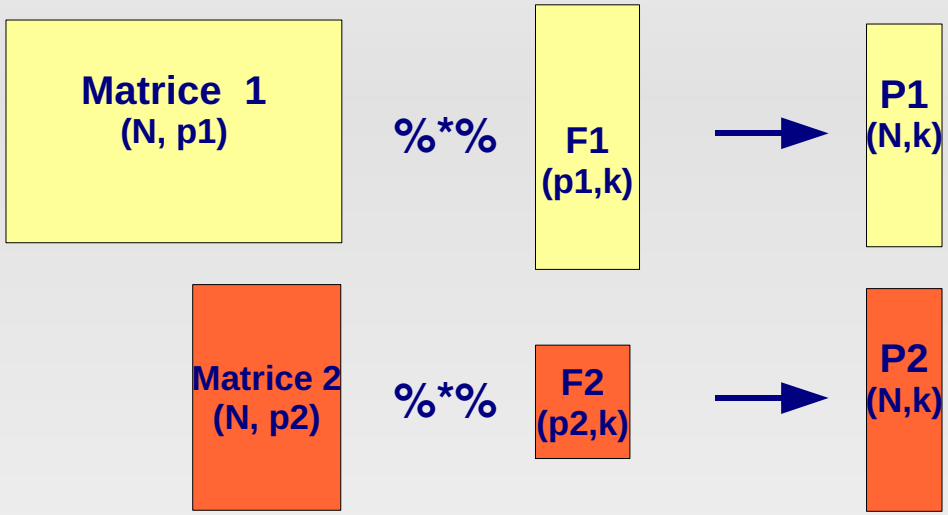
## Principe des 2B-PLS : analyse factorielle de V1.2 (R1.2)

- Si V1.2 (ou R1.2) sont les seules données disponibles, l'analyse s'arrête là.
- Si on dispose des données initiales, on peut par contre calculer les projections P1 et P2 sur les directions données par F1 et F2




P1 = M1 . F1
et
P2 = M2 . F2

## Principe des 2B-PLS : analyse factorielle de V1.2 (R1.2)




P1 = M1 . F1
et
P2 = M2 . F2

- Interprétations possibles en calculant les corrélations entre les variables et les projections pour chacun des blocs :

```
> cor(cbind(M1, P1))
```
et
```
> cor(cbind(M2,P2))
```

similitude avec les corrélations variables-facteurs en ACP

## 2B-PLS : tests

**pas de tests statistiques de signification associés aux variables latentes**

- mais test possible par permutations :
  - **H0** : les % de variances ne diffèrent pas de % au hasard
    - on permute aléatoirement l'ordre des objets dans l'un ou l'autre des deux blocs. On stocke les % de variances obtenus
    - on répète le processus permutation + SVD 999 fois par exemple si Nrepet = 1000
    - on calcule le nombre de fois Nsup où le % de variance obtenu par permutation est >= à celui observé avec les données initiales
    - Nsup/Nrepet donne le risque alpha du test

## Exemple Fowls (data modifiées Marcus)

```
f <- read.table('~/Desktop/2B-PLS_Semin-R/fowldata.txt', header=TRUE)
M1 <- log(f[,2:3])
M2 <- log(f[,5:8])

args(pls)
fpls <- pls(M1,M2)
fplsr <- pls(M1,M2, plstype= »cor »)

summary(fpls)
summary(fpls, short=FALSE)
summary(fplsr, short=FALSE)

args(plot.pls)
plot(fpls, group=f$CODE)
plot(fpls, group=f$CODE, vecnum=2)
```

## Corrélations canoniques

- buts:
  - tester l'indépendance entre deux blocs (= matrices)
  - trouver un ensemble des combinaisons de variables de chacuns des blocs (variables canoniques) qui sont :
    - maximalement corrélées
    - indépendantes (=orthogonales) les unes des autres
  - maximisation corrélations entre variables canoniques, pas entre les variables initiales : les variables canoniques n'expliquent pas nécessairement une part importante de la covariation entre blocs

## Corrélations canoniques

- min(p1,p2) axes canoniques (comme PLS)
- calcul plus complexe faisant appel à R12 mais aussi à R1 et R2 (matrices intra-blocs). Sommairement (notation matricielle) :
  - F1 : diagonaliser $R1^{-½} * R12 * R2^{-1} * R21 * R1^{-1/2}$
  - F2 : diagonaliser $R2^{-½} * R21 * R1^{-1} * R12 * R2^{-1/2}$
- valeurs propressont identiques dans chaque bloc et sont égales aux carrés des coefficients de corrélations canoniques
  - analogie avec $R^2$ pour chaque paire d'axes
  - analogie stricte avec $R^2$ est le redundancy index
- tests associés : test de signification des corrélations canoniques pour chacun des axes
- test global de signification de l'ensemble des axes

## PLS et corrélations canoniques

- PLS : variables canoniques (variables latentes) expliquant la covariance entre blocs
- CC : couples de variables canoniques maximalement corrélées (mais qui ne prennent pas en compte le max de covariance)
- Dans les 2 cas : relations linéaires
- Interprétation des axes
  - directe (ou presque) avec PLS : corr variables-F1 ou F2 (loadings)
  - difficile avec CC : plusieurs possibilités (weights, loadings) mais
    - instabilité notable des coefficients
    - magnitude liée à plusieurs causes possibles

## PLS, corrélations canoniques et régression multiple

**Si un des bloc n'a qu'une seule variable :**

- **avec les PLS : forme de régression multiple particulière (régression PLS)**
- **Rôle symétrique des deux blocs (pas de var. dépendante et indépendante)**
- **= relation structurale ou symétrique**
- **voir library(pls)**

- **avec les CC : modèle de régression multiple classique :**
- **coefficients = coef. régression partielle normalisés**

- **CC comme une généralisation de la régression multiple**