

# Apport de l'ACP probabiliste pour la gestion des données manquantes en ACP

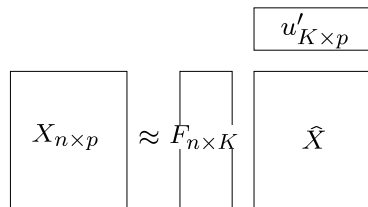
Julie Josse, François Husson, Jérôme Pagès

Laboratoire de mathématiques appliquées, Agrocampus Ouest, Rennes

Société Française de Statistique (SFdS)  
Bordeaux, 27 mai 2009

## Minimiser l'erreur de reconstitution

⇒ Approximation de  $X$  par une matrice de rang  $K < p$



$$\begin{aligned} \mathcal{F} &= \|X_{n \times p} - F_{n \times K} u'_{K \times p}\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \sum_{k=1}^K F_{ik} u_{jk})^2 \end{aligned}$$

- $u$  axes principaux (normés à 1)
- $F$  composantes principales (normées à la valeur propre)

⇒ Diagonalisation de la matrice de variance-covariance

⇒ Algorithmes itératifs

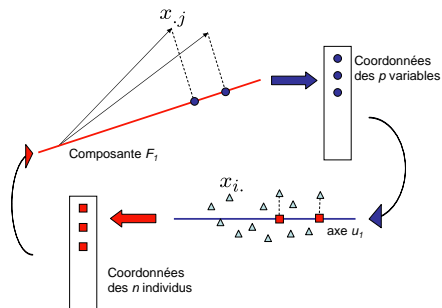
# ACP via NIPALS (Non linear Iterative PARTial Least Squares)

- Wold (1966, 1969) : méthode séquentielle  $\Rightarrow$  meilleure approximation de rang 1

$$\mathcal{F}_1 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - F_{i1} u_{j1})^2$$

$$\frac{\partial \mathcal{F}_1}{\partial F_{i1}} = 0 \rightarrow u_{j1} = \frac{\sum_i (x_{ij} \times F_{i1})}{\sum_i F_{i1}^2}$$

$$\frac{\partial \mathcal{F}_1}{\partial u_{j1}} = 0 \rightarrow F_{i1} = \frac{\sum_j (x_{ij} \times u_{j1})}{\sum_j u_{j1}^2}$$



- Déflation : une fois  $(\hat{F}_1, \hat{u}_1)$  trouvé, on cherche  $(\hat{F}_2, \hat{u}_2)$  premier axe et première composante de  $\tilde{X} = X - \hat{F}_1 \hat{u}_1'$

# ACP via la recherche directe du sous-espace

$$\mathcal{F} = \|X_{n \times p} - F_{n \times K} u'_{K \times p}\|^2$$

⇒ Recherche directe du sous-espace de dimension  $K$  avec  $K > 1$

$$\begin{cases} \frac{\partial \mathcal{F}}{\partial u} = 0 & \Rightarrow u = X'F(F'F)^{-1} \\ \frac{\partial \mathcal{F}}{\partial F} = 0 & \Rightarrow F = Xu(u'u)^{-1} \end{cases}$$

## Moindres carrés pondérés

$$\mathcal{F} = \|W * (X - Fu')\|^2 = \sum_{i=1}^n \sum_{j=1}^p (w_{ij}x_{ij} - \sum_{k=1}^K F_{ik}w_{ij}u_{jk})^2,$$

avec  $W$  matrice de poids,  $w_{ij} = 0$  si  $x_{ij}$  manquant,  $w_{ij} = 1$  sinon.

⇒ Mêmes algorithmes mais on "saute" les données manquantes

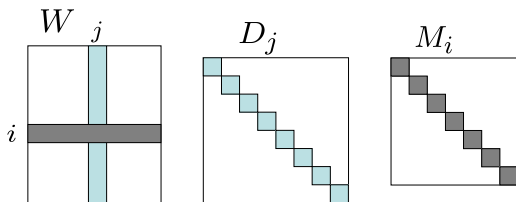
- NIPALS :  $u_{j1} = \frac{\sum_i (w_{ij}x_{ij}F_{i1})}{\sum_i w_{ij}F_{i1}^2}$ ;  $F_{i1} = \frac{\sum_j (w_{ij}x_{ij}u_{j1})}{\sum_j w_{ij}u_{j1}^2}$
- Recherche directe du sous-espace (Gabriel & Zamir, 1979)

## Interprétation géométrique

⇒ Données complètes :  $(X_{n \times p}, M, D); (\mathbb{R}^p, M); (\mathbb{R}^n, D)$

- NIPALS :  $u_1 = P_{F_1}^D(X) \Rightarrow u_{j1} = P_{F_1}^D(x_{.j})$   
 $F_1 = P_{u_1}^M(X') \Rightarrow F_{i1} = P_{u_1}^M(x_{i.})$
- Directe : idem projection sur  $F$  et  $u$

⇒ Données manquantes :



$$\Rightarrow u_{j1} = P_{F_1}^{D_j}(x_{.j})$$

$$\Rightarrow F_{i1} = P_{u_1}^{M_i}(x_{i.})$$

## ACP itérative

⇒ Kiers (1997) : minimiser  $\|W * (X - \mathcal{M})\|^2$  par minimisation itérative de  $\|X - \mathcal{M}\|^2$ .

- 1 initialisation  $\ell = 0 : X^0$
- 2 itération  $\ell$  :
  - (a)  $(F^\ell, u^\ell)$  minimisent (ou diminuent)  $\|X^{\ell-1} - Fu'\|^2$ ;  $K$  dimensions sont retenues
  - (b)  $\hat{X}^\ell = \hat{F}^\ell \hat{u}^{\ell'}$   $\Rightarrow X^\ell = W * X + (1 - W) * \hat{X}^\ell$
- 3 les étapes (a) et (b) sont répétées jusqu'à convergence

⇒ Estimation/Imputation

## ACP itérative = ACP-EM

Modèle :  $x_{ij} = \sum_{k=1}^K F_{ik} u_{jk} + \varepsilon_{ij}$ , avec  $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ .

Vraisemblance :  $L_c(X|F, u, \sigma^2) = -\frac{np}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|X - Fu'\|^2$ .

- Etape E :  $\mathbb{E}(x_{ij}|X_{obs}, F^\ell, u^\ell, \sigma^\ell) \Rightarrow$  Imputation
- Etape M : Maximise l'espérance de  $L_c \Rightarrow$  ACP
- Etape M' : Augmente l'espérance de  $L_c \Rightarrow$  MCA

$$\begin{aligned}\hat{u}^\ell &= X^{\ell-1'} F^{\ell-1} (F^{\ell-1'} F^{\ell-1})^{-1} \\ \hat{F}^\ell &= X^{\ell-1} u^\ell (u^{\ell'} u^\ell)^{-1}\end{aligned}$$

$\Rightarrow$  ACP-GEM

# Propriétés

Deux algorithmes = deux points de vues  
(imputation; "saute"  $\Rightarrow$  imputation implicite)

- Décentrage : recentrage
- Minima locaux : plusieurs solutions initiales
- Solutions non emboîtées : choix du nombre d'axes
- Réduction de la variabilité
  
- Surajustement
  - Diminuer  $K$
  - Pénaliser les régressions

## ACP Probabiliste (*Bishop & Tipping, 1999; Roweis, 1998*)

⇒ modèle d'analyse en facteurs communs et spécifiques

$$x = \mu + \Gamma z + \varepsilon, \text{ avec } z \sim \mathcal{N}(0, I_K), \varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$$

- Les variables sont indépendantes sachant les variables latentes :  $x|z \sim \mathcal{N}(\mu + \Gamma z, \sigma^2 I_p)$
- Décomposition de la matrice de variance-covariance :  $x \sim \mathcal{N}(\mu, \Sigma)$  avec  $\Sigma_{p \times p} = \Gamma_{p \times K} \Gamma'_{K \times p} + \sigma^2 I_p$
- Solution explicite :
  - $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$
  - $\hat{\sigma}^2 = \frac{1}{p-K} \sum_{i=K+1}^K \lambda_i$
  - $\hat{\Gamma} = u_K (\Lambda_K - \sigma^2 I_K)^{1/2}$

## ACP Probabiliste via l'algorithme EM

Vraisemblance complète :  $L_c(X, Z | \Gamma, \mu, \sigma^2)$

- Etape E : loi des variables latentes sachant les observations

$$z|x \sim \mathcal{N}((\Gamma'\Gamma + \sigma^2 I)^{-1} \Gamma'x, V)$$

⇒ L'espérance conditionnelle correspond à une régression ridge:

$$\hat{Z}' = (\hat{\Gamma}'\hat{\Gamma} + \hat{\sigma}^2 I)^{-1} \hat{\Gamma}'X'$$

- Etape M : maximise  $\mathbb{E}[L_c]$  par rapport à  $\Gamma$  et  $\sigma^2$

$$\hat{\Gamma}' = (\hat{Z}'\hat{Z} + n\hat{V})^{-1} \hat{Z}'X'$$

$$\hat{\sigma}^2 = \frac{1}{np} \|X - \hat{\Gamma}Z\|^2$$

# Utilisation de l'ACPP pour limiter le surajustement

- 1 initialisation  $\ell = 0 : X^0$
- 2 itération  $\ell$  :
  - (a) Estimation :  $\hat{Z}$  et  $\hat{\Gamma} \Rightarrow$  régressions régularisées
  - (b) Imputation  $\hat{X}_{acpp}^K = \hat{Z}\hat{\Gamma}' \simeq \sum_{k=1}^K \frac{\lambda_k - \sigma^2}{\lambda_k} \hat{X}_{acp}^k$
- 3 les étapes (a) et (b) sont répétées jusqu'à convergence

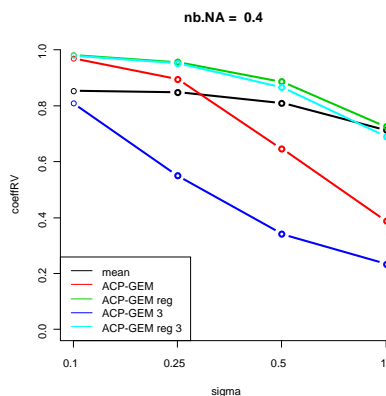
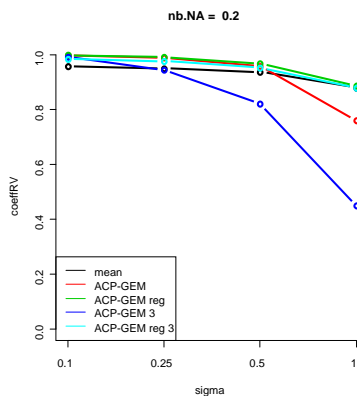
$\Rightarrow$  ACP-GEM régularisée

## Simulations (1)

- 21 individus, 10 variables
- ACP et données reconstituées avec 2 dimensions :  $\hat{X} = \hat{F}\hat{u}'$
- Ajout de bruit sur chaque variable ( $\sigma = 0.1, 0.2, 0.5, 1$ )  $\Rightarrow$  différentes structures de données
- Nombre de données manquantes : 10%, 20% 50%
- 100 simulations pour chaque jeu de paramètres
  
- ACP construite sur données complètes
- ACP construite sur données incomplètes avec les différents algorithmes : imputation par la moyenne, NIPALS, ACP-GEM et ACP-GEM régularisée

## Simulations (2)

- Deux critères :
  - Erreur de reconstitution
  - Coefficient  $RV$  entre configurations (vraie / incomplète)



# Conclusion

- Réduction de variabilité
- Visualisation de l'incertitude due aux données manquantes
- Bootstrap ou imputation multiple adapté à l'ACP

Du 7 au 10 juillet 2009 :



<http://www.agrocampus-ouest.fr/math/useR-2009>