

Données manquantes en Analyse Factorielle Multiple

Julie Josse, François Husson, Jérôme Pagès

Laboratoire de mathématiques appliquées, Agrocampus Rennes

Nancy, 13 juin 2008

Plan

- 1 Contexte de l'étude
- 2 Données manquantes en ACP
- 3 Données manquantes en AFM
- 4 Simulations, résultats
- 5 Conclusion et perspectives

Plan

- 1 Contexte
- 2 Données manquantes en ACP
 - ACP données complètes
 - ACP données manquantes
- 3 Données manquantes en AFM
 - AFM itérative
 - Minimisation de distance entre tableaux
- 4 Simulations
- 5 Conclusion
 - Conclusion
 - Perspectives : ACP probabiliste

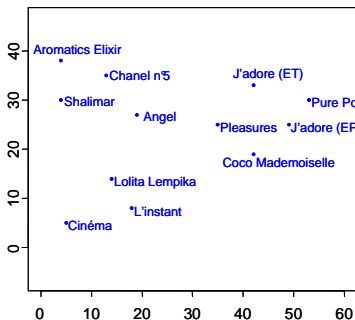
Les données

- Evaluation sensorielle :

99 juges, 12 parfums



- Recueil de données par Napping (*Pagès 2003*)



	X	Y
Angel	19	27
Chanel	13	35
Coco	42	19
...
Shalimar	4	30

Problématique sensorielle

	1	p_1				1	p_J
1	i	X_1				X_J	
n							

Figure: Tableaux multiples.

- Obtenir une carte compromis des parfums
- Est-ce que les juges perçoivent les parfums de la même façon (consensus entre juges) ?
- Confronter l'espace produit des experts à celui des consommateurs

⇒ Problématique classique en analyse multi-tableaux : Analyse Factorielle Multiple (Escofier & Pagès, 1982)

Une configuration de données manquantes particulière

- Problème : chaque juge ne peut évaluer qu'un petit nombre de produits → difficulté et saturation
- Construire un plan d'expériences

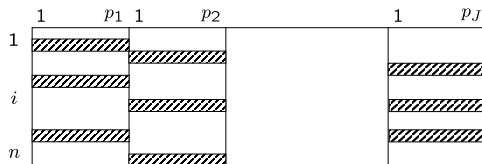


Figure: Structure de données manquantes.

⇒ **But** : obtenir la configuration compromis à partir du tableau de données incomplet (par une AFM avec données manquantes).

Comment aborder ce problème de données manquantes ?

- Contexte d'analyse de données
- Structure des données particulière
- Structure spécifique des données manquantes

- Estimation (hypothèses, distributions) ?
- Imputation ?

Plan

- 1 Contexte
- 2 Données manquantes en ACP
 - ACP données complètes
 - ACP données manquantes
- 3 Données manquantes en AFM
 - AFM itérative
 - Minimisation de distance entre tableaux
- 4 Simulations
- 5 Conclusion
 - Conclusion
 - Perspectives : ACP probabiliste

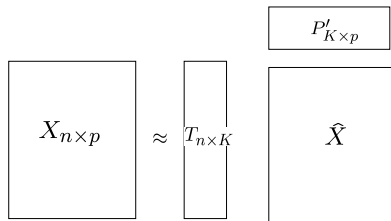


ACP

- Deux points de vues :
 - Maximisation de la variance des points projetés
 - Minimisation de l'erreur de reconstitution

 - Diagonalisation de la matrice de variance covariance (ou de produit scalaire)
- ⇒ Coûteux quand n et p sont très grands simultanément

Minimiser l'erreur de reconstitution



$$\begin{aligned}
 \mathcal{F} &= \|X_{n \times p} - T_{n \times K} P'_{K \times p}\|^2 \\
 &= \sum_i \sum_j (x_{ij} - \sum_{k=1}^K t_{ik} p_{jk})^2.
 \end{aligned}$$

\Rightarrow meilleure approximation de X par une matrice de rang inférieur K au sens des moindres carrés (Eckart-Young)

ACP via NIPALS

- Wold (1966, 1969) : Non linear Iterative PArTial Least Squares
- Meilleure approximation de rang 1 :

$$\mathcal{F}_1 = \sum_i \sum_j (x_{ij} - t_{i1}p_{j1})^2.$$

⇒ On annule les dérivées partielles :

$$\begin{cases} \frac{\partial \mathcal{F}_1}{\partial t_i} = 0 \rightarrow p_j = \frac{\sum_i (x_{ij} \times t_i)}{\sum_i t_i^2}, \\ \frac{\partial \mathcal{F}_1}{\partial p_j} = 0 \rightarrow t_i = \frac{\sum_j (x_{ij} \times p_j)}{\sum_j p_j^2}. \end{cases}$$

⇒ Méthode séquentielle : une fois (\hat{t}_1, \hat{p}_1) trouvé, on cherche (\hat{t}_2, \hat{p}_2) premier axe et première composante de $\tilde{X} = X - \hat{t}_1 \hat{p}'_1$.

Algorithme NIPALS (Données complètes)

- Soit t_1 une variable (au hasard) de \mathbb{R}^n (une composante principale), soit p_1 un vecteur de \mathbb{R}^p (un axe).
- Tant que continue :
 - Pour $j = 1, \dots, p$,

$$p_1[j] = \left\langle x_j, \frac{t_1}{\|t_1\|^2} \right\rangle .$$

- On normalise p_1

$$p_1 = \frac{p_1}{\|p_1\|} .$$

- Pour $i = 1, \dots, n$,

$$t_1[i] = \left\langle x_i, \frac{p_1}{\|p_1\|^2} \right\rangle .$$

- Stop

Algorithme NIPALS (Données complètes)

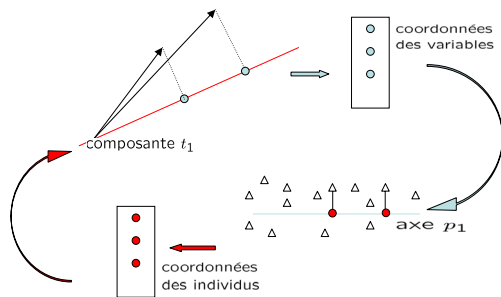


Figure: Nipals.

Algorithme de diagonalisation

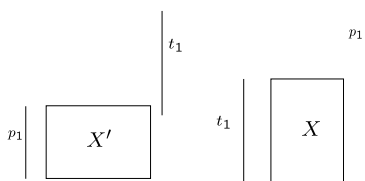


Figure: Puissance itérée.

A convergence, les vecteurs t_1 et p_1 vérifient les équations :

$$\begin{cases} X'Xp_1 = \lambda_1 p_1, \\ XX't_1 = \lambda_1 t_1. \end{cases}$$

NIPALS (Données complètes)

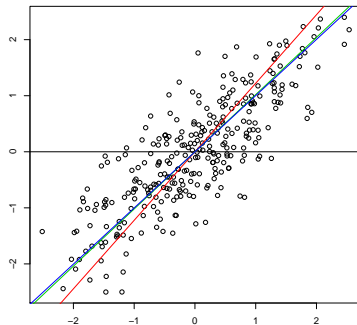


Figure: Recherche du premier axe.

ACP via Alternating Least Squares

$$\mathcal{F} = \|X_{n \times p} - T_{n \times K} P'_{K \times p}\|^2.$$

⇒ Alternating Least Squares (ALS)

$$\begin{cases} \frac{\partial \mathcal{F}}{\partial P} = -2X'T + 2PT'T = 0 & \Rightarrow P = X'T(T'T)^{-1}, \\ \frac{\partial \mathcal{F}}{\partial T} = -2XP + 2TP'P = 0 & \Rightarrow T = XP(P'P)^{-1}. \end{cases}$$

Et avec des données manquantes ?

- Méthodes ad-hoc : méthodes d'imputation (moyenne, régression, etc.)
- Weighted low rank approximation : critère WPCA

$$\mathcal{F} = \|W * (X - TP')\|^2 = \sum_i \sum_j (w_{ij}x_{ij} - \sum_k t_{ik}w_{ij}p_{jk})^2,$$

avec W une matrice de poids, $w_{ij} = 0$ si x_{ij} est manquant et $w_{ij} = 1$ sinon et $*$ le produit d'Hadamard (élément par élément).

⇒ T et P obtenus en minimisant le critère sur les données présentes seulement

⇒ Les données manquantes sont complétées par $\hat{X} = \hat{T}\hat{P}'$

⇒ Méthodes itératives pour résoudre ce problème

Et avec des données manquantes ?

⇒ Mêmes algorithmes que précédemment mais on saute les données manquantes.

- Nipals : Christoffersen (1970); Gabriel et Zamir (1979) : Criss Cross regression. Alternance de deux régressions simples pondérées :

$$\begin{cases} p_j = \frac{\sum_i w_{ij} x_{ij} t_i}{\sum_i w_{ij} t_i^2}, \\ t_i = \frac{\sum_j w_{ij} x_{ij} p_j}{\sum_j w_{ij} p_j^2}. \end{cases}$$

- Weighted Alternating Least Squares : alternance de deux régressions multiples pondérées. Gabriel et Zamir (1979) : Criss Cross Multiple regression; Kiers (1997) : Weighted Least Squares (WLS).

ACP itérative

- 1 Initialisation $i = 0$: remplacer les valeurs manquantes par la moyenne de chaque variable (par exemple) : X^0 ;
- 2 Réalisation de l'ACP (SVD ou une étape de ALS) sur le tableau de données complété X^i (on retient K axes)
- 3 On obtient X^{i+1} en imputant les valeurs manquantes dans X^i par les valeurs reconstituées : $\hat{X} = \hat{T}\hat{P}'$
- 4 $i \leftarrow i + 1$ et on répète les étapes 2 et 3 jusqu'à convergence

⇒ A convergence, les données imputées n'influent pas sur la construction des axes.

Re-centrer (Re-réduire)

⇒ La moyenne (l'écart-type) doit être ré-estimée avec les composantes et les axes dans la minimisation du critère WPCA.

- A l'étape d'initialisation, on centre les données par $M^{(0)}$
- A l'itération i :
 - ACP sur le tableau de données complété centré
 - Imputation des données manquantes → les données ne sont plus centrées
 - Rajout de la moyenne $M^{(i-1)}$ et calcul de la nouvelle moyenne $M^{(i)}$
 - Re-centrage des données (on enlève $M^{(i)}$)
- A la fin, les données imputées sont obtenues en rajoutant $M^{(l)}$

⇒ Même procédure pour la réduction

ACP itérative; WLS : propriétés

- Problème de convergence : autres initialisations possibles
- Solutions non emboîtées → choix du nombre d'axes ? (Wold, validation croisée)
- Augmente les liaisons entre les variables (optimiste) $\hat{X} = \hat{T}\hat{P}'$.

Bilan

- Données complètes : toutes les méthodes permettent de réaliser une ACP à moindre coût (calcul de K axes)
- Données manquantes : idée simple qui consiste à "sauter" les données manquantes pour obtenir les axes et les composantes avec les données présentes (poids nul aux données manquantes).
 - Nipals
 - WLS, ACP itérative, ALS itérative : différences algorithmiques.

Algorithme aussi appelé EM-PCA

(Expectation-Maximisation) : initialisation des paramètres θ_0

→ on estime les données manquantes \hat{X}_{miss} → avec

$(\hat{X}_{miss}, X_{obs})$ on obtient θ_1 → on réestime \hat{X}_{miss} , et on itère.

Plan

- 1 Contexte
- 2 Données manquantes en ACP
 - ACP données complètes
 - ACP données manquantes
- 3 Données manquantes en AFM
 - AFM itérative
 - Minimisation de distance entre tableaux
- 4 Simulations
- 5 Conclusion
 - Conclusion
 - Perspectives : ACP probabiliste

AFM (1)

	1	p_1				1	p_J
1	X_1					X_J	
i							
n							

Figure: Tableaux multiples.

- Questionnaire : santé des étudiants (consommation de drogues, état psychologique, qualité du sommeil, signalétique)
- Analyse sensorielle : données sensorielles, mesures physico-chimiques
- Données génomiques : protéine, ADN

AFM (2)

- Equilibrer l'influence des groupes
- Le cœur de l'AFM est une ACP pondérée :
 - ACP sur chaque groupe de variables $\rightarrow \lambda_1^1, \dots, \lambda_1^J$
 - ACP globale sur : $(\frac{X_1}{\sqrt{\lambda_1^1}}, \dots, \frac{X_J}{\sqrt{\lambda_1^J}})$
- Résultats classiques de l'analyse factorielle (représentation des individus, représentation des variables)
- Problématique enrichie : résultats spécifiques de la structure en groupes de variables (représentation superposée, représentation globale des groupes)

Données manquantes en AFM

⇒ Avec deux groupes de variables :

- AFM itérative /Weighted AFM/ EM-AFM
- Minimisation de distance entre tableaux :
 - Imputation par maximisation du coefficient R^2
 - Imputation avec des rotations procustes;



Algorithme de l'AFM itérative

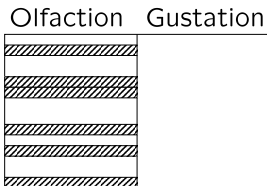
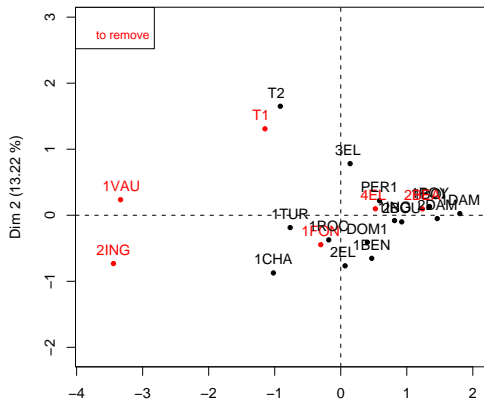
⇒ Pondération comme la réduction (ou le centrage) : une étape de re-pondération après imputation des données

- Initialisation : imputation par la moyenne $\rightarrow X^0; i = 0$
- Calcul des poids $\Lambda^{(0)}$ puis pondération ($X = X/\sqrt{\Lambda^{(0)}}$)
- Soit $T_{n \times K}$ une matrice au hasard
- Iteration i :
 - $P = X' T (T' T)^{-1}; T = X P (P' P)^{-1}$
 - $\hat{X} = \hat{T} \hat{P}'$
 - Multiplication par $\Lambda^{(i-1)}$ et calcul des nouveaux poids $\Lambda^{(i)}$
 - Calcul du tableau global ($X = X/\sqrt{\Lambda^{(i)}}$)
- A la fin les données imputées sont obtenues en multipliant par $\Lambda^{(l)}$

Exemple

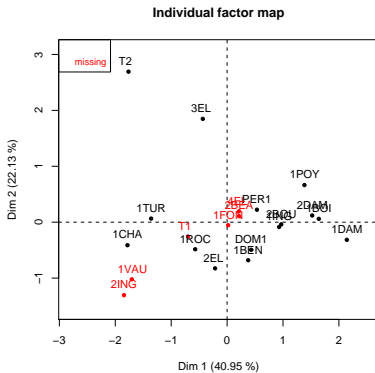
- Données vins (FactoMineR) : 21 vins décrits par deux groupes de variables (olfaction/ gustation) → 6 lignes manquantes dans olfaction

Individual factor map





Exemple



Le coefficient RV , Escoufier 1973 (1)

⇒ Un coefficient de corrélation entre deux ensembles de variables.

- Soit $X_{n \times p_1}$ et $Y_{n \times p_2}$, si X et Y sont centrés par colonnes, le coefficient RV est défini par :

$$RV(X, Y) = \frac{\text{tr}(XX'YY')}{\sqrt{\text{tr}(XX')^2 \text{tr}(YY')^2}} = \frac{\langle W_X, W_Y \rangle}{\|W_X\| \|W_Y\|}.$$

- Propriétés :
 - $0 \leq RV(X, Y) \leq 1$
 - $RV(X, Y) = 0$, si et seulement si $X^T Y = 0$
 - $RV(X, XB + \mathbb{1}_n c) = 1$, $B_{p_1 \times p_2}$ une matrice orthogonale ($BB' = I_{p_1}$) et $c_{1 \times p_2}$ un vecteur constant
 - si $p_1 = p_2 = 1$, $RV(X, Y) = r^2(X, Y)$

Le coefficient RV , Escoufier 1973 (2)

- Dissimilarité entre deux matrices :

$$\begin{aligned} d(X, Y) &= \left\| \frac{XX'}{(tr(XX'))^{1/2}} - \frac{YY'}{(tr(YY'))^{1/2}} \right\|, \\ &= \sqrt{2} \sqrt{1 - RV(X, Y)}. \end{aligned}$$

- La proximité entre matrices indique que la position relative des n points dans \mathbb{R}^{p_1} et des n points dans \mathbb{R}^{p_2} est similaire

⇒ Minimisation d'une dissimilarité entre nuage de points = maximiser le RV

Maximisation du coefficient RV

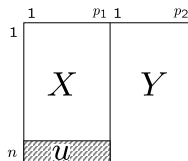


Figure: Imputation avec le coefficient RV .

- On cherche u qui maximise $RV(u)$
- Si r individus ont des données manquantes, on traite séparément chacun des r individus incomplets avec les $n - r$ individus complets
- Problème de maximisation si p_1 grand

Analyse procustéenne

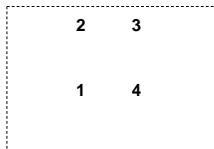
L'analyse procustéenne consiste à ajuster un nuage de points sur un autre représentant les mêmes individus par des rotations, réflexions, translations et dilatations (Gower, 1971)

⇒ Ajustement par moindres carrés

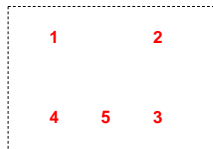
Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration



Complete configuration



Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration

2	3
1	4

Complete configuration

1		2
4	5	3

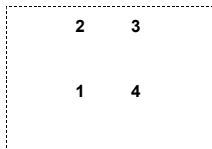
2	3
	x
1	4

1		2
	x	
4	5	3

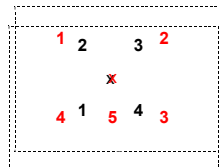
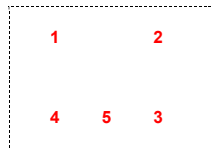
Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration



Complete configuration

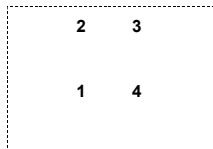


Translation

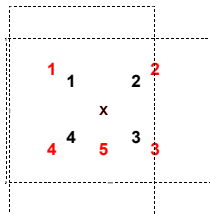
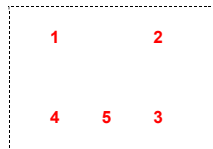
Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration



Complete configuration

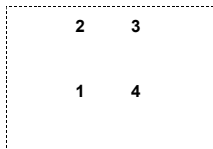


Rotation

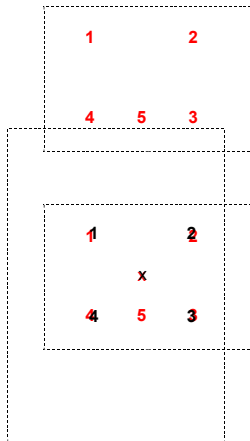
Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration



Complete configuration



Dilatation

Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration

2	3
1	4

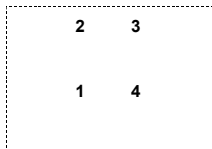
Complete configuration

1		2
4	5	3
1		2
	x	
4	5	3

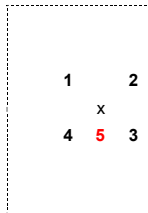
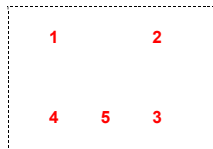
Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration



Complete configuration

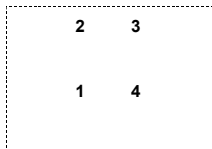


Inverse dilatation

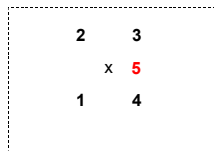
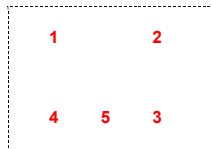
Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration



Complete configuration



Inverse rotation

Rotations procustes pour gérer les données manquantes

⇒ Compléter, imputer une configuration en s'aidant d'une autre

Incomplete configuration

2	3
1	4

Complete configuration

1	2	
4	5	3

2	3
	x 5
1	4

Inverse translation

Plan

- 1 Contexte
- 2 Données manquantes en ACP
 - ACP données complètes
 - ACP données manquantes
- 3 Données manquantes en AFM
 - AFM itérative
 - Minimisation de distance entre tableaux
- 4 Simulations
- 5 Conclusion
 - Conclusion
 - Perspectives : ACP probabiliste

Simulations (1)

- 21 individus
- Deux groupes de 9 variables assez liés
- AFM effectuée puis données reconstituées avec 2 axes :
 $\hat{X} = \hat{T}\hat{P}' \rightarrow 2$ dimensions.
- Ajout de bruit sur chaque variable
($\sigma = 0.025, 0.05, 0.1, 0.2, 0.5, 1$) \rightarrow différentes structures de données
- Nombre de données manquantes : 10%, 20% 50% des lignes d'un groupe
- 100 simulations pour chaque jeu de paramètres

Simulations (2)

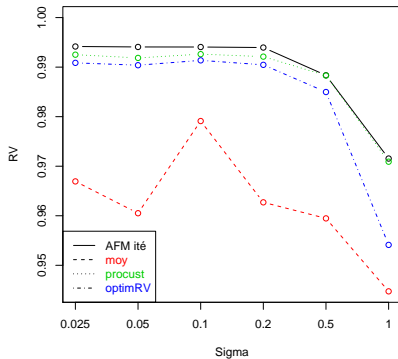
- AFM construite sur données complètes
- AFM construite sur données incomplètes avec les différents algorithmes

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

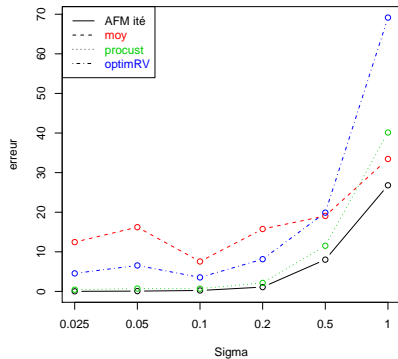


Resultats (1)

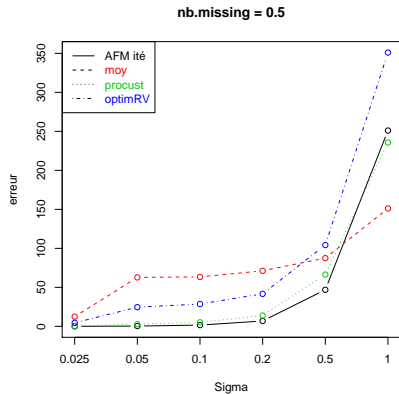
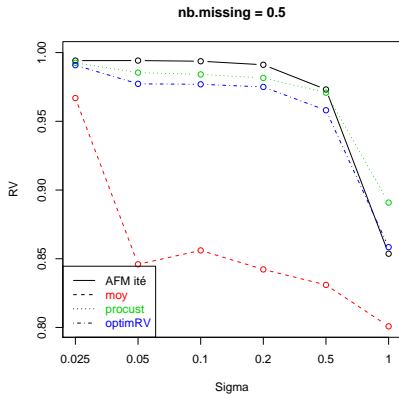
nb.missing = 0.1



nb.missing = 0.1



Resultats (2)



Plan

- 1 Contexte
- 2 Données manquantes en ACP
 - ACP données complètes
 - ACP données manquantes
- 3 Données manquantes en AFM
 - AFM itérative
 - Minimisation de distance entre tableaux
- 4 Simulations
- 5 Conclusion
 - Conclusion
 - Perspectives : ACP probabiliste

J configurations avec des données manquantes

- AFM
 - Prise en compte simultanément de toutes les configurations
- Rotations procustes, maximisation du RV : deux par deux
 - Pour chaque ligne manquante $\rightarrow J - 1$ estimations possibles
 - Moyenne, médiane (si outliers), sélection ?
- Besoin de sélectionner ? \rightarrow test sur la significativité de la liaison entre deux matrices (Josse, Husson, Pagès, 2008)

Conclusion

- AFM itérative
 - Estimation simultanée des données manquantes et des paramètres
 - Bon résultats

 - Choix du nombre de composantes ?
 - Surajustement : pénalisation ?
 - Augmente les corrélations entre variables

- Minimisation de distance
 - Augmente les corrélations entre variables
 - Plus de deux configurations ?

Perspectives

- Kernel matrix completion (Kernel matrix regression,...)
- Choix du nombre d'axes
- Etudier les propriétés des algorithmes
- ACP probabiliste



ACP probabiliste (*Bishop and Tipping, Roweis 1998*) : modèle d'analyse en facteurs communs et spécifiques

$$x = \mu + \Gamma z + \varepsilon.$$

- x (p); z les variables latentes (q); $q \ll p$
- $z \sim \mathcal{N}(0, I_q)$, $\varepsilon \sim \mathcal{N}(0, \Psi)$, Ψ diagonale

$$p(x|z) \sim \mathcal{N}(\mu + \Gamma z, \Psi).$$

⇒ Les variables sont indépendantes sachant les variables latentes.

$$p(x) \sim \mathcal{N}(\mu, \Sigma) \text{ avec } \Sigma_{p \times p} = \Gamma_{p \times q} \Gamma'_{q \times p} + \Psi_{p \times p}.$$

⇒ Décomposition de la matrice de variance-covariance particulière.

⇒ ACP probabiliste : $\Psi = \sigma^2 I_p$



Réduction de la dimension et estimation des paramètres

- Loi des variables cachées sachant les observations :

$$p(z|x) \sim \mathcal{N}(M^{-1}\Gamma'(x - \mu), \sigma^2(M)^{-1}),$$

avec $M = \Gamma'\Gamma + \sigma^2I$.

- Estimation des paramètres : maximum de vraisemblance

$$\ln L(X|\Gamma, \mu, \sigma^2) = -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Sigma|) - \frac{1}{2} \text{trace}(n\Sigma^{-1}S).$$



Réduction de la dimension et estimation des paramètres

⇒ Solution explicite :

- $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$: la moyennes des données,
- $\hat{\sigma}^2 = \frac{1}{p-q} \sum_{i=q+1}^p \lambda_i$: la variance perdue en projection,
- $\hat{\Gamma} = U_q(\Lambda_q - \sigma^2 I_q)^{1/2} R,$

⇒ ACP cas particulier de l'ACPP :

- On fait tendre le bruit vers 0 : $\sigma \rightarrow 0$
- Réduction de la dimension : $\mathbb{E}[z|x] = (\Gamma' \Gamma)^{-1} \Gamma' X$

⇒ Algorithme Expectation-Maximisation



Probabilistic PCA : EM

- Vraisemblance complète des observations : $\ln L(X, Z | \Gamma, \mu, \sigma^2)$
- E-step : calcul de l'espérance de la vraisemblance complète en utilisant l'espérance et la variance de $p(z|x)$;
 - $\langle z_i \rangle = M^{-1} \Gamma' (x_i - \bar{x})$
 - $\langle z_i z_i' \rangle = \sigma^2 (M)^{-1} + \langle z_i \rangle \langle z_i' \rangle$
- M-step : maximise $\ln L$ par rapport aux paramètres Γ et σ^2 :

$$\Gamma_{new} = \left[\sum_{i=1}^n (x_i - \bar{x}) \langle z_i \rangle \right] \left[\sum_{i=1}^n \langle z_i z_i' \rangle \right]^{-1}$$

$$\sigma_{new}^2 = \left\{ \frac{1}{np} \sum_{i=1}^n \|x_i - \bar{x}\|^2 - 2 \langle z_i \rangle \Gamma'_{new} (x_i - \bar{x}) + \text{Tr}(\langle z_i z_i' \rangle \Gamma'_{new} \Gamma_{new}) \right\}$$

- Itération jusqu'à convergence



ACP : EM

- E-step : $\hat{Z} = (\Gamma'\Gamma)^{-1}\Gamma'X$
- M-step : $\hat{\Gamma} = X'Z(ZZ')^{-1}$

⇒ Alternative Least Squares

Données manquantes ACPP et ACP

- 1 Remplacement des données manquantes par la moyenne de chaque variable
- 2 On effectue les deux étapes : E et M
- 3 Imputation des données manquantes par $\hat{X} = \hat{Z}\hat{\Gamma}'$
- 4 Itération des étapes 2 et 3 jusqu'à convergence

- Avec 1 axe sans données manquantes : $\hat{X}_{acpp} = \hat{X}_{acp} \frac{\lambda_1 - \sigma^2}{\lambda_1}$
- Package pcaMethods (Bioconductor)

Bibliographie

- Schafer & Graham (2002). Missing data the view of the state of the art, *Psychological Methods*.
- Little & Rubin (1987, 2002). Statistical analysis with missing data *Wiley*.
- Kiers (1997). Weighted least squares. *Psychometrica*.
- Grung & Manne (1998). Missing values in Principal Component Analysis *Chem.Intell.Lab.Sys*.
- Gabriel & Zamir (1979). Lower Rank Approximation of Matrices by Least Squares with Any Choice of Weights, *Technometrics*.
- Bishop, Tipping (1999). Probabilistic Principal Component Analysis, *JRSS*.
- Roweis (1998). EM algorithms for PCA and Sensible PCA, *Advances in Neural Information Processing Systems*.
- Crettaz de rotten & Helbling (1991). Une estimation de données manquantes basée sur le coefficient RV. *RSA*.

Ces analyses sont incluses dans

FACTOMINER **SensominER**

librairies R dédiées à

l'Analyse des données

la sensométrie

écrit par

Laboratoire de mathématiques appliquées d'Agrocampus

Rennes France



<http://factominer.free.fr>

<http://sensominer.free.fr>





8 au 10 juillet 2009



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