

# Analyse de $(K + 1)$ tableaux avec le logiciel ade4

## Application en épidémiologie

S. Bougeard<sup>(1)</sup> & S. Dray<sup>(2)</sup>

(1) *Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail, Unité d'Epidémiologie*

(2) *Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive*



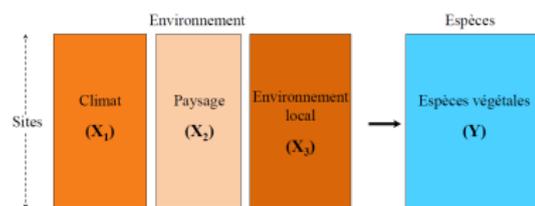
# Plan de la présentation

- 1 Contexte
  - Données multiblocs
  - Outils disponibles sur le logiciel R
  
- 2 Méthodes
  - Exemple de données
  - Méthodes multiblocs
  
- 3 Application des méthodes multiblocs avec ade4
  - Import et calculs
  - Interprétation descriptive
  - Interprétation explicative
  
- 4 Conclusion & perspectives

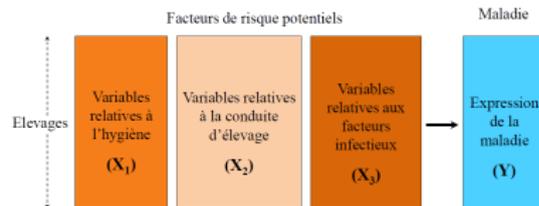
# Données organisées en $(K + 1)$ blocs de variables

## Données multiblocs

- Nombreuses variables explicatives organisées en  $K$  blocs  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ ,
- Plusieurs variables à expliquer  $\mathbf{Y}$ ,
- Données rencontrées dans de nombreux domaines.



(a) Données d'écologie



(b) Données d'épidémiologie vétérinaire

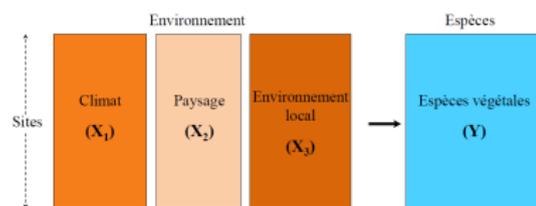
## Objectifs du traitement statistique

- Descriptif : décrire des liens entre variables explicatives  $\mathbf{X}$  et à expliquer  $\mathbf{Y}$ ,
- Explicatif : déterminer les variables  $\mathbf{X}$  et les blocs  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$  ayant une influence significative dans l'explication de  $\mathbf{Y}$ .

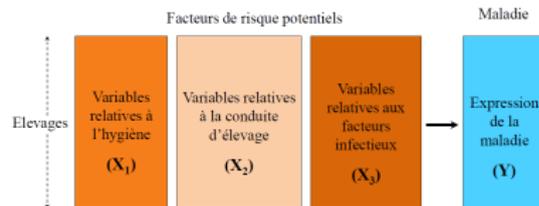
## Données organisées en $(K + 1)$ blocs de variables

### Données multiblocs

- Nombreuses variables explicatives organisées en  $K$  blocs  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ ,
- Plusieurs variables à expliquer  $\mathbf{Y}$ ,
- Données rencontrées dans de nombreux domaines.



(a) Données d'écologie



(b) Données d'épidémiologie vétérinaire

### Objectifs du traitement statistique

- Descriptif : décrire des liens entre variables explicatives  $\mathbf{X}$  et à expliquer  $\mathbf{Y}$ ,
- Explicatif : déterminer les variables  $\mathbf{X}$  et les blocs  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$  ayant une influence significative dans l'explication de  $\mathbf{Y}$ .

## Outils disponibles sur le logiciel R

### Méthodes trop simples ...

- Méthodes développées pour deux blocs de variables **X** et **Y**, *e.g.* analyse canonique, régression PLS, ACP sur variables instrumentales,
- Packages associés : stats (fonction `cancor`), pls, ade4 (fonction `pcaiv`),
- Limite : ignorance de la structure des variables **X** en blocs.

### Méthodes trop compliquées ...

- Méthodes développées pour un grand nombre de blocs de variables pouvant être reliés entre eux, *e.g.* PLS Path Modeling,
- Package associé : `pls`, `plspm`,
- Limite : algorithme itératif dont la convergence n'est pas prouvée.

→ Objectif : Proposer des méthodes multiblocs pour le traitement de données organisées en  $(K + 1)$  tableaux sur le logiciel R (intégration au package `ade4`).

## Outils disponibles sur le logiciel R

### Méthodes trop simples ...

- Méthodes développées pour deux blocs de variables **X** et **Y**, *e.g.* analyse canonique, régression PLS, ACP sur variables instrumentales,
- Packages associés : `stats` (fonction `cancor`), `pls`, `ade4` (fonction `pcaiv`),
- Limite : ignorance de la structure des variables **X** en blocs.

### Méthodes trop compliquées ...

- Méthodes développées pour un grand nombre de blocs de variables pouvant être reliés entre eux, *e.g.* PLS Path Modeling,
- Package associé : `pls`, `plspm`,
- Limite : algorithme itératif dont la convergence n'est pas prouvée.

→ Objectif : Proposer des méthodes multiblocs pour le traitement de données organisées en  $(K + 1)$  tableaux sur le logiciel R (intégration au package `ade4`).

## Outils disponibles sur le logiciel R

### Méthodes trop simples ...

- Méthodes développées pour deux blocs de variables **X** et **Y**, *e.g.* analyse canonique, régression PLS, ACP sur variables instrumentales,
- Packages associés : `stats` (fonction `cancor`), `pls`, `ade4` (fonction `pcaiv`),
- Limite : ignorance de la structure des variables **X** en blocs.

### Méthodes trop compliquées ...

- Méthodes développées pour un grand nombre de blocs de variables pouvant être reliés entre eux, *e.g.* PLS Path Modeling,
- Package associé : `pls`, `plspm`,
- Limite : algorithme itératif dont la convergence n'est pas prouvée.

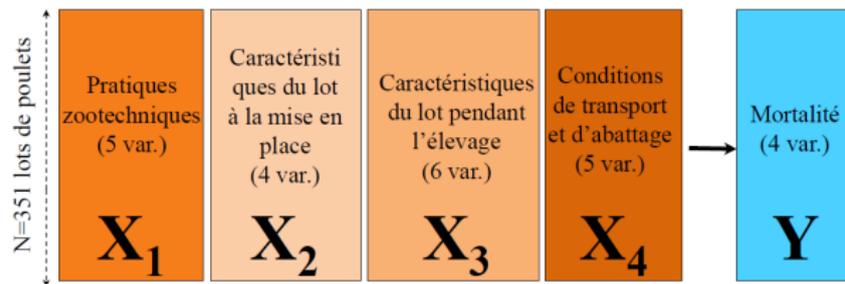
→ Objectif : Proposer des méthodes multiblocs pour le traitement de données organisées en  $(K + 1)$  tableaux sur le logiciel R (intégration au package `ade4`).

# Plan de la présentation

- 1 Contexte
  - Données multiblocs
  - Outils disponibles sur le logiciel R
- 2 Méthodes
  - Exemple de données
  - Méthodes multiblocs
- 3 Application des méthodes multiblocs avec ade4
  - Import et calculs
  - Interprétation descriptive
  - Interprétation explicative
- 4 Conclusion & perspectives

# Données d'épidémiologie vétérinaire

Enquête visant à déterminer les facteurs de risques des mortalités dans la filière poulet de chair [Lupo, 2008]



## Données

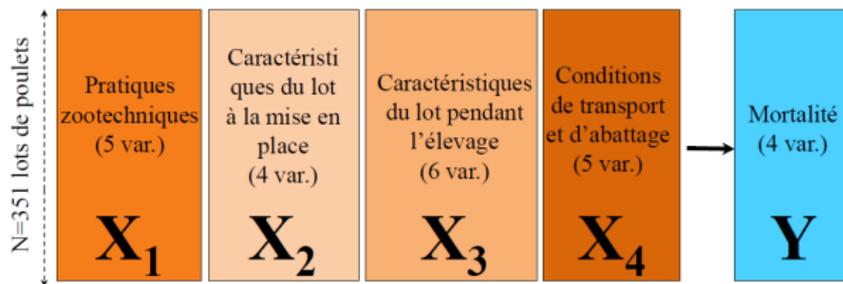
- **Y**: mortalité (4 var.),
- **X**: 20 facteurs de risques potentiels organisés en 4 blocs (*i.e.*, les étapes de production).

## Objectifs du traitement statistique

- **Descriptif** : Décrire les liens entre les variables **X**, les étapes de production ( $X_1, \dots, X_4$ ) et la mortalité **Y**,
- **Explicatif** : Déterminer les facteurs de risques de la mortalité **Y** au niveau des variables **X** et des étapes de production ( $X_1, \dots, X_4$ ).

# Données d'épidémiologie vétérinaire

Enquête visant à déterminer les facteurs de risques des mortalités dans la filière poulet de chair [Lupo, 2008]



## Données

- **Y**: mortalité (4 var.),
- **X**: 20 facteurs de risques potentiels organisés en 4 blocs (*i.e.*, les étapes de production).

## Objectifs du traitement statistique

- Descriptif : Décrire les liens entre les variables **X**, les étapes de production ( $X_1, \dots, X_4$ ) et la mortalité **Y**,
- Explicatif : Déterminer les facteurs de risques de la mortalité **Y** au niveau des variables **X** et des étapes de production ( $X_1, \dots, X_4$ ).

# Méthodes multiblocs développées sous R [1]

ACP sur variables instrumentales multibloc [Bougeard et al., 2011] et PLS multibloc [Wold, 1984]

## Analyse de $(K + 1)$ triplets

- $(K + 1)$  triplets à analyser :  $(\mathbf{X}_k, \mathbf{Q}_{\mathbf{X}_k}, \mathbf{D})$  pour  $k = (1, \dots, K)$  et  $(\mathbf{Y}, \mathbf{Q}_{\mathbf{Y}}, \mathbf{D})$
- Analyse du triplet synthétique :  $(\sum_k \mathbf{X}'_k \mathbf{D} \mathbf{Y}, \mathbf{Q}_{\mathbf{Y}}, \mathbf{Q}_{\mathbf{X}})$  avec  $\mathbf{Q}_{\mathbf{Y}} = \mathbf{I}$ ,  $\mathbf{D} = \frac{1}{N} \mathbf{I}$ 
  - $\mathbf{Q}_{\mathbf{X}} = \mathbf{I}$  pour mbpls
  - $\mathbf{Q}_{\mathbf{X}}$  matrice bloc-diagonale avec pour éléments diagonaux  $(\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1}$  pour mbpcaiv

## Problème de maximisation

$\lambda^{(1)} = \mathbf{v}^{(1)'} (\sum_k \mathbf{Q}'_{\mathbf{Y}} \mathbf{Y}' \mathbf{D} \mathbf{X}_k \mathbf{Q}_{\mathbf{X}_k} \mathbf{X}'_k \mathbf{D} \mathbf{Y} \mathbf{Q}_{\mathbf{Y}}) \mathbf{v}^{(1)}$  sous la contrainte  $\|\mathbf{v}^{(1)}\| = 1$

- $\lambda^{(1)}$  première valeur propre,
- $\mathbf{u}^{(1)} = \mathbf{Y} \mathbf{v}^{(1)}$  variable latente associée au tableau  $\mathbf{Y}$ ,
- $\mathbf{Q}_{\mathbf{X}_k} = \mathbf{I}$  pour mbpls et  $\mathbf{Q}_{\mathbf{X}_k} = (\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1}$  pour mbpcaiv.

# Méthodes multiblocs développées sous R [1]

ACP sur variables instrumentales multibloc [Bougeard et al., 2011] et PLS multibloc [Wold, 1984]

## Analyse de $(K + 1)$ triplets

- $(K + 1)$  triplets à analyser :  $(\mathbf{X}_k, \mathbf{Q}_{\mathbf{X}_k}, \mathbf{D})$  pour  $k = (1, \dots, K)$  et  $(\mathbf{Y}, \mathbf{Q}_{\mathbf{Y}}, \mathbf{D})$
- Analyse du triplet synthétique :  $(\sum_k \mathbf{X}'_k \mathbf{D} \mathbf{Y}, \mathbf{Q}_{\mathbf{Y}}, \mathbf{Q}_{\mathbf{X}})$  avec  $\mathbf{Q}_{\mathbf{Y}} = \mathbf{I}$ ,  $\mathbf{D} = \frac{1}{N} \mathbf{I}$ 
  - $\mathbf{Q}_{\mathbf{X}} = \mathbf{I}$  pour mbpls
  - $\mathbf{Q}_{\mathbf{X}}$  matrice bloc-diagonale avec pour éléments diagonaux  $(\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1}$  pour mbpcaiv

## Problème de maximisation

$\lambda^{(1)} = \mathbf{v}^{(1)'} (\sum_k \mathbf{Q}'_{\mathbf{Y}} \mathbf{Y}' \mathbf{D} \mathbf{X}_k \mathbf{Q}_{\mathbf{X}_k} \mathbf{X}'_k \mathbf{D} \mathbf{Y} \mathbf{Q}_{\mathbf{Y}}) \mathbf{v}^{(1)}$  sous la contrainte  $\|\mathbf{v}^{(1)}\| = 1$

- $\lambda^{(1)}$  première valeur propre,
- $\mathbf{u}^{(1)} = \mathbf{Y} \mathbf{v}^{(1)}$  variable latente associée au tableau  $\mathbf{Y}$ ,
- $\mathbf{Q}_{\mathbf{X}_k} = \mathbf{I}$  pour mbpls et  $\mathbf{Q}_{\mathbf{X}_k} = (\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1}$  pour mbpcaiv.

## Méthodes multiblocs développées sous R [2]

ACP sur variables instrumentales multibloc [Bougeard et al., 2011] et PLS multibloc [Wold, 1984]

### Solution d'ordre 1 pour la variable latente $\mathbf{u}$

- $\lambda^{(1)} = \sum_k \text{var}(\mathbf{P}_{X_k} \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)})$  pour mbpcaiv ( $\|\mathbf{v}^{(1)}\| = 1$ ),
- $\lambda^{(1)} = \sum_k \text{var}(\mathbf{X}'_k \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)}) = \text{var}(\mathbf{X}' \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)})$  pour mbpls ( $\|\mathbf{v}^{(1)}\| = 1$ ).

### Solutions d'ordre 1 pour les autres variables latentes $\mathbf{t}$ et $(\mathbf{t}_1, \dots, \mathbf{t}_k)$

- Pour mbpls :  $\lambda^{(1)} = \text{var}(\mathbf{Y}' \mathbf{D} \mathbf{X} \mathbf{w}^{(1)})$  sous la contrainte  $\|\mathbf{w}^{(1)}\| = 1$
- Pour mbpcaiv (contrainte  $\|\mathbf{t}_k^{(1)}\| = 1$ ) :
  - $\mathbf{t}_k^{(1)} = \mathbf{X}_k \mathbf{w}_k^{(1)} = \mathbf{P}_{X_k} \mathbf{u}^{(1)} / \|\mathbf{P}_{X_k} \mathbf{u}^{(1)}\|$  avec  $\mathbf{P}_{X_k} = \mathbf{X}_k (\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1} \mathbf{X}'_k$
  - $\mathbf{t}^{(1)} = \mathbf{X} \mathbf{w}^{(1)} = \sum_k \mathbf{P}_{X_k} \mathbf{u}^{(1)} / \sqrt{\sum_k \|\mathbf{P}_{X_k} \mathbf{u}^{(1)}\|^2}$

### Solutions d'ordres supérieurs

- Déflation des tableaux  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$  sur la variable latente globale  $\mathbf{t}^{(1)}$ ,
- Application du problème de maximisation aux résidus de ces tableaux, etc.

## Méthodes multiblocs développées sous R [2]

ACP sur variables instrumentales multibloc [Bougeard et al., 2011] et PLS multibloc [Wold, 1984]

### Solution d'ordre 1 pour la variable latente $\mathbf{u}$

- $\lambda^{(1)} = \sum_k \text{var}(\mathbf{P}_{\mathbf{X}_k} \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)})$  pour mbpcaiv ( $\|\mathbf{v}^{(1)}\| = 1$ ),
- $\lambda^{(1)} = \sum_k \text{var}(\mathbf{X}'_k \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)}) = \text{var}(\mathbf{X}' \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)})$  pour mbpls ( $\|\mathbf{v}^{(1)}\| = 1$ ).

### Solutions d'ordre 1 pour les autres variables latentes $\mathbf{t}$ et $(\mathbf{t}_1, \dots, \mathbf{t}_K)$

- Pour mbpls :  $\lambda^{(1)} = \text{var}(\mathbf{Y}' \mathbf{D} \mathbf{X} \mathbf{w}^{(1)})$  sous la contrainte  $\|\mathbf{w}^{(1)}\| = 1$
- Pour mbpcaiv (contrainte  $\|\mathbf{t}_k^{(1)}\| = 1$ ) :
  - $\mathbf{t}_k^{(1)} = \mathbf{X}_k \mathbf{w}_k^{(1)} = \mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)} / \|\mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)}\|$  avec  $\mathbf{P}_{\mathbf{X}_k} = \mathbf{X}_k (\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1} \mathbf{X}'_k$
  - $\mathbf{t}^{(1)} = \mathbf{X} \mathbf{w}^{(1)} = \sum_k \mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)} / \sqrt{\sum_k \|\mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)}\|^2}$

### Solutions d'ordres supérieurs

- Déflation des tableaux  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$  sur la variable latente globale  $\mathbf{t}^{(1)}$ ,
- Application du problème de maximisation aux résidus de ces tableaux, etc.

## Méthodes multiblocs développées sous R [2]

ACP sur variables instrumentales multibloc [Bougeard et al., 2011] et PLS multibloc [Wold, 1984]

### Solution d'ordre 1 pour la variable latente $\mathbf{u}$

- $\lambda^{(1)} = \sum_k \text{var}(\mathbf{P}_{\mathbf{X}_k} \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)})$  pour mbpcaiv ( $\|\mathbf{v}^{(1)}\| = 1$ ),
- $\lambda^{(1)} = \sum_k \text{var}(\mathbf{X}'_k \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)}) = \text{var}(\mathbf{X}' \mathbf{D} \mathbf{Y} \mathbf{v}^{(1)})$  pour mbpls ( $\|\mathbf{v}^{(1)}\| = 1$ ).

### Solutions d'ordre 1 pour les autres variables latentes $\mathbf{t}$ et $(\mathbf{t}_1, \dots, \mathbf{t}_K)$

- Pour mbpls :  $\lambda^{(1)} = \text{var}(\mathbf{Y}' \mathbf{D} \mathbf{X} \mathbf{w}^{(1)})$  sous la contrainte  $\|\mathbf{w}^{(1)}\| = 1$
- Pour mbpcaiv (contrainte  $\|\mathbf{t}_k^{(1)}\| = 1$ ) :
  - $\mathbf{t}_k^{(1)} = \mathbf{X}_k \mathbf{w}_k^{(1)} = \mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)} / \|\mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)}\|$  avec  $\mathbf{P}_{\mathbf{X}_k} = \mathbf{X}_k (\mathbf{X}'_k \mathbf{D} \mathbf{X}_k)^{-1} \mathbf{X}'_k$
  - $\mathbf{t}^{(1)} = \mathbf{X} \mathbf{w}^{(1)} = \sum_k \mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)} / \sqrt{\sum_k \|\mathbf{P}_{\mathbf{X}_k} \mathbf{u}^{(1)}\|^2}$

### Solutions d'ordres supérieurs

- Déflation des tableaux  $(\mathbf{X}_1, \dots, \mathbf{X}_K)$  sur la variable latente globale  $\mathbf{t}^{(1)}$ ,
- Application du problème de maximisation aux résidus de ces tableaux, etc.

# Plan de la présentation

- 1 Contexte
  - Données multiblocs
  - Outils disponibles sur le logiciel R
  
- 2 Méthodes
  - Exemple de données
  - Méthodes multiblocs
  
- 3 Application des méthodes multiblocs avec ade4
  - Import et calculs
  - Interprétation descriptive
  - Interprétation explicative
  
- 4 Conclusion & perspectives

## Import et utilisation des méthodes multiblocs avec ade4

### Spécificités des tableaux en tant qu'objets ade4

- Le tableau **Y** est défini comme un objet `dudi`,
- Le tableau **X** =  $[X_1 | \dots | X_K]$  est défini comme un objet `ktab`.

### Code R pour l'import

```
R> library(ade4)
R> data(chickenk)
R> mortalité <- chickenk[[1]]
R> dudiY_chik <- dudi.pca(mortalité, scannf=FALSE)
R> ktabX_chik <- chickenk[2:5]
```

### Code R pour l'utilisation des méthodes multiblocs

```
R> res<-mbpcaiv(dudiY_chik, ktabX_chik, scale=T, option="uniform")
```

## Import et utilisation des méthodes multiblocs avec ade4

### Spécificités des tableaux en tant qu'objets ade4

- Le tableau **Y** est défini comme un objet `dudi`,
- Le tableau **X** =  $[X_1 | \dots | X_K]$  est défini comme un objet `ktab`.

### Code R pour l'import

```
R> library(ade4)
R> data(chickenk)
R> mortalité <- chickenk[[1]]
R> dudiY_chik <- dudi.pca(mortalité, scannf=FALSE)
R> ktabX_chik <- chickenk[2:5]
```

### Code R pour l'utilisation des méthodes multiblocs

```
R> res<-mbpcaiv(dudiY_chik, ktabX_chik, scale=T, option="uniform")
```

## Import et utilisation des méthodes multiblocs avec ade4

### Spécificités des tableaux en tant qu'objets ade4

- Le tableau **Y** est défini comme un objet `dudi`,
- Le tableau **X** =  $[X_1 | \dots | X_K]$  est défini comme un objet `ktab`.

### Code R pour l'import

```
R> library(ade4)
R> data(chickenk)
R> mortalité <- chickenk[[1]]
R> dudiY_chik <- dudi.pca(mortalité, scannf=FALSE)
R> ktabX_chik <- chickenk[2:5]
```

### Code R pour l'utilisation des méthodes multiblocs

```
R> res<-mbpcaiv(dudiY_chik, ktabX_chik, scale=T, option="uniform")
```

# Inertie et variances expliquées des tableaux par les variables latentes t

Aide à la sélection du nombre de dimensions présentant un intérêt pour l'interprétation

## Code R pour l'obtention des inerties et variances expliquées

```
R> summary(res)
```

```
> summary(res)
Multiblock Principal Component Analysis with Instrumental Variables
rows: 351          Y columns: 4
Xk blocks: 4      X columns: 20
```

	Eig	%Iner	%Iner+	%VarYT	%VarYT+	%VarXT	%VarXT+
1	44	35.1	35.1	41.2	41.2	8.58	8.58
2	26.3	20.9	56	21.5	62.7	9.05	17.6
3	23.7	18.9	74.9	19.2	81.8	7.36	25
4	19.6	15.6	90.6	11.1	92.9	6.49	31.5
5	5.35	4.27	94.8	3.02	95.9	6.97	38.5
6	1.97	1.57	96.4	1.51	97.4	6.71	45.2
7	1.68	1.34	97.8	1.04	98.4	5.91	51.1
8	1.04	0.831	98.6	0.598	99	7.16	58.2
9	0.559	0.446	99	0.412	99.5	4.88	61
10	0.484	0.386	99.4	0.224	99.7	5.3	68.4

10 premières dimensions ↓

	X1.%VarXkT	X1.%VarXkT+	X2.%VarXkT	X2.%VarXkT+	X3.%VarXkT	X3.%VarXkT+	X4.%VarXkT	X4.%VarXkT+
1	6.15	6.15	7.9	7.9	14.5	14.5	5.11	5.11
2	6.59	12.7	10.6	18.5	4.85	19.4	14.5	19.6
3	3.22	16	5.21	23.7	9.81	29.2	11.2	30.8
4	3.5	19.5	11	34.7	6.14	35.4	4.75	35.6
5	6.7	26.2	5.25	39.9	9.82	45.2	6.02	41.6
6	11.1	37.2	7.66	47.6	2.44	47.6	6.03	47.6
7	9.52	46.8	4.52	52.1	4.32	49.9	5.65	53.3
8	5.65	52.4	13.5	65.6	6.62	58.5	2.02	55.3
9	6.51	58.9	3.05	68.6	7.65	66.1	2.21	57.5
10	7.89	66.8	8.85	77	1.98	68.1	2.39	59.9

10 premières dimensions ↓

**X<sub>1</sub>**      **X<sub>2</sub>**      **X<sub>3</sub>**      **X<sub>4</sub>**



## Graphiques descriptifs

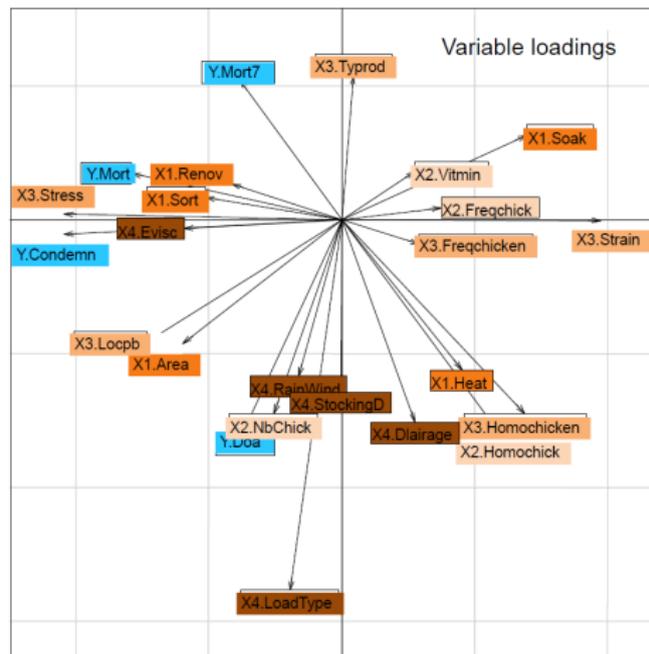
Description des liens entre les variables  $X$ , les étapes de productions ( $X_1, \dots, X_4$ ) et la mortalité  $Y$

### Code R pour les graphes descriptifs

```
plot(res)
```

### Interprétation

- Les quatre éléments caractérisant la mortalité sont peu liés entre eux,
- Chaque élément est lié à des facteurs de risque spécifiques,
- La première composante de la mortalité (liée à la "mortalité durant l'élevage" et les "saisies à l'abattoir") est surtout liée aux étapes de production  $X_1, X_2, X_3$ .



## Graphiques descriptifs

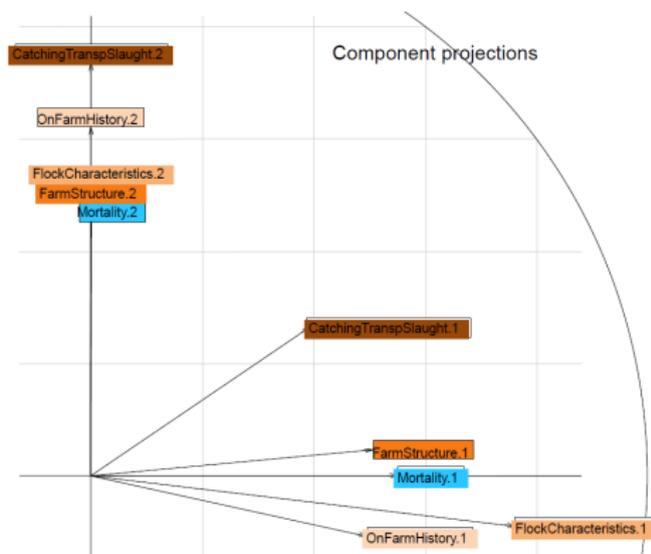
Description des liens entre les variables  $X$ , les étapes de productions ( $X_1, \dots, X_4$ ) et la mortalité  $Y$

### Code R pour les graphes descriptifs

```
plot(res)
```

### Interprétation

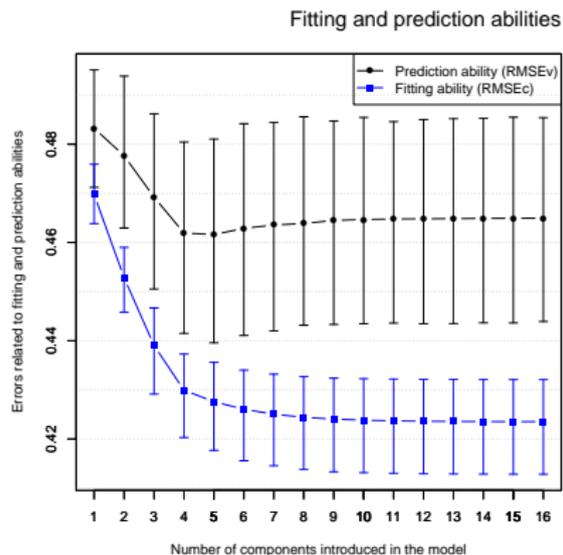
- Les quatre éléments caractérisant la mortalité sont peu liés entre eux,
- Chaque élément est lié à des facteurs de risque spécifiques,
- La première composante de la mortalité (liée à la "mortalité durant l'élevage" et les "saisies à l'abattoir") est surtout liée aux étapes de production  $X_1, X_2, X_3$ .



## Sélection du modèle optimal par validation croisée

### Code R pour la validation croisée

```
rescrossval <- testdim.multiblock(res, nbcross=500)  
plot(rescrossval)
```



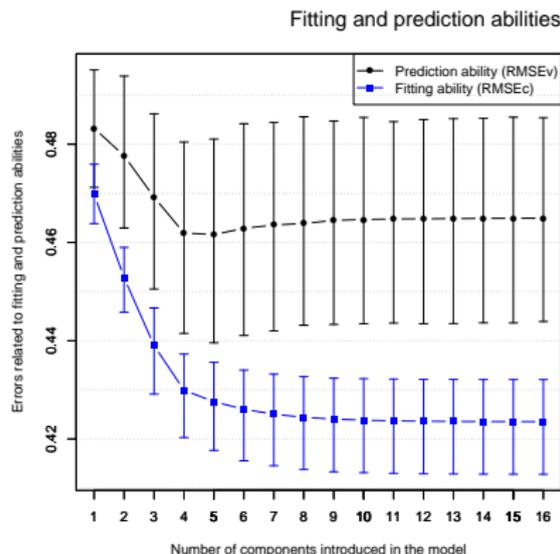
### Interprétation

- Minimisation des erreurs de modélisation et de prédiction,
- Sélection d'un modèle à quatre variables latentes.

## Sélection du modèle optimal par validation croisée

### Code R pour la validation croisée

```
rescrossval <- testdim.multiblock(res, nbcross=500)  
plot(rescrossval)
```



### Interprétation

- Minimisation des erreurs de modélisation et de prédiction,
- Sélection d'un modèle à quatre variables latentes.

## Interprétation du modèle optimal

Coefficients de régression, indices d'importance des variables et des blocs. Intervalles de tolérance bootstrapés.

### Code R pour les simulations bootstrap et les graphes explicatifs

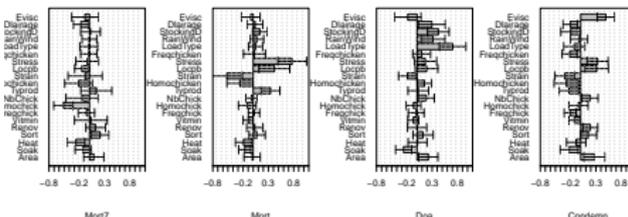
```
resboot<-bootstrap.multiblock(res,
nboot=500, dimopt=4)
plot(resboot)
```

### Interprétation

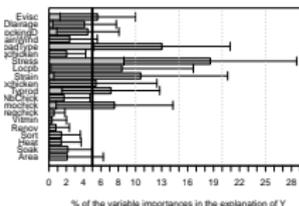
La mortalité est principalement liée à :

- Niveau variable : Stress des animaux pendant la période d'élevage (pb d'alimentation, de chauffage), ainsi qu'au type de ramassage des animaux (mécanique vs manuel).
- Niveau bloc : Caractéristiques du lot pendant la période d'élevage ( $X_3$ ).

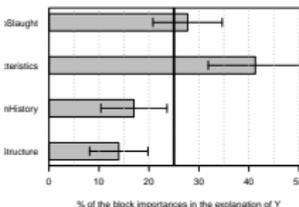
Coefficients (4 dim.)



Variable importances (4 dim.)



Block importances (4 dim.)



## Interprétation du modèle optimal

Coefficients de régression, indices d'importance des variables et des blocs. Intervalles de tolérance bootstrapés.

### Code R pour les simulations bootstrap et les graphes explicatifs

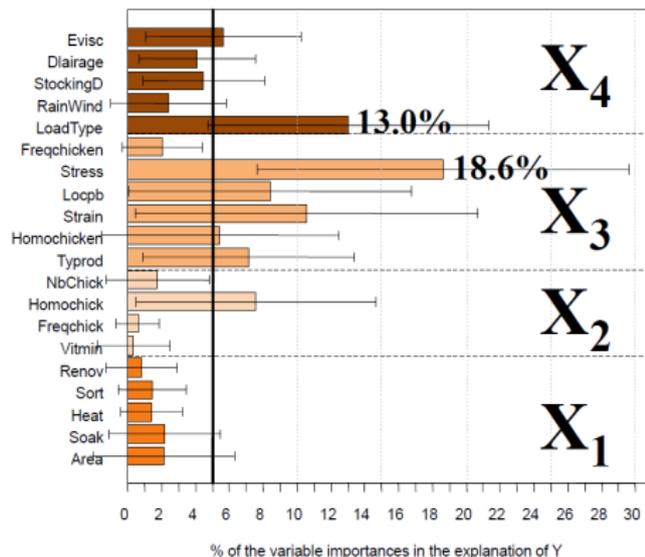
```
resboot<-bootstrap.multiblock(res,  
nboot=500, dimopt=4)  
plot(resboot)
```

### Interprétation

La mortalité est principalement liée à :

- Niveau variable : Stress des animaux pendant la période d'élevage (pb d'alimentation, de chauffage), ainsi qu'au type de ramassage des animaux (mécanique vs manuel).
- Niveau bloc : Caractéristiques du lot pendant la période d'élevage ( $X_3$ ).

Variable importances (4 dim.)



## Interprétation du modèle optimal

Coefficients de régression, indices d'importance des variables et des blocs. Intervalles de tolérance bootstrapés.

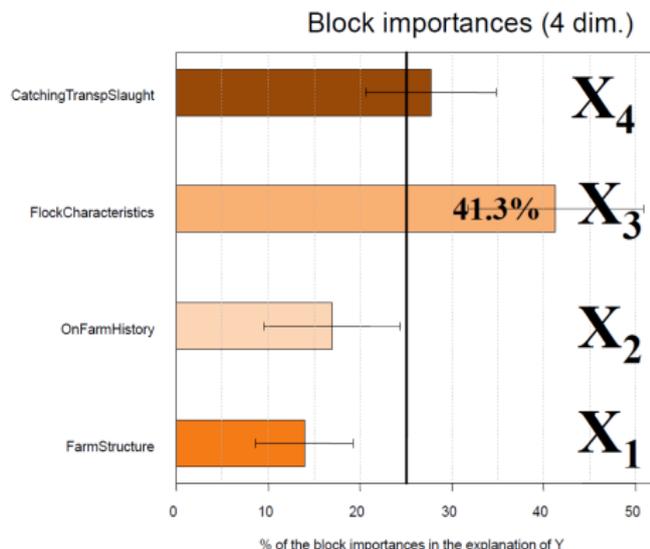
### Code R pour les simulations bootstrap et les graphes explicatifs

```
resboot<-bootstrap.multiblock(res,  
nboot=500, dimopt=4)  
plot(resboot)
```

### Interprétation

La mortalité est principalement liée à :

- Niveau variable : Stress des animaux pendant la période d'élevage (pb d'alimentation, de chauffage), ainsi qu'au type de ramassage des animaux (mécanique vs manuel).
- Niveau bloc : Caractéristiques du lot pendant la période d'élevage ( $X_3$ ).



# Plan de la présentation

- 1 Contexte
  - Données multiblocs
  - Outils disponibles sur le logiciel R
  
- 2 Méthodes
  - Exemple de données
  - Méthodes multiblocs
  
- 3 Application des méthodes multiblocs avec ade4
  - Import et calculs
  - Interprétation descriptive
  - Interprétation explicative
  
- 4 Conclusion & perspectives

## Conclusion & perspectives sur les méthodes multiblocs avec ade4

### Conclusion

- Proposition de deux nouvelles méthodes pour le traitement de  $(K + 1)$  tableaux dans le package `ade4` sur le logiciel R,
- Bénéficie de l'environnement `ade4` (autres méthodes, graphes, forum),
- Proposition de nombreux graphiques et indices d'aide à l'interprétation ( + rénovation actuelle des graphes d'`ade4`),
- Intégration dans la prochaine version d'`ade4`.

### Perspectives de développement

- Extension possible au cas de plusieurs tableaux à expliquer ( $\mathbf{Y}_1, \dots, \mathbf{Y}_{K'}$ ),
- Traitement de mélange de variables quantitatives et qualitatives (extension de la fonction `dudi.mix`),
- Extension possible au cas d'individus structurés en groupes (extension des fonctions `ktab.within` et `ktab.between`).

## Conclusion & perspectives sur les méthodes multiblocs avec `ade4`

### Conclusion

- Proposition de deux nouvelles méthodes pour le traitement de  $(K + 1)$  tableaux dans le package `ade4` sur le logiciel R,
- Bénéficie de l'environnement `ade4` (autres méthodes, graphes, forum),
- Proposition de nombreux graphiques et indices d'aide à l'interprétation ( + rénovation actuelle des graphes d'`ade4`),
- Intégration dans la prochaine version d'`ade4`.

### Perspectives de développement

- Extension possible au cas de plusieurs tableaux à expliquer ( $\mathbf{Y}_1, \dots, \mathbf{Y}_{K'}$ ),
- Traitement de mélange de variables quantitatives et qualitatives (extension de la fonction `dudi.mix`),
- Extension possible au cas d'individus structurés en groupes (extension des fonctions `ktab.within` et `ktab.between`).

# Analyse de $(K + 1)$ tableaux avec le logiciel ade4

## Application en épidémiologie

S. Bougeard<sup>(1)</sup> & S. Dray<sup>(2)</sup>

(1) *Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail, Unité d'Epidémiologie*

(2) *Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive*

