UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE CANONIQUE, GRAPHIQUE BIPLOT ET APPLICATION

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

SYLVAIN FRANCOEUR

AOÛT 2006

REMERCIEMENTS

Tout d'abord, je remercie ma directrice, madame Pascale Rousseau pour son intérêt, son support et ses précieux conseils qui ont assuré la réussite de ce projet.

Je remercie mes parents qui m'ont toujours encouragé dans la poursuite de mes études.

Je veux remercier toute ma famille et mes amis, ça fait du bien de vous avoir près de moi. Un merci particulier aux trois mousquetaires : Daunais, Pascal et Pierre. Je souhaite à tous d'avoir des amis comme eux. Je remercie également Alex, notamment pour son aide tout au long de la maîtrise et du baccalauréat.

Merci à ma complice Mélanie, sa patience, ses encouragements et sa joie de vivre m'ont aidé à passer à travers cette expérience. Un merci spécial à sa famille.

Finalement, je tiens à remercier mes collègues de travail et en particulier Jérôme, Dominique, Carole, Guilain, Geneviève et Mario. Je veux aussi remercier ceux qui m'ont engagé, René et Sylvie, ce qui m'a permis de combiner travail et études dans de bonnes conditions.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Dans ce mémoire, nous présentons le détail mathématique de la technique de l'analyse des corrélations canoniques et du graphique biplot associé à cette technique. L'interprétation des résultats est mise en relief afin de montrer l'utilité du biplot. À l'aide du logiciel SAS, nous utilisons cette technique et celle du graphique biplot sur un ensemble de données provenant d'une enquête de Statistique Canada. Nous concluons que le graphique biplot permet de comprendre facilement la structure de la matrice des corrélations ainsi que les résultats des corrélations canoniques. Nous présentons également l'analyse canonique de redondance qui parachève l'analyse des corrélations canoniques. De plus, nous donnons un bref résumé de toutes les autres méthodes qui ont des buts similaires à l'analyse des corrélations canoniques. Ces autres méthodes sont regroupées sous le thème général de l'analyse canonique.

mots-clés : analyse canonique, corrélations canoniques, graphique biplot, coefficient de corrélation, matrice de corrélation, intra-corrélation, inter-corrélation, analyse canonique de redondance, régression à rang réduit

INTRODUCTION

De nos jours, grâce à l'avancement de la technologie et aux grandes capacités informatiques, les banques de données disponibles pour la recherche sont de plus en plus volumineuses en terme de variables et d'observations. Plus ces bases de données sont importantes, plus il devient difficile de bien connaître les relations entre les variables observées. La recherche de relations, de liens, d'associations ou de corrélations entre les variables d'un ensemble de données sont les sujets qui nous intéressent particulièrement. La méthode statistique de l'analyse canonique fait partie des techniques utilisées pour résumer adéquatement cette information : elle cherche à trouver des relations entre deux ou plusieurs groupes de variables qui sont mesurées sur les mêmes individus d'un échantillon.

Pour visualiser ces multiples relations, la technique de représentation graphique du biplot, est devenue un outil important pour faciliter cette analyse. J.C. Gower (1996) a élaboré les mathématiques des graphiques biplot pour l'analyse en composantes principales, les modèles biadditifs, le «multidimmentionnal scaling» et l'analyse canonique. Le morphème «bi» indique qu'il y a double représentation dans un graphique. Un article de Cajo J. F. ter Braak, «Interpreting canonical correlation analysis through biplots of structure correlations and weights», discute des premiers développements mathématiques du graphique biplot dans le contexte des corrélations canoniques. Gower présente plutôt le graphique biplot de l'analyse canonique dans le contexte de l'analyse des variables canoniques (« Canonical variate analysis ») qui est un cas particulier des corrélations canoniques. De plus, Gower développe les mathématiques pour établir une échelle sur les axes biplot.

Tout d'abord au premier chapitre, nous étudierons l'analyse des corrélations canoniques qui est la principale méthode en analyse canonique. Dans l'ancienne littérature,

l'analyse canonique est souvent associée à l'analyse des corrélations canoniques. Nous présenterons les détails mathématiques de la théorie et les mesures reliées aux corrélations canoniques. Nous présenterons également les détails mathématiques permettant d'obtenir le graphique biplot dans le contexte de l'analyse des corrélations canoniques (ter Braak).

Au deuxième chapitre, nous appliquerons la technique des corrélations canoniques et du biplot sur un ensemble de données provenant d'une enquête de Statistique Canada disponible sur le site Internet Sherlock (http ://sherlock.crepuq.qc.ca). Nous présenterons l'ensemble de données ainsi que l'application des calculs des corrélations canoniques faites avec le logiciel SAS. Nous tracerons le graphique biplot qui aidera à bien comprendre les relations entre les variables de l'ensemble de données et à interpréter les résultats des corrélations canoniques. Nous allons voir qu'il est facile de visualiser les relations dans l'ensemble de données à l'aide de ces méthodes.

Au troisième chapitre, nous présenterons une méthode assez récente qui est l'analyse canonique de redondance (appelée aussi régression à rang réduit). Cette méthode est une suite de l'analyse des corrélations canoniques. Nous présenterons sommairement un graphique biplot du modèle. Nous appliquerons la méthode sur notre ensemble de données que nous avons étudié au chapitre II.

Au dernier chapitre, nous ferons une synthèse des méthodes reliées au thème de l'analyse canonique. Tous les auteurs mentionnent l'aspect important de généralité que propose l'analyse canonique. Nous allons donc présenter un résumé de nos recherches et de nos lectures qui nous ont permis d'éclaircir ce domaine aussi vaste. Nous allons voir, sans entrer dans les détails théoriques, toutes les méthodes se rapportant à l'analyse canonique et les cas particuliers qui sont souvent des techniques d'analyse multidimensionnelles bien connues. Nous jugeons qu'il est important de faire cette mise au point car elle mettra en perspective plusieurs méthodes d'analyse de données et permettra peut-être de mieux les appliquer.

L'analyse canonique présente donc un cadre théorique intéressant car elle généra-

lise plusieurs techniques statistiques. Afin de montrer son utilité, nous mettrons l'emphase dans ce mémoire sur l'interprétation des résultats et ce, autant dans la présentation de la théorie que dans l'application. Même si l'analyse canonique tend à être utilisée davantage ces dernières années, elle est encore peu appliquée en analyse de données probablement par méconnaissance des graphiques biplot (la mathématique ayant été développée récemment). Sans ces graphiques, les auteurs prétendaient souvent que les résultats étaient difficiles à interpréter. Nous verrons ainsi que le graphique biplot aide à comprendre, interpréter et résumer les résultats plus facilement et c'est une des raisons pourquoi nous lui apportons tant d'intérêt.

CHAPITRE I

L'ANALYSE DES CORRÉLATIONS CANONIQUES ET BIPLOT

1.1 Introduction

La théorie de la corrélation canonique a été développée initialement par Hotelling en 1935-1936. Elle repose essentiellement sur la recherche d'une combinaison linéaire de variables d'un premier groupe et d'une combinaison linéaire de variables d'un autre groupe (variables quantitatives mesurées sur les mêmes individus) de sorte que la corrélation entre ces combinaisons linéaires soit maximale. La recherche d'autres combinaisons linéaires avec des corrélations maximales et des propriétés orthogonales aux combinaisons linéaires précédentes peut se poursuivre. Le but de l'analyse canonique est de résumer, le plus adéquatement possible, les relations linéaires entre les groupes de variables. L'*analyse* proprement dite des corrélations canoniques est l'analyse des résultats des calculs : les coefficients canoniques, les corrélations entre les variables canoniques et les variables originales auxquels s'ajoutent maintenant le graphique biplot. L'analyse des corrélations canoniques est utilisée spécialement pour décrire et comprendre les relations entre deux groupes de variables dont les rôles sont symétriques, c'est-à-dire, qu'un groupe n'est pas nécessairement dépendant de l'autre. Les chercheurs doivent explorer les résultats de l'application de la méthode pour tirer des conclusions sur l'ensemble de données étudié. La méthode ne donne pas un résultat instantané comme en régression ou en classification et c'est peut-être une des raisons qui fait qu'elle est moins présente en analyse de données. L'écologie et la biologie sont les domaines qui utilisent le plus l'analyse des corrélations canoniques. Par exemple, des écologistes sont intéressés à connaître

les liens entre les caractéristiques d'un type d'arbre et des caractéristiques décrivant la situation géographique ou géologique, des biologistes désirent connaître les relations entre des caractéristiques des os de la tête et des caractéristiques des os des jambes. En science sociale, des chercheurs veulent étudier les performances des étudiants du cégep face aux réalisations qu'ils ont accomplies au secondaire.

Dans ce chapitre, nous allons tout d'abord présenter les méthodes de calcul des corrélations canoniques et les tests statistiques qui sont possibles d'effectuer. Ensuite, une section sera consacrée à l'interprétation des résultats obtenus par la théorie des corrélations canoniques. Nous étudierons notamment les corrélations entre les variables canoniques et les variables originales (intra/inter-corrélations). Cette section est très importante lorsque vient le temps d'appliquer la méthode dans un cas réel. Nous développerons par la suite l'article de ter Braak, «Interpreting canonical correlation analysis through biplots of structure correlations and weights», que nous trouvons très intéressant sur le graphique biplot dans le contexte des corrélations canoniques. Ce graphique permettra d'estimer les corrélations des variables originales avec les intra/inter-corrélations. Nous allons discuter également des extensions en lien avec la corrélation canonique que nous avons lues lors de nos recherches. Les livres de Krzanowski (2000, chapitre 14), Johnson et Wichern (1998, chapitre 10), Gittins (1985, chapitre 2 et 3) ainsi que les articles de ter Braak (1990) et Gabriel (1971) nous ont servis de références pour élaborer la théorie de ce chapitre. Le concept de la présentation générale est inspiré du livre de Dillon et Golstein (1984, chapitre 9) et de Rencher (1998, chapitre 8).

1.2 Technique et théorie

Nous supposons que les variables observées sur les mêmes individus permettent de se séparer naturellement en deux groupes. Nous cherchons une combinaison linéaire des variables du premier groupe (U_1) et une autre combinaison linéaire des variables du deuxième groupe (V_1) qui possèdent le plus grand coefficient de corrélation. Ces deux nouvelles variables, U_1 et V_1 appelées variables canoniques, sont construites en imposant une variance unitaire. Puis, nous cherchons une deuxième combinaison linéaire calculée

sur chacun des groupes de variables qui ont la deuxième plus grande corrélation. Les nouvelles variables trouvées, U_2 et V_2, doivent être indépendantes des premières. Nous continuons jusqu'à ce que s combinaisons linéaires seront trouvées qui seront associées à s variables canoniques et à s coefficients de corrélation, appelées coefficients de corrélation canonique. s est le minimum du nombre de variables du premier ou du deuxième groupe. Il faut cependant s'assurer que les observations sur chaque groupe de variables forment deux matrices de rang plein. Si elles ne le sont pas, nous pourrons les rendre de rang plein aisément car il n'y a pas de perte d'informations.

1.2.1 Calcul des corrélations canoniques sur la population

Soient les variables du premier et du deuxième groupe qui sont respectivement

$$\underset{p\times 1}{\mathbf{X}} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \text{ et } \underset{q\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_q \end{pmatrix}$$

dont la matrice de variance-covariance est :

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{XX} & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{YY} \end{pmatrix}.$$

Nous voulons trouver la corrélation maximale entre une combinaison linéaire du vecteur $\mathbf{X}$ et une combinaison linéaire du vecteur $\mathbf{Y}$. Les nouvelles variables, $U_1 = \mathbf{a}_1'\mathbf{X}$ et $V_1 = \mathbf{b}_1'\mathbf{Y}$, doivent être de variance unitaire. Les vecteurs $\mathbf{a}_1' = (a_{11} \ldots a_{1p})$ et $\mathbf{b}_1' = (b_{11} \ldots b_{1q})$ sont appelés les vecteurs canoniques et leurs éléments sont les poids canoniques. Nous avons donc :

$$\begin{aligned} Corr(U_1, V_1) &= \frac{Cov(U_1, V_1)}{\sqrt{Var(U_1)}\sqrt{Var(V_1)}} \\ &= \frac{Cov(\mathbf{a}_1'\mathbf{X}, \mathbf{b}_1'\mathbf{Y})}{\sqrt{Var(\mathbf{a}_1'\mathbf{X})}\sqrt{Var(\mathbf{b}_1'\mathbf{Y})}} \\ &= \frac{\mathbf{a}_1'\mathbf{\Sigma}_{XY}\mathbf{b}_1}{\sqrt{\mathbf{a}_1'\mathbf{\Sigma}_{XX}\mathbf{a}_1}\sqrt{\mathbf{b}_1'\mathbf{\Sigma}_{YY}\mathbf{b}_1}} \end{aligned} \tag{1.1}$$

Étant donné que la corrélation est invariante sous les changements d'échelles, nous pouvons donc imposer la contrainte que les variances de U_1 et V_1 sont unitaires. Pour que

cette corrélation soit maximale, nous devons maximiser $Cov(U_1, V_1)$ sous la contrainte $Var(U_1) = Var(V_1) = 1$. Le Lagrangien de la fonction à maximiser sous contrainte est donc :

$$V = \mathbf{a}_1' \mathbf{\Sigma}_{XY} \mathbf{b}_1 - \lambda_a(\mathbf{a}_1' \mathbf{\Sigma}_{XX} \mathbf{a}_1 - 1) - \lambda_b(\mathbf{b}_1' \mathbf{\Sigma}_{YY} \mathbf{b}_1 - 1)$$

où λ_a et λ_b sont les multiplicateurs de Lagrange. Le maximum sera atteint lorsque :

$$\frac{\partial V}{\partial \mathbf{a}_1} = 0 \quad \text{et} \quad \frac{\partial V}{\partial \mathbf{b}_1} = 0.$$

Nous devons donc résoudre les équations suivantes :

$$\frac{\partial V}{\partial \mathbf{a}_1} = \mathbf{\Sigma}_{XY} \mathbf{b}_1 - 2\lambda_a \mathbf{\Sigma}_{XX} \mathbf{a}_1 = 0 \tag{1.2}$$

$$\frac{\partial V}{\partial \mathbf{b}_1} = \mathbf{\Sigma}_{YX} \mathbf{a}_1 - 2\lambda_b \mathbf{\Sigma}_{YY} \mathbf{b}_1 = 0 \tag{1.3}$$

En multipliant 1.2 par $\mathbf{a}_1'$ et 1.3 par $\mathbf{b}_1'$ et en sachant que $\mathbf{a}_1' \mathbf{\Sigma}_{XX} \mathbf{a}_1 = \mathbf{b}_1' \mathbf{\Sigma}_{XX} \mathbf{b}_1 = 1$ et que $\mathbf{a}_1' \mathbf{\Sigma}_{XY} \mathbf{b}_1 = \mathbf{b}_1' \mathbf{\Sigma}_{YX} \mathbf{a}_1$, nous obtenons :

$$\begin{aligned} 2\lambda_a &= \mathbf{a}_1' \mathbf{\Sigma}_{XY} \mathbf{b}_1 \\ 2\lambda_b &= \mathbf{a}_1' \mathbf{\Sigma}_{XY} \mathbf{b}_1 \\ \Rightarrow \quad 2\lambda_a &= 2\lambda_b. \end{aligned}$$

Posons $\lambda_1 = 2\lambda_a = 2\lambda_b$. Avec 1.1 nous savons que

$$\begin{aligned} corr(U_1, V_1) &= \mathbf{a}_1' \mathbf{\Sigma}_{XY} \mathbf{b}_1 \\ &= 2\lambda_a \\ &= 2\lambda_b \\ &= \lambda_1 \end{aligned} \tag{1.4}$$

Alors de 1.2 et de 1.3 nous avons

$$\mathbf{\Sigma}_{XY} \mathbf{b}_1 = \lambda_1 \mathbf{\Sigma}_{XX} \mathbf{a}_1 \tag{1.5}$$

$$\mathbf{\Sigma}_{YX} \mathbf{a}_1 = \lambda_1 \mathbf{\Sigma}_{YY} \mathbf{b}_1 \tag{1.6}$$

Donc de 1.5, $\mathbf{a}_1 = \lambda_1^{-1}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\mathbf{b}_1$ et en substituant dans 1.6 nous avons

$$\lambda_1^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}b_1 = \lambda_1\boldsymbol{\Sigma}_{YY}\mathbf{b}_1.$$

En multipliant par λ_1 et par $\boldsymbol{\Sigma}_{YY}^{-1}$ des deux côtés nous obtenons

$$(\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} - \lambda_1^2)\mathbf{b}_1 = 0. \tag{1.7}$$

De la même façon, de 1.6, $\mathbf{b}_1 = \lambda_1^{-1}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\mathbf{a}_1$ et en substituant 1.5 nous avons

$$\lambda_1^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\mathbf{a}_1 = \lambda_1\boldsymbol{\Sigma}_{XX}\mathbf{a}_1.$$

En multipliant par λ_1 et par $\boldsymbol{\Sigma}_{XX}^{-1}$ des deux côtés nous obtenons

$$(\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} - \lambda_1^2)\mathbf{a}_1 = 0. \tag{1.8}$$

Alors de 1.7 et 1.8, nous sommes en présence d'un problème de valeurs et vecteurs propres d'une matrice carrée. λ_1^2 est valeur propre commune de la matrice $\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$ et de la matrice $\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}$ et les vecteurs $\mathbf{b}_1$ et $\mathbf{a}_1$ sont les vecteurs propres des matrices respectives associés à la même valeur propre λ_1^2. Pour simplifier l'écriture, posons

$$\mathbf{E}_1 = \boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}, \quad \mathbf{E}_2 = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}. \tag{1.9}$$

Étant donné que $\lambda_1 = \mathbf{a}_1'\boldsymbol{\Sigma}_{XY}\mathbf{b}_1$ était la quantité à maximiser, λ_1^2 est donc la plus grande valeur propre commune possible des matrices $\mathbf{E}_1$ et $\mathbf{E}_2$. Le plus grand coefficient de corrélation entre une combinaison linéaire de $\mathbf{X}$ et une combinaison linéaire de $\mathbf{Y}$ est donc la racine carrée de la plus grande valeur propre commune des matrices $\mathbf{E}_1$ et $\mathbf{E}_2$ en vertu de l'équation 1.4.

La seconde plus grande corrélation est la racine carrée de la deuxième plus grande valeur propre commune des matrices $\mathbf{E}_1$ et $\mathbf{E}_2$. Les vecteurs canoniques $\mathbf{a}_2$ et $\mathbf{b}_2$ sont les vecteurs propres de ces matrices associés à la deuxième plus grande valeur propre.

Comme les observations des deux groupes de variables (p variables et q variables) forment deux matrices de rang plein, $\mathbf{\Sigma}_{XY}$ sera de rang s où $s = min(p, q)$. Donc $\mathbf{E}_1$ et $\mathbf{E}_2$ auront au plus s valeurs propres communes (λ_i^2). Nous aurons s coefficients de corrélations canoniques, s paires de variables canoniques (U_i, V_i) et $2 \times s$ combinaisons linéaires associées (vecteurs canoniques $\mathbf{a}_1$ et $\mathbf{b}_1$).

En posant les matrices $\underset{p\times s}{\mathbf{A}} = (\mathbf{a}_1 \ ... \ \mathbf{a}_s)$ et $\underset{q\times s}{\mathbf{B}} = (\mathbf{b}_1 \ ... \ \mathbf{b}_s)$ composées des vecteurs canoniques, les variables canoniques U_i et V_i s'écrivent donc sous la forme matricielle suivante :

$$\begin{aligned} \underset{s\times 1}{\mathbf{U}} &= \begin{pmatrix} U_1 \\ \vdots \\ U_s \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1'\mathbf{X} \\ \vdots \\ \mathbf{a}_s'\mathbf{X} \end{pmatrix} \\ &= \mathbf{A}'\mathbf{X} \end{aligned} \tag{1.10}$$

$$\begin{aligned} \underset{s\times 1}{\mathbf{V}} &= \begin{pmatrix} V_1 \\ \vdots \\ V_s \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{b}_1'\mathbf{Y} \\ \vdots \\ \mathbf{b}_s'\mathbf{Y} \end{pmatrix} \\ &= \mathbf{B}'\mathbf{Y} \end{aligned} \tag{1.11}$$

Nous avons donc trouvé s paires de combinaisons linéaires sur $\mathbf{X}$ et $\mathbf{Y}$. La première paire (U_1, V_1) est de corrélation maximale, la deuxième paire (U_2, V_2) est de corrélation maximale orthogonale à (U_1,V_1) et ainsi de suite. Étant donné que ces U_i et V_i sont construites avec les vecteurs propres de $\mathbf{E}_1$ et $\mathbf{E}_2$ associés aux mêmes valeurs propres, il serait facile de démontrer (voir Johnson et Wichern p.590) les propriétés suivantes :

Propriétés des variables canoniques 1.2.1.1 *Pour U_i et V_i, $i, j = 1, ..., s$:*

1. *Les U_i sont non corrélées pour $i \neq j$*
2. *Les V_i sont non corrélées pour $i \neq j$*
3. *Les U_i sont non corrélées avec les V_j sauf pour $i = j$*
4. *La corrélation entre U_i et $V_i = \lambda_i$*

L'information sur l'association entre les U_i et V_i $i = 1, ..., s$, peut être facilement visualisée par la matrice de variance-covariance suivante

$$\begin{pmatrix} 1 & \dots & 0 & \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & \lambda_s \\ \lambda_1 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_s & 0 & \dots & 1 \end{pmatrix}$$

qui nous permet d'isoler quelques covariances bien choisies. Il est alors beaucoup plus simple de se faire une idée générale des relations entre les groupes qu'avec la matrice de covariance $\mathbf{\Sigma}_{XY}$, surtout si p et q sont grands.

1.2.2 Autres méthodes pour obtenir les corrélations canoniques

La méthodologie expliquée précédemment n'est pas la seule possible pour trouver les variables canoniques. Celles-ci peuvent être trouvées à partir de la décomposition en valeurs singulières d'une matrice. Nous présentons cette deuxième façon de procéder car elle sera utilisée dans la technique du graphique biplot.

Décomposition en valeurs singulières

Certains auteurs comme Johnson et Wichern (p. 590), démontrent que les corrélations canoniques sont les racines carrées des valeurs propres des matrices

$$\mathbf{M}_1 = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1/2} \quad \text{et} \quad \mathbf{M}_2 = \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}.$$

Nous pouvons démontrer que les vecteurs canoniques correspondent dans ce cas à

$$\mathbf{a}_i = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{e}_i,$$

où $\mathbf{e}_i$ est vecteur propre de $\mathbf{M}_1$ associé à λ_i^2 et à

$$\mathbf{b}_i = \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{f}_i,$$

où $\mathbf{f}_i$ est vecteur propre de $\mathbf{M}_2$ associé à la même valeur propre λ_i^2, $i = 1, ..., s$.

Nous allons démontrer que les vecteurs $\mathbf{e}_i$ et $\mathbf{f}_i$ peuvent être obtenus par la décomposition en valeurs singulières de la matrice

$$\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}.$$

Décomposition en valeurs singulières d'une matrice 1.2.2.1 *Une matrice* $\mathbf{Z}$ *quelconque peut être exprimée sous la forme de produit de matrices,*

$$\mathbf{Z} = \mathbf{K\Gamma L}',$$

où $\mathbf{K}$ *est la matrice des vecteurs propres de* $\mathbf{ZZ}'$,
$\mathbf{L}$ *est la matrice des vecteurs propres de* $\mathbf{Z}'\mathbf{Z}$ *et*
$\mathbf{\Gamma}$ *est la matrice diagonale des racines carrées des valeurs propres de* $\mathbf{ZZ}'$ *et de* $\mathbf{Z}'\mathbf{Z}$.

En effet, nous avons alors que la décomposition spectrale de $\mathbf{ZZ}'$ *est :*

$$\mathbf{ZZ}' = \mathbf{K\Gamma L}'\mathbf{L\Gamma K}' = \mathbf{K\Gamma}^2\mathbf{K}'$$

c'est-à-dire, que les vecteurs propres de $\mathbf{ZZ}'$ *sont donnés dans la matrice* $\mathbf{K}$ *et les valeurs singulières sont les carrés des valeurs prores de la matrice diagonale* $\mathbf{\Gamma}$ *(idem pour* $\mathbf{Z}'\mathbf{Z}$*).*

■

La décomposition en valeurs singulières de

$$\mathbf{W} = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2},$$

est,

$$\mathbf{W} = \mathbf{E}\mathbf{\Lambda}\mathbf{F}' \tag{1.12}$$

où $\underset{p\times s}{\mathbf{E}}$ est la matrice des s vecteurs propres de $\mathbf{WW'}$.
Comme $\mathbf{WW'} = \mathbf{E}\mathbf{\Lambda}\mathbf{F}'\mathbf{F}'\mathbf{\Lambda}\mathbf{E}' = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1/2} = \mathbf{M}_1$,
alors la matrice $\underset{p\times s}{\mathbf{E}}$ est la matrice des vecteurs propres de $\mathbf{M}_1$.
De même, $\underset{q\times s}{\mathbf{F}}$ est la matrice des s vecteurs propres de $\mathbf{W}'\mathbf{W} = \mathbf{M}_2$.
Nous avons aussi que $\mathbf{\Lambda} = diag(\lambda_i)$, où $\lambda_1 \geq ... \geq \lambda_s \geq 0$, sont les racines carrées des valeurs propres communes à $\mathbf{M}_1$ et $\mathbf{M}_2$.

Comme nous l'avons mentionné précédemment, les vecteurs canoniques $\mathbf{a}_i$ et $\mathbf{b}_i$ correspondent respectivement à $\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{e}_i$ ($\mathbf{e}_i$ est la $i^{\text{ième}}$ colonne de $\mathbf{E}$) et à $\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{f}_i$ ($\mathbf{f}_i$ est la $i^{\text{ième}}$ colonne de $\mathbf{F}$). Les matrices des vecteurs canoniques sont donc

$$\mathbf{A} = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{E} \quad \text{et} \quad \mathbf{B} = \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{F}.$$

À partir des expressions 1.10 et 1.11, les s paires de variables canoniques sont données par les expressions matricielles suivantes

$$\begin{aligned} \underset{s\times 1}{\mathbf{U}} &= \mathbf{A}'\mathbf{X} \\ &= (\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{E})'\,\mathbf{X} \end{aligned} \tag{1.13}$$

$$\begin{aligned} \underset{s\times 1}{\mathbf{V}} &= \mathbf{B}'\mathbf{Y} \\ &= (\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{F})'\,\mathbf{Y} \end{aligned} \tag{1.14}$$

Variables standardisées

Il est également possible de travailler à partir de variables standardisées. Certains préfèrent cette méthode car elle permet de mieux interpréter les poids canoniques. Nous remplaçons $\mathbf{\Sigma}$ par

$$\boldsymbol{\rho} = \begin{pmatrix} \boldsymbol{\rho}_{XX} & \boldsymbol{\rho}_{XY} \\ \boldsymbol{\rho}_{YX} & \boldsymbol{\rho}_{YY} \end{pmatrix}$$

dans le calcul de coefficients de corrélations canoniques présenté à la section précédente (1.2.1) ou dans la décomposition en valeurs singulières que nous venons d'aborder. Nous n'avons qu'à utiliser les groupes de variables standardisées $(\mathbf{Z}_X, \mathbf{Z}_Y)$ au lieu des groupes de variables originales $(\mathbf{X}, \mathbf{Y})$ dans le calcul des valeurs des variables canoniques.

À noter que les valeurs des corrélations canoniques et des variables canoniques ne sont jamais affectées par ces changements et que toutes les propriétés énumérées sont conservées. Cependant, les poids des vecteurs canoniques ne sont pas les mêmes et cela influence l'interprétation que nous pourrons en faire. Toutes les sections qui suivent dans ce chapitre s'appliquent également aux variables standardisées. Nous verrons que dans certains cas, il est préférable d'utiliser la standardisation afin de simplifier les calculs et d'obtenir une interprétation des résultats plus claire.

1.3 Échantillon, population et tests

Nous supposons que les variables des deux groupes mesurées sur les individus sont quantitatives. Nous pouvons présumer une distribution normale pour le vecteur joint des deux groupes de variables avec

$$\boldsymbol{\mu}' = (\boldsymbol{\mu}_X \quad \boldsymbol{\mu}_Y) \quad \text{et} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}.$$

Il est donc possible d'estimer $\boldsymbol{\Sigma}$ par $\mathbf{S}$ ou $\boldsymbol{\rho}$ par $\mathbf{R}$,

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{XX} & \mathbf{S}_{XY} \\ \mathbf{S}_{YX} & \mathbf{S}_{YY} \end{pmatrix} \quad \text{et} \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_{XX} & \mathbf{R}_{XY} \\ \mathbf{R}_{YX} & \mathbf{R}_{YY} \end{pmatrix}$$

qui sont les estimateurs à vraisemblance maximale. En utilisant ces estimateurs dans la procédure de calcul des corrélations canoniques, nous trouverons des estimateurs pour $\mathbf{A} = (\mathbf{a}_1 \ \ldots \ \mathbf{a}_s)$, $\mathbf{B} = (\mathbf{b}_1 \ \ldots \ \mathbf{b}_s)$ et $\lambda_1 \ \ldots \ \lambda_s$ qui seront notés

$$\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1 \ \ldots \ \hat{\mathbf{a}}_s),\ \hat{\mathbf{B}} = (\hat{\mathbf{b}}_1 \ \ldots \ \hat{\mathbf{b}}_s) \text{ et } \hat{\lambda}_1 \ \ldots \ \hat{\lambda}_s.$$

Il est possible d'obtenir des tests statistiques inférentiels sur ces paramètres. Les distributions exactes sont très difficiles à trouver et nous devons faire appel à la théorie

asymptotique. En pratique, la seule hypothèse qu'il est souhaitable de tester est que les coefficients de corrélation canonique de la population soient nuls. Si nous établissons que plusieurs coefficients peuvent être nuls, nous pourrons alors nous concentrer que sur quelques-uns et par conséquent, réduire la dimension du problème (nombre d'associations entre les groupes de variables).

1.3.1 Tests séquentiels

Nous testons s'il n'y a pas d'association entre les groupes de variables originales de la population, c'est-à-dire que $\mathbf{\Sigma}_{XY} = \mathbf{0}$ (H_0). Pour une grande taille d'échantillon, le test statistique du ratio de vraisemblance maximale est donné par :

$$-2\ln\Lambda = -n\sum_{i=1}^{s}\ln(1-\hat{\lambda}_i^2)$$

Si H_0 est vraie (équivalent à tous les $\lambda_i = 0$), $-2\ln\Lambda$ suit une distribution asymptotique χ^2 avec pq degrés de liberté. Nous pouvons améliorer l'approximation de la χ^2 avec le facteur de correction de Barlett en remplaçant n par $n' = n - \frac{1}{2}(p+q+3)$, n étant la taille de l'échantillon.

Si nous rejetons H_0, il y a au moins un des $\hat{\lambda}_i$ différent de 0. Étant donné que les $\hat{\lambda}_i$ sont en ordre décroissant $(\hat{\lambda}_1 > \hat{\lambda}_2 > ... > \hat{\lambda}_s)$, nous sommes assurés que la plus grande valeur $\hat{\lambda}_1$ est différente de 0. Nous pouvons donc faire un autre test $H_0 : \lambda_1 \neq 0, \lambda_2 = ... = \lambda_s = 0$. Il est alors possible de supprimer la plus grande valeur dans Λ et la statistique suivra une loi χ^2 avec $(p-1)(q-1)$ degrés de liberté. Ainsi, nous testons si la deuxième plus grande valeur propre n'est pas différente de 0. Nous pouvons faire des tests séquentiels, c'est-à-dire continuer ces tests jusqu'à ce qu'il soit impossible de rejeter H_0. Nous présentons dans le tableau 1.1 la série de tests qu'il est possible d'effectuer en considérant $\hat{\lambda}_1 > \hat{\lambda}_2 > ... > \hat{\lambda}_s$.

C'est à l'aide de cette série de tests que nous trouvons k "importantes" relations entre les variables et c'est avec ces k relations qu'il est possible de réduire la dimension du problème. Il est bon de mentionner que si les tests sont tous effectués avec un niveau α,

Tableau 1.1 Tests séquentiels d'association entre les variables

Test	Ho $(s = min(p, q))$	Statistique observée $(n' = n - \frac{1}{2}(p + q + 3))$	Loi et degré de liberté
Test 0 :			
Pas d'association entre les variables	$\lambda_i = 0, i = 1..s,$	$-n' \sum_{i=1}^{s} \ln(1 - \hat{\lambda}_i^2)$	χ^2_{pq}
Test 1 :			
Une association via la première paire de variables canoniques	$\lambda_1 \neq 0, \lambda_2 = ... = \lambda_s = 0$	$-n' \sum_{i=2}^{s} \ln(1 - \hat{\lambda}_i^2)$	$\chi^2_{(p-1)(q-1)}$
.	.	.	.
.	.	.	.
.	.	.	.
Test k :			
k associations via les k premières paires de variables canoniques	$\lambda_1 \neq 0, ..., \lambda_k \neq 0,$ $\lambda_{k+1} = ... = \lambda_s = 0$	$-n' \sum_{i=k+1}^{s} \ln(1 - \hat{\lambda}_i^2)$	$\chi^2_{(p-k)(q-k)}$

le niveau global de l'ensemble des tests n'est pas α et il est difficile à déterminer. Il est donc préférable de choisir un α assez petit (ex : 0.01). Nous devons remarquer également que nous ne continuons plus les tests seulement lorsqu'il est impossible de rejeter H_0. Cela ne veut pas nécessairement dire que cette hypothèse est vraie.

Cette série de tests nous **guide** seulement à sélectionner un nombre important de relations, de variables canoniques. Nous devons continuer le travail d'exploration des données.

1.4 Interprétation des résultats

En analyse des corrélations canoniques, l'interprétation des résultats (corrélations canoniques et variables canoniques) n'est pas aussi simple que dans les autres méthodes statistiques en général et c'est pourquoi nous lui consacrons cette section. Les résultats ne sont pas évidents à interpréter car ils ont un sens particulier et nous devons faire appel à d'autres mesures afin d'obtenir des conclusions intéressantes de la théorie appliquée à un ensemble de données. Plusieurs auteurs jugent difficile cette étape et c'est probablement une des raisons qui fait que l'analyse des corrélations canoniques n'a pas été très utilisée par le passé. De plus, bien que certaines mesures (que nous présenterons) sont plus importantes que d'autres, les auteurs ne s'entendent pas tous sur celles à utiliser afin de donner la meilleure interprétation. À la prochaine section, nous verrons qu'il existe une façon différente d'interpréter les résultats avec le graphique biplot.

Dans cette section, nous expliquerons d'abord comment nous pouvons interpréter directement les résultats obtenus du calcul des corrélations canoniques. Nous mentionnerons les limites et verrons pourquoi il est essentiel d'aborder d'autres mesures. Les mesures les plus importantes que nous présenterons sont les corrélations entre les variables canoniques et les variables originales (appelées inter-corrélation et intra-corrélation) et la variance expliquée. Au prochain chapitre nous verrons également l'indice de redondance. Toutes ces mesures nous aident à atteindre le but de l'analyse des corrélations canoniques qui est de comprendre et de simplifier les relations entre deux groupes de

variables mesurées sur les mêmes individus.

1.4.1 Coefficient de corrélation canonique (λ_i)

Contrairement au coefficient de corrélation simple ou multiple dont le carré détermine la proportion de variance expliquée d'une variable par la (les) variable (s) indépendante (s), le coefficient de corrélation canonique établit la force des relations entre les groupes de variables à l'étude. Le coefficient de corrélation canonique est souvent vu comme la variance partagée par deux combinaisons linéaires de chaque groupe et non la variance expliquée d'un groupe par l'autre groupe. Alors ce coefficient n'a pas nécessairement une bonne capacité de prédiction (il n'avait pas la prétention de l'avoir non plus !). Afin de combler ce manque et pour améliorer son interprétation, nous verrons au chapitre III, l'indice de redondance qui propose une moyenne de la variance expliquée d'un groupe de variables originales par une variable canonique de l'autre groupe.

Un coefficient de corrélation élevé doit toujours être bien analysé car cette forte corrélation peut provenir d'une forte corrélation entre une seule variable d'un groupe et une seule variable de l'autre groupe. Le graphique biplot que nous présenterons nous permettra de constater rapidement ce fait et nous aidera à identifier un coefficient de corrélation canonique anormalement élevé.

1.4.2 Vecteurs canoniques ($\mathbf{A}, \mathbf{B}$)

De la façon dont les variables canoniques ($\mathbf{U}, \mathbf{V}$) sont construites, elles sont la plupart du temps artificielles, c'est-à-dire que nous ne pouvons leur donner un sens physique. Nous tenterons donc d'interpréter les résultats à l'aide des vecteurs canoniques $\mathbf{A}$ et $\mathbf{B}$ qui déterminent les transformations linéaires des $\mathbf{X}$ et $\mathbf{Y}$ pour obtenir les variables canoniques $\mathbf{U}$ et$\mathbf{V}$. Les éléments qui composent les vecteurs canoniques (poids canoniques) nous aident à trouver la nature des relations entre les variables originales. Lorsque les variables originales sont standardisées, les poids canoniques n'ont pas d'unité de mesure et peuvent être comparés entre eux. Ces poids reflètent les contributions conjointes des

variables originales des groupes dans la corrélation canonique entre **U** et **V**. Si les variables ne sont pas standardisées, les poids canoniques sont proportionnels aux variables originales et ne peuvent qu'être interprété de cette façon.

Cependant, il arrive souvent dans un grand ensemble de données que les variables d'un même groupe soient très corrélées entre elles (multicolinéarité dans le même groupe de variables). Cela fait en sorte que les poids canoniques peuvent changer énormément si les observations sont légèrement perturbées dans l'ensemble de données. Les poids canoniques deviennent alors très difficiles à interpréter car ils ne sont pas stables. Ce problème arrive également en régression linéaire. Afin de bien valider ces poids, il est suggéré de calculer la corrélation entre les variables canoniques et chacune des variables originales du même groupe (intra-corrélation). Il est important que la valeur de ce coefficient de corrélation soit dans le même sens que le poids canonique associé à cette variable pour s'assurer de la stabilité.

1.4.3 Corrélation entre les variables originales et les variables canoniques (intra/inter-corrélation)

Corrélation à l'intérieur du même groupe de variable (intra-corrélation)

Ces corrélations (**X** avec **U**, **Y** avec **V**) nous permettent de voir la contribution de chaque variable originale par rapport à la variable canonique de son groupe. Elles nous aident à interpréter les poids canoniques et bien que plusieurs corrélations seront calculées, nous pourrons identifier les plus importantes dans le but de réduire la dimension du problème. Certains auteurs comme Rencher (1995), n'approuvent pas cette façon de faire d'utiliser une statistique unidimensionnelle dans un problème multidimensionnel car on ne voit pas la contribution conjointe des variables. Ils préfèrent juger des contributions des variables originales par les poids canoniques en utilisant des variables originales standardisées.

L'intra-corrélation exprime la contribution de chaque variable d'un groupe sur les

variables canoniques du même groupe. Elle sert à caractériser les variables canoniques U_i et V_i par rapport aux variables originales avec lesquelles nous les avons définies. Plus la corrélation est forte, plus la variable originale est représentée dans la variable canonique. Nous cherchons donc :

Pour U_i et X_j, $i = 1, ..., s$, $j = 1, ..., p$

$$\begin{aligned} Corr(U_i, X_j) &= \frac{Cov(U_i, X_j)}{\sqrt{Var(X_j)}} \\ &= \frac{Cov(\mathbf{a}_i'\mathbf{X}, X_j)}{\sigma_{X_j}^{1/2}} \\ &= \frac{\mathbf{a}_i' Cov(\mathbf{X}, X_j)}{\sigma_{X_j}^{1/2}} \\ &= \mathbf{a}_i' \, \boldsymbol{\sigma}_{XX_j} \, \sigma_{X_j}^{-1/2} \end{aligned}$$

Cette corrélation sera estimée par

$$\hat{\mathbf{a}}_i' \, \mathbf{s}_{XX_j} \, s_{X_j}^{-1/2}$$

où $\mathbf{s}_{XX_j}$ = la $j^{\text{ième}}$ colonne de $\mathbf{S}_{XX}$ et s_{X_j} l'élément (j,j) de $\mathbf{S}_{XX}$.

Pour U_i et le vecteur de variables originales $\mathbf{X}$ cela correspond pour $i = 1, ..., s$:

$$\begin{aligned} \underset{(1 \times p)}{Corr(U_i, \mathbf{X})} &= Cov(\mathbf{a}_i'\mathbf{X}, \mathbf{X}) \; diag(\sigma_{X_1} ... \; \sigma_{X_p})^{-1/2} \\ &= \mathbf{a}_i' Cov(\mathbf{X}, \mathbf{X}) \; diag(\sigma_{X_1} ... \; \sigma_{X_p})^{-1/2} \\ &= \mathbf{a}_i' \boldsymbol{\Sigma}_{XX} \; diag(\sigma_{X_1} ... \; \sigma_{X_p})^{-1/2} \end{aligned}$$

Cette corrélation sera estimée par

$$\hat{\mathbf{a}}_i' \mathbf{S}_{XX} \; diag(s_{X_1} \; ... \; s_{X_p})^{-1/2}$$

Alors pour le vecteur des variables canoniques de $\mathbf{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_s \end{pmatrix}$ et le vecteur $\mathbf{X}$,

$$\begin{aligned} \underset{(s \times p)}{Corr(\mathbf{U}, \mathbf{X})} &= Cov(\mathbf{A}'\mathbf{X}, \mathbf{X})\ diag(\sigma_{X_1} ... \sigma_{X_p})^{-1/2} \\ &= \mathbf{A}' Cov(\mathbf{X}, \mathbf{X})\ diag(\sigma_{X_1} ... \sigma_{X_p})^{-1/2} \\ &= \mathbf{A}' \mathbf{\Sigma}_{XX}\ diag(\sigma_{X_1} ... \sigma_{X_p})^{-1/2} \end{aligned}$$

Cette corrélation sera estimée par

$$\hat{\mathbf{A}}' \mathbf{S}_{XX}\ diag(s_{X_1} ... s_{X_p})^{-1/2} \tag{1.15}$$

De la même façon pour l'autre groupe de variables :

Pour V_i et Y_j, $i = 1, ..., s$, $j = 1, ..., q$

$$Corr(V_i, Y_j) = \mathbf{b}_i'\ \boldsymbol{\sigma}_{YY_j}\ \sigma_{Y_j}^{-1/2}$$

sera estimée par

$$\hat{\mathbf{b}}_i'\ \mathbf{s}_{YY_j}\ s_{Y_j}^{-1/2}$$

où $\mathbf{s}_{YY_j}$ = la $j^{\text{ième}}$ colonne de $\mathbf{S}_{YY}$ et s_{Y_j} l'élément (j, j) de $\mathbf{S}_{YY}$.

Pour V_i et le vecteur de variables originales $\mathbf{Y}$,

$$\underset{(1 \times q)}{Corr(V_i, \mathbf{Y})} = \mathbf{b}_i' \mathbf{\Sigma}_{YY}\ diag(\sigma_{Y_1} \ ... \ \sigma_{Y_q})^{-1/2}$$

sera estimée par

$$\hat{\mathbf{b}}_i' \mathbf{S}_{YY}\ diag(s_{Y_1} \ ... \ s_{Y_q})^{-1/2}$$

Alors pour le vecteur des variables canoniques de $\mathbf{V} = \begin{pmatrix} V_1 \\ \vdots \\ V_s \end{pmatrix}$ et le vecteur $\mathbf{Y}$,

$$\underset{(s \times q)}{Corr(\mathbf{V}, \mathbf{Y})} = \mathbf{B}' \mathbf{\Sigma}_{YY}\ diag(\sigma_{Y_1} \ ... \ \sigma_{Y_q})^{-1/2}$$

sera estimée par

$$\hat{\mathbf{B}}'\mathbf{S}_{YY}\ diag(s_{Y_1} \ldots s_{Y_q})^{-1/2} \tag{1.16}$$

Ces matrices de corrélations nous informent sur la contribution de chaque variable originale sur leurs variables canoniques. Similaires aux poids des vecteurs canoniques, ces corrélations nous aident à porter attention aux variables originales les plus importantes dans le groupe via le calcul des variables canoniques. Étant donné que les poids canoniques ne sont pas tellement stables (valeurs influencées par la multicolinéarité), nous devons les comparer avec les valeurs de l'intra-corrélation afin de s'assurer qu'ils soient similaires et surtout dans le même sens (positif ou négatif). Une différence importante nous porterait à croire qu'il peut y avoir de la multicolinéarité entre les variables originales. Ce phénomène devra être analysé dans l'ensemble de données de départ et les actions nécessaires doivent être prises afin d'éviter ce problème.

Indice de la variance expliquée

Lorsqu'un coefficient de corrélation canonique est élevé, cela signifie qu'il y existe une forte relation linéaire entre les deux groupes. Comme nous l'avons mentionné, cette relation peut ne provenir que de certaines variables dans les groupes. En analysant les poids canoniques et les intra-corrélations, nous sommes dans une certaine mesure capable de porter un jugement sur cette situation. Afin de faire cette analyse pour toutes les variables canoniques à la fois, nous calculerons l'indice de la variance expliquée. Cet indice nous permettra de mieux apprécier rapidement un coefficient de corrélation canonique à valeur élevée sans nécessairement comparer **tous** les poids canoniques avec les coefficients de l'intra-corrélation de chaque variable.

Cet indice se veut le degré de représentation d'un groupe de variables originales dans une de ses variables canoniques. C'est la proportion de la variance totale de $\mathbf{X}$ expliquée par U_k ou de $\mathbf{Y}$ expliquée par V_k. Lorsque les variables originales sont standardisées, l'indice correspond alors à la **moyenne** des coefficients de l'intra-corrélation (au carré) des variables originales pour sa variable canonique k :

$$U_k^2 = \sum_{i=1}^{p} Corr(\mathbf{U}_k, \mathbf{X}_i)^2/p$$

$$V_k^2 = \sum_{i=1}^{q} Corr(\mathbf{V}_k, \mathbf{Y}_i)^2/q$$

L'indice de la variance expliquée totale est donc

$$\sum_{k=1}^{s} U_k^2 \text{ et } \sum_{k=1}^{s} V_k^2$$

qui correspond à la contribution des variables canoniques dans leurs variables originales.

Nous savons très bien que les combinaisons linéaires des $\mathbf{X}_i$ et $\mathbf{Y}_j$ (U_k et V_k) ne sont pas définis de façon à maximiser l'indice de la variance expliquée. Cependant, pour contrer les lacunes des corrélations canoniques anormalement élevées qui peuvent provenir d'une corrélation très forte entre deux variables, nous devrons analyser cet indice. Si nous observons un coefficient élevé et un indice de variance expliquée très faible, nous devons nous poser des questions sur la signification de ce coefficient de corrélation. Cet indice nous guide dans le choix des variables canoniques à sélectionner et à diminuer la dimension du problème.

Corrélation entre les deux groupes (inter-corrélation)

Comme nous l'avons mentionné précédemment, le coefficient de corrélation canonique n'a pas une bonne capacité de prédiction. Nous allons donc calculer la corrélation entre une variable canonique d'un groupe et une variable originale de l'autre groupe ($\mathbf{X}$ avec $\mathbf{V}$, $\mathbf{Y}$ avec $\mathbf{U}$). Le carré de ce coefficient de corrélation représente la proportion de la variance d'une variable originale qui est expliquée par la variable canonique de l'autre groupe. Nous explorons ainsi la variance couverte par une variable canonique en terme de prédiction (mesure asymétrique). Ce coefficient nous informe donc sur l'étendue de la force de U_i et V_i entre les groupes. C'est le premier indice qui nous aidera à trouver des modèles de prédiction des variables originales à partir des variables canoniques.

Pour V_i et X_j, $i = 1, ..., s$, $j = 1, ..., p$

$$Corr(V_i, X_j) = Corr(\mathbf{b}_i'\mathbf{Y}, X_j)$$

Ce coefficient est comparable au coefficient de corrélation multiple. Étant donné que les poids des $\mathbf{b}_i$ ne sont pas construits comme en régression linéaire, ce coefficient n'est pas maximal (le maximal étant atteint lorsque les $\mathbf{b}_i$ sont obtenus avec la méthode des moindres carrés en régression multiple). Cependant nous pouvons l'interpréter de la même façon : la variance des X_j est expliquée par les $V_i = \mathbf{b}_i'Y_i$ dans une proportion de $[Corr(Vi, Xj)^2]$ %.

Alors,

$$Corr(\mathbf{b}_i'\mathbf{Y}, X_j) = \mathbf{b}_i' \, \boldsymbol{\sigma}_{YX_j} \, {\sigma_{X_j}}^{-1/2}$$

sera estimée par

$$\hat{\mathbf{b}}_i' \, \mathbf{s}_{YX_j} \, {s_{X_j}}^{-1/2}$$

où $\mathbf{s}_{YX_j}$ = la $j^{\text{ième}}$ colonne de $\mathbf{S}_{YX}$ et s_{X_j} l'élément (j, j) de $\mathbf{S}_{XX}$.

Pour V_i et le vecteur des variables originales $\mathbf{X}$,

$$\begin{aligned} \underset{(1 \times p)}{Corr(V_i, \mathbf{X})} &= \mathbf{b}_i' Cov(\mathbf{Y}, \mathbf{X}) \; diag(\sigma_{X_1} ... \; \sigma_{X_p})^{-1/2} \\ &= \mathbf{b}_i' \boldsymbol{\Sigma}_{YX} \; diag(\sigma_{X_1} ... \; \sigma_{X_p})^{-1/2} \end{aligned}$$

qui sera estimée par

$$\hat{\mathbf{b}}_i' \mathbf{S}_{YX} \; diag(s_{X_1} ... \; s_{X_p})^{-1/2} \tag{1.17}$$

Alors pour le vecteur des variables canoniques $\mathbf{V}$ et le vecteur $\mathbf{X}$,

$$\underset{(s \times q)}{Corr(\mathbf{V}, \mathbf{X})} = \mathbf{B}' \boldsymbol{\Sigma}_{YX} \; diag(\sigma_{X_1} ... \; \sigma_{X_p})^{-1/2}$$

sera estimée par

$$\hat{\mathbf{B}}' \mathbf{S}_{YX} \; diag(s_{X_1} ... \; s_{X_p})^{-1/2} \tag{1.18}$$

De la même façon nous obtenons les corrélations entre les variables du deuxième groupe $(\mathbf{Y})$ et U_i :

Pour U_i et Y_j, $i = 1, ..., s$, $j = 1, ..., q$

$$Corr(U_i, Y_j) = \mathbf{a}_i' \, \boldsymbol{\sigma}_{XY_j} \, {\sigma_{Y_j}}^{-1/2}$$

sera estimée par

$$\hat{\mathbf{a}}_i' \, \mathbf{s}_{XY_j} \, {s_{Y_j}}^{-1/2}$$

où $\mathbf{s}_{XY_j}$ = la $j^{\text{ième}}$ colonne de $\mathbf{S}_{XY}$ et s_{Y_j} l'élément (j, j) de $\mathbf{S}_{YY}$.

Pour U_i et le vecteur des variables originales $\mathbf{Y}$,

$$\underset{(1 \times q)}{Corr(U_i, \mathbf{Y})} = \mathbf{a}_i' \boldsymbol{\Sigma}_{XY} \; diag(\sigma_{Y_1} \; ... \; \sigma_{Y_q})^{-1/2}$$

sera estimée par

$$\hat{\mathbf{a}}_i' \mathbf{S}_{XY} \; diag(s_{Y_1} \; ... \; s_{Y_q})^{-1/2} \tag{1.19}$$

Alors pour le vecteur des variables canoniques $\mathbf{U}$ et le vecteur $\mathbf{Y}$,

$$\underset{(s \times q)}{Corr(\mathbf{U}, \mathbf{Y})} = \mathbf{A}' \boldsymbol{\Sigma}_{XY} \; diag(\sigma_{Y_1} \; ... \; \sigma_{Y_q})^{-1/2}$$

sera estimée par

$$\hat{\mathbf{A}}' \mathbf{S}_{XY} \; diag(s_{Y_1} \; ... \; s_{Y_q})^{-1/2} \tag{1.20}$$

Cette mesure montre comment une variable peut être expliquée à partir d'une certaine combinaison linaire (variable canonique) des variables disponibles de l'autre groupe. Elle nous permet donc de remarquer rapidement les variables originales les plus pertinentes face à la variable canonique de l'autre groupe. Le désavantage est que nous devons observer chacune des valeurs pour tenter d'avoir une idée globale de la prédiction des variables originales d'un groupe à partir d'une variable canonique de l'autre groupe. L'indice de la redondance que nous présenterons au chapitre III nous fournira une appréciation globale de la capacité de prédiction d'une variable canonique d'un groupe sur **l'ensemble** des variables originales de l'autre groupe (et non sur une seule variable à la fois).

1.5 Matrice $\mathbf{R}_{XY}$ et matrices des intra/inter-corrélations

1.5.1 Introduction

En 1990, ter Braak discute de l'estimation de la matrice de corrélation entre $\mathbf{X}$ (p variables) et $\mathbf{Y}$ (q variables) standardisées ($\mathbf{R}_{XY}$) à l'aide de la technique graphique du biplot (Gabriel, 1971). Pour ce faire, il utilise les matrices de corrélations entre les variables originales et les variables canoniques que nous avons présentées à la section 1.4.3. La décomposition de $\mathbf{R}_{XY}$ en produit de matrices permettra de tracer un graphique afin de visualiser la meilleure estimation de $\mathbf{R}_{XY}$ au sens des moindres carrés (pondérés). Nous verrons que cette décomposition contient des valeurs des matrices des intra/inter-corrélations. L'auteur propose également une façon efficace de lire le graphique. Nous aurons donc ici un outil visuel plus intéressant que la matrice $\mathbf{R}_{XY}$ de dimension $p \times q$. Nous verrons au prochain chapitre que ce graphique nous permettra de comprendre plus facilement les relations entre $\mathbf{X}$ et $\mathbf{Y}$ car il nous amènera à restructurer $\mathbf{R}_{XY}$.

Nous détaillerons les résultats présentés dans l'article de ter Braak concernant le biplot dans le contexte des corrélations canoniques et des matrices des intra/inter-corrélations. Notre intérêt marqué pour les corrélations et le graphique biplot motive la présentation détaillée de cet article en tant qu'une des techniques d'**interprétation** des résultats de l'analyse des corrélations canoniques. C'est d'ailleurs en partie grâce à cet article que nous nous sommes portés à l'étude de l'analyse canonique dans le cadre de ce mémoire.

1.5.2 Approximation de la matrice $\mathbf{R}_{XY}$

Nous voulons approximer la matrice $\mathbf{R}_{XY}$ de dimension $p \times q$ de rang $s = min(p, q)$ ($\mathbf{R}_{XX}$ et $\mathbf{R}_{YY}$ sont de rang plein), par des matrices $\underset{p\times r}{\mathbf{C}}$ et $\underset{q\times r}{\mathbf{D}}$ de rang inférieur. Dans l'approximation, nous prenons comme pondération les racines carrées des inverses des matrices $\mathbf{R}_{XX}$ et $\mathbf{R}_{YY}$ afin de rendre cette approximation indépendante des transformations linéaires sur $\mathbf{X}$ et $\mathbf{Y}$. Les matrices $\underset{p\times r}{\mathbf{C}}$ et $\underset{q\times r}{\mathbf{D}}$ seront solutions de la méthode des

moindres carrés (pondérés) appliquée à $\mathbf{R}_{XY}$. Le problème se formule ainsi :

$$\underset{\mathbf{C},\,\mathbf{D}}{Min} \parallel \mathbf{R}_{XX}^{-1/2} \left(\mathbf{R}_{XY} - \underset{p\times r}{\mathbf{C}} \underset{r\times q}{\mathbf{D}}{}'\right) \mathbf{R}_{YY}^{-1/2} \parallel^2. \tag{1.21}$$

Remarquons que le terme à minimiser (la distance entre deux matrices) est égal à

$$\parallel \mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} - (\mathbf{R}_{XX}^{-1/2} \underset{p\times r}{\mathbf{C}}) \, (\mathbf{R}_{YY}^{-1/2} \underset{q\times r}{\mathbf{D}})' \parallel^2. \tag{1.22}$$

Le problème 1.21 revient à trouver des matrices $\underset{p\times r}{\mathbf{C}}$ et $\underset{q\times r}{\mathbf{D}}$ qui minimisent l'expression 1.22. Nous allons maintenant énoncer le théorème d'Eckart-Young qui nous sera utile dans la solution de ce problème (Eckart et Young, 1936).

Théorème d'Eckart-Young 1.5.2.1 *Soit la décomposition en valeurs singulières de la matrice* $\mathbf{W}$ *de rang* m

$$\mathbf{W} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}' \tag{1.23}$$

où

$\mathbf{P}$ *contient les vecteurs propres de* $\mathbf{W}\mathbf{W}'$,

$\mathbf{Q}$ *contient les vecteurs propres de* $\mathbf{W}'\mathbf{W}$ *et*

$\mathbf{\Lambda} = diag(\lambda_i)$ *où* $\lambda_1 \geq ... \geq \lambda_m \geq 0$ *sont les valeurs singulières, c'est-à-dire, les racines carrées des valeurs propres communes à* $\mathbf{W}\mathbf{W}'$ *et* $\mathbf{W}'\mathbf{W}$.

Soit $\mathbf{W}_{[r]}$ *une matrice de rang* $r < m$. *Posons,*

$$\mathbf{W}_{[r]} = \mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]}\mathbf{Q}'_{[r]}$$

où

$\mathbf{P}_{[r]}$ = *les* r *premières colonnes de* $\mathbf{P}$,

$\mathbf{Q}'_{[r]}$ = *les* r *premières lignes de* $\mathbf{Q}'$ *et*

$\mathbf{\Lambda}_{[r]}$ = *la matrice diagonale des* r *premières valeurs singulières de* $\mathbf{\Lambda}$.

Alors $\mathbf{W}_{[r]}$ *est la meilleure approximation de la matrice* $\mathbf{W}$ *au sens des moindres carrés, c'est-à-dire que* $\mathbf{W}_{[r]}$ *est solution à*

$$\underset{\mathbf{M}_{[r]}}{Min} \parallel \mathbf{W} - \mathbf{M}_{[r]} \parallel^2 \tag{1.24}$$

où $\mathbf{M}_{[r]}$ *est une matrice de rang* r.

Ce minimum est donné par

$$\lambda_{r+1}^2 + ... + \lambda_m^2$$

car $\|\mathbf{W}\|^2 = \lambda_1^2 + ... + \lambda_m^2$. *La mesure absolue de l'approximation peut être définie par*

$$\sum_{\alpha=1}^{r} \lambda_\alpha^2 / \sum_{\alpha=1}^{m} \lambda_\alpha^2.$$

■

Afin de trouver la solution aux problèmes 1.21 et 1.22, nous allons appliquer le théorème d'Eckart-Young à la matrice à $\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2}$. La décomposition en valeurs singulières de $\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2}$ est

$$\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}'. \tag{1.25}$$

En appliquant le théorème d'Eckart-Young, nous obtenons que

$$(\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2})_{[r]} = \mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]}\mathbf{Q}'_{[r]}$$

est la meilleure approximation de $\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2}$ au sens des moindres carrés.
En posant

$$\mathbf{R}_{XX}^{-1/2} \underset{p\times r}{\mathbf{C}} = \mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]} \tag{1.26}$$

$$\mathbf{R}_{YY}^{-1/2} \underset{q\times r}{\mathbf{D}} = \mathbf{Q}_{[r]} \tag{1.27}$$

Nous avons que

$$(\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2})_{[r]} = \mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]}\mathbf{Q}'_{[r]} = (\mathbf{R}_{XX}^{-1/2} \underset{p\times r}{\mathbf{C}})(\mathbf{R}_{YY}^{-1/2} \underset{q\times r}{\mathbf{D}})'$$

Mais,

$$\begin{aligned} \underset{\mathbf{M}_{[r]}}{Min} \| \mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} - \mathbf{M}_{[r]} \|^2 &= \| \mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} - \mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]}\mathbf{Q}'_{[r]} \|^2 \\ &= \| \mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} - (\mathbf{R}_{XX}^{-1/2} \underset{p\times r}{\mathbf{C}})(\mathbf{R}_{YY}^{-1/2} \underset{q\times r}{\mathbf{D}})' \|^2 \end{aligned}$$

Nous avons donc que

$$\underset{p\times r}{\mathbf{C}} = \mathbf{R}_{XX}^{1/2}\mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]} \tag{1.28}$$

$$\underset{q\times r}{\mathbf{D}} = \mathbf{R}_{YY}^{1/2}\mathbf{Q}_{[r]} \tag{1.29}$$

sont solutions aux problèmes 1.21 et 1.22.

Nous avons ainsi obtenu la meilleure approximation de rang r de la matrice $\mathbf{R}_{XY}$ (de rang $s > r$ et pondérée par $\mathbf{R}_{XX}^{-1/2}$ et $\mathbf{R}_{YY}^{-1/2}$) de la forme $\underset{p\times r}{\mathbf{C}}\ \underset{r\times q}{\mathbf{D}}{}'$ au sens des moindres carrés.

1.5.3 Lien entre $\mathbf{R}_{XY}$ et les matrices des intra/inter-corrélations

En introduction à cette section, nous avons affirmé que ter Braak utilise les matrices des coefficients de l'intra/inter-corrélation afin d'approximer $\mathbf{R}_{XY}$. Nous allons maintenant faire le lien entre ces matrices et la composition des matrices $\underset{p\times r}{\mathbf{C}}$ et $\underset{q\times r}{\mathbf{D}}$ qui nous permettent d'approximer $\mathbf{R}_{XY}$.

Dans la section 1.2.2, nous avons démontré qu'il est possible d'obtenir les corrélations canoniques via la décomposition en valeurs singulières de la matrice $\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}$. Lorsque nous appliquons cette méthode avec des variables standardisées, cela revient à faire la décomposition en valeurs singulières de $\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2}$:

$$\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}' \tag{1.30}$$

qui est la décomposition présentée à l'équation 1.25 et utilisée dans le théorème d'Eckart-Young. Les variables canoniques données par les expressions 1.13 et 1.14 correspondent alors à

$$\begin{aligned}\underset{s\times 1}{\mathbf{U}} &= \hat{\mathbf{A}}'\mathbf{X} \\ &= (\mathbf{R}_{XX}^{-1/2}\mathbf{P})'\,\mathbf{X}\end{aligned}$$

$$\begin{aligned}\underset{s\times 1}{\mathbf{V}} &= \hat{\mathbf{B}}'\mathbf{Y} \\ &= (\mathbf{R}_{YY}^{-1/2}\mathbf{Q})'\,\mathbf{Y}\end{aligned}$$

Nous allons maintenant appliquer ces expressions aux équations de l'intra-corrélation entre $\mathbf{V}$ et $\mathbf{Y}$ et de l'inter-corrélation entre $\mathbf{V}$ et $\mathbf{X}$ lorsque $\mathbf{X}$ et $\mathbf{Y}$ sont standardisées :

Avec l'équation 1.16 de l'intra-corrélation, nous avons que la matrice

$$\begin{aligned} Corr(\mathbf{V},\mathbf{Y}) &= \hat{\mathbf{B}}'\mathbf{R}_{YY} \\ &= (\mathbf{R}_{YY}^{-1/2}\mathbf{Q})'\mathbf{R}_{YY} \\ &= \mathbf{Q}'\mathbf{R}_{YY}^{1/2} \\ \Rightarrow \quad Corr(\mathbf{Y},\mathbf{V}) &= \mathbf{R}_{YY}^{1/2}\mathbf{Q} \end{aligned} \tag{1.31}$$

est la matrice des coefficients de l'intra-corrélation entre $\mathbf{Y}$ et $\mathbf{V}$ de dimensions $q \times s$.

Avec l'équation 1.18 de l'inter-corrélation, nous avons que la matrice

$$\begin{aligned} Corr(\mathbf{V},\mathbf{X}) &= \hat{\mathbf{B}}'\mathbf{R}_{YX} \\ &= (\mathbf{R}_{YY}^{-1/2}\mathbf{Q})'\mathbf{R}_{YX} \\ &= \mathbf{Q}'\mathbf{R}_{YY}^{-1/2}\mathbf{R}_{YX}. \end{aligned} \tag{1.32}$$

En multipliant les membres de gauche et de droite de l'expression 1.30 par $\mathbf{Q}$ (à droite) et par $\mathbf{R}_{XX}^{1/2}$ (à gauche), nous obtenons que

$$\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2}\mathbf{Q} = \mathbf{R}_{XX}^{1/2}\mathbf{P}\Lambda. \tag{1.33}$$

Comme $\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2}\mathbf{Q}$ est la transposée de $Corr(\mathbf{V},\mathbf{X})$ (1.32), alors

$$Corr(\mathbf{X},\mathbf{V}) = \mathbf{R}_{XX}^{1/2}\mathbf{P}\Lambda \tag{1.34}$$

est la matrice des inter-corrélations entre $\mathbf{X}$ et $\mathbf{V}$ de dimensions $p \times s$.

Nous avons donc que

$$\begin{aligned} \mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} &= \mathbf{P}\mathbf{\Lambda}\mathbf{Q}' \\ \mathbf{R}_{XY} &= (\mathbf{R}_{XX}^{1/2}\mathbf{P}\mathbf{\Lambda})(\mathbf{Q}'\mathbf{R}_{YY}^{1/2}) \\ &= (Corr(\mathbf{X},\mathbf{V}))(Corr(\mathbf{V},\mathbf{Y})) \end{aligned}$$

par les expressions 1.34 et 1.31.

1.5.4 Liens entre les matrices des intra/inter-corrélations et les matrices $\underset{p\times r}{\mathbf{C}}$ et $\underset{q\times r}{\mathbf{D}}$

Revenons maintenant aux matrices $\underset{p\times r}{\mathbf{C}}$ et $\underset{q\times r}{\mathbf{D}}$. Avec l'expression 1.28, nous avions que

$$\underset{p\times r}{\mathbf{C}} = \mathbf{R}_{XX}^{1/2}\mathbf{P}_{[r]}\mathbf{\Lambda}_{[r]}.$$

Cette expression correspond aux r premières colonnes de la matrice $\mathbf{R}_{XX}^{1/2}\mathbf{P}\Lambda$, donc aux r premières colonnes de la matrice des inter-corrélations entre $\mathbf{X}$ et $\mathbf{V}$ (1.34). Cela signifie que $\underset{p\times r}{\mathbf{C}}$ contient les vecteurs des corrélations entre les variables de $\mathbf{X}$ et les r premières variables canoniques de $\mathbf{V}$:

$$\begin{aligned} \underset{p\times r}{\mathbf{C}} &= Corr(\mathbf{X}, Vi)\ ,\ i = 1..r. \\ &= (\mathbf{r}_{XV_1},\ \mathbf{r}_{XV_2},\ \ldots\ ,\ \mathbf{r}_{XV_r}) \end{aligned} \tag{1.35}$$

où $\mathbf{r}_{XV_i} = (r_{X_1V_i},\ r_{X_2V_i},\ \ldots\ ,\ r_{X_pV_i})\ ',\ i = 1, ..., r.$

De même, à partir de l'expression 1.29, nous avions que

$$\underset{q\times r}{\mathbf{D}} = \mathbf{R}_{YY}^{1/2}\mathbf{Q}_{[r]}.$$

Cette expression correspond aux r premières colonnes de la matrice $\mathbf{R}_{YY}^{1/2}\mathbf{Q}$, donc aux r premières colonnes de la matrice des intra-corrélations entre $\mathbf{Y}$ et $\mathbf{V}$ (1.31). Cela signifie que $\underset{q\times r}{\mathbf{D}}$ contient les vecteurs des corrélations entre les variables de $\mathbf{Y}$ et les r premières variables canoniques de $\mathbf{V}$:

$$\begin{aligned} \underset{q\times r}{\mathbf{D}} &= Corr(\mathbf{Y}, Vi)\ ,\ i = 1, ..., r. \\ &= (\mathbf{r}_{YV_1},\ \mathbf{r}_{YV_2},\ \ldots\ ,\ \mathbf{r}_{YV_r}) \end{aligned} \tag{1.36}$$

où $\mathbf{r}_{YV_i} = (r_{Y_1V_i},\ r_{Y_2V_i},\ \ldots\ ,\ r_{Y_qV_i})\ ',\ i = 1, ..., r.$

À noter que $\mathbf{\Lambda}_{[r]}$ peut être intégré à $\underset{q\times r}{\mathbf{D}}$ (1.29) au lieu qu'à $\underset{p\times r}{\mathbf{C}}$ (1.28) sans affecter la fonction 1.22. $\underset{p\times r}{\mathbf{C}}$ contiendra alors les r premières colonnes de $corr(\mathbf{X}, \mathbf{U})$, c'est-à-dire les corrélations entre les variables de $\mathbf{X}$ et les r premières variables canoniques de $\mathbf{U}$. $\underset{q\times r}{\mathbf{D}}$ pour

sa part contiendra les corrélations entre les variables de $\mathbf{Y}$ et les r premières variables canoniques de $\mathbf{U}$. Cette décomposition nous donnera également la même estimation de $\mathbf{R}_{XY}$ mais la représentation par le biplot sera toutefois différente. Par symétrie, il est également possible d'approximer la transposée de $\mathbf{R}_{XY}$ $\quad$ $\mathbf{R}_{YX}$ comme utilisé par ter Braak dans son article qui nous donnera les mêmes résultats (transposés).

1.6 Graphique biplot

Nous avons donc que

$$\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1/2} \simeq (\mathbf{R}_{XX}^{-1/2} \underset{p\times r}{\mathbf{C}})\,(\mathbf{R}_{YY}^{-1/2} \underset{q\times r}{\mathbf{D}})'$$

au sens des moindres carrés (pour $r < s$), d'où

$$\underset{p\ \times\ q}{\mathbf{R}_{XY}} \simeq \underset{p\times r}{\mathbf{C}}\ \underset{r\times q}{\mathbf{D}}{}'$$

au sens des moindres carrés pondérés.

$$\begin{aligned}\underset{p\ \times\ q}{\mathbf{R}_{XY}} \simeq \underset{[\ r\ ]}{\mathbf{R}_{XY}} &= \underset{p\times r}{\mathbf{C}}\ \underset{r\times q}{\mathbf{D}}{}' \\ &= \begin{pmatrix} c_{11} & \dots & c_{1r} \\ \vdots & & \vdots \\ c_{p1} & \dots & c_{pr} \end{pmatrix} \begin{pmatrix} d_{11} & \dots & d_{1q} \\ \vdots & & \vdots \\ d_{r1} & \dots & d_{rq} \end{pmatrix}\end{aligned}$$

Soient les points $P_i,\ i = 1, ..., p$, de coordonnées $(c_{i1},\ c_{i2},\ \dots\ ,\ c_{ir})$ notés $\underset{[\ r\ ]}{\mathbf{c_i'}}$ et les points $Q_j,\ j = 1, ..., q$, de coordonnées $(d_{1j},\ d_{2j},\ \dots\ ,\ d_{rj})$ notés $\underset{[\ r\ ]}{\mathbf{d_j}}$, alors

$$r_{x_iy_j} \simeq \underset{[\ r\ ]}{r_{x_iy_j}} = \underset{[\ r\ ][\ r\ ]}{\mathbf{c_i'}\ \mathbf{d_j}}$$

Or, la projection du point P_i sur le vecteur $\overrightarrow{OQ_j}$ où O désigne l'origine est :

$$Proj\ (P_i\ sur\ \overrightarrow{OQ_j}\ ) = \frac{\mathbf{c}_i'\mathbf{d}_j}{\|\mathbf{d}_j\|^2}$$

donc

$$r_{x_iy_j} \simeq \underset{[\ r\ ]}{r_{x_iy_j}} = \|\mathbf{d}_j\|^2 Proj\ (P_i\ sur\ \overrightarrow{OQ_j}\ ).$$

La projection de chaque point P_i sur le vecteur $\overrightarrow{OQ_j}$ nous renseigne sur la corrélation entre X_i et Y_j à un facteur constant $\|\mathbf{d}_j\|^2$ près.

1.6.1 Deux dimensions ($r = 2$)

Afin d'être en mesure de représenter cette corrélation dans un graphique, le choix naturel de l'approximation est de poser $\mathbf{r = 2}$ car les points et la projection se tracent dans le plan.

Nous avons donc

$$\underset{p \times q}{\mathbf{R}_{XY}} \simeq \underset{[\,2\,]}{\mathbf{R}_{XY}} = \underset{p\times 2}{\mathbf{C}} \; \underset{2\times q}{\mathbf{D}}{}'$$

où $\underset{[\,2\,]}{\mathbf{R}_{XY}}$ est l'estimation de rang 2 de la matrice $\underset{p \times q}{\mathbf{R}_{XY}}$.

Pour $i = 1, ..., p$ et $j = i, ..., q$,

$$r_{x_i y_j} \simeq \underset{[\,2\,]}{r_{x_i y_j}} = \underset{[\,2\,]}{\mathbf{c}_\mathbf{i}'} \; \underset{[\,2\,]}{\mathbf{d_j}}$$

Alors la projection dans le graphique de chaque point $P_i = (c_{i1},\ c_{i2})$, $i = 1, ..., p$ sur le vecteur $\overrightarrow{OQ_j}$ où $Q_j = (d_{1j},\ d_{2j})$, $j = 1, ..., q$, nous renseigne sur la corrélation entre X_i et Y_j à un facteur constant $\|\mathbf{d}_j\|^2$ près.

Étant donné que nous avons déjà les valeurs de $\underset{p\times 2}{\mathbf{C}}$ et $\underset{q\times 2}{\mathbf{D}}$ (1.35 et 1.36),

$$\begin{aligned} \underset{p\times 2}{\mathbf{C}} &= (\mathbf{r}_{XV_1},\ \mathbf{r}_{XV_2}) \\ &= \begin{pmatrix} r_{X_1V_1} & r_{X_1V_2} \\ \vdots & \vdots \\ r_{X_pV_1} & r_{X_pV_2} \end{pmatrix} \end{aligned}$$

sont les points P_i de coordonnées $P_i = (r_{X_iV_1}\ \ r_{X_iV_2})$, $i = 1, .., p$ et

$$\begin{aligned} \underset{q\times 2}{\mathbf{D}} &= (\mathbf{r}_{YV_1},\ \mathbf{r}_{YV_2}) \\ &= \begin{pmatrix} r_{Y_1V_1} & r_{Y_1V_2} \\ \vdots & \vdots \\ r_{Y_qV_1} & r_{Y_qV_2} \end{pmatrix} \end{aligned}$$

sont les points Q_j de coordonnées $Q_j = (r_{Y_jV_1}\ \ r_{Y_jV_2})$, $j = 1, .., q$ que nous pouvons tracer directement dans le plan. Il ne reste qu'à projeter orthogonalement le point P_i sur

le vecteur $\overrightarrow{OQ_j}$ et multiplier la longueur de la projection par la longueur de $\overrightarrow{OQ_j}$ pour obtenir l'approximation de $r_{x_i y_j}$.

À noter que l'approximation de $r_{x_i y_j}$ sera négative si la projection de P_i sur $\overrightarrow{OQ_j}$ est faite sur le prolongement de $\overrightarrow{OQ_j}$ (direction opposée). Il faut donc multiplier le résultat par -1.

Soit Θ_{ij}, l'angle entre $\underset{[2]}{\mathbf{c_i'}}$ et $\underset{[2]}{\mathbf{d_j}}$:

- Si Θ_{ij} est obtus, $\cos\Theta_{ij} < 0$ alors $\underset{[2]}{\mathbf{c_i'}}\ \underset{[2]}{\mathbf{d_j}} < 0$,

$$r_{x_i y_j} \simeq \underset{[2]}{r_{x_i y_j}} < 0\ .$$

- Si Θ_{ij} est aigu, $\cos\Theta_{ij} > 0$ alors $\underset{[2]}{\mathbf{c_i'}}\ \underset{[2]}{\mathbf{d_j}} > 0$,

$$r_{x_i y_j} \simeq \underset{[2]}{r_{x_i y_j}} > 0\ .$$

- Si $\underset{[2]}{\mathbf{c_i'}} \perp \underset{[2]}{\mathbf{d_j}}$ alors $\underset{[2]}{\mathbf{c_i'}}\ \underset{[2]}{\mathbf{d_j}} = 0$,

$$r_{x_i y_j} \simeq \underset{[2]}{r_{x_i y_j}} = 0\ .$$

car

$$\|\mathbf{c}_i\|^2 \cos\Theta_{ij} = Proj\ (P_i\ sur\ \overrightarrow{OQ_j}\ ).$$

Ce qui est intéressant dans ce graphique est que nous sommes en mesure d'obtenir rapidement un ordre de grandeur des corrélations entre les variables de $\mathbf{X}$ et un Y_j donné ($r_{x_i y_j},\ i = 1, ..., p$) en comparant les projections des $P_i,\ i = 1, ..., p$, sur un des vecteurs $\overrightarrow{OQ_j}$ sans nécessairement calculer les vraies valeurs numériques. La longueur des vecteurs est donc très importante et permet d'analyser les corrélations et d'identifier très facilement les variables les plus corrélées.

Ce graphique biplot est un donc outil visuel très intéressant et **facile** à exploiter qui nous permettra de comprendre rapidement les relations entre les deux groupes de variables. Nous présenterons comment bien l'utiliser dans un cas réel au prochain chapitre.

1.7 Extensions

Comme nous l'avons constaté dans les sections précédentes, les statistiques produites par les corrélations canoniques comportent certaines problématiques. C'est pour cela que nous devons analyser d'autres mesures afin de bien interpréter les résultats. Plusieurs articles proposent ou contestent les mesures à utiliser pour interpréter les résultats. Les statistiques produites en analyse des corrélations canoniques ne sont pas tellement stables et beaucoup d'études actuelles portent sur la recherche de statistiques qui sont moins influencées par une légère perturbation dans les données. De plus, cette méthode est souvent comparée avec d'autres méthodes d'analyse multivariée et utilisée conjointement avec ces dernières. Nous présentons ici quelques constats des recherches et des lectures que nous avons effectuées sur le sujet.

1.7.1 Validation

Lorsque l'échantillon disponible est très grand, plusieurs auteurs suggèrent de le séparer en deux et d'appliquer la méthode sur une de ces deux parties. Nous pourrons par la suite valider les résultats obtenus en recalculant les mêmes statistiques avec l'autre partie. Cela nous permettra d'évaluer la robustesse des statistiques produites et de s'assurer des interprétations que nous ferons.

1.7.2 Valeurs extrêmes et robustesse

L'existence et la recherche des valeurs extrêmes (outliers) est un important sujet à l'étude avec la méthode des corrélations canoniques.

Rencher (1998) propose de tracer le graphique des (U_{1i}, V_{1i}) afin de détecter des valeurs extrêmes ou des relations non-linéaires.

Dans la pratique, la détection des valeurs extrêmes est très utile pour connaître un ensemble de données. Elle permet de comprendre des phénomènes spéciaux et d'être en mesure de les expliquer. Pemajayantha (1995, 2002) a étudié la détection de valeurs ex-

trêmes avec les modèles de corrélations canoniques appliqués aux techniques de contrôle statistique des processus (popularisé par Deming) qui est un sujet très intéressant en statistique appliquée. Ce sujet porte particulièrement sur la détection de valeurs extrêmes afin d'être en mesure de contrôler un processus. L'auteur a également appliqué les modèles de corrélations canoniques dans le contexte des processus stochastiques et déterministe afin de trouver également des valeurs extrêmes.

La présence de valeurs extrêmes (fréquent dans la pratique) implique que les données ne suivent pas nécessairement une loi normale multivariée, ce qui était une condition à l'application de la théorie des corrélations canoniques de l'échantillon à la population. Dans un tel cas, l'estimateur $\mathbf{S}$ ne représentera pas bien la matrice $\mathbf{\Sigma}$ (il n'est pas optimal) et les résultats de l'analyse des corrélations canoniques pourraient être totalement erronés car les calculs reposent beaucoup sur $\mathbf{\Sigma}$. Si les valeurs extrêmes font effectivement parties de l'ensemble de données et ne sont pas des "erreurs", nous devrons opter pour des estimateurs plus robustes de $\mathbf{\Sigma}$. Nous obtiendrons alors une analyse des corrélations canoniques robustes. Campbell (1982) utilise le M-estimateur et des fonctions de poids basées sur des distances robustes tandis que Karnel (1992) utilise seulement le M-estimateur pour estimer de façon robuste $\mathbf{\Sigma}$. Croux et Dehon (2002) proposent deux estimateurs robustes de $\mathbf{\Sigma}$ à plus haut point de rupture, le "Minimum Covariance Determinant" (MCD) et le "Reweighted Minimum Covariance Determinant" (RMCD).

CHAPITRE II

APPLICATION DE L'ANALYSE DES CORRÉLATIONS CANONIQUES ET DU BIPLOT À UN ENSEMBLE DE DONNÉES

Pour montrer l'utilité de l'application de l'analyse canonique et de son interprétation à l'aide des corrélations et du biplot, il nous fallait un ensemble de données réel (non simulé) et volumineux en terme de variables et d'observations. De plus, nous voulions des données autres qu'en biologie ou en écologie qui sont des domaines où cette méthode est la plus appliquée.

Nous avons obtenu des données qui répondaient à ces critères à partir du site Internet Sherlock (http ://sherlock.crepuq.qc.ca) qui est un système d'accès coopératif aux données numériques dans les bibliothèques universitaires québécoises. Ce site permet l'accès à des **microdonnées** (observations de variables qui n'ont pas été résumées) de différents organismes publics. Sherlock est conçu pour des utilisateurs ayant une certaine expérience des traitements des microdonnées. En plus des fichiers de microdonnées, le site contient également la description de l'enquête (but, questionnaire, etc.), un guide de l'utilisateur pour comprendre et traiter adéquatement les données (définition, fréquence, clés, etc.) ainsi que certains programmes qui permettent de lire les données avec les logiciels SAS et SPSS.

Dans ce chapitre, nous allons présenter l'ensemble de données retenu extrait du site Sherlock et appliquer sur cet ensemble les méthodes de l'analyse des corrélations canoniques et du graphique biplot décrites au chapitre précédent. Nous allons donc voir

qu'il est simple d'explorer un ensemble de données avec ces méthodologies. Nous avons utilisé le logiciel SAS pour faire cette application.

2.1 Présentation de l'ensemble de données

À partir du site Internet Sherlock, nous avons choisi comme ensemble de micro-données l'Enquête sur la dynamique du travail et du revenu (EDTR) de 2002 provenant de Statistique Canada. Cette enquête est très intéressante car elle contient différentes variables des domaines du revenu, du travail et de la famille. L'application de l'analyse des corrélations canoniques est donc très appropriée dans ce cas-ci car elle suppose que les variables se séparent naturellement en deux groupes. Ainsi, nous regrouperons d'une part les variables sociodémographiques (caractéristiques sur les personnes, leur famille et leur travail) et d'une autre part, les variables à caractère fiscal (salaire, impôt, cotisation REER, etc.). L'application de l'analyse des corrélations canoniques nous aidera donc à comprendre la dynamique qui relie le travail au revenu qui est le sujet principal de l'enquête.

Nous avons choisi le fichier des personnes (EC2002PR.DAT) qui contient un enregistrement pour chaque individu faisant partie de l'échantillon de l'enquête (individu âgé de 16 ans et plus). Ce fichier contient au total 56 216 enregistrements (individus) et 130 variables.

2.1.1 Individus retenus

La réalité socio-économique du Canada est différente d'une province à l'autre et nous ne connaissons pas nécessairement toutes ces réalités. Afin de rendre l'interprétation des résultats plus concrète, nous avons décidé de ne conserver que les individus de la province du **Québec**.

Comme nous nous intéressons à la dynamique entre le travail et le revenu, seules les personnes ayant déclaré que leur activité principale durant l'année 2002 était le *travail pour un emploi ou en affaires* ont été conservées. Nous rendons ainsi les données plus

homogènes en excluant les personnes à la retraite et les étudiants. À noter qu'il y a tout de même des personnes dans la catégorie *travail pour un emploi ou en affaires* qui sont devenus retraités ou qui sont étudiants (temps partiel) mais ce n'était pas leur activité principale durant l'année 2002.

Pour des raisons évidentes, nous avons exclu quelques individus (moins de 1%) dont certaines valeurs de variables étaient exorbitantes par rapport au reste de l'échantillon. Par exemple, des revenus de placements supérieurs à 100 000 $, des pensions alimentaires supérieures à 20 000 $, etc. Après tous ces traitements, nous obtenons un fichier final à exploiter de 3 681 individus.

2.2 Présentation des variables

Le fichier total contient 130 variables qui sont séparées en trois grands thèmes :

1. **Caractéristiques personnelles** : Démographie, famille, instruction
2. **Travail** : Emploi, expérience, employeur
3. **Revenu** : Source de revenu, prestation, impôt

Plusieurs variables dans le fichier sont utilisées à des fins descriptives, c'est-à-dire qu'elles servent à subdiviser les individus en quelques sous-groupes. Bien qu'intéressante pour des statistiques descriptives générales, ces variables catégoriques sont moins pertinentes dans notre genre d'analyse. D'ailleurs, nous nous intéressons à des variables de type quantitatif afin d'observer des liens qui peuvent en ressortir. Nous conserverons tout de même quelques variables de type catégorique que nous recoderons 1 ou 0 (présence/absence). Nous ferons les applications avec et sans ces variables binaires pour en observer les impacts sur les corrélations canoniques.

De plus, plusieurs variables quantitatives sont fonction d'autres variables dans le fichier. Par exemple, le revenu après impôt représente le revenu total moins les impôts à payer qui sont tous deux des variables déjà présentes dans le fichier. Nous avons donc exclus ces variables qui sont linéairement dépendantes car elles ne nous apportent pas d'information supplémentaire.

Comme la majorité des variables catégoriques se retrouvent dans les thèmes 1 et 2 (caractéristiques et travail), nous les avons regroupées. Nous appellerons ce premier groupe de variables le groupe socio-démographique et ils seront représentées par les $\mathbf{X}$. Nous avons ainsi retenu **18** variables dont 10 variables quantitatives et 8 variables binaires présentées respectivement aux tableaux 2.1 et 2.2. À noter que les définitions des variables correspondent à la situation de l'individu interrogé à la **fin** de l'année 2002 sauf dans les cas indiqués.

Le deuxième groupe de variables, les $\mathbf{Y}$, contiendra **16** variables relatives au revenu. Ce groupe sera appelé le groupe des variables fiscales. Certaines de ces variables sont des sommes de quelques variables d'un même sujet. Par exemple, les crédits pour enfant sont la somme des pensions alimentaires payées plus les frais de garde. Toutes ces variables correspondent à des montants en dollars. Les définitions, moyennes et écart-types de ce deuxième groupe sont présentés dans le tableau 2.3.

2.2.1 Corrélations

Nous présentons au tableau 2.4 la matrice des corrélations entre les $\mathbf{X}$ et les $\mathbf{Y}$ ($\mathbf{R}_{XY}$). Afin d'expliquer et comprendre cette matrice de corrélation, nous utiliserons le graphique biplot qui estime $\mathbf{R}_{XY}$.

Tableau 2.1 Variables quantitatives du premier groupe **X** (sociodémographiques)

X	**Nom**	**Définition**	**Moy**	**ET**
X_1	Âge	Âge de la personnes	41.21	11.59
X_3	Nb ménage	Nombre de personnes dans le ménage	2.74	1.29
X_4	Nb étude	Nombre total d'années d'études complétées	13.26	3.50
X_5	Nb étude ps	Nombre d'années d'études postsecondaires	2.34	2.40
X_6	Nb étude uni	Nombre d'années d'études universitaires	0.88	1.75
X_9	Nb emp	Nombre d'emplois occupés pendant l'année	1.18	0.49
X_{11}	Hres trav	Nombre total d'heures rémunérées durant l'année	1 845.6	662.2
X_{12}	Hres tmais	Nombre d'heures par semaine travaillées à la maison	1.77	6.89
X_{15}	Nb sem	Nombre de semaines travaillées	49.44	10.45
X_{16}	Expérience	Nombre d'années d'expérience sur le marché du travail	17.58	10.86

Tableau 2.2 Variables binaires du premier groupe **X** (sociodémographiques)

X	**Nom**	**Définition**	**Fréquence relative**	
			1 (présence)	**0** (absence)
X_2	Couple	Égale à 1 si la personne était en couple, 0 sinon	0.60	0.40
X_7	Rurale	Égale à 1 si la personne habite en région rurale (< 30 000 pers.), 0 sinon	0.20	0.80
X_8	Loué	Égale à 1 si la personne habite dans un logement loué, 0 sinon	0.31	0.69
X_{10}	Arrêt	Égale à 1 si la personne a arrêté de travailler durant l'année, 0 sinon	0.16	0.84
X_{13}	Hor irr	Égale à 1 si l'horaire de travail est irrégulier (autre que le jour), 0 sinon	0.36	0.64
X_{14}	T part	Égale à 1 si la personne a travaillé à temps partiel durant l'année, 0 sinon	0.14	0.86
X_{17}	Reg ret	Égale à 1 si la personne adhère à un régime de retraite, 0 sinon	0.40	0.60
X_{18}	Public	Égale à 1 si l'employeur de la personne est du secteur public, 0 sinon	0.21	0.79

Tableau 2.3 Variables du deuxième groupe **Y** (fiscales)

Y	**Nom**	**Définition**	**Moy** ($)	**ET** ($)
Y_1	Salaire	Salaire et traitements (revenu d'emploi)	31 557	23 524
Y_2	Trav auto	Revenu d'un travail autonome	1 924	11 389
Y_3	Autre rev	Autre revenu	244.85	1 923
Y_4	Âgées	Retrait d'un REER + pension de retraite + pension de la sécurité vieillesse	614.5	3 641
Y_5	Prestation	Indemnité pour accident + assurance emploi + assistance sociale	1 047	2 658
Y_6	Transfert	Transferts gouvernementaux	1 967	3 480
Y_7	Enfant	Pension alimentaire reçue + prestations pour enfant	561.2	1874
Y_8	Placement	Gains en capital + revenu de placements	831.5	4 840
Y_9	Imp fed	Impôt fédéral	3 421	3 802
Y_{10}	Imp prov	Impôt provincial	3 668	4 375
Y_{11}	Cred TPS	Crédit de la TPS	116.7	186.7
Y_{12}	Cred prov	Crédit d'impôt provincial (TVQ)	85.9	178.5
Y_{13}	C enfant	Pension alimentaire payé + frais de garde	470.0	1 432.5
Y_{14}	Frais med	Frais médicaux	415.0	784.5
Y_{15}	Cotisation	Cotisations syndicales/association + cotisation régimes pensions agrées	803.1	1 237
Y_{16}	Cotis oblig	Cotisations au régime des rentes du Québec + cotisation à l'assurance emploi	1 643.0	861.6

Tableau 2.4 Corrélations entre les $\mathbf{X}$ (sociodémographiques) et $\mathbf{Y}$ (fiscales), $\mathbf{R}_{XY}$

Rxy	Y01	Y02	Y03	Y04	Y05	Y06	Y07	Y08
	Salaire	Trav auto	Autre rev	Agees	Prestation	Transfert	Enfant	Placement
X01 Age	0.10	0.03	0.03	0.22	-0.03	0.09	-0.07	0.12
X02 Couple	0.09	0.07	-0.01	0.04	-0.02	-0.08	-0.13	0.05
X03 Nb menage	0.02	0.04	-0.00	-0.05	0.02	0.06	0.19	-0.01
X04 Nb etude	0.27	0.09	0.02	-0.02	-0.14	-0.20	-0.01	-0.01
X05 Nb etude ps	0.25	0.10	0.01	-0.00	-0.12	-0.16	-0.03	-0.00
X06 Nb etude uni	0.26	0.14	0.03	0.06	-0.09	-0.13	-0.06	0.01
X07 Rurale	-0.09	-0.01	-0.03	-0.02	0.08	0.09	0.01	0.01
X08 Loue	-0.14	-0.06	-0.00	-0.04	0.04	0.06	0.08	-0.07
X09 Nb emp	-0.09	-0.01	-0.01	-0.06	0.10	0.02	-0.01	-0.05
X10 Arret	-0.28	-0.05	0.04	0.06	0.27	0.28	0.04	-0.04
X11 Hres trav	0.26	0.08	-0.03	-0.09	-0.16	-0.24	-0.09	0.03
X12 Hres tmais	-0.07	0.21	0.01	0.06	-0.04	-0.04	0.00	0.10
X13 Hor irr	-0.26	0.21	0.02	0.07	-0.03	0.03	-0.00	0.10
X14 T part	-0.24	0.05	0.02	0.09	0.01	0.14	0.08	0.04
X15 Nb sem	0.29	0.03	-0.02	-0.09	-0.26	-0.33	-0.06	0.02
X16 Experience	0.18	0.01	0.00	0.18	-0.05	0.01	-0.10	0.12
X17 Reg ret	0.45	-0.13	-0.00	-0.05	-0.08	-0.15	-0.04	-0.05
X18 Public	0.26	-0.08	0.01	-0.03	-0.06	-0.10	-0.03	-0.03
	Y09	Y10	Y11	Y12	Y13	Y14	Y15	Y16
	Imp fed	Imp prov	Cred tps	Cred prov	C enfant	Frais med	Cotisation	Cotis oblig
X01 Age	0.15	0.17	-0.14	-0.05	-0.05	0.11	0.13	0.01
X02 Couple	0.12	0.10	-0.32	-0.07	-0.02	0.08	0.04	0.10
X03 Nb menage	0.01	-0.04	0.00	0.04	0.10	0.07	0.01	0.03
X04 Nb etude	0.26	0.27	-0.16	-0.17	0.06	0.05	0.27	0.29
X05 Nb etude ps	0.26	0.28	-0.17	-0.16	0.05	0.05	0.27	0.27
X06 Nb etude uni	0.31	0.33	-0.17	-0.14	0.04	0.04	0.28	0.23
X07 Rurale	-0.09	-0.10	0.06	0.09	0.01	-0.01	-0.07	-0.08
X08 Loue	-0.16	-0.16	0.22	0.12	-0.03	-0.09	-0.12	-0.14
X09 Nb emp	-0.10	-0.10	0.10	0.06	-0.03	-0.03	-0.06	-0.08
X10 Arret	-0.22	-0.22	0.19	0.16	-0.06	-0.03	-0.19	-0.37
X11 Hres trav	0.22	0.21	-0.08	-0.04	0.07	-0.03	0.08	0.37
X12 Hres tmais	0.06	0.07	-0.04	-0.02	0.02	0.02	-0.01	-0.03
X13 Hor irr	-0.11	-0.10	0.07	0.08	-0.01	-0.00	-0.14	-0.23
X14 T part	-0.14	-0.13	0.04	0.04	-0.06	0.03	-0.10	-0.34
X15 Nb sem	0.21	0.20	-0.15	-0.14	0.06	0.03	0.18	0.41
X16 Experience	0.20	0.21	-0.14	-0.06	-0.03	0.08	0.15	0.15
X17 Reg ret	0.29	0.30	-0.24	-0.22	0.07	0.04	0.58	0.46
X18_Public	0.16	0.18	-0.19	-0.16	-0.01	0.05	0.47	0.27

2.3 Calculs et résultats des corrélations canoniques

Nous allons maintenant appliquer la technique de l'analyse des corrélations canoniques à notre ensemble de données à l'aide du logiciel SAS et de sa procédure *CANCORR*. La programmation de cette procédure et les paramètres que nous avons utilisés sont donnés dans l'appendice A.

L'option *ALL* de la procédure *CANCORR* nous donne les statistiques suivantes :

1. Statistiques descriptives des deux groupes de variables : moyenne, écart-type et matrices des corrélations ;
2. Coefficients de corrélations canoniques et tests séquentiels sur ce coefficient ;
3. Poids canoniques et poids canoniques des variables originales standardisées ;
4. Intra/inter-corrélations des deux groupes de variables
5. Indices de variance expliquée et indice de redondance (prochain chapitre) des variables originales standardisées et non standardisées ;
6. Coefficient de l'inter-corrélation au carré des deux groupes de variables.

Les valeurs des variables canoniques $\mathbf{U}$ et $\mathbf{V}$ pour chaque observation sont disponibles dans le fichier de sortie *out cancor*.

Nous allons maintenant présenter les résultats de cette procédure qui sont disponibles dans le *Output* SAS ou dans le fichier de sortie créé *out stat cancor*. Nous interpréterons plus concrètement ces résultats à la prochaine section avec nos techniques présentées au chapitre précédent (graphique biplot et intra/inter-corrélation).

2.3.1 Coefficients de corrélation canonique

Au tableau 2.5, nous présentons les coefficients de corrélation canonique calculés par SAS. Comme nous l'avons mentionné dans le chapitre précédent, nous avons au plus $s = min(p, q)$ coefficients de corrélations canoniques, c'est à dire $16 = min(18, 16)$ coefficients de corrélations canoniques qui correspondent aux racines carrées des valeurs

Tableau 2.5 Coefficients de corrélation canonique, output SAS

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.738364	0.735484	0.007497	0.545181
2	0.537446	0.530253	0.011723	0.288848
3	0.478715	0.471071	0.012707	0.229168
4	0.439207	0.433522	0.013305	0.192903
5	0.385865	0.380167	0.014030	0.148892
6	0.312419	0.304805	0.014876	0.097606
7	0.240907	0.230962	0.015528	0.058036
8	0.181107	.	0.015944	0.032800
9	0.132882	.	0.016193	0.017658
10	0.099817	.	0.016320	0.009963
11	0.070788	.	0.016402	0.005011
12	0.057346	.	0.016430	0.003289
13	0.055782	.	0.016433	0.003112
14	0.046721	.	0.016449	0.002183
15	0.041135	.	0.016457	0.001692
16	0.011719	.	0.016482	0.000137

propres communes (triées dans l'ordre décroissant) des matrices $\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$ et $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}$.

Ces coefficients représentent la force des relations entre les groupes de variables. Nous pouvons remarquer rapidement qu'il existe au moins une forte relation entre les groupes avec le premier coefficient, $\lambda_1 = 0.738$. Nous avons dans la dernière colonne le carré de ces coefficients qui sont les valeurs propres. Certains auteurs interprètent ces valeurs comme étant la variance partagée par les variables canoniques.

Dans la deuxième colonne, la procédure nous fournit le coefficient de corrélation ajusté qui est un coefficient asymptotiquement moins biaisé que le premier mais qui ne peut être toujours calculé (Lawley, 1959). Il y a très peu de différence entre ce coefficient et le premier car notre taille d'échantillon est grande. La troisième colonne nous donne l'erreur-standard approximative du coefficient. Elle est assez petite due également à notre taille d'échantillon importante.

2.3.2 Réduction de dimension et test d'association

Le but de l'analyse des corrélations canoniques est de réduire les $p \times q$ (18×16) relations en au plus $s = min(p, q)$ relations (16) entre les variables. Comme nous l'avons vu au tableau 1.1, il est possible de tester séquentiellement les coefficients de corrélations canoniques. Les résultats de ces tests séquentiels sont fournis au tableau 2.6. Nous remarquons qu'à 0.05, il y a seulement 10 coefficients de corrélations canoniques statistiquement significatifs.

Bien que 10 coefficients soient statistiquement significatifs, nous nous concentrerons sur les relations dont le coefficient possède une valeur concrètement intéressante. Ainsi, nous étudierons en premier lieu les coefficients supérieurs à 0.40 qui correspondent aux quatre premières relations. Ces quatre relations semblent se démarquer des autres. Nous réduisons donc dans un premier temps le problème de 288 relations (18×16) entre les variables originales à 4 relations entre les variables canoniques :

Tableau 2.6 Tests de signification, output SAS

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.13484097	27.64	288	43844	<.0001
2	0.29647205	18.38	255	41268	<.0001
3	0.41688988	14.87	224	38691	<.0001
4	0.54083100	11.88	195	36112	<.0001
5	0.67009417	8.90	168	33532	<.0001
6	0.78731991	6.20	143	30950	<.0001
7	0.87247889	4.19	120	28366	<.0001
8	0.92623384	2.84	99	25780	<.0001
9	0.95764439	1.98	80	23190	<.0001
10	0.97485795	1.48	63	20597	0.0079
11	0.98466864	1.18	48	17998	0.1851
12	0.98962767	1.09	35	15390	0.3267
13	0.99289289	1.09	24	12766	0.3470
14	0.99599201	0.98	15	10104	0.4726
15	0.99817084	0.84	8	7322	0.5687
16	0.99986266	0.17	3	3662	0.9182

1. $\lambda_1 = 0.738364$
2. $\lambda_2 = 0.537446$
3. $\lambda_3 = 0.478715$
4. $\lambda_4 = 0.439207$

À noter que les tests séquentiels supposent la normalité des variables ce qui n'est pas le cas pour toutes nos variables (binaires, 1/0). Cependant, certains auteurs mentionnent qu'avec une taille d'échantillon élevée comme dans notre cas, les tests sont valables de façon approximative (Kshirsagar, 1972).

2.3.3 Vecteurs des variables canoniques et intra/inter-corrélations

Comme nous l'avons mentionné au chapitre précédent, les vecteurs canoniques peuvent être obtenus en utilisant les variables originales standardisées et non standardisées. Ces deux résultats sont disponibles par la procédure *CANCORR* de SAS ainsi que les coefficients de l'intra/inter-corrélation qui ne changent pas d'échelle en fonction de la standardisation.

Pour alléger la lecture de cette section, tous les tableaux des résultats suivants sont disponibles à la **fin** de ce chapitre (section 2.8).

Variables originales non standardisées

Les poids des vecteurs canoniques dans ce cas sont proportionnels aux unités de mesures des variables originales et de leur distribution. Ils peuvent donc être assez difficiles d'interprétation car il faut toujours avoir en tête comment les variables originales se distribuent (tableaux 2.1, 2.2 et 2.3). Nous présentons ces résultats aux tableaux 2.10 et 2.11 (quatre premières relations seulement) qui permettront aux lecteurs de remarquer les différences avec les vecteurs canoniques programmés à partir des variables originales standardisées.

À noter qu'il est très difficile d'interpréter les poids canoniques du deuxième groupe car

ils possèdent tous de très petites valeurs.

Variables originales standardisées

Pour faciliter la comparaison des poids des vecteurs canoniques entre les variables canoniques, il est préférable d'utiliser les résultats obtenus avec les variables originales standardisées. Ceci est particulièrement vrai dans un cas comme le nôtre où les variables originales ont des unités de mesures et des distributions très différentes. Nous privilégierons plutôt les résultats présentés aux tableaux 2.12 et 2.13 que ceux présentés aux tableaux 2.10 et 2.11.

2.3.4 Intra-corrélation

Nous fournissons également à la fin de ce chapitre (tableaux 2.14 et 2.15 de la section 2.8) les coefficients de l'intra-corrélation (corrélation entre X_i et U_k ou entre Y_j et V_k). Ces coefficients nous permettront de déterminer les variables originales qui sont les mieux représentées par leurs propres variables canoniques. Ces coefficients sont utilisés dans le graphique biplot et seront très utiles pour l'interprétation.

2.3.5 Inter-corrélation

Le **carré** de ce coefficient représente la proportion de variance d'une variable originale expliquée par une variable canonique de l'autre groupe. Il nous aidera donc à comprendre davantage nos relations dans l'ensemble de données avec une meilleure compréhension des variables canoniques en terme de prédiction. Ces coefficients, présentés aux tableaux 2.16 et 2.17 de la section 2.8, seront également utilisés dans le graphique biplot.

2.3.6 Validation

Comme nous l'avons déjà mentionné, les coefficients canoniques et les vecteurs canoniques ne sont pas tellement stables. Une légère perturbation dans l'ensemble de

Tableau 2.7 Comparaison des corrélations canoniques

	Corrélations canoniques			
nombre d'observations	λ_1	λ_2	λ_3	λ_4
3 681 (total)	0.738364	0.537446	0.478715	0.439207
2 000 (sous-échantillon)	0.729839	0.549007	0.477532	0.430584
1 681 (validation)	0.750871	0.540381	0.498726	0.462138

données peut générer des résultats assez différents. Pour contrer cette instabilité, il est conseillé d'utiliser les coefficients de l'intra-corrélation afin de mieux comprendre les relations canoniques.

Étant donné que la taille de l'échantillon disponible est très grande, nous avons séparé l'échantillon en deux pour valider les résultats qui ne seraient pas stables tel qu'il est proposé par certains auteurs.

Nous avons créé un sous-échantillon de 2 000 individus choisis aléatoirement et un autre sous-ensemble de validation avec le reste des individus (1681). Nous avons comparé les quatre premières corrélations canoniques des deux ensembles à notre ensemble total (voir tableau 2.7). Nous obtenons des valeurs assez proches ce qui nous permet de croire que les relations semblent se reproduire dans les trois ensembles.

Cependant, nous avons aussi comparé les vecteurs canoniques de ces trois ensembles pour s'assurer de la stabilité. Les couples de vecteurs canoniques (U_3, V_3) se comportent très différemment surtout pour l'ensemble de 1 681 données. Il y a également quelques écarts trouvés pour le couple (U_4, V_4) dans les deux nouveaux ensembles. Cette comparaison nous rappelle donc la vigilance que nous devons apporter pour interpréter ces deux relations (3 et 4) et l'importance d'une bonne validation des résultats. De plus, nous pouvons croire que moins la relation est forte (coefficient de corrélation canonique faible), plus il est difficile de dégager des tendances générales dans l'ensemble de données.

2.3.7 Corrélations canoniques sans les variables binaires

Nous avons conservé dans notre application des variables catégoriques que nous avons recodées en binaire 1/0 pour présence ou absence. L'application sans les variables binaires nous donne les quatre premiers coefficients de corrélations canoniques suivants :

1. $\lambda_1 = 0.591859$
2. $\lambda_2 = 0.491493$
3. $\lambda_3 = 0.431998$
4. $\lambda_4 = 0.269041$

Nous pouvons donc constater que la présence des variables binaires apporte de l'information supplémentaire et semble renforcir les liens canoniques entre les groupes de variables originales (coefficients de corrélations canoniques plus élevés).

Nous avons également étudié les coefficients de l'intra-corrélation de ces quatre premières relations canoniques sans les variables binaires. Nous avons constaté que les 2 premières relations nous donnent des résultats similaires mais que les relations 3 et 4 sont différentes. Étant donné que les coefficients de corrélations canoniques de ces deux relations sont plus faibles et que nous venons de voir que ces relations sont moins stables, nous pouvons croire que la présence des variables binaires peut facilement influencer ces relations dans l'ensemble de données. Il serait intéressant de comparer ces relations avec les relations 5, 6 ou 7 dans l'application avec toutes les variables et tenter d'en ressortir quelques constats.

2.3.8 Valeurs extrêmes et robustesse

À la figure 2.1, nous avons tracé le graphique des (U_1, V_1) de chaque individu comme le propose Rencher (1998) afin de détecter des valeurs extrêmes ou des relations non-linéaires.

Nous pouvons constater rapidement qu'il y semble y avoir quelques valeurs ex-

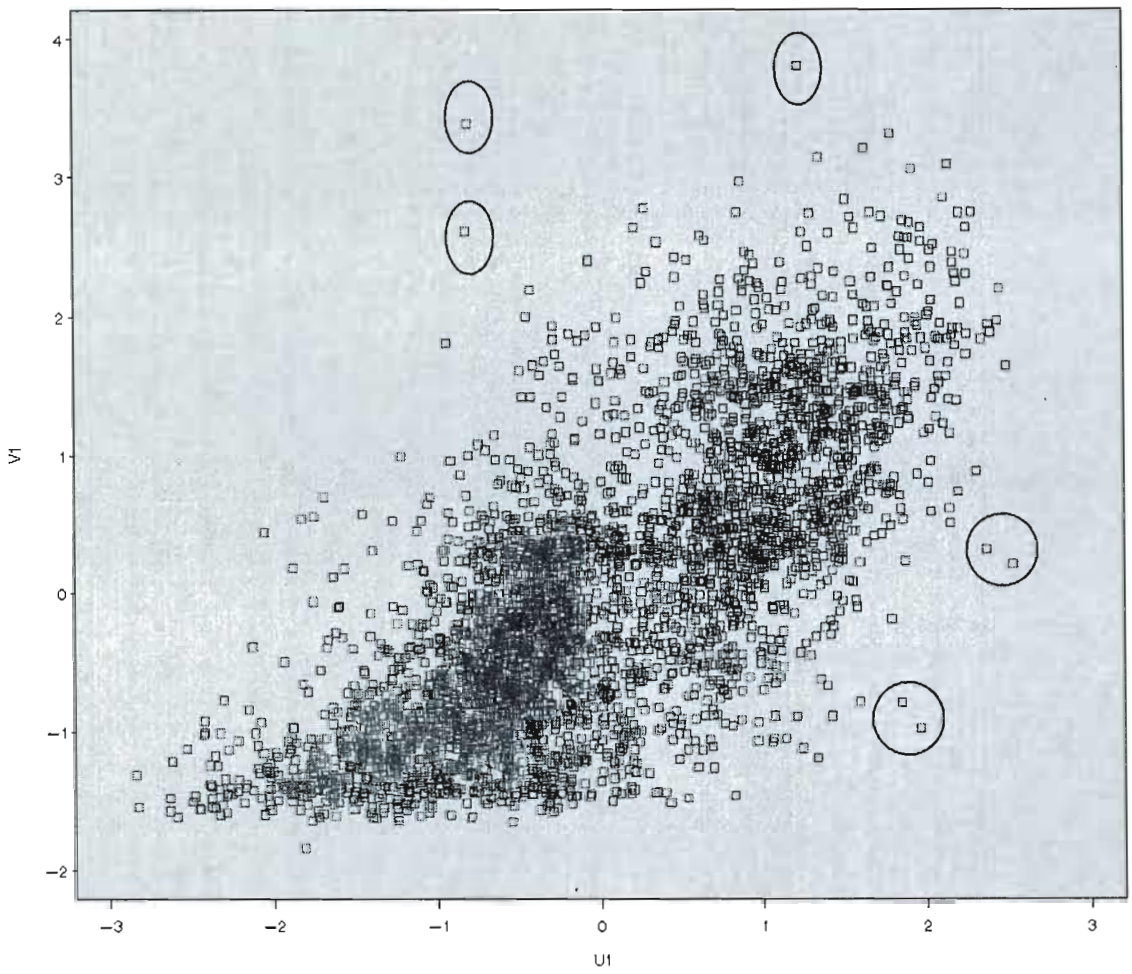


Figure 2.1 Graphique des (U_1, V_1) pour chaque individu

trêmes que nous avons encerclées sur le graphique. La présence de valeurs extrêmes est très fréquente dans la pratique et le calcul des variables canoniques nous permet de les identifier rapidement.

Nous avons calculé les corrélations canoniques sans ces valeurs extrêmes. Étant donné que la taille de notre échantillon est élevée, les corrélations canoniques sans ces valeurs extrêmes sont presque identiques pour les deux premières relations mais un peu différentes pour les relations 3 et 4. Cela confirme encore une fois l'instabilité des relations 3 et 4. Il aurait été intéressant de comparer les résultats avec une analyse canonique robuste comme discuter à la section 1.7.2 qui permet de mieux estimer $\mathbf{\Sigma}$.

Nous pouvons remarquer également dans le graphique (figure 2.1) qu'il semble y avoir une légère relation non-linéaire. Il aurait intéressant d'appliquer des transformations sur les variables originales qui influencent le plus cette tendance et observer les impacts sur les corrélations canoniques.

2.4 Estimation de $\mathbf{R}_{XY}$ par le graphique biplot

Avant de discuter des relations canoniques trouvées précédemment, nous allons voir par l'application de la théorie de l'article de ter Braak que nous avons détaillée au chapitre précédent (section 1.5), comment il est possible de restructurer la matrice $\mathbf{R}_{XY}$ présentée au tableau 2.4. Cette restructuration est faite à partir du graphique biplot construit avec les coefficients de l'intra/inter-corrélation de V_1 et V_2 (tableaux 2.15 et 2.17). Ce graphique nous donne des estimations des $r_{X_iY_J}$. Nous allons voir qu'il est alors très facile d'étudier et de comprendre les relations entre les deux groupes de variables dans le graphique biplot au lieu d'essayer d'extraire cette information directement à partir de la matrice $\mathbf{R}_{XY}$ de dimension 18×16. Dans la prochaine section, nous allons tenter d'interpréter les résultats des relations canoniques à l'aide de ce graphique biplot.

Au chapitre I, nous avions déterminé que

$$\begin{aligned}
\underset{p \times q}{\mathbf{R}_{XY}} &= (\underset{p \times s}{\mathbf{R}_{XX}^{1/2}} \mathbf{P} \mathbf{\Lambda}) (\underset{s \times q}{\mathbf{Q}' \mathbf{R}_{YY}^{1/2}}) \\
&= (\underset{p \times s}{Corr(\mathbf{X}, \mathbf{V})}) (\underset{s \times q}{Corr(\mathbf{V}, \mathbf{Y})}) \\
&\simeq \underset{[2]}{\mathbf{R}_{XY}} = \underset{p \times 2}{\mathbf{C}} \underset{2 \times q}{\mathbf{D}}{}'
\end{aligned}$$

où $\underset{[2]}{\mathbf{R}_{XY}}$ est une matrice de dimension $p \times q$ de rang 2,

$$\underset{p \times 2}{\mathbf{C}} = \begin{pmatrix} r_{X_1 V_1} & r_{X_1 V_2} \\ \vdots & \vdots \\ r_{X_p V_1} & r_{X_p V_2} \end{pmatrix}$$

et

$$\underset{q \times 2}{\mathbf{D}} = \begin{pmatrix} r_{Y_1 V_1} & r_{Y_1 V_2} \\ \vdots & \vdots \\ r_{Y_q V_1} & r_{Y_q V_2} \end{pmatrix}$$

Nous avons ainsi les coordonnées des points $P_i = (r_{X_i V_1} \quad r_{X_i V_2})$, $i = 1, .., p$ et les coordonnées des points $Q_j = (r_{Y_j V_1} \quad r_{Y_j V_2})$, $j = 1, .., q$.

Donc à partir des coefficients de l'inter-corrélation du tableau 2.17 et des coefficients de l'intra-corrélation du tableau 2.15, nous obtenons :

$$\underset{18\times 2}{\mathbf{C}} = \begin{pmatrix} 0.09885 & 0.36056 \\ 0.09944 & 0.17563 \\ 0.01714 & -0.15984 \\ 0.32583 & 0.00006 \\ 0.30724 & 0.06290 \\ 0.29515 & 0.14026 \\ 0.09237 & -0.01831 \\ 0.16096 & -0.13022 \\ 0.09022 & -0.09451 \\ 0.33504 & 0.08051 \\ 0.25706 & -0.19036 \\ 0.03624 & 0.10388 \\ 0.23142 & 0.15006 \\ 0.24942 & 0.23766 \\ 0.34630 & -0.17165 \\ 0.18537 & 0.23482 \\ 0.60351 & 0.01090 \\ 0.42840 & 0.09071 \end{pmatrix} \quad \text{et} \quad \underset{16\times 2}{\mathbf{D}} = \begin{pmatrix} 0.80690 & -0.05989 \\ -0.07925 & 0.21342 \\ -0.00956 & 0.06499 \\ -0.05890 & 0.43758 \\ -0.24856 & 0.02348 \\ -0.39038 & 0.16735 \\ -0.13623 & -0.30095 \\ -0.02225 & 0.22313 \\ 0.60276 & 0.18996 \\ 0.61307 & 0.26579 \\ -0.46348 & -0.40664 \\ -0.37869 & -0.13835 \\ 0.10975 & -0.18475 \\ 0.09143 & 0.15555 \\ 0.84527 & 0.20580 \\ 0.86725 & -0.24763 \end{pmatrix}$$

Ces matrices $\mathbf{C}$ et $\mathbf{D}$ contiennent les coordonnées des points P_1 à P_{18} d'une part et les points Q_1 à Q_{16} d'autre part. Les $\mathbf{X}$ seront représentés par P_1 à P_{18} (corrélations des $\mathbf{X}$ avec V_1 et V_2) et les $\mathbf{Y}$ par Q_1 à Q_{16} (corrélations $\mathbf{Y}$ des avec V_1 et V_2).

À la figure 2.2, nous avons tracé les points P_i et Q_j à l'aide de la procédure *GPLOT* de *SAS/GRAPH*. Il est très facile de tracer ce graphique car nous avons déjà les points P_i et Q_j (intra/inter-corrélations) dans un fichier obtenu par la procédure *CANCORR*. Notre programme est disponible dans l'appendice A.

Pour obtenir les estimations de $r_{x_i y_j}$, il ne reste qu'à projeter orthogonalement le point P_i sur le vecteur $\overrightarrow{OQ_j}$ et multiplier la longueur de la projection par la longueur de $\overrightarrow{OQ_j}$.

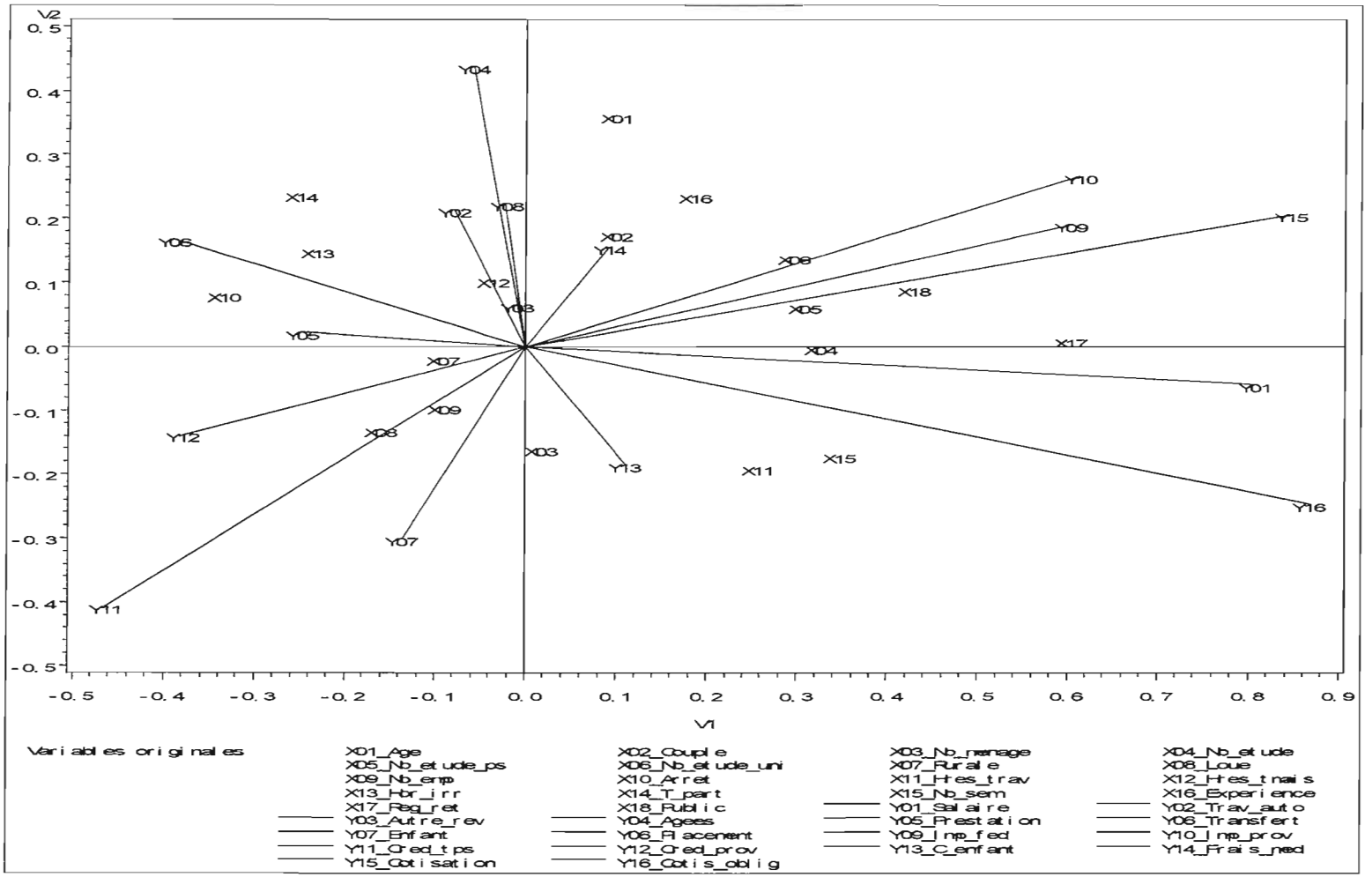
La valeur de la corrélation est négative si l'angle entre $\overrightarrow{OP_j}$ et $\overrightarrow{OQ_j}$ est obtus (direction opposée).

$$r_{x_i y_j} \simeq \underset{[\,2\,]}{r_{x_i y_j}} = \|\overrightarrow{OQ_j}\|^2 Proj\ (P_i\ sur\ \overrightarrow{OQ_j}\ ).$$

où $Proj\ (P_i\ sur\ \overrightarrow{OQ_j}\ ) = \|\overrightarrow{OP_i}\|^2 \cos\Theta_{ij}$, donc < 0 si Θ_{ij} est obtus (direction opposée).

Dans ce graphique, nous ne recherchons pas nécessairement à obtenir les vrais valeurs de $r_{x_i y_j}$ car elles se calculent très facilement avec les logiciels statistiques. Nous voulons plutôt **comparer** les variables entre elles et rechercher une structure des corrélations données par ce graphique. Nous pouvons donc **observer facilement** le comportement des variables dans nos deux groupes de données par l'étude du graphique.

Figure 2.2 Graphique biplot, estimation de $\mathbf{R}_{XY}$ par intra/inter-corrélation de V_1 et V_2



2.4.1 Réorganisation de la matrice $\mathbf{R}_{XY}$

À l'aide du graphique biplot de la figure 2.2, nous avons pu restructurer la matrice $\mathbf{R}_{XY}$ du tableau 2.4. Nous allons expliquer comment nous sommes arrivés facilement à la restructurer.

En observant le biplot de la figure 2.2, il se dégage des tendances évidentes dans les longueurs des vecteurs $\overrightarrow{OP_i}$ et $\overrightarrow{OQ_j}$ qui représentent respectivement les X_i et les Y_j. Nous avons ordonné les Y_j selon leur ordre d'importance donné par la longueur des vecteurs $(\overrightarrow{OQ_j})$ et leur positionnement dans le graphique. Nous avons débuté avec Y_{16} (le plus long vecteur) et nous avons continué dans le sens anti-horaire. Lorsque les vecteurs sont d'une longueur proche, nous avons conservé l'ordre anti-horaire pour les ordonner. Nous avons ordonné les X_i selon leur ordre d'importance donné par la projection des P_i sur les $\overrightarrow{OQ_j}$ les plus longs (Y16, Y15 et Y01). Nous avons priorisé les projections positives et ensuite les projections négatives pour dégager facilement les structures des corrélations.

Nous pouvons ainsi constater au tableau 2.8, que des structures se dégagent naturellement de la matrice des corrélations $\mathbf{R}_{XY}$. Nous pouvons voir immédiatement les sous-groupes de variables les plus et les moins corrélées (voir blocs identifiés par les lettres encadrées dans le tableau dans le tableau 2.8). Cette simple réorganisation nous **résume rapidement les relations** entre les $\mathbf{X}$ et les $\mathbf{Y}$ ce qui aurait pris énormément de temps sans la présence du graphique biplot.

Tableau 2.8 Matrice restructurée des corrélations $\mathbf{R}_{XY}$

Rxy	Y16	Y01	Y15	Y09	Y10	Y06	Y12	Y11
	Cotis oblig	Salaire	Cotisation	Imp fed	Imp prov	Transfert	Cred prov	Cred tps
X17_Reg_ret	0,46	0,45	0,58	0,29	0,30	-0,15	-0,22	-0,24
X18_Public	0,27	0,26	0,47	0,16	0,18	-0,10	-0,16	-0,19
X15_Nb_sem	0,41	0,29	0,18	0,21	0,20	-0,33	-0,14	-0,15
X11_Hres_trav	0,37	0,26	**A1** 0,08	0,22	0,21	**B1** -0,24	-0,04	-0,08
X04_Nb_etude	0,29	0,27	0,27	0,26	0,27	-0,20	-0,17	-0,16
X05_Nb_etude_ps	0,27	0,25	0,27	0,26	0,28	-0,16	-0,16	-0,17
X06_Nb_etude_uni	0,23	0,26	0,28	0,31	0,33	-0,13	-0,14	-0,17
X16_Experience	0,15	0,18	0,15	0,20	0,21	0,01	-0,06	-0,14
X01_Age	0,01	0,10	**A2** 0,13	0,15	0,17	0,09	-0,05	**B2** -0,14
X02_Couple	0,10	0,09	0,04	0,12	0,10	-0,08	-0,07	-0,32
X03_Nb_menage	0,03	0,02	0,01	0,01	-0,04	0,06	0,04	0,00
X14_T_part	-0,34	-0,24	-0,10	-0,14	-0,13	0,14	0,04	0,04
X10_Arret	-0,37	-0,28	-0,19	-0,22	-0,22	0,28	0,16	0,19
X13_Hor_irr	-0,23	-0,26	**C1** -0,14	-0,11	-0,10	0,03	**D** 0.08	0,07
X08_Loue	-0,14	-0,14	-0,12	-0,16	-0,16	0,06	0,12	0,22
X09_Nb_emp	-0,08	-0,09	-0,06	-0,10	-0,10	0,02	0,06	0,10
X07_Rurale	-0,08	-0,09	**C2** -0,07	-0,09	-0,10	0,09	0,09	0,06
X12_Hres_tmais	-0,03	-0,07	-0,01	0,06	0,07	-0,04	-0,02	-0,04
	Y04	Y07	Y13	Y14	Y08	Y05	Y02	Y03
	Agees	Enfant	C enfant	Frais med	Placement	Prestation	Trav auto	Autre rev
X17_Reg_ret	-0,05	-0,04	0,07	0,04	-0,05	-0,08	**J** -0,13	0,00
X18_Public	-0,03	-0,03	-0,01	0,05	-0,03	-0,06	-0,08	0,01
X15_Nb_sem	-0,09	-0,06	0,06	0,03	0,02	-0,26	0,03	-0,02
X11_Hres_trav	-0,09	-0,09	0,07	-0,03	0,03	**I** -0,16	0,08	-0,03
X04_Nb_etude	-0,02	-0,01	0,06	0,05	-0,01	-0,14	0,09	0,02
X05_Nb_etude_ps	0,00	-0,03	0,05	0,05	0,00	-0,12	**K** 0.10	0,01
X06_Nb_etude_uni	0,06	-0,06	0,04	0,04	0,01	-0,09	0,14	0,03
X16_Experience	**E** 0.18	-0,10	-0,03	0,08	**H** 0.12	-0,05	0,01	0,00
X01_Age	0,22	**F** -0,07	-0,05	0,11	0,12	-0,03	0,03	0,03
X02_Couple	0,04	-0,13	-0,02	0,08	0,05	-0,02	0,07	-0,01
X03_Nb_menage	-0,05	0,19	**G** 0.18	0,07	-0,01	0,02	0,04	0,00
X14_T_part	0,09	0,08	-0,06	0,03	0,04	0,01	0,05	0,02
X10_Arret	0,06	0,04	-0,06	-0,03	-0,04	**L** 0.27	-0,05	0,04
X13_Hor_irr	0,07	0,00	-0,01	0,00	0,10	-0,03	0,21	0,02
X08_Loue	-0,04	0,08	-0,03	-0,09	-0,07	0,04	-0,06	0,00
X09_Nb_emp	-0,06	-0,01	-0,03	-0,03	-0,05	0,10	-0,01	-0,01
X07_Rurale	-0,02	0,01	0,01	-0,01	0,01	0,08	-0,01	-0,03
X12_Hres_tmais	0,06	0,00	0,02	0,02	0,10	-0,04	**M** 0.21	0,01

2.4.2 Comportement des variables d'un même groupe

L'information donnée par le biplot indique que la proximité des variables d'un même groupe signifiera qu'elles se comportent de la même façon par rapport aux variables de l'autre groupe : les corrélations de ces variables avec les variables de l'autre groupe devraient être assez semblables. Nous remarquons que la proximité des variables suivantes semble faire ressortir une certaine logique dans leur comportement par rapport à l'autre groupe :

Les sous-groupes suivant des **X** se comporteront de la même façon par rapport aux Y_j :

X04, X05 et X06 : Tous des variables reliées au nombre d'années d'étude.

X14, X13 et X10 : Le travail à temps partiel, les horaires irréguliers et l'arrêt de travail se retrouvent assez proches ce qui semble logique.

X11 et X15 : Les heures et le nombre de semaines travaillées semblent également se comporter façon similaire par rapport aux Y_j ce qui est très cohérent.

Les sous-groupes suivants des **Y** se comporteront de la même façon par rapport aux X_i :

Y09, Y10 : L'impôt fédéral et provincial se retrouvent très proches l'une de l'autre. Il n'est pas étonnant de remarquer que ces deux variables se comportent de la même façon par rapport aux X_i.

Y01 et Y16 : Le salaire et les cotisations obligatoires sont assez proches l'une de l'autre.

Y11 et Y12 : Les crédits fédéraux (TPS) et provinciaux semblent aussi très similaires. Nous retrouvons un peu la même logique que l'impôt fédéral et provincial mais avec un comportement inverse (cadrant négatif) par rapport aux X_i.

En identifiant tout de suite ces comportements à l'intérieur d'un même groupe de variables, il sera plus facile de résumer les relations entre les deux groupes.

2.4.3 Description des corrélations

Nous allons décrire les comportements des $r_{x_i y_j}$ à l'aide du graphique biplot et nous rappellerons que les constatations faites avec le graphique sont très bien mises en évidence dans la matrice restructurée (tableau 2.8).

En observant le graphique, il est possible de connaître rapidement les corrélations les plus fortes et les plus faibles entre les groupes de variables. Pour approximer $r_{x_i y_j}$, il faut projeter orthogonalement le point P_i (représentation des $\mathbf{X}$) sur le vecteur $\overrightarrow{OQ_j}$ (représentation des $\mathbf{Y}$) et multiplier la longueur de la projection par la longueur de $\overrightarrow{OQ_j}$. Ainsi, plus le vecteur $\overrightarrow{OQ_j}$ et la projection de P_i sur $\overrightarrow{OQ_j}$ sont grands, plus la corrélation $r_{x_i y_j}$ sera importante. Remarquons que les points P_i et Q_i près de l'origine relèveront des corrélations pratiquement nulles.

Corrélations fortes (positives)

Un grand vecteur $\overrightarrow{OQ_j}$ et une longue projection de P_i sur $\overrightarrow{OQ_j}$ dans le **même** sens indique une forte corrélation $r_{x_i y_j}$.

Une classe à part : Y15, Y01 et Y16. Tout de suite, nous pouvons constater que Y15 (cotisations), Y01 (salaire) et Y16 (cotisations obligatoires) se démarquent par la longueur des vecteurs qui les représentent. Les plus grandes projections des $\mathbf{X}$ sur ces vecteurs nous donnerons les plus grandes corrélations entre les deux groupes de variables. En effet,

- La projection de X17 (participation à un régime de retraite) nous donne la plus forte corrélation suivie de X18 (employeur public).
- Les projections des variables reliées aux études (X04, X05 et X06) et du temps travaillé (X11 et X15) viennent par la suite comme des corrélations importantes avec ces trois variables dans l'ensemble de donnée.

Groupe impôt : Y09 et Y10. Les corrélations de certains des X_i avec ces deux variables ressortent également comme étant importantes par la longueur de leur

vecteur. Par la projection des points P_i représentant les X_i, les corrélations sont importantes avec

- X17 (participation à un régime de retraite) suivi de X18 (employeur public).
- Les variables reliées aux études : X04, X05 et X06.
- Le temps travaillé : X11 et X15.

► Ces constatations correspondent au bloc supérieur gauche **A1** de la matrice du tableau 2.8.

Lorsque nous observons le graphique et que nous traçons une droite verticale à 0.2, toutes ces corrélations positives importantes (bloc A1) que nous venons de décrire se retrouvent à la droite de cette ligne verticale. Ce groupe de variables est donc très corrélé dans l'ensemble de données et nous avons trouvé cette structure à partir du graphique (très visuel).

Pour les variables à gauche de cette ligne verticale, les vecteurs qui représentent Y04, Y06, Y11 et Y12 semblent avoir une longueur intéressante. Par les projections des points P_i, nous détectons des corrélations assez importantes entre Y04 (revenu de gens plus âgés) et X01 (âge), Y06 (transferts gouvernementaux) et X10 (arrêt de travail) ainsi qu'entre Y11 (crédit TPS) et X08 (logement loué). Par la nature de la description de ces variables, ces corrélations sont très cohérentes. Ces corrélations se présentent dans les blocs **D** et **E** de la matrice $\mathbf{R}_{XY}$ restructurée.

Corrélations fortes (négatives)

Une forte corrélation négative est déterminée par une grande projection de P_i sur le prolongement de $\overrightarrow{OQ_j}$ (sens inverse) avec des vecteurs $\overrightarrow{OQ_j}$ d'une longueur assez grande. Nous devons donc reprendre les vecteurs les plus importants que nous avons détectés précédemment pour les **Y** et observer les plus grandes projections des vecteurs des **X** dans le sens opposé.

Salaire et cotisations obligatoires : Y01 et Y16. Les grands vecteurs que nous avons

identifiés (classe à part) qui les représentent auront des projections opposées intéressantes avec les variables suivantes :

- X10 (arrêt de travail)
- Le travail irrégulier : X13 (Horaire irrégulier) et X14 (temps partiel)

Les corrélations avec ces variables sont donc importantes et négatives.

Groupe impôt : Y09 et Y10. Une corrélation négative importante et logique ressort avec ces deux variables par la projection dans le sens opposé de X10 (arrêt de travail).

► Ces fortes corrélations négatives sont presque tous incluses dans bloc gauche central **C1** de la matrice du tableau 2.8.

Transfert et crédits : Y06, Y11 et Y12. Les trois vecteurs qui représentent ces variables ont une longueur assez importante. Nous avons alors des corrélations négatives importantes entre ces trois variables et

- X17 (participation à un régime de retraite) ainsi que X18 (employeur public).
- Le travail irrégulier (X11 et X15).
- Le nombre d'années d'études (X04, X05 et X06).
- Nous remarquons également que par la projection de X01 et X16 sur le prolongement de Y11, le crédit TPS (Y11) est assez corrélé négativement avec l'âge (X01) et l'expérience (X16).

► Ces corrélations négatives correspondent au bloc supérieur droit **B** de la matrice du tableau 2.8.

Corrélations faibles et nulles

Les projections des P_i sur des vecteurs $\overrightarrow{OQ_j}$ qui sont près de l'origine nous donnent des corrélations $r_{x_i y_j}$ très faibles. Alors d'un premier coup d'oeil, nous pouvons connaître les variables les moins corrélées aux autres dans l'ensembles de données. Nous remarquons tout de suite que les variables X_i et Y_i dans le cercle intérieur (de rayon 0.2) seront les

moins corrélées entre elles par les courtes longueurs des vecteurs qui les représentent. Nous pouvons constater que

- Y03 (Autre revenu) est très près de l'origine donc les corrélations des X_i avec Y03 seront pratiquement nulles.
- Les corrélations des Y_i avec X12 (heures travaillées à la maison), X07 (région rurale) et X09 (nombre d'emplois) seront tous très faibles car ces X_i sont près de l'origine.
- Y02 (travaille autonome), Y08 (placement) et Y14 (frais médicaux) possèderont également de faibles corrélations avec les X_i.

► Les faibles corrélations sont identifiés dans les sections grisées de la matrice restructurée (tableau 2.8). Nous pouvons remarquer qu'il y a beaucoup de faibles corrélations entre les X_i et Y_i. Celles que nous venons de mettre en évidence pas le graphique biplot correspondent surtout à la section inférieur de la matrice du tableau 2.8.

En résumé, plus les vecteurs sont grands et proches les uns des autres (même sens ou sens inverse), plus les corrélations seront importantes. Il est donc facile de « voir » les corrélations les plus fortes et les plus faibles dans le graphique. Le graphique nous permet de comprendre plus rapidement le comportement et les relations des variables au lieu d'étudier les 288 entrées de la matrice $\mathbf{R}_{XY}$. Grâce au biplot, nous pouvons restructurer cette matrice afin de la rendre beaucoup plus significative pour un lecteur et d'étudier plus facilement les relations entre les X_i et Y_i.

En comparant la matrice $\mathbf{R}_{XY}$ avec les constatations du biplot, nous remarquons que nous pouvons quand même échapper certaines corrélations importantes. Ceci vient de l'erreur d'estimation car il ne faut pas oublier que ce biplot **estime** la matrice $\mathbf{R}_{XY}$. Alors, en réorganisant $\mathbf{R}_{XY}$ (à l'aide du biplot), **aucune** corrélation importante ne nous échappe et la compréhension des relations est tout aussi simple.

2.4.4 Ordre de grandeur et direction opposée

Ce qui est également intéressant dans le graphique est que nous sommes en mesure d'obtenir **rapidement** un ordre de grandeur des corrélations des $\mathbf{X}$ avec un Y_j donné. En débutant par l'extrémité d'un Y_j, nous n'avons qu'à observer les projections des $\mathbf{X}$ sur ce Y_j et nous obtenons ainsi les corrélations entre ce Y_j et les $\mathbf{X}$ en ordre décroissant.

Par exemple pour Y01 (salaire), la plus forte corrélation de cette variable avec les $\mathbf{X}$ est avec X17 (participation à un régime de retraite). Suivent dans l'ordre décroissant X18 (employeur public), le temps travaillé (X15 et X11), le groupe des études (X04, X05 et X06) et X16 (expérience). Par la suite, nous avons des corrélations très faibles mais positives avec X02 (couple), X03 (nombre de personnes dans le ménage) et X01 (l'âge). Ensuite, il y a des corrélations négatives faibles avec X09 (nombre d'emplois), X07 (région rurale) et X08 (logement loué). Pour terminer, Y01 possède des corrélations négatives plus fortes avec X13 (horaire irrégulier), X14 (temps partiel) et X10 (arrêt de travail).

Autre fait intéressant à remarquer : deux vecteurs représentant les $\mathbf{Y}$ qui sont alignés sur une droite dans des directions opposées auront des comportements inverses avec les $\mathbf{X}$. Si les vecteurs des $\mathbf{Y}$ sont de même distance, les corrélations avec les $\mathbf{X}$ seront similaires mais de signes opposés (+ /-). S'ils ne sont pas de même distance, leurs comportements avec les $\mathbf{X}$ seront quand même opposés mais dans un ordre de grandeur différent. Par exemple, Y16 (cotisations obligatoires) et Y06 (transferts gouvernementaux) s'alignent presque sur une droite dans des directions différentes. Leurs corrélations avec les $\mathbf{X}$ sont donc complètement à l'opposé mais dans un ordre de grandeur quelque peu différent (plus petit pour Y06 car vecteur moins grand).

2.5 Interprétation des relations canoniques avec le biplot

Le graphique biplot qui estime $\mathbf{R}_{XY}$ et qui permet de réorganiser cette matrice est obtenu grâce aux calculs des corrélations canoniques (λ_i) de la section 2.3 sur nos

deux groupes de variables.

Nous allons voir quel genre d'information ces relations canoniques peuvent nous donner, bien qu'en principe, les relations entre les $\mathbf{X}$ et les $\mathbf{Y}$ se trouvent dans $\mathbf{R}_{XY}$. Nous allons mettre en lien les relations canoniques avec le biplot et la matrice $\mathbf{R}_{XY}$ restructurée (tableau 2.8). De cette façon, nous pourrons tenter de donner un sens concret à ces relations canoniques.

2.5.1 Poids canoniques vs intra-corrélations

Auparavant, certains auteurs mentionnaient qu'il fallait utiliser les poids des vecteurs canoniques (tableaux 2.12 et 2.13) pour interpréter les relations canoniques. Remarquons tout de suite que la stabilité des poids canoniques peut être influencée par la multicolinéarité dans les groupes de variables à l'étude.

Par exemple, nous savons par l'étude du comportement des variables dans le biplot (section 2.4.2) que dans le groupe des $\mathbf{X}$, les variables X04, X05 et X06, relatives aux nombre d'années d'études, ont des corrélations similaires avec les $\mathbf{Y}$. Dans le groupe $\mathbf{Y}$, le salaire (Y01) et les cotisations obligatoires (Y16) ont un comportement similaire avec les $\mathbf{X}$ tout comme les deux variables des impôts (Y09, Y10) et des crédits (Y11, Y12). Dus à ces comportements, les poids canoniques relatifs à ces variables peuvent être "faussées" comme en régression multiple. Nous opterons plutôt pour les coefficients de l'intra-corrélation afin d'interpréter les relations canoniques.

Nous allons donner un exemple avec la première relation canonique où ces comportements viennent influencer l'interprétation des résultats (tableau 2.9) .

Exemple : V_1

Par les poids canoniques de V_1, nous constatons que Y15 (cotisations) et Y16 (cotisations obligatoires) sont les variables de $\mathbf{Y}$ les plus importantes dans la relation 1 (poids = 0.50978 et 0.51395). Lorsque nous observons les coefficients de l'intra-corrélation de

Tableau 2.9 Poids canoniques vs intra-corrélations de V_1

	Poids canoniques	Intra-corrélations
Y01 Salaire	0.15878	0.80690
Y09 Imp fed	-0.19985	0.60276
Y10 Imp prov	0.13791	0.61307
Y11 Cred tps	-0.06703	-0.46348
Y12 Cred prov	0.01545	-0.37869
Y15 Cotisation	0.50978	0.84527
Y16 Cotis oblig	0.51395	0.86725

V_1, nous remarquons que V_1 est très corrélée avec ces deux variables (Y15=0.84527 et Y16=0.86725) ainsi qu'avec le salaire (Y01=0.80690) qui pourtant, ne semblait pas être très important par son poids canonique (0.115878, tableau 2.9). Le fait que le salaire (Y01) est peut-être corrélé aux cotisations obligatoires (Y16) entraîne donc un problème de multicolinéarité ce qui cause que son poids canonique est faible (0.15878). L'interprétation de cette relation canonique peut donc être erronée si nous n'observons pas attentivement les coefficients l'intra-corrélation. Nous avons également mis en évidence dans le tableau 2.9 qu'avec les intra-corrélations, les impôts (Y09 et Y10) semblent être des variables importantes dans V_1 (0.60276 et 0.61307) mais leur poids se comporte différemment (-0.19985 et 0.13791). Le même phénomène se produit avec les crédits (Y11 et Y12) mais dans l'ordre inverse (intra-corrélation négative).

Ainsi, pour tenter d'interpréter les relations canoniques, nous préférerons travailler avec les coefficients de l'intra-corrélation dans un premier temps plutôt que de donner des interprétations plus ou moins fiables aux poids canoniques. Par la suite, il serait possible de tenter de corriger la muticolinéarité en supprimant des variables dans l'ensemble de départ et peut-être que les poids canoniques seront alors plus stables et plus facile d'interprétation.

2.5.2 Graphique biplot pour l'interprétation des relations canoniques

En travaillant avec les intra-corrélations pour interpréter nos relations canoniques, nous pouvons alors utiliser notre graphique biplot (figure 2.2) afin de constater les variables les plus corrélées avec V_1 et V_2, car rappelons que les Q_j (Y_j) du graphique correspondent à $(r_{Y_jV_1}, r_{Y_jV_2})$. Il est alors très facile de constater visuellement les Y_j forts de V_1 et V_2. Ainsi, pour

V_1 :

- Relations fortes positives avec Y16, Y01, Y15, Y10 et Y09 : Salaire, impôts et cotisations.
- Relations fortes négatives avec Y11, Y12 et Y06 : Crédits et transferts gouvernementaux.

V_2 :

- Relations fortes positives avec Y04 et Y10 : Revenu relié aux personnes plus âgées (pension, REER, etc.) et impôt provincial.
- Relations fortes négatives avec Y11, Y07 : Crédit TPS et revenu relié aux enfants.

Graphique biplot pour (U_1, U_2)

Le graphique biplot que nous avons tracé pour estimer $\mathbf{R}_{XY}$ est construit à partir des coefficients de l'intra/inter-corrélation de V_1 et V_2 seulement. Nous n'avons donc pas encore les intra-corrélations de U_1 et U_2 qui nous permettront d'interpréter les variables fortes des X_j dans la relation canonique 1 et 2.

Comme nous l'avons mentionné à la fin de la section 1.5.4 du chapitre précédent, $\mathbf{\Lambda}_{[r]}$ peut être intégré à $\underset{q\times r}{\mathbf{D}}$ (1.29) au lieu qu'à $\underset{p\times r}{\mathbf{C}}$ (1.28). Dans ce cas,

$$\underset{p\times r}{\mathbf{C}} = \underset{p\ \times\ r}{(Corr(\mathbf{X}, \mathbf{U}))}$$

et

$$\underset{q\times r}{\mathbf{D}} = \underset{q\ \times\ r}{(Corr(\mathbf{Y}, \mathbf{U}))}\,.$$

Les coefficients de l'intra-corrélation de U_1 et U_2 sont donc contenus dans la matrice $\underset{p\times 2}{\mathbf{C}}$. Cette décomposition nous donnera également la même estimation de $\mathbf{R}_{XY}$ mais la représentation par le biplot sera basée sur les axes canoniques U_1 et U_2 au lieu de V_1 et V_2. Il sera donc possible de détecter de fortes corrélations $r_{x_i y_j}$ qui n'étaient pas nécessairement évidente à voir dans l'autre biplot (figure 2.2). Nous ne ferons pas tous les détails théoriques, ils sont disponibles dans l'article de ter Braak (il estime $\mathbf{R}_{YX}$ au lieu de $\mathbf{R}_{XY}$ ce qui revient à faire la même chose). Alors

$$\begin{aligned}
\underset{p\ \times\ q}{\mathbf{R}_{XY}} &= \underset{p\ \times\ s}{(\mathbf{R}_{XX}^{1/2}\mathbf{P})}\,\underset{s\ \times\ q}{(\mathbf{\Lambda}\mathbf{Q}'\mathbf{R}_{YY}^{1/2})} \\
&= \underset{p\ \times\ s}{(Corr(\mathbf{X},\mathbf{U}))}\,\underset{s\ \times\ q}{(Corr(\mathbf{U},\mathbf{Y}))} \\
&\simeq \underset{[\,2\,]}{\mathbf{R}_{XY}} = \underset{p\times 2}{\mathbf{C}}\ \underset{2\times q}{\mathbf{D}}{}'
\end{aligned}$$

où

$$\underset{p\times 2}{\mathbf{C}} = \begin{pmatrix} r_{X_1U_1} & r_{X_1U_2} \\ \vdots & \vdots \\ r_{X_pU_1} & r_{X_pU_2} \end{pmatrix}$$

et

$$\underset{q\times 2}{\mathbf{D}} = \begin{pmatrix} r_{Y_1U_1} & r_{Y_1U_2} \\ \vdots & \vdots \\ r_{Y_qU_1} & r_{Y_qU_2} \end{pmatrix}$$

Nous avons ainsi les coordonnées des points $P_i = (r_{X_iU_1} \quad r_{X_iU_2})$, $i = 1,..,p$ et les coordonnées des points $Q_j = (r_{Y_jU_1} \quad r_{Y_jU_2})$, $j = 1,..,q$. Les points P_1 à P_{18} représentent les corrélations entre U_1 et U_2 avec les 18 variables du premier groupe $\mathbf{X}$. C'est en observant ces points que nous pourrons interpréter facilement les variables des $\mathbf{X}$ les plus importantes dans la relation canoniques 1 et 2.

Tout comme le graphique précédent, il est très facile de d'obtenir ce graphique car nous utilisons le même fichier (*out stat cancor*) pour les P_i et les Q_j (intra/inter-corrélations de U_1 et U_2). À la figure 2.3, nous pouvons constater visuellement les relations fortes des $\mathbf{X}$ avec U_1 et U_2 :

Pour U_1,

- Relations fortes positives avec X17, X18, X15 et X04 : Participation à un régime de retraite, employeur public, nombre de semaines travaillées et nombre d'années d'études.
- Relations fortes négatives avec X10, X14 et X13 : Période d'arrêt de travail, horaire irrégulier et travail à temps partiel.

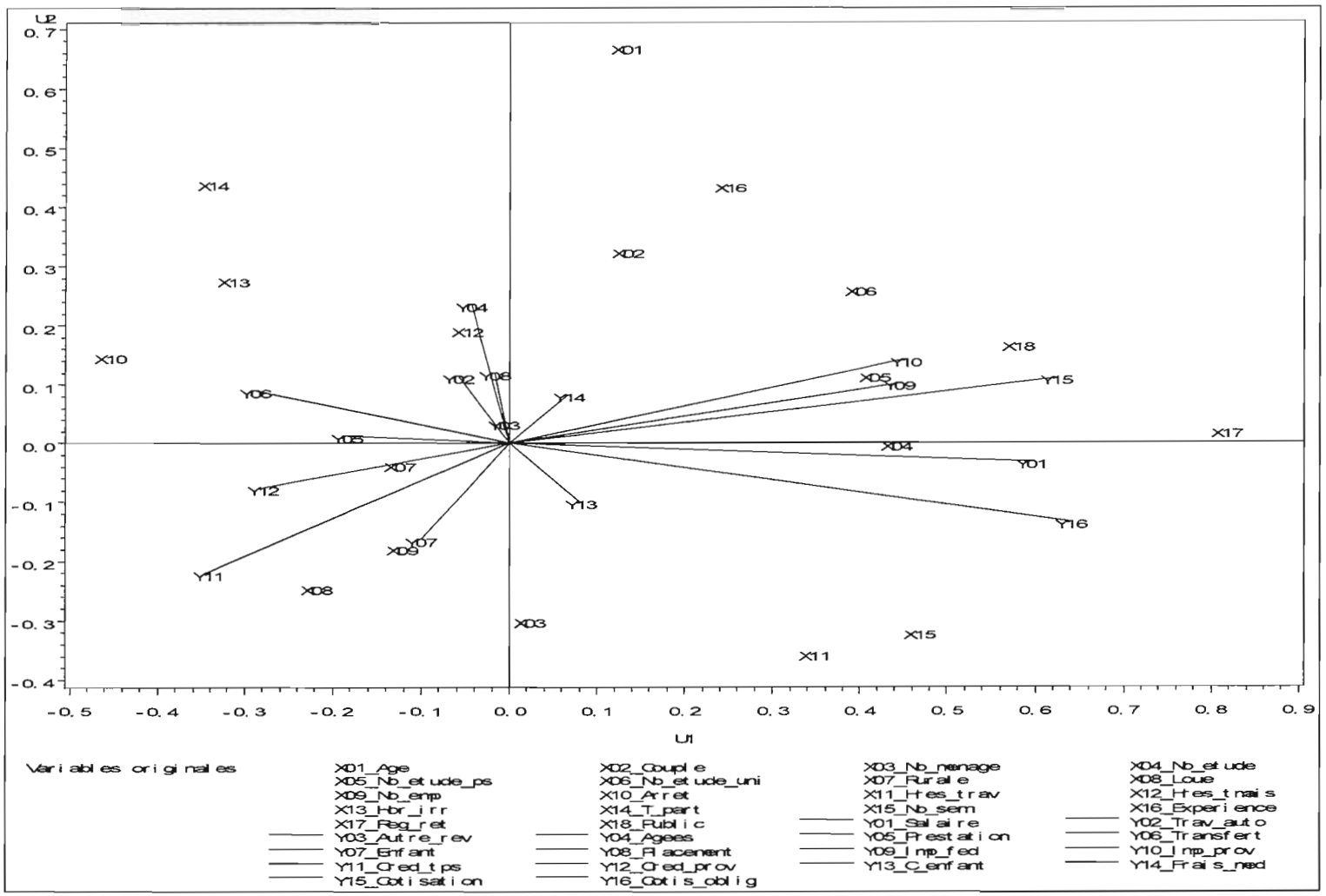
Pour U_2,

- Relations fortes positives avec X01, X14, X16 : L'âge, travail à temps partiel et le nombre d'années d'expérience.
- Relations fortes négatives avec X11, X15, X03 : Les heures et le nombre de semaines travaillées et le nombre de personnes dans le ménage.

Remarque

Si nous observons attentivement les deux graphiques biplot des figures 2.2 et 2.3, nous remarquons que la disposition des vecteurs qui représentent les $\mathbf{X}$ et $\mathbf{Y}$ est **très semblable** dans chacun des graphiques. Ainsi dans le premier graphique 2.2, bien que nous n'avons pas les vraies valeurs des corrélations avec U_1 et U_2, nous remarquons que les X_i mis en évidence précédemment par rapport à U_1 et U_2 sont les mêmes qui se démarquent dans les axes V_1 et V_2. Donc, seul le graphique par rapport à V_1 et V_2 (2.2) est suffisant pour décrire et comprendre les deux premières relations canoniques.

Figure 2.3 Graphique biplot, estimation de $\mathbf{R}_{XY}$ par les intra/inter-corrélations de U_1 et U_2



2.5.3 Interprétation des deux premières relations canoniques

Nous allons maintenant tenter d'interpréter dans le contexte les deux premières relations canoniques à l'aide des constatations du biplot et de la matrice $\mathbf{R}_{XY}$ restructurée (tableau 2.8). Nous aurons ainsi une interprétation concrète des résultats de l'analyse des corrélations canoniques par l'utilisation du biplot.

Relation canonique 1 ($\lambda_1 = 0.738364$)

Cette forte corrélation canonique $\lambda_1 = 0.738364$ nous indique qu'il existe une assez forte relation entre les variables canoniques U_1 et V_1 que nous verrons dans la matrice restructurée. L'étude du biplot nous a amené à dégager quelques constats sur les variables originales $\mathbf{X}$ et $\mathbf{Y}$ les plus importantes (corrélations fortes) avec leur variable canonique U_1 et V_1.

Ainsi, V_1 représentait positivement le salaire, les impôts et les cotisations tandis que U_1 représentait positivement les études, le temps travaillé, la participation à un régime de retraite et l'employeur public. Ces deux liens positifs importants se reflètent bien dans le coin supérieur gauche de la matrice de corrélations restructurée (bloc **A1**). Il y a donc une relation positive importante entre le nombre d'années d'études et le salaire mais également avec les impôts et les cotisations payés. Cela montre la priorité d'investir en éducation pour le Gouvernement afin d'obtenir plus de revenus pour l'État (impôts et cotisations). Cette relation positive peut être également est très motivante pour les jeunes.

Ce bloc **A1** dans la matrice montre que le temps travaillé est relié positivement aux impôts et cotisations, ce qui n'est pas surprenant. Nous voyons aussi qu'avec un employeur public (bonnes conditions de travail), le salaire, les impôts et les cotisations ont une tendance à la hausse. Il serait intéressant également de vérifier si la corrélation entre le nombre d'années études et l'employeur public ne serait pas, par le fait même, assez élevée. Cela encouragerait probablement plus nos jeunes à poursuivre des études dans le

but d'avoir de bonnes conditions de travail.

U_1 possède une forte relation négative avec un travail atypique (horaire irrégulier, arrêt et temps partiel). La corrélation de ce type de travailleur est décroissante avec les salaires, les impôts et les cotisations (importantes dans V_1). Cela correspond aux trois premières lignes du bloc **C1** de la matrice des corrélations restructurées. Il serait intéressant de mieux comprendre le profil de ce type de travailleur atypique afin de soit les aider dans le but de faire augmenter leur salaire, ou mieux prévoir les revenus de l'État (ils payent moins d'impôts et de cotisations) car c'est peut-être un choix qu'ils ont fait.

La relation des **Y** avec V_1 est aussi forte négativement avec les transferts gouvernementaux et les crédits d'impôt. Il n'est donc pas surprenant que ces variables soient corrélées négativement avec le temps travaillé (X11 et X15 positives pour U_1). Ces liens correspondent à une partie du bloc **B1** de la matrice restructurée. Ces transferts et crédits sont plus importants chez les travailleurs atypiques (relation négative avec U_1) ce qui correspond au bloc **D** de la matrice.

En résumé, plus les gens travaillent et plus ils ont des années d'études, plus le salaire, les impôts et cotisations seront élevés et moins ils recevront de crédits et de transferts. Pour les travailleurs ayant un profil particulier (temps partiel, arrêt ou horaire irrégulier), c'est l'inverse qui se passe.

Cette première relation canonique dans notre ensemble de données reflète une réalité évidente et il est très facile de l'étudier dans les blocs A1, B1, C1 et D de la matrice du tableau 2.8. Nous pensons que les jeunes et l'État devraient s'y intéresser davantage.

Relation canonique 2 ($\lambda_2 = 0.537446$)

Cette corrélation canonique λ_2 nous indique qu'il existe une bonne relation entre les variables canoniques U_2 et V_2. Par l'étude du biplot, tout porte à croire que cette relation est reliée au vieillissement de la population.

En effet, V_2 est fortement corrélée positivement avec les revenus relatifs aux personnes plus âgées tandis que U_2 est corrélée positivement avec l'âge et l'expérience. Ces deux constatations sont mises en évidence dans le bloc **E** de la matrice restructurée. Il y a également une corrélation négative pour V_2 avec le crédit TPS et les revenus relatifs aux enfants qui se reflète avec l'âge et l'expérience (positifs pour U_2) dans les blocs **B2** et **F**.

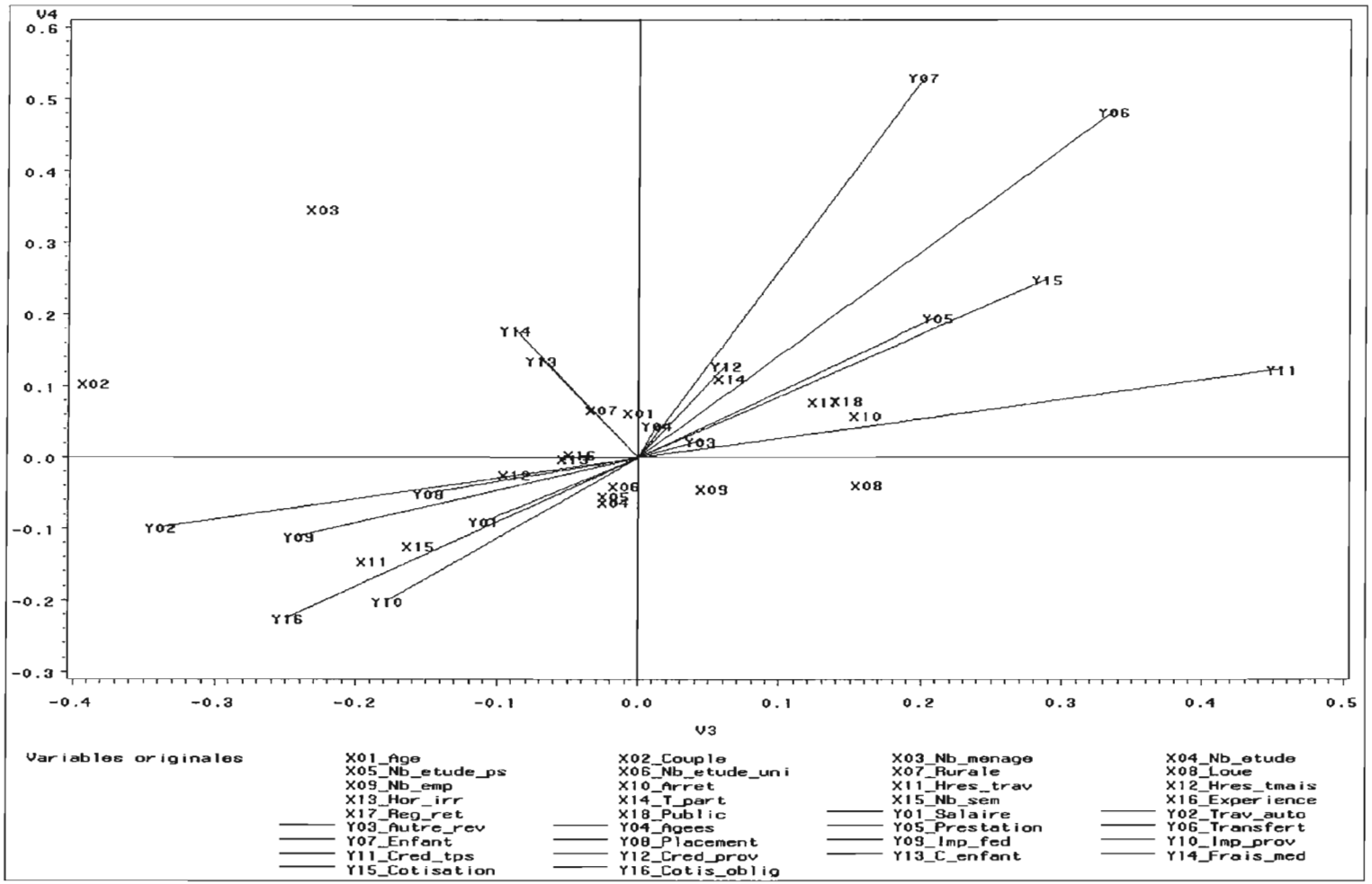
Ce qui est intéressant et qu'il serait possible d'approfondir est que V_2 est corrélée avec les impôts (bloc **A2** à droite pour X16 et X01) et que U_2, est corrélée avec le travail à temps partiel. Nous pouvons nous poser la question si les personnes plus âgées continuent à travailler à temps partiel. Combiné à un revenu de personnes plus âgées (REER ou autres), ces gens continuent à payer de l'impôt ce qui est important pour l'État.

Nous croyons donc que le Gouvernement devrait étudier ce phénomène (le deuxième en importance dans l'ensemble de données) plus en détails car cette portion de travailleurs plus âgés qui payent des impôts n'est sûrement pas négligeable. Avec l'espérance de vie qui augmente, nous pouvons croire que ces individus continueront à travailler de cette façon encore quelques années. Le Gouvernement pourra ainsi prendre en considération cette réalité dans les prévisions de ses revenus et peut-être de cette façon, mieux absorber les coûts futurs du vieillissement de la population.

2.5.4 Graphique biplot pour (V_3, V_4)

Afin d'étudier les relations canoniques 3 et 4, nous avons tracé à la figure 2.4 le biplot en fonction des coefficients de l'intra/inter-corrélation des vecteurs V_3 et V_4. Ce biplot ne sera par utilisé pour estimer la matrice $\mathbf{R}_{XY}$ dans l'ensemble car les estimations seront beaucoup moins précises que celles obtenues par le biplot de V_1 et V_2 (voir Théorème d'Eckart-Young 1.5.2.1). Il est tout de même intéressant de noter qu'il est possible de détecter de fortes corrélations $r_{x_i y_j}$ qui nous avaient échappé dans le biplot précédent. Par exemple, nous pouvons remarquer rapidement que par ce biplot, X02 et Y02 semblent assez corrélés ce qui n'était pas aussi clair dans le premier biplot. Nous avons cette corrélation dans le bloc **M** de la matrice restructurée.

Figure 2.4 Graphique biplot, estimation de $\mathbf{R}_{XY}$ par les intra/inter-corrélations de V_3 et V_4



À partir de ce graphique, nous observons que pour V_3 :

- La corrélation est forte et positive avec le crédit TPS (Y11), les transferts gouvernementaux (Y06), et les cotisations (Y15).
- La corrélation est forte et négative avec le revenu de travail autonome (Y02), les cotisations obligatoires (Y16) et les impôts (Y09 et Y10).

Nous ne ferons pas le graphique en fonction de U_3 et U_4 car comme nous l'avons constaté, il sera similaire à celui de V_3 et V_4. Ainsi, pour les $\mathbf{X}$ reliées à U_3 :

- La corrélation est importante et positive avec l'arrêt de travail (X10) et le fait d'avoir un logement loué (X08).
- La corrélation est importante et négative avec le fait d'être en couple (X02), le nombre de personne dans le ménage (X03) et le temps travaillé (X11 et X15).

De même, nous observons que pour V_4 :

- Il y a une corrélation importante et positive avec les transferts gouvernementaux (Y06) et les revenus relatifs aux enfants (Y07).
- Il y a une corrélation importante et négative avec les cotisations obligatoires (Y16) et l'impôt provincial (Y10).

De même, pour les $\mathbf{X}$ reliées à U_4 :

- Il y a une corrélation positive avec le nombre de personnes dans le ménage (X03) et le fait d'être en couple (X02).
- Il y a une corrélation négative avec le temps travaillé (X11 et X15).

2.5.5 Interprétation des relations canoniques 3 et 4

Comme nous l'avons constaté, les relations canoniques 3 et 4 doivent être interprétées avec vigilance car elles semblaient très instables avec les validations que nous avons effectuées. Nous allons tenter de les résumer brièvement mais nous leur apporterons moins d'intérêt à cause de cette instabilité et des coefficients de corrélation canonique

plus faibles.

Relation 3 ($\lambda_3 = 0.478715$)

Par l'étude du biplot, la relation 3 tend à montrer que les personnes qui semblent avoir eu plus souvent une période d'arrêt de travail et qui habitent plus souvent un logement loué (positives pour U_3) auront des montants plus élevés de transferts gouvernementaux et de crédits (positives pour V_3). Ce fait correspond aux corrélations fortes du bloc **D** de la matrice restructurée. Également, la relation avec V_3 est négative avec les impôts (payent moins d'impôts). Ce sont les corrélations fortes négatives de la partie droite de bloc **C1**.

Relation 4 ($\lambda_4 = 0.439207$)

Par le biplot, les variables U_4 et V_4 de la relation 4 semblent identifier les individus en famille (avec enfant). En effet, V_4 montre un lien positif avec les montants de revenus pour enfants et de transferts gouvernementaux. L'élément familial ressort également (positivement) dans U_4 avec le nombre de personnes dans le ménage et le fait d'être en couple. Le temps travaillé est corrélé négativement avec V_4 ce qui peut sembler normal pour des individus ayant des enfants.

2.6 Modèles de prédiction

L'analyse des corrélations canoniques nous aide à comprendre un ensemble de données et peut-être ainsi, à trouver des modèles de prédiction. C'est le coefficient de l'inter-corrélation qui nous guidera dans cette recherche. Le **carré** de ce coefficient représente la proportion de variance d'une variable originale expliquée par une variable canonique. Ces coefficients sont présentés aux tableaux 2.16 et 2.17.

Inter-corrélations U_1 à U_4

La variable canonique U_1 semble être la meilleure pour expliquer la variance de quelques variables originales du groupe $\mathbf{Y}$. En effet, U_1 explique respectivement 41 % et 39 % des cotisations obligatoires (Y16) et autres cotisations (Y15).

Pour ce qui est de U_2, U_3 et U_4, pas plus 5 % de variance de chacun des Y_i est expliquée par ces variables canoniques.

Inter-corrélations V_1 à V_4

Seule la variable V_1 semble expliquer la variance de quelques X_i mais dans un ordre beaucoup moins important que U_1 est en mesure de le faire avec quelques Y_i. V_1 explique 36 % de la variance de la participation à des régimes de retraite (X18) qui est le maximum que V_1 est en mesure d'expliquer.

L'étude de ces coefficients de l'inter-corrélation nous guide vers des modèles de prédiction. Nous pourrions continuer l'analyse de l'ensemble de données en tentant d'établir des modèles de prédictions de Y16 et Y15 comme variables dépendantes par les X_i comme variables indépendantes. Il semble logique d'établir des variables fiscales (Y_j) dépendantes des variables sociodémographiques ($\mathbf{X}$).

La prédiction des cotisations obligatoires (régime des rentes du Québec et assurance emploi, Y16) est un exemple concret de prédiction intéressante pour le Gouvernement afin de comprendre comment ces cotisations peuvent varier.

2.7 Conclusion

Avec l'analyse des corrélations canoniques et le graphique biplot, nous avons montrer qu'il est simple et facile de mettre en évidence et de comprendre les relations dans un ensemble de données complexes (nombre élevé de variables et d'observations). Avec le graphique biplot, nous pouvons restructurer la matrice de corrélations $\mathbf{R}_{XY}$ qui facilite la compréhension de ces relations. Le graphique biplot, les coefficients de l'intra/inter-

corrélation et la réorganisation de la matrice $\mathbf{R}_{XY}$ nous aident également à interpréter les corrélations canoniques ce qui, par le passé, ne semblaient pas aussi évident.

La connaissance de l'ensemble de données étant acquise, il serait maintenant intéressant pour des chercheurs d'obtenir des statistiques descriptives en fonction des relations trouvées. Par exemple, nous désirons mieux connaître les travailleurs mis en évidences dans la relation canonique 2 (plus âgées, temps partiel, etc.). Nous pourrons alors sortir des statistiques de ce groupe face à la population totale par ville, par groupe d'âges, par tranche de revenus ou autres variables pour en faire ressortir les caractéristiques.

De plus, en comprenant mieux les relations entre toutes les variables, il devient ainsi plus facile et intéressant d'étudier des modèles de prédiction.

En terminant, Statistique Canada nous mentionne qu'il est important de rappeler que cette analyse est fondée sur les microdonnées à grande diffusion de l'Enquête sur la dynamique du travail et du revenu de Statistique Canada, qui contiennent des données anonymes de l'Enquête sur la dynamique du travail et du revenu. Tous les calculs effectués à l'aide de ces microdonnées sont la responsabilité de l'auteur. L'utilisation et l'interprétation de ces résultats sont uniquement la responsabilité de l'auteur.

2.8 Tableaux

Voici les tableaux 2.10 à 2.17 dont nous avons discutés à la section 2.3.3 qui portait sur les résultats obtenus.

Tableau 2.10 Vecteurs canoniques U_1 à U_4 des variables originales non standardisées

NAME	U1	U2	U3	U4
X01_Age	-0.00281	0.03543	0.02061	0.03078
X02_Couple	0.09416	0.76971	-1.48952	-0.34491
X03_Nb_menage	-0.00079	-0.32863	-0.06567	0.75455
X04_Nb_etude	0.03451	-0.08325	-0.00699	-0.03367
X05_Nb_etude_ps	0.00801	0.10540	-0.03033	0.02078
X06_Nb_etude_uni	0.09369	0.12554	-0.02132	-0.03160
X07_Rurale	-0.13128	-0.09989	0.00480	0.29219
X08_Loue	-0.20905	-0.29982	0.21243	0.50268
X09_Nb_emp	-0.17264	0.03572	0.21724	0.04203
X10_Arret	-0.16622	-0.08570	0.30791	-0.18535
X11_Hres_trav	0.00023	-0.00015	-0.00027	-0.00032
X12_Hres_tmais	-0.00886	0.01399	-0.00906	0.00022
X13_Hor_irr	-0.16586	0.46833	-0.05024	0.03706
X14_T_part	-0.37226	0.69233	0.03626	0.32840
X15_Nb_sem	0.01218	-0.02648	-0.01324	-0.01753
X16_Experience	0.01480	0.00662	-0.00919	-0.00564
X17_Reg_ret	1.14610	0.06966	0.52934	0.46358
X18_Public	0.51114	0.27791	0.60611	0.31739

Tableau 2.11 Vecteurs canoniques V_1 à V_4 des variables originales non standardisées

NAME	V1	V2	V3	V4
Y01_Salaire	0.00000675	-.00000075	-.00003173	0.00002066
Y02_Trav_auto	-.00000251	0.00001836	-.00005016	0.00001767
Y03_Autre_rev	0.00001117	-.00000262	-.00001158	0.00001526
Y04_Agees	-.00000103	0.00005910	-.00002378	0.00002915
Y05_Prestation	-.00001653	-.00017199	-.00002486	-.00009836
Y06_Transfert	0.00000522	0.00018785	0.00006642	0.00016259
Y07_Enfant	0.00000649	-.00017050	-.00006345	0.00019944
Y08_Placement	0.00000625	0.00001630	-.00003067	0.00000185
Y09_Imp_fed	-.00005257	-.00045409	-.00057812	0.00083454
Y10_Imp_prov	0.00003152	0.00043637	0.00063608	-.00085430
Y11_Cred_tps	-.00035909	-.00348227	0.00593681	-.00260144
Y12_Cred_prov	0.00008655	0.00147769	-.00406473	0.00185131
Y13_C_enfant	-.00001046	-.00010925	-.00003329	0.00008631
Y14_Frais_med	0.00005351	0.00014608	-.00012664	0.00022515
Y15_Cotisation	0.00041191	0.00025703	0.00044820	0.00039542
Y16_Cotis_oblig	0.00059653	-.00060233	0.00001234	-.00043772

Tableau 2.12 Vecteurs canoniques U_1 à U_4 des variables originales standardisées

NAME	U1	U2	U3	U4
X01_Age	-0.03255	0.41064	0.23890	0.35677
X02_Couple	0.04617	0.37741	-0.73035	-0.16912
X03_Nb_menage	-0.00101	-0.42382	-0.08469	0.97311
X04_Nb_etude	0.12086	-0.29158	-0.02449	-0.11793
X05_Nb_etude_ps	0.01918	0.25255	-0.07267	0.04980
X06_Nb_etude_uni	0.16378	0.21947	-0.03727	-0.05524
X07_Rurale	-0.05243	-0.03990	0.00192	0.11670
X08_Loue	-0.09674	-0.13875	0.09831	0.23262
X09_Nb_emp	-0.08448	0.01748	0.10631	0.02057
X10_Arret	-0.06161	-0.03176	0.11413	-0.06870
X11_Hres_trav	0.15427	-0.09658	-0.17649	-0.21010
X12_Hres_tmais	-0.06108	0.09642	-0.06249	0.00153
X13_Hor_irr	-0.07956	0.22463	-0.02410	0.01778
X14_T_part	-0.12884	0.23961	0.01255	0.11366
X15_Nb_sem	0.12730	-0.27674	-0.13840	-0.18317
X16_Experience	0.16074	0.07188	-0.09983	-0.06127
X17_Reg_ret	0.56152	0.03413	0.25935	0.22713
X18_Public	0.20900	0.11364	0.24784	0.12978

Tableau 2.13 Vecteurs canoniques V_1 à V_4 des variables originales standardisées

NAME	V1	V2	V3	V4
Y01_Salaire	0.15878	-0.01773	-0.74645	0.48605
Y02_Trav_auto	-0.02860	0.20905	-0.57123	0.20120
Y03_Autre_rev	0.02148	-0.00503	-0.02227	0.02935
Y04_Agees	-0.00373	0.21521	-0.08660	0.10614
Y05_Prestation	-0.04392	-0.45713	-0.06593	-0.26141
Y06_Transfert	0.01817	0.65379	0.23117	0.56586
Y07_Enfant	0.01217	-0.31967	-0.11897	0.37393
Y08_Placement	0.03025	0.07887	-0.14843	0.00895
Y09_Imp_fed	-0.19985	-1.72641	-2.19793	3.17284
Y10_Imp_prov	0.13791	1.90909	2.78282	-3.73756
Y11_Cred_tps	-0.06703	-0.65001	1.10819	-0.48560
Y12_Cred_prov	0.01545	0.26373	-0.72545	0.33041
Y13_C_enfant	-0.01499	-0.15650	-0.04769	0.12365
Y14_Frais_med	0.04198	0.11458	-0.09935	0.17663
Y15_Cotisation	0.50978	0.31810	0.55469	0.48936
Y16_Cotis_oblig	0.51395	-0.51894	0.01063	-0.37712

Tableau 2.14 Intra-corrélations U_1 à U_4

NAME	U1	U2	U3	U4
X01_Age	0.13388	0.67087	0.00329	0.14706
X02_Couple	0.13468	0.32679	-0.80283	0.24054
X03_Nb_menage	0.02322	-0.29740	-0.46778	0.79295
X04_Nb_etude	0.44129	0.00011	-0.03517	-0.13890
X05_Nb_etude_ps	0.41610	0.11704	-0.03424	-0.11913
X06_Nb_etude_uni	0.39973	0.26098	-0.01839	-0.08630
X07_Rurale	-0.12510	-0.03407	-0.05316	0.15816
X08_Loue	-0.21800	-0.24229	0.33908	-0.08439
X09_Nb_emp	-0.12219	-0.17584	0.11202	-0.09492
X10_Arret	-0.45376	0.14980	0.33665	0.13711
X11_Hres_trav	0.34814	-0.35419	-0.39221	-0.32671
X12_Hres_tmais	-0.04908	0.19329	-0.18093	-0.05063
X13_Hor_irr	-0.31342	0.27921	-0.09575	0.00052
X14_T_part	-0.33780	0.44221	0.13701	0.25685
X15_Nb_sem	0.46900	-0.31939	-0.32288	-0.27969
X16_Experience	0.25106	0.43692	-0.08506	0.01431
X17_Reg_ret	0.81736	0.02027	0.27527	0.18410
X18_Public	0.58020	0.16878	0.30877	0.18479

Tableau 2.15 Intra-corrélations V_1 à V_4

NAME	V1	V2	V3	V4
Y01_Salaire	0.80690	-0.05989	-0.10849	-0.08901
Y02_Trav_auto	-0.07925	0.21342	-0.33740	-0.09730
Y03_Autre_rev	-0.00956	0.06499	0.04432	0.02416
Y04_Agees	-0.05890	0.43758	0.01379	0.04655
Y05_Prestation	-0.24856	0.02348	0.21193	0.19639
Y06_Transfert	-0.39038	0.16735	0.33693	0.48377
Y07_Enfant	-0.13623	-0.30095	0.20170	0.53073
Y08_Placement	-0.02225	0.22313	-0.14780	-0.04909
Y09_Imp_fed	0.60276	0.18996	-0.24036	-0.11011
Y10_Imp_prov	0.61307	0.26579	-0.17681	-0.20026
Y11_Cred_tps	-0.46348	-0.40664	0.45628	0.12415
Y12_Cred_prov	-0.37869	-0.13835	0.06239	0.13036
Y13_C_enfant	0.10975	-0.18475	-0.06848	0.13683
Y14_Frais_med	0.09143	0.15555	-0.08707	0.17866
Y15_Cotisation	0.84527	0.20580	0.28967	0.24953
Y16_Cotis_oblig	0.86725	-0.24763	-0.24748	-0.22351

Tableau 2.16 Inter-corrélations U_1 à U_4

NAME	U1	U2	U3	U4
Y01_Salaire	0.59578	-0.03219	-0.05194	-0.03909
Y02_Trav_auto	-0.05851	0.11470	-0.16152	-0.04274
Y03_Autre_rev	-0.00706	0.03493	0.02122	0.01061
Y04_Agees	-0.04349	0.23517	0.00660	0.02045
Y05_Prestation	-0.18353	0.01262	0.10145	0.08625
Y06_Transfert	-0.28824	0.08994	0.16129	0.21247
Y07_Enfant	-0.10059	-0.16174	0.09656	0.23310
Y08_Placement	-0.01643	0.11992	-0.07075	-0.02156
Y09_Imp_fed	0.44506	0.10209	-0.11506	-0.04836
Y10_Imp_prov	0.45267	0.14285	-0.08464	-0.08795
Y11_Cred_tps	-0.34222	-0.21855	0.21843	0.05453
Y12_Cred_prov	-0.27961	-0.07436	0.02987	0.05726
Y13_C_enfant	0.08104	-0.09929	-0.03278	0.06010
Y14_Frais_med	0.06751	0.08360	-0.04168	0.07847
Y15_Cotisation	0.62411	0.11061	0.13867	0.10960
Y16_Cotis_oblig	0.64034	-0.13309	-0.11847	-0.09817

Tableau 2.17 Inter-corrélations V_1 à V_4

NAME	V1	V2	V3	V4
X01_Age	0.09885	0.36056	0.00158	0.06459
X02_Couple	0.09944	0.17563	-0.38433	0.10565
X03_Nb_menage	0.01714	-0.15984	-0.22393	0.34827
X04_Nb_etude	0.32583	0.00006	-0.01684	-0.06100
X05_Nb_etude_ps	0.30724	0.06290	-0.01639	-0.05232
X06_Nb_etude_uni	0.29515	0.14026	-0.00880	-0.03790
X07_Rurale	-0.09237	-0.01831	-0.02545	0.06946
X08_Loue	-0.16096	-0.13022	0.16232	-0.03706
X09_Nb_emp	-0.09022	-0.09451	0.05363	-0.04169
X10_Arret	-0.33504	0.08051	0.16116	0.06022
X11_Hres_trav	0.25706	-0.19036	-0.18776	-0.14350
X12_Hres_tmais	-0.03624	0.10388	-0.08662	-0.02224
X13_Hor_irr	-0.23142	0.15006	-0.04584	0.00023
X14_T_part	-0.24942	0.23766	0.06559	0.11281
X15_Nb_sem	0.34630	-0.17165	-0.15457	-0.12284
X16_Experience	0.18537	0.23482	-0.04072	0.00629
X17_Reg_ret	0.60351	0.01090	0.13178	0.08086
X18_Public	0.42840	0.09071	0.14781	0.08116

CHAPITRE III

ANALYSE CANONIQUE DE REDONDANCE

3.1 Introduction - Indice de redondance

L'analyse des corrélations canoniques nous aide à explorer un ensemble de données et nous pouvons ainsi tenter de trouver des modèles de dépendance intéressants. Jusqu'à présent dans la corrélation canonique, le seul indice asymétrique qui peut nous aider à cet effet est l'inter-corrélation (section 1.4.3). Cet indice nous permettait de constater dans quelle mesure une variable originale peut être expliquée par une variable canonique de l'autre groupe.

En 1968, Stewart et Love propose un indice basé sur la moyenne des coefficients de l'inter-corrélation d'un groupe par rapport à une variable canonique de l'autre groupe lorsque les variables sont **standardisées**. Le but est d'observer dans quelle mesure, un groupe de variables originales peut être expliqué par une variable canonique opposée (une combinaison linéaire de l'autre groupe). C'est donc un indice asymétrique entre les deux ensembles qui fournira une base afin de trouver des modèles de prédiction entre les groupes. L'asymétrie vient également du fait que la variance expliquée change selon le groupe de variables originales que nous désirons expliquer.

Cet indice est la proportion de la variance **totale d'un groupe** qui est expliquée par une combinaison linéaire (variable canonique) de l'autre groupe. Nous avions le carré de l'inter-corrélation qui signifiait la proportion de variance d'une variable originale

expliquée par une variable canonique. Alors,

$$RU_k^2 = \sum_{j=1}^{q} Corr(U_k, Y_j)^2/q \tag{3.1}$$

$$RV_k^2 = \sum_{i=1}^{p} Corr(V_k, X_i)^2/p \tag{3.2}$$

nous donnent la part de variabilité des variables originales d'un groupe expliquée par la variable canonique k, $k = 1, ..., s$, de l'autre groupe (en considérant les variables disponibles de cet ensemble). La redondance **totale** est donc la somme de ces k coefficients :

$$RU^2 = \sum_{k=1}^{s} RU_k^2 \tag{3.3}$$

$$RV^2 = \sum_{k=1}^{s} RV_k^2 \tag{3.4}$$

En 1976, Gleason propose une forme plus générale de cet indice pour des variables qui ne sont pas nécessairement standardisées. L'indice est basé sur la trace des matrices de variances-covariances. L'indice représente toujours une proportion de la variance totale d'un groupe qui est expliquée par une combinaison linéaire (variable canonique) de l'autre groupe.

Comme nous l'avons mentionné dans le chapitre précédent, la procédure *CANCORR* du logiciel SAS nous fournit la statistique de redondance. Nous l'étudierons dans la section de l'application de ce chapitre (section 3.7) . Plus l'indice de redondance est élevé, mieux les variables originales d'un groupe sont bien représentées par la variable canonique de l'autre groupe. Étant donné que le but de l'analyse des corrélations canoniques n'est pas de maximiser cet indice, il est possible qu'il soit faible même avec un coefficient de corrélation canonique très élevé.

Après avoir bien exploré les deux groupes de variables **X** et **Y** à l'aide de l'analyse des corrélations canoniques, nous pouvons maintenant tenter de prédire certaines variables d'un groupe (dites dépendantes) par des variables de l'autre groupe (dites indépendantes) et analyser la précision de cette prédiction. L'analyse canonique de redondance établit de **nouvelles** variables canoniques qui **maximisent** l'indice de redon-

dance. L'analyse canonique de redondance (Rao, 1964 ; van den Wollenberg, 1977) est aussi appelée régression à rang réduit dans un contexte de régression multi-variables (plusieurs variables explicatives et expliquées) ou analyse en composantes principales avec variables instrumentales. Dans ce chapitre, nous présenterons la théorie reliée à cette méthode à partir de l'article de van den Wollenberg (1977) et différentes approches qui permettront d'obtenir les mêmes résultats. Nous discuterons des façons d'interpréter les résultats grâce entre autres au graphique biplot appliqué à ce contexte. Par la suite, nous tenterons d'appliquer cette méthode sur nos deux groupes de variables étudiées avec l'analyse de la corrélation canonique. Le livre de Legendre et Legendre (1998, Chapitre 11) a été une importante source de référence pour la rédaction de ce chapitre. Nous avons utilisé également les livres de Reinsel et Velu (1998, chapitre 2), de Gower et Hand (1996, chapitre 11) ainsi que l'article de ter Braak (1994).

3.2 Théorie

Voici les principaux éléments de la théorie de l'analyse canonique de redondance tirés de l'article de van den Wollenderg (1977) qui utilisent des variables originales **standardisées**.

Pour la **première** paire de variables canoniques (U_1,V_1), l'indice de redondance nous donne les expressions suivantes :

$$RU_1^2 = \sum_{j=1}^{q} Corr(U_1, Y_j)^2/q \tag{3.5}$$

$$RV_1^2 = \sum_{i=1}^{p} Corr(V_1, X_i)^2/p \tag{3.6}$$

où X_i et Y_j sont les variables originales standardisées appartenant respectivement aux premier et deuxième groupes.

Pour l'équation 3.5, nous voulons trouver une nouvelle variable $U_1^* = \boldsymbol{\alpha'}_1\mathbf{X}$ (au lieu de U_1) de variance unitaire de sorte que $\sum_{j=1}^{q} Corr(U_1^*, Y_j)^2$ sera maximale. La redondance $RU_1^{*^2}$ sera alors maximale.

Remarque

Si $q = 1$, alors $Corr(U_1^*, Y_1)^2$ est maximale si $\boldsymbol{\alpha}_1 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y_1$. Le problème de redondance est difficile car c'est la somme (au carré) des corrélations de U_1^* avec chacune des variables de $\mathbf{Y}$ qui est maximisée. Nous verrons à la section suivante le lien avec la régression multiple.

Nous devons donc maximiser la fonction suivante :

$$\Phi = \sum_{j=1}^{q} Corr(Y_j, U_1^*)^2 - \lambda_1' \ (Var(U_1^*) - 1).$$

Comme $Corr(\mathbf{Y}, U_1^*) = R_{YX}\boldsymbol{\alpha}_1$ (voir 1.19) et $Var(U_1^*) = \boldsymbol{\alpha}_1' R_{XX}\boldsymbol{\alpha}_1$ alors,

$$\begin{aligned} \Phi &= (\mathbf{R}_{YX}\boldsymbol{\alpha}_1)'(\mathbf{R}_{YX}\boldsymbol{\alpha}_1) - \lambda_a' \ (\boldsymbol{\alpha}'\mathbf{R}_{XX}\boldsymbol{\alpha}_1 - 1) \\ &= \boldsymbol{\alpha}_1'\mathbf{R}_{XY}\mathbf{R}_{YX}\boldsymbol{\alpha}_1 - \lambda_a' \ (\boldsymbol{\alpha}'\mathbf{R}_{XX}\boldsymbol{\alpha}_1 - 1). \end{aligned} \tag{3.7}$$

Inversement, nous aurons pour 3.6 une nouvelle variable $V_1^* = \boldsymbol{\beta}_1'\mathbf{Y}$ (au lieu de V_1) de variance unitaire de sorte que $\sum_{i=1}^{p} Corr(V_1^*, X_i)^2$ sera maximale. La redondance $RV_1^{*^2}$ sera alors maximale. Ceci revient à maximiser la fonction suivante :

$$\Psi = \sum_{i=1}^{p} Corr(X_i, V_1^{*^2}) - \lambda_b' \ (Var(V_1^*) - 1).$$

Comme $Corr(\mathbf{X}, V_1^*) = \mathbf{R}_{XY}\boldsymbol{\beta}_1$ (voir 1.17) et $Var(V_1^*) = \boldsymbol{\beta}_1'\mathbf{R}_{YY}\boldsymbol{\beta}_1$,

$$\begin{aligned} \Psi &= (\mathbf{R}_{XY}\boldsymbol{\beta}_1)'(\mathbf{R}_{XY}\boldsymbol{\beta}_1) - \lambda_b' \ (\boldsymbol{\beta}_1'\mathbf{R}_{YY}\boldsymbol{\beta}_1 - 1) \\ &= \boldsymbol{\beta}_1'\mathbf{R}_{YX}\mathbf{R}_{XY}\boldsymbol{\beta}_1 - \lambda_b' \ (\boldsymbol{\beta}_1'\mathbf{R}_{YY}\boldsymbol{\beta}_1 - 1). \end{aligned} \tag{3.8}$$

Les dérivés partielles de 3.7 et 3.8 par rapport à $\boldsymbol{\alpha}_1$ et $\boldsymbol{\beta}_1$ égales à 0 nous donnent :

$$\frac{\partial \Phi}{\partial \boldsymbol{\alpha}_1} = \mathbf{R}_{XY}\mathbf{R}_{YX}\boldsymbol{\alpha}_1 - \lambda_a'\mathbf{R}_{XX}\boldsymbol{\alpha}_1 = 0$$

$$\frac{\partial \Psi}{\partial \boldsymbol{\beta}_1} = \mathbf{R}_{YX}\mathbf{R}_{XY}\boldsymbol{\beta}_1 - \lambda_b'\mathbf{R}_{YY}\boldsymbol{\beta}_1 = 0$$

qui peuvent être écrite sous la forme suivante :

$$(\mathbf{R}_{XY}\mathbf{R}_{YX} - \lambda'_a\mathbf{R}_{XX})\,\boldsymbol{\alpha}_1 = 0 \tag{3.9}$$

$$\Rightarrow\ (\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX} - \lambda'_a)\,\boldsymbol{\alpha}_1 = 0, \tag{3.10}$$

$$(\mathbf{R}_{YX}\mathbf{R}_{XY} - \lambda'_b\mathbf{R}_{YY})\,\boldsymbol{\beta}_1 = 0 \tag{3.11}$$

$$\Rightarrow\ (\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}\mathbf{R}_{XY} - \lambda'_b)\,\boldsymbol{\beta}_1 = 0. \tag{3.12}$$

Comme dans la théorie de la corrélation canonique, nous sommes en présence de deux problèmes de valeurs propres et de vecteurs propres d'une matrice carrée de rang plein. L'équation 3.10 correspond à l'équation 1.8 avec $\mathbf{\Sigma}_{YY}^{-1} = I$ et 3.12 correspond à 1.7 avec $\mathbf{\Sigma}_{XX}^{-1} = I$. Cependant, en analyse canonique de redondance, les valeurs propres associées aux matrices $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$ et $\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}\mathbf{R}_{XY}$ ne sont pas les mêmes ($\lambda'_a \neq \lambda'_b$, d'où l'asymétric).

En multipliant 3.9 par $\boldsymbol{\alpha}'_1$ et en sachant que $\boldsymbol{\alpha}'_1 R_{XX}\boldsymbol{\alpha}_1 = 1$, nous obtenons que

$$\begin{aligned}\boldsymbol{\alpha}'_1\mathbf{R}_{XY}\mathbf{R}_{YX}\boldsymbol{\alpha}_1 &= \lambda'_a\\ \sum_{j=1}^{q} Corr(Y_j, U_1^*)^2 &= \lambda'_a\end{aligned}$$

qui correspond à la plus grande valeur propre de $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$. De même, en mutipliant 3.11 par $\boldsymbol{\beta}'_1$ et en sachant que $\boldsymbol{\beta}'_1 R_{YY}\boldsymbol{\beta}_1 = 1$, nous obtenons que

$$\begin{aligned}\boldsymbol{\beta}'_1\mathbf{R}_{YX}\mathbf{R}_{XY}\boldsymbol{\beta}_1 &= \lambda'_b\\ \sum_{i=1}^{q} Corr(X_i, V_1^*)^2 &= \lambda'_b\end{aligned}$$

qui correspond à la plus grande valeur propre de $\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}\mathbf{R}_{XY}$. Par 3.5 et 3.6, les redondances des premières variables canoniques correspondent respectivement pour les groupes $\mathbf{Y}$ et $\mathbf{X}$ à

$$RU_1^{*^2} = \lambda'_a/q \ \text{ et } \ RV_1^{*^2} = \lambda'_b/p.$$

Étant donné que l'analyse canonique de redondance nous fournit la meilleure prédiction possible d'un groupe de variables via une variable canonique de redondance définie sur l'autre groupe (une combinaison linéaire), cela implique qu'il existe une relation de dépendance entre les groupes que nous aurons établie au préalable (asymétrie dans les données). Il n'est donc pas toujours nécessaire de calculer les **deux** redondances $RU_1^{*^2}$ ($\mathbf{Y}$ via $\mathbf{X}$) et $RV_1^{*^2}$ ($\mathbf{X}$ via $\mathbf{Y}$). Par convention, les $\mathbf{X}$ sont habituellement les variables indépendantes et les $\mathbf{Y}$ les variables que nous tentons de prédire. Nous pourrons alors calculer la redondance $RU_1^{*^2}$ qui est associée est l'équation 3.10.

Pour obtenir les redondances suivantes des $\mathbf{Y}$ via les $\mathbf{X}$ ($RU_k^{*^2}$), nous n'avons qu'à prendre la $k^{\text{ième}}$ plus grande valeur propre de la matrice $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$ qui sera associée au vecteur propre $\boldsymbol{\alpha}_k$, $j = 2, ..., s$ où $s = min(p, q)$. Il ne reste qu'à normaliser $\boldsymbol{\alpha}_k$ pour que $\boldsymbol{\alpha}_k'\mathbf{R}_{XX}\boldsymbol{\alpha}_k = 1$. La redondance totale ($RU^{*^2}$) correspond donc à la somme des valeurs propres de la matrice $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$ divisée par q :

$$\begin{aligned} RU^{*^2} &= \sum_{k=1}^{s} RU_k^{*^2} \\ &= \sum_{k=1}^{s} \lambda'_{a_k} \ / \ q \end{aligned} \tag{3.13}$$

Nous discuterons de l'interprétation de ces résultats un peu plus loin.

Si nous voulons calculer toutes les variables canoniques de deux groupes, les U_i^* et V_j^*, $i \neq j$, elles ne sont pas en général orthogonales étant donné qu'elles sont déterminées séparément sans contrainte d'orthogonalité. Wollenderg propose une transformation afin de rendre les U_i^* et V_j^* orthogonales. Plusieurs articles ont été écrits afin de proposer des alternatives à cette problématique.

3.3 Autres méthodes pour obtenir l'analyse canonique de redondance

3.3.1 Régression multiple et ACP

Il est intéressant de mentionner que dans leur livre *Numerical ecology*, Pierre Legendre et Louis Legendre démontrent que l'analyse canonique de redondance est une

régression multiple sur chaque variable de $\mathbf{Y}$ suivi d'une analyse en composante principale. À noter qu'ils ne considèrent qu'une seule redondance ($\mathbf{Y}$ via les $\mathbf{X}$) étant donné le caractère asymétrique de la méthode.

Pour $\underset{n\times p}{\mathbf{X}}$ et $\underset{n\times q}{\mathbf{Y}}$ standardisées, la régression multiple sur chaque Y_j est donné par l'expression suivante :

$$\underset{n\times q}{\widehat{\mathbf{Y}}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times q}{\mathbf{B}} \quad \text{où} \quad \mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Étant donné que $\mathbf{X}$ et $\mathbf{Y}$ sont standardisées (ou centrées), la matrice $\mathbf{B}$ ne contient pas d'ordonnées à l'origine. La matrice de covariance de $\widehat{\mathbf{Y}}$ correspond à

$$\begin{aligned}
\mathbf{R}_{\widehat{Y}'\widehat{Y}} &= [1/(n-1)]\ \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} \\
&= [1/(n-1)]\ \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \mathbf{R}_{YX}\ \mathbf{R}_{XX}^{-1}\ \mathbf{R}_{XY}
\end{aligned}$$

Lorsque nous appliquons une analyse en composantes principales sur l'ensemble $\widehat{\mathbf{Y}}$, ceci correspond à résoudre le problème de vecteur propre suivant :

$$(\mathbf{R}_{\widehat{Y}'\widehat{Y}} - \lambda_k\ \mathbf{I})\ \mathbf{u}_k = 0$$

qui revient à résoudre

$$(\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY} - \lambda_k\ \mathbf{I})\ \mathbf{u}_k = 0. \tag{3.14}$$

où λ_k sont les valeurs propres de $\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$ et $\mathbf{u}_k$ sont les vecteurs propres associés, $k = 1, ..., s$, $s = min(p, q)$ (si $\mathbf{X}$ et $\mathbf{Y}$ de rang plein).

Lien avec la corrélation canonique et redondance canonique

Remarquons que l'équation 3.14 correspond à l'équation 1.7 de la corrélation canonique avec $\mathbf{\Sigma}_{YY}^{-1} = I$. Rappelons également que l'équation de la redondance canonique 3.10 des $\mathbf{Y}$ via les $\mathbf{X}$,

$$(\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX} - \lambda_a')\ \boldsymbol{\alpha}_1 = 0,$$

correspondait à l'équation 1.8 de la corrélation canonique avec $\mathbf{\Sigma}_{YY}^{-1} = I$.

Alors les deux matrices :

$$\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY} \text{ (3.14, Legendre et Legendre) et}$$

$$\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX} \text{ (3.10, van den Wollenberg),}$$

correspondent respectivement aux matrices $\mathbf{E}_1$ et $\mathbf{E}_2$ (1.9) avec $\mathbf{\Sigma}_{YY}^{-1} = I$ dans la théorie du calcul des corrélations canoniques. Comme nous l'avons déjà démontré dans cette théorie, ces deux matrices ont s **valeurs propres communes** (λ_i, $i = 1, ..., s$).

Comme nous l'avons démontré à la section précédente, λ'_{a_k}, $k = 1, ..., s$ sont les valeurs propres de $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$ (van den Wollenberg) et la somme de ces λ'_{a_k} divisée par q est la redondance totale RU^{*^2} des $\mathbf{Y}$ via les $\mathbf{X}$ (3.13).

Alors, ces valeurs propres sont communes à la matrice $\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$ (Legendre et Legendre) et nous obtenons les **mêmes redondances** des $\mathbf{Y}$ via les $\mathbf{X}$ ($RU_k^{*^2}$ et RU^{*^2}). Par conséquent, la méthode de van den Wollenberg et celle présentée par Legendre et Legendre sont des méthodes équivalentes.

La matrice $\underset{q\times s}{\mathbf{U}} = (\mathbf{u}_1 ... \mathbf{u}_s)$ est donc la matrice des vecteurs propres (combinaisons linéaires) de $\mathbf{R}_{\widehat{Y}\widehat{Y}}$. Les nouvelles coordonnées de l'**ACP** sont alors données par

$$\underset{n\times s}{\mathbf{Z}} = \widehat{\mathbf{Y}}\mathbf{U} = \mathbf{XBU}. \tag{3.15}$$

Les résultats de chacune des deux méthodes ne s'interpréteront pas nécessairement de la même façon. Pour la méthode de van den Woolenberg, nous utiliserons les mesures d'interprétation présentées en analyse des corrélations canoniques. Legendre et Legendre quant à eux suggèrent surtout l'interprétation avec certains graphiques biplot. Nous en présenterons un qui sera un graphique biplot du modèle.

3.3.2 Analyse en composantes principales

L'analyse en composantes principales peut être vue comme une analyse canonique de redondance sur le même groupe de variables. En effet, si $\mathbf{X} = \mathbf{Y}$ l'équation 3.10

revient à

$$(\mathbf{R}_{XX}^{-1}\mathbf{R}_{XX}\mathbf{R}_{XX} - \lambda_a'\mathbf{I})\ \boldsymbol{\alpha} = 0,$$

qui correspond à l'équation caractéristique de l'analyse en composante principale

$$(\mathbf{R}_{XX} - \lambda_a'\mathbf{I})\ \boldsymbol{\alpha} = 0.$$

L'analyse en composante principale peut donc être vue comme un cas particulier de l'analyse canonique de redondance avec $\mathbf{X} = \mathbf{Y}$.

3.3.3 Régression à rang réduit

L'analyse canonique de redondance est rarement présentée comme une des techniques multidimentionnelles dans les manuels statistiques. Elle n'est souvent mentionnée qu'à titre d'extension de l'analyse des corrélations canoniques. Ce sont surtout les références statistiques de biologie et d'écologie qui présentent et exploitent le mieux cette méthode. Les références statistiques parlent plutôt de **régression à rang réduit** dans un contexte plus général de régression à plusieurs variables explicatives et expliquées (multi-variables). Cette méthode est proposée lorsque les variables réponses sont corrélées. Afin de tenir compte de ces corrélations, la matrice de régresseurs est contrainte par un rang inférieur. Nous présentons ici un bref aperçu de cette méthode.

Le modèle habituel de régression linéaire multi-variables (et non multiple) pour les $\mathbf{X}$ et les $\mathbf{Y}$ centrés se présente de la façon suivante :

$$\underset{n\times q}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times q}{\mathbf{M}} + \underset{n\times q}{\mathbf{E}}, \tag{3.16}$$

où la matrice $\mathbf{M}$ contient $p \times q$ coefficients de régression. Afin de réduire la dimension de la matrice des coefficients $\mathbf{M}$, nous la décomposons de la façon suivante :

$$\underset{p\times q}{\mathbf{M}} = \underset{p\times r}{\mathbf{A}}\ \underset{r\times q}{\mathbf{B}},$$

où les matrices $\mathbf{A}$ et $\mathbf{B}$ sont de rang r. Le modèle de régression à rang réduit peut maintenant s'écrire sous la forme suivante

$$\underset{n\times q}{\mathbf{Y}} = (\ \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{A}}\ )\ \underset{r\times q}{\mathbf{B}} + \underset{n\times q}{\mathbf{E}},$$

où $\underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{A}}$ est de dimension réduite en r "composantes". En pratique, ces r combinaisons linéaires des prédicteurs sont suffisantes pour modéliser la variation des $\mathbf{Y}$.

Plusieurs méthodes d'analyse statistiques sont des cas particuliers de ce modèle. L'historique de cette méthode et sa théorie sont très bien présentés dans le livre de Reinsel et Velu (1998). Comme le font remarquer les auteurs (p.34), l'estimation par les moindres carrés de ce modèle pour des $\mathbf{X}$ et $\mathbf{Y}$ **standardisées** nous ramène à l'analyse canonique de redondance décrite par van den Wollenberg et également par Legendre et Legendre (Fortier, 1966). En effet, il serait possible de démontrer que

$$\underset{n\times q}{\hat{\mathbf{Y}}} = (\ \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\hat{\mathbf{A}}}\ )\ \underset{r\times q}{\hat{\mathbf{B}}} \tag{3.17}$$

est donné par

$$\underset{p\times r}{\hat{\mathbf{A}}} = \mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{U},\quad \underset{r\times q}{\hat{\mathbf{B}}} = \mathbf{U}'$$

où $\underset{q\times r}{\mathbf{U}}$ est la matrice des r vecteurs propres (normalisés) de $\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$.

La portion $\underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\hat{\mathbf{A}}}$ est suffisante pour expliquer la variation des $\mathbf{Y}$ et correspond exactement à l'équation 3.15 du modèle d'analyse canonique de redondace présenté par Legendre et Legendre.

3.4 Inférence sur les composantes successives

En 2002, Lazraq et Cléroux proposent, sous l'hypothèse de multinormalité, des tests inférentiels reliés aux variables canoniques de redondance (U_k^*, $k = 1, ..., s$, $s = min(p, q)$). Le but est de réduire la dimension du problème en ne conservant qu'un sous-ensemble significatif de r variables canoniques de redondance afin de prédire $\mathbf{Y}$ ($r < s$).

Les auteurs utilisent l'indice de redondance généralisé proposé par Gleason en 1976. Cet

indice de redondance totale des $\mathbf{Y}$ via les $\mathbf{X}$ pour la population sera noté :

$$\begin{aligned} \rho U^{*^2} &= \sum_{k=1}^{s} \rho U_k^{*^2} \\ &= \frac{\sum_{k=1}^{s} \lambda_k}{tr(\mathbf{\Sigma}_{YY})} \\ &= \frac{tr(\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY})}{tr(\mathbf{\Sigma}_{YY})} \end{aligned}$$

où λ_k sont les valeurs propres non nulles de $\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}$ triées en ordre décroissant.

Les auteurs démontrent que

$$(\rho U_1^{*^2} + ... + \rho U_{k-1}^{*^2} + \rho U_k^{*^2}) = (\rho U_1^{*^2} + ... + \rho U_{k-1}^{*^2}) \Leftrightarrow \rho U_k^{*^2} = 0$$

ce qui implique que $U_k^* = \boldsymbol{\alpha'}_k\mathbf{X}$ n'a pas de contribution dans la prédiction de $\mathbf{Y}$ en terme de redondance. Alors,

$$\rho U_k^{*^2} = 0 \Leftrightarrow \lambda_k = 0 \Leftrightarrow \lambda_k = \lambda_{k+1} = ... = \lambda_s = 0.$$

Nous obtenons alors la signification de la variable canonique de redondance U_k^* dans la prédiction de $\mathbf{Y}$ comme une fonction de λ_k, la $k^{\text{ième}}$ valeur propre de $\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}$. Le test pour U_k^* se présente de la façon suivante :

$H_0 : \rho U_k^{*^2} = 0$ (contre $H_1 : \rho U_k^{*^2} > 0$) est rejetée à un niveau α si

$$\frac{RU_k^{*^2}}{1 - RU_k^{*^2}} > g_\alpha$$

où g_α est le $100(1-\alpha)^{\text{ième}}$ percentile de la distribution de $\sum_{i=1}^{p(n-1)} \gamma_i W_i^2$ qui est une fonction linéaire de variables aléatoires indépendantes d'une χ^2 centrale. Les valeurs de g_α sont obtenues en utilisant l'algorithme de Imhof (1961) dont la description et le code sont donnés dans Koertz et Abrahamse (1969). À noter que ce test n'est valide que pour des grandes valeurs de n étant donné que la distribution est dépendante de $\mathbf{\Sigma}_{YY}$ (inconnue) qui est estimée par $\mathbf{S}_{YY}$.

ALGORITHME

Une fois que toutes les redondances $RU_k^{*^2}$ sont obtenues (par ordre décroissant) :

- **Étape 0** : $k = 1$
- **Étape 1** : Si $\frac{RU_k^{*2}}{1-RU_k^{*2}} \leq g_\alpha$, impossible de rejeter H_0 alors STOP.
- **Étape 2** : $k = k + 1$, si $k \leq s$, retour à l'étape 1 sinon STOP.

Nous obtenons alors les k redondances les plus significatives et par conséquent, les k variables canoniques de redondance les plus significatives dans la relation entre $\mathbf{Y}$ et $\mathbf{X}$. La redondance totale des $\mathbf{Y}$ via les $\mathbf{X}$ sera donc la somme des k redondances les plus significatives. Si aucune n'est significative (ne rejette pas H_0 à la première redondance), il est alors impossible d'expliquer linéairement les $\mathbf{Y}$ par les $\mathbf{X}$ en terme de redondance.

3.5 Interprétation des résultats

Étant donné que les références sur l'analyse canonique de redondance sont rares et peu détaillées, la discussion sur l'interprétation des résultats est souvent négligée. Comme cette méthode provient de l'analyse des corrélations canoniques, nous allons reprendre quelques mesures associées à l'interprétation des résultats (section 1.4) et les adapter à l'analyse canonique de redondance. Les résultats et les mesures associés ne seront présentés qu'en fonction d'un seul groupe de variables (les $\mathbf{X}$ par convention) ce qui allège beaucoup l'interprétation face à l'analyse des corrélations canoniques.

3.5.1 Vecteurs canoniques (α_k') et intra-corrélations

Les variables canoniques de redondance U_k^*, $k = 1, ..., s$ sont des combinaisons linéaires des $\mathbf{X}$ où la corrélation entre $\mathbf{Y_j}$ et U_k^* est la plus grande. Les poids $\boldsymbol{\alpha}_k$ d'un vecteur canonique représentent donc la contribution de chaque variable du groupe $\mathbf{X}$ qui explique le plus de variabilité du groupe $\mathbf{Y}$. Étant donné que les variables originales sont standardisées, nous pourrons comparer les poids entre eux et déterminer quelles variables du groupe $\mathbf{X}$ sont les plus importantes dans la relation.

Comme nous l'avons vu, les poids canoniques ne sont pas des plus stables (similaires aux β en régression). Pour en faire une interprétation plus fiable, il faut que les intra-corrélation présentée à la section 1.4.3 soient dans le même sens que ces poids. Nous

calculerons alors pour la variable canonique U_k^*, $k = 1, ..., s$:

$$\begin{aligned} \underset{(1 \times p)}{Corr(U_k^*, \mathbf{X})} &= \frac{Cov(U_k^*, \mathbf{X})}{\sqrt{Var(\mathbf{X})}} \\ &= Cov(\boldsymbol{\alpha}_k' \mathbf{X}, \mathbf{X}) \\ &= \boldsymbol{\alpha}_k' Cov(\mathbf{X}, \mathbf{X}) \\ &= \boldsymbol{\alpha}_k' \mathbf{R}_{XX} \end{aligned}$$

Ces valeurs devront être comparées à $\boldsymbol{\alpha}_k'$ pour s'assurer de la stabilité et elles pourront alors être bien interprétées.

3.5.2 Indice de redondance et inter-corrélation

En analyse des corrélations canoniques, l'indice de redondance n'était pas nécessairement maximal. La technique de l'analyse canonique de redondance maximise cet indice. Nous avons donc la **plus grande** proportion de la variance totale des $\mathbf{Y}$ expliquée par une combinaison linéaire ($\boldsymbol{\alpha}_k'$) des $\mathbf{X}$. Plus cet indice est élevé, plus la prédiction des $\mathbf{Y}$ par la variable canonique est bonne. Nous les analyserons de la façon suivante :

L'indice de redondance sélectionné

$$RU_k^{*2} = \sum_{j=1}^{q} Corr(U_k^*, Y_j)^2 / q,$$

est composé de la somme des coefficients de l'**inter-corrélation** au carré (section 1.4.3) et correspondent à

$$\begin{aligned} Corr(U_k^*, Y_j) &= \frac{Cov(\boldsymbol{\alpha}_k' \mathbf{X}, Y_i)}{\sqrt{Var(Y_j)}} \\ &= \boldsymbol{\alpha}_k' Cov(\mathbf{X}, Y_j) \\ &= \boldsymbol{\alpha}_k' \mathbf{r}_{XY_j} \end{aligned}$$

où $\mathbf{r}_{XY_j}$ = la $j^{\text{ième}}$ colonne de $\mathbf{R}_{XY}$.

Ce coefficient est similaire au coefficient de corrélation multiple. Son carré représente le pourcentage de la dispersion de Y_j expliqué par le calcul de la variable canonique

U_k^*. Comme nous avons maximisé la somme des carrés de ces coefficients (définition de l'analyse canonique de redondance) pour un k donné, il sera facile d'observer les coefficients qui contribuent le plus à l'indice de redondance maximal. Ils nous permettront de comprendre les Y_j qui sont les plus expliquées par U_k^*. Ce sont ces variables qui seront les plus importantes dans la relation de dépendance avec $\mathbf{X}$.

3.6 Graphique biplot

Le graphique biplot que nous présenterons sommairement est obtenu à partir des modèles présentés par Legendre et Legendre et par la régression à rang réduit. Il représente comment les combinaisons linéaires de $\mathbf{X}$ expliquent la variation des $\mathbf{Y}$ à l'aide des axes canoniques (vecteurs propres) qui maximisent la variance. C'est un biplot du **modèle** étudié qui aide à l'interprétation des capacités de prédiction.

3.6.1 Modèle présenté par Legendre et Legendre

Sans utiliser la décomposition en valeurs singulières (qui démontre l'optimalité), Legendre et Legendre présentent les graphiques biplot comme étant des techniques simples d'interprétation des résultats. Comme les auteurs ont présenté le modèle avec une analyse en composantes principales (ACP) sur $\hat{\mathbf{Y}}$, un biplot peut s'effectuer de la même façon, c'est-à-dire que la matrice $\hat{\mathbf{Y}}$ se décompose en produit de deux matrices :les valeurs observées de l'ACP (scores) et les vecteurs propres. Il correspond donc à

$$\underset{n\times q}{\hat{\mathbf{Y}}} = \underset{n\times s}{\mathbf{Z}}\ \underset{s\times q}{\mathbf{U}}\,' \tag{3.18}$$

À noter que ce modèle de décomposition de $\hat{\mathbf{Y}}$ correspond exactement au modèle complet de prédiction de la régression à rang réduit (3.17).

Nous obtenons une estimation de $\hat{\mathbf{Y}}$ lorsque nous posons $s < min(p, q)$. Lorsque $s = 2$, l'estimation de $\hat{\mathbf{Y}}$ se représente dans un graphique car la projection de $\underset{n\times 2}{\mathbf{Z}}$ sur $\underset{2\times q}{\mathbf{U}}\,'$ se trace dans le plan. Nous traçons alors les lignes de $\underset{n\times 2}{\mathbf{Z}}$ comme n vecteurs, notés $\mathbf{z}_i'$, $i = 1, ..., n$, et les colonnes de $\underset{2\times q}{\mathbf{U}'}$ comme q vecteurs, notés $\mathbf{u}_j$, $j = 1, ..., q$.

Nous avons donc

$$\underset{n \times q}{\hat{\mathbf{Y}}} \simeq \underset{[2]}{\hat{\mathbf{Y}}} = \underset{n \times 2}{\mathbf{Z}} \; \underset{2 \times q}{\mathbf{U}}{}'$$

où $\underset{[2]}{\hat{\mathbf{Y}}}$ est l'estimation de rang 2 de la matrice $\underset{n \times q}{\hat{\mathbf{Y}}}$.

Le produit scalaire est obtenu en projetant le vecteur $\mathbf{z}_i'$ sur le vecteur de $\mathbf{u}_j$ et en multipliant les longueurs de $\mathbf{u}_j$ par la projection du vecteur $\mathbf{z}_i'$:

$$\begin{aligned} \hat{y}_{ij} \simeq \underset{[2]}{\hat{y}_{ij}} &= \mathbf{z}_i' \mathbf{u}_j \\ &= \frac{\mathbf{z}_i' \mathbf{u}_j}{\|\mathbf{u}_j\|^2} \|\mathbf{u}_j\|^2 \\ &= Proj(\mathbf{z}_i \; sur \; \mathbf{u}_j) \|\mathbf{u}_j\|^2. \end{aligned}$$

Le biplot est intéressant car comme nous l'avons mentionné en introduction, nous sommes en mesure de représenter dans un graphique comment les observations mesurées sur les $\mathbf{X}$ (par $\underset{n \times 2}{\mathbf{Z}}$) expliquent les q variables de $\mathbf{Y}$ (par $\underset{2 \times q}{\mathbf{U}'}$). Il devient alors plus facile d'étudier la prédiction entre les deux groupes de variables.

Rappelons que plus les valeurs propres de $\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$ sont élevées, plus la prédiction de $\mathbf{Y}$ par $\mathbf{X}$ en utilisant le modèle est bonne (redondance élevée). L'approximation par le biplot sera efficace si les **deux** premières valeurs propres sont élevées.

3.6.2 Modèle de la régression à rang réduit

Dans le livre *Biplots* (1996), Gower présente sommairement le biplot du modèle plus général de régression à rang réduit. Il reprend principalement les travaux de ter Braak (1994) qui présente l'estimation du modèle par la décomposition en valeurs singulières, d'où l'optimalité du biplot. Ces deux auteurs expliquent également que l'analyse canonique de redondance n'est qu'un cas spécifique de la régression à rang réduit. Ils démontrent essentiellement que l'estimation du modèle de régression à rang réduit (3.16) s'obtient par la décomposition en valeurs singulières généralisée de rang $r \leq min(p, q)$. Par le théorème d'Eckart-Young (1.5.2.1), l'estimation de rang r sera la meilleure approximation au sens des moindres carrés. Par cette décomposition, le graphique biplot

du modèle avec $r = 2$ sera alors la meilleure approximation de rang 2.

Le minimum de $||\underset{n\times q}{\mathbf{Y}} - \underset{n\times p}{\mathbf{X}}\,\underset{p\times q}{\mathbf{M}}||$ de rang $r < min(p,q)$ s'obtient par la décomposition en valeurs singulières **généralisées** de $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ et est donné par

$$\underset{p\times q}{\hat{\mathbf{M}}_\mathbf{r}} = \mathbf{S}_{[r]}\Sigma_{[r]}\mathbf{T}'_{[r]}$$

de rang r où

$\mathbf{S}_{[r]}$ = les r premières colonnes de $\underset{p\times s}{\mathbf{S}}$,
$\mathbf{T}'_{[r]}$ = les r premières lignes de $\underset{s\times q}{\mathbf{T}'}$ et
$\Sigma_{[r]}$ = la matrice diagonale des r premières valeurs singulières de $\underset{s\times s}{\mathbf{\Sigma}}$, $s = min(p,q)$.
Comme

$$\underset{p\times q}{\mathbf{M_0}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

est l'estimation par les moindres carrés du modèle de régression multi-variables alors,

$$\underset{p\times q}{\mathbf{M_0}} = \underset{p\times s}{\mathbf{S}}\,\underset{s\times s}{\mathbf{\Sigma}}\,\underset{s\times q}{\mathbf{T}}\,'.$$

Étant donné que $\underset{p\times q}{\hat{\mathbf{M}}_\mathbf{r}}$ approxime $\underset{p\times q}{\mathbf{M_0}}$, nous pouvons réécrire $\underset{p\times q}{\hat{\mathbf{M}}_\mathbf{r}}$ de la façon suivante :

$$\begin{aligned}\underset{p\times q}{\hat{\mathbf{M}}_\mathbf{r}} &= \mathbf{M_0}\,\mathbf{T}_{[r]}\mathbf{T}'_{[r]} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\;\mathbf{T}_{[r]}\mathbf{T}'_{[r]}.\end{aligned}$$

Donc $\underset{n\times p}{\mathbf{X}}\,\underset{p\times q}{\hat{\mathbf{M}}_\mathbf{r}}$ correspond (à une constante près) à l'estimation du modèle de la régression à rang réduit présenté à 3.17. Lorsque les $\mathbf{X}$ et $\mathbf{Y}$ sont standardisées, nous obtenons le modèle d'analyse canonique de redondance. Le modèle correspond à l'équation 3.18 utilisée par Legendre et Legendre pour tracer le biplot.

En utilisant la décomposition en valeurs singulières **généralisée**, nous sommes en mesure d'obtenir une estimation de rang r de $\mathbf{XM}$ qui est **optimale** au sens des moindres carrés.

Pour $r = 2$,

$$\begin{aligned}\underset{n\times p}{\mathbf{X}}\,\underset{p\times q}{\hat{\mathbf{M}}_\mathbf{2}} &= \mathbf{X}\mathbf{S}_{[2]}\Sigma_{[2]}T'_{[2]} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\;\mathbf{T}_{[2]}\mathbf{T}'_{[2]}.\end{aligned}$$

et nous pouvons tracer cette estimation comme un biplot avec les lignes de $X\mathbf{S}_{[2]}\Sigma_{[2]}$ comme n vecteurs et les colonnes de $T'_{[2]}$ comme q vecteurs. L'interprétation se fera comme nous l'avons expliquée précédemment dans le biplot de Legendre et Legendre.

3.7 Application à l'aide des données à l'étude

Nous venons de présenter la méthode de l'analyse canonique de redondance. Cette méthode explore la **prédiction** entre deux groupes de variables plutôt que les **associations** entre elles comme le suggérait l'analyse des corrélations canoniques. Nous tenterons maintenant d'appliquer cette méthodologie sur notre ensemble de données. Rappelons que nous devons utiliser les variables originales standardisées.

3.7.1 Indice de redondance

Comme suggéré par Stewart et Love, il est possible de calculer l'indice de redondance à partir des coefficients de l'inter-corrélation obtenus en analyse des corrélations canoniques (tableaux 2.16 et 2.17). La procédure *CANCORR* de SAS nous donne l'indice de redondance pour chacun des groupes de variables. Cet indice est la proportion de la variance totale d'un groupe qui est expliquée par une variable canonique (combinaison linéaire) de l'autre groupe.

Au tableau 3.1, nous avons l'indice de redondance des $\mathbf{Y}$ via U_k (RU_k^2 de l'équation 3.1). La colonne `Proportion` nous donne la variabilité des $\mathbf{Y}$ expliquée par U_k. Seule la variable U_1 semble expliquer légèrement la variance des $\mathbf{Y}$ avec 11,9 %. La redondance totale correspond à 17,89 % de la variance des $\mathbf{Y}$ expliquée par les U_k.

Le tableau 3.2 nous donne les résultats de l'indice de redondance des $\mathbf{X}$ via V_k (RV_k^2 de l'équation 3.2). Nous remarquons que la redondance totale correspond à seulement 15,70 % de la variance des $\mathbf{X}$ expliquée par les V_k.

À la lumière de ces résultats, nous pouvons remarquer qu'il est difficile d'expliquer beaucoup de variance des groupes de variables originales avec nos variables canoniques.

Tableau 3.1 Indice de redondance des $\mathbf{Y}$ via U_k (RU_k^2), Output SAS

Canonical Redundancy Analysis

Standardized Variance of the WITH Variables Explained by The Opposite Canonical Variables

Canonical Variable Number	Proportion	Cumulative Proportion
1	0.1185	0.1185
2	0.0157	0.1342
3	0.0116	0.1457
4	0.0099	0.1557
5	0.0110	0.1667
6	0.0043	0.1710
7	0.0040	0.1749
8	0.0018	0.1768
9	0.0009	0.1777
10	0.0006	0.1782
11	0.0002	0.1785
12	0.0001	0.1786
13	0.0001	0.1787
14	0.0001	0.1788
15	0.0001	0.1789
16	0.0000	0.1789

Tableau 3.2 Indice de redondance $\mathbf{X}$ via V_k (RV_k^2), Output SAS

Canonical Redundancy Analysis

Standardized Variance of the VAR Variables Explained by The Opposite Canonical Variables

Canonical Variable Number	Proportion	Cumulative Proportion
1	0.0749	0.0749
2	0.0257	0.1005
3	0.0205	0.1210
4	0.0121	0.1331
5	0.0092	0.1423
6	0.0075	0.1497
7	0.0036	0.1533
8	0.0023	0.1556
9	0.0005	0.1561
10	0.0004	0.1564
11	0.0002	0.1566
12	0.0001	0.1567
13	0.0001	0.1568
14	0.0001	0.1569
15	0.0001	0.1570
16	0.0000	0.1570

Seules les variables canoniques de notre relation la plus forte (relation 1) semble expliquer un peu de variance des groupes de variables originales.

3.7.2 Indice de redondance maximal

Nous appliquons maintenant la méthode de l'analyse canonique de redondance qui établit de **nouvelles** variables canoniques qui **maximisent** l'indice de redondance (la variance expliquée de l'autre groupe).

Pour appliquer cette méthode de prédiction, nous devons avoir un groupe de variables dépendantes et un groupe de variables indépendantes. Nous venons de voir que les $\mathbf{X}$ semblent mieux expliquer la variation des $\mathbf{Y}$ par les U_k. De plus, nous avons mentionné à la section 2.6 qu'il est logique d'avoir des variables fiscales ($\mathbf{Y}$) dépendantes des variables sociodémographiques ($\mathbf{X}$). Nous conserverons cette relation de dépendance pour appliquer l'analyse canonique de redondance.

Il est assez laborieux d'appliquer cette méthode avec le logiciel SAS car aucune procédure nous fournit directement les résultats. Nous devons faire de la programmation et quelques calculs manuels pour y arriver. Nous fournissons dans les appendices B et C les programmes qui nous ont permis d'y arriver. Nous avons appliqué l'analyse canonique de redondance selon les calculs de van den Wollenberg et selon la méthode présentée par Legendre et Legendre (régression multiple et ACP) afin de s'assurer qu'elles sont bien équivalentes.

Méthode van den Wollenberg

Pour appliquer cette méthode, nous avons utilisé la procédure *TRANSREG* de SAS avec l'option *METHOD=REDUNDANCY*. La procédure ne nous fournit pas directement en *Output* la redondance maximale pour chacune des nouvelles variables canoniques (RU_k^{*2}). Nous devons calculer RU_k^{*2} par programmation. La procédure nous fournit seulement la redondance totale qui est la même que celle des corrélations canoniques (17,89 %). C'est l'indice de redondance de chaque variable canonique RU_k^{*2}

qui sera maximal comparativement à celui obtenu par les corrélations canoniques RU_k^2. Nous avons produit les résultats au tableau 3.3.

Nous remarquons que la première variable canonique U_1^* explique 12,48 % de la variance des $\mathbf{Y}$ ce qui n'est pas un gain très appréciable par rapport à U_1 de la corrélation canonique qui expliquait 11,85 % de la variance des $\mathbf{Y}$ (tableau 3.1). Le reste des variances expliquées sont marginales mais nous pouvons remarquer que les premiers indices RU_k^{*2} sont supérieurs (maximaux) à ceux trouvés par les corrélations canoniques RU_k^2 au tableau 3.1.

Régression multiple et ACP

Il n'existe pas non plus de procédure directe en SAS pour appliquer l'analyse canonique de redondance selon cette façon de procéder. Il est cependant assez simple de la programmer. Nous avons tout d'abord effectué une régression multiple sur chacune des Y_j (avec *PROC REG*) et appliqué une analyse en composantes principales (ACP) (avec *PROC PRINCOMP*) sur les valeurs prédites de cette régression. Les valeurs propres de l'ACP sur les valeurs prédites sont disponibles dans le *Output SAS* de la procédure (voir tableau 3.4).

Nous remarquons que les valeurs propres de la matrice $\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$ (voir les détails de la méthode à la section 3.3.1) qui se retrouvent dans la colonne `Eigenvalue` du tableau 3.4 correspondent **exactement** aux valeurs propres λ'_{a_k} du tableau 3.3. Ce sont les valeurs propres communes de cette matrice et celles de la matrice $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$ (méthode van den Wollenberg) comme nous l'avons démontré à la section 3.3.1. Les redondances $RU_k^{*2} = \lambda'_{a_k} \,/\, q$ sont alors les mêmes et les deux méthodes sont équivalentes.

Tableau 3.3 Analyse canonique de redondance, méthode van den Wollenberg

	Valeurs propres de $\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}\mathbf{R}_{YX}$	**Indice de redondance** (RU_k^{*2})	**Indice de redondance total** $\sum(RU_k^{*2})$
k	λ'_{a_k}	λ'_{a_k} / q	$\sum(\lambda'_{a_k}$ / $q)$
1	1,9971233	0,1248202	0,1248202
2	0,2913521	0,0182095	0,1430297
3	0,2155257	0,0134704	0,1565001
4	0,1597213	0,0099826	0,1664827
5	0,0813366	0,0050835	0,1715662
6	0,0426572	0,0026661	0,1742323
7	0,0366093	0,0022881	0,1765203
8	0,0168745	0,0010547	0,1775750
9	0,0087264	0,0005454	0,1781204
10	0,0057016	0,0003563	0,1784767
11	0,0025716	0,0001607	0,1786375
12	0,0018683	0,0001168	0,1787542
13	0,0015905	0,0000994	0,1788536
14	0,0009984	0,0000624	0,1789160
15	0,0000854	0,0000053	0,1789214
16	0,0000128	0,0000008	0,1789222

Tableau 3.4 ACP sur les valeurs prédites, Output SAS

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.99712327	1.70577115	0.6976	0.6976
2	0.29135211	0.07582641	0.1018	0.7994
3	0.21552571	0.05580439	0.0753	0.8747
4	0.15972132	0.07838471	0.0558	0.9305
5	0.08133661	0.03867942	0.0284	0.9589
6	0.04265719	0.00604792	0.0149	0.9738
7	0.03660927	0.01973476	0.0128	0.9866
8	0.01687451	0.00814814	0.0059	0.9925
9	0.00872637	0.00302482	0.0030	0.9955
10	0.00570155	0.00313000	0.0020	0.9975
11	0.00257155	0.00070324	0.0009	0.9984
12	0.00186831	0.00027785	0.0007	0.9991
13	0.00159047	0.00059204	0.0006	0.9996
14	0.00099843	0.00091300	0.0003	1.0000
15	0.00008543	0.00007266	0.0000	1.0000
16	0.00001277		0.0000	1.0000

3.7.3 Sélection de variables et prédiction

Il y a beaucoup de variables dans le groupe des **Y**. Le groupe des **X** ne semble pas expliquer une très grande proportion de variance des Y_j. Il est intéressant de faire quelques sous-groupes dans les **Y** et d'observer dans quelle mesure nous pouvons maintenant prédire ces **Y**.

Par exemple, nous avons créé une nouvelle variable de revenu de travail :

Y00 (revenu travail) = Y01 (salaire) + Y02 (travail autonome) + Y03 (autres revenus)

Le graphique biplot du chapitre précédent nous a fait remarqué que dans la matrice de corrélation, les variables Y08 (revenu de placement), Y13 (crédits relatifs aux enfants) et Y14 (frais médicaux) n'ont pratiquement pas de relations avec les X_i.

Nous avons refait l'analyse canonique de redondance avec la nouvelle variable Y00 et éliminé les variables Y08, Y14 et Y13. Le tableau 3.5 fournit les résultats de cette analyse canonique de redondance avec une régression multiple et une ACP sur les nouvelles valeurs prédites.

La redondance totale de ce modèle est de 24,52 % ($\sum$ `Eigenvalue` / 11). La première redondance est un peu plus intéressante :

$$\begin{aligned} RU_1^{*^2} &= 2.00514325/\ 11 \\ &= 0.18228575 \end{aligned}$$

Le modèle de l'analyse canonique de redondance commence à devenir est plus pertinent lorsque l'indice de redondance tend à être plus élevé. Pour nos deux groupes de variables à l'étude dans notre ensemble de données, elle semble être moins appropriée.

Il serait intéressant de poursuivre la recherche de meilleures prédictions de la variance du groupe des **Y** en regroupant ou supprimant d'autres variables dans le groupe des **Y** ou des **X**. Nous pourrions également appliquer certaines transformations comme nous l'avons mentionné à la section 2.3.8 pour tenter d'améliorer la linéarité de certaines

Tableau 3.5 ACP sur les valeurs prédites, sous-groupes de $\mathbf{Y}$ ($q = 11$)

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.00514325	1.79949557	0.7434	0.7434
2	0.20564768	0.00173134	0.0762	0.8197
3	0.20391634	0.05185993	0.0756	0.8953
4	0.15205641	0.08887858	0.0564	0.9517
5	0.06317783	0.03191127	0.0234	0.9751
6	0.03126655	0.01237694	0.0116	0.9867
7	0.01888961	0.00830362	0.0070	0.9937
8	0.01058599	0.00589711	0.0039	0.9976
9	0.00468888	0.00309136	0.0017	0.9994
10	0.00159751	0.00148232	0.0006	1.0000
11	0.00011519		0.0000	1.0000

variables afin d'obtenir une meilleure prédiction. L'étude serait intéressante à poursuivre.

Lorsque nous sommes en mesure de prédire un pourcentage appréciable de la variance des $\mathbf{Y}$, c'est à ce moment que la visualisation du modèle avec le graphique biplot devient pertinente.

CHAPITRE IV

SYNTHÈSE DE L'ANALYSE CANONIQUE

Pour ce mémoire, il a été difficile de trouver des références qui résument et présentent **bien** l'analyse canonique en général et tous les liens qui en découlent. Les terminologies et les présentations sont souvent très différentes d'un auteur à l'autre. L'analyse canonique est souvent négligée dans les manuels ou elle est présentée en dernier lieu comme méthode alternative et difficile à appliquer (interprétation des résultats complexes). Pourtant, presque tous les auteurs s'entendent pour dire que l'analyse canonique contient un contexte théorique général de plusieurs méthodes d'analyse de données.

Nous venons d'étudier les deux principales méthodes en analyse canonique, soit l'analyse des corrélations canoniques et l'analyse canonique de redondance. Nous savons cependant que l'analyse canonique est beaucoup plus vaste que ces deux méthodes. Il reste beaucoup de matière à présenter et à exploiter dans ce domaine. Ainsi, nous croyons qu'il est important de présenter une synthèse de l'analyse canonique qui se veut un résumé de nos recherches et lectures qui nous ont permis d'éclaircir ce domaine souvent négligé. En présentant cette synthèse, cela offrira une perspective des autres méthodes d'analyse de données face à l'analyse canonique qui donne souvent le contexte théorique général. Le lecteur obtiendra ainsi une vision de plusieurs méthodes d'analyse de données qui facilitera la compréhension des traitements statistiques à faire en fonction d'un type de données et du but recherché.

Nous allons faire un survol rapide du thème de l'analyse canonique en présentant

le principe de base, les différentes méthodes et les cas particuliers qui sont des techniques bien connues. Nous n'allons pas présenter toute la théorie car ce n'est pas notre but. Le livre de Legendre et Legendre (1998, chapitre 11) est une très bonne référence pour présenter de façon générale l'analyse canonique. Un lecteur doit consulter d'autres références s'il veut avoir plus de détails théoriques (il s'adresse à des biologistes) mais ce livre nous a beaucoup aidé à obtenir une vision générale de l'analyse canonique.

4.1 Introduction

Le terme canonique en mathématique veut dire simplicité, régularité, structure fondamentale et de base. C'est la forme réduite la plus simple et la plus compréhensive de relations ou de fonctions. Par exemple, la forme canonique d'une matrice de covariance est la matrice des vecteurs propres. L'analyse canonique s'applique donc très bien à un ensemble de données complexes où les relations sont souvent entrecroisées pour nous ramener à quelques relations simples comme la matrice de variances-covariances des (U_i, V_j) (voir 1.2.1) en corrélation canonique.

L'analyse canonique s'intéresse à l'étude des relations. Les relations sont contenues implicitement dans les matrices de variances-covariances des groupes de variables. Le principe général en analyse canonique est d'analyser les vecteurs propres des différentes structures de matrices de variances et de covariances. Les relations sont donc résumées dans les diverses variables canoniques. Ces variables sont souvent définies par des vecteurs propres comme combinaisons linéaires des variables originales.

Dans l'ancienne littérature, l'analyse canonique est fréquemment associée à l'analyse des corrélations canoniques. Cette méthode est en fait le point central de l'analyse canonique. C'est à partir de cette méthode que d'autres méthodes d'analyse canonique ont été développées au cours des années. C'est pour cette raison qu'il faut parler plutôt d'analyse canonique avant de parler d'analyse des corrélations canoniques bien que certains auteurs n'apportent pas nécessairement de différence.

Nous présenterons succinctement les principales méthodes utilisées en analyse ca-

nonique. Ceci représente un résumé de nos lectures qui nous ont permis de mieux comprendre tout le contexte de l'analyse canonique.

4.2 Analyse des corrélations canoniques

Auteur et année : Hotelling, 1935-1936

Synonymes : Analyse canonique

Type de données : Deux groupes de variables quantitatives mesurées sur les mêmes individus.

But : Résumer les relations entre les groupes de variables

Principaux domaines d'application : L'écologie et la biologie

Description : La théorie repose essentiellement sur la recherche de combinaisons linéaires (variables canoniques) de variables d'un premier groupe et de combinaisons linéaires de variables d'un autre groupe de sorte que la **corrélation** entre ces combinaisons linéaires soit **maximale**. Les résultats de cette procédure doivent être bien analysés pour obtenir des conclusions sur les relations entre les ensembles de données.

Cas particuliers : Régression linéaire simple et multiple

Les corrélations canoniques ont comme cas particuliers "directs" la régression simple et la régression multiple. En effet, lorsque le premier groupe de variables contient une seule variable (expliquée) et le deuxième groupe est soit :

- une seule variable explicative, nous avons la régression linéaire simple,
- plusieurs variables explicatives, nous avons la régression linéaire multiple.

Il existe alors une seule combinaison linéaire entre les variables dont la **corrélation** est **maximale**. Les variables sont également quantitatives et le but n'est pas nécessairement de résumer les relations entre les groupes (il n'en existe qu'une seule). Nous pouvons voir clairement la relation de dépendance entre les deux groupes de variables (explicative (s) et expliquée) ce qui n'est pas nécessairement le cas en corrélation canonique.

Il est possible d'utiliser le principe des corrélations canoniques avec des variables qualitatives. Cela nous amène à la prochaine méthode où l'analyse des corrélations canoniques est le point de départ. Certains auteurs la présentent même comme un cas particulier de l'analyse des corrélations canoniques.

4.2.1 Analyse canonique discriminante

Auteur et année : Fisher, Mahalanobis, 1936, Rao, 1948, 1952

Synonymes : Analyse discriminante, Analyse des variables canoniques (canonical variate analysis)

Type de données : Un groupe de variables quantitatives et un second groupe de variables représentant **un caractère qualitatif** répartit en g classes (g variables 0/1) mesurées sur les mêmes individus.

But : Séparer et classer des individus

Principaux domaines d'application : Médical et l'épidémiologie

Description : Recherche de combinaisons linéaires (variables canoniques ou fonctions discriminantes) de variables quantitatives du premier groupe qui séparent le mieux les g classes (deuxième groupe de variables). Ces combinaisons linéaires sont construites selon le même principe que l'analyse des corrélations canoniques. Les résultats permettront d'attribuer une des g classes à un nouvel individu qui possède les caractéristiques du premier groupe de variables.

Cas particuliers : Analyse des correspondances (Benzécri, 1965)

Comme Lebart, Morineau et Fénélon (1982) le mentionnent, lorsque les deux groupes de variables représentent un caractère qualitatif répartit en h et g classes respectivement, nous avons alors une double analyse discriminante qui correspond à l'analyse des correspondances.

Les deux prochaines méthodes d'analyse canonique ne maximisent pas nécessairement la corrélation. Elles respectent cependant l'idée générale de l'analyse canonique

en se basant sur les vecteurs propres d'une matrice de variance-covariance légèrement différente.

4.3 Analyse de la redondance canonique

Auteur et année : Rao, 1964, van den Wollenberg, 1977

Synonymes : Régression à rang réduit, Analyse en composantes principales avec variables instrumentales

Type de données : Deux groupes de variables quantitatives mesurées sur les mêmes individus.

But : Trouver des modèles de prédiction entre les groupes de variables

Principaux domaines d'application : L'écologie et la biologie

Description : Recherche de combinaisons linéaires (variables canoniques) des variables du premier groupe de sorte que la variance expliquée (redondance) de l'autre groupe de variables soit **maximale**. Les résultats doivent être bien analysés pour bien comprendre les variables des deux groupes qui sont les plus importantes dans le modèle de prédiction. Nous supposons au préalable un lien de dépendance entre les données.

Cas particulier : Analyse en composantes principales

Comme van den Wollenberg le mentionne dans son article (voir 3.3.2 chapitre précédent), lorsque le premier groupe de variables est identique au second, cela revient à rechercher des combinaisons linéaires de ce groupe de sorte que la variance est **maximale**. Cela correspond donc à la définition de l'analyse en composantes principales.

La prochaine méthode peut être vue également comme un cas spécial de l'analyse de la redondance canonique pour des variables discrètes et qualitatives. Elle est aussi présentée comme une extension de l'analyse des correspondances. Elle a été développée spécifiquement pour le domaine de l'écologie.

4.3.1 Analyse des correspondances canoniques

Auteur et année : ter Braak, 1986

Synonymes : Forme canonique de l'analyse des correspondances, cas particulier de l'analyse de la redondance canonique avec des variables qualitatives

Type de données : Deux groupes de variables qualitatives (fréquence des espèces et les milieux) et un groupe de variables quantitatives (l'environnement).

But : Étude des relations entre les espèces dans des milieux différents et les variables environnementales

Principal domaine d'application : Écologie

Description : Cette méthode a été développé par un biométricien dans un contexte précis de l'écologie. La théorie repose sur la recherche de combinaisons linéaires des variables environnementales disponibles dans les milieux qui expliquent le mieux la distribution des espèces (maximise la variance moyenne des espèces).

4.4 Analyse canonique généralisée

Auteur et année : Caroll, 1968, Kettenring, 1971

Synonymes : Analyse canonique à plusieurs groupes

Type de données : Trois groupes ou plus de variables quantitatives mesurées sur les mêmes individus.

But : Étude des relations linéaires entre les groupes de variables

Principaux domaines d'application : Très peu utilisée, modèle théorique intéressant

Description : La théorie repose essentiellement sur la recherche de combinaisons linéaires (variables canoniques) de sorte que certaines **mesures généralisées de corrélation** (il en existe plusieurs) entre ces combinaisons linéaires soient **maximales**. Ces corrélations sont toujours fonctions de valeurs propres. Les variables canoniques seront également toujours fonctions des vecteurs propres.

4.5 Analyse des corrélations canoniques non linéaires

Auteur et année : Gifi 1990, Shi et Tamm (1992)

Synonymes : Analyse canonique non linéaire

Type de données : Deux groupes variables quantitatives mesurées sur les mêmes individus.

But : Étude des relations non linéaires entre les groupes de variables

Principaux domaines d'application : Quelques applications en météorologie et en finance.

Description : La théorie repose essentiellement sur la recherche de combinaisons **non** linéaires (variables canoniques) de variables d'un premier groupe et de combinaisons **non** linéaires de variables d'un autre groupe de sorte que la corrélation entre ces combinaisons soit maximale.

4.6 Application croisée de l'analyse canonique

Nous avons lu beaucoup de travaux de recherches qui croisent les différentes méthodes d'analyse canonique avec d'autres méthodes d'analyse de données. Comme nous l'avons mentionné, l'analyse canonique est utile pour **comprendre** un ensemble de données. Lorsque l'analyse canonique est appliquée **préalablement** à d'autres méthodes d'analyses de données, ces dernières s'avèrent beaucoup plus efficace. Par exemple, l'analyse des corrélations canoniques est utilisée pour déterminer un voisinage de variables hydrologiques et ensuite une régression multiple est appliquée sur ce voisinage (Ouarda et al., 2001). Nous avons également lu des applications de l'analyse canonique croisée avec les réseaux de neurones, les contrôles statistiques des processus et les processus stochastiques. Par une plus grande maîtrise de l'analyse canonique, ces méthodes d'applications croisées se développeront davantage au cours des prochaines années.

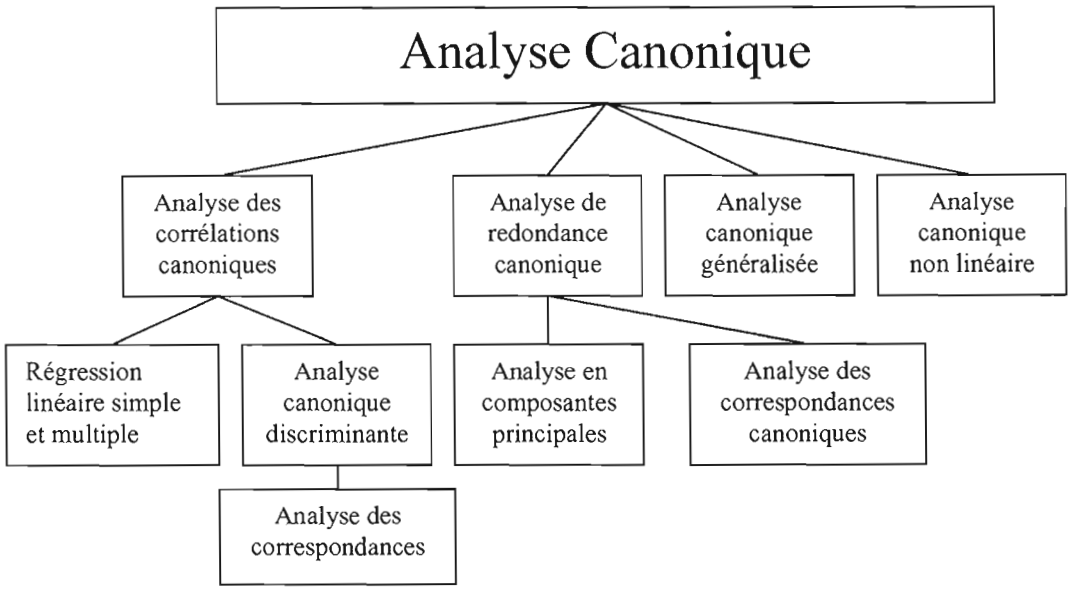


Figure 4.1 L'analyse canonique

4.7 Conclusion

L'analyse canonique est essentielle en analyse de données et devra être de plus en plus à considérer dans l'avenir dû au nombre important de variables et d'observations qui sont maintenant disponibles pour les chercheurs. Étant donné l'aspect général qu'elle possède, nous croyons qu'il est important de la présenter dans les cours d'introduction à l'analyse de données. À la figure 4.1, nous présentons la synthèse des méthodes que nous venons de discuter.

CONCLUSION

Notre intérêt marqué pour la recherche de relations, de structures, de liens ou de corrélations dans des ensembles de données nous a amené à étudier dans ce mémoire l'analyse canonique et le graphique biplot qui sont deux méthodes qui se prêtent bien à ce genre de sujet.

Au premier chapitre, nous avons présenté toute la théorie de l'analyse des corrélations canoniques qui est la technique la plus importante en analyse canonique. Nous avons également présenté en détails le graphique biplot qu'il est possible d'obtenir dans le contexte de l'analyse des corrélations canoniques. Nous nous sommes appuyés surtout sur un article de Cajo J.F. ter Braak qui discute de ce biplot.

Au deuxième chapitre, nous avons appliqué la technique des corrélations canoniques et tracé le graphique biplot décrit au chapitre I. L'ensemble de données étudié provient d'une enquête de Statistique Canada sur la dynamique entre le travail et le revenu. Nous avons constaté qu'il est possible de visualiser rapidement les relations de l'ensemble de données à l'aide du graphique biplot. Ce graphique nous a permis de réorganiser la matrice des corrélations entre nos deux groupes de variables ($\mathbf{R}_{XY}$). Ainsi, nous avons pu dégager de cette matrice des structures intéressantes qui auraient été difficile à trouver sans le biplot. Nous avons également tenté de donner une interprétation des corrélations (relations) canoniques simplement avec le biplot et la matrice des corrélations restructurées.

Au troisième chapitre, nous avons étudié l'analyse canonique de redondance (régression à rang réduit). Cette méthode, assez récente, est une suite de l'analyse des corrélations canoniques. Nous avons présenté les différentes façons théoriques d'obtenir cette méthode ainsi qu'un graphique biplot du modèle. Nous avons appliqué la méthode de deux

façons sur notre ensemble de données, mais elle représentait un intérêt assez limité pour cet ensemble. Il faudrait trouver des sous-groupes de variables dans notre ensemble ou appliquer des transformations pour tenter d'obtenir des résultats plus intéressants.

Au dernier chapitre, nous avons présenté une synthèse (non théorique) des diverses méthodes se rapportant à l'analyse canonique. Ce résumé de nos recherches et lectures sur le domaine a peut-être permis aux lecteurs de mieux se situer dans toutes ces méthodes d'analyse de données. Peu d'auteurs présentent cette vision de l'analyse canonique même si plusieurs s'entendent sur le fait qu'elle possède un aspect important de généralité.

Nous connaissions déjà quelques propriétés intéressantes du graphique biplot. Dans ce mémoire, nous avons vu toute son importance dans le contexte des corrélations canoniques. Son application aide grandement la compréhension des corrélations canoniques et des relations dans un ensemble de données. Il serait intéressant, suite à ce mémoire, de poursuivre l'étude du graphique biplot des corrélations canoniques en y introduisant une échelle qui permettrait d'obtenir directement les valeurs approximatives des corrélations (selon les idées de Gower). Il serait également intéressant d'approfondir l'analyse canonique de redondance avec une application plus concrète et développer son graphique biplot que nous avons présenté sommairement.

APPENDICE A

PROGRAMME SAS DES CORRÉLATIONS CANONIQUES ET DU GRAPHIQUE BIPLOT

```
*Calcul des corrélations canoniques;
PROC CANCORR data=z.donnees_finales /*Notre fichier de données*/
ALL /*Toutes les statistiques disponibles de CANCORR*/ Vprefix=U
Wprefix=V /*Nom des variables canoniques*/ out=out_cancor /*Scores
des variables canoniques dans un fichier*/ outstat=out_stat_cancor
/*Statistiques dans un fichier*/;

WITH /*DEUXIÈME GROUPE Y*/ Y01_Salaire Y02_Trav_auto Y03_Autre_rev
Y04_Agees Y05_Prestation Y06_Transfert Y07_Enfant Y08_Placement
Y09_Imp_fed Y10_Imp_prov Y11_Cred_tps Y12_Cred_prov Y13_C_enfant
Y14_Frais_med Y15_Cotisation Y16_Cotis_oblig ;

VAR /*PREMIER GROUPE X*/ X01_Age X02_Couple X03_Nb_menage
X04_Nb_etude X05_Nb_etude_ps X06_Nb_etude_uni X07_Rurale X08_Loue
X09_Nb_emp X10_Arret X11_Hres_trav X12_Hres_tmais X13_Hor_irr
X14_T_part X15_Nb_sem X16_Experience X17_Reg_ret X18_Public ;
RUN;
```

```
*Préparer les données pour faire le biplot;
data inter_intra_v1_v2(DROP=_TYPE_ );
set out_stat_cancor; *fichier de la procédure CANCORR;
*prendre seulement inter_intra corr de v1_v2;
if _TYPE_='STRUCTUR' and _NAME_ in ('V1', 'V2');
*ici nous pouvons changer V1, V2 pour U1, U2 ou V3, V4;
run;

*Transposer;
proc transpose data=inter_intra_v1_v2 out=inter_intra_v1_v2_t
(drop=_LABEL_);
run;

*Mettre l'origine pour tracer avec join;
data inter_intra_v1_v2_0;
set inter_intra_v1_v2_t;
V1=0;
V2=0;
run;
data inter_intra_v1_v2_biplot;
set inter_intra_v1_v2_t  inter_intra_v1_v2_0;
label X="V1" Y="V2" _name_="Variables originales";
*Ces paramètres sont obligatoires pour annoter dans proc gplot: ;
xsys='2';
ysys='2';
size=1;
x=V1;
y=V2;
*Mettre des Xi et Yj au lieu des noms car beaucoup de variables;
*Il faut prendre la variable TEXT absolument pour annoter dans proc gplot,
```

```
reconnu par la procédure;
text=substr(_name_,1,3);
if V1=0 then text='  ';
*pour ne pas annoter le centre utilisé pour la jointure;
run;


*Tracer le graphique;
goptions reset=all reset=symbol colors=(black) border;
PROC GPLOT DATA=inter_intra_v1_v2_biplot;
SYMBOL1  V=NONE ;
SYMBOL2  V=none ;
SYMBOL3  V=none ;
SYMBOL4  V=none ;
SYMBOL5  V=none ;
SYMBOL6  V=none ;
SYMBOL7  V=none ;
SYMBOL8  V=none ;
SYMBOL9  V=none ;
SYMBOL10 V=none ;
SYMBOL11 V=none ;
SYMBOL12 V=none ;
SYMBOL13 V=none ;
SYMBOL14 V=none ;
SYMBOL15 V=none ;
SYMBOL16 V=none ;
SYMBOL17 V=none ;
SYMBOL18 V=none ;
SYMBOL19 V=none i=join;
SYMBOL20 V=none i=join;
```

```
SYMBOL21 V=none i=join;
SYMBOL22 V=none i=join;
SYMBOL23 V=none i=join;
SYMBOL24 V=none i=join;
SYMBOL25 V=none i=join;
SYMBOL26 V=none i=join;
SYMBOL27 V=none i=join;
SYMBOL28 V=none i=join;
SYMBOL29 V=none i=join;
SYMBOL30 V=none i=join;
SYMBOL31 V=none i=join;
SYMBOL32 V=none i=join;
SYMBOL33 V=none i=join;
SYMBOL34 V=none i=join;
SYMBOL35 V=none i=join;
SYMBOL36 V=none i=join;
plot Y*X=_name_ / ANNOTATE=inter_intra_v1_v2_biplot frame
href=0 vref=0;
run;quit;
```

APPENDICE B

PROGRAMME SAS DE L'ANALYSE CANONIQUE DE REDONDANCE, MÉTHODE VAN DEN WOLLENBERG

```
*Pour STANDARDISÉES les données;
PROC STDIZE data= z.donnees_finales out=donnees_finales_std;
RUN;

*Transreg;
proc transreg  data=donnees_finales_std;
model
identity (Y01_Salaire Y02_Trav_auto Y03_Autre_rev Y04_Agees
Y05_Prestation Y06_Transfert Y07_Enfant Y08_Placement Y09_Imp_fed
Y10_Imp_prov Y11_Cred_tps Y12_Cred_prov Y13_C_enfant Y14_Frais_med
Y15_Cotisation Y16_Cotis_oblig)
=
identity (X01_Age X02_Couple X03_Nb_menage X04_Nb_etude
X05_Nb_etude_ps X06_Nb_etude_uni X07_Rurale X08_Loue X09_Nb_emp X10_Arret
X11_Hres_trav X12_Hres_tmais X13_Hor_irr X14_T_part X15_Nb_sem X16_Experience
X17_Reg_ret X18_Public)
/ method=redundancy   ;
output out=result_rda   red=sta MREDUNDANCY;
run;
```

```
*Ne conserver que les Yj et les variables canoniques (Red) pour
calculer manuellement les indices de redondance RU*k;
data yu(keep=TY01_Salaire
TY02_Trav_auto TY03_Autre_rev TY04_Agees TY05_Prestation TY06_Transfert
TY07_Enfant TY08_Placement TY09_Imp_fed TY10_Imp_prov TY11_Cred_tps
TY12_Cred_prov TY13_C_enfant TY14_Frais_med TY15_Cotisation
TY16_Cotis_oblig Red1     Red2     Red3     Red4     Red5     Red6
Red7     Red8     Red9     Red10 Red11       Red12       Red13
Red14       Red15       Red16);
set result_rda;
run;

*Calcul des corrélations entre les variables canoniques et les Yj;
PROC CORR  data=yu PEARSON OUTP=corr ;
VAR
TY01_Salaire TY02_Trav_auto TY03_Autre_rev TY04_Agees TY05_Prestation
TY06_Transfert TY07_Enfant TY08_Placement TY09_Imp_fed TY10_Imp_prov
TY11_Cred_tps TY12_Cred_prov TY13_C_enfant TY14_Frais_med
TY15_Cotisation TY16_Cotis_oblig;
WITH
Red1   Red2    Red3   Red4   Red5   Red6   Red7   Red8   Red9   Red10
 Red11      Red12      Red13      Red14      Red15      Red16;
run;

*Mettre les corrélations du fichier corr au carré, divisé par q et faire
la somme (facile en Excel). Nous avons alors les indices de redondance de
chaque variable canoniques, la somme donne la redondance totale trouvée en
 analyse des corrélations canoniques;
```

APPENDICE C

PROGRAMME SAS DE L'ANALYSE CANONIQUE DE REDONDANCE, RÉGRESSION MULTIPLE ET ACP

```
*REGRESSION MULTIPLE;
PROC REG DATA=donnees_finales_std;
 MODEL Y01_Salaire Y02_Trav_auto Y03_Autre_rev Y04_Agees Y05_Prestation
Y06_Transfert Y07_Enfant Y08_Placement  Y09_Imp_fed Y10_Imp_prov
Y11_Cred_tps Y12_Cred_prov Y13_C_enfant  Y14_Frais_med Y15_Cotisation
Y16_Cotis_oblig
=
X01_Age X02_Couple X03_Nb_menage X04_Nb_etude X05_Nb_etude_ps
X06_Nb_etude_uni X07_Rurale X08_Loue X09_Nb_emp X10_Arret X11_Hres_trav
X12_Hres_tmais X13_Hor_irr X14_T_part X15_Nb_sem X16_Experience
X17_Reg_ret X18_Public;
output out=ychapeau /*METRE DANS UN FICHIER*/ predicted=
Y01_Salaire_chap Y02_Trav_auto_chap Y03_Autre_rev_chap Y04_Agees_chap
Y05_Prestation_chap Y06_Transfert_chap Y07_Enfant_chap Y08_Placement_chap
Y09_Imp_fed_chap Y10_Imp_prov_chap Y11_Cred_tps_chap Y12_Cred_prov_chap
Y13_C_enfant_chap Y14_Frais_med_chap Y15_Cotisation_chap
Y16_Cotis_oblig_chap;
RUN;quit;
```

```
*Valeurs prédites seulement;
data ychap
(keep=Y01_Salaire_chap Y02_Trav_auto_chap Y03_Autre_rev_chap
Y04_Agees_chap Y05_Prestation_chap Y06_Transfert_chap Y07_Enfant_chap
Y08_Placement_chap Y09_Imp_fed_chap Y10_Imp_prov_chap Y11_Cred_tps_chap
Y12_Cred_prov_chap Y13_C_enfant_chap Y14_Frais_med_chap
Y15_Cotisation_chap Y16_Cotis_oblig_chap);
set ychapeau;
run;


*Prendre ychap et faire ACP;
Proc princomp data=ychap cov standard out=ychapacp;
run;
*Les valeurs propres du Output divisée par q nous donne les mêmes redondances;
```

BIBLIOGRAPHIE

Anderson, T. 1984. *An introduction to Multivariate Statistical Analysis.* Coll. « Wiley series in probability and mathematical statistics ». New York : John Wiley et Sons, 2^e édition. 675 p.

Barraud, M. 2002. « Analyse non linéaire en composantes canoniques ». Mémoire présenté comme exigence partielle de la maîtrise en mathématiques, Montréal, Université du Québec à Montréal. 80 p.

Berthier, P. et J.-M. Bouroche. 1977. *Analyse de données multidimensionnelles.* Paris : Presse Universitaires de France, 2^e édition. 270 p.

Caillez, F. et J. Pagès. 1976. *Introduction à l'analyse de données.* Paris : Société de Mathématiques Appliquées et de Sciences Humaines. 616 p.

Campbell, N. A. 1982. « Robust procedures in multivariate analysis : Robust canonical variate analysis », *Applied Statistics*, vol. 31, p. 1–8.

Casin, P. et J. C. Turlot. 1986. « Une présentation de l'analyse canonique généralisée dans l'espace des individus », *Revue de statistiques appliquées*, vol. 35, no. 3, p. 65–75.

Croux, C. et C. Dehon. 2002. « Analyse canonique basée sur des estimateurs robustes de la matrice de covariance », *Revue de la Statistique Appliquée*, vol. 2, p. 5–26.

Dillon, W. R. et M. Goldstein. 1984. *Multivariate Analysis, Methods and Applications.* Coll. « Wiley series in probability and mathematical statistics ». New York : John Wiley et Sons. 587 p.

Eckart, C. et G. Young. 1936. « The approximation of one matrix by another of lower rank », *Psychometrika*, vol. 1, no. 3, p. 211–218.

Fortier, J. 1966. « Simultaneoous linear prediction », *Psychometrika*, vol. 31, p. 369–381.

Gabriel, K. R. 1971. « The biplot graphic display with application to principal component analysis », *Biometrika*, vol. 58, no. 3, p. 453–467.

Gifi, A. 1990. *Nonlinear multivariate analysis.* Chichester : Wiley. 579 p.

Gittins, R. 1985. *Canonical Analysis : a review with applications in ecology.* Coll. « Biomathematics », no 12. Berlin : Springer-Verlag. 351 p.

Gleason, T. C. 1976. « On redundancy in canonical analysis », *Psychological Bulletin*, vol. 83, no. 6, p. 1004–1006.

Gower, J. C. et D. J. Hand. 1996. *Biplots*. Coll. « Monographs on statistics and applied probability », no 54. London : Chapman and Hall. 277 p.

Graffelman, J. 2001. « Quality statistics in canonical correspondence analysis », *Environmetrics*, vol. 12, p. 485–497.

Haber, M. et K. Gabriel. 1976. Weighted least squares approximation of matrices and its application to canonical correlations and biplot display. Technical report, University of Rochester, Rochester. Department of Statistics.

Harville, D. A. 1997. *Matrix Algebra from a Statistician's Perspective*. New York : Springer. 630 p.

Imhof, P. 1961. « Computing the distribution of quadratic forms in normal variates », *Biometrika*, vol. 48, p. 419–426.

Johnson, R. A. et D. W. Wichern. 1998. *Applied Multivariate Statistical Analysis*. New Jersey : Prentice Hall, 4^e édition. 816 p.

Karnel, G. 1991. « Robust canonical correlation and correspondence analysis », *The Frontiers of Statistical Scientific and Industrial Applications*, vol. 2, p. 335–354.

Kettenring, J. R. 1971. « Canonical analysis of several sets of variables », *Biometrika*, vol. 58, no. 3, p. 433–450.

Koerts, J. et A. Abrahamse. 1969. *On the theory and application of the general linear model*. Rotterdam : Rotterdam university press.

Krzanowski, W. J. 1995. *Recent Adavances in Descriptive Multivariate Analysis*. Coll. « Royal Statistical Society Lectures Notes Series », no 2. Oxford : Oxford University Press. 444 p.

———. 2000. *Principles of Multivariate Analysis : a user's perspective*. Coll. « Oxford statistical science series », no 22. New York : Oxford University Press, 2^e édition. 586 p.

Kshirsagar, A. 1972. *Multivariate Analysis*. New York : Dekker. 534 p.

Lawley, D. 1959. « Test of significance in canonical analysis », *Biometrika*, vol. 46, p. 59–66.

Lazraq, A. et R. Cléroux. 2002. « Testing the significance of the successive components in redundancy analysis », *Psychometrika*, vol. 67, p. 411–419.

Lebart, L., A. Morineau, et J.-P. Fénélon. 1982. *Traitement des données statistiques*. Paris : Dunod, 2^e édition. 510 p.

Legendre, P. et L. Legendre. 1998. *Numerical Ecology*. Coll. « Developments in environmental modelling », no 20. Amsterdam : Elsevier, 2^{e} édition. 853 p.

Lindeman, R. H., P. F. Merenda, et R. Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis*. Glenview : Scott, Foresman and Compagny. 444 p.

Muller, K. E. 1981. « Relationshiphs between redundancy analysis, canonical correlation and multivaraite regression », *Psychometrika*, vol. 46, no. 2, p. 139–142.

Nzobounsana, V. et T. Dhorne. 2003. « Ecart : Une nouvelle méthode d'analyse canonique généralisée (acg) », *Revue de la Statistique Appliquée*, vol. LI, no. 4, p. 57–82.

Ouarda, T. B. J., C. Girard, G. S. Cavadias, et B. Bobée. 2001. « Regional flood frequency estimation with canonical correlation analysis », *Journal of Hydrology*, vol. 254, p. 157–173.

Palm, R. 2003. La corrélation canonique : Principes et application. Note de statistiques et informatique, Faculté Universitaire des Sciences Agronomiques de Gembloux, Belgique. réédition.

Pemajayantha, V. 2002. « Special canonical models for multidimensional data analysis with applications and implications ». p. 24, University of Singapore. International conference on stochastics and applications.

Pilotte, A. 2005. « Utilisation du clustering pour trouver et décrire des profils de dépenses de ménages québécois et validation avec le biplot ». Mémoire présenté comme exigence partielle de la maîtrise en mathématiques, Montréal, Université du Québec à Montréal. 194 p.

Reinsel, G. C. et R. P. Velu. 1998. *Multivariate Reduced-Rank Regression*. Coll. « Lecture Notes in Statistics », no 136. New York : Springer. 258 p.

Rencher, A. C. 1995. *Methods of Multivariate Analysis*. Coll. « Wiley series in probability and mathematical statistics ». New York : John Wiley et Sons. 627 p.

———. 1998. *Multivariate Statistical Inference and Applications*. Coll. « Wiley series in probability and statistics ». New York : John Wiley et Sons. 559 p.

Shi, S. et W. Tamm. 1992. « Non-linear canonical correlation analysis with a simulated annealing solution », *Journal of Applied Statistics*, vol. 19, p. 155–165.

Stewart, D. et W. Love. 1968. « A general canonical correlation index », *Psychological*, vol. 70, no. 3, p. 160–163.

ter Braak, C. J. F. 1986. « Canonical correspondence analysis : A new eigenvector technique for multivariate direct gradient analysis », *Ecology*, vol. 67, p. 1167–1179.

———. 1990. « Interpreting canonical correlation analysis through biplots of structure

correlations and weights », *Psychometrika*, vol. 55, no. 3, p. 519–531.

ter Braak, C. J. F. et C. W. N. Looman. 1994. « Biplots in reduced-rank regression », *Biometrical Journal*, vol. 36, no. 8, p. 983–1003.

Thompson, B. 1984. *Canonical correlation analysis : Uses and interpretation.* Coll. « Quantitative applications in the social sciences », no 47. Beverly Hills : Sage University Papers. 71 p.

Tso, M.-S. 1981. « Reduced-rank regression and canonical analysis », *Journal of Royal Statistics Society B*, vol. 43, no. 2, p. 183–189.

van den Wollenberg, A. L. 1977. « Redundancy analysis, an alternative for canonical correlation analysis », *Psychometrika*, vol. 42, p. 207–219.