UNIVERSIDADE ABERTA

ANÁLISE DE CORRELAÇÃO CANÓNICA:

EXTENSÕES e APLICAÇÕES

Idálio Esperança Luís

Mestrado em Estatística Matemática e Computação

Área de Especialização em Estatística Computacional

Lisboa, 2015

UNIVERSIDADE ABERTA

ANÁLISE DE CORRELAÇÃO CANÓNICA:

EXTENSÕES e APLICAÇÕES

Idálio Esperança Luís

Mestrado em Estatística Matemática e Computação

Área de Especialização em Estatística Computacional

Dissertação orientada por:
Professora Doutora Maria do Rosário Ramos

Lisboa, 2015

À minha esposa
e filhas

AGRADECIMENTOS

À Professora Doutora Maria do Rosário Ramos, minha orientadora de dissertação e docente de Estatística, a quem agradeço, em primeiro lugar, o seu interesse em orientar o tema, o desejo de continuar a pesquisa e aprofundamento dos meus conhecimentos. Agradeço sua inspiração, a confiança, e contribuição na transmissão dos seus conhecimentos para a elaboração desta dissertação.

Ao Doutor Amílcar Manuel do Rosário Oliveira, Professor Auxiliar do Departamento de Ciências e Tecnologia da Universidade Aberta, pela motivação direcionada no sentido de continuar os estudos mesmo no momento difícil e terminar este trabalho.

A minha esposa, e filhas, pelo apoio incondicional, pela sua compreensão e o apoio nos finais de semana e as noites, ao meu lado enquanto ia escrevendo.

Aos técnicos do Instituto Nacional de Estatística de São Tomé e Príncipe, pela colaboração em disponibilizar as informações que serviram de apoio a este trabalho.

O meu agradecimento a todos os colegas que juntos tornaram possível a realização deste trabalho.

RESUMO

Neste trabalho apresentamos a teoria da análise de correlação canónica, uma técnica de análise estatística multivariada para o estudo da relação, simultânea, entre dois, três ou mais grupos de variáveis.

Descrevemos a natureza da correlação canónica com três ou mais variáveis, com modelos matemáticos, fazendo uma síntese dos métodos de generalização de correlação canónica nomeadamente o método Ssqcor, método Sumcor, método Ecart, método Maxvar, método Minvar, e o método de Carroll.

Apresentamos uma aplicação utilizando dados provenientes do cálculo do Índice de Preços no Consumidor IPC, produzido pelo INE - STP (Instituto Nacional de Estatística de São Tomé e Príncipe), referente ao período 2010 a 2014.
Estamos interessados em conhecer as correlações canónicas entre grupos de variáveis relacionadas com o cabaz de produtos pré-estabelecido para o cálculo do índice de preços no consumidor, concretamente os produtos alimentares (PA), produtos para bebidas (PB) e produtos não alimentares (PNA), constituindo assim os três grandes grupos de variáveis da nossa pesquisa.

ABSTRACT

In this master thesis we present the theory of canonical correlation analysis, a multivariate statistical technique to study the relation between two, three or more groups of variables.

We describe the nature of canonical correlation with three or more variables with mathematical models by making the syntax of canonical correlation generalization methods including Ssqcor method Sumcor method Ecart method Maxvar method Minvar method and Carroll method.

We present an application of Canonical Correlation Analysis, using data from the Consumer Price Index CPI, produced by INE - STP (National Institute of Statistics of São Tome and Principe), for the period of 2010-2014.

We are interested in knowing and interpret the canonical correlations between groups of variables related to food (PA), products for drinks (PB) and non-food products (PNA), thus constituting three groups of variables of our research.

RESUME

Dans cette dissertation vous est présentée la théorie de l'analyse de corrélation canonique, une technique d'analyse statistique multivariée pour deux, trois ou plusieurs groupes de variables.

Nous décrivons la nature de la corrélation canonique avec trois ou plusieurs variables avec des modèles mathématiques, faisant une syntaxe des méthodes de généralisation de corrélation canonique y compris la méthode Ssqcor, méthode Sumcor, méthode Ecart, méthode Maxvar, méthode Minvar et la méthode Carroll.

Nous présentons des exemples pratiques, en utilisant les données du calcul IPC (Indice des prix à la consommation), produite par l'INE - STP (Institut national de la statistique de Sao Tomé-et-Principe), pour l'année de 2010 à 2014.
Nous sommes intéressés à connaître les corrélations canoniques entre les groupes de variables liées à l'alimentation (PA), les produits pour les boissons (PB) et les produits non alimentaires (PNA), constituant ainsi trois groupes de variables de notre recherche.

LISTA DAS TABELAS

LISTA DE ABREVIATURAS

INE-STP: Instituto Nacional de Estatística de São Tome e Príncipe.

IPC: Índice do preço no consumidor

PB: Produtos destinados para Bebida

PA: Produtos Alimentares

PNA: Produtos Não Alimentares

Índice

INTRODUÇÃO

A análise de correlação canónica é um método estatístico proposto em 1936 por Hotelling, conhecido sobretudo pelas suas qualidades teóricas, pois ele engloba outros métodos de estatísticas multivariada. Atualmente, graças ao avanço da tecnologia e das capacidades dos computadores, é possível tratar bases de dados provenientes de pesquisas que são compostas por grandes quantidades de variáveis e de observações. Estas informações são de estrema importância, mas, frequentemente, é difícil conhecer as relações entre as variáveis. A necessidade de conhecer estas associações ou correlações entre variáveis de um conjunto de dados ou entre conjuntos de variáveis pode ser do interesse do pesquisador. A técnica de análise canónica é uma técnica de análise multivariada, utilizada para resumir adequadamente estas informações. O método permite descrever as relações que existem entre dois grupos de variáveis, maximizando a correlação entre os vetores de variáveis que são consideradas de alguma forma dependentes e independentes.

Nesta dissertação será aprofundado o estudo da Análise de Correlação Canónica e investigados os desenvolvimentos e aplicações, nomeadamente a extensão da Análise Canónica a mais de dois grupos, a sua relação com outras técnicas, a sua implementação no software estatístico SPSS e apresentada uma aplicação a dados reais de variáveis económicas de São Tomé e Príncipe.

Vários métodos têm sido propostos para generalizar a análise de correlação canónica. Estes métodos são agrupados sob o nome de análise canónica generalizada, quando eles dão, no caso de dois grupos de variáveis, Hotelling de 1936.

Em geral, nós olhamos para combinações lineares das variáveis em cada grupo e chamamos de variáveis canónicas que otimizam funções internas da sua matriz de correlação ou de covariância.

No primeiro capítulo descrevemos a natureza de correlação canónica de duas variáveis, com modelos matemáticos, e uma descrição das referências de suporte.

No segundo capítulo descrevemos a natureza de correlação canónica com três ou mais variáveis, com modelos matemáticos, fazendo uma síntese dos métodos de generalização de correlação canónica nomeadamente o método Ssqcor, método Sumcor, método Ecart, método Maxvar, método Minvar, e o método de Carroll.

No terceiro capítulo será apresentada a relação entre a análise de correlação canónica e outras técnicas multivariadas clássicas.

No quarto capítulo será apresentado uma aplicação para ilustrar o tema para dois ou mais conjuntos de variáveis utilizando dados económicos do Instituto Nacional de Estatística de São Tomé e Príncipe.

No quinto capítulo serão apresentadas as conclusões do trabalho desenvolvido.

CAPITULO I

ANÁLISE DA CORRELAÇÃO CANÓNICA PARA DOIS GRUPOS DE VARIÁVEIS

I.1. Introdução

A análise de correlação canónica é um método estatístico ou análise canónica simples, proposto em 1936 por Hotelling, conhecido sobretudo pelas suas qualidades teóricas, pois ele engloba os outros métodos estatísticos.

A técnica de análise canónica é uma técnica de análise multivariada, utilizada para resumir adequadamente estas informações. O método permite descrever as relações que existem entre dois grupos ou conjuntos de variáveis, maximizando a correlação entre os vetores de variáveis dependentes e independentes.

Neste capítulo descrevemos a natureza de correlação canónica de duas variáveis, descrevendo: Formulação de análise canónica; Variáveis canónicas e Correlação canónica; Variáveis canónicas amostrais e Correlação canónica amostral; Validação da análise de correlação canónica; Algoritmo baseado por formulação algébrica.

I.2. Formulação da Analise Canónica

A análise canónica pode ser formulada de diversas maneiras. A formulação clássica, de natureza algébrica (Hotelling, 1936 ; Anderson, 1958), é um problema de otimização sobre as restrições. Uma outra formulação pode ser feita em termos de ângulos entre subespaços de um espaço Euclidiano (Dempster, 1969 ; Björch et Golub, 1973).

Existem ainda outras formulações baseados sobre o princípio dos mínimos quadrados, James (1979), van der Burg et de Leeuw (1983) et ter Braak (1990).

Neste trabalho serão consideradas as formulações em problemas de otimização

I.3. A Variável canónica e correlação canónica

O objetivo da correlação canónica é determinar uma combinação linear para cada grupo de variáveis (dependentes e independentes) que maximize a correlação entre os dois grupos.

O primeiro grupo de p variáveis é representado por $(p\times 1)$ vetor aleatório $X^{(1)}$.O segundo grupo de variáveis é representado por $(q\times 1)$ vetor aleatório $X^{(2)}$.

Seja X um vetor de dimensão $(p+q\times 1)$, o qual possui matriz de co - variância $\sum$ e a matriz da média μ. Sejam os vetores $X^{(1)}(p\times 1)$ e $X^{(2)}(q\times 1)$ definidos como sendo originados de uma partição do vetor original X, representando um grupo com p variáveis e outro com q, respetivamente. Sem perda de generalidade é assumido que $p\le q$. Pressupõe-se, também, que $\sum$ possui elementos finitos e é definida positiva. Para o vetor aleatório X, os seguintes resultados são apresentados.

$$\underset{(p+q)\times 1}{X}=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}=\begin{bmatrix}X_1^{(1)}\\X_2^{(1)}\\\vdots\\X_p^{(1)}\\X_1^{(2)}\\X_2^{(2)}\\\vdots\\X_q^{(2)}\end{bmatrix}$$

Para o vetor $X^{(1)}$ e $X^{(2)}$ tem-se

$$E\left(X^{(1)}\right)=\mu^{(1)}\ Cov\left(X^{(1)}\right)=\sum_{X^{(1)}}$$

$$E\left(X^{(2)}\right)=\mu^{(2)}\ Cov\left(X^{(2)}\right)=\sum_{X^{(2)}}$$

$$Cov\left(X^{(1)},X^{(2)}\right)=\sum_{X^{(1)}X^{(2)}}=\sum_{X^{(2)}X^{(1)}}^{t}$$

As médias dos vetores estão definidas como

$$\underset{(p+q)\times 1}{\mu}=E\left(X\right)=E\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}=\begin{bmatrix}\mu^{(1)}\\\mu^{(2)}\end{bmatrix}$$

E a matriz de covariância de dimensão $p+q$ representada por $\sum$

$$\sum = E\left(X-\mu\right)\left(X-\mu\right)^{'} =$$
$$= \begin{bmatrix} E\left(X^{(1)}-\mu^{(1)}\right)\left(X^{(1)}-\mu^{(1)}\right)^{'} & E\left(X^{(1)}-\mu^{(1)}\right)\left(X^{(2)}-\mu^{(2)}\right)^{'} \\ E\left(X^{(2)}-\mu^{(2)}\right)\left(X^{(1)}-\mu^{(1)}\right)^{'} & E\left(X^{(1)}-\mu^{(2)}\right)\left(X^{(2)}-\mu^{(2)}\right)^{'} \end{bmatrix}$$

$$\underset{(p+q)\times(p+q)}{\sum} = E\left(X-\mu\right)\left(X-\mu\right)^{'} = \begin{pmatrix} \Sigma_{X^{(1)}} & \Sigma_{X^{(1)}X^{(2)}} \\ \Sigma_{X^{(2)}X^{(1)}} & \Sigma_{X^{(2)}} \end{pmatrix}$$

As covariâncias entre pares de variáveis pertencentes aos dois grupos, uma de $X^{(1)}$ e outra de $X^{(2)}$, estão contidas em $\Sigma_{X^{(1)}X^{(2)}}$. Dessa forma, os pq elementos de $\Sigma_{X^{(1)}X^{(2)}}$ medem a associação entre os dois grupos. Se ambos os valores de p e q forem grandes, a interpretação simultânea desse conjunto de covariâncias é uma tarefa difícil e na maioria das vezes infrutífera. Como a finalidade, é de realizar comparação, o interesse pode ser focado em combinações lineares das variáveis originais. A ideia é, portanto, concentrar a atenção em algumas poucas combinações lineares de variáveis pertencentes a $X^{(1)}$ e a $X^{(2)}$, ao invés de utilizar todas as pq covariâncias contidas em $\Sigma_{X^{(1)}X^{(2)}}$.

As covariâncias entre variáveis de diferentes conjuntos, uma variável de $X^{(1)}$ e outra de $X^{(2)}$ estarão contidas em $\Sigma_{X^{(1)}X^{(2)}}$ ou $\Sigma_{X^{(2)}X^{(1)}}$. Analisar as covariâncias em $\Sigma_{X^{(1)}X^{(2)}}$ ou $\Sigma_{X^{(2)}X^{(1)}}$ pode ser extremamente trabalhoso, ainda mais se p e q forem grandes. Porém, o principal objetivo da correlação canónica é resumir as associações entre $X^{(1)}$ e $X^{(2)}$ em função de algumas poucas correlações escolhidas, ao invés das p x q correlações.

A ideia inicial de Hotelling (1936) é de pesquisar a combinação linear entre duas variáveis em que por um lado $X^{(1)}$ define primeiro um vetor com p componentes $a_{p\times 1}$, notado neste trabalho como vetor a e por outro $X^{(2)}$ define o primeiro vetor com q componentes $b_{q\times 1}$ notado neste trabalho como vetor b, tal que $a^t = \left[a_{11}\cdots a_{1k}\cdots a_{1p}\right]$ e $b^t = \left[b_{11}\cdots b_{1k}\cdots b_{1q}\right]$ maximizando o coeficiente de correlação entre as combinações lineares entre $U = a^t X^{(1)}$ e $V = a^t X^{(2)}$ em que os vetores a^t e $b^t \in \mathbb{R}^n$ são chamados ***vetores canónicos*** e os seus elementos $a \in \mathbb{R}^p$ e $b \in \mathbb{R}^q$ chamados ***de peso canónicos*** ou ***fatores canónicos***. As variáveis U e V são chamados de ***variáveis canónicas***.

A análise de correlação canónica é feita de forma interativa, o primeiro par de variáveis canónicas são as combinações lineares $\left(U_1, \mathrm{V}_1\right)$, e que maximiza a correlação $Corr\left(U_1, V_1\right)$.

O segundo par de variáveis canónicas são as combinações lineares (U_2, V_2), U_2 combinação linear que não esteja correlacionada com U_1 e V_2 combinação linear do vetor $X^{(2)}$ não correlacionada com V_1, que maximiza a correlação $Corr(U_2, V_2)$.

Na k-ésima etapa, o k-ésima par de variáveis canónicas são as combinações lineares (U_k, V_k), que maximizam a correlação $Corr(U_k, V_k)$ entre aquelas que não são correlacionadas com os $k-1$ primeiros pares de variáveis canónicas definidos.

A correlação entre o k-ésima par de variáveis canónicas é denominada por k-ésima correlação canónica.

Como as observações dos dois grupos de variáveis p-variáveis e q-variáveis formam duas matrizes completas, a matriz de covariância $\Sigma_{X^{(1)}X^{(2)}}$ será de tamanho s onde $s = \min(p, q)$ O valores próprios resultam de s tamanho das duas matrizes. Também se tem s coeficientes de correlação canónica e variáveis canónicas (U_k, V_k) , e sendo $2 \times s$ combinações lineares associadas entre os vetores canónicos a e b .

Matematicamente para encontrar os pesos canónicos a e b tem-se:

$$\{a, b\} = \arg.\max_{a,b} \left\{ Corr\left(a^t X^{(1)}, \mathrm{b}^t X^{(2)}\right) \right\}$$

e pelas variáveis de correlação canónica U e V é definida por

$$\rho_i = Corr(\mathrm{U}_i, \mathrm{V}_i),\ i = 1, .., \mathrm{s}$$

Sobre as restrições de $Var(U_i) = Var(V_i) = 1$ e $Cor(U_s, \mathrm{V}_k) = Cor(\mathrm{V}_k, U_s) = 0$, $i = 1, 2, .., (\mathrm{s}-1)$

Sejam as combinações lineares $U = a^t X^{(1)}$ e $V = a^t X^{(2)}$, o coeficiente de correlação canónica entre as variáveis canónicas é dada pela,

$$\rho_{U,V} = Corr(U, V) = \frac{Cov(U, V)}{\sqrt{Var(U)}\sqrt{Var(V)}}$$

com,

$$Var(U) = Cov\left(a^t X^{(1)}\right) = a^t \Sigma_{X^{(1)}} a$$
$$Var(V) = Cov\left(b^t X^{(2)}\right) = b^t \Sigma_{X^{(2)}} b$$
$$Cov(U,V) = a^t Cov\left(X^{(1)}, X^{(2)}\right) b = a^t \Sigma_{X^{(1)}X^{(2)}} b$$

A correlação entre U e V é definida por:

$$\rho_{U,V} = Corr(U,V) = \frac{a^t \Sigma_{X^{(1)}X^{(2)}} b}{\sqrt{a^t \Sigma_{X^{(1)}} a}\ \sqrt{b^t \Sigma_{X^{(2)}} b}} \qquad (1.1)$$

Utilizando o método de multiplicadores de Lagrange da função a maximizar com as restrições é portanto:

$$V = a^t \Sigma_{X^{(1)}X^{(2)}} b - \lambda_a \left(a^t \Sigma_{X^{(1)}X^{(1)}} a - 1\right) - \lambda_b \left(b^t \Sigma_{X^{(2)}X^{(2)}} b - 1\right),$$

onde λ_a e λ_b ,são os multiplicadores de Lagrange. O máximo desta função resulta da sua derivação em que:

$$\frac{\partial V}{\partial a} = 0 \text{ e } \frac{\partial V}{\partial b} = 0.$$

Resolvendo as equações seguintes:

$$\frac{\partial V}{\partial a} = \Sigma_{X^{(1)}X^{(2)}} b - 2\lambda_a \Sigma_{X^{(1)}X^{(1)}} a = 0 \qquad (1.2)$$

$$\frac{\partial V}{\partial a} = \Sigma_{X^{(1)}X^{(2)}} b - 2\lambda_a \Sigma_{X^{(1)}X^{(1)}} a = 0 \qquad (1.3)$$

$$\frac{\partial V}{\partial b} = \Sigma_{X^{(2)}X^{(1)}} a - 2\lambda_b \Sigma_{X^{(2)}X^{(2)}} b = 0 \qquad (1.4)$$

Multiplicando (1.2) por a^t e (1.3) por b^t e sabendo que $a^t \Sigma_{X^{(1)}X^{(1)}} a = b^t \Sigma_{X^{(2)}X^{(2)}} b = 1$ e que $a^t \Sigma_{X^{(1)}X^{(2)}} b = b^t \Sigma_{X^{(2)}X^{(1)}} a$ temos:

$$a^t \Sigma_{X^{(1)}X^{(2)}} b = 2\lambda_a$$

$$b^t \Sigma_{X^{(2)}X^{(1)}} a = 2\lambda_b$$

$$\Rightarrow 2\lambda_a = 2\lambda_b.$$

Considerando $\lambda_1 = 2\lambda_a = 2\lambda_b$ e $a^t \Sigma_{X^{(1)}X^{(1)}} a = b^t \Sigma_{X^{(2)}X^{(2)}} b = 1$ sabemos que

$$Corr(U,V) = a^t \Sigma_{X^{(1)}X^{(2)}} b = 2\lambda_a = \lambda_1 \ .$$

A equação $\Sigma_{X^{(1)}X^{(2)}} b = 2\lambda_a \Sigma_{X^{(1)}X^{(1)}} a$

$$\Leftrightarrow \Sigma_{X^{(1)}X^{(2)}} b = \lambda_1 \Sigma_{X^{(1)}X^{(1)}} a \qquad (1.5)$$

$$\Sigma_{X^{(2)}X^{(1)}} a = 2\lambda_b \Sigma_{X^{(2)}X^{(2)}} b$$

$$\Leftrightarrow \Sigma_{X^{(2)}X^{(1)}} a = \lambda_1 \Sigma_{X^{(2)}X^{(2)}} b \qquad (1.6)$$

Portanto a partir de (1.5) $a = \lambda_1^{-1} \Sigma_{X^{(1)}X^{(1)}}^{-1} \Sigma_{X^{(1)}X^{(2)}} b$ e substituindo em (1.6) vem

$$\lambda_1^{-1} \Sigma_{X^{(2)}X^{(1)}} \Sigma_{X^{(1)}X^{(1)}}^{-1} \Sigma_{X^{(1)}X^{(2)}} b = \lambda_1 \Sigma_{X^{(2)}X^{(2)}} b$$

Multiplicando por λ_1 e por $\Sigma_{X^{(2)}X^{(2)}}^{-1}$ os dois membros obtemos

$$\left(\Sigma_{X^{(2)}X^{(1)}} \Sigma_{X^{(1)}X^{(1)}}^{-1} \Sigma_{X^{(1)}X^{(2)}} \Sigma_{X^{(2)}X^{(2)}}^{-1}\right) b = \lambda_1^2 b \qquad (1.7)$$

Da mesma forma a partir de (1.6) $b = \lambda_1^{-1} \Sigma_{X^{(2)}X^{(2)}}^{-1} \Sigma_{X^{(2)}X^{(1)}} a$ e substituindo em (1.5) vem

$$\lambda_1^{-1} \Sigma_{X^{(1)}X^{(2)}} \Sigma_{X^{(2)}X^{(2)}}^{-1} \Sigma_{X^{(2)}X^{(1)}} a = \lambda_1 \Sigma_{X^{(1)}X^{(1)}} a$$

Multiplicando por λ_1 e por $\Sigma_{X^{(1)}X^{(1)}}^{-1}$ aos ambos os membros obtemos

$$\left(\Sigma_{X^{(1)}X^{(2)}} \Sigma_{X^{(2)}X^{(2)}}^{-1} \Sigma_{X^{(2)}X^{(1)}} \Sigma_{X^{(1)}X^{(1)}}^{-1}\right) a = \lambda_1^2 a \qquad (1.8)$$

Portanto por (1.7) e (1.8) estamos na presença de um problema de valores e vetores próprios da matriz quadrada, onde λ_1^2 é o valor próprio da matriz de $\Sigma_{X^{(2)}X^{(1)}} \Sigma_{X^{(1)}X^{(1)}}^{-1} \Sigma_{X^{(1)}X^{(2)}} \Sigma_{X^{(2)}X^{(2)}}^{-1}$ e da $\Sigma_{X^{(1)}X^{(2)}} \Sigma_{X^{(2)}X^{(2)}}^{-1} \Sigma_{X^{(2)}X^{(1)}} \Sigma_{X^{(1)}X^{(1)}}^{-1}$ e os vetores a e b são os vetores próprios das referidas matrizes.

Tendo a matriz $\underset{p\times s}{A} = \left[a_1, a_2, ..., \mathrm{a_s}\right]$ e $\underset{p\times s}{B} = \left[b_1, \mathrm{b_2}, ..., \mathrm{b_s}\right]$ onde $s = \min(p,q)$ composto por vetores canónicos, as variáveis canónicas (U,V) podemos escrever sobre a forma de matriz:

$$\underset{s\times 1}{U} = \left[u_1, \mathrm{u_2}, ..., \mathrm{u_s}\right]^t = \left[a_1^t \mathrm{X}^{(1)}, a_2^t \mathrm{X}^{(1)}, ..., a_s^t \mathrm{X}^{(1)}\right]^t \quad (1.9)$$

$$\underset{s\times 1}{V} = \left[v_1, \mathrm{v_2}, ..., \mathrm{v_s}\right]^t = \left[b_1^t \mathrm{X}^{(2)}, \mathrm{b_2^t X^{(2)}}, ..., \mathrm{b_s^t X^{(2)}}\right]^t \quad (1.10)$$

Podemos encontrar s por de combinação linear entre $X^{(1)}$ *e* $X^{(2)}$. O primeiro par (U_1, V_1) tem a correlação máxima e o segundo par (U_2, V_2) é de correlação máxima e ortogonal a (U_1, V_1) e assim sucessivamente.

Outro método para obter a correlação canónica.

A metodologia explicada anteriormente não é a única possível para encontrar variáveis canónicas. Elas podem ser encontradas a partir da decomposição de matriz.

O resultado a seguir, de acordo com Johnson e Wichern (2005), fornece detalhes de como as variáveis canónicas e suas correlações são obtidas.

Supomos que $p \le q$ e seja os vetores aleatórios $\mathrm{X}^{(1)}$ e $\mathrm{X}^{(2)}$ com $Cov(\mathrm{X}^{(1)}) = \Sigma_{X^{(1)}}$, $Cov(\mathrm{X}^{(2)}) = \Sigma_{X^{(2)}}$ e $Cov(\mathrm{X}^{(1)}, \mathrm{X}^{(2)}) = \Sigma_{X^{(1)}X^{(2)}} = \Sigma_{X^{(2)}X^{(1)}}$ .Sejam as combinações lineares $U = a^t X^{(1)}$ e $V = b^t X^{(2)}$. Então

$$\max_{a,b} Corr\left(U, V\right) = \rho_1^* \ ,$$

é obtida pelas combinações lineares (primeiro par de variáveis canónicas)

$$U_1 = e_1^{'} \sum_{X^{(1)}}^{-\frac{1}{2}} X^{(1)} \text{ e } V_1 = f_1^{'} \sum_{X^{(2)}}^{-\frac{1}{2}} X^{(2)} \ ,$$

o k-ésimo par de variáveis canónicas, k = 1,2,…,p

$$U_k = e_k^{'} \sum_{X^{(1)}}^{-\frac{1}{2}} X^{(1)} \text{ e } V_k = f_k^{'} \sum_{X^{(2)}}^{-\frac{1}{2}} X^{(2)} \ ,$$

maximiza

$$Corr\left(U_k, V_k\right) = \rho_k^* \ ,$$

entre as combinações lineares não correlacionadas com as precedentes $1, 2, ..., \mathrm{k}-1$ variáveis canónicas.

Considerando aqui $\rho_1^{*2} \ge \rho_2^{*2} \ge ... \ge \rho_p^{*2}$ como sendo os valores próprios de $\sum_{X^{(1)}}^{-\frac{1}{2}} \sum_{X^{(1)}X^{(2)}} \sum_{X^{(2)}}^{-1} \sum_{X^{(2)}X^{(1)}} \sum_{X^{(1)}}^{-\frac{1}{2}}$ e $e_1, e_2, ..., e_p$ são os $(p \times 1)$ valores próprios associados. Os valores $\rho_1^{*2}, \rho_2^{*2}, ..., \rho_p^{*2}$ são também os p valores próprios da matriz $\sum_{X^{(2)}}^{-\frac{1}{2}} \sum_{X^{(2)}X^{(1)}} \sum_{X^{(1)}}^{-1} \sum_{X^{(1)}X^{(2)}} \sum_{X^{(2)}}^{-\frac{1}{2}}$ com

os (q×1) valores próprios correspondentes $f_1, f_2, ..., f_p$ .Cada vetor f_i é proporcional á matriz $\sum_{X^{(2)}}^{-1} \sum_{X^{(2)}X^{(1)}} \sum_{X^{(1)}}^{-\frac{1}{2}} e_i$, logo, a correlação entre U_k e V_k é definida por:

$$Corr\left(U_k, V_k\right) = \rho_k^* = \sqrt{\rho_k^{*2}}$$

As variáveis canónicas têm as seguintes propriedades:

$Var(U_k) = Var\left(V_k\right) = 1$

$Cov(U_k, \mathrm{U}_l) = Corr(U_k, \mathrm{U}_l) = 0, k \neq l;$

$Cov(\mathrm{V}_k, \mathrm{V}_l) = Corr(\mathrm{V}_k, \mathrm{V}_l) = 0, k \neq l;$

$Cov(\mathrm{U}_k, \mathrm{V}_l) = Corr(\mathrm{U}_k, \mathrm{V}_l) = 0, k \neq l, \forall_k, l = 1, 2, ..., p$

Quando utilizado apenas uma amostra da população, as matrizes de covariância Σ podem se substituídas pelas matrizes de correlação ρ . A matriz de covariância e de correlação serão R e $\hat{\rho}$, respetivamente. Assim, as correlações e variáveis canónicas serão da mesma forma, a partir da matriz de covariância ou de correlação.

I.4. Coeficientes padronizados

Os coeficientes nas variáveis canónicas $U_i = a_i^t X^{(1)}$ $V_i = b_i^t X^{(2)}$ podem refletir diferenças no dimensionamento das variáveis, bem como diferenças na contribuição das variáveis a correlação canónica. Para remover o efeito de escala, a_i^t e b_i^t pode ser padronizado multiplicando pelos desvios-padrão das variáveis correspondentes

$c_i = D_{x^{(1)}} a_i^t$, e $d_i = D_{x^{(2)}} b_i^t$, onde $D_{x^{(1)}} = diag\left(s_{x_1^{(1)}}, s_{x_2^{(1)}}, .., s_{x_p^{(1)}}\right)$ e $D_{x^{(2)}} = diag\left(s_{x_1^{(2)}}, s_{x_2^{(2)}}, .., s_{x_q^{(2)}}\right)$. Todavia c_i e d_i podem ser obtidos diretamente como vetores próprios de $R_{X^{(1)}X^{(1)}}^{-1} R_{X^{(1)}X^{(2)}} R_{X^{(2)}X^{(2)}}^{-1} R_{X^{(2)}X^{(1)}}$ e $R_{X^{(2)}X^{(2)}}^{-1} R_{X^{(2)}X^{(1)}} R_{X^{(1)}X^{(1)}}^{-1} R_{X^{(1)}X^{(2)}}$ respetivamente.

A partir dos valores próprios c_i e d_i são os coeficientes dos vetores padronizados. Para mostrar isso, note que em termos de variáveis centradas $x_1^{(1)} - \bar{X}^{(2)}$.

$$U = a^t\left(x_1^{(1)} - \bar{X}^{(1)}\right) = a^t D_{x_1^{(1)}} D^{-1}_{x_1^{(1)}}\left(x_1^{(1)} - \bar{X}^{(1)}\right)$$

$$= c^t D^{-1}_{x_1^{(1)}}\left(x_1^{(1)} - \bar{X}^{(1)}\right)$$

$$= c_1 \frac{\left(x_1^{(1)} - \bar{X}^{(1)}\right)}{s_{x_1^{(1)}}} + c_2 \frac{\left(x_2^{(1)} - \bar{X}^{(1)}\right)}{s_{x_2^{(1)}}} + \ldots + c_p \frac{\left(x_p^{(1)} - \bar{X}^{(1)}\right)}{s_{x_p^{(1)}}}$$

e

$$V = b^t\left(x_1^{(2)} - \bar{X}^{(2)}\right) = b^t D_{x_1^{(2)}} D^{-1}_{x_1^{(2)}}\left(x_1^{(2)} - \bar{X}^{(2)}\right)$$

$$= c^t D^{-1}_{x_1^{(2)}}\left(x_1^{(2)} - \bar{X}^{(2)}\right)$$

$$= d_1 \frac{\left(x_1^{(2)} - \bar{X}^{(2)}\right)}{s_{x_1^{(2)}}} + d_2 \frac{\left(x_2^{(2)} - \bar{X}^{(2)}\right)}{s_{x_2^{(2)}}} + \ldots + d_q \frac{\left(x_p^{(2)} - \bar{X}^{(2)}\right)}{s_{x_q^{(2)}}}$$

Assim, o efeito das diferenças de tamanho ou dimensionamento das variáveis é removido, e os coeficientes $c_{i1}, c_{i2}, \ldots c_{ip}$ em c_i reflete a relativa contribuição de cada $x_1^{(1)}, x_2^{(1)}, \ldots, x_p^{(1)}$ para U_i . A forma similar pode ser feito em relação a d_i

Os coeficientes padronizados demonstram a contribuição das variáveis na presença de cada um. Assim, se algumas das variáveis são excluídos e outros acrescentados, os coeficientes poderão mudar. Este é precisamente o comportamento que desejamos a partir dos coeficientes num ambiente multivariado.

I.5. Correlação entre cada variável original e a variável canónica

Muitos autores recomendam o passo adicional de converter os coeficientes padronizados para correlações.

Essas correlações são por vezes referidas como loadings ou coeficientes de estrutura, e é amplamente alegado que eles fornecem uma interpretação mais válida das variáveis canónicas. Rencher (1988; 1992 b; 1998, Secção 8.6.3) mostrou, no entanto, que a soma ponderada das correlações entre $X_i^{(1)}, i \in \{1, 2, \ldots p\}$ e as variáveis canónicas, u_i , $i \in \{1, 2, \ldots, s\}$ é igual a $R^2_{x_i^{(1)} x^{(2)}}$ a correlação múltipla ao quadrado entre cada variável de $X_i^{(1)}, i \in \{1, 2, \ldots p\}$ com as variáveis de $X^{(2)}$. Não há informações sobre como o $X^{(1)}$ de contribuir conjuntamente a correlação canónica com os $X^{(2)}$ no seu todo. Portanto, as correlações são inúteis na aferição da importância de dada variável no contexto.

I.6. Analise Canónica de redundância

A análise de correlação canónica ajuda-nos a explorar um conjunto de dados e pode tentar ajudar a encontrar modelos de dependências interessantes. Este índice nos permite constatar em que medida as variáveis originais podem ser explicadas por uma variável canónica de outro grupo.

Em 1968, Stewart e Love propuseram um índice baseado nas médias dos coeficientes de inter - correlação dum grupo em relação as variáveis são standard. O princípio é observar em que medida, um grupo de variáveis originais podem ser explicadas por outras variáveis canónicas opostas (uma combinação linear de outro grupo). È portanto um índice assimétrico entre os dois conjuntos que formam uma base a fim de encontrar o modelo de predição entre os grupos. A assimetria vem igualmente na certeza que a variância explicada muda segundo o grupo de variáveis originais que desejamos explicar.

Este índice é a proporção de variância total de um grupo que é explicada por uma combinação linear (variável canónica) de outro grupo. Sabe-se que o quadrado de inter-correlação explica a proporção de variâncias de uma variável original por uma variável original.

A redundância é uma medida de associação entre os grupos com base nas correlações entre variáveis originais e variáveis canónicas. Uma vez que estas correlações fornecem informações apenas univariadas, a redundância transforma e acaba por ser uma medida de relação univariada, em vez de uma medida da relação multivariada

Se o quadrado de correlação múltipla entre cada uma das variáveis do grupo de variáveis $X_i^{(2)}$ e os do grupo $X^{(1)}$ tem a notação $R^2_{X^{(2)}X_i^{(1)}}$ então a média do quadrado da correlação múltipla de $X_i^{(2)}$ e os $V_i \;\; i=1,2,...,\mathrm{p}$ é dada por:

$$Rd\left(X \mid \mathrm{V}\right)=\frac{\sum_{i=1}^{p} R^2_{X^{(2)}X_i^{(1)}}}{p}.$$

e da forma similar para a redundância de variáveis $X_i^{(1)}$, dado o grupo U , é a média

$$Rd\left(X \mid \mathrm{U}\right)=\frac{\sum_{i=1}^{q} R^2_{X_i^{(2)}X^{(1)}}}{q}.$$

onde $R^2_{X_i^{(2)}X^{(1)}}$ é o quadrado de correlação múltipla de $x_i^{(2)}$ com $x^{(1)}$.

As duas redundâncias medidas em cima não são simétricas, isto é, $Rd\left(X^{(1)} \mid \mathrm{U}\right) \neq Rd\left(X^{(2)} \mid V\right)$.

I.7. Variáveis Canónicas Amostrais e Correlações Canónicas Amostrais

Considerando uma amostra aleatória composta por n observações dos vetores $X^{(1)}$, p-dimensional e o $X^{(2)}$, q-dimensional, tem-se a matriz de dados de dimensão $n\times(p+p)$ representada por:

$$Z=\left[X^{(1)}\mid X^{(2)}\right]=\begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)}\cdots x_{1p}^{(1)} & x_{11}^{(2)} & x_{12}^{(2)}\cdots x_{1q}^{(2)} \\ x_{21}^{(1)} & x_{22}^{(1)}\cdots x_{2p}^{(1)} & x_{11}^{(2)} & x_{12}^{(2)}\cdots x_{1q}^{(2)} \\ \vdots & \vdots\ \ddots\ \vdots & \vdots & \vdots\ \ddots\ \vdots \\ x_{n1}^{(1)} & x_{n2}^{(1)}\cdots x_{np}^{(1)} & x_{n1}^{(2)} & x_{n2}^{(2)}\cdots x_{nq}^{(2)} \end{bmatrix}$$

e o vetor de médias com dimensão $p+q\times 1$

$$\bar{Z}=\begin{bmatrix}\bar{X}^{(1)}\\ \bar{X}^{(2)}\end{bmatrix}$$

Tendo como $\bar{X}^{(1)}=\frac{1}{n}\sum_{n=1}^{n}x_n^{(1)}$ e $\bar{X}^{(2)}=\frac{1}{n}\sum_{n=1}^{n}x_n^{(2)}$

A matriz de covariâncias amostral pode ser representada por :

$$S=\begin{bmatrix}S_{X^{(1)}} & S_{X^{(1)}X^{(2)}}\\ S_{X^{(2)}X^{(1)}} & S_{X^{(2)}}\end{bmatrix},$$

e as combinações lineares por

$\hat{U}=\hat{a}^{t}X^{(1)}$ e $\hat{V}=\hat{b}^{t}X^{(2)}$,

pelo que as correlações canónicas amostrais serão dadas por:

$$r_{\hat{U},\hat{V}}=\frac{\hat{a}^{t}S_{X^{(2)}X^{(1)}}\hat{b}}{\sqrt{\hat{a}^{t}S_{X^{(1)}}\hat{a}}\sqrt{\hat{b}^{t}S_{X^{(2)}X^{(1)}}\hat{b}}} \qquad (1.11)$$

O primeiro par de variáveis canónicas amostrais é formado pelas combinações lineares $\hat{U}_1$ e $\hat{V}_1$, com variância amostral unitária, e que maximizam a correlação em (1.11)

O segundo par de variáveis canónicas amostrais são as combinações lineares $\hat{U}_2$ e $\hat{V}_2$, com variância amostral unitária, e que maximizam a correlação em (1.11) dentre aquelas que não são correlacionadas com o primeiro par de variáveis canónicas.

Na k-ésima etapa, o k-ésimo par de variáveis canónicas amostrais são as combinações lineares $\hat{U}_k$ e $\hat{V}_k$, com variância amostral unitária, e que maximizam a correlação (1.11) entre aquelas que não são correlacionadas com os k-1 primeiros pares de variáveis canónicas definidos.

I.8. Teste de Significância para Validação da Analise de Correlação Canónica

Para verificar a aplicabilidade da análise de correlação canónica ao conjuntos de dados do problema, deve-se verificar se os vetores $X^{(1)}$ e $X^{(2)}$ são independentes entre si, ou não correlacionados (Mingoti, 2007). Caso isso aconteça, a aplicação do método será inútil, pois $a^t X^{(1)}$ e $b^t X^{(2)}$ terão correlação zero para qualquer escolha de a^t e b^t. Portanto, para validar a análise de correlação canónica é necessária uma análise da matriz de covariância ou de correlações a fim de determinar se elas são próximas ou não da matriz nula.

Assim vamos testar as seguintes hipóteses:

$$H_0 : \Sigma_{X^{(1)}X^{(2)}} = 0_{p\times q}$$

contra

$$H_1 : \Sigma_{X^{(1)}X^{(2)}} \neq 0_{p\times q}$$

ou, equivalente para a matriz das correlações e a estatística do teste é definida por:

$$-2\ln(\Lambda) = -\left(n-1-\frac{1}{2}(p+q+1)\right)\ln\left(\prod_{i=1}^{p}(1-\lambda_i)\right)$$

onde n é o tamanho da amostra e Λ é a estatística de Wilk's dada por $\Lambda = \frac{1}{\prod_{i=1}^{p}(1-\lambda_i)}$ A estatística do teste está associada a uma distribuição Qui-quadrado χ^2_v com pq graus de liberdade. Se o valor obtido na equação acima for maior ou igual ao respetivo valor da distribuição qui-quadrado, com o nível de significância igual a 5%, rejeitamos a hipótese nula. Existe também uma aproximação d estatística pela distribuição F que contudo não foi utilizada neste trabalho.

I.9. Algoritmo baseado por formulação algébrica

Na sua formulação clássica, a analise de Correlação Canónica é a procura da solução de um problema sobre valores próprios (Hotelling, 1936; Anderson, 1958, citados em IGNACIO GONZÁLEZ

$$S_{X^{(1)}}^{-1} S_{X^{(1)}X^{(2)}} S_{X^{(2)}}^{-1} S_{X^{(2)}X^{(1)}} a = \lambda^2 a$$

$$S_{X^{(2)}}^{-1} S_{X^{(2)}X^{(1)}} S_{X^{(1)}}^{-1} S_{X^{(1)}X^{(2)}} b = \lambda^2 b$$

Pela decomposição de Cholesky

$$S_{X^{(1)}} = L_{x^{(1)}} L'_{x^{(1)}} \quad \text{e} \quad S_{X^{(2)}} = L_{x^{(2)}} L'_{x^{(2)}}$$

Onde $L_{x^{(1)}}$ e $L_{x^{(2)}}$ são as matrizes invisíveis, triangulares superiores de ordem p e q respetivamente,

depois de realizar a decomposição em valores da matriz M com $M = L_{X^{(2)}}^{-1} S_{X^{(1)}X^{(2)}} L_{X^{(1)}}^{-1}$

Preposição: Sejam $L_{x^{(1)}}$ e $L_{x^{(2)}}$ duas matrizes invisíveis onde

$$S_{X^{(1)}} = L_{x^{(1)}} L'_{x^{(1)}} \quad \text{e} \quad S_{X^{(2)}} = L_{x^{(2)}} L'_{x^{(2)}} \ ,$$

e seja $M = L_{X^{(2)}}^{-1} S_{X^{(1)}X^{(2)}} L_{X^{(1)}}^{-1}$. Estão o problema de valores próprios

$$S_{X^{(1)}}^{-1} S_{X^{(1)}X^{(2)}} S_{X^{(2)}}^{-1} S_{X^{(2)}X^{(1)}} a = \lambda^2 a \ ,$$

é equivalente à decomposição em valores da matriz M.

A proposição dá-nos um algoritmo para resolver a análise canónica.

Algoritmo sobre a formulação algébrica [IGNACIO GONZÁLEZ**]**

- Sejam $S_{x^{(1)}}$, $S_{x^{(2)}}$, $S_{x^{(1)},x^{(2)}}$.
- Calcular matrizes triangulares superiores $L_{X^{(1)}}$ e $L_{X^{(2)}}$ onde

  $$S_{X^{(1)}} = L_{x^{(1)}} L'_{x^{(1)}} \quad \text{e} \quad S_{X^{(2)}} = L_{x^{(2)}} L'_{x^{(2)}}$$

  Através da decomposição de Cholesky
- Calcular $M = L_{X^{(2)}}^{-1} S_{X^{(1)}X^{(2)}} \left(L'_{X^{(1)}} \right)^{-1}$
- Calcular a decomposição em valores singulares de M

  $$M = U_M D_M V'_M \ ,$$

  onde s variáveis canónicas são : $U = X^{(1)} (\mathrm{L}'_{X^{(1)}})^{-1} V_M$ e $V = X^{(2)} (\mathrm{L}'_{X^{(2)}})^{-1} U_M$

  As correlações canónicas são: $\rho_j = \left[D_M \right]_{jj}$, $j = 1, ..., \mathrm{p}$.

CAPITULO II

GENERALIZAÇÃO DE ANÁLISE DA CORRELAÇÃO CANÓNICA PARA MAIS DE DOIS GRUPOS DE VARIÁVEIS

II.1. Introdução

O termo canónico em matemática quer dizer a simplicidade, regularidade, estrutura fundamental e de base. É desta forma reduzida e mais simples é que se pretende as relações ou funções. Por exemplo a forma canónica de uma matriz de covariância é a matriz de valores próprios.

O principal objetivo deste estudo é investigar e descrever maneiras de estender a teoria da correlação canónica ou análise canónica para lidar com mais do que dois conjuntos de variáveis aleatórias, que otimizam as funções construídas a partir das matrizes de correlação ou de variância.

Vários métodos são propostos para Generalizar a Análise de Correlação Canónica (GACC). Estes diferentes métodos são reagrupados em mesmos resultados da análise de correlação canónica simples [Hotelling, 1936]. Apresentaremos uma síntese dos métodos que permitem Generalizar a Analise de Correlação Canónica nomeadamente (método Ssqcor, método Sumcor, método Ecart, método Maxvar, método Minvar, e o método de Carroll, método Genvar)

II.2. Síntese dos métodos em (GACC).

Para o estudo de análise de correlação canónica é difícil encontrar as referências bibliográficas que resumem e apresentam a GACC. As termologias e as apresentações são muitos diferentes de um autor para outro. Portanto muitos autores, entendem que a análise canónica contém um contexto teórico de vários métodos de análise de dados.

Em algumas literaturas, analise canónica é frequentemente associado a análise de correlação canónica. Este método é com certeza o ponto central da análise canónica. É a partir desta método que outros autores de métodos de análise canónica foram desenvolvendo ao longo do tempo. É por essa razão que é necessário falar da análise canónica antes de análise de correlação canónica, o que se mostra indiferente para muitos autores.

Vamos apresentar a generalização da análise de correlação canónica de forma sintetizada como se segue.

Supomos que despomos de $p(p>2)$ grupos de variáveis em que $X^{(1)}, X^{(2)}, ..., X^{(p)}$ onde elas são compostas por n indivíduos e $m_1, m_2, ..., m_p$ o número total das variáveis de cada grupo.

O método de análise de correlação canónica generalizada tem como objetivo descrever e resumir a relação existente entre vários grupos de variáveis medidos para mesmo indivíduo.

Supomos também que a matriz de variâncias e covariâncias ou de correlação de $X = \left[X^{(1)} \mid X^{(2)} \mid ... \mid X^{(p)} \right]$ é decomposto segundo p grupos de variáveis da seguinte forma:

$$\Sigma_X = \begin{bmatrix} \Sigma_{X^{(1)}} & \Sigma_{X^{(1)}X^{(2)}} & \cdots & \Sigma_{X^{(1)}X^{(p)}} \\ \Sigma_{X^{(2)}X^{(1)}} & \Sigma_{X^{(2)}} & \cdots & \Sigma_{X^{(2)}X^{(p)}} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{X^{(p)}X^{(1)}} & \Sigma_{X^{(p)}X^{(2)}} & \cdots & \Sigma_{X^{(p)}} \end{bmatrix}$$

Sejam $Z_1, Z_2, ..., Z_p, p$ combinações lineares definidos a partir de $X^{(1)}, X^{(2)}, ..., X^{(p)}$ ou $Z_i = X^{(i)} \mathrm{P}_i$ e $\mathrm{P}_i \in R^{m_i}$ para todo $i \in \{1, 2, ..., p\}$. O Z é uma combinação linear de variáveis de X definido por $Z = XP, P \in R^m$ e $m = m_1 + m_2 + ... + m_p$ que representam o numero total de variáveis X .

Vamos apresentar funções para otimização e para a generalização de análise de correlação canónica.

Método Genvar

Este método consiste em minimizar o valor próprio de matriz de correlação do conjunto de variáveis canónicas. Ele foi proposto por Steel, 1951e consiste em encontrar vetores $P_1, P_2, ..., P_p$ que minimizam a função:

$$f\left(P_1, P_2, ..., P_p\right) = \min \prod_{j=1}^{m} \lambda_j \qquad (2.1)$$

Método Minvar

Este método consiste em minimizar o valor próprio de matriz de correlação de conjunto de variáveis canónicas. Ele foi proposto por Kenttenring, 1971 e consiste em encontrar vetores $P_1, P_2, ..., P_p$ que minimizam a função:

$$f\left(P_1, P_2, ..., P_p\right) = \min \left\{1 - cor(Z_i, Z_j\right\} \qquad (2.2)$$

Método Maxvar

Este método consiste em maximizar o valor próprio de matriz de correlação de conjunto de variáveis canónicas. Ele foi proposto por Horst, 1961 e consiste em encontrar vetores $P_1, P_2, ..., P_p$ que maximizam a função:

$$f\left(P_1, P_2, ..., P_p\right) = \max\left\{1 + cor(Z_i, Z_j\right\} \qquad (2.3)$$

Método Ecart

Este método consiste em minimizar a diferença entre o maior e menor o valor próprio de matriz de correlação de conjunto de variáveis canónicas. Ele foi proposto por Nzobounsana, 2001 e Nzobounsana et Dhorne, 2003 e consiste em encontrar vetores $P_1, P_2, ..., P_p$ que maximiza a função:

$$f\left(P_1, P_2, ..., P_p, U\right) = \arg\max_{P_1, P_2, ..., P_p} \max_{U \in H_P} \left\{\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}(\alpha^*)\,\mathrm{Cor}\left(Z_i, Z_j\right)\right\} \qquad (2.4)$$

onde,

- $\Delta_\alpha = \text{Diagonal}\left(\alpha_1, \alpha_2, ..., \alpha_p\right)$ a matriz diagonal associada ao vetor $\alpha = \left(\alpha_1, \alpha_2, ..., \alpha_p\right)$
- $H_p = \left\{U : {}^tUU = Id_p\right\}$, Conjunto de matrizes ortogonais
- $m_{ij}\left(\alpha\right)$: termo geral da matriz $U\Delta_\alpha\,{}^tU$
- α^* o vetor tal que $\alpha_1^* = 2, \alpha_P^* = 0$ e para $i \in \left\{2, 3, ..., p-1\right\}$ $\alpha_1^* = 1$

Método Sumcor

Este método consiste em maximizar a soma das correlações de grupos de variáveis canónicas. Ele foi proposto por Host em 1961 e consiste em encontrar vetores $P_1, P_2, ..., P_p$ que maximizam a função:

$$f\left(P_1, P_2, ..., P_p\right) = \max \sum_{i=1}^{p}\sum_{j=1}^{p} cor\left(Z_i, Z_j\right) \qquad (2.5)$$

Método Ssqcor

Este método consiste em maximizar a soma dos quadrados das correlações de grupos de variáveis canónicas. Ele foi proposto por Kettenring em 1971 e consiste em encontrar vetores $P_1, P_2, ..., P_p$ que maximizam a função:

$$f\left(P_1, P_2, ..., P_p\right) = \max \sum_{i=1}^{p} \sum_{j=1}^{p} cor^2\left(Z_i, Z_j\right) \qquad (2.6)$$

Método Carroll

Este método consiste em maximizar a soma das correlações múltiplas entre os vetores canónicos (combinação linear das variáveis de X) e as variáveis de cada grupo. Ele foi proposto por Carroll em 1968 que consiste em encontrar um vetor P que maximize a função:

$$f\left(P\right) = \max \sum_{i=1}^{p} cor^2\left(Z, X^i\right) \qquad (2.7)$$

II.3. Tipos de restrições impostas à variável canónica

A maior parte do problemas de otimização, o cálculo de variáveis e os vetores canónicos $Z_1, Z_2, ..., Z_p$ e $P_1, \mathrm{P}_2, ..., \mathrm{P}_p$ se faz passo a passo.

A cada passo, as variáveis canónicas obtidas são conservados com vista a construir uma base ortonormada por cada subespaço vetorial, $\mathrm{Im}\left(X^i\right)$ por $i = \{1, 2, ..., p\}$ composto por variáveis de cada X^i .

Seja $Z_1^s, Z_2^s, ..., Z_p^s$, o conjunto de variáveis obtidos no passo s, no passo $(s+1)$, estes são obtidos segundo uma família de restrição:

Restrições gerais: Para cada variável X^i, pretende-se ter uma base ortonormado de $\mathrm{Im}\left(X^i\right)$, otimizam-se as funções definidas sobre as restrições

$$cor\left(Z_i^{(s+1)}, Z_i^{(\mathrm{k})}\right) = 0, \quad i \in \{1, ..., p\} \quad e \ k \in \{1, ..., \mathrm{s}\}$$

As restrições particulares : Pretende-se uma base ortonormada dos vetores canónicos associados às variáveis canónicos as pesquisar, para tal utiliza-se a restrição de ortogonalidade seguinte:

$${}^t P_i^{(\mathrm{k})} P_i^{(s)} = 0 \ , k \in \{1, 2, ..., s\} \quad e \quad i\{1, 2, ..., \mathrm{p}\}$$

II.4. Solução de diferentes métodos

II.4.1 Solução do Método de Mínima Variância (Método Minvar) e de Máxima Variância (Método Maxvar)

Consideramos p grupos de variáveis $X^1, X^2, ...X^p$ medidos sobre o mesmo indivíduo. Notamos m_i o número total de variáveis de X^i , $Z_i = X_i P_i$ uma combinação linear das variáveis de X^i , P_i é um vetor R^{m_i}. Designamos $cor\left(Z_i, Z_j\right)$ a correlação linear entre Z_i, Z_j e Φ a matriz $p \times p$ de correlações lineares de Z_i. Designamos $\lambda_1(\Phi), \lambda_2(\Phi), ..., \lambda_p(\Phi)$ os valores próprios de matriz Φ de ordem decrescente $\lambda_1(\Phi) \geq \lambda_2(\Phi) \geq ... \geq \lambda_p(\Phi)$ e Σ_{ij} , a matriz de inter - correlação entre variáveis X^i e X^j.

O Métodos MINVAR [Kettering 71] e MAXVAR [Host 61a] consiste em primeiro lugar, em encontrar, P combinação lineares $Z_1, Z_2, ..., Z_p$ (soluções de ordem um) de variáveis de cada grupo tal que a menor (respetivamente o maior) valor próprio da matriz de correlação canónica seja minimizada (respetivamente maximizado).

As soluções de ordem $s\left(s > 1\right)$ consistem em encontrar igualmente p combinações lineares de variáveis de cada grupo tal que os mesmos critérios são otimizados, sobre as restrições adicionais que, as variáveis canónicas d'ordem s sejam ortogonais as variáveis canónicas de ordem $(s-1)$ por $s \in \{2, 3, ..., r\}$ e $r = \min\left(m_1, m_2, ..., m_p\right)$ e m é o numero de variáveis X.

Definição [V.NZOBOUNSANA, T.DHORNE] Seja Φ uma matriz simétrica e definida positivo. Seja $\lambda_1(\Phi) \geq \lambda_2(\Phi) \geq ... \geq \lambda_p(\Phi) > 0$ os p valores próprios não nulos onde $\lambda_1(\Phi)$ é o maior valor próprio e $\lambda_p(\Phi)$ é o menor valor próprio. Tem-se os seguintes resultados (cf [Anderson 84]) $\lambda_1(\Phi) = \max_w \left\{ {}^t w \Phi w \right\}$ *e* $\lambda_p(\Phi) = \min_u \left\{ {}^t u \Phi u \right\}$,

onde w e u são vetores normados, definidos por $w = {}^t\left\{w_1, w_2, ..., w_p\right\} \in R^p$ e $u = {}^t\left\{u_1, u_2, ..., u_p\right\} \in R^p$

Preposição[Kenttenring 71]: Sobre as restrições $Var\left(Z_i\right) = 1$ para todo i os vetores canónicos P_i são soluções de método *MAXVAR* ou, *MINVAR* si e só se elas verificam as seguintes igualdades:

$$P_i^{(1)} = \frac{w_{\max}^i}{\left({}^t w_{\max}^i w_{\max}^i\right)^{\frac{1}{2}}} \text{ para MAXVAR}$$

$$P_i^{(1)} = \frac{u_{\min}^i}{\left({}^t u_{\min}^i u_{\min}^i\right)^{\frac{1}{2}}} \text{ para MINVAR}$$

onde, $u_{\min}$ é o ultimo valor próprio e primeiro $w_{\max}$ vetor próprio de Σ ;

$u_{\min}^i$ e $w_{\max}^j$ os i, i-ésimos vetores de $u_{\min}$ e $w_{\max}$ de tamanho m_i ;

Σ é a matriz de correlação dos grupos de variáveis $Y^{(1)}, \mathrm{Y}^{(2)}, ..., Y^{(\mathrm{p})}$ para todo i $Y^{(\mathrm{i})} = \mathrm{X}^i \Sigma_{ii}^{-1/2}$ e Σ_{ii} a matriz de correlação de variáveis de X^i

II.4.2 Solução do Método (Método Ecart)

O Método Ecart, tem como objetivo encontrar para cada grupo de variáveis, as variáveis canónicas de diferentes ordem, e tal que em cada ordem dada a diferença entre o maior e o menor valor próprio.

Consideramos p grupos de variáveis $X^{(1)}, X^{(2)}, ..., X^{(\mathrm{p})}$ medido sobre os mesmos indivíduos. Seja m_i o número total de variável de X^i , $Z_i = X^i P_i$ uma combinação linear de variáveis de X^i , P_i é um vetor R^{m_i}. Notamos $Cor\left(Z_i, Z_j\right)$ a combinação linear entre Z_i e Z_j e Φ a matriz $p \times p$ de correlação linear Z_i. Notamos $\lambda_1(\Phi), \lambda_2(\Phi), ..., \lambda_p(\Phi)$ o valor próprio de Φ em ordem decrescente (isto é $\lambda_1(\Phi) \geq \lambda_2(\Phi) \geq ... \geq \lambda_p(\Phi)$ e Σ_{ij} a matriz de inter - correlação entre a matriz X^i e X^j.

Matematicamente o método de Ecart reduz-se função escrita em,

$$\left(P_1, P_2, P_3, ..., P_p\right) = \underset{P_1, P_2, ..., P_p}{\arg\max} \left\{ \frac{1}{p} \left\{ \lambda_1(\Phi) - \lambda_p(\Phi) \right\} \right\} \qquad (2.8)$$

Calculo variáveis canónicas de ordem um

A preposição que se segue dá-nos a solução para o método Ecart.

Preposição: As soluções do método Ecart sobre as restrições que $Z_1, Z_2, ..., Z_p$ são normados pela seguinte igualdade:

$$\lambda_i^* P_i = \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) \Sigma_{ij} P_j \qquad (2.9)$$

para todo $i \in \{1, 2, ..., p\}$ e $\lambda_1^*, \lambda_2^*, ..., \lambda_p^*$ são multiplicadores de Lagrange. Esta preposição define um método iterativo de obtenção de soluções do método Ecart.

Demonstração: Pela preposição precedente as soluções do método Ecart são obtidos pela, procura simultanea os vetores canónicos $P_1, P_2, ..., \mathrm{P}_p$ e de uma matriz ortogonal U tal que:

$$\frac{1}{p} tr\left[{}^t U \Phi U \Delta_\alpha \right] = \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} m_{ij}(\alpha) Cor\left(Z_i, Z_j\right), \quad (2.10)$$

seja maximizada sobre as restrições que $Var\left(Z_i\right) = {}^t P_i \Sigma_{ii} P_i = {}^t P_i P = 1 \quad \forall_i$.

Fixando a matriz U e introduzindo os multiplicadores de Lagrange $\lambda_1^*, \lambda_2^*, ..., \lambda_p^*$ deve-se portanto maximizar a função:

$$L\left(P_1, P_2, ..., P_p, \mathrm{U}\right) = \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} m_{ij}(\alpha) Cor\left(Z_i, Z_j\right) - \sum_{i=1}^{p} \lambda_i \left(Var(\mathrm{Z}_i) - 1\right). \qquad (2.11)$$

Derivando L em relação a cada P_i e igualando o resultado a zero, deduzimos a igualdade (2.9):

$$\lambda_i^* P_i = \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) \Sigma_{ij} P_j \ , \qquad (2.12)$$

onde $\lambda_i^* = \lambda_i - \frac{m_{ii}(\alpha)}{p}$

Portanto pela restrição de normalização canónica de variáveis e de igualdade (2.10) que:

$$\lambda_i = \frac{m_{ii}(\alpha)}{p} + \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha)^t P_i \Sigma_{ij} P_j = \frac{1}{p} \sum_{j=1}^{p} m_{ij}(\alpha) Cor\left(Z_i, Z_j\right) \qquad (2.13)$$

Calculo variáveis canónicas de ordem s

Definição: [V.NZOBOUNSANA, T.DHORNE*]*

Chama-se a solução de ordem k $(k-1)$ do problema , $\left(P_1,P_2,P_3,...,P_p\right)=\underset{P_1,P_2,...,P_p}{\arg\max}\left\{\frac{1}{p}\left\{\lambda_1(\Phi)-\lambda_p(\Phi)\right\}\right\}$

para encontrar p combinações lineares $Z_1^{(k)},Z_2^{(k)},...,Z_p^{(k)}$ onde para todo i, $Z_i^{(k)}=X^i\,\mathrm{P}_i^{(k)}$,

$\mathrm{P}_i^{(k)}\in R^{m_i}$ e $\Phi=\left(Cor\left(Z_i^k,Z_j^k\right)\right)_{ij}$ tal que:

$$\left(P_1^{(k)},P_2^{(k)},...,P_p^{(k)}\right)=\arg\max_{P_1^{(k)},P_2^{(k)},...,P_p^{(k)}}\max_{U}\left\{tr\left[{}^tU\Phi U\Delta_\alpha\right]\right\} \qquad (2.14)$$

sobre as restrições

$$Var\left(Z_1^{(k)}\right)=Var\left(Z_2^{(k)}\right)=...=Var\left(Z_p^{(k)}\right)=1$$

$Cor\left(Z_i^k,Z_j^k\right)=0$ para $s\in\left\{1,2,...,(k-1)\right\}$ e $k\in\left\{1,2,...,\mathrm{r}\right\}$

onde Δ_α é uma matriz diagonal definida por $\Delta_\alpha=\left\{2,1,1,1,...,1,0\right\}$

A preposição seguinte, dá-nos a igualdade que deve verificar para obtermos a solução de ordem superior do método Ecart.

Preposição:[V.NZOBOUNSANA, T.DHORNE*]*

Sejam $Z_i^{(k)}$ as soluções de ordem k do problema

$$\left(P_1,P_2,P_3,...,P_p\right)=\underset{P_1,P_2,...,P_p}{\arg\max}\left\{\frac{1}{p}\left\{\lambda_1(\Phi)-\lambda_p(\Phi)\right\}\right\} \qquad (2.15)$$

sobre as restrições de normalização $\mathrm{var}\left(Z_1^k\right)=\mathrm{var}\left(Z_2^k\right)=...=\mathrm{var}\left(Z_p^k\right)=1$ e as restrições adicionais $corr\left(Z_i^k,Z_i^s\right)=0$ por $s\in\left\{1,2,...,(k-1)\right\}$ e $s\in\left\{2,...,\mathrm{r}\right\}$ as soluções do problema (2.15) a ordem s $(s>1)$, $k\in\left\{1,2,3,...,(s-1)\right\}$ verifica as p seguintes equações:

$$\lambda_i^*P_i^{(s)}=\sum_{j=1,\,j\neq i}m_{ij}(\alpha)\left[Id_{m_i}-\sum_{k=1}^{s-1}P_i^{(k)t}P_i^{(k)}\right]\Sigma_{ij}\left[Id_{m_j}-\sum_{k=1}^{s-1}P_j^{(k)t}P_j^{(k)}\right]P_j^{(\mathrm{s})}\quad\forall_i \qquad (2.16)$$

onde $\lambda_1^*,\lambda_2^*,...\lambda_p^*$ são as constantes de normalização (multiplicadores de Lagrange) e $m_{ij}(\alpha)$, para todo $i\,e\,j$,é o termo geral da matriz $M=U\Delta_\alpha{}^tU$.

Demonstração: Tanto para o cálculo as soluções de ordem um, como de ordem s (s >1), as soluções de método Ecart são em maximizando a função seguinte:

$$L\left(P_1^{(s)},P_2^{(s)},...,P_p^{(s)}\right)=\frac{1}{p}\sum_{i=1,j=1}^{p}m_{ij}\left(\alpha\right)Cor\left(Z_i^{(s)},Z_j^{(s)}\right)-\sum_{i=1}^{p}\lambda_i\left(Var\left(Z_i^{(s)}\right)-1\right)- \\ -2\sum_{i=1}^{p}\sum_{k=1}^{s-1}\beta_i^k Cor\left(Z_i^{(k)},Z_i^{(s)}\right) \tag{2.17}$$

onde $\lambda_1,\lambda_2,...,\lambda_p$ e os $\beta_1^k,\beta_2^k,...,\beta_p^k$ para todo o $k\in\left\{1,2,...,\left(s-1\right)\right\}$ são os multiplicadores de Lagrange.

È possível demonstrar que através da restrição $Var\left(Z_i^l\right)={}^tP_i^{(l)}P_i^{(l)}=1$ para todo i e para $l\in\left\{1,2,...,(s-1)\right\}$ em que a função L é igual á:

$$L\left(P_p^{(s)}\right)=\sum_{i=1,j=1,i\neq j}^{p}\frac{m_{ij}}{p}\left(\alpha\right){}^tP_i^{(s)}\Sigma_{ij}P_i^{(s)}-\sum_{i=1}^{p}\lambda_i^*\,{}^tP_i^{(s)}P_i^{(s)}+ \\ +\sum_{i=1}^{p}\lambda_i-2\sum_{i=1}^{p}\sum_{k=1}^{s-1}\beta_i^{kt}P_i^{(s)}P_i^{(s)} \tag{2.18}$$

onde $P^{(s)}=\left(P_1^{(s)},P_2^{(s)},...,P_p^{(s)}\right)$ e $\lambda_i^*=\left\{\lambda_i^*-\frac{m_{ii}\left(\alpha\right)}{p}\right\}$

Derivando L em relação à $P_i^{(s)}$ e igualando o resultado a zero obtém-se:

$$\frac{1}{p}\sum_{j=1,i\neq j}^{p}m_{ij}\left(\alpha\right)\Sigma_{ij}P_j^{(s)}-\lambda_i^*P_i^{(s)}-\sum_{i=1}^{s-1}\beta_i^kP_i^{(k)}=0 \tag{2.19}$$

Multiplicando sucessivamente $\forall_k$ a equação (2.19) por ${}^tP_i^{(k)}$, obtemos

$$\beta_i^{(k)}=\frac{1}{p}\sum_{j=1,i\neq j}^{p}m_{ij}\left(\alpha\right){}^tP_i^{(k)}\Sigma_{ij}P_j^{(s)}\ ,\ \forall_k\in\left\{1,2,...,(s-1)\right\} \tag{2.20}$$

Substituindo $\forall_{\beta_i^{(k)}}$ na equação (2.19)

$$\frac{1}{p}\sum_{j=1,i\neq j}^{p}m_{ij}\left(\alpha\right)\Sigma_{ij}P_j^{(s)}-\lambda_i^*P_i^{(s)}-\frac{1}{p}\sum_{j=1,i\neq j}^{p}m_{ij}\left(\alpha\right)\left[\sum_{k=1}^{s-1}P_i^{(k)t}P_i^{(k)}\right]\Sigma_{ij}P_j^{(s)}=0 \tag{2.21}$$

Ou

$$\frac{1}{p}\sum_{j=1,i\neq j}^{p}m_{ij}\left(\alpha\right)\left[Id_{m_i}-\sum_{k=1}^{s-1}P_i^{(k)t}P_i^{(k)}\right]\Sigma_{ij}P_j^{(s)}-\lambda_i^*P_i^{(s)}=0 \tag{2.22}$$

Portanto $\forall_i$ e $P_i^{(s)}$, para cada U e Δ_α fixos verificando as p igualdades seguintes:

$$\lambda_i^* P_i^{(s)} = \frac{1}{p} \sum_{j=1, i \neq j}^{p} m_{ij}(\alpha) \left[Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)t} P_i^{(k)} \right] \Sigma_{ij} P_j^{(s)} \qquad (2.23)$$

Ou de forma equivalente

$$\lambda_i^* P_i^{(s)} = \frac{1}{p} \sum_{j=1, i \neq j}^{p} m_{ij}(\alpha) \left[Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)t} P_i^{(k)} \right] \Sigma_{ij} \left[Id_{m_i} - \sum_{k=1}^{s-1} P_j^{(k)t} P_j^{(k)} \right], \quad (2.24)$$

para $\forall_i \in \{1, 2, ..., p\}$ pois $\left[Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)t} P_i^{(k)} \right] P_i^{(s)} = P_i^{(s)}$.

II.4.3 Solução do Método Soma das Correlações (Método Sumcor)

O Método Sumcor, tem como objetivo estudar a relação entre vários grupos de variáveis que maximiza,

$$f(P_1, P_2, ..., P_p) = \max \sum_{i=1}^{p} \sum_{j=1}^{p} cor(Z_i, Z_j)$$

Ela explica, por um lado a relação entre dois a dois p grupos de variáveis e por outro, a relação dentro de cada grupo e (p-1) dos restantes grupos. As soluções (variáveis e vetores canónicos) deste método são obtidas passo a passo.

Cálculo de variáveis de ordem um. As soluções de ordem 1 deste método são obtidos procurando para todo $i \in \{1, 2, 3, ..., p\}$, os valores $Z_1, Z_2, Z_3, ..., Z_p$ que maximiza a função:

$$f(P_1, P_2, ..., P_p) = \max \sum_{i=1}^{p} \sum_{i \neq j}^{p} cor(Z_i, Z_j) \qquad (2.25)$$

com as restrições

$$Var(Z_1) = Var(Z_2) = ... = Var(Z_p) = 1 \qquad (2.26)$$

Como é suposto que as variáveis sejam ortogonais duas a duas, a maximização de (2.25) sobre restrição de (2.26) é equivalente a encontrar p vetores canónicos tal que:

$$f(P_1, P_2, ..., P_p) = \arg.\max_{P_1, P_2, ..., P_p} \left\{ \sum_{i=1}^{p} \sum_{i \neq j}^{p} {}^t P_i \sum_{x^i x^j} P_j \right\} \quad (2.27)$$

sobre as restrições

$${}^t P_1 P_1 = {}^t P_2 P_2 = ... = {}^t P_p P_p = 1,$$

o que quer dizer que cada vetor deve ter a norma igual a 1.

Preposição: As soluções de análise canónica generalizada segundo o método Sumcor verificam a seguinte igualdade

$$\lambda_i P_i = \sum_{j=1, i\neq j}^{p} \Sigma_{x^i x^j} P_j$$

(para todo $i \neq j$ e $i, j \in \{1,2,...,\mathrm{p}\}$) onde λ_i, são as constantes de normalização de $P_1, P_2, ..., P_p$. Estes resultados são obtidos utilizando o método de multiplicadores de Lagrange. Ele consiste em obter $P_1, P_2, ..., P_p$ soluções de sistema de equações:

$$\begin{pmatrix} I_{m_1} & \Sigma_{X^{(1)}X^{(2)}} \cdots \Sigma_{X^{(1)}X^{(\mathrm{p})}} \\ \Sigma_{X^{(2)}X^{(1)}} & I_{m_2} \quad \cdots \Sigma_{X^{(2)}X^{(\mathrm{p})}} \\ \vdots & \vdots \quad \ddots \quad \vdots \\ \Sigma_{X^{(\mathrm{p})}X^{(1)}} & \Sigma_{X^{(\mathrm{p})}X^{(2)}} \cdots I_{m_p} \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix} = \begin{pmatrix} \lambda_1 + 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 + 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 \cdots & & \lambda_p + 1 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix}$$

Logo que as constantes de normalização, $\lambda_1, \lambda_2, ..., \lambda_p$ sejam, a primeira solução do método Sumcor é obtido fazendo a decomposição de matriz

$$\begin{pmatrix} I_{m_1} & \Sigma_{X^{(1)}X^{(2)}} \cdots \Sigma_{X^{(1)}X^{(\mathrm{p})}} \\ \Sigma_{X^{(2)}X^{(1)}} & I_{m_2} \quad \cdots \Sigma_{X^{(2)}X^{(\mathrm{p})}} \\ \vdots & \vdots \quad \ddots \quad \vdots \\ \Sigma_{X^{(\mathrm{p})}X^{(1)}} & \Sigma_{X^{(\mathrm{p})}X^{(2)}} \cdots I_{m_p} \end{pmatrix}$$

Os vetores canónicos $P_1, P_2, ..., P_p$, são os seus vetores $P = \left({}^t P_1, ..., {}^t P_p\right)^t$ (primeiro valor próprio de Σ) associado ao maior valor próprio.

No caso geral, as soluções do problema de análise canónica generalizada segundo o método Sumcor, de obter um par de processo interativo (interativos sucessivos) de seguinte modo:

A partir de uma solução inicial $P_1^{(0)}, P_2^{(0)}, ..., P_p^{(0)}$ qualquer par pertence a $\mathrm{Im}\left(X^1\right), \mathrm{Im}\left(X^2\right), ..., \mathrm{Im}\left(X^p\right)$.

A solução de ordem 1 do problema é resultante de um conjunto de vetores $P_1^{(1)}, P_2^{(1)}, ..., P_p^{(1)}$

da solução do sistema de equação: $s_1 = \begin{cases} \sum_{j=1 j\neq 1}^{p} \Sigma_{x^1 x^j} P_j = \lambda_1 P_1 \ , \lambda_1 = \dfrac{P_1}{{}^t P_1 P_1} \\ \sum_{j=1 j\neq 1}^{p} \Sigma_{x^2 x^j} P_j = \lambda_2 P_2 \ , \lambda_2 = \dfrac{P_2}{{}^t P_2 P_2} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{j=1 j\neq p}^{p} \Sigma_{x^p x^j} P_j = \lambda_p P_p \ , \lambda_p = \dfrac{P_p}{{}^t P_p P_p} \end{cases}$

Sejam $P_1^{(1)}, P_2^{(1)}, ..., P_p^{(1)}$ conjunto de p vectores no $R^{m_1}, R^{m_2}, ..., R^{m_p}$ a solução do sistema de equações de s_1 tem a seguinte propriedade:

Preposição: As variáveis canónicas de ordem um do método Sumcor são normado (por todo $i, \mathrm{Var}(\mathrm{Z}_i^{(1)} = 1)$. Elas verificam as seguintes igualdade:

$$\begin{pmatrix} a_{11} I_{m_1} & \Sigma_{X^{(1)} X^{(2)}} \cdots \Sigma_{X^{(1)} X^{(p)}} \\ \Sigma_{X^{(2)} X^{(1)}} & a_{22} I_{m_2} \quad \cdots \Sigma_{X^{(2)} X^{(p)}} \\ \vdots \quad\quad \vdots \quad \ddots \quad \vdots \\ \Sigma_{X^{(p)} X^{(1)}} & \Sigma_{X^{(p)} X^{(2)}} \cdots a_{pp} I_{m_p} \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix} = \lambda \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix}$$

onde, $a_{ii} = \sum_{k=1, k\neq i}^{p} \sum_{j=1, j\neq k}^{p} cor\left(Z_k, Z_j\right), \lambda = \lambda_1 + \lambda_2 + ... + \lambda_p = a_{ii} + \lambda_i$

$$\lambda_i Z_i^{(1)} = \Pi_{x^i} \left(\sum_{i\neq j, j=1}^{p} Z_j^{(1)} \right) \quad i \in \{1, 2, ..., p\}$$

onde

Π_{x^i} é a matriz de projeção ortogonal associado ao sob espaço vetorial $I_m\left(X^i\right)$ formado pela variáveis de X^i e $Z_i^{(1)}$ uma variável canonica de X^i.

Assim os multiplicadores de Lagrange, $\lambda_1, \lambda_2, ... \lambda_p$ verificam as igualdades:

$$\lambda_i = \mathrm{cov}\left(Z_i^{(1)}, \sum_{j=1, i\neq j}^{p} Z_j^{(1)} \right) \quad i \in \{1, 2, ..., p\}$$

Cálculo de variáveis canónicas de ordem 2. O procedimento de cálculo de variáveis canónicas de ordem dois é similar ao de ordem um. Consiste em maximizar a função $f(\mathrm{P}_1, \mathrm{P}_2, ..., \mathrm{P}_p)$ em relação a cada vetor P_i ,com as restrições que as variáveis canónicas são normado e de restrições

adicionais $Cor\left(Z_i, Z_i^{(1)}\right) = 0$ (a solução de ordem dois é ortogonal a solução de ordem um). Mais precisamente, resulta em encontrar soluções do problema:

$$\left(P_1, P_2, ..., P_p\right) = \arg.\max\left[\sum_{i=1}^{p}\sum_{j=1}^{p} cor\left(Z_i, Z_j\right)\right],$$

com variância,

$$\text{var}\left(Z_1\right) = \text{var}\left(Z_2\right) = ... = \text{var}\left(Z_p\right) = 1$$

e as correlações entre as variáveis canónicas,

$$\text{Cor}\left(Z_1, Z_1^{(1)}\right) = \text{Cor}\left(Z_2, Z_2^{(1)}\right) = ... = \text{Cor}\left(Z_p, Z_p^{(1)}\right) = 0$$

Preposição*: A* análise de ordem dois da análise de correlação canónica generalizada segundo o método Sumcor verificam as seguintes igualdades:

$$\lambda_i P_i = \sum_{j=1}^{p}\left[\mathrm{I}_{\mathrm{m}_i} - \Pi_{p_i^{(1)}}\right]\Sigma_{x^i x^j}\left[\mathrm{I}_{\mathrm{m}\,j} - \Pi_{p_i^{(1)}}\right]P_j \text{ para todo } i \neq j$$

onde $\Pi_{p_i^{(1)}}$ (respectivamente λ_i) desiguinado de projeção ortogonal associado ao sub espaço $I_m\left(P_i^{(1)}\right)$ (respectivamente as constantes de Lagrange)

Demonstração: Seja $\lambda_1, \lambda_2, ..., \lambda_p$ e $\mu_1, \mu_2, ..., \mu_p$, $2p$ multiplicadores de Lagrange. Para encontrar as soluções de ordem dois do método Sumcor, deve-se resolver (porque para todo $i, \Sigma_{x^i} = Id_{m_i}$) o seguinte sistema de equação:

$$\frac{\partial}{\partial P_i}\left\{\sum_{i=1}^{p}\left\{\sum_{i\neq j}^{p} {}^t P_i \Sigma_{x^i x^j} P_j - \lambda_i\left({}^t P_i P_i - 1\right) - \mu_i {}^t P_i P_i^{(1)}\right\}\right\} = 0 \qquad (2.28)$$

ao qual depois de simplificação resulta:

$$\lambda_i P_i = \sum_{j=1}^{p}\left[\mathrm{I}_{\mathrm{m}_i} - \Pi_{p_i^{(1)}}\right]\Sigma_{x^i x^j}\left[\mathrm{I}_{\mathrm{m}\,j} - \Pi_{p_i^{(1)}}\right]P_j \text{ para todo } i \neq j$$

onde $\Pi_{p_i^{(1)}}$ é a matriz (respetivamente λ_i) designado de projeção ortogonal associado ao sub espaço $I_m\left(P_i^{(1)}\right)$ (respetivamente as constantes de Lagrange)

Como para a soluções de ordem 1, os vetores canónicos $P_1^{(2)}, P_2^{(2)}, ..., P_p^{(2)}$ solução do problema (2.28 são obtidos por um processo de iterativo similar a S_1. Desde que faça a substituição da matriz do sistema $\Sigma_{x^i x^j}$ do sistema S_1 para novas matrizes $\left[\mathrm{I}_{\mathrm{m}_i} - \Pi_{P_i^{(1)}}\right]\Sigma_{x^i x^j}$

Cálculo variáveis canónicas de ordem s $(s>2)$. Seja $\left(P_1^{s-1}, P_2^{s-1}, ..., P_p^{s-1}\right)$ p vetores de $R^{m_1}, R^{m_2}, ..., R^{m_p}$ que maximizam a função $\left(P_1, P_2, ..., P_p\right)$ a ordem $(s-1)$.

A ordem s $(s>3)$, procura-se encontrar p novos vetores canónicos $\left(P_1^s, P_2^s, ..., P_p^s\right)$ que maximiza a função $f_1\left(P_1, P_2, ..., P_p\right) = \sum_{i=1}^{p}\sum_{j=1}^{p} cor\left(Z_i, Z_j\right)$,

com as restrições condicionais que ,

$$^tP_i^s P_i^k = 0 \text{ para todo } k \in \{1, 2, ..., (s-1)\}$$

Utilizando os multiplicadores de Lagrange para maximizar a função

$$L_{(\lambda,\beta)}(P) = \sum_{i=1}^{p}\left\{\sum_{j=1}^{p} corr\left(Z_i^{(s)}, Z_j^{(s)}\right) - \lambda_i\left(^tP_i^{(s)}P_i^{(s)} - 1\right) - \sum_{k=1}^{(s-1)}\beta_i^k\, ^tP_i^{(s)}P_i^{(k)}\right\} \qquad (2.29)$$

onde $\lambda = \left\{\lambda_1, \lambda_2, ..., \lambda_p\right\}$ e $\beta = \left\{\beta_1^k, \beta_2^k, ..., \beta_p^k\right\}$, chamado multiplicadores de Lagrange. , e p é um numero real. Portanto o resultado seguinte permite resolver o problema.

Preposição: A ordem s sobre as restrições condicionais, a função (2.25) é maximizado se e só se os vetores canónicos procurado verifica-se as seguintes sistema:

$$S_1^{(s)}\begin{cases} \sum_{j=1, j\neq 1}^{p} M_1^{(s)} \sum_{X^{(1)}X^{(j)}} M_j^{(s)} P_j^{(s)} = \lambda_1 P_1^{(s)} \\ \sum_{j=1, j\neq 2}^{p} M_1^{(s)} \sum_{X^{(2)}X^{(j)}} M_j^{(s)} P_j^{(s)} = \lambda_2 P_2^{(s)} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{j=1, j\neq p}^{p} M_p^{(s)} \sum_{X^{(p)}X^{(j)}} M_j^{(s)} P_j^{(s)} = \lambda_p P_p^{(s)} \end{cases}$$

onde a matriz M_i^s para todo i é definido por:

$M_i^s = I_{m_i} - \prod_A$ e $A = \mathrm{I_m}\left\{P_1^{(1-s)}, P_2^{(1-s)}, ..., P_p^{(1-s)}\right\}$ e $\prod_A$ chamado a matriz de projeção ortogonal associado ao A.

Demostração: Para ter o resultado da preposição precedente, deriva-se a igualdade (2.29) em ordem a p vetores $P_1^{(s)}, P_2^{(s)}, ..., P_p^{(s)}$ e igualando a zero obtém-se o seguinte resultado:

$$\sum_{j=1, i\neq j}^{p} \Sigma_{x^i x^j} P_j^{(s)} - \lambda_i P_i^{(s)} - \sum_{k=1}^{(s-1)} \beta_i^k P_i^{(k)} = 0 \quad \forall_i \qquad (2.30)$$

Multiplicando a equação (2.30) por $^tP_i^{(k)}$ obtemos

$$\sum_{j=1,i\neq j}^{p} {}^{t}P_i^{(k)}\Sigma_{x^i x^j}P_j^{(s)} = \beta_i^k \quad \forall_i \;\; e \;\; k \qquad (2.31)$$

Substituindo β_i^k em (2.30), obtemos:

$$\sum_{j=1,i\neq j}^{p}\left[\mathrm{I}_{m_i} - \sum_{k=1}^{(s-1)} P_i^{(kt)}P_i^{(k)}\right]\Sigma_{x^i x^j}P_j^{(s)} = \lambda_i P_j^{(s)} \qquad (2.32)$$

o que permite dizer que o vector $P_1^{(s)}, P_2^{(s)}, ..., P_p^{(s)}$ verifica bem o sistema de equação $S^{(s)}$ (pois que $M_j^{(s)}P_j^{(s)} = P_j^{(s)}$)

II.4.3 Solução do Método Soma dos Quadrados das Correlações (Método Ssqcor)

Esta generalização de analise canónica foi proposta por [Kettering, 1971]. Ele consiste em encontrar p combinações lineares $Z_1, Z_2, .., Z_p$ chamado variáveis canónicas que maximizam a função:

$$f_2\left(P_1, P_2, ..., P_p\right) = \sum_{i=1}^{p}\sum_{j\neq i}^{p} cor^2\left(X^i P_i, X^j P_j\right) \qquad (2.33)$$

Calculo variáveis canónicas de ordem um

O primeiro passo implica encontrar p vetores canónicos $P_1, P_2, ..., P_p$ (pois se supõe que as variáveis são dois a dois ortogonal e normado.) associado às p variáveis canónicas $Z_1, Z_2, .., Z_p$ que maximizam a função (2.33).

Sejam $\lambda_1\lambda_2, ..., \lambda_p$ p multiplicadores de Lagrange à resolução de sistema de equação:

$$\frac{\partial L\left(P_1, ..., P_p\right)}{\partial P_i} = 0 \quad \text{onde } L\left(P_1, ..., P_p\right) = \sum_{i=1}^{p}\sum_{j\neq i} cor^2\left(Z_i, Z_j\right) - \sum_{i=1}^{p}\lambda_i\left({}^{t}P_i P_i - 1\right)$$

De forma análoga ao Sumcor, permite verificar que sobre a restrição ${}^{t}P_i P_i = 1$ os vetores canónicos que maximizam a função (2.33) pode ser escrita em forma de matriz

$$\begin{pmatrix} I_{m_1} & r_{12}\Sigma_{X^{(1)}X^{(2)}} & \cdots & r_{1p}\Sigma_{X^{(1)}X^{(p)}} \\ r_{21}\Sigma_{X^{(2)}X^{(1)}} & I_{m_2} & \cdots & r_{2p}\Sigma_{X^{(2)}X^{(p)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1}\Sigma_{X^{(p)}X^{(1)}} & r_{p2}\Sigma_{X^{(p)}X^{(2)}} & \cdots & I_{m_p} \end{pmatrix}\begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix} = \begin{pmatrix} \lambda_1^* & 0 & \cdots & 0 \\ 0 & \lambda_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p^* \end{pmatrix}\begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix}$$

onde para todo i, j $r_{ij} = cor\left(Z_i, Z_j\right)$ e $\lambda_i^* = \lambda_i + 1$

De maneira explícita, a igualdade matricial precedente é equivalente a p sistema de equações que a seguir se apresente:

$$S_2 = \begin{cases} \sum_{j=1, j\neq 1}^{p} r_{1j}\Sigma_{x^1x^j} P_j = \lambda_1^* P_1 \\ \sum_{j=1, j\neq 1}^{p} r_{2j}\Sigma_{x^2x^j} P_j = \lambda_2^* P_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{j=1, j\neq p}^{p} r_{jp}\Sigma_{x^jx^p} P_j = \lambda_p^* P_1 \end{cases}$$

É mostrado também que se $P_1, P_2, ..., P_p$ são soluções de s_2 então as variáveis canónicas associados de $Z_1, Z_2, ..., Z_p$.

Preposição:

Sejam $P_i^{(1)}, P_i^{(2)}, ..., P_i^{(s-1)}$ as $(s-1)$ primeiras variáveis canónicas de X^i que maximizam a função $f_2\left(P_1^{(s)}, P_2^{(s)}, ..., P_p^{(s)}\right)$ em ordem $1, 2, ..., \left(s-1\right)$.

Para ordem s, procura-se encontrar p combinações lineares $Z_1^{(s)}, Z_2^{(s)}, ..., Z_p^{(s)}$, onde $\forall_i$ os vetores de $Z_i^{(s)}$ são definidos por $Z_i^{(s)} = Y^i P_i^{(s)}$, que maximiza a função $f_2\left(P_1^{(s)}, P_2^{(s)}, ..., P_p^{(s)}\right)$ sobre as restrições ${}^t P_i^{(s)} P_i^{(s)} = 1$ e uma restrição adicional ${}^t P_i^{(s)} P_i^{(k)} = 0$ para todo $k \in \left\{1, 2, ..., \left(s-1\right)\right\}$.

II.4.4 Solução do Método (Método Carroll)

Carroll (1968) introduziu uma generalização da análise de correlação canónica de dois conjuntos de variáveis para n conjuntos de variáveis.

A generalização de análise canónica proposta por Carroll tem por objetivo medir e caraterizar as combinações lineares existentes entre p grupos de variáveis. Ela apoia-se na noção de proximidade entre um sub-espaço vetorial dito global (sub espaço vetorial) formado pelas variáveis de $X = \left[X^1, X^2, ..., X^p\right]$ e dos sub - espaços vetorial ditos parciais de $\mathrm{Im}(\mathrm{X}^1) ... \mathrm{Im}(\mathrm{X}^p)$ onde $\mathrm{Im}(\mathrm{X}^i)$ é o sub espaço vetorial formado pelas variáveis X^i.

Definição: Seja $X^1, X^2, ..., X^p$, p grupos de variáveis medidos sobre o mesmo indivíduo. Seja m_i para todo i, o número de variáveis do grupo X^i e $m = \sum_{i=1}^{p} m_i$ o número total de variáveis de $X = \left[X^1, X^2, ..., X^p \right]$. Segundo Carrol chamamos a generalização de analise de correlação canónica a um $q = \inf \left\{ m_1, m_2 ... m_p \right\}$ combinações lineares, notado como $Z^{(k)} = XP^{(k)}$ com $k \in \left\{ 1, 2, ..., p \right\}$ e $P^{(k)} \in R^p$, $p = m_1 + m_2 + ... + m_p$ .A solução é a seguinte:

$$p^{(k)} = \arg.\max_{P} \left\{ \sum_{i=1}^{p} Cor^2 \left(Z^{(k)}, X^i \right) \right\},$$

onde $Var(Z^{(k)}) = 1$ e $Cor(Z^{(s)}, Z^{(k)}) = 0$ para $\forall_k \in \left\{ 1, 2, ..., q \right\}$ e $s \neq 1$

A Solução da GACC segundo Carroll, consiste em encontrar o máximo da soma do quadrado das correlações múltiplas entre um vetor, as variáveis de cada grupo sobre a condição que $\text{var}(z) = 1$ é resolvida pela matemática clássica por meio dos multiplicadores de Lagrange. Verifica-se que pela ordem um, $Z^{(1)} = Z = XP$ (respetivamente, $P^{(1)} = \text{P}$), é solução da equação, $\left\{ \sum_{i=1}^{p} \prod_{xi} - \lambda I_{n \times n} \right\} Z = 0 \quad \Leftrightarrow \left\{ \sum_{i=1}^{p} \prod_{xi} \right\} Z = \lambda Z = 0 \Leftrightarrow XM^t XZ = \lambda Z = 0,$ onde $M = diag.bloc\left(\sum_{x^1}^{-1}, ..., \sum_{x^p}^{-1} \right)$ (respetivamente, da equação $M \sum_x P = \lambda P = 0$) e que as variáveis canónicas notado Z_i , para todo i ,verificando as igualdades: $Z_i = kk \prod_{x^i} Z$ onde kk é um parâmetro de normalidade e $\prod_{x^i} = X \sum_{x^i}^{-1 t} X^i$.

As variáveis canónicas de GACC denotado por $Z_i^{(1)}, Z_i^{(2)}, ..., Z_i^{(k)}$ onde para todo i $Z_i^{(k)} = X^i P_i^{(k)}$, são as projeções dos vetores $Z_i^{(1)}, Z_i^{(2)}, ..., Z_i^{(k)}$ num sub - espaço vetorial formado pelas variáveis de X^i.

CAPITULO III

RELAÇÃO ENTRE ANALISE DE CORRELAÇÃO CANONICA E OUTRAS TECNICAS MULTIVARIADAS

III.1. Introdução

A Análise de Correlação Canónica (ACC) é um método de estatística descritiva multivariada que apresenta analogia com Analise de Componente Principais (ACP), pela construção e interpretação de gráficos e com Análise de Regressão Múltipla (ARM) pela natureza dos dados.

Na análise de correlação canónica pretende-se explorar as relações que podem existir entre dois grupos de variáveis quantitativas observada pelo mesmo individuo. O facto de estudar a relação entre dois grupos de variáveis constitui a principal particularidade de análise de correlação canónica em relação a Analise de Componente Principais.

A ACC pode ser vista como uma extensão da análise de regressão linear múltipla. Lembre que ARM envolve uma única variável dependente métrica e diversas variáveis independentes métricas e não só. A ACC envolve várias variáveis dependentes métricas.

III.2. Análise de Regressão Múltipla (ARM)

Há uma ligação direta entre os coeficientes de variáveis canónicas e o coeficiente regressão múltipla.

A matriz de coeficientes de regressão de ($X^{(1)}$ como variável dependente) e ($X^{(2)}$ como variável independente) é obtido com $\hat{B}_1 = S_{x^{(1)}x^{(1)}} S_{x^{(1)}x^{(2)}}$.

Esta matriz pode ser usada para obtenção de a_i e b_i:

$$b_i = \hat{B}_1 a_i \qquad (3.1)$$

desde que a_i e b_i sejam vetores próprios de $b_i = \hat{B}_1 a_i$, pode-se escrever $b_i = c\hat{B}_1 a_i$, onde c é um fator escalar arbitrário.

No coeficiente de correlação canónica, o vetor b_i é expresso como combinação linear das colunas de $\hat{B}_i$. A expressão similar para a_i pode-se obter pela regressão de $X^{(2)}$ no $X^{(1)}: a_i = S_{x^{(1)}x^{(1)}} S_{x^{(1)}x^{(2)}} b_i$.

A correlação canónica pode ser definida como uma extensão de correlação múltipla. Do mesmo modo, a correlação canónica reduz-se a correlação múltipla quando um dos dois conjuntos de variáveis tem apenas uma variável.

Quando $p=1$, por exemplo $R_{x^{(1)}x^{(1)}}$ reduz-se a um, e pela expressão $\left|R^{-1}_{X^{(2)}X^{(2)}} R_{X^{(2)}X^{(1)}} R^{-1}_{X^{(1)}X^{(1)}} R_{X^{(1)}X^{(2)}} - r^2 I\right| = 0$ a correlação canónica reduz-se para,

$$R^2 = R_{X^{(2)}X^{(1)}} R^{-1}_{X^{(1)}X^{(1)}} R_{X^{(1)}X^{(2)}} \qquad (3.2)$$

As duas estatísticas do teste de Wilks em teste de analise multivariada [Reencher, Alvin C ,(1934), secção 10.5.1 e 10.5.2] isto é, o teste para a regressão total e o teste de um subconjunto de $X^{(1)}$ e $X^{(2)}$, podem ambas ser expressa em termos de correlações canónicas. Por [Reencher, Alvin C ,(1934), secção 10.55 e 11.17], o teste estatístico para toda a regressão com a hipótese nula $H_0: B_1 = 0$ pode ser escrito como:

$$\Lambda_f = \frac{\left|\left(X^{(1)}\right)^t X^{(1)} - \hat{B}^t \left(X^{(2)}\right)^t X^{(1)}\right|}{\left|\left(X^{(1)}\right)^t X^{(1)} - n\left(\bar{X}^{(1)}\right)^t \bar{X}^{(1)}\right|} \qquad (3.3)$$

$$= \prod_{i=1}^{s}\left(1 - r_i^2\right) \qquad (3.4)$$

onde r_i^2 é a i-ésima correlação canónica ao quadrado. Um teste estatístico para $H_0: B_1 = 0$ hipótese de $X^{(1)}$'s não depende do último h das $X^{(2)}$'s, e pelo [Reencher, Alvin C ,(1934), secção 10.65], como

$$\Lambda\left(x_{q-h+1}, ..., x_p \mid x_1, ..., x_{q-h}\right) = \frac{\Lambda_f}{\Lambda_r} \qquad (3.5)$$

onde $\Lambda_f = \prod_{i=1}^{s}\left(1 - r_i^2\right)$ e

$$\Lambda_r = \frac{\left|\left(X^{(1)}\right)^t X^{(1)} - \hat{B}_r^t \left(X^{(2)}\right)^t X^{(1)}\right|}{\left|\left(X^{(1)}\right)^t X^{(1)} - n\left(\bar{X}^{(1)}\right)^t \bar{X}^{(1)}\right|} \qquad (3.6)$$

O Λ_r pode ser expresso em termos de quadrado de correlação canónica $c_1^2, c_2^2, ..., c_t^2$ entre $X_1^{(1)}, X_2^{(1)}, ...X_p^{(1)}$ e $X_1^{(2)}, X_2^{(2)}, ...X_{q-h}^{(2)}$:

$$\Lambda_r = \prod_{i=1}^{t} \left(1 - c_i^2\right) \qquad (3.7)$$

onde $t = \min(\mathrm{p}, \mathrm{q} - \mathrm{h})$ e h o numero de parâmetro em B

Usamos a notação c_i^2 ao invés de r_i^2 para enfatizar a correlação canónica para diferenciar o modelo reduzido do modelo completo.

Pelo (3.4) e (3.7) o teste do modelo reduzido e o modelo completo de $H_0 : B_d = 0$ em (3.5) pode agora ser expresso em termo de correlação canónica como,

$$\Lambda\left(x_{q-h+1}, ..., x_p \mid \mathrm{x}_1, ..., x_{q-h}\right) = \frac{\Lambda_f}{\Lambda_r} = \frac{\prod_{i=1}^{s}\left(1 - r_i^2\right)}{\prod_{i=1}^{t}\left(1 - c_i^2\right)} \qquad (3.8)$$

Se $p = 1$,como em regressão múltipla, então $s = t = 1$ e (3.8) reduz a,

$$\Lambda = \frac{1 - R_f^2}{1 - R_r^2} \qquad (3.9)$$

onde R_f^2 e R_r^2 são o quadrado de correlação múltipla para o modelo completo e o modelo reduzido. A distribuição de Λ em (3.9) é $\Lambda_{1,h,n-q-1}$ quando H_0 é verdadeira.

Desde que $p = 1$ será uma F-transformação exata

$$F = \frac{(1 - \Lambda)(n - q - 1)}{\Lambda h} \ ,$$

que segue uma distribuição F com h e n-q-1 grau de liberdade.

Substituindo o $\Lambda = \frac{\left(1 - R_f^2\right)}{\left(1 - R_r^2\right)}$ a estatística F será expressado em termo de R^2 ,

$$F = \frac{\left(R_f^2 - R_r^2\right)(n-q-1)}{\left(1-R_f^2\right)h} \qquad (3.10)$$

A análise de correlação canónica pode ser tratada como método de regressão multivariada. Muller (1982) discute a relação entre a análise de correlação canónica e a regressão múltipla e análise de componentes principais.

III.3. MANOVA e análise discriminante

Segundo [Reencher, Alvin C ,(1934), secção 6.1.8 e 8.4.2] é notado que em MANOVA ou na analise discriminante, $\lambda_i / (1+\lambda_i)$ é igual a r_i^2 onde λ_i é o i-ésimo valor próprio de $E^{-1}H$ e r_i^2 é o i-ésimo quadrado de correlação canónica entre o p variável dependente e o $k-1$ grupo de variáveis.

Sejam as variáveis dependentes designados por $X_i^{(1)} \quad i = 1, 2, ..., p$. Vamos representar os K grupos de variáveis pelo $K-1$ dummy variáveis $X_i^{(2)} \quad i = 1, 2, ..., (\mathrm{k}-1)$ definido para cada membro de i-ésimo grupo, $i \le k-1$ como $X_1^{(2)} = 0, ..., X_{i-1}^{(2)} = 0, X_i^{(2)} = 1, X_{i+1}^{(2)} = 0, X_{k-1}^{(2)} = 0.$ Para o k-ésimo grupo, todos os $X_i^{(2)}$ *são zero*.

O modelo MANOVA é equivalente à regressão multivariada de $X_i^{(1)} \quad i = 1, ..., \mathrm{p}$ variáveis dependentes com as variáveis dummy $X_i^{(2)} \quad i = 1, 2, ..., (\mathrm{k}-1)$. O teste de MANOVA o $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ é equivalente ao teste de regressão multivariada $H_0 : \mathrm{B}_1 = 0$ [Reencher, Alvin C ,(1934), secção 11.17],

$$\Lambda = \prod_{i=1}^{s} \left(1 - r_i^2\right) \qquad (3.11)$$

Quando comparamos o teste (3.11) e o teste estatístico MANOVA

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \qquad (3.12)$$

Obtém-se

$$r_i^2 = \frac{\lambda_i}{1+\lambda_i} \qquad (3.13)$$

e

$$\lambda_i = \frac{r_i^2}{1-r_i^2} \qquad (3.14)$$

Para estabelecer esta relação mais formal, pode-se escrever como:

$$Ha = \lambda Ea \qquad (3.15)$$

e $\left(\Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(1)}X^{(1)}}\Sigma_{X^{(1)}X^{(2)}}\Sigma^{-1}_{X^{(2)}X^{(2)}} - r^2 I\right)a = 0$ poderá também ser escrito como

$\Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(1)}X^{(1)}}\Sigma_{X^{(1)}X^{(2)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}a = r^2 a$ e multiplicando ambos os membros por $\Sigma^{-1}_{X^{(1)}X^{(1)}}$ obtemos

$$\Sigma_{X^{(2)}X^{(1)}}\Sigma_{X^{(1)}X^{(2)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}a = r^2\Sigma^{-1}_{X^{(1)}X^{(1)}}a \qquad (3.16)$$

[Reencher, Alvin C ,(1934)] afirma que uma vez que a MANOVA é equivalente a regressão multivariada o grupo de variáveis dummy , pode ser substituído pelos valores de E e H em (3.15) pode-se obter

$$\Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}\Sigma_{X^{(1)}X^{(2)}}a = \lambda\left(\Sigma_{X^{(2)}X^{(2)}} - \Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}\Sigma_{X^{(1)}X^{(2)}}\right)a \qquad (3.17)$$

e subtraindo $r^2\Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}\Sigma_{X^{(1)}X^{(2)}}a$ em (3.16) temos:

$$\Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}\Sigma_{X^{(1)}X^{(2)}}a = \frac{r^2}{1-r^2}\left(\Sigma_{X^{(2)}X^{(2)}} - \Sigma_{X^{(2)}X^{(1)}}\Sigma^{-1}_{X^{(2)}X^{(2)}}\Sigma_{X^{(1)}X^{(2)}}\right)a \qquad (3.18)$$

Uma comparação de (3.17 e 3.18) resulta

$$\lambda = \frac{r^2}{1-r^2}$$

CAPITULO IV

APLICACÃO DA ANÁLISE DA CORRELAÇÃO CANÓNICA.

IV.1. Introdução

Passamos agora a uma avaliação das informações contidas nas correlações canónicas entre variáveis canónicas. Apresentamos e discutimos três instrumentos comuns para a interpretação das variáveis canónicas: (1) coeficientes padronizados, (2) a correlação entre cada variável e a variável canónica, e (3) realizar a análise de redundância.

Na aplicação prática sobre dados reais, apenas apresentaremos a análise de correlação canónica de 2 x 2 grupos de produtos. A dificuldade na implementação dos métodos de GACC, resulta do facto de não encontrarmos à data software com pacotes disponíveis para a sua aplicação e não ter sido possível desenvolvê-los. Por isso continuaremos a desenvolver pesquisa nesta técnica de forma a desenvolver algoritmos para dar respostas ao uso de três ou mais grupos de variáveis em simultâneo. Alguns autores, como já referido anteriormente, apresentam comparações entre diferentes métodos da análise de correlação canónica. Kettering,1969 apresenta comparação dos métodos (SUMCOR, SSQCOR, GENVAR), NZOBOUNSANA V., DHORNE T, 2003 ; apresenta comparação dos métodos (Ecart, MAXVAR, MINVAR) e NZOBOUNSANA V., GAYMARD S.,2010 apresenta comparação do métodos (SUMCOR,SSQCOR,CARROL).

IV.2. Introdução à análise de dados

Para aplicação da Análise Canónica para mais de dois grupos utilizamos uma base de dados real, não simulada, proveniente do Índice do Preços no Consumidor (IPC), relativa a um período de 60 meses, ou seja no período de 2010 a 2014, com informações produzidas pelo INE-STP (Instituto Nacional de Estatística de São Tome e Príncipe). As informações são produzidas mensalmente através de recolha de preços nos mercados locais.

Neste capítulo estamos interessados em conhecer a correlação canónica o entre o primeiro grupo de variáveis relacionadas com produtos alimentares, o segundo grupo de variáveis as bebidas e terceiro grupo, produtos não alimentares.

O Índice de Preços no consumidor é um indicador que mede a evolução no tempo dos Preços dos bens e Serviços considerados como representativos de uma estrutura de consumo, num determinado espaço geográfico.

O índice de preços de S. Tomé apresenta-se como sendo um índice unificado que representa os diversos tipos de estabelecimentos e todas as formas de venda. Para a elaboração deste índice utilizou-se a estrutura de despesa e um cabaz de produtos representativos da despesa das famílias, resultante do Inquérito aos Agregados Familiares, realizado em 1995 no distrito de Água-Grande.

Os resultados do IPC96 são analisados, segundo os principais grupos de bens e serviços que integram a despesa média do consumo familiar.

IV.3. Descrição do método de recolha dos dados

Recolhem-se preços na cidade de S. Tomé, nos diferentes pontos de venda situados em dois mercados (Mercado Municipal e Feira de Ponto) e em cerca de trinta lojas ou estabelecimentos diversos, o que perfaz um total de 160 locais de recolha de preços. Nos mercados a recolha é semanal o que dá uma média de dois a quatro preços observados por ponto de venda e por produto durante o mês, enquanto nas lojas ou noutros tipos de estabelecimento se observam três ou quatro preços por produto.

A dimensão da amostra é de cerca de 185 produtos, selecionados a partir da análise dos resultados do Inquérito aos Agregados Familiares, efetuado no período de Junho/Agosto de 1995.

Os produtos selecionados estão classificados de acordo com a COICOP (Classificação do Consumo Individual por Objetivo), nomenclatura que permite a comparabilidade quer com outros domínios de informação estatística (Contas Nacionais) quer a nível internacional.

Para o presente estudo procedemos à partição de variáveis em três grupos

Produtos Alimentares (PA)

PA1--Pão e cereais

PA2--Carnes

PA3--Peixe

PA4--Leite, Queijo e Ovos

PA5--Óleo e Gorduras

PA6--Frutas e Legumes

PA7--Batata, Mandioca e Outros Tubérculos

PA8--Açúcar

PA9--Café

PA10--Outros produtos alimentares

Produtos para Bebidas (PB)

PB1--Agua mineral

PB2--Gazoza

PB3--Cerveja

PB4--Vinho

PB5--Espirituosas

Produtos Não Alimentares (PNA)

PNA1--Vestuário e Reparações

PNA2--Sapatos e Reparações

PNA3--Mobiliários e Acessórios

PNA4--Artigos em têxtil

PNA5--Aparelhos

PNA6--Vidraçaria, Louça e Outros

PNA7--Artigos não duráveis

PNA8--Serviços domésticos

IV.4. Análise de correlação canónica entre os grupos de variáveis

O objetivo deste trabalho é realizar análise de correlações canónicas entre três grupos de variáveis.

Esta análise está diretamente relacionada com diversos métodos de dependência. Semelhante à análise de regressão, a meta da correlação canónica é quantificar a força da relação, neste caso entre dois conjuntos de variáveis. Como abordado anteriormente na parte teórica, a análise de correlação canónica lida com a associação entre composições de conjuntos de variáveis dependentes e independentes. Ao fazer isso devolve diversas funções canónicas independentes que maximizam a correlação entre as combinações lineares, também conhecidas como variáveis estatísticas, as quais podem ser tratadas como conjuntos de variáveis independentes e dependentes. A correlação canónica não termina com a derivação de uma única relação entre os conjuntos de variáveis. Em vez disso, diversas funções canónicas (pares de variáveis estatísticas canónicas) podem ser obtidas.

De acordo com a Tabela 1, os resultados permitem verificar que não existe uma correlação entre IPC dos diferentes grupos.

As correlações entre as variáveis iniciais apresentam uma correlação superior ou igual a 0.85 nos produtos como:

(PA1 Pão e cereais e PA3 Peixe com 0.89)

(PA1 Pão e cereais e PB3 Cerveja com 0.87)

(PA2 Carne e PA3 Peixe com 0.86)

(PA2 Carne e PNA2 Sapatos e Reparações com 0.92)

(PA2 Carne e PNA4 Artigos em têxtil com 0.91)

(PA3 Peixe e PNA1 Vestuário e Reparações com 0.87)

(PA3 Peixe e PNA6 Vidraçaria, Louça e Outros com 0.86)

(PA3 Peixe e PNA8 Serviços domésticos com 0.85)

(PA7 Batata, Mandioca e Outros Tubérculos e PB5 Espirituosas com 0.87)

(PA7 Batata, Mandioca e PNA1 Vestuário e Reparações com 0.87)

(PA7 Batata, Mandioca e PNA4 Artigos em têxtil com 0.90)

(PA7 Batata, Mandioca e PNA6 Vidraçaria, Louça e Outros com 0.86)

(PB3 Cerveja e PNA1 Vestuário e Reparações com 0.85)

(PB4 Vinho e PB5 Espirituosas com 0.87)

(PB4 Vinho e PNA8 Serviços domésticos com 0.86)

(PB5 Espirituosas e P416 e Serviços domésticos com 0.90)

(PNA1 Vestuário e Reparações e PNA4 Artigos em têxtil com 0.93)

(PNA1 Vestuário e Reparações e PNA6 Vidraçaria, Louça e Outros com 0.91)

(PNA2 Sapatos e Reparações e PNA4 Artigos em têxtil com 0.85)

(PNA4 Artigos em têxtil e PNA6 Vidraçaria, Louça e Outros com 0.91).

É de salientar que, o IPC relativo ao produto PA8 Açúcar e PB2 Gazoza tiveram a correlação muito baixa em relação a maioria dos produtos.

Podemos também verificar que os produtos como Vestuário e Reparações e Artigos em têxtil tiveram a correlação mais forte 0.93 de acordo com a Tabela 1.

Tabela1:Correlação entre as variáveis originais

	PA1	PA2	PA3	PA4	PA5	PA6	PA7	PA8	PA9	PA10	PB1	PB2	PB3	PB4	PB5	PNA1	PNA2	PNA3	PNA4	PNA5	PNA6	PNA7	PNA8
PA1	1																						
PA2	0.67	1																					
PA3	0.89	0.86	1																				
PA4	0.55	0.35	0.51	1																			
PA5	0.53	0.3	0.45	0.68	1																		
PA6	0.81	0.6	0.8	0.57	0.5	1																	
PA7	0.72	0.82	0.82	0.49	0.54	0.59	1																
PA8	-0.12	-0.15	-0.2	0.32	0.41	-0.03	0.03	1															
PA9	0.7	0.4	0.58	0.49	0.66	0.62	0.57	0.12	1														
PA10	0.61	0.54	0.62	0.25	0.16	0.43	0.57	-0.22	0.23	1													
PB1	0.45	0.01	0.32	0.58	0.6	0.46	0.33	0.22	0.56	0.08	1												
PB2	-0.07	-0.42	-0.26	0.36	0.63	-0.05	-0.08	0.43	0.32	-0.32	0.53	1											
PB3	0.87	0.64	0.81	0.58	0.61	0.79	0.83	0.03	0.69	0.57	0.56	0.06	1										
PB4	0.82	0.59	0.8	0.65	0.62	0.8	0.75	0.02	0.7	0.43	0.59	0.1	0.84	1									
PB5	0.74	0.7	0.79	0.61	0.67	0.71	0.87	0.11	0.67	0.42	0.52	0.14	0.8	0.87	1								
PNA1	0.81	0.83	0.87	0.47	0.41	0.72	0.87	-0.05	0.51	0.7	0.26	-0.28	0.85	0.71	0.73	1							
PNA2	0.58	0.92	0.79	0.17	0.05	0.53	0.69	-0.25	0.26	0.57	-0.14	-0.61	0.51	0.47	0.53	0.77	1						
PNA3	0.67	0.61	0.77	0.25	0.1	0.59	0.49	-0.33	0.39	0.52	0.2	-0.44	0.51	0.61	0.5	0.55	0.64	1					
PNA4	0.74	0.91	0.84	0.42	0.41	0.62	0.9	-0.01	0.48	0.66	0.14	-0.3	0.76	0.66	0.75	0.93	0.85	0.56	1				
PNA5	0.6	0.36	0.44	0.49	0.56	0.35	0.53	0.27	0.46	0.33	0.34	0.13	0.61	0.55	0.48	0.52	0.22	0.24	0.51	1			
PNA6	0.76	0.84	0.86	0.5	0.42	0.7	0.86	-0.01	0.53	0.6	0.23	-0.26	0.77	0.74	0.78	0.91	0.77	0.6	0.91	0.49	1		
PNA7	0.53	0.21	0.47	0.45	0.54	0.44	0.28	0.02	0.46	0.11	0.42	0.24	0.44	0.46	0.38	0.2	0.02	0.37	0.21	0.45	0.3	1	
PNA8	0.76	0.74	0.85	0.63	0.6	0.83	0.84	0.01	0.64	0.4	0.46	0.01	0.81	0.86	0.9	0.8	0.58	0.52	0.75	0.43	0.81	0.37	1

De acordo com Johnson & Wichern (2007), para validação da análise de correlação canónica se faz necessária uma análise da matriz de covariâncias Σ ou de correlações R a fim de determinar se elas são próximas ou não da matriz nula.

Para validação da aplicação da análise de correlação canónica sobre estes dados, verificou-se que o valor da estatística tem um nível de significância inferior a 5 % para as três primeiras funções, relativamente à análise entre o grupo de variáveis do grupo de produtos alimentares e bebidas.

No que toca aos grupos de produtos alimentares e não alimentares podemos verificar que existem quatro funções com a significância inferior a 5% de acordo a Tabela 2.

Para o grupo de variáveis relativo a bebida e produtos não alimentares podemos verificar que existem quatro funções com o nível de significância inferior a 5%, ou seja, a matriz das covariância dos grupos $x^{(1)}$ e $x^{(2)}$ são estatisticamente diferentes de zero ao nível de 5% de significância, indicando que se pode fazer análise de correlação canónica.

Tabela 2: Testes de hipóteses sobre as correlações

Função	Wilk's	Chi-SQ	DF	Sig.
Produtos alimentares e para bebidas				
1	0.002	307.663	50	0
2	0.047	155.496	36	0
3	0.376	49.878	24	0.001
4	0.723	16.534	14	0.282
5	0.934	3.476	6	0.747
Produtos alimentares e não alimentares				
1	0	430.225	80	0
2	0.012	219.722	63	0
3	0.066	134.853	48	0
4	0.234	71.936	35	0
5	0.539	30.588	24	0.166
6	0.808	10.527	15	0.785
7	0.9	5.213	8	0.735
8	0.986	0.713	3	0.87
Produtos não alimentares e para bebidas				
1	0.005	276.782	40	0
2	0.07	138.279	28	0
3	0.208	81.653	18	0
4	0.497	36.311	10	0
5	0.92	4.342	4	0.362

Correlação entre o grupo de Produtos Alimentares (PA) e os Produtos para Bebidas (PB)

A Tabela 3 refere aos pesos canónicos de três funções, onde a primeira função contém uma correlação canónica 0.974, a segunda com correlação canónica de 0.935 e a terceira com a correlação canónica de 0.693. Podemos ordenar as variáveis em função de sua magnitude. Assim sendo, a variável mais importante do primeiro grupo é PA5 (Óleo e Gorduras), tanto na primeira como na segundo função, enquanto que a variável do segundo grupo a mais importante é PB2 (Gazoza). Relativamente a terceira função o destaque vai para PA3 (Peixe), PB4(vinho). Este tipo de análise não é recomendado, dado que pode haver problema de multicolinearidade, afetando estes resultados.

Analisando as cargas canónicas (loadings), podemos verificar que a primeira função contém altos valores negativos para a maioria dos variáveis. As cargas são otimizadas para a correlação, e não para a interpretação.

Já com as cargas canónicas cruzadas (cross-loadings), podemos verificar que as correlações entre as variáveis do primeiro grupo como PA5 (Óleo e Gorduras) e PA8 (Açúcar), e o segundo grupo de variáveis do segundo grupo PB1 (Agua mineral) e PB2 (Gazoza) são as que apresentam menores correlação na primeira função.
Para essas mesmas variáveis, podemos observar que tem uma forte carga canónica cruzada.
Elevando cargas canónicas cruzadas ao quadrado, podemos calcular quanto da variância destas variáveis pode ser explicado pela primeira função. Note-se PA1 (Pão e cereais), PA7 (Batata, Mandioca e Outros Tubérculos), PB3 (Peixe) tem valores superiores a 80% em cada grupo de variáveis.

Tabela 3:Produtos alimentares e para bebida, pesos canónicos, cargas canónicas, cargas canónicas cruzadas das três funções canónicos

	Pesos canónicos			Cargas canónicas			Cargas canónicas cruzadas		
	1	2	3	1	2	3	1	2	3
PA1	-0.26	0.03	0.40	-0.89	0.14	0.25	-0.87	0.13	0.18
PA2	0.00	-0.72	-1.30	-0.84	-0.34	-0.29	-0.82	-0.31	-0.20
PA3	-0.08	-0.21	1.24	-0.93	-0.09	0.12	-0.91	-0.08	0.08
PA4	0.04	0.23	0.28	-0.53	0.52	0.11	-0.51	0.49	0.08
PA5	0.11	0.72	-0.77	-0.50	0.76	-0.27	-0.49	0.71	-0.19
PA6	-0.29	0.03	0.22	-0.82	0.15	0.29	-0.80	0.14	0.20
PA7	-0.60	0.24	-0.32	-0.93	0.05	-0.29	-0.90	0.05	-0.20
PA8	-0.05	-0.15	0.05	0.06	0.42	-0.28	0.06	0.39	-0.20
PA9	0.03	0.19	0.00	-0.63	0.49	0.04	-0.62	0.46	0.02
PA10	0.00	-0.12	-0.10	-0.63	-0.22	0.06	-0.61	-0.21	0.04
PB1	*0.06*	*0.13*	*0.81*	*-0.39*	*0.70*	*0.38*	*-0.38*	*0.65*	*0.27*
PB2	*0.26*	*0.89*	*-0.53*	*0.18*	*0.96*	*-0.20*	*0.18*	*0.90*	*-0.14*
PB3	*-0.58*	*0.05*	*-0.14*	*-0.93*	*0.26*	*0.09*	*-0.91*	*0.24*	*0.06*
PB4	*-0.05*	*0.34*	*1.26*	*-0.86*	*0.33*	*0.21*	*-0.84*	*0.30*	*0.15*
PB5	*-0.44*	*-0.25*	*-1.55*	*-0.89*	*0.28*	*-0.21*	*-0.86*	*0.26*	*-0.15*

Correlação entre o grupo de Produtos Alimentares (PA) e os Produtos Não Alimentares

A Tabela 4 refere aos pesos canónicos, cargas canónicas, cargas canónicas cruzadas de quatros funções, onde a primeira função contém uma correlação canónica 0.993, a segunda com correlação canónica de 0.906, a terceira com a correlação canónica de 0.848 e a quarta função tem a correlação canónica de 0.753. Ordenado as variáveis em função de sua magnitude, podemos verificar que ma primeira função, a maioria das variáveis tem o peso negativo, tanto no primeiro grupo de variáveis como no segundo grupo, com maior destaque para as variáveis PA9 (Café) e PA10 (Outros produtos alimentares) tem pesos aproximadamente iguais a zero.
Relativamente à segunda função as variáveis PA7(Batata, Mandioca e Outros Tubérculos), PNA4 (Artigos em têxtil) e PNA8 (Serviços domésticos), são as que apresentam o maior peso.
A terceira função as variáveis mais importante PA2 (Carnes), PA7 (Batata, Mandioca e Outros Tubérculos) e PNA4 (Artigos em têxtil) são as que apresentam maior peso tanto no primeiro grupo como no segundo grupo. Este tipo de análise não é recomendado, dado que pode haver problema de multicolinearidade, afetando estes resultados.

Analisando as cargas canónicas (loadings), podemos verificar que a primeira função contém altos valores negativos para à maioria dos variáveis a semelhança da correlação canónica entre grupos de variáveis alimentares e bebidas.

Em relação as cargas canónicas cruzadas (cross-loadings), pode-se verificar que a maioria das funções apresentam valores negativos e as variáveis como PA3 (Peixe) e PNA3 (Mobiliários e Acessórios) apresentam valores negativos simultaneamente nas três primeiras funções.
Elevando cargas canónicas cruzadas ao quadrado, podemos calcular quanto os valores da variância podem ser explicadas na primeira função. As variáveis PA2 (Carnes) tem o valor 83%, PA3 (Peixe) tem o valor de 94%, PNA1 (Vestuário e Reparações), PNA4 (Artigos em têxtil) e PNA6 (Vidraçaria, Louça e Outros) têm valores de 83% respetivamente.

Tabela 4:Produtos alimentares e não alimentares, pesos canónicos, cargas canónicas, cargas canónicas cruzadas das três funções canónicos

	Pesos canônicos				Cargas canónicas				Cargas canónicas cruzadas			
	1	2	3	4	1	2	3	4	1	2	3	4
PA1	0.01	0.32	0.40	1.39	-0.85	0.21	-0.20	0.34	-0.85	0.19	-0.17	0.26
PA2	-0.25	-0.87	1.36	-0.49	-0.92	-0.22	0.28	-0.10	-0.91	-0.20	0.24	-0.07
PA3	-0.42	-0.56	-2.04	-0.12	-0.97	-0.02	-0.16	0.08	-0.97	-0.01	-0.14	0.06
PA4	-0.01	0.18	-0.05	-0.14	-0.51	0.55	-0.11	-0.06	-0.50	0.50	-0.09	-0.04
PA5	0.07	0.53	0.08	-0.27	-0.44	0.76	0.02	-0.09	-0.44	0.69	0.02	-0.07
PA6	-0.24	0.13	-0.25	-0.73	-0.82	0.18	-0.34	-0.08	-0.81	0.17	-0.28	-0.06
PA7	-0.25	0.91	0.60	-0.21	-0.88	0.27	0.30	0.03	-0.88	0.24	0.25	0.02
PA8	-0.01	-0.01	0.20	0.33	0.14	0.51	0.34	-0.03	0.14	0.46	0.29	-0.02
PA9	0.00	-0.10	-0.09	-0.05	-0.57	0.44	-0.16	0.03	-0.57	0.40	-0.13	0.02
PA10	0.00	-0.17	0.24	0.69	-0.61	-0.11	0.11	0.60	-0.61	-0.10	0.09	0.45
PNA1	*-0.29*	*-0.21*	*-1.03*	*1.63*	*-0.92*	*0.07*	*0.17*	*0.24*	*-0.91*	*0.07*	*0.14*	*0.18*
PNA2	*-0.33*	*-1.42*	*-0.01*	*-0.91*	*-0.84*	*-0.47*	*0.26*	*-0.01*	*-0.84*	*-0.43*	*0.22*	*-0.01*
PNA3	*-0.10*	*-0.14*	*-0.62*	*0.67*	*-0.72*	*-0.28*	*-0.38*	*0.24*	*-0.71*	*-0.26*	*-0.32*	*0.18*
PNA4	*-0.07*	*0.97*	*1.74*	*0.19*	*-0.91*	*0.02*	*0.38*	*0.14*	*-0.91*	*0.01*	*0.32*	*0.10*
PNA5	*0.04*	*0.21*	*0.26*	*0.27*	*-0.45*	*0.48*	*0.19*	*0.42*	*-0.44*	*0.43*	*0.16*	*0.31*
PNA6	*0.06*	*-0.05*	*0.13*	*-0.32*	*-0.91*	*0.07*	*0.16*	*0.08*	*-0.91*	*0.06*	*0.14*	*0.06*
PNA7	*-0.16*	*-0.01*	*-0.43*	*-0.07*	*-0.38*	*0.38*	*-0.46*	*0.06*	*-0.38*	*0.34*	*-0.39*	*0.05*
PNA8	*-0.38*	*0.63*	*-0.29*	*-1.33*	*-0.89*	*0.33*	*-0.08*	*-0.24*	*-0.89*	*0.30*	*-0.07*	*-0.18*

Correlação entre o grupo de Produtos Não Alimentares (PNA) e os Produtos para Bebidas (PB)

Na Tabela 5 estão presentes os pesos canónicos, cargas canónicas, cargas canónicas cruzadas relativamente a quatros funções, onde a primeira função contém uma correlação canónica 0.965, a segunda com correlação canónica de 0.815, a terceira com a correlação canónica de 0.763 e a quarta função tem a correlação canónica de 0.678. Ordenando as variáveis em função de sua magnitude, podemos verificar que na primeira função, apenas a variável PB1 (Agua mineral) e PB2 (Gazoza) têm valores positivos e as restantes com o peso negativo. É de salientar que a variável PB4 (Vinho) tem o peso positivo na segunda, terceira e quartas funções.

Analisando as cargas canónicas (loadings), podemos verificar que a primeira função contém altos valores negativos para a maioria dos variáveis a semelhança da correlação canónica entre grupos de variáveis alimentares e bebidas.

A relação as cargas canónicas cruzadas (cross-loadings), a variável PB2 (Gazoza) tem valor negativo referente a primeira função e para a segunda função todas as variáveis relativo a produtos não alimentares tem o valor positivo contra valores negativos de produtos para bebida.

Tabela 5:Produtos não alimentares e para bebidas, pesos canónicos, cargas canónicas, cargas canónicas cruzadas das três funções canónicos

	Pesos canônicos				Cargas canónicas				Cargas canónicas cruzadas			
	1	2	3	4	1	2	3	4	1	2	3	4
PB1	0.05	-0.07	0.44	0.63	-0.38	-0.59	0.48	0.33	-0.36	-0.48	0.37	0.23
PB2	0.30	-0.85	0.12	-0.36	0.21	-0.94	0.23	-0.04	0.20	-0.77	0.17	-0.02
PB3	-0.37	0.32	1.31	-1.35	-0.89	-0.16	0.39	-0.14	-0.86	-0.13	0.30	-0.10
PB4	-0.22	0.09	0.04	1.81	-0.89	-0.28	0.10	0.32	-0.86	-0.22	0.08	0.21
PB5	-0.48	-0.57	-1.50	-0.80	-0.90	-0.40	-0.16	-0.02	-0.87	-0.32	-0.12	-0.01
PNA1	*-0.20*	*1.05*	*2.62*	*-0.50*	*-0.92*	*0.16*	*0.17*	*-0.26*	*-0.89*	*0.13*	*0.13*	*-0.17*
PNA2	*-0.01*	*1.15*	*-0.57*	*-0.01*	*-0.76*	*0.53*	*-0.31*	*-0.19*	*-0.74*	*0.43*	*-0.24*	*-0.13*
PNA3	*-0.22*	*0.25*	*0.16*	*0.96*	*-0.71*	*0.36*	*-0.03*	*0.46*	*-0.68*	*0.29*	*-0.02*	*0.31*
PNA4	*-0.16*	*-1.17*	*-1.02*	*-1.58*	*-0.90*	*0.14*	*-0.11*	*-0.36*	*-0.87*	*0.12*	*-0.08*	*-0.24*
PNA5	*-0.09*	*-0.02*	*0.13*	*0.35*	*-0.54*	*-0.20*	*0.36*	*-0.05*	*-0.52*	*-0.16*	*0.28*	*-0.03*
PNA6	*-0.04*	*-0.02*	*-0.84*	*0.75*	*-0.92*	*0.08*	*-0.08*	*-0.11*	*-0.89*	*0.07*	*-0.06*	*-0.08*
PNA7	*0.01*	*-0.04*	*0.37*	*-0.32*	*-0.36*	*-0.34*	*0.30*	*0.16*	*-0.35*	*-0.28*	*0.23*	*0.11*
PNA8	*-0.46*	*-1.01*	*-0.65*	*0.49*	*-0.93*	*-0.28*	*-0.07*	*0.04*	*-0.89*	*-0.23*	*-0.06*	*0.03*

Análise de redundância

De acordo com Mingoti (2007), a percentagem da variância total explicada por cada variável é dada por:

$$PVTE_{U_k} = \frac{\sum_{i=1}^{p} corr\left(U_k, X_i^{(1)}\right)^2}{p} \times 100 \quad \text{e} \quad PVTE_{V_k} = \frac{\sum_{i=1}^{q} corr\left(V_k, X_i^{(2)}\right)^2}{q} \times 100$$

A proporção da variação total explicada por cada variável canónica está expressa de acordo a Tabela 6. Observamos que a percentagem da variação total explicada a primeira função é de 52,3% para produtos alimentares e 49,1% para a bebidas, referente a análise canónica entre produtos alimentares e bebidas.

Relativamente a análise de correlação canónica entre as variáveis de produtos alimentares e não alimentares a primeira função tem uma variância explicada de 51,4% para os produtos alimentares e 60% para produtos não alimentares.

Para produtos destinados para bebidas, na primeira função tem a variância explicada de 51,7 % contra 56,7% para produtos não alimentares, e as demais pares apresentam valores para variância explicada inferiores para as restantes funções.

Tabela 6 : Proporção da variação total explicada por cada variável

	Proporção da variância explicada de primeiro grupo pela própria variável canónica	Proporção da variância explicada de primeiro grupo pela segunda variável canónica	Proporção da variância explicada pelo segundo grupo pela própria variável canónica	Proporção da variância explicada de segundo grupo pela primeira variável canónica
	Prop	Prop	Prop	Prop
Produtos Alimentares e para Bebidas				
CV1-1	0.523	0.496	0.517	0.491
CV1-2	0.148	0.129	0.332	0.290
CV1-3	0.050	0.024	0.057	0.027
CV1-4	0.031	0.007	0.047	0.011
CV1-5	0.044	0.003	0.047	0.003
Produtos Alimentares e não Alimentares				
CV1-1	0.514	0.507	0.609	0.600
CV1-2	0.155	0.127	0.098	0.081
CV1-3	0.050	0.036	0.083	0.059
CV1-4	0.052	0.029	0.047	0.027
CV1-5	0.029	0.010	0.080	0.027

CV1-6	0.021	0.002	0.027	0.003
CV1-7	0.038	0.003	0.030	0.003
CV1-8	0.057	0.001	0.026	0.000
Produtos para bebidas e não Alimentares				
CV1-1	0.517	0.481	0.609	0.567
CV1-2	0.300	0.199	0.086	0.057
CV1-3	0.094	0.055	0.047	0.027
CV1-4	0.046	0.021	0.060	0.028
CV1-5	0.042	0.003	0.060	0.005

CAPITULO V

CONCLUSÕES FINAIS.

A técnica de correlação canónica pode ser muito útil em problemas que possuam mais de uma variável métrica dependente e quando se pretende estudar correlações simultâneas entre vários conjuntos de variáveis. O uso da correlação canónica pode simplificar o problema e determinar quais variáveis são mais importantes na análise. Desta forma, podemos realizar a análise em duas etapas, primeiro determinando os fatores relevantes, e posteriormente realizando regressões simples entre os mesmos.

Os resultados permitiram verificar quais as correlações fortes em diferentes grupos de variáveis, sabendo que o IPC mede o custo de um conjunto alargado de bens típicos do consumo das famílias. O cálculo da variação do custo desse cabaz de bens traduz a elevação ou não do custo de vida da generalidade de uma população.

Quanto ao grupo de produtos alimentares e bebidas, verificou-se que, das cinco funções, três são significativas inferior a 5%, onde a primeira função contém uma correlação canónica 0.974, a segunda com correlação canónica de 0.935 e a terceira com a correlação canónica de 0.693. Sendo assim, a variável mais importante do primeiro grupo é PA5 (Óleo e Gorduras), tanto na primeira como na segundo função, enquanto que a variável do segundo grupo a mais importante é PB2 (Gazoza). Relativamente a terceira função o destaque vai para PA3 (Peixe), PB4(vinho).

O grupo de produtos não alimentares e o grupo de produtos para bebidas, é composto por oito funções canónicas, dentre elas onde a primeira função contém uma correlação canónica 0.993, a segunda com correlação canónica de 0.906, a terceira com a correlação canónica de 0.848 e a quarta função tem a correlação canónica de 0.753.

Na primeira função, a maioria das variáveis tem o peso negativo, tanto do primeiro grupo de variáveis como no segundo grupo, com maior destaque para as variáveis PA9 (Café) e PA10 (Outros produtos alimentares) tem pesos aproximadamente iguais a zero.

Relativamente à segunda função as variáveis PA7(Batata, Mandioca e Outros Tubérculos), PNA4 (Artigos em têxtil) e PNA8 (Serviços domésticos), são as que apresentam o maior peso.

O grupo de produtos não alimentares e os produtos para bebidas, das cinco funções canónicas apenas uma não é significativa, ou seja as quatros funções, a primeira função contém uma correlação canónica 0.965, a segunda com correlação canónica de 0.815, a terceira com a correlação canónica de 0.763 e a quarta função tem a correlação canónica de 0.678. Apenas a variável PB1 (Agua mineral) e PB2 (Gazoza) têm valores positivos e as restantes com o peso negativo. É de salientar que a variável PB4 (Vinho) tem o peso positivo na segunda, terceira e quartas funções.

É de salientar que durante estes cinco anos, de 2010 a 2014, os produtos que tiveram maior contribuição, foram Peixes, Cervejas, Pão e cereais e os que menos contribuíram foram Gazoza, Açúcar, Agua mineral para inflação em São Tomé e príncipe.

Com este estudo foi possível aprofundar os conhecimentos na técnica de Análise de Correlação Canónica. Permitiu também estudar de forma prática, usando dados recolhidos pelo Instituto Nacional de Estatística de São Tomé e Príncipe, a correlação canónica entre os três grupos de variáveis acima mencionados. Pretendo continuar a aprofundar este estudo, criando o algoritmo computacional para realizar a correlação canónica com três ou mais grupos de variáveis, tomando alguns trabalhos[1][2] que serão uteis para a execução da desta tarefa. Pretendo também analisar os dados com outras técnicas de análise de estatística multivariada, para dar respostas ou tentar a inúmeras situações existentes na minha atividade profissional.

1 "Multiset Canonical Correlation Analysis using specified cost function. Cost functions implemented are {'genvar'}, 'maxvar', 'minvar', 'ssqcor' This version is from an original implementation by Yiou Li in Dec 2007 % based on [Kettenring, Biometrica 1971]. Em http://mlsp.umbc.edu/codes/mcca.m".

2 "Multiset Canonical Correlation Analysis, Jan de Leeuw Version 0.08, December 01, 2015"

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ANDERSON, T. W. An Introduction to Multivariate Statistical Analysis. New York : Wiley., 1958.

[2] BJÖRCH, A. and GOLUB, G. H. Numerical methods for computing angles between linear subspaces. Mathematics of Computation 27, 579-594, 1973.

[3] CARROLL J.D., Generalization of canonical correlation analysis to three or more sets of colonnes, Proceedings of the 76th Annual Convention of the American Psychological Association, P. 227-228, 1968.

[4] DEMPSTER, A. P. Elements of continuous multivariate analysis. Massachusetts : Addison_Wesley, 1969.

[5] GONZALEZ I., Analyse Canonique Régularisée pour des données fortement multidimensionnelles, Institut de Mathématiques de Toulouse UMR CNRS 5219, Thése doctorat de L'université Paul Sabatier Toulouse III, 2007.

[6] HAIR J, ANDERSON R. E., TATHAM R. L., BLACK W. C., Análise Multivariada de Dados, Bookman, 2005.

[7] HARDLE W., SIMAR L., Applied Multivariate Statistical Analysis, Version: 29th April 2003.

[8] HARLOW, L. L. Essence of Multivariate Thinking- Basic Themes and Methods. Taylor & Francis e-Library, 2010.

[9] HARRIS,R.J..A primer of Multivariate Statistics. University of New Mexico 3rd edition, Lawrence Erlbaum associates, publishers Mahwah London, 2001.

[10] HOST P., Relations among m sets of measures, Psychometrika 26(2), p. 129-149, 1961.

[11] HOST P., Generalized canonical correlations and their applications to experimental data, Journal of clinical psychology, Monograph supplement 14, p. 331-347, 1961.

[12] HOTELLING H., Relations between two sets variables, Biometrika, 28, P. 321-377, 1936.

[13] INSTITUTO NACIONAL DE ESTATISTICA, Índice de Preço no Consumidor, Folha de informação Rápida, I TRIMESTRE 2015, ABRIL 2015.

[14] JAMES, M. The generalized inverse form of canonical correlation. Communications in Statistics-Theory and Methods A8, 561-568, 1979.

[15] JOHNSON, R. A., WICHERN D. W. Applied Multivariate Statistical Methods, Prentice Hall, 2005.

[16] KETTENRING J. R., Canonical Analysis of Several sets of Variables, Department of Statistics, University of North Carolina at Chapel Hill, Institute of Statistics Mimeo Series No. 615, 1969.

[17] KETTENRING J. R., Canonical analysis of several sets of variables, Biometrika 58(3), P. 433-451, 1971.

[18] KETTENRING J. R., Relations of Several Sets of Variates, by University of North Carolina Institute of Statistics Mimeo Series No. 538. 1967.

[19] MINGOTI, S. A. Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada. Belo Horizonte: Editora UFMG, 2007.

[20] MULLER, K. E. , "Understanding Canonical Correlation through the General Linear Model and Principal Components," *American Statistician,* **36**, 342–354., 1982.

[21] NZOBOUNSANA V., DHORNE T., Ecart: Une nouvelle méthode d'analyse canonique généralisée, revue statistique appliqué, tome 51, nº4, p.57-82, 2003.

[22] NZOBOUNSANA V., GAYMARD S., Les Analyses Canonique Simple et Généralisée Linéaires: Applications à des Données Psychosociales, Math. & Sci. hum. / Mathematics and Social Sciences (48e année, nº 189,p. 69–101), 2010.

[23] RENCHER A.C., Methods of multivariate analysis, Second Edition, Brigham Young University., 1934.

[24] STEEL R.G.D ,Minimum Generalized Variance for set of linear functions, University of Wisconsin Ann Math Statist, 22, 1951.

[25] STEWART,D and LOVE,W A general canonical index. Psychological, 70,160-163, 1968.

[26] TER BRAAK, C. J. F. Interpreting canonical correlation analysis through biplots of structure correlations and weights. Psychometrika 55, 519-531,1990.

[27] VAN DE VELDEN M., On Generalized Canonical Correlation Analysis, Inst.: Proc. 58th World Statistical Congress, Dublin (Session IPS042), p. 758–765, Econometric Institute, Erasmus University Rotterdam P.O.Box 1738, 3000DR Rotterdam, The Netherlands, 2011.

[28] VAN DER BURG, E. and DE LEEUW, J. Non-linear canonical correlation. British Journal of Mathematical and Statistical Psychology 36, 54-80, 1983.